Deep learning for real-time traffic signal control on urban networks

Junwoo Song

Imperial College London

Centre for Transport Studies
Department of Civil & Environmental Engineering
Imperial College London

A thesis submitted for the degree of Doctor of Philosophy and Diploma of Membership of Imperial College London

This work is dedicated to my god and family.

Declaration of originality

All the work presented in this thesis is that of the author. Any work or contributions made by others is referenced appropriately.

Copyright declaration

Abstract

Real-time traffic signal controls are frequently challenged by (1) uncertain knowledge about the traffic states; (2) need for efficient computation to allow timely decisions; (3) multiple objectives such as traffic delays and vehicle emissions that are difficult to optimize; and (4) idealized assumptions about data completeness and quality that are often made in developing many theoretical signal control models. This thesis addresses these challenges by proposing two real-time signal control frameworks based on deep learning techniques, followed by extensive simulation tests that verifies their effectiveness in view of the aforementioned challenges.

The first method, called the Nonlinear Decision Rule (NDR), defines a nonlinear mapping between network states and signal control parameters to network performances based on prevailing traffic conditions, and such a mapping is optimized via off-line simulation. The NDR is instantiated with two neural networks: feedforward neural network (FFNN) and recurrent neural network (RNN), which have different ways of processing traffic information in the near past. The NDR is implemented and tested within microscopic traffic simulation (S-Paramics) for a real-world network in West Glasgow, where the off-line training of the NDR amounts to a simulation-based optimization procedure aiming to reduce delay, CO_2 and black carbon emissions. Extensive tests are performed to assess the NDR framework, not only in terms of its effectiveness in optimizing different traffic and environmental objectives, but also in relation to local vs. global benefits, trade-off between delay and emissions, impact of sensor locations, and different levels of network saturation.

The second method, called the Advanced Reinforcement Learning (ARL), employs the potential-based reward shaping function using Q-learning and 3^{rd} party advisor to enhance its performance over conventional reinforcement learning. The potential-based reward shaping in this thesis obtains an opinion from the 3^{rd} party advisor when calculating reward. This technique can resolve the problem of sparse reward and slow learning speed. The ARL is tested with a range of existing reinforcement learning methods. The results clearly show that ARL outperforms the other models in almost all the scenarios.

Lastly, this thesis evaluates the impact of information availability and quality on different real-time signal control methods, including the two proposed ones. This is driven by the observation that most responsive signal control models in the literature

tend to make idealized assumptions on the quality and availability of data. This research shows the varying levels of performance deterioration of different signal controllers in the presence of missing data, data noise, and different data types. Such knowledge and insights are crucial for real-world implementation of these signal control methods.

Acknowledgements

Firstly, I would like to thank my supervisor Dr.Ke Han. You are my excellent mentor, researcher and educator. You have treated me with all your heart and given me a lot of motivations for improvement of my academic capability and insight. Your advice gives me invaluable challenges. I look forward to many more year of fruitful collaborations with you in the future.
"韩科教授, 谢谢您指导. 非常感谢!"

I also would like to thank Prof. Washington Yotto Ochieng who have supervised me for PhD period. Without his support and guidance, this PhD work would have not been completed.

I am very grateful to my research group members; Yang Yu, Shiming Xu, Jun Song, Peeranut Jeammaneeporn, Suwan(Krystal) Yin and Jianan Yin. They cheered me up whenever I was having a hard time during my PhD, and I can learn much knowledge by discussing with you. I always pray for you to achieve outstanding academic achievements and finish your PhD very well.
"能一起学习是我的荣幸, 祝你们学业有成."

Moreover, I would like to thank my korean friends at Imperial College London. You always give me life advices. I am also grateful for your help in the college adaptation. I always wish you your PhD researches and academic achievements.

I would like to very thank you my parents, parents-in-law for their unconditional love and support. Without you, none of this would have been possible.

Lastly, to my wife, Meesook. Thank you very much for always supporting me for the PhD period. Without your dedication, sacrifice and love, I wouldn't have made anything for the Phd period. Thank you for always praying to me whenever I am mentally tired and hard. To my son, Sion. Although taking care of you during my PhD has been a great challenge, your bright smile always made me feel good. I hope you grow up healthily and happily. I will do my best for you. I love you so much!
"정말 감사하고 모두 사랑합니다."

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Intelligent Transportation System (ITS) for congestion mitigation and environmental impact

Intelligent Transport Systems (ITS) aim to leverage recent developments in information and communication technologies (ICT) and advanced control & machine learning algorithms to improve the efficiency of traffic operation and reduce congestion externalities such as emission and fuel consumption, as well as accidents in urban traffic networks (D'Acierno et al., 2012). By combining sensing technologies (such as inductive loops, radio frequency identification, automatic number plate recognition, blue tooth, and global navigation satellite systems), ITS can measure and evaluate traffic characteristics or states on dynamic urban traffic networks. In particular, with sensors deployed in urban traffic networks, traffic monitoring systems can collect, transmit, process and fuse heterogeneous and real-time traffic data (e.g. vehicle velocity, road occupancy, and traffic flow). The data collected via these means may be used for the following purposes:

- Road traffic control and management for improved travel time and reliability (Cambridge Systematics, 2005), and alleviated exhaust emissions and their impacts on public health (He et al., 2013),
- optimal dispatch of emergency vehicles such as police, ambulance, and fire truck (Chakraborty et al., 2015);

- Information provision for travel guidance, private trip scheduling, policy appraisal, and data visualization (Smith et al., 2001);
- Information for predicting traffic congestion, traffic accident, and travel demand (Chakraborty et al., 2015).

However, despite the wide development and deployment of ITS infrastructure and technology, traffic congestion remains a major challenge in dense urban areas and produces staggering social, economic, and environmental costs. This is further exacerbated by factors such as traffic accidents, road works, weather, and special events (e.g. strikes, sporting events) (Cambridge Systematics, 2005). Regarding the environmental impact of the traffic congestion, increased number of vehicles in congested areas, as a result of rapid urbanization and motorization especially in developing countries, produce significant amount of exhaust emissions. Barth & Boriboonsomsin (2008) pointed out that, in the United States, exhaust emissions comprise about 33 % of all air pollutants. From an economic perspective, the traffic congestion brings negative economic impacts in the form of lost labour, among others, which attenuate employment growth and worker productivity (Sweet, 2014). Furthermore, in the United Kingdom, traffic-driven environmental issues, investment in traffic infrastructure, and fuel consumption amount to a total of £20 bn per year impediments to the national economy. In the United State, the Texas transportation institute estimated a total loss of $115 bn due to traffic congestion during the year 2009 (Ellis, 2009, Schrank et al., 2010).

1.2 Research scope and objectives

1.2.1 Traffic control and management

Within the context of traffic control and management, different ITS strategies have been developed and tested over the past two decades, which include congestion charging (de Palma & Lindsey, 2011, Evans, 2007), dynamic route guidance (Papageorgiou, 1990, Watling & Van Vuren, 1993), variable message signs (Liu et al., 2016, Peeta & Gedela, 2001, Zuurbier et al., 2006), variable speed limits (Yu & Fan, 2019, Frejo et al., 2019), and traffic signal control (Liu et al., 2015, Liang et al., 2018, Wei et al., 2018). This PhD research focuses on traffic signal control systems since traffic

intersections are known to be the main source of vehicle queuing and travel delays, and a sufficiently optimized traffic signal control tends to reduce vehicle delays and stop-and-go frequencies, and such effects are more pronounced when a number of signalized intersections are centrally controlled and coordinated. Moreover, traffic controls are critical for ensuring the safety of pedestrians and vehicles moving in conflict directions. Furthermore, the utilization of traffic network capacity can be optimized using signal controls with a balance between centralization and decentralization of local traffic controls (e.g. at individual intersections).

1.2.2 Real-time traffic signal control

In the real world, traffic flows are dynamically changing and varying on both within-day and day-to-day scales. In this situation, real-time (responsive) traffic controls should be able to process real-time traffic network states and offer timely decisions. Depending on the control architecture and optimization techniques involved, there are two approaches for real-time traffic signal control: centralized and decentralized. Centralized traffic signal control is coordinated by a single central agent, which, through sophisticated optimization methods by taking into account the global effect of localized control measures and coordination among different local controllers, has the potential to achieve more efficient (sometimes global optimal) traffic control policies. However, given the highly demanding computational requirement and limited data communication and processing capacities (Han, 2017), centralization of responsive signal controls on a network tends to be difficult to implement in large scale in practice (Chow et al., 2019). On the other hand, decentralized system requires multi-agents which control local traffic at interconnected sub-networks. This system can minimize computational efforts but may cause conflicts for traffic signal control policies among local agents (LIU et al., 2017, El-Tantawy et al., 2013, Aziz et al., 2018, Arel et al., 2010). In addition, the decentralized system requires accurate and frequent measurements so that it might be vulnerable to sensor failure (Manolis et al., 2018). In this research, our frameworks focus on centralized traffic signal control for its potential to achieve superior performance, stability and robustness, while investigating techniques to overcome its shortcomings such as computational burden and slow convergence towards optimal policies.

1.2.3 Content of this research

Traffic signal controls play a vital role as traffic intersections are the most stringent bottlenecks of traffic networks, and are the main locations where congestions and queues emerge, which could propagate through links to the whole network. Traffic signal controls are important for the safety of pedestrians, and when properly optimized, reduce traffic congestion and alleviate exhaust emissions and fuel consumptions of vehicles. The main goal of this research is to develop and evaluate a novel real-time road traffic signal control framework in different operational environments with a balanced view of different traffic and environmental objectives to be optimized. The proposed framework has the potential to yield robust and near-optimal real-time signal timing solutions for managing urban traffic in peak times. This is achieved by integrating traffic control theory with machine learning (ML) models, including Feedforward Neural Network (FFNN), Recurrent Neural Network (RNN) and Reinforcement Learning(RL). Figure 1.1 shows that the real-time traffic signal control framework consists of the machine learning module, an optimization model, and a traffic model.

A hallmark of this research is its emphasis on global optimality, online computational efficiency, multiple (including environmental) objectives, and performance with different levels of information availability and quality. Among these four, the last issue has not been addressed in the literature, as most of the proposed real-time signal control strategies (especially based on ML techniques) assume perfect and complete information on traffic flow, occupancy or speed with no error or missing data, and that real-time information are readily available without time delays. In fact, in a real-world operating environment, limited sensing penetration that only covers a portion of the network of interest, sensing error, or transmission delays could compromise the performance of such models developed under these idealised assumptions. This research is the first to systematically and thoroughly investigate the impact of sensor location, information availability, and signal-to-noise ratio of the data on the performance of the signal controls. Using data collected from a real-world traffic network, the proposed traffic control framework is demonstrated to be effective in mitigating traffic congestion, reducing travel time and vehicle emissions, even under some of the aforementioned imperfect operation environment.

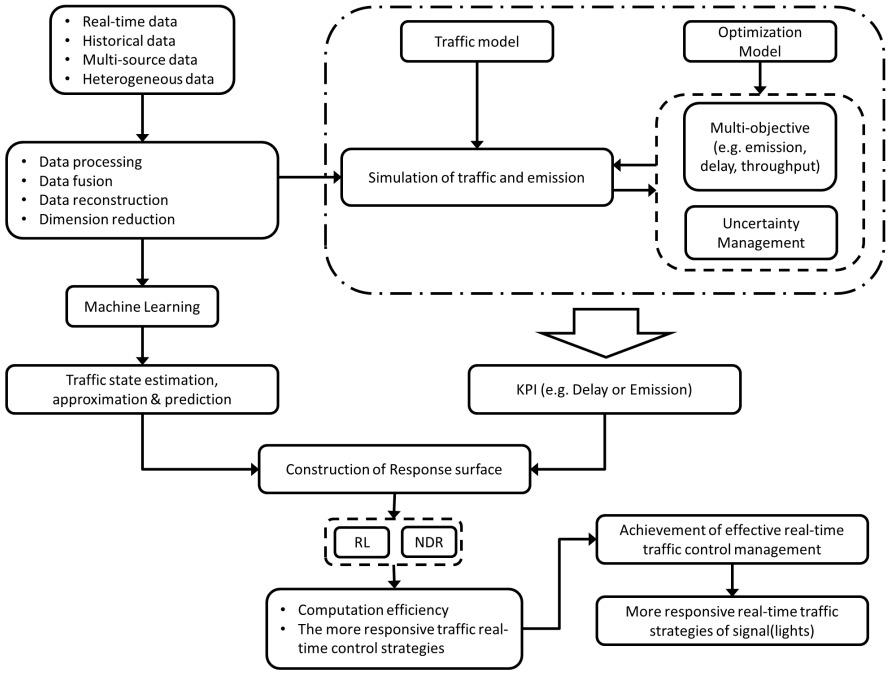


Figure 1.1: Construction of real-time traffic signal control system framework

1.2.4 Real-world data

To test and validate the proposed signal control strategies, this research will apply real-time traffic data collected from the West End of Glasgow, Scotland, where network and traffic data were used to establish and calibrate simulation models. Data collection for our research is performed via the CARBOTRAF project (http://www.carbotraf.eu). These data include dynamic traffic demand, traffic flows at intersections, local pollutant concentration, and vehicle fleet composition (e.g. cars, mini-buses, light goods vehicles, heavy goods vehicles, coaches). Details of the test site will be introduced later in Sections 4.1 and 4.1.1

1.3 Research Challenges

Based on the literature review presented later in Chapter 2, this research has identified four main challenges in developing and deploying an effective real-time signal control framework; see Figure 1.2.

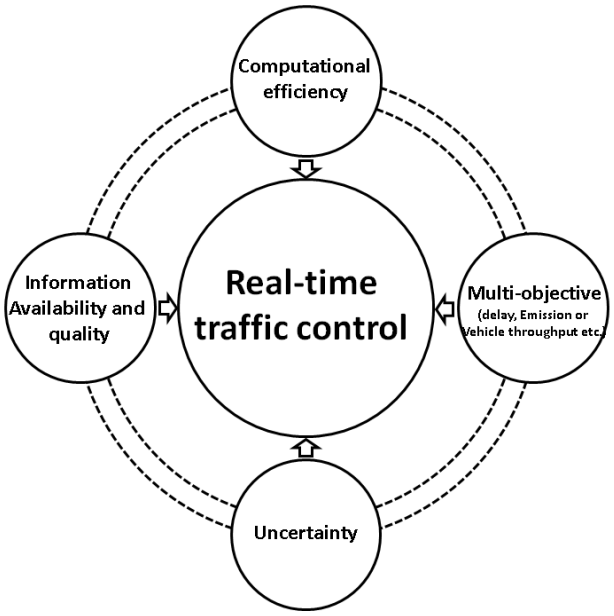


Figure 1.2: Four challenges for development of real-time traffic control framework

1.3.1 Uncertainty of traffic network dynamics

The meaning of traffic uncertainty in the context traffic network awareness and control is two-fold: (1) it may refer to unpredictable variations in traffic states, which is akin to stochasticity; (2) it may also refer to unknown traffic states in part of the network, in which case the control algorithm need to take into account possible realizations of the unknown states. Depending on how the uncertainty set is characterized and analyzed, the effectiveness of the controls can be varied (Pengra & Dillman, 2009). In the field of transportation, traffic flow may vary significantly at road intersections even during the similar times of the day or on the same day of the week (Li et al., 2016b, Song et al., 2017). As a result, it is crucial that the design of the real-time traffic signal control framework allows the handling of uncertain and unexpected traffic flow patterns (Srinivasan et al., 2006, Dong & Chen, 2010). This can ensure satisfactory performance in real-life traffic situations and assist decision makers in constructing feasible solution in uncertain traffic environments.

1.3.2 Multi-objective optimization of network traffic

Specific objectives considered in the literature of traffic control and management include the minimization of (weighted) vehicle/pedestrian delay (He et al., 2014, Sun et al., 2006, Zhang et al., 2010), minimization of passenger delay (Christofa &

Skabardonis, 2011, Christofa et al., 2016), minimization of number of stops (Lucas et al., 2000), maximization of total throughput (Chang & Sun, 2004, Han et al., 2014). Furthermore, there are also numerous studies that incorporate environmental objectives such as emission and fuel consumption.

In many cases, traffic congestion/delay and exhausted emission objectives are not completely aligned with each other, especially when traffic network configuration is non-trivial and traffic dynamics are highly nonlinear. Additionally, striking a balance between the two objectives is very important for effective sustainable traffic management. Thus, in order to keep trade-off between traffic and environmental objectives, multi-objective optimization model for traffic control in a timely fashion will be considered and developed in this research. In addition, the emission of each pollutant has different mechanisms. For example, the CO_2 emission is highly dependent on the engine load and vehicle speed. The emission of Black Carbon tends to increase at low driving speeds as consequences of congestion and stop-and-go episodes. Considering CO_2 and Black Carbon separately, this research investigates their relationships with vehicle delays at intersections and across the entire network. Lastly, the environmental objectives are relatively indirect to mitigate traffic congestion. Therefore, our framework additionally finds the optimal traffic control policy and investigates the potential impact among the traffic objectives by balancing between different traffic objectives.

1.3.3 Computational efficiency

In a realistic and complex traffic environment, decisions should be made in a few seconds to realize real-time and adaptive signal controls. In addition, traffic flow and emission may dynamically vary every time. As shown in Figure 1.1, continuous interpretation of response surface consisting state, control(action) and objective (e.g. KPIs of delay or emissions) is very crucial for real-time traffic signal control. The efficient traffic signal control strategies require the timely and accurate interpretation of the response surface, in order to reduce air pollutants and alleviate traffic congestion.

To tackle the challenge of excessive computational demands in an on-line signal optimization environment, many traffic control architectures (e.g. SCOOT) resort to decentralized controls (Chow et al., 2019, Manolis et al., 2018), which rely on locally

optimized control rules without guaranteeing global optimality; the resulting controls tend to be sub-optimal, and it is difficult to optimize sophisticated objectives such as vehicle delay and emissions in a decentralised manner. However, centralized controls can guarantee global optimality so that it can outperform the decentralized controls by reaching the optimal value (Chow et al., 2019, Manolis et al., 2018). To overcome the computational effort on the centralized controls, Liu et al. (2015) propose offline and online stage to optimize traffic signal control policies and reduce the computational efforts. In the same vein, as the characteristic of machine learning(ML), ML consists of training and testing procedure for learning and evaluating, respectively. Based on training procedure(offline), ML efficiently shows the performance on online procedure (El-Tantawy et al., 2014, Liang et al., 2018, Arel et al., 2010, Liang et al., 2018). Therefore, this research proposes two frameworks: (1) NDR (Nonlinear Decision Rule)-based framework; and (2) ARL (Advanced Reinforcement learning)-based framework.

1.3.4 Information availability and quality

Sensors on road networks generate spatial and/or temporal data with different sparsity, granularity, and reliability. This has a major impact on theoretically derived or tested signal control models. Many existing studies make idealized assumptions on data availability and quality. For example, most previous literatures use real traffic demand or synthetic demand for the experiment (Wiering, 2000, El-Tantawy et al., 2014, Chin et al., 2011, Jin & Ma, 2018, Schultz & Sokolov, 2018, Khamis & Gomaa, 2014, Genders & Razavi, 2016, Gao et al., 2017). But, they did not investigate the effect of the sensor failure on the performance. In particular, some literatures using image-based data from sensors (like cameras) do not consider data incompleteness caused by bad weather (LIU et al., 2017, Mousavi et al., 2017, Liang et al., 2018, Arel et al., 2010, Van der Pol & Oliehoek, 2016, Lin et al., 2018, Mannion et al., 2015, Gao et al., 2017, Genders & Razavi, 2016, Wei et al., 2018). In reality, communication failure (such as data noise and data missing) of the sensor can deteriorate the data quality (Kaisler et al., 2013), causing issues such as missing data, transmission delays, and data noises. Traffic decisions that are made based on the erroneous input by the sensor failure might not be practical or even feasible, and they might result in unsatisfactory performance in real-life traffic situations (Uselton et al., 1998). They

can even have negative effects on computational efficiency (Kaisler et al., 2013).

1.4 Contributions

In view of the four main research challenges mentioned in Section 1.3:

1. Uncertainty in the traffic network;
2. Multi-objective optimization;
3. Computational efficiency; and
4. Data availability and quality

this thesis develops two real-time traffic signal control frameworks for optimizing traffic network performance in terms of vehicle delay, vehicle throughput and reduction of exhausted emissions. The main contributions are articulated as follows.

- This thesis develops a Nonlinear Decision Rule (NDR) approach for real-time signal control based on two types of neural networks: feedforward neural network and recurrent neural network. The NDR allows the optimization of responsive signal control policies/rules to be trained and optimized in an off-line environment, saving the need for real-time optimization and hence addressing challenge **3** by allowing efficient on-line computations.
- This thesis further develops an Advanced Reinforcement Learning (ARL) framework for real-time signal control, by introducing potential-based reward shaping function and a 3^{rd} party advisor. As a model-free approach, Q-learning is employed to reduce computational expenses and generate an immediate and proper traffic signal timings (challenge **3**). A number of state variables and their combinations are considered to target optimization objectives such as average network delay and throughput (challenge **2**). To avoid sparse reward and improve the learning speed, potential-based reward shaping function with the 3^{rd} party advisor is proposed to better handle network uncertainty (challenge **1**).
- The proposed NDR is tested in a microsimulation environment, which is built using real-world data from Glasgow. The training of the NDR framework

considers both vehicle delay and CO_2, black carbon emissions as objectives. The effectiveness of the proposed framework is demonstrated with simulation results, followed by in-depth analysis of the relationship among different objectives (challenge **2**).

- The proposed ARL is tested using hydrodynamic traffic simulations (the Lighthill-Whitham-Richards model) on the Glasgow network. A number of baseline reinforcement learning models are considered in a comparative study in terms of vehicle delay and network throughput, which shows that the ARL has superior performance not only in normal cases (i.e. with good data quality) but also under deteriorated data quality.

- This thesis investigates the impact of data incompleteness and imperfectness on the performance of a range of real-time signal control methods. This is driven by the observation that most responsive signal control models in the literature tend to make idealized assumptions on the quality of data. These assumptions are challenged in this thesis, and we show the varying levels of performance deterioration of different signal controllers (challenge **4**). The impact of sensor locations is also analyzed within the traffic microsimulation of Glasgow. Such findings are not previously seen in the literature, and yield crucial insights regarding the expected level of performance of real-time signal controls in a realistic traffic data collection paradigm.

More findings specific to the test results are mentioned at the end of each relevant chapter, which provide managerial insights regarding managing dynamic traffic networks. The two real-time signal control frameworks put forward in this thesis are derived based on machine learning techniques integrated with an optimization-based training procedure that fully takes into account the possible realization of uncertain traffic states. The resulting real-time controls take full advantage of current information and communication technologies to collect, represent, and process traffic information in order to achieve timely, robust (against uncertainties) and near-optimal decisions for a range of traffic control objectives.

1.5 Thesis structure

In this section, we present the outline of the thesis. The thesis consists of eight chapters which have subsections to support.

Chapter 2 describes the state-of-the-art in real-time adaptive traffic signal control. this chapter introduces the research background and scientific literature with respect to real-time adaptive traffic signal control.

Chapter 3 shows the application of the real-time adaptive traffic signal control framework based on nonlinear decision rule(NDR). With the core knowledge of the framework, this thesis addresses how to efficiently deal with nonlinear traffic dynamics on the proposed framework.

Chapter 4 demonstrates the effectiveness and applicability of the proposed NDR simulation-based optimization framework by applying the proposed framework to real-word traffic network in west Glasgow.

Chapter 5 proposes reinforcement learning-based traffic signal control framework including a novel potential reward shaping function with 3^{rd} party advisor. In addition, in this chapter, the thesis addresses how to apply the ARL to real-life traffic environment, by newly defining three main ARL components(state, action and reward).

Chapter 6 shows the efficient application to real-life traffic network. The benchmark models and our proposed ARL-based framework are tested in equivalent environment and constraints. Therefore, the thesis shows the effectiveness of our proposed model.

Chpater 7 describes the impact of information availability and quality. Based on the experiment settings in chapter 6, the thesis checks the efficiency of both NDR- and ARL- based framework with other benchmark models.

Lastly, Chapter 8 concludes our thesis by completing four challenges; 1) Data availability and quality, 2) Uncertainty of traffic network dynamics, 3) Multi-objective problem and 4) Computational efficiency. In addition, the research contributions and limitations are shown in the Chapter 8. The thesis suggests future works to improve this research.

Chapter 2

State-of-the-art in real-time adaptive traffic signal control

This chapter outlines the previous research on real-time traffic signal control using various state-of-the-art, such as machine learning, mathematical optimisation, simulation-base traffic model as well as traffic signal control with environmental concern.

2.1 Application of machine learning to traffic signal control

2.1.1 The characteristic of machine learning

Machine learning, as a subfield of artificial intelligence techniques, focuses on developing computer-learning algorithms. Representation and generalization are the centrepieces of machine learning. The representation is defined as data assessment, and the generalization is the capability to process unknown and heterogeneous data or parameters. Hence, machine learning can be used 1) to recognize data patterns in order to extract hidden information from enormous amounts of data, 2) to deduce novel information from historical data, 3) to provide the most appropriate solutions, and 4) to recover and understand missing, noisy, or ambiguous data. Machine learning is comprised of three main categories, i.e., supervised learning, unsupervised learning and reinforcement learning. The aim of supervised learning is to detect a

rule and to predict a target attribute with labelled data by directly providing feedback. Supervised learning incorporates classification and regression. On the contrary, although there are no labels and feedbacks, unsupervised learning can explain and summarize the main characteristics of the data by identifying hidden structure. The related techniques include clustering, density estimation, outlier detection and dimension reduction. As the third category, reinforcement learning typically is used to solve learning problem to obtain the maximized reward from specific actions on the given environments. As a typical example, reinforcement learning usually is used in various academic and industrial fields, such as humanoid robotics, management of financial investments and power station control.

2.1.2 Computational issues in machine learning

In the field of transport, the machine learning techniques have been widely used to widen the range of existing research and solve realistic problems. First, the management of continuously increasing traffic data still remains as a critical issue of the analysis and realization of the data, due to the limitation of computer's CPU and memory.

Wibisono et al. (2016) proposed a Fast Incremental Model Trees-Drift Detection (FIMT-DD) algorithm to process time-varying big data generated by 2,500 local sensors deployed in the United Kingdom. The algorithm, based on binary search tree, finds the best split for individual attributes and then split each attribute by standard deviation reduction. The proposed algorithm can efficiently use the computer memory and effectively cope with continuous increment of data over time.

Hoplaros et al. (2014) focused on traffic data summarization technique derived from K-mean clustering, which is a way to minimise the variance in the distance between clusters. The technique defines and simplifies huge traffic raw data. In addition, through minimizing information loss, it assists with not only exact analysis of enormous traffic data, but also fast and effective traffic information processing. The proposed summarisation technique thereby is able to reduce the running time for traffic data analysis.

Guardiola et al. (2014) focused on data dimensionality reduction and the long-term monitoring of traffic flow patterns from day-to-day, in order to find the evolution of traffic patterns over time. As a solution to computational issue, principal compo-

nent analysis (PCA) is employed to reduce the dimensionality of the dataset, which induces and interpret principal components that are independent of each other by analyzing correlation among components.

Smith et al. (2001) constructed a framework to set up and monitor traffic signal plans by using hierarchical clustering and classification and regression trees(CART) which can search for the significant traffic patterns and relationship between a set of traffic data, in order to decide the optimal tree. After training procedure where is computationally expensive to learn historical traffic data, the framework shows computational efficiency by testing the validity of the framework.

Above these researches, some researchers have taken note of other artificial intelligent techniques. Wang et al. (2016) proposed new intelligent transportation system with RFID sensors. To alleviate system load, fuzzy control rules are used. In addition, the fuzzy control query set is updated by genetic algorithm which is able to improve the overall performance of the proposed system. That is, through fuzzy control rule, this research minimize computational issue.

Song et al. (2017) proposed real-time adaptive traffic signal control for trade-off between traffic and environment objectives. This framework employs linear decision rule (LDR) with distributionally robust optimisation(DRO). The implementation of the LDR method consists of two stages, such as off-line module and on-line module. The off-line module is to train the LDR model using historical traffic data, which amounts to a data-driven distributionally robust optimization(DRO) with traffic and environmental objectives. The performance of the LDR method on the on-line module is guaranteed by the off-line module, and the computationally expensive optimization problem is solved by on-line module.

2.1.3 Effective countermeasures of unstable and dynamic traffic data

Machine learning techniques, through prediction analysis, can determine the traffic strategies that can be used to appropriately react to future events and to control traffic flows. Many studies focus on how to efficiently handle unstable and dynamic traffic data with machine learning technique. For example, in the environmental perspective, Yan et al. (2012) suggested a novel environmental monitoring methodology using compressive sensing. As the reconstruction model handling unstable and

dynamic environmental data, the goal of this research is to monitor large-scale environmental symptoms by using a small number of the sensors. The novel methodology, which is based on Bayesian theory, is cost-effective and accurately recovers a high resolution of environmental symptoms (including temperature changes and lever of atmosphere pollution) with under-sampling measurements.

He et al. (2013) proposed a novel artificial neural network model to extract and recognize potential and specific patterns in huge and dynamical traffic data. The proposed model consists of two parts; self-organized feature map (SOFM), GA (Genetic Algorithm)-chaos optimized radial basis function(RBF) neural network. In particular, as an unsupervised learning, SOFM is based on competitive learning in which output neurons compete against each other to be activated (Bullinaria, 2004). The result is that the activated neurons organize themselves. Turning to RBF neural network, the RBF neural network is conceptually similar with K-Nearest Neighbor (k-NN) method. Gaussian activation function is used in the neural network with both the standard deviation(or radius setting) of input data and the average of input data which are provided by the reseach of Hájek (2011). Based on these parts, this research clusters dynamic traffic data and recognises hidden patterns. As a result, the proposed neural network model using chaos optimisation and genetic algorithm performs better compared to other models.

McHugh (2015) illustrated an prediction approach using huge traffic data, weather data and twitter data. This research focused on how to deal with highly volatility of both traffic data and weather data, in order to achieve traffic prediction and accurate analysis for real-time traffic events. In addition, McHugh (2015) mentioned that different area has different characteristics and patterns in traffic and weather data, and each prediction model(including linear regression, bayesian ridge regression and support vector machine) then has different performances according to the each nature of different areas. Therefore, McHugh (2015) applied different prediction models to different areas. Based on this, real-time visualisation of the employed prediction models has been implemented by combining numerous real-time traffic tweets.

Polson & Sokolov (2016) proposed a deep learning architecture to interpret nonlinear spatio-temporal variation in traffic flow. The learning architecture is based on feed-forward neural network(FFNN) with $tanh$ and rectifier function ($f(x) = max(0, x)$). This research tests and evaluates the performance of the proposed ar-

chitecture by comparing with the performance of sparse linear vector sutoregressive (VAR) combining with data pre-filtering techniques(such as median filtering and trend filtering). Therefore, the proposed architecture using FFNN can provide better solutions to predict unexpected traffic events suddenly occurred on traffic roads.

Yang et al. (2016) focuses on nonlinear interactions control between vehicles in congested areas, by using machine learning that captures and counts vehicles on the road from image sensors. The proposed deterministic microscopic model investigates the dependence of accelerations on vehicle velocity, relative velocity and headway. The model is formulated by using nonlinear least-square regression which minimizes the sum of squares of the error and fits a set of observed data with nonlinearity in unknown parameters. After implementing the proposed model, the results show that vehicle headway have strong influence on the dependence of acceleration on vehicle velocity relative velocity.

Kianfar & Edara (2013) analysed similarity among traffic data (including flow, occupancy, and speed) in each flow condition and partitioned the traffic data using three clustering techniques; K-means clustering, general mixture model (GMM) and hierarchical clustering. The fundamental traffic flow diagrams and macroscopic traffic stream models are created by partitioning traffic data. The results indicate that the performance of hierarchical clustering and K-means clustering outperform GMM clustering. Lastly, this research investigates the effect of input variable for the clustering technique. As a result, as input variable(s), using the combination of occupancy and speed, or only speed show the best performance for clustering technique.

2.1.4 Applications to traffic signal control

In recent year, responsive traffic signal control has been important on urban signalized road intersections because traffic signals can directly manage traffic flows and keep balance between centralization and dispersion on the traffic networks. In addition, it can minimize numerous traffic problems, such as long delay of vehicle drivers, traffic accident, exhaust emissions from vehicles on roads and energy consumption (Liang et al., 2018). In particular, recent advancement in articial intelligence, both in theory and computational architecture, has led to the emergence of a number of machine learning (ML) based approaches for traffic signal controls, such as neural

networks (NNs) and reinforcement learning (RL).

Multi-agent approach using NNs has been applied to minimize average vehicular delay time and average stoppage time (Srinivasan et al., 2006), improve the reactivity of traffic control and capacity of traffic network (Castro et al., 2017), alleviate traffic congestion (El-Tantawy et al., 2013), and improve traffic control decision-making (Hauser & Scherer, 2001). On the other hand, RL is used to develop multi-agent traffic control architecture to optimize phase timing (Balaji et al., 2010), reduce queue length and the number of stops (Li et al., 2016a), and minimize the average delay and congestion at intersections (Arel et al., 2010).

Chiu (1992) proposes a distributed architecure using fuzzy logic to control traffic at multiple intersections in a signalised road network consisting of two-way streets. Each intersection independently changes signal traffic parameters (such as cycle time, phase split, and offset) based on traffic data. A set of 40 fuzzy decision rules is used for the adjustment of the cycle time, phase split and offset. The output of the fuzzy decision rules is the proportional level of adjustment to the traffic signal parameters. This research performs the simulation on signalised road network consisting of 9 intersections. In a simulation, the control scope is changed accroding to traffic control strategy which has two strategies consisting of fixed cycle time and fuzzy-based cycle time. Through this simulation, this paper evaluates the performance of both average waiting time and number of vehicle stops according to the definition of the initial cycle time.

For development of adaptive traffic signal control, Castro et al. (2017) employed biologically-inspired neural network. The network is different from artificial neural network because it does not require training stage to achieve a desired control action by exploring biological natures of real neuron to keep improved performances of overall control action in the model. This model has two different types of inhibitions between each layer, such as feed forward and feedback. Through the communication with each neural in each layer, the bi-neural network provides a phase green time in fixed control cycle time. From the experiment, this paper shows that the proposed model has an efficiency of controlling traffic.

Srinivasan et al. (2006) designed multi-agent traffic signal control framework. The framework consists of two models such as Simultaneous perturbation stochastic approximation in fuzzy neural networks (NN) and Hybrid neural network-based

multi-agent system. The goal is to manage the traffic signal control efficiently. So, in order to evaluate the performance of the developed both multi-agent systems, Srinivasan et al. (2006) considered the mean delay and mean stoppage time of the vehicles. Both models were performed by PARAMICS with JAVA script. With three scenarios such as Three-Hour simulations, Six-Hours simulations and Long Extreme simulation with multiple peaks(24 hours), the models are compared with banch-mark model based on Sydney Coordinated Adaptive Traffic System (SCATS). As a result, the longer the simulation time is, the more total mean delay and vehicle mean speed increase in the banch-mark model and SPSA-NN. However, hybrid NN has a superior performance rather than other models and keeps numerical-stability even under extremely long simulation running time. Therefore, this paper contributed to apply the intelligence techniques in real world, with application of neural network technique.

Hauser & Scherer (2001) constructs a procedure in order to develop, implement and monitor traffic signal plans by using both hierarchical clustering analysis and Classification And Regression Trees (CART). In United States, the term of Time-Of-Day (TOD) is widely used to select and implement the traffic plan, which is an effective way to come up with specific time intervals per day. So, in order to design the TOD system, this research identifies the proper intervals for the traffic plan, decides the occurrence of the traffic volume counts and builds the effective plans to apply to each intervals. Hierarchical cluster analysis classifies the similar cases by grouping together. This method is able to be used for identifying TOD intervals, but it is not possible to analyse traffic data pattern. The proposed CART is a prediction model which is able to cover what cluster analysis does not cover. It is able to search for the significant patterns and relationships in a set of traffic data and then decide the optimal tree by using non-stopping rule. Therefore, Hauser & Scherer (2001) suggests the efficient method of how to manage and analyse a large number of data by hierarchical cluster analysis and CART. Through automatically analysing the traffic data sample, this research provides traffic engineers with more useful information and aids them to determine traffic timing plans.

With recent successful application using deep Q-network(DQN) at Atari game Mnih et al. (2013, 2015), RL is more popular among researchers and developers in the various field. The RL is the method of optimising action policy, which aims at maximizing the reward by interacting with the given environments. In particular, in

the field of transportation, many researchers have been finding the most appropriate traffic signal control policy corresponding to real-time traffic environments, which can mitigate traffic congestion. Therefore, the related research shows the efficiency of the responsive traffic signal control by comparing fixed-time traffic signal or different methods (Arel et al., 2010, Aslani et al., 2018a,b, 2017, Balaji et al., 2010, El-Tantawy et al., 2013, Gao et al., 2017, Jin & Ma, 2015, Junchen & Xiaoliang, 2016, Genders & Razavi, 2019).

Balaji et al. (2010) used reinforcement learning to develop multi-agent traffic control architecture to optimise green timing. Based on the proposed architecture, the published paper analysed how overall average travel time was minimised according to the optimised green time. For validation of the proposed architecture, other traffic control systems (including cooperative ensemble (CE), hierarchical multi-agent system (HMS), and actuated control) were compared with the proposed architecture. The results showed that the proposed architecture had better performance for minimising total mean travel time than the other existing systems.

In the same vein, Arel et al. (2010) focused on finding the efficient 8 phase combinations with the regular time interval, which is non-conflict and compatible in multi-intersection network. Reinforcement learning with feedforward neural network(FFNN) is employed to find the efficient traffic signal control strategy which minimizes average delay per vehicle and average cross blocking. Compared with benchmark model(longest-queue-first (LQF) algorithm), Arel et al. (2010) demonstrates the advantage of the proposed model for efficient traffic control management.

Chin et al. (2011) analyzed the stability and robustness of the reinforcement learning by decreasing or increasing traffic demand at specific time interval in a 4-way intersection. The state is vehicle queue length. the defined action distributes 1 second/5 seconds to green time in each phase according to the vehicle queue length. In addition, the actions are rewarded if the additional green time(1 second/5 seconds) is distributed to the phase when the vehicles are in the queue. On the contrary, the actions are penalized when the additional green time(1 second/5 seconds) is unnecessarily distributed when the vehicle queue is not in the corresponding phase. The result shows that the reinforcement learning algorithm efficiently reduce the vehicle queue length when traffic demand is increased in the specific time interval. Even when traffic demand is reverted to original traffic demand after the traffic

demand has been reduced in a specific time interval, the reinforcement learning approach effectively controls the reverted traffic flows and maintains the vehicle queue length. But, it is not clear that this paper did not mention how much increase and decrease the traffic demand numerically.

To efficiently handle non-stationary of traffic environment, Abdoos et al. (2011) developed an reinforcement learning for traffic signal control, which is based on model-free approach(Q-learning). To accurately describe the traffic state, this research creates 24 state which has four approaching links. Each approach is ranked by the vehicle queue length, and then green times is assigned to ranked approaching links. The proposed algorithm is tested on large-scale traffic network including 50 intersections which are four-way junctions. The simulation results demonstrate that the proposed algorithm outperforms the standard reinforcement learning and effectively reduces the vehicle delay.

Li et al. (2016a) focused on how to develop more responsive traffic signal control algorithm. This research proposed the deep reinforcement learning approach to design the efficient traffic signal timing plans with deep neural network(DNN) which learns the Q-function interacting with the given traffic environments and also find the appropriate traffic signal timing strategies(policies) by clearly defining the traffic state and traffic control actions. In addition, as the DNN, the deep stacked auto-encoders(SAE) neural network is employed, which is one of unsupervised learning techniques and reduces the dimensionality of traffic image data by compressing the input value to extract specific features which closely match original input data. The experiment performed by PARAMICS traffic simulation software. By comparing with the base RL model, this research shows the efficiency of traffic signal control on the proposed RL model.

The reinforcement learning algorithm with Q-table has a limitation of state-action space because Q-table has limited size. To overcome this problem, Wiering (2000) focused on communicating traffic signal controllers with cars by using multi-agent reinforcement learning, in order to optimize vehicle driving policy and minimize the cumulative waiting time until when vehicles arrive to their destinations. For evaluation, this paper set 6 traffic signal control strategies; fixed traffic signal controller, random traffic signal controller, TC-1, TC-2 and TC-3. TC-1 uses only a central traffic information without sharing local traffic information to control traffic. TC-2

applies global information to control only the first car approaching at an intersection and uses local information to control from the second car. Lastly, TC-3 considers global information to control all cars. The experimental results show that TC-3 effectually handles the traffic flow and then outperforms other controllers. However, the agents in this research cannot individually coordinate their traffic signal. This can cause global inefficiencies. To solve this problem, Kuyer et al. (2008) uses max-plus algorithm for the explicit coordination of neighboring traffic signals and the cooperative learning. The max-plus algorithm estimates the optimized action by sharing locally optimized information with neighboring agents.

Expand the above researches, Khamis & Gomaa (2014) used multi-objective function with reinforcement learning algorithm proposed by Wiering (2000) since the algorithm proves its superiority and efficiency when Wiering (2000) apply to large-scale traffic network. By using cooperative hybrid exploration which can be interchangeably used ϵ-exploration(low traffic congestion) with softmax exploration(high traffic congestion), the proposed algorithm can efficiently handle the dynamical changes in traffic network.

Van der Pol & Oliehoek (2016) focused on instability of the reinforcement learning for multi-agent. To avoid this, this research proposed coordinated reinforcement learning using transfer planning which calculates heuristic value function $\hat{Q}$ by computing similar tasks with neighbor agents. In particular, the reinforcement learning algorithm is based on Wiering's algorithm (Wiering, 2000) and Kuyer's algorithm (Kuyer et al., 2008) to learn traffic signal control policies for the multi-agent and combine with max-plus coordination algorithm. Through proposed algorithm, Van der Pol & Oliehoek (2016) evaluate the performance according to the number of agent. Therefore, this research conclude that the proposed algorithm efficiently manages the instability caused by the base reinforcement learning.

Liang et al. (2018) focused on how to define the traffic signal's duration corresponding to the traffic situations. To deal with the complicated traffic status, a double dueling deep Q network (3DQN) with prioritized experience replay is proposed in this research. The proposed model is able to deal with the overestimation of Q-value and achieve better convergence performance in training module. In addition, by estimating the advantage of all possible actions at certain state, the model finds and uses the most beneficial action which is able to make better performance(related

to average vehicle waiting time). For evaluation of the proposed model, simulation of urban mobility(SUMO) is simulated in the proposed model. The result shows that the proposed model can efficiently adjust the duration of signal control rather than the fixed time signal control's duration.

Wei et al. (2018) also proposed the reinforcement learning approach to efficiently manage traffic and minimize queue length and vehicle delay by adjusting traffic signal's duration. In particular, unlike the previous research, Wei et al. (2018) focused on two issues; how to exactly describe traffic environment and how to maintain the balanced memory applicable for different traffic situations. To achieve the both, this research simultaneously considers 6 state variables(such as queue length, number of vehicle, updated waiting time of vehicle, vehicle position, current phase status and next phase status). In addition, this research pointed out that traffic on different lanes might be really different and imbalanced so that a memory cannot cover all traffic situation. Thus, Wei et al. (2018) employed the concept of memory palace, which is that training traffic data for different traffic signal control strategies are stored into different memories. According to the traffic status on different lane, the RL can manage traffic more accurately.

El-Tantawy et al. (2013) proposed a coordinated multi-agent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC). The state and reward are defined as queue length and total cumulative delay, respectively. This approach is tested on a large-scale urban network including 59 intersections. The results clearly show that the proposed model efficiently improves traffic status by minimising average queue length, average vehicle delay, average stop time, travel time, average and CO_2 emission and maximising vehicle throughput.

In the same vein, LIU et al. (2017) proposed cooperative reinforcement learning approach for efficient traffic management. Different agents have different traffic control policies. So, a traffic agent learns its own traffic control policies and considers traffic control policies of other agnets as a part of traffic environments. This can cause serious conflict and traffic accidents. So, this research focused on the problems caused by conflicts between multi-agents. To prevent the above problem, only one agent at central intersection is trained in training procedure. The optimal traffic control policy generated from training procedure will be shared with the other agents. Compared to base RL model, this research shows that each agent collaboratively

performs traffic signal control at each intersection.

Teo et al. (2014) designed robust and accurate dynamic traffic algorithm using the Q-learning which is a model-free approach in reinforcement learning. In addition, for state definition, the level of vehicle queue length is categorized to clearly describe traffic state. The action is to adjust the green time duration at each intersection. Here, reward is defined by identifying current greent signal and the current level of vehicles in queue. If there is no vehicle queue and green signal is still running, the proposed algorithm gets the penalty. Therefore, the proposed algorithm minimizes vehicle queue length at each intersection by adjusting green time.

Unlike the existing reinforcement learning approaches, Gao et al. (2017) do not use human-crafted features(such as average vehicle delay and vehicle queue length) for traffic control decision, which might guide to sub-optimal traffic signal control policies. Gao et al. (2017) propose the reinforcement learing approach using raw real-time traffic data(including vehicle speed, vehicle position and traffic signal status). This approach extracts useful traffic state information from the raw data by using convolutional neural network(CNN). The superiority and efficiency of the proposed approach is demonstrated through comparison with longest queue first algorithm and fixed time control algorithm.

Lin et al. (2018) pointed out that previous machine learning approaches requires vast state-action spaces to learn and describe the complicated characteristic of traffic dynamics in all traffic environments, and control traffic in the complex multiple signalized intersections. This can cause low learning speed(convergence speed) which makes the approaches very slowly reach optimal solution in large-scale urban traffic control problem. To solve these problems, Lin et al. (2018) proposed parallel reinforcement learning approach in which all agents are synchronously trained and learn different type of traffic states. In addition, in order to accelerate learning speed, this research uses general advantage estimation(GAE) which can minimize variance of estimating the overall sum of the reward. Therefore, compared to the performances(such as the number of arrival vehicle and average waiting time) of the fixed-time controller and actuated controller, this research shows that the performances of the proposed approach outperforms other benchmark models.

Genders & Razavi (2016) focused on how to improve representation of traffic, as the traffic state space. In order to extract relevant traffic information, convolutional

neural network is employed. The traffic state space consists of three vectors(including presence of vehicle, vehicle speed and the current traffic signal phase). The action is defined as traffic signal phase, and the reward is defined as the variation in the cumulative vehicle delay among actions. Based on these components, reinforcement learning algorithm with experience replay is developed and tested by traffic microsimulator(Simulation of Urban MObility, SUMO). The performance of the proposed algorithm outperforms that of shallow traffic signal control agent. Therefore, Genders & Razavi (2016) conclude that well-reflected traffic state can affect the output of the reinforcement learning.

Mousavi et al. (2017) approach both technical and practical issues of the deep reinforcement learning at the same time. That is, Mousavi et al. (2017) first focused on how to solve instabilities and oscillations of the reinforcement learning during training procedure and how to fully utilize traffic data to describe traffic environment in detail. Second, by achieving these issues, this research shows the improved performance and the efficiency of the proposed algorithm. Through policy gradient approach, the proposed algorithm can smoothly update traffic signal control policy at each epoch by just following the gradient to find the optimal parameters. In addition, the policy gradient approach can provide a guarantee to coverage on the best case and the worst case during the training process. The state is overall traffic status of one intersection generated by image sensors. The action and reward is defined as the green time of two phases(such as North/South and East/West) and the total cumulative delays between two consecutive actions, respectively. To evaluate the performance, the traffic micro-simulator(Simulation of Urban MObility, SUMO) is employed, and the benchmark model is defined as the classical reinforcement learning using shallow neural network(SNN) consisting of one hidden layer. Compared to the benchmark model, this research clearly shows the proposed algorithm has quicker convergence speed(learning speed) and also efficiently handle the traffic at a synthetic traffic intersection.

Touhbi et al. (2017) investigated the feasibility of the reinforcement learning for adaptive traffic signal control. In particular, this research analyze the performance according to the different reward functions separately considering queue length, cumulative delay and vehicle throughput. The state and action are defined as queue length and green time at each phase, respectively. The classical reinforcement learn-

ing is developed and tested on a four-way signalized intersection(used by El-Tantawy et al. (2013)) by using microscopic traffic simulator named Paramics. The result shows that different traffic volume can affect the performance of the reinforcement learning algorithm with different reward function but the developed algorithm still outperform the fixed time traffic signal control policy(Webster).

Genders & Razavi (2018) evaluate three different state representations of reinforcement learning for adaptive traffic signal control. The state definition has great influence on the performance of the Reinforcement learning agent. Three state representations consist of 1) occupancy and speed, 2) queue and density and 3) discrete cell encoding that describes the presence/absence of the vehicle. Action and reward are defined as green traffic phases and change in cumulative delay between previous and current time, respectively. With traffic micro-simulator(simulation of urban mobility, SUMO), the result shows that all reinforcement learning traffic signal control outperforms the actuated traffic signal control and efficiently handles traffic dynamics by reducing vehicle delay and queue length.

Chu et al. (2019) proposed multi-agent advantage actor critic(A2C) to overcome learning difficulty and improve the stability, observability and robustness of the existing reinforcement learning algorithm for traffic signal control. A spatial discount factor is first employed to weaken the influence of local agents. So, the discounted global reward makes the algorithm keep a balance between cooperative control and greedy control, and efficiently estimates the "advantages" of local traffic signal control policies. The action is defined as green time phase. State consists of two vector, such as the cumulative vehicle delay and the total number of vehicle approaching at intersections. The reward is calculated by queue length and cumulative vehicle delay. This research uses fully connected layers and LSTM layers on Deep neural network. Through synthetic traffic grid(5 X 5 traffic grid) and Monaco traffic network, this research evaluates the performance of the proposed model, and the result shows the proposed algorithm is more stable and robust than other benchmark models.

For good traffic representations, Muresan et al. (2019) proposed a novel state space definition on the reinforcement learning algorithm for traffic signal control, which consists of vehicle queue length, traffic signal status, time of data and day of the week. To evaluate the performance, unexpected traffic demand at a specific time interval is set in the experiment. As a result, compared to fixed time and

actuate traffic controller, this method outperforms and even is more stable when unexpected traffic demand is occurred. The result addresses that appropriate state space definition can affect the performance and stability of reinforcement learning algorithm.

In general, reinforcement learning has slow learning speed during training process. To solve this, Mannion et al. (2015) developed adaptive traffic signal control with potential-based advice where agents are advised by using specific state extracted from the traffic environment. To achieve this, potential-based advice(PBA) technique, based on potential-based reward shaping(PBRS), is employed to improve learning speed and improve the performance of the reinforcement learning agent. In addition, like the PBRS, the developers in the PBA can help the RL agent reach to the optimal state by shaping reward function with the discount factor. To evaluate the proposed algorithm, this research checks the performance of the proposed algorithm on three different phase(2, 3, 4 phases) junctions and compares it with classical reinforcement learning algorithm. As a result, this research shows that the proposed algorithm rapidly converges to the optimal output rather than the conventional algorithm and improve the performance. But, according to the the definition of the potential function in reward shaping function, the performance can be improved more. In addition, this research did not apply to real traffic network. So, through applying to real traffic network, the proposed algorithm is validated.

Aziz et al. (2018) developed reinforcement learning algorithm with R-Markov Average Reward Technique(RMART) which is a new temporal-difference method in which value function is defined by averaging the expected reward. In addition, traffic control agents in the proposed learning algorithm can share traffic information with the surrounding agents. The simulation is performed by VISSIM with different random seeds. Compared to four benchmark models(such as fixed-time control, adaptive control, standard Q-learning and standard SARSA), the results from the simulation show that the proposed algorithm reduces vehicle delay and emission(such as CO, CO_2, NO_X, VOC, PM_{10}) and alleviates traffic congestion.

2.2 Application of mathematical optmisation (Math programming) to traffic signal control

2.2.1 The characteristics of math programming

Mathematical optimisation (math programming), which includes linear, convex, and nonlinear optimisation, is one of problem-solving methodologies that aims to improve the decision-making and efficiency of various systems, including transportation systems (Boyd & Vandenberghe, 2004). In general, the mathematical optimisation technique minimises or maximises a real function in a certain problem by calculating input variables of the function with constraints. In the field of transportation, the math programming technique has been used for the minimisation of traffic delay, journey time, and travel cost and for the maximisation of vehicle throughput in traffic networks (Nagurney & Zhang, 2007). Turning to traffic flow theory, the traffic flow theory analyses the relations between infrastructure (including traffic signal controller, loop detector, road, and highway) and travellers (including motorcyclists, vehicle drivers, cyclists, and pedestrians). Its main purpose is to develop optimal feasible traffic network to achieve efficient traffic flow and to alleviate traffic congestion (Garavello et al., 2016). Adjunctively, dynamic traffic assignment (DTA) in the traffic flow theory is a mathematical model to describe the dynamic evolution of traffic networks with travel demands and travel behaviours. Under constraints, such as travel demand and delay, flow propagation, and link dynamics, the DTA minimises travel cost caused by travellers (Friesz, 2010).

2.2.2 Traffic optimisation models

In the real world, traffic flows may vary significantly at road intersections even in the same time period of the day and day of the week. As a result, the capability to handle uncertain and unexpected flow patterns on a network level is crucial in the design of adaptive signal controls. Numerous studies have attempted to design the adaptive signal control algorithm in the past. From an optimization point of view, a well-defined function is required to relate the traffic signal parameters to specific objective being optimized. Specific objectives in the literature include the minimization of (weight) vehicle/pedestrian delay (Chang & Lin, 2000, He et al.,

2014, Sun et al., 2006, Zhang et al., 2010, D'Acierno et al., 2012), minimization of passenger delay (Christofa & Skabardonis, 2011, Christofa et al., 2016), minimization of number of stops (Lucas et al., 2000), maximization of total throughput (Chang & Sun, 2004, Han et al., 2014).

Yin (2008) developed a pre-timed signal control model by aiming to minimize the average delay and maintain sound performance against the worst-case scenario. The robustness and efficiency of the resultant traffic signal timings are tested and validated by a macroscopic Monte-Carlo simulation which provides decision-makers with the range of all possible outcomes and the probability that can be occurred by any action choice. Moreover, based on 'Webster' split algorithm (Webster, 1958), Dong & Chen (2010) developed real-time signal timing model with non-fixed cycle and split. The model targets to minimise total delay time and maximise traffic capacity by determining the optimised parameters with objective value.

However, the above proposed models do not fully explain traffic dynamics. They inefficiently handle the uncertainty and fluctuation of traffic demand with high computational load. To overcome these limitations, Zhang et al. (2010) consider daily variations of the traffic demand in the optimization of pre-timed signal controls, by using a stochastic programming model that is informed by a range of demand scenarios and their corresponding probabilities of occurrence. In addition, Ukkusuri et al. (2010) proposes a robust system optimal signal control model with an embedded cell transmission model(CTM), in order to account for uncertainty of future transportation demand and capture traffic flow dynamics in a traffic network. The CTM evaluates the fundamental diagram of traffic density and flow by approximating a meso-scopic traffic behaviours in the given test network where is used for developing robust traffic signal control plans. The resulting performance illustrates that the proposed model is robust in a complicated traffic environment, and efficiently deals with the traffic uncertainty over the different level of traffic demand.

Li et al. (2016b) proposed a framework combining signal optimization model with simulation method which optimizes the duration of the green time at each phase. The simulation method can not only forecast the future traffic states but also evaluate current traffic condition. In addition, the proposed simulation method can appropriately find the evolution or changes on a traffic network and interact with the proposed optimization model based on simultaneous perturbation stochastic ap-

proximation algorithm(SPSA) which has a iterative property to reach at the optimal value. The experimental result shows that the proposed framework can mitigate traffic congestion and improve average travel velocity on the test-bed network.

He et al. (2014) formulated a request-based mixed-integer linear program (MILP) to accommodate priority-optimal signal timings and to avoid conflict of requests between priority eligible vehicles and pedestrians, with consideration for signal coordination and real-time vehicle actuations. The generated signal timing is responsive to real-time signal control. Other literature proposed real-time adaptive signal phase allocation algorithm to optimise duration and signal phase sequence by using two-level optimisation problem of which two objective functions minimise queue length and total vehicle delay (Feng et al., 2015).

Jiao et al. (2015) considered a pedestrian factor in the adaptive traffic signal control and proposes a signal timing optimization model to minimize the average delay time per person at an intersection. For realistic application, field survey is performed to obtain the information of average passenger load of the vehicle including bus and car. As the factors for evaluation, the average delay time per person and average queue length in certain time interval in each phase and each direction are considered to compare with each performance between the proposed model and the current signal time plan. In order to apply to real traffic signal control system, Dotoli et al. (2003) modified Jiao et al. (2015)'s traffic signal optimization model considering the different types of vehicle, pedestrian movement, the number of vehicle entering or leaving an intermediate link, etc. For evaluation, a case study is performed to evaluate and analyse the performance of the adapted optimization model. In the same vein, Dotoli et al. (2003) show the efficiency of the adapted optimization model to minimize the queue length at a real life intersection. However, existing research papers try to find and suggest efficient ways in order to avoid traffic congestion. Although the existing researches propose a variety of optimisation models, most optimization model has a difficulty to coordinate local signal timing affecting to the whole network. So, Dotoli et al. (2004) focused on synchronization of the local intersections in order to improve traffic flow in the signalised area and proposed a heuristic method to minimise the average number of vehicles per cycles. The proposed framework is based on the macroscopic traffic model developed by Barisone et al. (2002). The results show the efficiency of minimizing average number of vehicle

per cycles.

Lee et al. (2005) developed real-time adaptive traffic signal optimization with genetic algorithm(GA) which evaluates fitness value of the candidate traffic signal plans calculated by GA. With three scenarios using three different levels(high, medium and low) of traffic demand, the proposed algorithm is tested by microsimulation software(PARAMICS) in online testing module and compared with fixed-time signal control plan generated by TRANSYT-7F. The resulting performances indicate that the proposed algorithm efficiently minimizes total vehicle delay in all scenario, compared to the fixed-time signal control plan.

On the network-wide level, Liu et al. (2015) proposed a linear decision rule approach for on-line signal control. The linear decision rule relies on closed-form transformation from the state space to the control space, which is feasible in a real-time decision environment. Such a transformation can be trained via an off-line procedure, which amounts to a distributionally robust optimization.

Papatzikou & Stathopoulos (2015) proposed an optimization model of traffic signalization by combining dynamic traffic assignment with traffic network control. The proposed model is formulated based on conditional Value-at-Risk (CVaR) and simulated by TRANSYT-7F. That research aims to minimize the risk of over-budget travel time from the traffic signal planning properly deployed in the traffic network.

Different from aforementioned literature, Christofa et al. (2016) took into account of person delay in assessing performance of traffic signal setting. A mixed integer linear program is employed and built to minimise person delay with passenger occupancy of vehicles. The proposed real-time signal control system optimises signal setting by solving the mixed integer linear program.

Zhai et al. (2018) focused on signal-stage phase to maximize traffic efficiency and vehicle throughput and minimize average vehicle delay and stops. To achieve this, Zhai et al. (2018) proposed a multi-stage optimal decision framework based on signal-stage optimization. This framework has two signal-stage phases such as primary and secondary compatible phase. Each compatible phase does not conflict with each other since the secondary compatible phase is defined after the primary compatible phase. To verify the performance of the proposed framework, this paper performs the simulation with an intersection and compares the proposed framework with three benchmarks; fixed timing control, segmented timing control and fully

actuated control. In addition, using balanced and unbalanced traffic demands, this paper shows that the framework keep a guarantee of the performance.

Ahmed et al. (2019) emphasized that exact representation of traffic state plays important role on the efficient traffic management. In addition, this research pointed out that the traffic flow prediction based on historical traffic data cannot cover the dynamical changes of traffic. To overcome this, Ahmed et al. (2019) developed a framework using Cell Transmission Model(CTM), Extended Kalman Filter(EKF) and Genentic Algorithm(GA). In the framework, the CTM estimates macroscopic traffic behaviours on a traffic network by evaluating the traffic density and flow at different time stages and optimizes the traffic network. The predicted output are complemented by EKF to gain more accurate estimates of traffic state, which work well for nonlinear state estimation. The optimized traffic signal timing plans generated by CTM-EKF with GA(CTM-EKF-GA). The proposed framework is evaluate in a synthetic intersection. Through estimating vehicle delay, the prediction accuracy of the proposed framework is higher than other benchmark models.

2.2.3 Signal control with environmental objectives

Real-time traffic management with environmental objectives has been a difficult challenge because of (1) the highly dynamic and uncertain nature of road traffic and their emission profile; (2) the need for generating timely and robust decisions for large-scale networks; and (3) the balance between traffic and environmental objectives. To address these challenges, literatures proposes various methodologies. The incorporation of environmental objectives such as emission and fuel consumption has been less widely studied. Therefore, this thesis will explain the relationship traffic with emission.

Li et al. (2004) focused on addressing the relation between average vehicle delay and signal cycle length on a metropolitan network. To achieve this, Li et al. (2004) first formulated a traffic signal timing optimization model to decrease the amount of the exhaust emission, fuel consumption of vehicles and average delay per vehicle. In the opimization model, the objectives are green time and signal cycle length. By optimizing both objectives, this research investigates how much both objectives can affect emissions, fuel consumption and average vehicle delay. In order to address this, a signalized intersection in Nanjing city is used with real data. The experimental

results show that the relationships are close.

Environmental objectives are directly related to the behaviours of vehicle. Rakha et al. (2004) used a microscopic energy and emission model developed by Virginia Tech. This emission model focuses on how much exhaust emission is emitted according to the acceleration and speeds of vehicles. The experimental results show that the relationship between emissions and vehicle behaviours is very close. Moreover, Lefebvre et al. (2011) mentioned that total CO_2 emission is highly dependent on the engine load and vehicle speed. The emission of CO_2 tends to increase at low driving speeds as consequences of congestion and stop-and-go episodes. In addition, as speed slightly increases, the emission reduces, which is due to the vehicle engine working at optimal load, this occurs during the moderate traffic speed. At higher speed, the emission increases significantly again. A similar trend can be observed with other pollutants, such as NOx, PM2.5, however, NOx and PM2.5 are more sensitive to the vehicle dynamics (such as acceleration and idle) and vehicle technology compared with CO_2 (Barth & Boriboonsomsin, 2009, Zhang et al., 2011).

Zhang et al. (2013) formulated a multi-objective optimization model to coordinate traffic signal timings for the minimization of vehicle delay and the reduction of exhaust emission from vehicles. In particular, in environmental perspective, the optimization model takes into account pollutant dispersion affected by weather conditions. To achieve this, cell transmission model is employed to capture the pollutant dispersion and calculate the amount of the roadside air pollutions. In addition, for computational efficiency, Genetic Algorithm(GA) is employed. Lastly, by optimizing green splits, offsets, cycle length and phase sequences, traffic system delay and average exhaust emission are minimized.

Ji et al. (2014) also have developed a method to optimize transit signal priority scheme by alleviating impact on exhaust emission and reducing traffic vehicle delay. However, it finds, in many cases, traffic and emission objectives are not aligned very well with each other, especially when traffic network is complicated and traffic dynamics are nonlinear. Thus, in order to keep trade-off between both objectives, developing a bi-objective optimization model for traffic signal setting has gained popularity.

Chen et al. (2012) mentioned that vehicle emissions are affected by a variety of factors; vehicle type, vehicle operation time and condition (idle speed, acceleration,

and deceleration). So, the instantaneous vehicle emission model based on detailed vehicle dynamics(such as vehicle speed, acceleration and deceleration) is proposed. In addition, by combining the proposed vehicle emission model, this research developed a traffic signal timing optimization model considering both pollutant emissions and average vehicle delay at signalized intersections.

Chang & Hui (2016) develop a traffic emission control model considering signal timing and emission pricing. The model, based on particle swarm optimization, is able to optimize intersection traffic and link-based emissions. Numerical results indicates that the optimization model shows the efficiency of minimizing traffic exhaust emissions. However, due to simplification of the complicated traffic environment, there are limitations to apply to real-world traffic environments.

Most of these aforementioned signal optimization strategies rely on either simplified vehicle dynamics (such as the kinematic wave model) or fleet composition (e.g. single commodity). It is widely known that an accurate depiction of traffic emissions requires extensive knowledge of the detailed vehicle movements, vehicle types, as well as relevant emission factors (Mascia et al., 2017). However, such information is very difficult to obtain especially on a network-wide scale during a real-time operational environment, and most signal optimization algorithms tend to resort to heuristics. In addition, the potential trade-off between traffic performance and environmental impact has not been properly understood in an on-line decision-making context.

The environmental impact of traffic signal control strategies has been investigated and accounted for in a number of recent studies. Han et al. (2016) proposed a MILP approach to optimize signal timings that reduce network congestion as well as vehicle emissions. The MILP is developed using a robust optimization approach based on a macroscopic approximation of the relationship between link dynamics and emission rates. Their study is based on the Lighthill-Whitham-Richards(LWR) kinematic wave model, from which vehicle-derived emissions are calculated. Through robust optimization, the authors are able to reformulate signal optimization problems with emission constraints/objectives as a mixed integer linear program.

Lin et al. (2013) consider vehicle mean speed and the number of vehicle stops to simultaneously reduce vehicle delay and traffic emissions for urban traffic networks by applying model predictive control (MPC) which has the promising capability of efficiently coordinating traffic flow and easily solving multi-objective optimization

problem. In addition, by analyzing individual vehicle at a specific time and location, this research tried to express the traffic situation in mode detail. Similarly, Jamshidnejad et al. (2018) also use the MPC with a gradient-based control optimization approach to smooth vehicle flows, in order to reduce traffic congestion and emissions simultaneously. In addition, in order to secure accuracy and efficient computational time, the traffic flow model 'S-model' developed by Lin et al. (2012) is employed. To describe traffic environment, four groups are categorized according to the different vehicle behaviours and queue positions. The result shows computational efficiency and the balance between vehicle delay and emissions.

2.3 Simulation-based traffic models

Traffic control management has a direct effect on vehicles' travel times, exhaust emissions, and fuel consumption (Spall & Chin, 1997, Chunxiao & Shimamoto, 2011, Rakha et al., 2004, Lefebvre et al., 2011, Chen et al., 2012). The research reported in the literature commonly uses math optimization models, but when applied to reality, it is difficult to check its validity of the outputs. Through simulation model which is a surrogate model to analyze real physical model, many research can check validation and performance of their traffic model (Hirschmann et al., 2010, So et al., 2018, Zhou & Cai, 2014, Osorio et al., 2015, 2017, Stevanovic et al., 2015, Chen et al., 2015, Osorio & Selvam, 2015). In the field of transportation, traffic model (so called simulation model for traffic) is used to help researchers predict traffic flow and patterns on urban traffic networks. There are a variety of simulation software for traffic modeling (including verkehr in stadten-simulations model (VISSIM), and S-paramics (SIAS, 2011)) and mathematical models.

In particular, with increased interest of environmental problem, many researchers have been focusing traffic model with emissions (CO_2, CO, NO, NOx, etc). Through communicating traffic simulation-based model with emission model, such as passenger car and heavy-duty emission model (PHEM), comprehensive modal emission model (CHEM), comprehensive modal emission model (CMEM), and analysis of instantaneous road emissions model (AIRE), many studies are used to estimate exhaust emission for each simulated road vehicle (Scotland, 2011a).

Stevanovic et al. (2009) proposed a simulation-based framework by integrating

VISSIM, CMEM, and VISGAOST. The proposed framework consist of VISSIM for traffic model, CMEM for generating traffic demand, and VISGAOST for signal timing optimization with GA. The traffic model, emission model, and optimization program minimizes vehicular emission (CO_2) and fuel consumption (Diesel), optimizes traffic signal timing, and validates and calibrates surrogate model based on Park city, Utah, respectively. Based on the framework, Stevanovic et al. (2015) additionally analyze the traffic safety by adding surrogate safety assessment model(SSAM). This framework uses 3-dimensional Pareto Fronts of traffic signal timing plans considering safety, mobility and traffic environment which are important factors for traffic operation and are optimized by modifying traffic signal timing plans. Such a framework can keep a balance between mobility, safety, and exhaust emission by communicating between models in the integration method. Similarly, Chen et al. (2015) focused on developing a simulation-based adaptive traffic signal control framework to mitigate traffic congestion by analyzing approximation of the objective function. With traffic simulation model(AIMSUM) and different demand scenarios, the proposed framework is evaluated. The result shows that the proposed framework efficiently handles the traffic flow in the urban traffic network.

Hirschmann et al. (2010) first developed traffic model by using VISSIM with different driving modes considering desired speed and acceleration, and the traffic model is based on a metropolitan arterial road of Graz city. By connecting PHEM with the developed traffic model, this paper calculated emission (including NOx, CO, HC, PM, PN, and NO) and fuel consumption. In addition, So et al. (2018) proposes an integrated simulation-based approach consisting of a traffic model, an emission estimation model and a vehicle dynamic model. The traffic model generates vehicle trajectory information as input data for the vehicle dynamic model. Based on outputs from vehicle dynamic model, this approach estimates and assesses exhaust emission.

Zhou & Cai (2014) developed a multi-objective optimization method based on microscopic traffic simulation at a single intersection. A modal emission and fuel consumption model is used in conjunction with the genetic algorithm to minimize vehicle delay, exhaust emission and fuel consumption at the same time.

Osorio et al. (2015) proposed a meta-model, simulation-based approach to optimize fixed timing for dynamic traffic networks by incorporating dynamic traffic as-

signment models. The response surface methodology is shown to significantly reduce the computational burden typically associated with microscopic traffic and emission models. In order to design more efficient simulation-based optimization(SO) model, Osorio et al. (2017) formulated and used analytical traffic network model which generates more accurate approximation of the link-based travel time.

Song et al. (2017) proposed a real-time adaptive traffic signal control framework based on linear decision rule (LDR), which is integrated with realistic traffic and emission modeling via micro-simulation. The goal of the research is to trade-off between traffic delay and emission. To implement online for testing and off-line for training module, this research used S-Paramics. In the same vein, Zheng et al. (2019) developed a bi-objective stochastic simulation-based optimization model to keep a balance between vehicular exhaust emissions and total vehicle delay under stochastic traffic environment and minimize the differences between real values and simulated objective values by using VISSIM and micro emission model. To evaluate the proposed model, large-scale urban traffic network(Changsha, China) is employed, which consists of 15 traffic intersections with 47 traffic signal phases. The experimental results address that the proposed model efficiently keeps a trade-off between bi-objective values and improve the traffic state more than the existing traffic state.

2.4 Summary

Urban traffic signal controls play an important role for traffic management to solve traffic congestion and diminish adverse environmental impacts. Many researches devised different traffic signal control algorithm, ranging from traditionally pre-timed signal control systems based on historical traffic information to fully responsive systems that frequently update traffic signal control parameters and/or phasing schemes according to real-time traffic conditions.

In many literatures, traffic optimization models are developed and proposed with various methodologies(such as genetic algorithm (Wang et al., 2016, He et al., 2013, Lee et al., 2005, Zhang et al., 2013, Zhou & Cai, 2014), cell transmission model(CTM) (Ukkusuri et al., 2010, Ahmed et al., 2019, Zhang et al., 2013), colony optimization algorithm (D'Acierno et al., 2012) and various simulation-based approach (Hirschmann et al., 2010, So et al., 2018, Zhou & Cai, 2014, Osorio et al., 2015, 2017, Stevanovic

et al., 2015, Chen et al., 2015, Osorio & Selvam, 2015)). However, most traffic optimization models cannot be applied to real traffic signal control system due to the generalization issue. If many constraints are considered in the model, the computation might become more expensive. So, the model cannot cover the heterogeneous data. In addition, the optimization model can offer valuable solutions but do not guarantee global optimality due to the lack of coordination. To solve these problems, the thesis proposes a novel nonlinear decision rule (NDR) approach based on feed-forward neural network and recurrent neural network. The key novelty is that all the expensive computations are performed in an off-line environment through simulation-based optimization based on traffic microsimulation (S-Paramics) and high-fidelity emission modeling using AIRE and COPERT IV models (Mascia et al., 2017). The aim of the off-line optimization is to train the NDR such that its on-line (i.e. real-time) operation can be continuously improved. In addition, the on-line operation of the NDR is computationally efficient as all the optimizations are performed off-line. As we shall see later, some other advantages of this framework include:

- flexible input structure: The system can accommodate a wide range of data types, spatial coverage and temporal resolution. This is a desirable feature for real-time signal control as most existing studies assume full knowledge of traffic states at all key intersections and their approaches, which is often not the case in real-world networks.

- flexible scope and resolution of controls: Different signal parameters (cycle time, green split, offset) at one or several intersections can be controlled simultaneously in real time;

- user defined objectives and priorities: As the training of the NDR is based on simulation, the proposed framework can include various traffic and environmental performance indicators; and

- explicit incorporation of uncertainties: Demand variations and uncertainties inherent in traffic dynamics can be accounted for during the training of the NDR, so that the resulting real-time controls are robust against traffic uncertainties.

Moreover, with recent successful application of reinforcement learning, many researchers have tried to apply machine learning algorithms to traffic systems, in order to overcome the limitations which existing studies have. As the learning-based

algorithm, the machine learning efficiently handles the traffic dynamics, and the generalization issue, which mathematical optimization methods suffer from, would be readily resolved through learning traffic situations. Hence, the machine learning technique, especially reinforcement learning, has been considered as the promising methodology in various industrial fields.

However, in the field of transportation, there are a few things that the researchers might overlook in the reinforcement learning. Here, as the second approach, the thesis proposes learning-based traffic signal control approach with 3^{rd} party advisor. In addition to the limitations of existing literatures, the contribution of the thesis is to be described as follows.

First, most researcher might not address which state variable is more influential to describe traffic environment in detail. This is very important because if less influential state variables are employed more, the machine learning algorithm is diverged and takes much time to reach at the optimal traffic signal control policy. Therefore, this thesis will investigate the impact of state variables.

Second, there might be a few papers considering data incompleteness(e.g., data noise and data missing). In real-life, due to tele-communication error or bad weather, sometimes sensors might not be working or generate erroneous data which cause bad performance on the reinforcement learning algorithm. Therefore, this thesis considers which state variable is more influential for exact description of real time traffic environment. Additionally, the data incompleteness will be considered in the thesis.

Third, as the main objective function in the reinforcement learning, reward function plays a pivotal role on the reinforcement learning algorithm. The reward can effect the learning speed and action policy of the RL algorithm (Ng et al., 1999, Chang, 2006, Mannion, 2017). In addition, the existing literatures consider a range of forms of the reward such as queue length, vehicle delay, relative reduction of total travel delay, total travel time and emergency stop (Aslani et al., 2018b,a, Aziz et al., 2018, Balaji et al., 2010, El-Tantawy et al., 2013, Gao et al., 2017, Jin & Ma, 2015, Lin et al., 2018, Teo et al., 2014, Van der Pol & Oliehoek, 2016, Wei et al., 2018). However, in early stage of reinforcement learning implementation on the existing literatures, state and action are usually unknown and reward can be sparse. This causes the long time training process to reach at optimal traffic control policy (Li

et al., 2016a). Therefore, to get fast convergence with better performance and lead correct convergence, this thesis will extend the basic idea of the reward function from the previous researches and address how to efficiently deal with reward function.

To sum up, the thesis provides two frameworks, such as nonlinear decision rule (NDR) approach and learning-based traffic signal control approach. Through two approaches, the thesis will solve limitations which optimization model and machine learning algorithm have.

Chapter 3

Nonlinear decision rule (NDR) based traffic signal control framework

The nonlinear decision rule (NDR) framework for real-time signal control problem is detailed in this section. In presenting the model we first employ a generic representation without relying on any specific network configuration or control preferences, which highlights the flexibility, computational efficiency and robustness of the proposed method. This is done in Section 3.1. Implementation details of the model pertaining to the case study of this thesis will be presented in Section 3.2. Finally, the off-line training of the NDR based on simulation-based optimization will be detailed in Section 3.3.

3.1 Non-linear decision rule

The decision rule approach is first applied to on-line signal optimization in Liu et al. (2015) to match actual traffic data and optimized traffic signal control strategies. However, linear decision rule(LDR) proposed by Liu et al. (2015) has a limitation of handling traffic dynamics although the LDR outperform existing signal control optimization model in the research of Song et al. (2017). Therefore, this PhD thesis develops and proposes nonlinear decision rule approach.

The dynamics of the traffic network of interest may be perceived by a state vector $\boldsymbol{q} = \boldsymbol{q}(t)$ that changes with time. For example, the vector $\boldsymbol{q}$ may be used to

express traffic quantities such as flow, density, speed, and travel time, which may be measured with different types of sensors (e.g. loop detectors, GPS, and cameras). In addition, this thesis allows $\boldsymbol{q}$ to encapsulate multiple time periods so that the resulting decisions may rely on past memories; see (3.4) for further detail. The NDR stipulates the following form of the control:

$$\mu = \Theta(x, \boldsymbol{q}), \quad u = \mathcal{P}_\Omega[\mu] \tag{3.1}$$

where Θ represents the NDR that maps the states $\boldsymbol{q}$ to the control variables μ; x is the set of parameters of the NDR, which is to be optimized in the off-line training. However, the feasibility of the control μ in a complex control environment cannot be guaranteed by the NDR, and therefore a projection operator $\mathcal{P}_\Omega[\,\cdot\,]$ is employed to further map μ to the feasible control u, where Ω denotes the set of feasible signal control parameters. Ω may be characterized by fixed cycle time, maximum/minimum green time, and signal offsets, all of which may be expressed linearly. In this case, the projection operator $\mathcal{P}_\Omega[\,\cdot\,]$ reduces to a quadratic program (see Section 3.2.3 for details).

A NDR of the form (3.1) can yield timely signal control decisions given inputs regarding current and past network states, which enables real-time operations as it involves analytical or closed-form transformations. The key step in the NDR approach, which directly impacts its on-line performance, is the optimization of the parameters x through off-line training.

We let $\Phi(\boldsymbol{q}, u) = \Phi\Big(\boldsymbol{q}, \mathcal{P}_\Omega[\Theta(x, \boldsymbol{q})]\Big)$ be a given network performance measure, which depends on the system state $\boldsymbol{q}$ and the control u, along with some inherent uncertainties in the traffic system. For example, Φ may be the vehicle delay/vehicle throughput at a particular junction, or the total emission along a certain corridor. Without loss of generality, we assume that Φ is subject to minimization.

The problem of optimal NDR can be formulated as

$$\min_x \Phi\Big(\boldsymbol{q}, \mathcal{P}_\Omega[\Theta(x, \boldsymbol{q})]\Big) \tag{3.2}$$

However, note that $\boldsymbol{q}$ is a stochastic variable that varies on a daily basis. For example, $\boldsymbol{q}$ can be the vector of time-varying demands of an arterial network, which vary from day to day. Therefore, a robust feedback control policy such as (3.2) must take into account the uncertainties in the system. With this in mind, the

off-line training of the decision rule may be formulated as the following stochastic optimization problem:

$$\min_{x} \mathbb{E}\Big[\Phi\left(\boldsymbol{q},\, \mathcal{P}_{\Omega}[\Theta(x,\, \boldsymbol{q})]\right)\Big] \tag{3.3}$$

where the objective is to minimize the expectation of the performance measure with uncertain network states $\boldsymbol{q}$.

3.2 Implementation details

Building on the generic model presented in Section 3.1, this section presents some implementation details pertaining to the case study of the real-world traffic network in Glasgow presented in Section 4.1.

3.2.1 Traffic network state variables

This PhD research begins with the state variable $\boldsymbol{q}$, which captures the network-wide traffic state in terms of different measurements (flow, density, speed, etc.) obtained from a network of sensors (like loop detectors). Given the discrete time step t (t is an integer) with step size δt, we express the state variable as

$$\boldsymbol{q}(t) = \begin{bmatrix} q_1(t-n) & q_1(t-n+1) & \dots & q_1(t-m-1) & q_1(t-m) \\ q_2(t-n) & q_2(t-n+1) & \dots & q_2(t-m-1) & q_2(t-m) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_N(t-n) & q_N(t-n+1) & \dots & q_N(t-m-1) & q_N(t-m) \end{bmatrix} \tag{3.4}$$

where $0 \leq m < n \leq t$ are prescribed integers. On the right hand side of (3.4), each row corresponds to one sensor, and each column represents one single time step. The integer n is used to indicate the number of past time steps considered when making decisions at the current time step t; m is used to account for the fact that data collected in the most recent time intervals may not be immediately available for decision making due to limited capacities of data transmission and computation (Han, 2017).

Remark 1. In the Glasgow case study presented in Section 4.1, the network state is captured by 41 loop detectors, which calculate cross-sectional traffic flows every 2 min (i.e. $N = 41$, $\delta t = 2$ min). The NDR updates signal timing parameters every 10 min based on the flow information collected in the past 10 min; that is, $m = 0$ and $n = 4$.

3.2.2 NDR based on feedforward and recurrent neural networks

This PhD thesis selects feedforward neural network (FFNN) and recurrent neural network (RNN) to instantiate the nonlinear decision rule $\Theta(\,\cdot\,,\,\cdot\,)$. Figure 3.1 illustrates the internal structures of both networks. Given Remark 1, in order to generate traffic control parameters at time step t, both neural networks receive traffic flow vectors in the past 5 consecutive time steps (with a step size of 2 min)

$$\boldsymbol{f}(t),\, \boldsymbol{f}(t-1),\, \boldsymbol{f}(t-2),\, \boldsymbol{f}(t-3),\, \boldsymbol{f}(t-4) \in \mathbb{R}^{41}, \tag{3.5}$$

each being the vector of flows measured at the 41 loop detectors in a 2-min period. (3.5) suggests that the signal control decision made at time t forward depends on the flows in the past five 2-min intervals. To decrease the sensitivity of the neural networks to such input variables, we apply normalization to the vectors $\boldsymbol{f}(t)$, $\boldsymbol{f}(t-1)\ldots,\boldsymbol{f}(t-4)$ before feeding them to the neural networks.

The FFNN has two hidden layers with 100 and 50 neurons, respectively. The fully connected neural network employs the Sigmoid activation function, and the weights of connections among the neurons are treated as the parameters x of the NDR in (3.1), to be optimized in the off-line training. In Figure 3.1(left), the output of the neuron N in the first hidden layer is given as

$$s\left(\sum_{i=0}^{4} \omega_{iN} \boldsymbol{f}(t-i)\right) \tag{3.6}$$

where ω_{iN}'s are the weights, and $s(\,\cdot\,)$ is the Sigmoid activation function. Finally, the output μ for every decision period is generated and used for computing signal control parameters via the projection operator elaborated in Section 3.2.3.

On the other hand, the RNN has one hidden layer with $N = 100$ neurons and one context layer with the same number of neurons as shown in Figure 3.1(right). The RNN reads the vectors $\boldsymbol{f}(t),\ldots,\boldsymbol{f}(t-4)$ from the input layer one by one in a

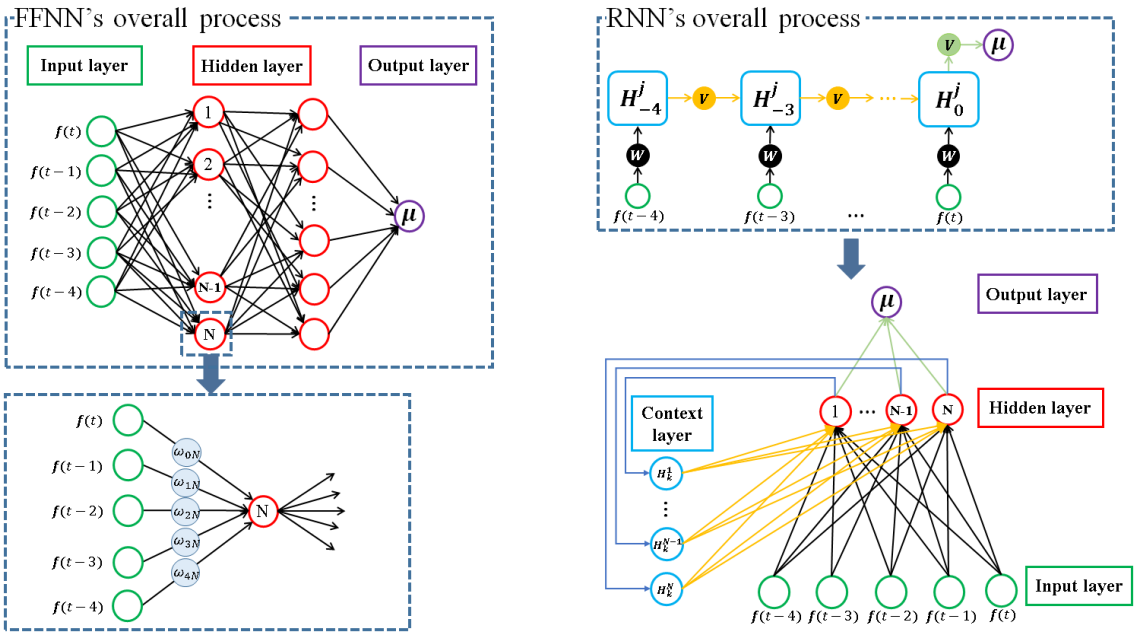


Figure 3.1: Structures of the FFNN (left) and RNN (right).

recursive way:

$$H_k^j = \begin{cases} s\left(w_{jk}\boldsymbol{f}(t+k)\right) & k = -4 \\ s\left(w_{jk}\boldsymbol{f}(t+k) + \sum_{i=1}^{N} v_{ik} H_{k-1}^i\right) & k = -3,\ -2,\ -1,\ 0 \end{cases} \tag{3.7}$$

for $j = 1, \ldots, N$. The RNN iteratively evaluates the quantities $\{H_k^j,\ j = 1, \ldots, N\}$ after reading a flow vector $\boldsymbol{f}(t-k)$.

In comparison with the FFNN, which perceives all the flow vectors $\boldsymbol{f}(t), \ldots, \boldsymbol{f}(t-4)$ at distinct time steps with a symmetric structure as in (3.6), the RNN processes these vectors in sequence following their chronological order. In this way, the RNN is able to capture the temporal dependencies among these state variables through composition of the activation functions.

3.2.3 Projection onto the feasible control set

The signal control parameters typically include cycle time, phasing plans, green times, all-red, and offset (Han & Gayah, 2015). Due to real-world safety considerations reported by (Mascia et al., 2015), the cycle time and phasing plans are fixed in this study. Nevertheless, this PhD research notes that the NDR framework can be easily extended to dynamically change these control variables.

This thesis focuses on the phase green times at each and every intersection, denoted $\boldsymbol{g} = (g_1, g_2, \ldots, g_N)^T$, where N is the number of phases. The green times g_i of all the phases must satisfy the following constraints:

$$g_{\min} \leq g_i \leq g_{\max} \quad \forall i, \qquad \sum_{i=1}^{N} g_i = T_{\text{cycle}} - \Delta \tag{3.8}$$

where $g_{\min}$ and $g_{\max}$ denote minimum and maximum green times, respectively; T_{cycle} is the fixed cycle time, and Δ includes amber (or all-red) time and pedestrian phase time, which are fixed for safety reasons.

Given the green times $\hat{\boldsymbol{g}} = (\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_N)^T$ as output of the neural network $\Theta(x, \boldsymbol{q})$, which do not necessarily satisfy (3.8), the minimum 2-norm projection $\mathcal{P}_\Omega$ onto the feasible set can be formulated as the following quadratic program:

$$\min_{\boldsymbol{g}} \frac{1}{2}\left\|\boldsymbol{g} - \hat{\boldsymbol{g}}\right\|^2 = \frac{1}{2}\left(\boldsymbol{g} - \hat{\boldsymbol{g}}\right)^T\left(\boldsymbol{g} - \hat{\boldsymbol{g}}\right) \tag{3.9}$$

subject to the linear constraints (3.8). Applying the Karush-Kuhn-Tucker conditions (Friesz, 2010) which is the first-order test required to achieve optimal solution in nonlinear programming when regularity conditions are satisfied, we can explicitly express the solution as:

$$\boldsymbol{g} = (g_1, g_2, \ldots, g_N)^T, \quad g_i = \left\{\hat{g}_i - \lambda\right\}_{g_{\min}}^{g_{\max}}$$

where we employ the notation

$$\left\{\hat{g}_i - \lambda\right\}_{g_{\min}}^{g_{\max}} \doteq \begin{cases} g_{\min} & \text{if} \quad \hat{g}_i - \lambda < g_{\min} \\ \hat{g}_i - \lambda & \text{if} \quad g_{\min} \leq \hat{g}_i - \lambda \leq g_{\max} \\ g_{\max} & \text{if} \quad \hat{g}_i - \lambda > g_{\max} \end{cases}$$

and the dual variable λ is such that

$$\sum_{i=1}^{N} \left\{\hat{g}_i - \lambda\right\}_{g_{\min}}^{g_{\max}} = T_{\text{cycle}} - \Delta, \tag{3.10}$$

which can be found by numerically solving the algebraic equation (3.10). Note that in reality the maximum and minimum green times may vary across different signal phases, in which case the formulae above remain valid.

3.3 Off-line optimization of the NDR

This section presents details of the simulation-based optimization procedure, which serves as the off-line module to train and optimize the NDR; i.e. the neural network presented in Section 3.2.2. The main purpose is to find the optimal (or near optimal) solutions of the optimization problem (3.3), which is recapped here:

$$\min_{x} \mathbb{E}\Big[\Phi\left(\boldsymbol{q},\, \mathcal{P}_{\Omega}[\Theta(x,\, \boldsymbol{q})]\right)\Big] \tag{3.11}$$

The inherent stochasticity in the network states $\boldsymbol{q}$ can be handled in different ways such as using robust optimization and stochastic optimization, with varying degrees of conservatism and computational complexity; see Bertsimas et al. (2011), Liu et al. (2015) and Han (2017) for more discussions. In this paper, due to the potentially expensive evaluation procedure, which is done through microscopic traffic and emission simulations, we propose a Monte-Carlo type evaluation method.

Specifically, the overall optimization procedure, which is viewed as the off-line module of the proposed signal control framework, can be divided into two levels; see Figure 3.2. The upper-level problem is to find the optimal parameters x to minimize the expectation shown in (3.11). The objective function involves traffic micro-simulation and high-fidelity emission modeling, whose dynamics and uncertainties are difficult to characterize analytically. Therefore, This thesis employs a heuristic method based on Particle Swarm Optimization method (PSO) to find optimal x. The PSO is chosen here as only $zero^{th}$-order information of the objective and the constraints are required. In addition, although the performance of PSO varies relying on the application or parameters, the research shows evidences of PSO or its variants outperforming other metaheuristic or evolutionary algorithms such as simulated annealing, tabu search, and genetic algorithm (Liu et al., 2015, Yin, 2006, Savsani et al., 2010, Sha & Hsu, 2008). On the other hand, the lower-level problem seeks to evaluate the expected network performance (in terms of traffic and emission indicators) with given parameters x, while taking into account stochasticity in the traffic states $\boldsymbol{q}$ and microscopic traffic dynamics such as driving behavior and route choices.

3.3.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) (Banks et al., 2007) offers an efficient and flexible trade-off between optimality of the solution and computational resources, which is based on the social behaviors in a group of animals, called a *swarm*. In a swarm, the animals are described as particles, and can share and collaborate their own information to adjust their positions in the search for a specific location. Their positions are adjusted by collective memory of the swarm on the best location achieved so far (hereafter referred to as "gbest"), and the individual memory of the best location that the particle has attained so far (hereafter referred to as "pbest"). As a result of the position adjustment, the particles tries to converge to either G or P_j. Although the performance of PSO varies according to the domain of applications or parameters chosen, this PhD research shows evidence of PSO outperforming well-established metaheuristics (e.g. genetic algorithm, simulated annealing, and tabu search) Liu et al. (2015).

Given the objective function to be minimized, denoted $f(\,\cdot\,)$, and the feasible domain S, the following pseudo code summarizes the PSO procedure.

Particle Swarm Optimization

Input. Population size N, $\{\omega_k : k \geq 0\} \subset (0, 1)$, c_1, $c_2 > 0$.

Step 0. Let $k = 0$. Randomly initialize the particles' positions X_i^0 and velocities V_i^0, $1 \leq i \leq N$. Initialize "pbest" P_i^0 and "gbest" G^0 as follows:

$$P_i^0 = X_i^0 \quad 1 \leq i \leq N, \qquad G^0 = P_{i^*}^0$$

where $i^* = \underset{1 \leq i \leq N}{\operatorname{argmin}} f(P_i^0)$.

Step 1. Update the velocities and positions: for all $1 \leq i \leq N$,

$$\begin{aligned} V_i^{k+1} =& \omega_k V_i^k + c_1 r_1 (P_i^k - X_i^k) + c_2 r_2 (G^k - X_i^k) \\ X_i^{k+1} =& \mathcal{P}_S[X_i^k + V_i^{k+1}] \end{aligned}$$

where r_1 and r_2 are random numbers uniformly generated within [0, 1]. ω_k is inertia weight, which can control the impact of previous velocity. c_1 and

c_2 are constants and determine the weights of P_i^k and G^k.

Step 2. Evaluate the objective values $f(X_i^{k+1})$ for all $1 \leq i \leq N$.

Step 3. Update "pbest" and "gbest":

$$P_i^{k+1} = \begin{cases} X_i^{k+1} & \text{if } f(X_i^{k+1}) < f(P_i^k) \\ P_i^k & \text{Otherwise} \end{cases} \qquad \forall 1 \leq i \leq N$$

$$G^{k+1} = \begin{cases} P_{i^*}^{k+1} & \text{if } \min\limits_{1 \leq i \leq N} f(P_i^{k+1}) < f(G^k) \\ G^k & \text{Otherwise} \end{cases}$$

where $i^* = \underset{1 \leq i \leq N}{\operatorname{argmin}} f(P_i^{k+1})$

Step 4. If the stopping criterion is met (e.g. no improvement in the objective within a given number of consecutive iterations), terminate the algorithm with output G^{k+1}. Otherwise, let $k = k + 1$, and go to Step 1.

3.3.2 Off-line training procedure

The off-line training of the NDR amounts to a simulation-based optimization procedure, which requires PSO to be carried out in conjunction with the Monte-Carlo approach that assesses the NDR with given parameters (for FFNN or RNN) via microsimulation and emission calculation. The work flow of the simulation-based optimization is outlined in Figure 3.2, with individual key components explained below.

3.3.2.1 PSO for solving optimization equation (3.3)

The PSO is an agent-based search method, which is detailed in Section 3.3.1. In each iteration of the PSO, a total of N agents, which interact with traffic environment in traffic network, conduct independent search by evaluating, for a given NDR, the corresponding objective value, which is defined as the expectation in Equation (3.3). The stochasticity arises from the microscopic traffic simulation where the departure rates and route choices are randomly sampled based on an origin-destination

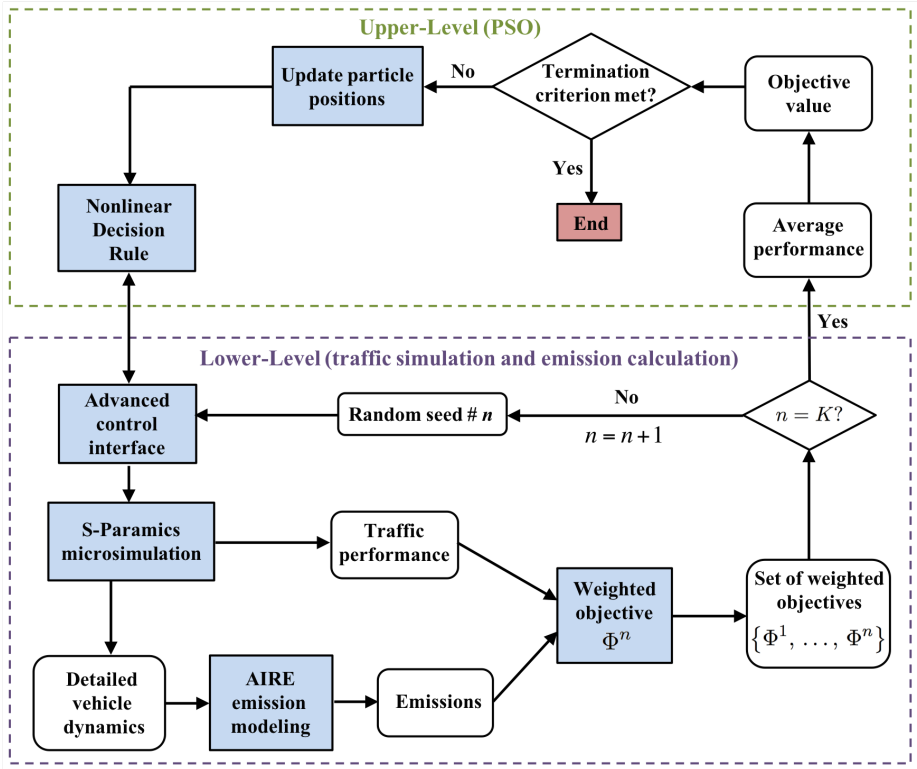


Figure 3.2: Off-line training (optimization) procedure of the nonlinear decision rule.

matrix describing travel demands. Another source of stochasticity comes from the microscopic driving dynamics, which involve car-following, lane-changing, and gap-acceptance behavior. All these stochasticities in each simulation run are populated by a random seed, and this research uses K distinct random seeds to represent the stochastic nature of the traffic states. The aforementioned expectation is then approximated as the average over K independent simulation runs.

In the case study presented in Section 4.1, the PSO employs a population size of $N = 5$, and the algorithm is terminated if no improvement is made on the objective within 20 iterations or when the total iteration number reaches 45.

3.3.2.2 Traffic simulation

Traffic simulation simulate complicated vehicle interactions realistically on a mesoscopic, macroscopic or microscopic perspectives, by modeling traffic demand, supply and vehicle's and pedestrian's behaviour. Based on traffic simulation with a variety of scenarios, researcher and engineers can design, plan and operate traffic system which efficiently improve traffic conditions in real-life. Our research focuses on microscopic

traffic simulation to manage traffic flow. The microscopic traffic simulation is performed using the S-Paramics software (SIAS, 2011), which not only calculates various traffic *key performance indicators* (KPIs) such as travel time, delay and throughput, but also produces detailed vehicle trajectories at a resolution of 0.5 second, which are used as input of emission modeling. For the case study presented in the next section, a microsimulation model is set up for the west end of Glasgow, and calibrated using a combination of macroscopic and microscopic data. See Section 4.1.1 for more details.

The number of traffic simulation (and emission estimation) that needs to be performed within one major PSO iteration is equal to $N \times K$ where N is the population size (independent search agents) and K is the number of random seeds used to populate stochastic parameters and dynamics in the simulation.

3.3.2.3 Emission calculation

A main feature of the proposed real-time signal control framework is the consideration of environmental impact caused by exhaust emissions from vehicles, which is directly impacted by vehicle dynamics and the signal control strategies. In this thesis, we focus on CO_2 and Black Carbon (BC). CO_2 is the primary greenhouse gas and contributes to global warming, while BC causes serious health concerns such as respiratory problems, heart attacks and lung caners.

This PhD research uses the AIRE (Analysis of Instantaneous Road Emissions) vehicle exhaust emission model (Scotland, 2011b) to calculate the instantaneous total carbon and particulate matter emissions resulting from the combustion of fuel throughout each journey of vehicles in the simulation. AIRE is a subsidiary software program that interfaces with S-Paramics and post-processes the output (including car position, vehicle type, direction of vehicle travel, angle of elevation of vehicles, link gradient, vehicle acceleration, vehicle speed, vehicle brake and right and left indicator) of the traffic simulation. Through built-in Instantaneous Emissions Modeling (IEM) tables, AIRE is able to generate estimated value of vehicle emission for each simulated vehicle (Scotland, 2011b).

As AIRE does not calculates CO_2 and BC directly, a post-processing tool developed by CARBOTRAF project is used to convert total carbon and particulate matters into CO_2 and BC emissions (Mascia et al., 2017). Through organic mass (OM) and Elemental Carbon (EC) in the emission inventory, PM is proportionally

assigned. COPERT model also includes emission factors that indicate the proportional assignment($PM_{2.5}/PM_{10}$ ratio) between $PM_{2.5}$ and PM_{10} for different traffic environments. These factors allow the model to estimate the EC fraction of PM_{10} exhaust emissions. For assessment of road vehicle emissions, it is assumed that EC is effectively equal to BC due to characteristics of the combustion processes in different type of vehicles. As a result, this combination of apportioned emission factors and assumptions can estimate BC emissions in our framework. The calculation flow is described in figure 3.3 (Mascia et al., 2017). Therefore, the following procedures are followed to achieve this.

- Based on the PHEM (Passenger Car and Heavy Duty Emission Model) fuel consumption metric, the total carbon metric consequently is able to be directly converted into a representative CO_2 emissions(Mascia et al., 2017). This is done by using the atomic weights of Carbon and Oxygen to generate a factor of 44/12 (one molecule CO_2 weighs 44, one atom carbon weighs 12)(Mascia et al., 2017).
- The calculation of BC is based on the estimated PM_{10} emission rates using the COPERT IV mothod for conversion (Gkatzoflias et al., 2007, Mascia et al., 2017).
- The PM_{10} predicted from the emission model contains emission exhausted from vehicles (Mascia et al., 2017).
- The emission model assumes that all exhaust PM_{10} is approximately equal to $PM_{2.5}$ (Gkatzoflias et al., 2007, Mascia et al., 2017).
- It is also assumed that EC is approximately equal to BC (Ntziachristos & Boulter, 2013, Mascia et al., 2017)

3.3.2.4 Weighted objective

In this thesis, NDR keeps a trade-off between traffic and environment objectives. From an optimization point of view, a well-defined function is required to relate the signal parameters to specific objective being optimized. Specific objectives in the literature include the minimization of (weighted) vehicle delay (He et al., 2014, Zhang

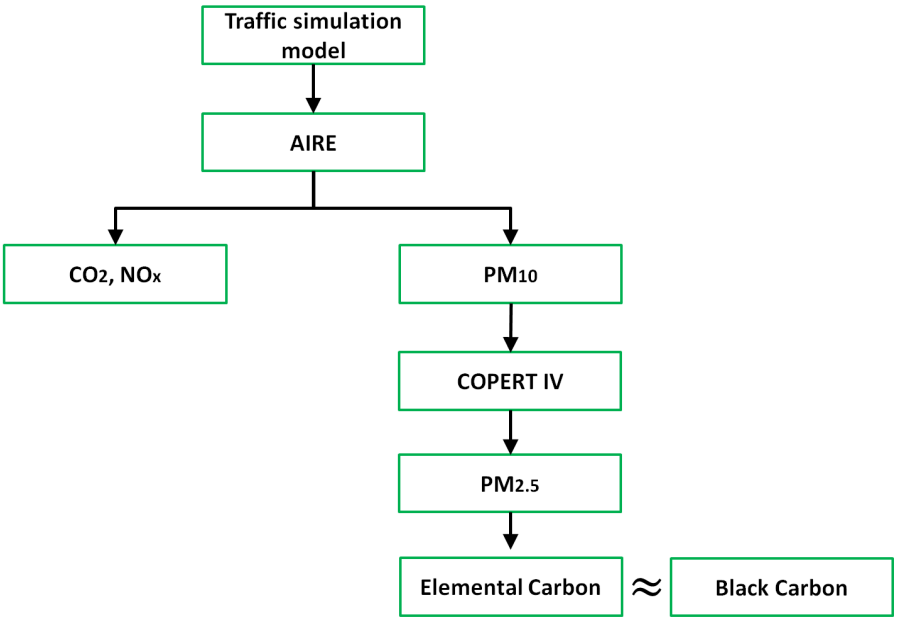


Figure 3.3: Emission procedure (Mascia et al., 2017)

et al., 2010, Sun et al., 2006), minimization of passenger delay (Christofa & Skabardonis, 2011), minimization of number of stop (Lucas et al., 2000), maximization of total throughput (Chang & Sun, 2004, Han et al., 2014). On the other hand, the incorporation of environmental objectives such as emission or fuel consumption has been less widely studied. Han et al. (2016) propose a signal optimization method that takes advantage of a macroscopic relationship between link occupancy and vehicle emission rate. Their research is based on the Lighthill-Whitham-Richards(LWR) kinematic wave model, from which vehicle-derived emissions are calculated. Through robust optimization, the authors are able to reformulate signal optimization problems with emission constraints/objectives as a mixed integer linear program. Ji et al. (2014) also have developed a method in optimising transit signal priority scheme by alleviating impact on exhaust emission and reducing traffic vehicle delay at the same time. However, it finds, in many cases, traffic and emission objectives are not aligned very well with each other, especially when traffic network is complicated and traffic dynamics are nonlinear. Thus, in order to keep trade-off between both objectives, developing a bi-objective optimisation model for traffic signal setting has gained popularity. Stevanovic et al. (2015) proposes a novel integration method in order to solve multi objective traffic signal optimization. The method can keep a balance between mobility, safety, and exhaust emission by communicating between models

in the integration method. Chen et al. (2012) mentioned that vehicle emissions are affected by a variety of factors; vehicle type, vehicle operation time and condition (idle speed, acceleration, and deceleration). So, the instantaneous vehicle emission model based on detailed vehicle dynamics is more appropriate for this kind of study. Most of these aforementioned signal optimization strategies rely on either simplified vehicle dynamics (such as the kinematic wave model) or fleet composition (e.g. single commodity). It is widely known that an accurate depiction of traffic emissions requires extensive knowledge of the detailed vehicle movements, vehicle types, as well as relevant emission factors (Mascia et al., 2017). However, such information is very difficult to obtain especially on a network-wide scale during a real-time operational environment, and most signal optimization algorithms tend to resort to heuristics. In addition, the potential trade-off between traffic performance and environmental impact has not been properly understood in an on-line decision-making context.

However, Song et al. (2017) integrate an on-line signal control framework based on linear decision rule (LDR) with traffic microsimulation and high-resolution emission computation, which serves as a proof of concept for the proposed signal control system to simultaneously reduce traffic congestion and derived vehicle emission in real-time operational environment. This research shows that equivalent value of the weight for multi-objectives efficiently keep a balance between traffic and environmental objective. Therefore, to simultaneously reduce traffic congestion and emissions based on Song et al. (2017), this PhD research reformulates the multi-objective optimization problem into a single-objective one through scalarization:

$$\min \text{Delay}, \quad \text{or} \quad \min\left(w_1 \cdot \frac{\text{Delay}}{n_1} + w_2 \cdot \frac{\text{CO}_2}{n_2}\right) \quad \text{or} \quad \min\left(w_1 \cdot \frac{\text{Delay}}{n_1} + w_3 \cdot \frac{\text{BC}}{n_3}\right) \tag{3.12}$$

where Delay refers to network-wide average delay per vehicle, CO_2 and BC are respectively the network-wide total CO_2 and BC emissions. Constants n_1, n_2 and n_3 are normalization factors to bring the three objective values to a comparable numerical scale. w_i's are positive weights and defined with equivalent value for the balance between traffic and environmental objectives. But, sensitive analysis of the weight w_i is important to analyze the characteristics of emission corresponding to traffic objective. Because our research focus on considering the trade-off between

traffic and environmental objective, this research consider the potential future work and leaves the sensitive analysis according to the weight w_i for near future.

3.3.2.5 Advanced control interface (ACI)

ACI is a method of accessing the traffic model via external program and exchanging information. In S-Paramics, it uses a component protocol called SNMP (Simple Network Management Protocol) to achieve that. Through this protocol, external program can organize, collect, and modify traffic information to change the condition of traffic model (SIAS, 2011). For example, in this paper, the ACI has two main functions; parameter/data exchange and simulation synchronization. First, the program will access real-time (in simulation) traffic data to monitor the performance of the traffic network. This information will be saved and used for the responsive signal optimization procedure. Second, the information will be sent to S-paramics for the synchronization purpose, as the right signal timings need to be implemented in the right time during the simulation. The ACI model has been developed by using visual basic application (VBA), which is an integral part of our experiment set up and facilitates the information exchange and control among different models.

3.4 Summary

In this section, this PhD research develops a real-time signal control framework based on nonlinear decision rule(NDR) to allow actuation of signal timing changes based on network traffic states. The NDR has been implemented with two neural networks: feedforward neural network (FFNN) and recurrent neural network (RNN). Through the NDR, the controller updates traffic signal parameters based on prevailing network states, and the performance of such mechanism can be optimized via off-line traning of the NDR.

Particle swarm optimization is employed to solve the off-line optimization problem, which is the computationally expensive part of the NDR framework, and the on-line implementation of the trained NDR, which only involves analytical and/or closed-form transformation, is quite efficient and can be fully compatible with real-time decision requirements. This is a key advantage of the NDR approach.

The practicality of this approach has been sufficiently demonstrated using mi-

croscopic traffic simulation. The NDR framework uses current/historical traffic data as input of the off-line training phase, which are provided by the microsimulation with different random seeds. The training procedure becomes a simulation-based optimization and is summarized in figure 3.2. A hallmark of this research is the simultaneous consideration of traffic and emission objectives, which can be easily incorporated in the objective function. This highlights a key advantage of the NDR framework over existing signal control methods: the signal control is fully informed by explicit and accurate depiction of emission obtained from microsimulation (AIRE and S-Paramics), and can still maintain high efficiency in an on-line decision making environment.

This real-time (responsive) signal control framework can be readily applied to a real-world environment with the help of a well-calibrated simulation model. In next section, this PhD research demonstrates the applicability and effectiveness of the proposed framework using microscopic traffic simulation and emission modeling based on a real-world traffic network in west Glasgow.

Chapter 4

Application to real traffic network: NDR-based framework

4.1 Case Study in Glasgow

4.1.1 Simulation of the test site

The proposed NDR-based framework has been applied to a real-world test network in Glasgow, Scotland. This PhD thesis employs the traffic simulation model performed by the EU-funded CARBOTRAF project, which aims to support adaptive traffic management for reducing urban congestion and associated environmental and health impacts. The study area is the west part of Glasgow (see Figure 4.1) with 14 signalized junctions and 478 links. There are 21 zones (Figure 4.1(a)), giving rise to 420 origin-destination pairs.

A typical demand scenario for the test network was generated within the S-Paramics simulation software for 7:30-9:30am, which represents morning peak of a typical working day (Monday-Thursday) in 2010; see Figure 4.2(a). The microscopic model has been built using the OS-ITN network to represent the supply, and a seeded demand matrix obtained from loop detectors that represent the within-day and day-to-day variability of traffic (Mascia et al., 2015). For the baseline control scenario, we consider the default traffic signal timing plans provided by the Glasgow City Council. The baseline model has been fully calibrated and validated using a combination of loop detector and floating car data (Mascia et al., 2015).

The vehicle fleet that has been simulated consists of private cars, taxis, buses,

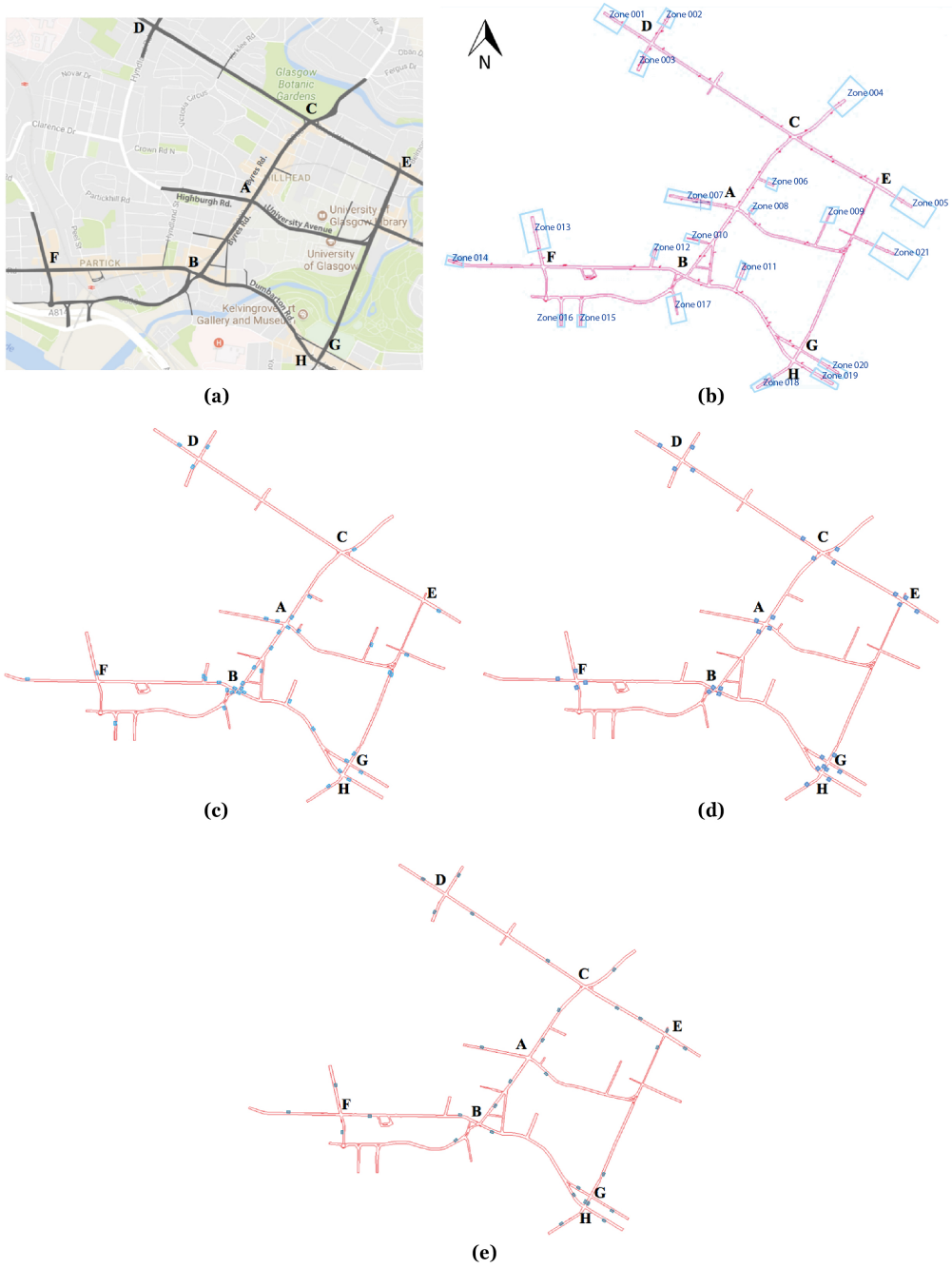
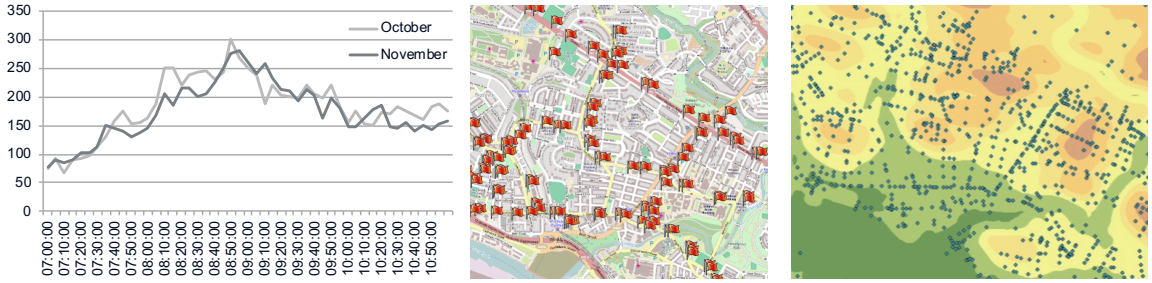


Figure 4.1: (a) The test area in Glasgow with 8 signalized intersections. (b) Road network with 21 Zones. (c) The locations of the 41 loop detectors in the real world. (d) Alternative 1 : First alternative locations of the loop detectors for the comparative study. (e) Alaternative 2 : Second alterative locations of the loop detectors for the comparative study.

vans, light goods vehicles, and heavy goods vehicles. The fleet composition is defined by the Annual Average Daily Flow (AADF) data for the Glasgow city provided by the Department for Transport between 2000 and 2010. This allows us to capture

realistic traffic dynamics with mixed vehicle types and to accurately estimate vehicle emissions with detailed emission factors for different vehicle types. Moreover, road gradient has been explicitly modeled based on the Digital Elevation Model as it has been shown to play a significant role in engine load and, subsequently, carbon emissions (Sobrino et al., 2016). Figure 4.2 shows the bus stops in and around the test network as well as the digital elevation information for the study area.



(a) (b) (c)

Figure 4.2: (a): 5-min average traffic flow (veh/hr) on Byres Road. (b): Bus stops in and around the test network. (b): Network nodes overlaid with Digital Elevation Model.

4.1.2 Signal control details

The network has eight major signal intersections, shown as intersections A-H in Figure 4.1. Other minor junctions in the network are either priority junctions/roundabouts or controlled by actuated signals, which are excluded from our control framework. The cycle times, inter-green (including amber and all-red), and phasing schemes of the eight main intersections are shown in Figure 4.3. These quantities are fixed in our NDR framework per real-world control and safety requirements imposed by the Glasgow City Council (GCC), and parameters subject to real-time optimization are the green times of all the vehicle-movement phases. Note that signal offsets could be easily included as additional decision variables in our control framework, but they are difficult to be adjusted dynamically within the microsimulation. For this reason the offsets are all fixed using the default setting (provided by the GCC) in our control framework.

The resolution for the adaptive signal control is 10 min, which means that the signal timings are adjusted every 10 minutes depending on the real-time traffic conditions. Accordingly, 10-min average traffic flow information collected by the 41 loop

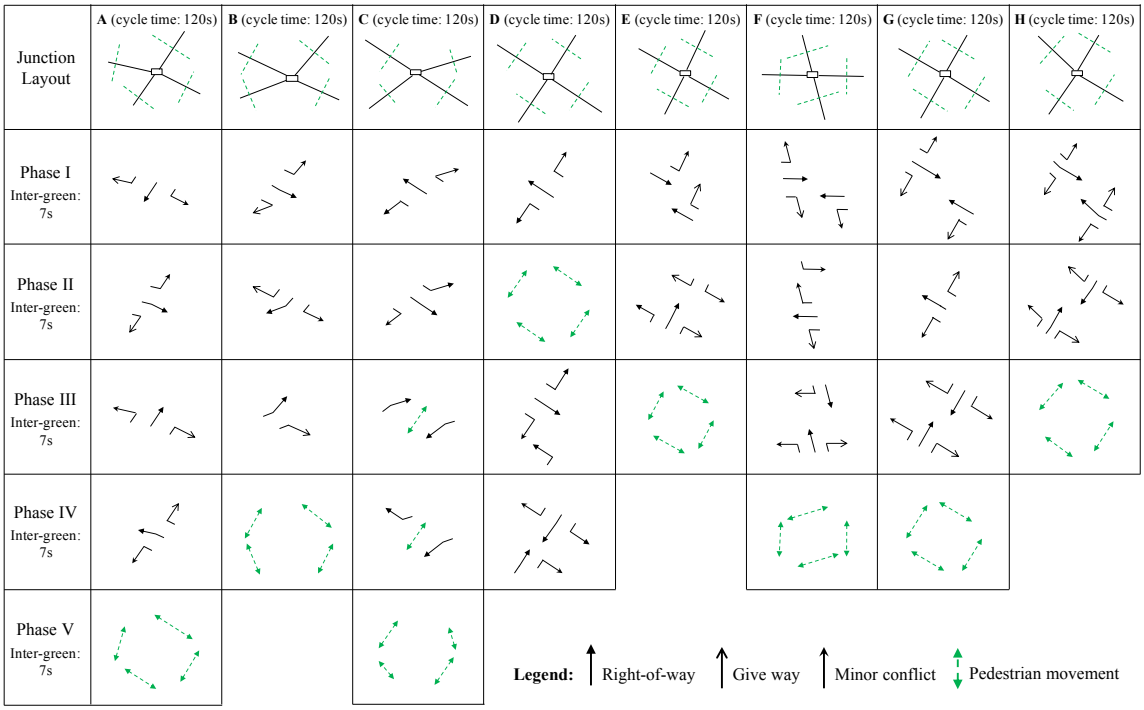


Figure 4.3: Phasing plans of the eight signal intersections.

detectors (see Figure 4.1(b)) is provided to the NDR framework to update signal controls for the next 10 min.

4.1.3 Signal control scenarios

To investigate the extent of traffic and environmental impact of the proposed real-time signal controls, and to make a case for coordinated signal controls on a network-wide level, this thesis considers three test scenarios with varying controllability.

(1) **Junction Level [JL]**: only intersection A is controlled dynamically by the proposed NDR framework; all the other seven intersections are controlled by the default signal timings (provided by the GCC). Intersection A is of critical importance as it connects traffic from the west to major local destinations including universities and hospitals. In the real-world, location A is most affected by traffic congestion and air pollution.

(2) **Corridor Level [CL]**: only intersections A, B and C are dynamically controlled by the proposed NDR framework; the other five intersections are controlled by the default timings. Intersections A-C are located along the Byres

road, which is a strategic corridor connecting the radial routes to the center of city for drivers approaching from the west of Glasgow.

(3) **Network Level [NL]**: all eight junctions in the network are simultaneously and dynamically controlled by the proposed NDR framework. In this way, the signal timings are coordinated in a centralized fashion. In other words, the control of any local intersection is informed by the traffic states and other signal timing plans on the entire network. This is in contrast to distributed controls, which seeks local efficiency over global optimality.

As a benchmark for comparing with the proposed signal control strategy, the thesis considers:

(4) **Glasgow City Council [GCC]**: the fixed signal timing plan provided by the Glasgow City Council, which is derived from static OD route flow information.

In accordance with the aforementioned control scenarios, this PhD research conducts off-line training (optimization) of the NDR by minimizing the average delay, CO_2 emission, BC emission through a weighted combination of these objectives. This allows us to understand the potential trade-off between traffic efficiency and environmental impact. Then, the on-line performance of the optimized NDR is tested in 30 independent simulation runs with 30 random seeds that are different from the ones used in the training. The resulting performance of the traffic network, measured in terms of delay, CO_2 emissions, BC emissions, queuing and throughput, is presented in the following sections.

4.1.4 Test results and discussion

The test results are evaluated against four key performance indicators (KPIs):

- network-wide average delay. The delay is defined as the difference between the actual journey time of a trip minus the free-flow time obtained by assuming little traffic;
- network throughput, defined to be the number of vehicles completing their trips by the end of the simulation period;
- average vehicles in queue, which is defined on the link level; and

- network-wide CO_2 and black carbon emissions.

4.1.4.1 Overall performance of the proposed signal controls

Figure 4.4 shows the average number of vehicles in queue on each link of the network, which is a direct indicator of network congestion. In the case of [GCC], significant congestion is seen along the Byres corridor, especially on the northern entrance. For the proposed methods, widening the scope of the signal controls ([JL] to [CL] to [NL]) tend to mitigate the congestion on the network level overall. However, minor spatial trade-offs of congestion can be seen, for example, between [CL] and [JL]. Through a coordinated control of the three intersections A, B and C, [CL] effectively reduces the congestion on the Byres corridor, especially on the northern entrance (including the Great Western Rd.) compared to [JL]. However, more significant queueing on the southeast part of the network results from the [CL], possibly due to (1) lack of direct control of that area; and (2) increased traffic flow on the Dumbarton Rd. as a result of improved Byres corridor. Such trade-off of congestion at different parts of the network reveals the complexity of network-wide adaptive signal control as drivers' route choices are affected by real-time traffic conditions (Han et al., 2015). Finally, [NL] eliminates all the major queuing on the network and achieves the highest efficiency in terms of vehicle queues. Nevertheless, even in this case some queuing still remains along the northern corridor (Great Western Rd.); this is due to the lack of sufficient sensor coverage along this main corridor; see Figure 4.1(b), and the proposed signal controls are not fully informed by the traffic states there.

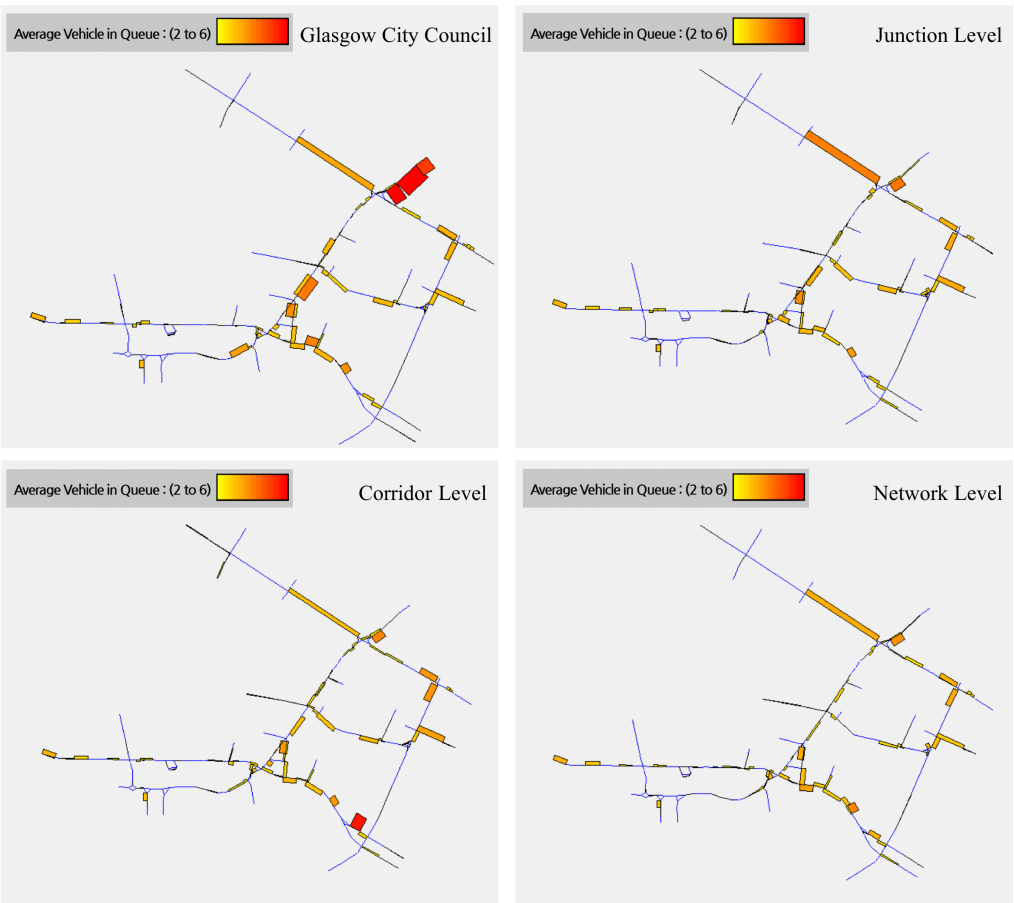


Figure 4.4: Average number of vehicles in queue.

Figure 4.5 shows the performances of the four control scenarios (GCC, JL, CL, NL) based on FFNN and RNN, in terms of delay, throughput, total carbon and black carbon emissions, followed by Table 4.1 summarizing the average improvements of the proposed signal controls over the baseline scenario (GCC). It can be seen that the proposed signal control methods significantly outperform the existing signal control (GCC). Among all four KPIs generated by multi-objectives vehicle delay has the most drastic improvement, from around 28 seconds per vehicle to below 10 seconds (NL). This is followed by CO_2 emission and network throughput, with up to 73 kg reduction and 74 veh increase, respectively. The decrease in CO_2 is likely caused by increased travel speeds as a result of reduced congestion, as CO_2 emissions tend to increase at low driving speeds (Lefebvre et al., 2011). The decrease of black carbon is comparatively less significant with 0.5-1.4 g reduction. Black carbon forms during incomplete combustion of carbonaceous fuels, and is primarily caused by sudden acceleration and brake of vehicle movements (Mascia et al., 2017). Therefore, the

reduction of BC is more significant at local intersections than on the network level (see Figure 4.7). It is also clear from Figure 4.5 and Table 4.1 that the benefits of the proposed real-time signal control are pronounced when more signalized intersections (e.g. network-wide control with 8 signals) are simultaneously controlled.

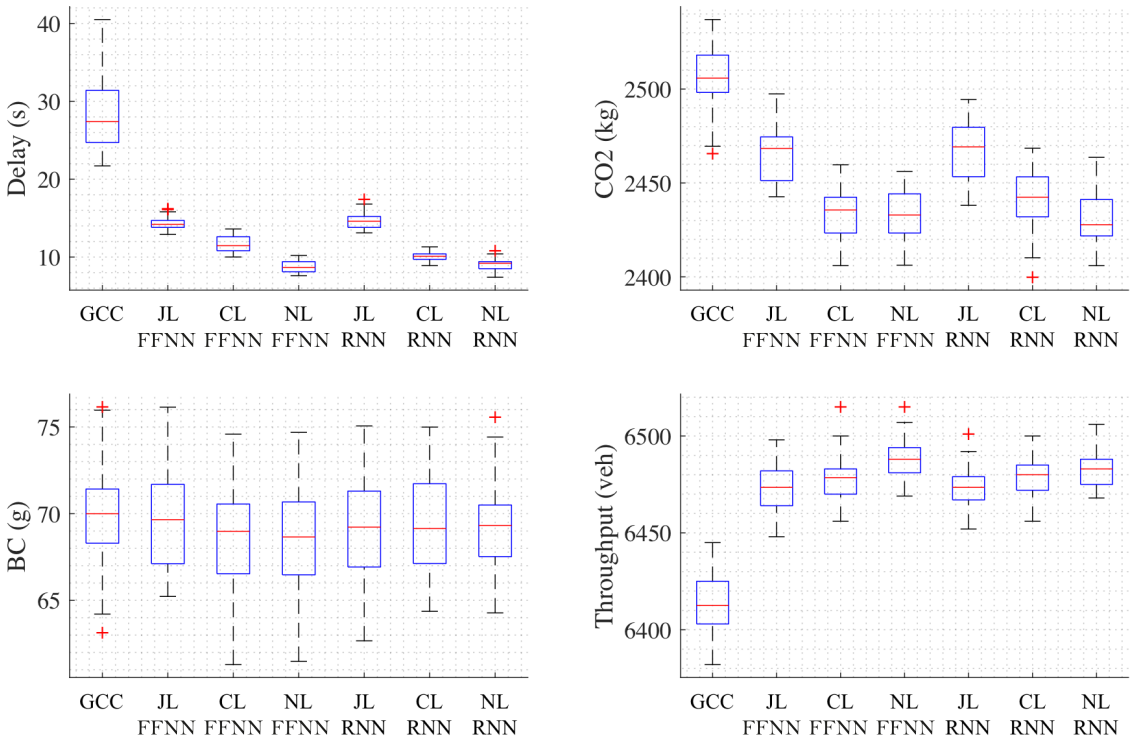


Figure 4.5: Box plot summary (with 30 random simulation runs) of the performance of the four control scenarios in terms of average network delay, total carbon emission, black carbon emission, and throughput.

Table 4.1: Statistical summary (with 30 random simulation runs) of the improvement of [JL], [CL] and [NL] over baseline [GCC]

		Scenario [JL]	Scenario [CL]	Scenario [NL]
Delay	FFNN	14.0 s (48.2%)	16.7 s (58.1%)	19.6 s (68.4%)
	RNN	13.7 s (47.2%)	18.3 s (63.7%)	19.3 s (67.2%)
CO_2	FFNN	38.5 kg (1.5%)	69.9 kg (2.8%)	71.3 kg (2.8%)
	RNN	35.3 kg (1.4%)	63.8 kg (2.5%)	73.2 kg (2.9%)
BC	FFNN	0.59 g (0.7%)	1.3 g (1.8%)	1.4 g (2.0%)
	RNN	0.81 g (1.2%)	0.72 g (0.8%)	0.72 g (0.8%)
Throughput	FFNN	59.6 veh (0.9%)	64.7 veh (1.0%)	73.8 veh (1.2%)
	RNN	59.8 veh (0.9%)	65.1 veh (1.0%)	68.2 veh (1.1%)

4.1.4.2 Improvement at junction level

To further examine the effects of the proposed controls, this PhD thesis evaluates the emission reduction at individual signalized intersections. The emission at an intersection is calculated as the sum of emissions at its incoming approaches, as shown in Figure 4.6.

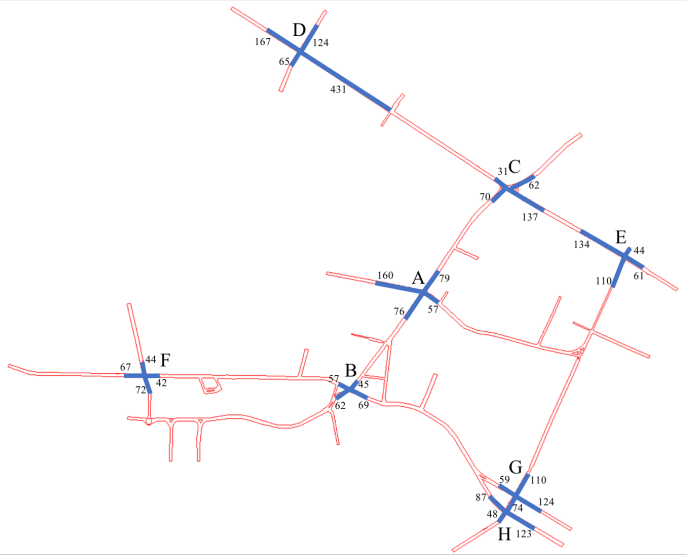


Figure 4.6: The emissions at signalized intersections are calculated on the highlighted incoming approaches, with their lengths shown (in meter).

In Figure 4.7, this PhD research shows the average absolute (left axis) and relative (right axis, in %) CO_2 and BC reductions at the eight signal intersections. The cases examined include: FFNN vs. RNN, and different objectives in the off-line training (i.e. optimizing delay only or combination of delay and CO_2/BC). Both absolute and relative CO_2/BC reductions at the junction level show far greater improvement compared to the network level (see Table 4.1 for comparison). The overall reduction of CO_2 (BC) at the junctions is above 80 kg (1.5 g), when the network-level reductions are up to 73 kg (1.4 g). This means that the network-wide reduction of emissions is almost entirely attributed to the improved signal controls at individual intersections, which indeed shows the effectiveness of the proposed controls. In addition, the majority of the savings occur at junction C, with over 30% reduction of both CO_2 and BC. As for rest of the intersections, the reductions of CO_2 are mostly positive except H, while the reductions of BC are mixed. Finally, in terms of the optimization objective, minimizing delay alone seems to yield similar emission reductions as the weighted sum of delay and emissions.

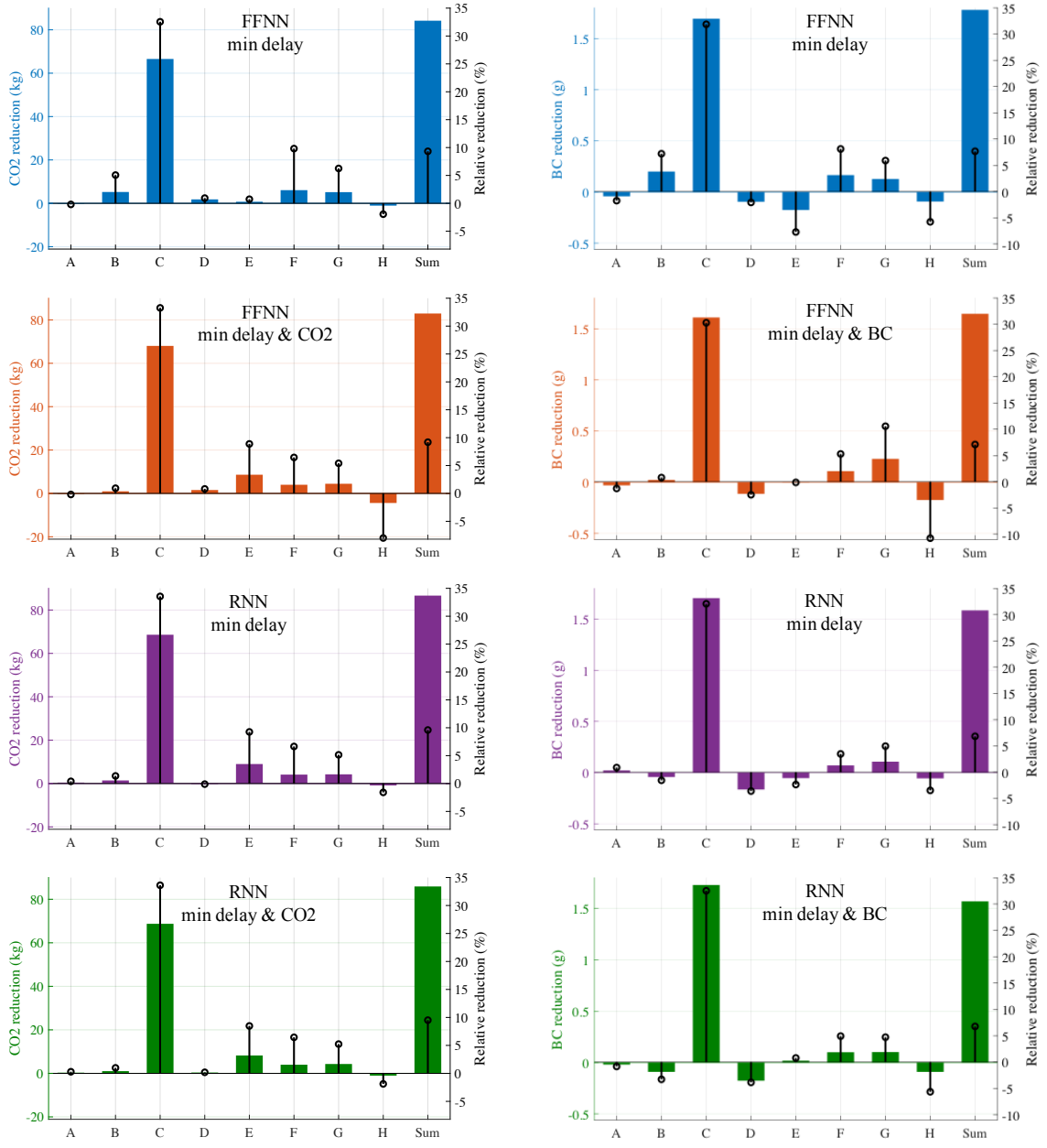


Figure 4.7: Reductions of CO_2 and BC emissions at individual intersections.

4.1.4.3 Trade-off between vehicle delay and emissions

As Figure 4.7 suggests, minimizing delay as the only objective seems to yield similar levels of emission reduction to the joint minimization of delay and emission. Intuitively, reducing vehicle delays leads to increase average speed, which could reduce CO_2 emissions, and reduce vehicle idling and acceleration/deceleration events. To further investigate the potential correlation between vehicle delays and emissions, in Figure 4.8 this thesis shows the scatter plots of delay reduction vs. total emission reduction at the junction level. These data points are obtained from a total of 90 independent on-line simulation runs, where the NDR was respectively optimized off line with the three objectives shown in (3.12). The figure shows that reductions of CO_2 are positively correlated with delay reductions, as indicated by the Pearson test ($p \approx 0$). This is consistent with the interpretation that CO_2 emissions are dependent on average vehicle speed, which is related to vehicle delays. On the other hand, BC reductions do not show meaningful correlation with delay reductions ($p = 0.44$). This is attributed to the fact that BC emissions are primarily caused by stop-and-go cycles and highly dependent on vehicle fleet composition (e.g. buses, HGVs), which are not directly related to average vehicle delays.

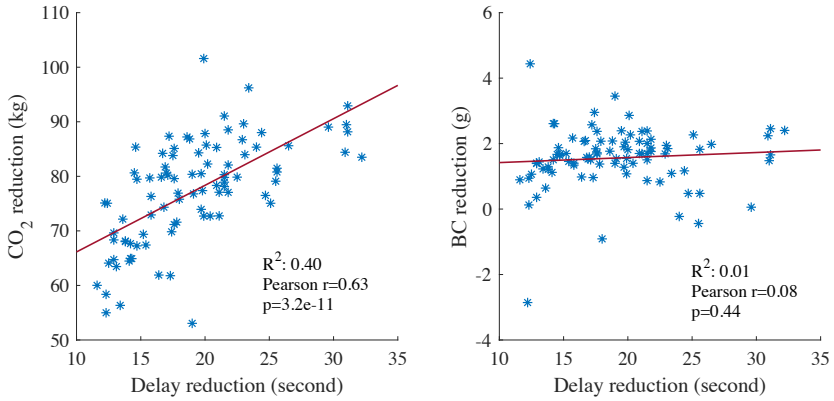


Figure 4.8: Correlation analysis of delay reductions vs. emission reductions at the junction level.

We also observe from Figure 4.7 that, when jointly minimizing delay and emissions (CO_2 or BC), the reduction of emissions does not improve compared to the case when only delay is minimized. Aside from the possibility that the PSO-based off-line heuristic optimization does not yield global optimal within the required computational resources, the lack of discernible trade-off (i.e. statistically significant negative correlation) between delay reduction and emission reduction, as seen in Figure 4.8,

also suggests that optimizing delays in this case seems to be sufficient in reducing emissions at signalized intersections. More effective measures for emission reduction may involve localized, actuated controls such as transit signal priority and offset optimization, which are beyond the scope of this paper.

4.1.4.4 Effect of sensor locations

The proposed NDR framework for real-time signal control perceives the traffic state via the input vector $\boldsymbol{q}$ in (3.1), which, in our case, represents the traffic flows at several key locations in the past 10 minutes. All the results presented to this point are based on the real-world configuration of 41 loop detectors as shown in Figure 4.1(c). To assess the impact of sensor locations on the performance of the NDR framework, this thesis considers two alternative configurations ("Alternative 1" and "Alternative 2") shown in Figure 4.1(d) and (e), respectively. In these hypothetical configurations, each incoming approach of a signalized intersection has a loop detector providing flow information, and there are 32 such detectors. The only difference between "Alternative 1" and "Alternative 2" is that the sensors are closer to the intersections in the former case, while the sensors are located in the middle of the relevant link in the latter.

It is noted that the real-world sensors are distributed unevenly across the network, with missing information on incoming traffic at several key intersections (C, D, E, F). The alternative sensor configurations make sure that the controller receives traffic information at all relevant incoming approaches.

Table 4.2: On-line performances of the NDR based on real-world and alternative sensor locations (respectively (c), (d) and (e) in Figure 4.1). Brackets mean standard error of the results

		Delay	CO_2	BC	Throughput
Real-world	GCC	28.3 s (4.77)	2,503 kg (4.84)	70 g (3.14)	6,413 veh (16.02)
	FFNN	8.7 s (0.71)	2,432 kg (3.69)	68.6 g (3.24)	6,487 veh (10.13)
	RNN	9.0 s (0.78)	2,431 kg (3.91)	69.3 g (2.53)	6,482 veh (8.96)
Alternative 1	FFNN	9.3 s (0.79)	2,432 kg (4.17)	68.7 g (3.28)	6,488 veh (11.76)
	RNN	9.6 s (0.76)	2,430 kg (3.98)	68.6 g (3.36)	6,489 veh (11.2)
Alternative 2	FFNN	14.4 s (1.36)	2,462 kg (4.47)	69.2 g (3.37)	6,463 veh (12.59)
	RNN	13.1 s (1.07)	2,456 kg (4.17)	69 g (3.28)	6,481 veh (12.1)

Table 4.2 shows the performances in the three cases of sensor locations, based on 30 independent on-line tests. Each case is run with both FFNN and RNN as the

NDR. We can see that the two sensor locations of both "Real-world" and "Alternative 1" yield similar performances regardless of the neural networks chosen. Surprisingly, the "Real-world" sensor configuration yields slightly lower delays than "Alternative 1", despite its uneven distribution of sensors and missing data at several key locations. The performance under "Alternative 2" is worse than the other two cases. Nevertheless, compared to the GCC, "Alternative 2" still can efficiently reduce the vehicle delay by about 50%.

One possible explanation of the results is that the neural networks in the NDR are sufficiently deep such that their performances are not sensitive to the dimension or the configuration of the state variables, as long as their parameters are sufficiently trained in the off-line environment. It also shows that the proposed NDR is quite robust against different configurations of the network of sensors.

4.1.4.5 Performances with demand increase

To further test the performance of the proposed signal controls under more congested network conditions, this thesis increases the dynamic travel demand in the network by uniformly scaling up the demand matrix by 10%, 20% and 30%. Tables 4.3, 4.4 and 4.5 show the corresponding performances of FFNN and RNN, where the off-line training aims to minimizes traffic delay, and compares them with the baseline scenario (original demand). In particular, Table 4.4 and 4.5 are based on changed sensor locations "Alternative 1" and "Alternative 2" as shown in Figure 4.1. All figures reported are based on 30 independent random simulation runs.

Table 4.3: On-line performances of the NDR with increased travel demand. The percentages indicate the relative increases compared to the baseline (based on the same type of neural network).

		Delay (s)	CO_2 (kg)	BC (g)	Throughput (veh)
Baseline (0% increase)	FFNN	8.7 / -	2,432 / -	68.6 / -	6,487 / -
	RNN	9.0 / -	2,431 / -	69.3 / -	6,482 / -
10% increase	FFNN	14.0 / 61%	2,627 / 8%	72.8 / 6%	7,068 / 9%
	RNN	12.6 / 40%	2,622 / 8%	73.2 / 6%	7,062 / 9%
20% increase	FFNN	20.3 / 133%	2,860 / 18%	76.3 / 11%	7,737 / 19%
	RNN	17.6 / 96%	2,853 / 17%	78.7 / 14%	7,743 / 19%
30% increase	FFNN	41.8 / 380%	3,219 / 32%	83.9 / 22%	8,403 / 30%
	RNN	36.2 / 302%	3,182 / 31%	83.3 / 20%	8,448 / 30%

In Tables 4.3 4.4 and 4.5, it can be seen that the vehicle throughputs are consistent with the demand increase (10%, 20% and 30%). Regarding the other three

Table 4.4: On-line performances of the NDR with increased travel demand in alternative configuration 1 (Figure 4.1(d)). The percentages indicate the relative increases compared to the baseline (based on the same type of neural network).

		Delay (s)	CO_2 (kg)	BC (g)	Throughput (veh)
Baseline (0% increase)	FFNN	9.3 / -	2,432 / -	68.7 / -	6,488 / -
	RNN	9.6 / -	2,430 / -	68.6 / -	6,489 / -
10% increase	FFNN	13.9 / 49%	2,645 / 8%	73.8 / 6%	7,131 / 9%
	RNN	13.1 / 36%	2,649 / 9%	74 / 6%	7,132 / 9%
20% increase	FFNN	19.6 / 110%	2,856 / 17%	77.8 / 14%	7,749 / 19%
	RNN	17.1 / 78%	2,851 / 17%	77.9 / 11%	7,766 / 19%
30% increase	FFNN	33.8 / 263%	3,118 / 28%	83.4 / 20%	8,322 / 28%
	RNN	26.2 / 172%	3,084 / 26%	82.5 / 22%	8,371 / 29%

Table 4.5: On-line performances of the NDR with increased travel demand in alternative configuration 2 (Figure 4.1(e)). The percentages indicate the relative increases compared to the baseline (based on the same type of neural network).

		Delay (s)	CO_2 (kg)	BC (g)	Throughput (veh)
Baseline (0% increase)	FFNN	14.4 / -	2,462 / -	69.2 / -	6,463 / -
	RNN	13.1 / -	2,456 / -	69 / -	6,481 / -
10% increase	FFNN	25.8 / 79%	2,714 / 10%	74.7 / 7%	7,036 / 8%
	RNN	20.5 / 56%	2,690 / 9%	74.8 / 8%	7,104 / 9%
20% increase	FFNN	35.2 / 144%	2,992 / 21%	80.1 / 15%	7,632 / 18%
	RNN	29 / 121%	2,962 / 20%	80.5 / 16%	7,738 / 19%
30% increase	FFNN	49.8 / 246%	3,267 / 32.6%	83.6 / 20%	8,170 / 26%
	RNN	43 / 228%	3,249 / 32.2%	84.9 / 23%	8,284 / 27%

performance indicators, there is a hyperlinear increase of delays as the network demand increases, which is caused by the nonlinear effect of network dynamics and vehicle congestion. In comparison, CO_2 emissions exhibit linear growth with the demand levels, which suggests that CO_2 emissions are proportional to the vehicle volumes, and are less sensitive to vehicle delays and traffic dynamics. Note that this does not contradict Figure 4.8, which shows high correlation between delay reduction and CO_2 emissions, for two reasons. Firstly, the differences of delay and CO_2 are due to different demand levels in Table 4.3 and 4.4, instead of different control strategies in Figure 4.8. Secondly, Figure 4.8 only shows local reductions at the junction level, when in fact vehicles emit majority (around 63%) of the CO_2 while traveling along the links. Finally, BC emissions grow sublinearly with demand. This is quite interesting as BC emissions are mainly produced during stop-and-go cycles near junctions, which suggests that the NDR approach is effective in reducing BC

emissions when the network demand increases.

Under the original sensor configuration, when the network becomes more congested, RNN starts to outperform FFNN in delay reduction, by 11%, 15% and 15% respectively, under 10%, 20% and 30% demand increase (see Table 4.3). This is because RNN takes into account the temporal precedence and chronological dependencies of the input variables when generating control parameters, and hence is capable of handling the highly nonlinear traffic dynamics under higher network loads.

Similarly, in the case that the loop detectors are re-located (Table 4.4), RNN still has better performance, by 6%, 14% and 29% respectively compared to FFNN. In particular, in the case of 30% demand increase, "Alternative 1" results in lower delays and emissions than the original sensor configuration. This suggests that while sensor re-location may not have a significant effect on performance under normal traffic demand, its impact becomes positive and more pronounced when the network demand increases.

Finally, Table 4.5 suggests that "Alternative 2" yields worse performance of the NDR than the other two sensor configurations. This indicates that placing the sensors away from the intersections compromises the performance of the proposed signal control frameworks.

4.2 Summary

In this section, with microscopic traffic simulation and emission model based on a real-world traffic network in west Glasgow, the applicability and effectiveness of the proposed framework are demonstrated. The traffic and emission models have been set up and calibrated based on an EU project (http://www.carbotraf.eu). Historical traffic flow data are used to reflect the levels of traffic demand and variability. The test phase is conducted in a simulation environment with different random seeds to populate stochasticity in the simulation. The performance of the proposed NDR approach is assessed in terms of travel delay, throughput, total carbon and black carbon emissions. The following findings are made.

- Compared with the fixed-timing plan used on the real-world site, the proposed NDR reduces network-wide delay by up to 68%, total carbon and black carbon emissions by 3% and 2%, respectively, and 1% increase of network throughput.

In addition, most emission reductions take place at signalized intersections, as a result of the proposed controls.

- Under the normal network demand level, the performances of FFNN and RNN are similar in terms of delay, CO_2 and BC emissions, and throughput. When the network demand increases (by 10%, 20% and 30% in this thesis), RNN begins to outperform FFNN. This is likely due to the internal structures of FFNN and RNN as we explained at the end of Section 3.2.2. Furthermore, there seems to be a mismatch between the depth of the neural networks and the nonlinearity of the traffic/control dynamics (i.e. 'depth' of the traffic network). The latter is dependent on the level of saturation of the traffic network, which causes the change in the relative performances of FFNN and RNN.

- There is a strong correlation between delay reductions and CO_2 emissions at local intersections. Such a correlation does not exist between delay reductions and BC emissions. This is because CO_2 emissions are highly dependent on vehicle average speeds, which are related to junction delays; BC emissions, on the other hand, are affected by stop-and-go cycles and vehicle type (such as buses and HGVs), which are not directly related to junction delays. Furthermore, minimizing delays in the off-line training tends to also minimize CO_2 and BC emissions.

- The NDR approaches with FFNN and RNN are both tested with a different set of loop detectors in the network, which offers relatively more complete information on all the incoming approaches of signalized intersections. Note that in the real-world network, some intersections have missing detectors on some incoming approaches. The test result, surprisingly, shows that the performances are very similar in these two cases, which means that the NDR approach is robust against different sensor locations. Lastly, compared to Table 4.3, the results in the Table 4.4 show that the location of the loop detectors might affect the robustness and performance of the proposed framework.

Chapter 5

Reinforcement learning based traffic signal control framework

Unlike pre-defined/fixed-time traffic control, responsive traffic control can be more flexible and meet the requirements of a dynamically changing traffic environment but in fact, it is difficult to handle the complicated characteristics of traffic network and to design appropriate signal timing plans (Lin et al., 2018). With increasing interests of reinforcement learning (RL), which has significant potential to control traffic through adaptive learning, many researchers are currently focusing on applying various RL algorithms to responsive traffic signal controls.

The RL approach can be categorized into model-based (such as prioritized sweeping (Moore & Atkeson, 1993), Dyna (Sutton, 1991) and policy-iteration (Puterman, 2014)) and model-free (such as SARSA and Q-learning). One notable distinction between the two is that model-free methods do not require the learning transition function T (Mannion, 2017, Liang et al., 2018); in the training phase, such methods directly learn from experiences faced by the agents. Then, cumulative rewards from a given environment are maximised by updating their value functions. On the other hand, the goal of the model-based methods is to construct a model interacting with the given environment and leaning transition function T, which can be used to select appropriate actions. In addition, model-based methods typically enjoy a good sampling efficiency as they may require much fewer samples to learn from the constructed model interacting with given environments if the traffic dynamics can be properly approximated. However, model-based methods are usually associated with high computational costs because in highly stochastic traffic environment, the

methods incur higher complexities than their model-free counterparts (El-Tantawy et al., 2013). Therefore, in this research, the RL-based traffic signal control framework follows one of the model-free approaches, by using Q-learning for computational efficiency.

For efficient applications of RL to traffic signal controls, defining three main components (state, action and reward) in the RL algorithm plays a pivotal role. The importance of these three components are briefly explained below.

1. **State** describes the environment in which RL agent faces. For example, in the field of transportation, according to the traffic state (environment), the RL agent can choose the appropriate action which can maximize the cumulative reward. Many candidate state variables have been attempted such as queue length (Balaji et al., 2010, Chin et al., 2011, El-Tantawy et al., 2013, Teo et al., 2014), total delay (Arel et al., 2010) and the number of vehicles on each signal phase (Aslani et al., 2018a).

2. **Actions** are used to implement certain traffic signal control policy by directly acting on the traffic of interest. In addition, the action helps the RL agent learn in the right direction to reach the optimal traffic signal policy. In many existing research, the actions are defined as green time duration (Arel et al., 2010, Aslani et al., 2018a, Balaji et al., 2010), green time extension (Chin et al., 2011, Jin & Ma, 2015) and phase plans (Gao et al., 2017, Wei et al., 2018, Van der Pol & Oliehoek, 2016, Lin et al., 2018). The action has a considerable impact on mitigating traffic congestion because it directly actuates on traffic flow. In this research, for safety considerations, the total cycle time and phasing sequence are assumed to be fixed, including the pedestrian phase, and the action is defined as the phase times.

3. **Reward** can be the signpost to reaching optimal traffic control policy (Li et al., 2016a). According to the reward, the RL agent can make appropriate actions given the current environment. The main objective of the RL is to maximize the reward. For example, given the RL agent's action followed by good performance in the previous stage, if the performance of the agent in the current stage is no better than the previous one, the difference between current and previous performances means 'regret (or negative)' as the reward. In that

case, the RL agent has to reason about appropriate actions in order to collect better rewards. As a result, if the agent continuously obtains better rewards with better performances, it is possible to reach the optimum quickly. That is, the reward can affect the learning speed (or convergence speed). However, in the early stage of the RL implementation, initial states and actions tend to be sub-optimal and there is no guarantee of improving reward/performance. As a result, the reward can be sparse in that stage. This causes prolonged training process to reach optimal(or nearly optimal) performance (Li et al., 2016a). In order to overcome this issue, this thesis proposes a method that uses reward shaping function, adding 3^{rd} party advisor, which combines the concept of potential based reward shaping (Ng et al., 1999) with that of expert advices (Chang, 2006). This allows the RL agent to not only search in the correct learning direction, but also reach optimal (or near optimal) performance as quick as possible. See Chapter 6 for test results.

This chapter details the proposed RL-based framework for real-time traffic signal control, in order to mitigate traffic congestion and maximize the throughput of the underlying traffic network. Alongside a detailed explanation of the basic RL framework, the thesis proposes a new potential reward shaping function, named the 3^{rd} party advisor.

5.1 Background of Reinforcement Learning (RL) - Q-learning

Before the thesis explains the fundamental theory of reinforcement learning, this thesis introduces a few terminologies and notations for understandable presentation below.

Table 5.1: key variables for Reinforcement learning in section 5.1

Symbol	Description
$s, s^{'}$	current state and next state
$a, a^{'}$	current action and next action
r	reward
α	learning rate
γ	discount rate
L()	loss function
$Q(s, a)$	Q-value at state s and action a
θ	primary neural network
θ^{-}	secondary(target) neural network
$\pi(s)$	control policy at state s
$V(s)$	value function at state s in markov decision process(MDP)
$\mathcal{M}$	prioritize experience replay memory
$\mathcal{B}$	minibatch
ϵ	greedy rate

In an uncertain control environment, without prior knowledge, the RL agents become increasingly more intelligent by interacting with the given environment to maximize reward obtained from actions. The overall RL process is illustrated in Figure 5.1.

RL is based on the Markov decision process (MDP) that finds the best policy(π^*). The MDP basically comprises of state(S), action(A), reward(R) and transition(T), i.e. a four-tuple $\langle S, A, R, T \rangle$. As the action executor, agents takes actions corresponding to the given environment, and the resulting state corresponding to the action returns a reward which can be either negative or positive. That is, while the state changes from s_t to s_{t+1}, the agent, which explores a certain environment, perceives the current state s_t and takes an appropriate action a_t (Li et al., 2016a). Here, T is transition function $T(s, a, s') \in (0, 1)$, which is a probability given by selected action $a \in A$ and moves from the current state $s \in S$ to the next state $s' \in S$. However, our proposed framework is based on the model-free approach(Q-learning) which does not require the transition function T (Mannion, 2017, Liang et al., 2018).

An agent in the MDP behaves based on the policy π, which is a mapping from the set of actions (which are selected by a RL agent in a given environment) to the set of states. Therefore, the MDP aims at finding the optimal policy(π^*) that maximizes the expected sum of the discounted rewards (Panait & Luke, 2005).

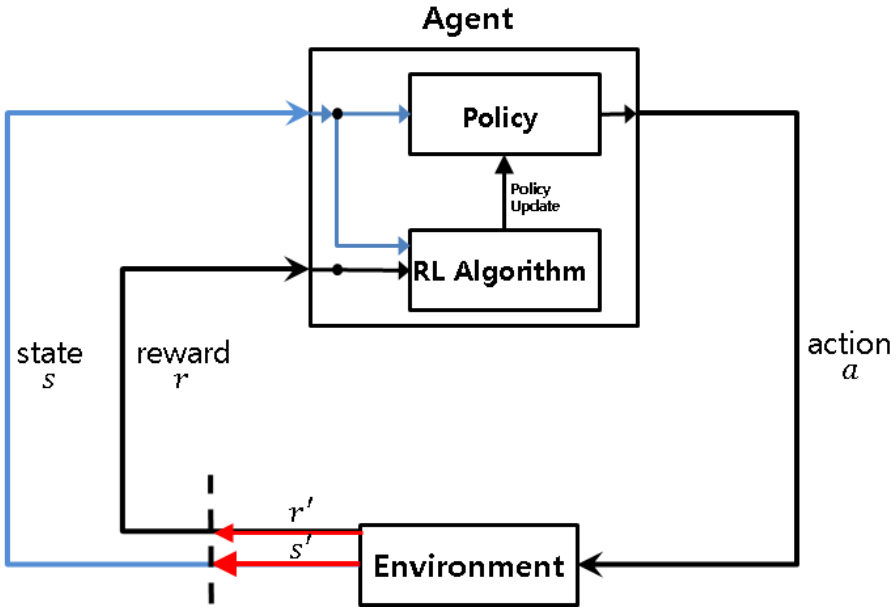


Figure 5.1: The overall process of Reinforcement Learning (RL) overall process

Due to the nature of the MDP, designing the reward function is very important because the RL agent tends to maximize the output generated from the reward, which defines the policy. The value function relies on the policy π which is used for the action selection. If the agent chooses the action by using a given policy, the value function is defined by:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t | s_t = s\right], \qquad \forall s \in S \tag{5.1}$$

where r_t is the designed reward function at time step t, the number of time steps is defined as T, and γ is the discount factor ranging in [0, 1], which describes how important rewards in the future are to current state.

In order to describe the importance to current state in RL, a reward, which occurs in the future N steps from the current state, is multiplied by γ^N. For example, if γ is 0.8, and a reward is 100 which is 2 steps ahead from the current state. The importance of this reward to the current state is (0.8^2)*100 (=64). Usually, in the field of machine learning, lower value of the discount factor γ encourages the action maximizing short-term reward, while higher value of γ makes the agent more forward-looking and maximizes long-term rewards because, in the case of long-term reward, reward can be sparse in the algorithm.

After the number of time steps T, the episodic domain finishes. The algorithm

finds an optimal value function in all possible value functions for all states:

$$V^*(s) = \max_{\pi} V^{\pi}(s) \qquad \forall s \in S \tag{5.2}$$

Here, π is the policy as the pair of perceived states and actions to be taken in those states of the environment. In eq. 5.3, the optimal policy π corresponding to the optimal value function can be expressed as:

$$\pi^*(s) = \arg\max_{\pi} V^{\pi}(s) \qquad \forall s \in S \tag{5.3}$$

Similarly, in order to define the value function taking action a with state s, the Q function in the RL, action-value function for policy π, is defined as:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t(s_t, a_t) | s_t = s, a_t = a, \right], \qquad \forall s \in S, \forall a \in A \tag{5.4}$$

In the same vein, in the Q-learning, Q-value function $Q^{\pi}(s,a)$ is used for value function, instead of $V(s)$ in MDP.

$$\pi^*(s,a) = \arg\max_{\pi} Q^{\pi}(s,a) \qquad \forall s \in S, \forall a \in A \tag{5.5}$$

Based on the nature of the Bellman optimality equation searching for optimal policy π, the optimal action policy π^* in eq. (5.5) can be recursively calculated. In addition, for the given state, the value of the optimal policy can be equal to the expected value of the optimal action. Therefore, the optimal action policy $Q^{\pi^*}(s,a)$ can be calculated by the optimal Q values of states and actions, which maximize the cumulative reward in each episode. Thus, $Q^{\pi^*}(s,a)$ can be calculated by the following equation:

$$Q^{\pi^*}(s,a) = \mathbb{E}_{s'}\left[r_t(s,a) + \gamma \max_{a'} Q^{\pi^*}(s^{'}, a^{'}) | s, a\right] \tag{5.6}$$

Lastly, learning from experiences, the agents in the Q-learning can update their Q-values by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s^{'}, a^{'}) - Q(s,a)] \tag{5.7}$$

where $\max_{a'} Q(s^{'}, a^{'})$ is an estimate of optimal future Q-value after selecting the next

action a' corresponding the next state s'. $\alpha \in [0, 1]$ is the learning rate which can control the extent to which Q value are updated at each time-step. γ is the discount factor ranging from 0 to 1. The overall process of Q-learning is presented in Algorithm 1. Moreover, a RL agent has to keep a balance between exploring the action for the next stage and exploiting the known actions resulting in good performance, in order to maximize the cumulative reward received during its simulation time in each episode and reach the optimal policy. To achieve this, this research employs decaying e-greedy with exploration rate, which randomly selects actions with probability e, or selects the actions defined by Q-value with probability $1 - e$. Through the decaying e-greedy, more random actions are selected at the early stages of the learning process, and more known actions resulting in good performance are selected at later stages of the learning process.

The estimates generated from the value function are simply stored to a look-up table, in which state-action pairs are associated with a Q-value. However, in more complicated environments, the number of state-action pairs that have to be stored increases exponentially. As a result, in real-life applications, learning a large number of state-action pairs requires requires a substantial amount of memory, data and computational time, which are unacceptable for real-time implementation (Sutton et al., 1998). To resolve this, function approximation using deep neural network (Q-network, with weight parameters θ) has been recently applied to mitigate the explosive growth of state-action space and generalize over the large state-action space by scaling linearly all computations without any loss of quality (Sutton et al., 1998). Therefore, table lookup representation is not used in this thesis and, instead, function approximation is employed for the proposed signal control framework. Eq. (5.7) can be changed with function approximation, as follows:

$$Q(s, a) \leftarrow Q(s, a; \theta) + \alpha[r + \gamma \max_{a'} Q(s', a'; \theta^{-}) - Q(s, a; \theta)] \tag{5.8}$$

However, the aforementioned Q-learning can be easily divergent due to strong correlation between samples and the frequent changes of secondary(target) Q-network θ in $\alpha[r + \gamma \max_{a'} Q(s', a'; \theta)$ calculating an estimate of optimal future Q-value. To avoid such divergence, this thesis employs Deep Q-Network (DQN). In the DQN algorithm, as the secondary neural network, the secondary(target) neural network is separately used to avoid the divergence problem;

The secondary network parameter Q value is fixed in the initial training phase and updated every C steps (see algorithm 1). In this research, C is equal to 2 that means the $Qvalue$ is updated twice in an episode. The primary network parameter θ for Q-value estimation is updated by the gradient back-propagation with the Mean Square Error (MSE) as the loss:

$$L(\theta) = \mathbb{E}\left[\left(r(s,a) + \gamma \max_{a'} Q(s^{'}, a; \theta^{-}) - Q(s,a;\theta)\right)^2\right] \tag{5.9}$$

where $r(s,a) + \gamma \max_{a'} Q(s^{'}, a^{'}; \theta^{-})$ and $Q(s,a;\theta)$ are the secondary (target) and estimated Q-values, respectively. Eq. (5.9) calculates the loss between the secondary(target) and the estimated value. Then, the loss is minimized by the gradient descent method.

5.2 Reinforcement learning(RL) structure

The environment model of the Q-learning algorithm is articulated based on the definitions of states and actions. As mentioned in Section 5.1, the three components (state, action and reward) are crucial for the implementation and performance of the RL.

Properly defined states and actions are crucial for the Q-learning system to ensure that the exploration process can be successfully implemented through all the possible states. Chin et al. (2012) mention that if the state is not properly defined, the Q-learning algorithm might lose itself in the process of exploration/exploitation or learning. In addition, if the state-and-action pairs are not set correctly, the whole Q-learning will not be able to get the optimum solution in the whole process (Chin et al., 2012). Therefore, the state-action pairs of the Q-learning play a key part in determining the robustness of the algorithm and its performance.

5.2.1 Agent Design

5.2.1.1 State variables

In RL, the state of the system is determined and returned by the current environment. According to the state observed, the RL agent defines appropriate actions that directly influence the reward.

In this thesis, the traffic state is defined for each link of the network, and may be measured from prevailing sensing infrastructure such as loop detectors, microwave detectors, ANPR cameras and GPS-enabled devices. In particular, the following three types of state variables are considered:

- **Average Relative Occupancy (ARO)**, which is defined to be the ratio between the link occupancy (number of cars on the link at a particular time) and the link's holding capacity (a constant representing the maximum number of cars the link can store), which is then averaged over at least one full signal cycle to filter the within-cycle effect;
- **Average Delay (AD)**, defined to be the average link traversal time (including wait time at the signal) averaged over at least a full signal cycle;
- **Average Speed (AS)**, defined to be the average vehicle speed on the link (link length over travel time) averaged over at least a full cycle.

These three state variables are commonly studied in the traffic engineering literature and can be easily obtained via loop detectors, ANPR cameras, and GPS devices. To further balance the three variables, we also consider the following derived state variable

- **Weighted Sum (WS)**, defined as

$$WS = \boldsymbol{w}_1 \mathrm{AD} + \boldsymbol{w}_2 \mathrm{ARO} + \boldsymbol{w}_3 \mathrm{AS} \tag{5.10}$$

WS is the sum of multiple states considering each state in the environment simultaneously. In this research, in order to efficiently describe traffic condition in the environment, this research uses WS in which the three states are considered equally, as an additional state. The structure of the defined state variables is illustrated in Figure 5.2.

5.2.1.2 Action

The complexity of traffic signal control is exemplified by the sophisticated phasing and timing plans. Although most research defines the action in relation to the green

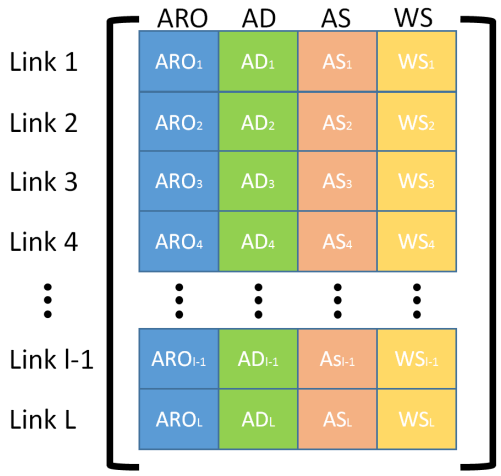


Figure 5.2: Four state definitions and conditions

time, the definition of the action policy might not be realistic and applicable to real-life traffic signal system with safety constraints for pedestrians and drivers.

To comply with real-world traffic signal control specifications (such as fixed cycle time and phasing plans for traffic stability and safety) while achieving more efficient traffic flow through real-time controls, we propose a new action policy to optimize the green time of each phase at every intersection. In particular, the agent monitors/calculates four traffic states described in Section 5.2.1.1: Average Relative Occupancy (ARO), Average Delay(AD), Average Speed(AS) and Weighted Sum(WS) for all the links. Based on these state variables, the Q-values at each phase ($\boldsymbol{P}$) of each intersection ($\boldsymbol{I}$) are estimated by the neural network (deep Q-network).

The signal control parameters usually include cycle time, phasing plans, green time, all red and offset (Han & Gayah, 2015). These parameters are subjected to real-life traffic safety considerations (Mascia et al., 2015). In this study, the phasing traffic plans and cycle time are fixed as they are predominantly influenced by safety regulations. The control variables amount to the green times of all the phases for every intersection. The agent decides which state should be given more consideration at each intersection, by using the ε-greedy strategy through exploration and exploitation. Thus, the action is defined as the discrete choice of one from the four state variables by the value of action ($\boldsymbol{a}$) that maximizes $Q(s, a; \theta)$ estimated from the primary neural network (deep Q-network) using four states as inputs:

$$\boldsymbol{a}\begin{cases} 1 : \text{Index considering Average Relative Occupancy (ARO)} \\ 2 : \text{Index considering Average Delay (AD)} \\ 3 : \text{Index considering Average Speed (AS)} \\ 4 : \text{Index considering Weighted Sum (WS)} \end{cases} \tag{5.11}$$

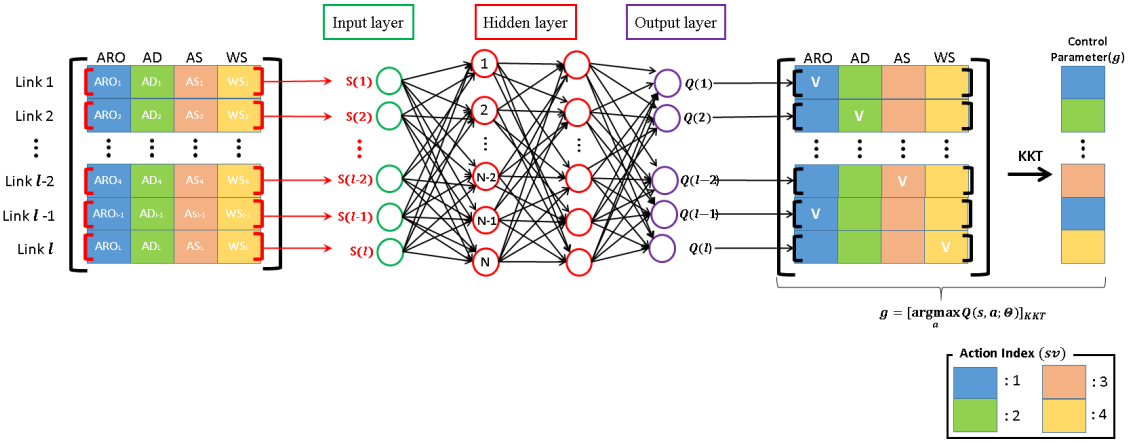


Figure 5.3: the definition process of the action

Based on the defined actions at each phase of each intersection, the phase green control parameters (g) take the Q-values estimated from the primary neural network using network parameter θ and four state inputs. Thus, the phase green control parameters (g) defined by the actions are denoted $g = (g_1, g_2, \ldots, g_N)^T$ where N is the number of phases. The control parameters g_p's must satisfy the following constraints:

$$g_{min} \leq g_p \leq g_{max} \quad \forall p, \quad \sum_{p=1}^{N} g_p = T_{cycle} - \triangle \tag{5.12}$$

where g_{min} and g_{max} denote minimum and maximum green times, respectively; T_{cycle} is the fixed cycle time, and $\triangle$ includes amber, all-red and pedestrian phase time, which are fixed for safety reasons. In this thesis, g_{min} and g_{max} is 5 and 50 sec, respectively. More detail of projection onto the feasible control set is explained in section 3.2.3.

5.2.1.3 Rewards

Reward is one of the main features that distinguish the RL from other learning-based algorithms.

Remark 2. Reinforcement learning(RL) is based on Markov decision process(MDP) which make a decision in stochastic and sequential environment. The RL agent interacts with the given environment by changing state in accordance with action choices generated by the RL agent and collecting temporal reward for a simulation(episode). Therefore, through trial-and-error learning, the RL agent **learns** the optimal policy maximizing the cumulative(total) reward in the given environment. On the other hand, NDR is the decision rule-based optimization model. The NDR agent **finds** the optimal policy based on heuristic method(Particle Swarm Optimization, PSO)

The reward is cast as feedback to the RL model about the performance obtained by the previous actions. Therefore, it plays a pivotal role in correctly and efficiently pitching the learning process towards the best action policy. To achieve this, the literature considers a range of forms of the reward such as queue length (Araghi et al., 2013, Aslani et al., 2018a, Aziz et al., 2018, Balaji et al., 2010, Teo et al., 2014), vehicle delay (El-Tantawy et al., 2013, Gao et al., 2017), relative reduction of total travel delay (Jin & Ma, 2015), outflow of the road network (Lin et al., 2018), vehicle waiting time (Van der Pol & Oliehoek, 2016), and the weighted sum using the queue length, delay, waiting time, light switches, and total travel time(Wei et al., 2018).

In this research, the main objective is to alleviate traffic congestion by minimizing delays and maximizing vehicle throughput. For the reward function, we consider three performances:

- Total network-wide delay over the past control period (**TP**);
- Average delay per vehicle over the past control period (**AP**); and
- Vehicle throughput over the past control period (**VT**).

Note that TP measures the total network-wide delay, which is influenced by the number of vehicles therein, and hence indirectly reflects traffic volume; AP is a more direct measure of the level of congestion by measuring the delay per vehicle. Lastly, VT describes the vehicle capacity of the urban traffic network. Here, by using TP and

AP, this thesis calculate rewards in the macro- and micro-perspective. To balance these three performances, we formulate the reward function as the linear combination of the relative differences of the performances in two consecutive time periods:

$$\textbf{Reward(t)} = \begin{cases} 0 & (t = 1) \\ w_1 \cdot \frac{\mathrm{TP}(t-1)-\mathrm{TP}(t)}{\mathrm{TP}(t)} + w_2 \cdot \frac{\mathrm{AP}(t-1)-\mathrm{AP}(t)}{\mathrm{AP}(t)} + w_3 \cdot \frac{\mathrm{VT}(t)-\mathrm{VT}(t-1)}{\mathrm{VT}(t)} & \\ & (w_1 + w_2 + w_3 = 1, t \geq 2) \end{cases} \tag{5.13}$$

where $w_1, w_2, w_3 \in (0,\ 1)$ are weights to keep a balance between the three normalized objectives. This reward function is aligned with the overall objective of minimizing the average vehicle delay and maximizing the vehicle throughput in the traffic network.

5.2.2 Potential-based reward shaping function with 3rd party advisor

Before this research addresses a novel potential-based reward shaping function, we introduce key variables and terminologies for the easy presentation in Table 5.1;

Table 5.2: key variables for Potential-based reward shaping function with 3^{rd} party advisor in Section 5.2.2

Symbol	Description
$Q(s, a)$	Q value corresponding the state s and the action a
$F(s, s^{'})$	potential-based reinforcement(or reward shaping) function moving from the current state s and the next state $s^{'}$
$\pi(s)$ or $\pi(s^{'})$	the potentials of the current or the next state, respectively
$\mathcal{P}_{advisor}(s)$	the probability of matching the action policy of the RL agent at the state s with that of the 3^{rd} party advisor
$sv^{advisor}$	action index when selecting one of four state variables by 3^{rd} advisor at each link
sv	action index when selecting one of four state variables by RL agent at each link shown in figure 5.3
β	Application rate of advice from $3^{rd} party advisor$
i	the $i^{th} link$ in traffic network.
I	total number of the link in traffic network.
$\Phi(\ \cdot\ )$	potential function for reward shaping

Reinforcement learning (RL) is known to effectively handle complex dynamics by interaction with a given environment (Grześ & Kudenko, 2010). However, conventional RL typically have slow learning speeds before converging to the optimal policy due to huge action space (Grześ & Kudenko, 2010), sparse reward and delayed reward (Grześ & Kudenko, 2010). To tackle these problems, reward shaping is one technique aiming to resolve the temporal credit assignment problem, which causes the delayed reward leading to slow convergence and undermining the influence of the reward in the learning process. Ng et al. (1999) devise Potential-Based Reward Shaping (PBRS) by proving the policy invariance under reward function transformations (the process of proving the optimal policy preservation is discussed in depth therein). That is, the optimal policy is not changed, even if the reward function is transformed.

Basically, in order to accelerate convergence speed of the RL approach, reward shaping provides additional reward, as a heuristic knowledge. Therefore, reward shaping function is added below.

$$Q(s,a) \leftarrow Q(s,a;\theta) + \alpha[\underbrace{r + \boldsymbol{F(s,s')}}_{\text{reward shaping}} + \gamma \max_{a'} Q(s',a';\theta^{-}) - Q(s,a;\theta)] \tag{5.14}$$

where $\boldsymbol{F(s,s')}$ is a potential-based reinforcement function (Ng et al., 1999, Chang, 2006), which can call an additional reward function moving from current state s to the next state s'. Moreover, by giving the agent frequent feedbacks on proper actions, the issue of sparse reward, which causes very slow convergence to the optimal policy, can be resolved. Here, the function $\boldsymbol{F(s,s')}$ with the potential Φ is defined by Ng et al. (1999):

$$\boldsymbol{F(s,s')} = \gamma\Phi(s') - \Phi(s) \tag{5.15}$$

where γ is the discount factor, and $\Phi(s)$ and $\Phi(s')$ denote the potentials of the current and next states, respectively. In addition, this heuristic technique is flexible to combine background knowledge (Grześ & Kudenko, 2010, Chang, 2006, Mannion, 2017). In particular, for fast learning, Chang (2006) tries to incorporate expert advices (or knowledge) into (5.15). The expert/advisor can be neural networks, decision trees (e.g. random forest (Mitchell, 1997)), expert experience information and model-free or model-based RL (Chang, 2006). This research employs another neural network as a 3^{rd} party advisor. So, the potential function for all states, which combine the advise of the 3^{rd} party, can be generated as follows:

$$\Phi(s) = \mathcal{P}_{advisor}(s) \tag{5.16}$$

In order to calculate the probability $\mathcal{P}_{advisor}(s)$, $Q^{advisor}$-network is additionally used for the potential-based reward shaping function because it might have a good shaping potential (Chang, 2006). As mentioned in figure 5.3, action index sv is used to calculate the probability with Q-network, which is based on which state variable is the most important among state variables at each link. $sv^{advisor}$ is similar to sv but uses different Q-network (named $Q^{advisor}$-network). Therefore, $\mathcal{P}_{advisor}$ is the probability of matching sv with $sv^{advisor}$ at each link. The probability $\mathcal{P}_{advisor}$ can be defined as:

$$\mathcal{P}_{advisor}(s) = \beta \left[\frac{1}{I} \sum_{i=1}^{I} M(sv_i, sv_i^{advisor}) \right] \tag{5.17}$$

where

$$M(sv, sv^{advisor}) = \begin{cases} 1, & sv = sv^{advisor} \\ 0, & \text{Otherwise} \end{cases} \tag{5.18}$$

$$\mathcal{P}_{advisor}(s) = \begin{cases} \mathcal{P}_{advisor}(s), & \mathcal{P}_{advisor}(s) = (1, 0) \\ 0, & \mathcal{P}_{advisor}(s) = 0 \ or \ 1 \end{cases} \tag{5.19}$$

where i $(i = 1, \ldots, I)$ indicates each link in the traffic network. $\beta(= [0, 1])$ is application rate which is about how much expert advice the agent applies and is calculated by $sv^{advisor} = \text{argmax}_a Q^{advisor}(s, a; \theta^{advisor})$ Lastly, M is the binary function of matching sv with $sv^{advisor}$. If the index sv of the RL agent at i^{th} link is equal to the index $sc^{advisor}$ suggested by the 3^{rd} party advisor, the matching function $M = 1$, and 0 otherwise. Lastly, this research does not allow the case that $\mathcal{P}_{advisor}(s)$ is equal to 0 or 1 because this research has an assumption that as people think differently, the traffic control policy generated through the $3^{rd} party advisor$ cannot be 100% matched with the traffic control policy produced by the RL agent. In addition, if the two agents are fully different (in the case $\mathcal{P}_{advisor}(s) = 0$), this research assumes that one of the two agents is learning in the incorrect way. Therefore, in that cases, our framework does not consider 3^{rd} party advisor.

5.2.3 Overall procedure for real-time traffic signal control

To sum up the previous sections, this research proposes a new process to address two main challenges: traffic congestion minimization and fast learning speed. The pseudo code is shown in Algorithm 1. As a non-linear approximator, fully-connected multi-layers deep neural network (Q-network) is applied in this thesis, in order to obtain the maximized reward generated by the action.

In addition, prioritized experience replay (PER) is employed for stable and quick learning in the Q-network and keeping a low degree of correlation between samples in the training process. The PER is important for stable learning with the optimal action by more efficiently using the previous experience for a simulation time (Schaul et al., 2015). The basic concept of PER is to store the experiences of RL agent in a limited size of re-training data named mini-batch ($\mathcal{B}$). (Here, $\mathcal{B}$ is the mini-batch which retrain our framework in the training procedure, in order to achieve fast

convergence.) Through iteratively training mini-batched data(or experience), the RL agent easily recap the previous experience in the mini-batch data, then can keep and improve its performance when facing new real-world experience. Unlike Experience Relay (ER), PER prioritizes the stored experiences by ranking the performance of the RL agent in the stored experience or using priority proportional to temporal difference (TD)-error in the stored experience. The superiority of the PER has shown in the research of (Schaul et al., 2015, Liang et al., 2018). Therefore, this thesis applies ranked-based method on the PER (Schaul et al., 2015, Liang et al., 2018), which stores the four-tuple $< s, a, r, s' >$ with performance into memory (M), which has a fixed size (mini-batch), and samples tuples from the memory based on the ranked-based priority.

Moreover, as mentioned in 5.1, deep neural network in this thesis uses two networks(primary neural network and secondary neural network), in order to get more stable training and quick learning process. To achieve fast convergence and quickly find the optimal policy, this research employs potential-based reinforcement function using 3^{rd} party advisor neural network. The following pseudo code in Algorithm 1 describes the proposed RL-based model.

Algorithm 1 Q-learning: Learn function $Q : s \times a \rightarrow r$

Require:

Prioritize Experience Replay memory $\mathcal{M}$
Minibatch $\mathcal{B}$
Greedy rate ϵ
States $s = \{1, \ldots, \mathcal{S}\}$
Actions $a = \{1, \ldots, \mathcal{A}\}$,
Reward function R :s×a$\rightarrow r$
Learning rate $\alpha \in [0, 1]$
Discount factor $\gamma \in [0, 1]$
Episode number $ep = \{1, \ldots, EP\}$

Notation

θ : the primary neural network.[3pt]
θ^- : the secondary(target) neural network.
$\theta_{advisor}$: the 3^{rd} party advisor's neural network.
$\phi()$: pre-process function (normalization of state variable)
$\hat{Q}(s^{'}, a^{'}; \theta^-)$: Q value for next state and next action
using secondary(target) NN
$Q(s, a; \theta)$: Q value for current state and current action using primary NN

for episode $ep = 1$ to EP **do**
Initialize parameters of $\theta, \theta^-, \theta_{advisor}$ with random values.
Initialize $\mathcal{M}$ to be empty at each episode before implementation
Initialize states space s with the starting scenario at the traffic network.
Initialize Actions space a.
while there exists a state s **do**
With probability ϵ select a random action a
Ohterwise select $a = \arg\max_a Q\ (\phi(s), a; \theta))$
Execute action a in traffic environment and Observe reward r and next state $s^{'}$
Add the four-tuple $\left\langle s, a, r, s^{'} \right\rangle$ into $\mathcal{M}$
Assign $s^{'}$ to $s(s \leftarrow s^{'})$
if size(M) > size(B) **then**
Select B samples from M based on the sampling rank-based priorities.
Set

$$y = \begin{cases} r, & \text{for terminal } s^{'} \\ r(s, a, s^{'}) + \boldsymbol{F(s, s'; \theta_{advisor})} + \gamma \cdot \max_{a'} \hat{Q}(s', a'; \theta^-)), & \text{for non-terminal } s^{'} \end{cases} \tag{5.20}$$

Perform a gradient descent step on $(y - Q(s, a; \theta))^2$
Every C steps update $\hat{Q} \leftarrow Q$
end if
end while
$i \leftarrow i + 1$
end for

The goal of our model is to develop a responsive/real-time traffic signal control strategy, which adaptively changes phase green times at relevant intersections to minimize average vehicle delay and maximize network throughput under uncertain traffic flows and demands. As the pre-train stage, the agent first randomly collects

tuples $< s, a, r, s' >$ of different performance until sufficient samples for mini-batch are collected. In that stage, the priorities of the collected samples are the same. From training, the priorities of the samples collected in the memory change and are chosen with different probabilities based on rank-based experience (Schaul et al., 2015). The network parameters θ are updated by performing gradient descent step on loss function defined as the square error between the Q-value and secondary(target) Q-value output from the Q-network. Lastly, the RL agent learns the optimal action policy maximizing reward based on different traffic demands.

5.3 Summary

In this chapter, the reinforcement learning framework is reviewed and the proposed Advanced Reinforcement Learning (ARL) framework is developed for real-time signal control. As a model-free approach, Q-learning is employed to reduce computational expenses and generate an immediate and proper traffic signal timings. Firstly, three types of state variables are considered (average relative occupancy, average delay, and average speed). In particular, their weighted sum is added as another state variable to keep a balance between the three. Secondly, in view of real-world traffic signal constraints, the phase green times, which are the primary control variables, are explicitly derived using the Karush-Kuhn-Tucker (KTT) conditions. Thirdly, the rewards are calculated based on total network-wide delay, average delay per vehicle and vehicle throughput over the paster control period. To avoid sparse reward and improve the learning speed, potential-based reward shaping function with 3^{rd} party advisor is proposed. Lastly, through prioritized experience replay, the proposed framework is capable of stable and rapid learning and keeping a low degree of correlation between samples in the training process.

Chapter 6

Assessment and comparative study of Advanced Reinforcement Learning in responsive traffic signal control

6.1 Experiment setting

In this chapter, the proposed Advanced Reinforcement Learning (ARL) framework will be applied to responsive traffic signal control based on the Glasgow test network. Through a quantitative evaluation by comparing with the performance of other benchmark models, we show the superiority of the proposed framework.

6.1.1 Traffic flow dynamics

The proposed framework is applied to a real-world traffic network in West end of Glasgow, Scotland. The test network consists of 5 signalized intersections and 35 directed links; see Figure 6.1(b). The simulation of traffic dynamics follows the Lighthill-Whitham-Richards (LWR) model extensively used in the traffic modeling and control literature (Han et al., 2014, Han & Gayah, 2015). For brevity, this chapter only highlights the key part of the model, while the rest of the modeling details can be found in Han et al. (2014) and Han & Gayah (2015). The LWR model is a macroscopic traffic simulation model that is based on the conservation of mass

Table 6.1: key variables for traffic flow dynamics.

Symbol	Description
$D(t)$	the demand at links
$S(t)$	the supply at links
$f_{\text{in}}(t)$	the inflow of each link
$f_{\text{out}}(t)$	the exit flow of each link
$u(t)$	signal control parameter
C	flow capacity
N_{in}	cumulative number of vehicle arriving at each link
N_{out}	cumulative number of vehicle leaving from each link
L	the length of each link
ρ^{jam}	the traffic jam density
i	link index ($i \in I$)

and a fundamental diagram describing the relationship between traffic density and flow:

$$\partial_t \rho(t,\, x) + \partial_x f\Big(\rho(t,\, x)\Big) \;=\; 0 \quad (t,\, x) \in [0,\, T] \times [a,\, b] \tag{6.1}$$

where $\rho(t,\, x)$ and $f\Big(\rho(t,\, x)\Big)$ denote density and flow, respectively. $f(\rho)$ is a concave function of the density, and in this thesis is simplified as the triangular fundamental diagram:

$$f(\rho) = \begin{cases} v\rho & \rho \in [0,\, \rho^c] \\ -w(\rho - \rho^{\text{jam}}) & \rho \in (\rho^c,\, \rho^{\text{jam}}] \end{cases} \tag{6.2}$$

where v and w are the positive forward and backward kinematic wave speeds; ρ^{jam} denotes the jam density, and ρ^c is the critical density at which the flow is maximized, i.e. the flow capacity $C = f(\rho^c)$.

The demand and supply of a link can be defined as

$$D(t) = \begin{cases} C & \text{if } \rho(t,\, b-) \geq \rho^c \\ f\Big(\rho(t,\, b-\Big) & \text{if } \rho(t,\, b-) < \rho^c \end{cases} \tag{6.3}$$

$$S(t) = \begin{cases} C & \text{if } \rho(t,\, a+) \leq \rho^c \\ f\Big(\rho(t,\, a+\Big) & \text{if } \rho(t,\, a+) > \rho^c \end{cases} \tag{6.4}$$

The demand and supply indicate the maximum flow that can be accommodated at the exit and entrance of a link, respectively. Based on such definition, the signalized intersection model follows that of Han et al. (2014) and Han & Gayah (2015), where the on-and-off effect is incorporated into the model via the following formula:

$$f_{out}(t) = \min\Big\{D(t),\ u(t)\ \cdot\ \min\{C,\, S_{dn}(t)\}\Big\} \tag{6.5}$$

where the signal control for the link of interest is expressed as a binary variable:

$$u(t) = \begin{cases} 1 & \text{if the signal is green at time } t \\ 0 & \text{if the signal is red at time } t \end{cases}, \tag{6.6}$$

and $S_{dn}(t)$ denotes the downstream supply, whose precise mathematical expression depends on the junction layout, signal phasing plan, and vehicle turning probabilities/proportions. The reader is referred to Han et al. (2014) and Han & Gayah (2015) for full details. Finally, to define the various state variables, we define the cumulative link entering and exiting counts:

$$N^i_{\text{in}}(t) = \int_0^t f^i_{\text{in}}(t)\, dt, \quad N^i_{\text{out}} = \int_0^t f^i_{\text{out}}(t)\, dt, \quad \forall i \in I \tag{6.7}$$

The simulation environment allows a number of traffic state variables such as average vehicle relative occupancy(ARO), average vehicle delay (AD) and average vehicle speed(AS) to be defined. The Instantaneous Relative Occupancy(IRO) is defined as follows;

$$\text{IRO}^i(t) = \frac{N^i_{\text{in}}(t) - N^i_{\text{out}}(t)}{L^i \rho_i^{\text{jam}}} \tag{6.8}$$

where L_i indicates the length of link i, and ρ_i^{jam} is the jam density. Clearly, Eqn. (6.8) represents a quantity that changes over time within a full signal cycle. In order to reasonably reflect the level of link occupancy that is independent of the on-and-off effect of signal control, we integrate the IRO over at least a full cycle to obtain the

Average Relative Occupancy (ARO):

$$\mathrm{ARO}^i(T_j) = \frac{1}{|T_j|}\int_{t\in T_j} \mathrm{IRO}^i(t)\,dt \qquad i \in I,\, T_j \subset \mathbb{R} \tag{6.9}$$

where T_j is a given time interval whose length is a multiple of the fixed signal cycle time.

The definition of vehicle delay (travel time) within a signal-controlled link relies on the notion of link entry time function $\tau^i(\,\cdot\,)$:

$$t^i_{\mathrm{in}} = \tau^i(t_{\mathrm{out}}) = \max\left\{t:\ N^i_{\mathrm{in}}(t) = N^i_{\mathrm{out}}(t_{\mathrm{out}})\right\} \tag{6.10}$$

where t^i_{in} is the link entry time of a vehicle that leaves the link at time t_{out}. Therefore, the Instantaneous Delay (ID) at link i is simply calculated as

$$\mathrm{ID}^i(t^i_{\mathrm{out}}) = t^i_{\mathrm{in}} - t^i_{\mathrm{out}} = \tau^i(t_{\mathrm{out}}) - t_{\mathrm{out}} \quad \forall t_{\mathrm{out}},\ i \in I \tag{6.11}$$

Similar to IRO, ID varies within a full signal cycle, hence we propose the Average Delay(AD) as the time-integral of ID over several cycles:

$$\mathrm{AD}^i(T_j) = \frac{1}{|T_j|}\int_{t\in T_j} \mathrm{ID}^i(t)\,dt \qquad i \in I,\, T_j \subset \mathbb{R} \tag{6.12}$$

where T_j is a given time interval whose length is a multiple of the fixed signal cycle time. Finally, the Average Speed (AS) within the signalized link can be defined as

$$\mathrm{AS}^i(T_j) = \frac{L^i}{\mathrm{AD}^i(T_j)} \qquad i \in I,\, T_j \subset \mathbb{R} \tag{6.13}$$

6.1.2 Configuration of ARL

In our experiments, Matlab(R2017b) is employed to execute the proposed framework and experiments. In accordance with Table 5.1, the discount factor γ is set to be 0.99 and all weights of the neural network (see details in figure 3.1) are updated by the mini-batch gradient descent with the learning rate $\alpha = 0.001$ and mini batch size $\mathcal{B} = 50$. Unlike the feedforward neural network (FFNN) described in section 3.2.2, the FFNN used in this section has five hidden layers with 100, 500, 500, 500 and 100 neurons. The fully connected neural network employs the sigmoid activation func-

tion and continuously updates the weights of connections among the neurons as the simulation continues (see more details in section 3.2.2). The networks is trained for 1,000 episodes and tested for another 50 episodes. Here, an episode is a simulation from the start state until termination state. In addition, each episode has different traffic demand. The main objectives of this framework are the average vehicle delay and the average vehicle throughput. The action is chosen by ϵ-greedy method alternating probability 1-ϵ(exploitation) with probability ϵ (exploration) every 12 min. Lastly, for the implementation of the prioritized experience replay (PER), the PER memory size is set to be 10,000.

6.1.3 Signal control details for experiment

The traffic network for experiment consists of five major signalized intersections, shown as intersections A-E in figure 6.1. The time step size is 5 s, and the cycle times at all the intersections are set to be 100 s. With the exception of junction B, which has three vehicle movement phases, all intersections have four vehicle movement phases.

The simulation horizon is morning commuting period (8:00 am - 9:00 am) of a typical working day. The ARL agent adaptively changes signal control parameters (the green times of all the phases except the pedestrian phase) every 12 min, which is the control resolution. In other words, during the course of 1 hr simulation, the ARL may adaptively change the phase green times of each junction for the next 12-min period, based on average relative occupancy(ARO), average delay(AD) and average speed (AS) in the past 12 min.

6.1.4 Evaluation

6.1.4.1 Research objective

To investigate the extent of traffic impact of the proposed real-time signal controls, this experiment considers two key performance indicators (KPIs):

1. average delay per vehicle.

2. network throughput.

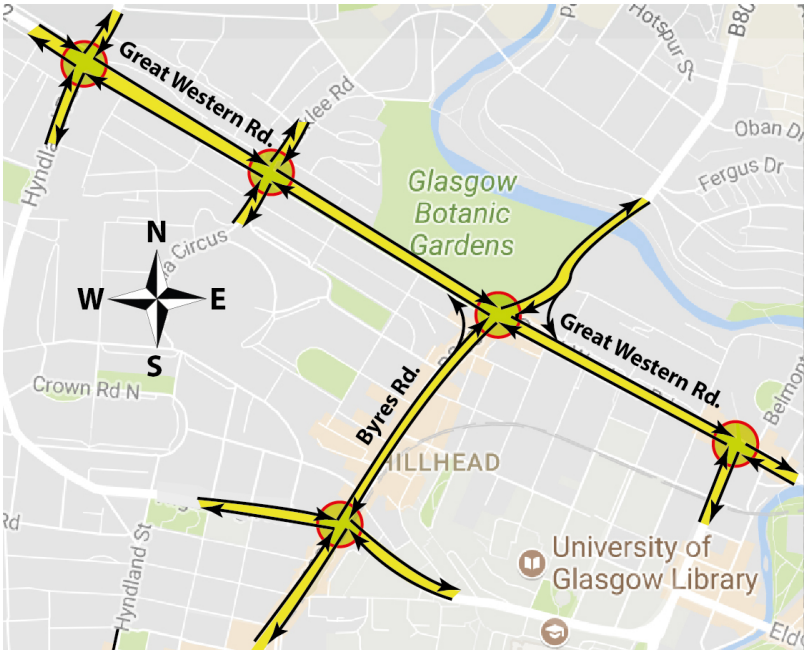


Figure 6.1: The test network in the West end of Glasgow, Scotland.

Average delay per vehicle and network throughput are the main objectives of the proposed responsive signal optimization framework. In particular, we approximate the total vehicle travel time on a link as:

$$\sigma^i = \int_0^T N^i_{\text{in}}(t) - N^i_{\text{out}}(t)\, dt \qquad i \in I \tag{6.14}$$

where σ^i represents the total vehicle travel time on link i. The network-wide average delay per vehicle is then approximated as

$$\text{AD} = \frac{\sum_{i \in I} \sigma^i}{\sum_{i \in I_{\text{entry}}} N^i_{\text{in}}(T)} \tag{6.15}$$

where I_{entry} is the set of source links of the network.

On the other hand, the network throughput is defined as

$$\text{NT} = \sum_{i \in I_{\text{exit}}} N^i_{\text{out}}(T) \tag{6.16}$$

where I_{exit} denotes the set of sink links in the network.

These two objectives are direct measurements of the network's efficiency and capacity under certain signal controls. In addition, the essential goal of the RL is to find the optimal action maximizing the cumulative reward in each episode. In

general, reward shaping function can not only avoid sparse reward, but also improve the learning speed. According to the definition of the potential in reward shaping function, the performance and the learning speed (or convergence speed) might be different (Mannion, 2017). In order to show the effectiveness of our model using the potential-based reward shaping function with 3^{rd} party advisor, three different benchmark models are set in the section 6.1.4.2, which use state variable ARO, AD and AS as potentials in the reward shaping function, respectively (see the details in eq.6.17). Therefore, this thesis additionally checks the performance according to reward shaping function using different potentials (see eq.6.17).

6.1.4.2 Benchmarks for evaluation

In section 6.1.4.1, the three objectives play a significant role on applying the Q-learning to real-life signal optimization. To evaluate the effectiveness of the proposed ARL in terms of the three objectives, we consider and compare the following methods:

1. [GCC] : Fixed signal timing
2. [ARL]: Base Q-learning + Reward shaping function using 3rd party advisor
3. [BA]: Base Q-learning
4. [R-ARO]: Base Q-learning + Reward shaping function using ARO
5. [R-AD]: Base Q-learning + Reward shaping function using AD
6. [R-AS]: Base Q-learning + Reward shaping function using AS

Here, [GCC] is provided by the Glasgow City Council (GCC) as an off-line approximation of the SCOOT system. [BA], [R-ARO], [R-AD] and [R-AS] are based on the Base Q-learning for the responsive traffic signal control. As the conventional RL, [BA] is a baseline model including the prioritized replay experience (PER) and Q-network. Building on [BA], the three RL variants [R-ARO], [R-AD] and [R-AS] employ different reward shaping functions. As the proposed model, [ARL] employs the potential-based reward shaping function with the 3^{rd} party advisor elaborated in Section 5.2.2. Reward shaping function basically uses the potential Φ; see Eq. (5.16). Therefore, [R-ARO], [R-AD] and [R-AS] define the potential as the average relative occupancy (ARO), average delay (AD) and average vehicle speed (AS) in

each control period, respectively. Therefore, the defined potentials in the reward shaping function are as follows:

$$\Phi(s) = \underbrace{\frac{ARO_a}{\sum ARO}}_{\text{[R-ARO]}} \quad \text{or} \quad \underbrace{\frac{AD_a}{\sum AD}}_{\text{[R-AD]}} \quad \text{or} \quad \underbrace{\frac{AS_a}{\sum AS}}_{\text{[R-AS]}} \tag{6.17}$$

6.2 Results and discussion

With the real-world traffic network in West of Glasgow simulated using the LWR hydrodynamic model, this thesis evaluates the performance of different RL methods under different scenarios. The proposed framework is trained in 1,000 episodes. The reward is cumulated in an episode. As mentioned in Section 6.1.4.1, the RL-based optimization has two goals: the theoretical objective (reward) for the RL and practical objective (vehicle delay and throughput) for the signal control problem. To achieve these objectives in our framework, through the experiments, the reward and the vehicle throughput are maximized, and the vehicle delay is minimized in each episode by adjusting the phase green times in a full cycle.

All the RL methods, including the proposed one, are tested in 50 episodes with different random seeds (which are different from those in the training episodes), the results are averaged and summarized in Table 6.2. In addition, through [R-ARO], [R-AD] and [R-AS], which use different state variables in the potential, the impact of the different reward shaping functions on their overall performances is analyzed.

6.2.1 Overall performance and comparison with benchmarks

As mentioned in section 5.2.1.1, four state variables, namely Average delay per vehicle (AD), Average Relative Occupancy(ARO), Average vehicle Speed(AS) and Weighted Sum(WS), are considered in this experiment. These variables can be computed based on the LWR hydrodynamic traffic simulation described in Section 6.1.1.

Table 6.2 shows the average performance (vehicle delay and throughput) of the five benchmark models ([GCC], [BA], [R-ARO], [R-AD] and [R-AS]), as well as the relative improvement (%) of the proposed model ([ARL]) compared to the benchmarks. The results clearly show that the proposed method outperforms the benchmarks in terms of both vehicle delay and throughput.

To evaluate different reward shaping functions, [R-ARO], [R-AD] and [R-AS] are analyzed. The performances of three scenarios are similar, and slightly better than that of [BA]. That is, well-designed reward shaping functions can indeed improve the performances (Mannion, 2017), although such improvement is not pronounced according to our experiment. On the other hand, comparing [ARL], which introduces the 3^{rd} party advisor, with [BA], [R-ARO], [R-AD] and [R-AS] reveals that the 3^{rd} party advisor makes a significant improvement in the model's performance.

Table 6.2: Statistical summary of the mean performances of different scenarios compared to the proposed [ARL] model.

Scenario	Vehicle Delay (Unit: sec)	Throughput(Unit: vehicle)
[ARL]	64.68 / -	2,473 / -
[GCC]	120.99 / 46.53%	2,349 / -5.28%
[BA]	88.84 / 27.18%	2,370 / -4.36%
[R-ARO]	71.38 / 9.38%	2,437 / -1.47%
[R-AD]	75.65 / 14.49%	2,416 / -2.35%
[R-AS]	88.06 / 26.54%	2,379 / -3.97%

In Table 6.3, we compare the performances of different signal control methods under varying levels of demand (i.e. network congestion). We achieve this by increasing the demand by 10% or 20% for all relevant source links of the network. It can be seen that the superiority of ARL over benchmarks, in terms of delay, is more pronounced when the network becomes more congested, but the benefit of ARL in network throughput decreases under 20% demand increase. This is due to the fact that in such a case the network becomes severely congested (oversaturated), and not much improvements can be done on the network throughput but local delays.

Table 6.3: Comparison analysis of demand increase (10% and 20%) over each models The percentages indicate the relative increases compared to the [ARL].

		0%	10%	20%
[ARL]	Vehicle Delay	64.68 / -	81.47 / -	99.95 / -
	Throughput	2,473 / -	2,735 / -	2,982 / -
[BA]	Vehicle Delay	88.84 / 27.18%	122.65 / 33.57%	143.06 / 30.13%
	Throughput	2,370 / -4.36%	2,669 / -2.45%	2,902 / -2.78%
[R-ARO]	Vehicle Delay	71.38 / 9.38%	102.21 / 20.28%	120.94 / 17.35%
	Throughput	2,437 / -1.47%	2,671 / -2.37%	2,916 / -2.26%
[R-AD]	Vehicle Delay	75.65 / 14.49%	88.34 / 7.77%	106.09 / 5.79%
	Throughput	2,416 / -2.35%	2,601 / -5.13%	2,828 / -5.44%
[R-AS]	Vehicle Delay	88.06 / 26.54%	102.23 / 20.3%	121.25 / 17.56%
	Throughput	2,379 / -3.97%	2,597 / -5.31%	2,848 / -4.7%

6.2.2 Convergence speed

In order to evaluate the learning speeds of different RL methods, we focus on cumulative reward in the training phase. If the cumulative rewards obtained in a sequence of training episodes reach the maximum level within a relatively small number of episodes, then the learning speed is considered faster. This has some important practical implications, as the real-world implementation of RL-based signal controller requires feedbacks from the traffic system of interest, and the faster the algorithm converges to the optimal level, the less running cost it imposes to the real-world network.

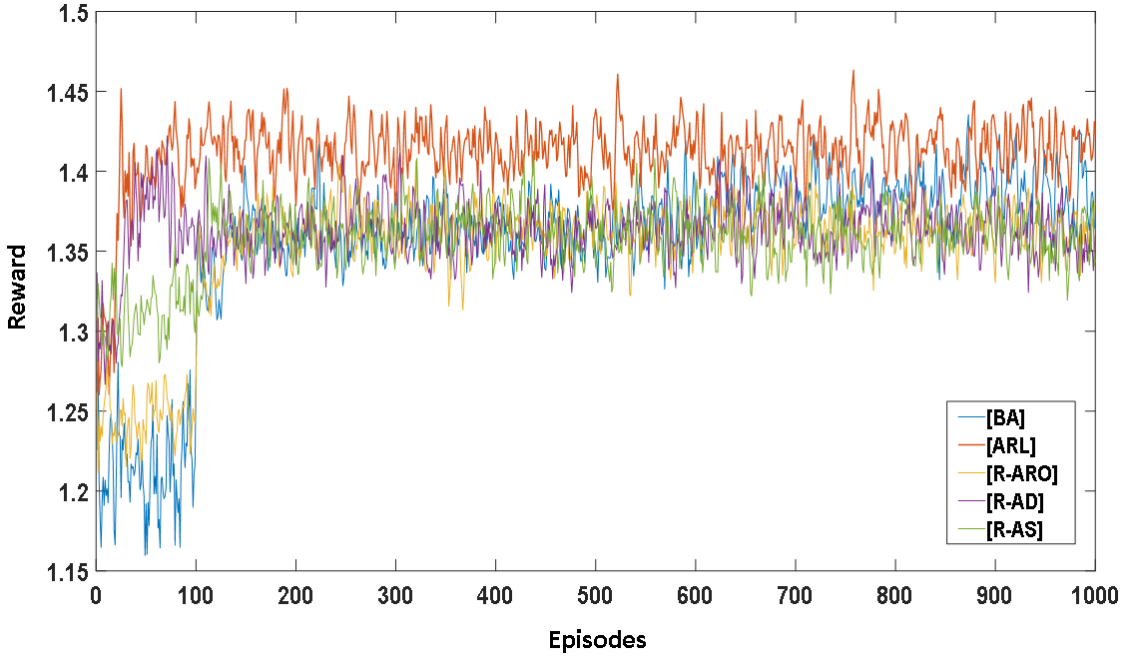


Figure 6.2: Cumulative rewards in all episodes.

Figure 6.2 shows the cumulative rewards obtained at different training episodes. For better visualization, we apply curve-smoothing technique (moving average) to these results and obtain Figure 6.3. From both figures we clearly see that ARL outperforms the other RL models in terms of mean cumulative rewards upon convergence, and reaching the level of optimal performance with minimum number of episodes. In particular, the cumulative reward is mostly greater than 1.4, and such performance is achieved within 50 episodes. After 100 episodes, the cumulative rewards are stabilized with local variations between 1.38 and 1.44. The rewards of [R-ARO], [R-AD] and [R-AS] stay in the range between 1.33 and 1.38, and it took

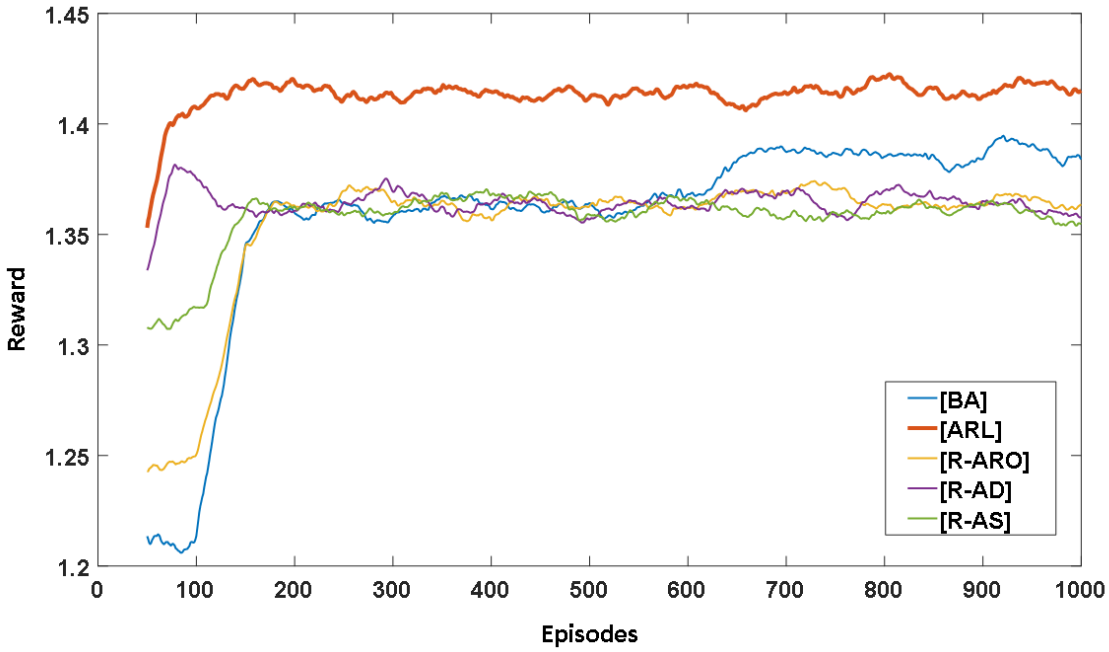


Figure 6.3: Cumulative rewards (smoothed via moving average with span= 50 for better visualization) in all episodes.

50+ more episodes to reach those levels.

6.3 Summary

In this chapter, the advanced reinforcement learning (ARL) signal control model, which incorporates reward shaping and 3^{rd} party advisor, is evaluated in terms of average vehicle delay and network throughput and compared with five benchmark models using LWR-based stimulations of a real-world test network in Glasgow. By comparing [BA] (a baseline RL model) with [R-ARO], [R-AD] and [R-AS] (RL models incorporating reward shaping), it is proven that reward shaping function works well and quickly learn the optimal (near optimal) action (phase green time duration). Moreover, ARL shows considerable superiority (in terms of vehicle delay and network throughput) over [BA], [R-ARO], [R-AD] and [R-AS], proving that the 3^{rd} party advisor significantly enhances the performance of the signal controller, which is also more effective than reward shaping. In addition, as shown in Figures 6.2 and 6.3, ARL achieves much faster convergence to the optimal level of rewards than the other methods, which demonstrates the potential benefits of the 3^{rd} party advisor in that it helps the signal controller to achieve the desired level of performance with much

fewer iterations, saving running cost in a real-world deployment scenario.

Chapter 7

Impact of information availability and quality

This chapter investigates the impact of information availability and quality on the performance of different machine learning based responsive signal controls, including the proposed NDR and ARL frameworks.

In recent year, with advanced technology including vehicular sensors and network communication, various state information on the urban traffic network has been generated, but data completeness is still not guaranteed due to environmental factors that cause sensor malfunction or failure. Relevant work in the literature address the effectiveness of signal controls by making idealized assumptions regarding information availability. For example, Gao et al. (2017), Aziz et al. (2018) and Lin et al. (2018) adapt signal control parameters based on vehicle position, vehicle speed, and the number of vehicles stopping at each link, that are instantaneously available without any uncertainty (e.g. arising from sensing error or sampling granularity) or missing data (e.g. arising from sparsely distributed sensors). Moreover, existing studies of signal control based on reinforcement learning assumes complete information on the state variables (e.g. queue length, flow, speed) throughout the network (El-Tantawy et al., 2013, Wei et al., 2018, Mannion et al., 2015).

In the real-world implementation of responsive signal control strategies, the aforementioned methods are challenged in three aspects:

- little understanding exists regarding the most appropriate type of variable (e.g. flow, speed, queue length, travel time, etc.) to capture real-time traffic states,

such that certain machine learning algorithms are most effective in managing traffic with such variables. While a wide range of traffic state variables have been investigated in the literature, no comparison exists between these state variables as the source of information for different responsive control strategies.

- traffic state variables are not uniformly available on all links/junctions, due to actual sensor locations and working conditions. In a real-world network, it is likely that the desired traffic data are missing from certain links or junctions, resulting in uncertainty when describing the network's state. The decision making capabilities of different signal control methods under such uncertainty has not been thoroughly investigated.

- traffic observations are associated with certain levels of uncertainty arising from either sensing error or systematic error (e.g. arising from sampling methods or communication capacities). Different machine learning architecture (e.g. deep neural networks) reacts differently to this kind of input uncertainties, leading to unknown performance of the signal controls. The performances of different real-time signal control algorithms are less understood in this regard.

To address these challenges, this chapter presents a comprehensive study of the impact of data availability and quality on different machine learning based signal control frameworks. This includes

1. a comparison between different state variables (ARO, AD, AS), see Section 7.1;
2. impact analysis of data noise with different signal-noise ratio, see 7.2; and
3. investigation of model performance with missing data, see Section 7.2.2.

The remainder of this chapter reports assessment and comparative results for the proposed ARL and NDR models, which are obtained via LWR-based simulation (see Section 6.1.1) performed on Matlab(R2017b) platform. The experiment setup follows that in Chapter 6.1. In particular, ARL and other RL-based methods are trained with 1,000 episodes and tested with 50 different episodes. The NDR methods (including NDR with FFNN and RNN) are trained with 45 iterations and tested with 50 different random seeds.

7.1 State selection

To generate the appropriate action corresponding to dynamic traffic states, selecting the right state variables is a critical issue that is related to the size of the state space and the learning efficiency of the model . Many researchers use various state variables (such as number of vehicle, delay time, vehicle position, queue length etc.) for state representation. However, if the state selection is not done in an optimal way, the large amount of state information might cause the divergence of learning of the traffic network (Casas, 2017). In addition, according to different state information such as Average Delay (AD), Average Relative Occupancy (ARO), Average Speed (AS) and their combinations, the performances of the RL models can vary significantly (Genders & Razavi, 2018). Therefore, by comparing the proposed NDR and ARL frameworks with other benchmarks mentioned in 6.1.4.2, this section evaluates their performances based on different selection of the state variable. Two versions (based on FFNN and RNN, respectively) of the proposed NDR framework are considered:

- [**NDR-1**] : the NDR framework based on FFNN
- [**NDR-2**] : the NDR framework based on RNN

Figure 7.1 shows the impacts of all-state selection and single-state selection. Overall, the proposed frameworks (including [ARL], [NDR-1] and [NDR-2]) outperform other benchmark models. In all four scenarios, the proposed models, including ARL and NDR, outperform the other benchmarks in terms of both vehicle delay and network throughput. Across all scenarios, ARL slightly outperform NDR, and the performances of NDR-1 and NDR-2 are similar. When it comes to the state selection, all the responsive signal controls perform best when considering all three states (AD, AS, ARO) simultaneously. However, their performances vary significantly when a single state variable is applied. In particular, BA, R-ARO, R-AD and R-AS perform the worst based on average delay as the state variable, which suggests that these models may not work well with data related to link travel times (e.g. data collected from GPS devices, cameras and virtual trip lines). When using average speed or average relative occupancy, the benchmarks have worse performance than the fixed signal timing (GCC). In contrast, the performances of ARL and NDR are stably when using single state variables, indicating that they are robust against different choices of state variables. In addition, AD and ARO are the most effective single

state variables for ARL and NDR, whose performances are comparatively worse when AS is involved.

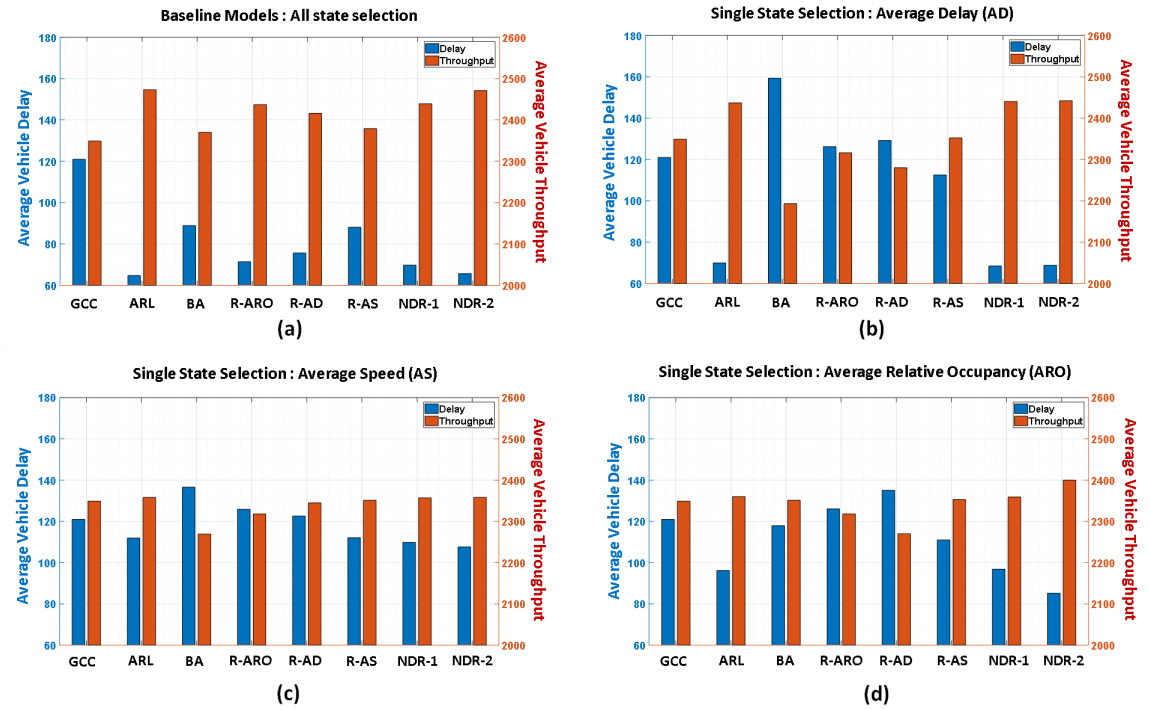


Figure 7.1: Performances of using all state variable(a) and single state variable(b,c,d) in our models and benchmark models (X-axis: Each scenario, Y-axis(Left): average delay per vehicle(unit: sec) and Y-axis(right): average vehicle throughput(unit: veh))

In Figure 7.2, we test different combinations of the state variables. The performances of the benchmarks are more steady and consistent when using two state variables instead of one, and all the responsive controllers outperform the fixed-timing control (GCC). The proposed ARL and NDR outperform BA, R-ARO, R-AD and R-AS in the case of AD+AS, and such improvement is positive yet less significant in ARO+AS and AD+ARO.

Overall, Figures 7.1 and 7.2 suggest that using more state variables tend to enhance the performances of all the responsive signal controls. The proposed ARL, NDR-1 and NDR-2 outperform the benchmarks regardless of the type of state variables or their combinations.

In Figure 7.3, we present, in each subfigure, the performance of a given signal controller with different combinations of state variables. We see that ARL, NDR-1 and NDR-2 experience the minimum deterioration of the KPIs when the state variables vary; namely the difference in delays are within 25 s (ARL), 22 s (NDR-1)

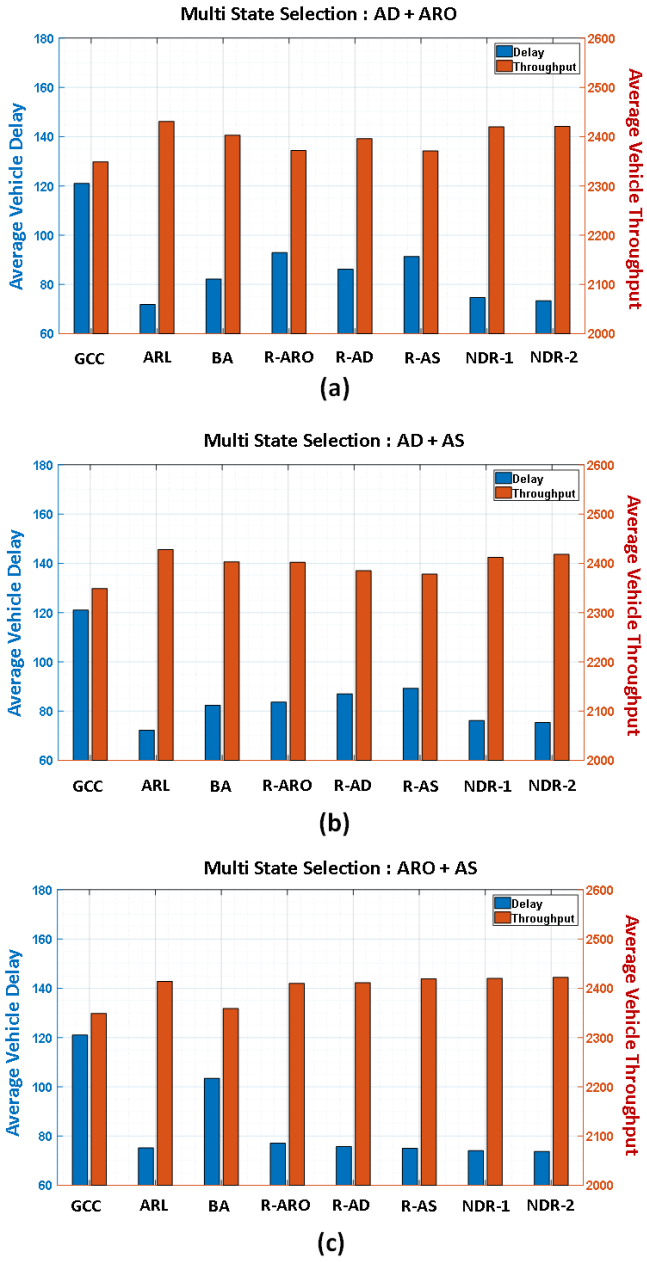


Figure 7.2: Performances of using two state variable in our models and benchmark models (X-axis: Each scenario, Y-axis(Left): average delay per vehicle(unit: sec) and Y-axis(right): average vehicle throughput(unit: veh))

and 23 s (NDR-2), while the difference in throughputs are within 170 veh (ARL), 130 veh (NDR-1) and 120 veh (NDR-2). On the other hand, the other methods result in either significant deterioration in delays (up to 84 s) or throughput (up to 300); see

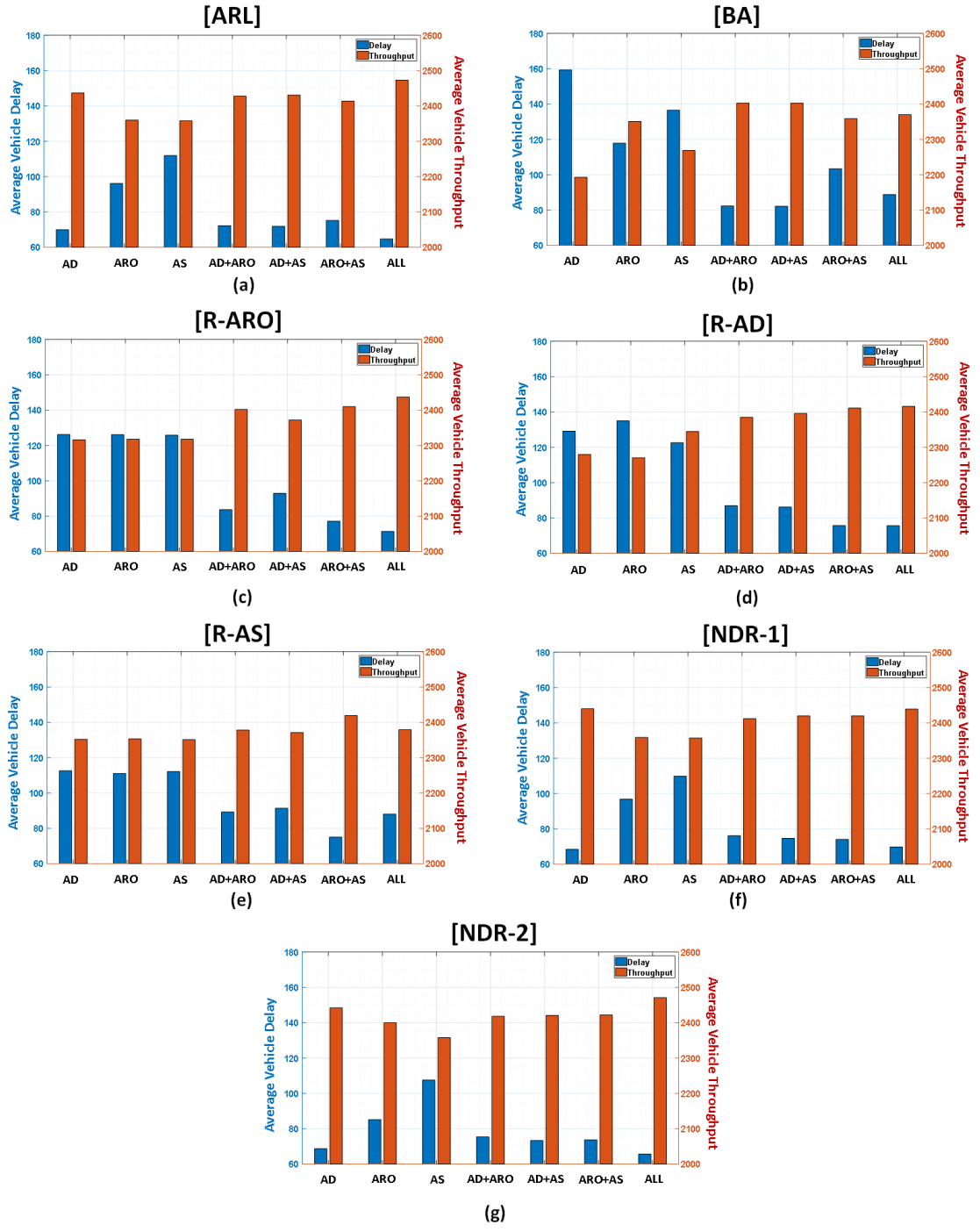


Figure 7.3: Re-arranged performances(including vehicle delay and vehicle throughput) according to the scenario (X-axis: Each case, Y-axis(Left): average delay per vehicle(unit: sec) and Y-axis(right): average vehicle throughput(unit: veh))

Table 7.1. In particular, NDR and BA seem to be the most and least robust signal controllers against different choices of the state variables.

	ARL	BA	R-ARO	R-AD	R-AS	NDR-1	NDR-2
Delay (s)	25	84	46	55	33	22	23
Throughput (veh)	170	300	170	210	120	130	120

Table 7.1: Maximum deterioration of average vehicle delay or network throughput when the state variable(s) vary.

For ARL and NDR, the most effective single state variable(s) are AD or ARO. When it comes to the combination of two state variables, no discernible difference can be seen for these proposed models, and the gap with all three state variables ('ALL' in Figure 7.3) is quite insignificant.

7.2 Information imperfectness and incompleteness

In a real-world operational environment of signal control, external conditions such as low spatial sensor coverage or sensing error could lead to low availability of traffic data that are crucial for calculating state variables. This gives rise to the notions of data completeness and perfectness, which are directly related to data quality and can present the totality of characteristic of the data to accomplish a given goal (Kaisler et al., 2013). Among the many instances of data incompleteness/imperfectness, we consider missing data (for certain links of the network) and data uncertainty (in the form of additive noises present in the data). Therefore, this section aims to test the proposed responsive signal control frameworks as well as several benchmarks in these conditions.

7.2.1 Data noise

To represent uncertainties in the traffic data that might be caused by systematic or random errors, we use an additive noise term to the three state variables:

$$\tilde{\text{ARO}} = \text{ARO} + \varepsilon_1, \quad \tilde{\text{AD}} = \text{AD} + \varepsilon_2, \quad \tilde{\text{AS}} = \text{AS} + \varepsilon_3$$

Here, each ε_i is a random vector of appropriate size comparable with the dimension of the relevant state variable, and the individual elements of ε_i follow i.i.d Normal

distribution $N(0, \sigma_i^2)$ where the standard deviation σ_i is chosen to be 5%, 10% and 20% of the mean value of the state variable (i.e. the signal). These different noise-to-signal ratios allow us to test the performance of different models under different levels of data uncertainty.

Figure 7.4 shows the performances (in terms of vehicle delay and network throughput) under different signal control measures in the presence of data noises (5%, 10% and 20% noise-to-signal ratios). Overall, the models [ARL], [NDR-1] and [NDR-2] outperform other benchmark models in almost all cases except one (AS with 20% noise), which shows that the proposed models have superior robustness over the benchmarks when the input of the controller is perturbed with noises. Among the three state variables, the proposed models have the least satisfactory performances under average speed (AS), which is consistent with the findings in Figure 7.1 (ARL and NDR perform better with AD and ARO than AS).

The performances of R-ARO, R-AD and R-AS are susceptible to significant (20%) perturbations of the input state variables. Interestingly, in the case where noise is added to state variable AS, the benchmark model [R-AD] has the worst performances rather than [R-AS]. That is, in the case of [R-AS], the reward shaping function in the benchmark model [R-AS] efficiently deals with the noise. As a result, the benchmark model [R-AS] can avoid the rapid decline in performances. On the other hand, [R-AD] does not have good performances due to inaccurate state information, even if the reward shaping function works properly.

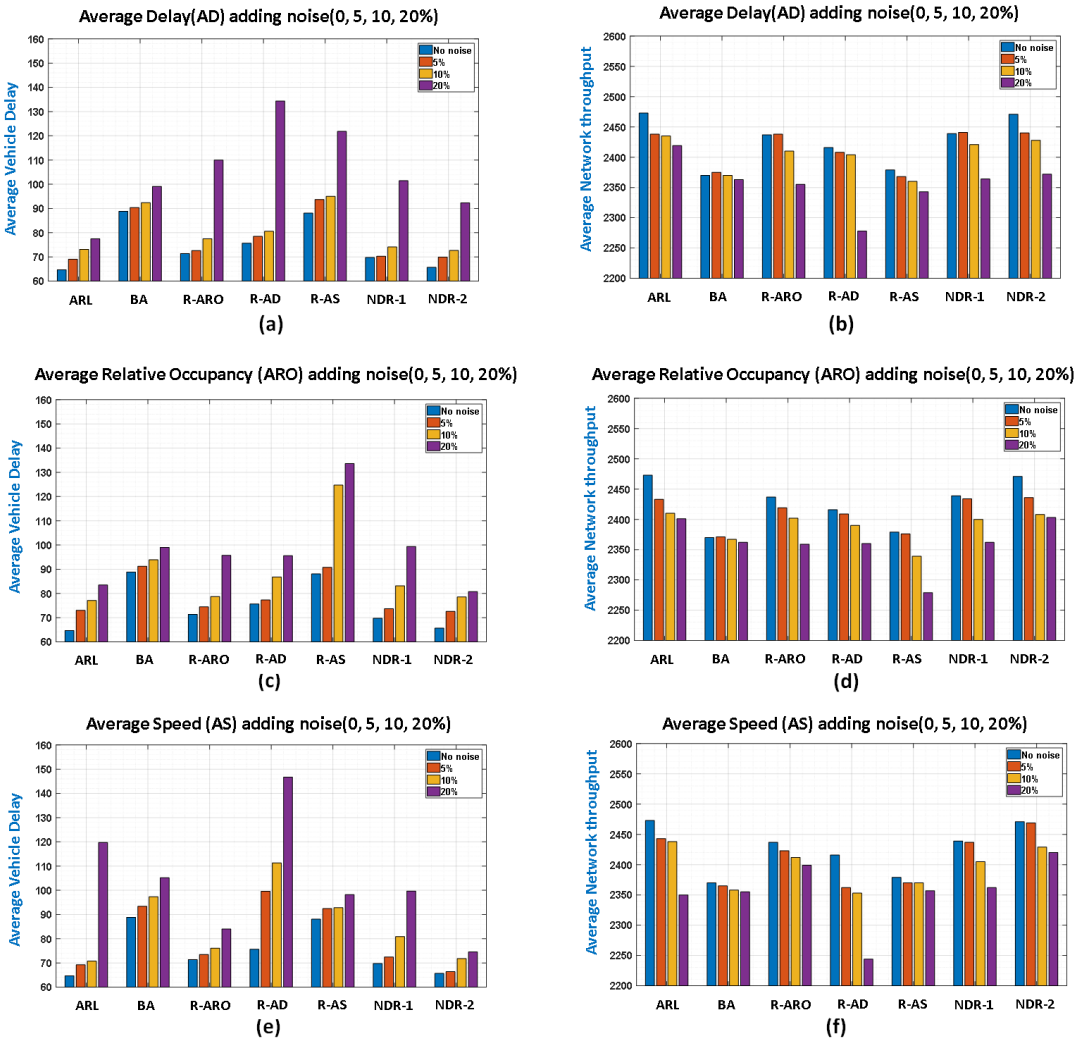


Figure 7.4: Performances of different signal control methods with noisy state variables (5%, 10% and 20% noise-to-signal ratio).

7.2.2 Missing information

In a real-world traffic network, due to limited sensor deployment or sensing failure, it is not possible to obtain traffic data on all the links in the network. When traffic information is not available on certain links in the network, the traffic states on those links become unknown, which poses challenges to responsive signal controllers.

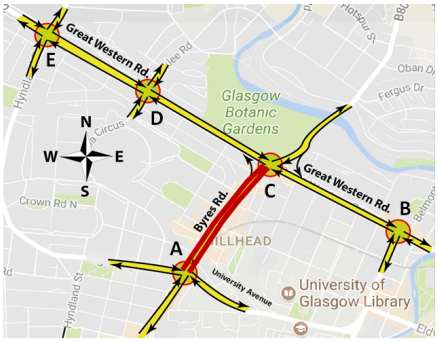
In this experiment, four state variables (Average Relative Occupancy, Average Delay, Average Speed and Weighted sum) mentioned in Section 5.2.1.1 are considered in the reinforcement learning. we consider three scenarios where traffic states are unknown on certain links. As illustrated in Figure 7.5(a-1), (b-1) and (c-1), data on Byres Road and Great Western Road are missing. In particular, we define the

following three scenarios:

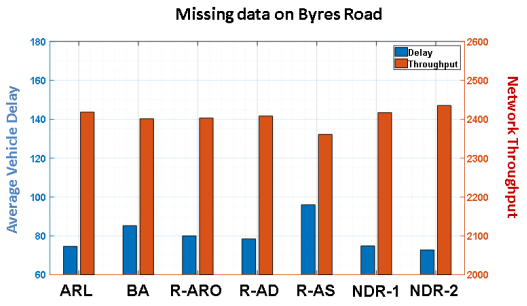
1. Missing data on Byres Road (both directions),
2. Missing data on Byres Road (south bound) and University Avenue (west bound),
3. Missing data on the Great Western Road (both directions between junctions C and D, and west bound direction between junctions D and E).

These scenarios are chosen as Byres Road and Great Western Road are considered highly congested, carrying significant amount of traffic approaching the city center from the west and south in morning peaks (see Section 4.1.4.1). A lack of state variables on these critical links therefore poses significant challenge to the responsive signal controllers. This section investigate the level of impact on the performances of different signal control methods in terms of vehicle delay and network throughput.

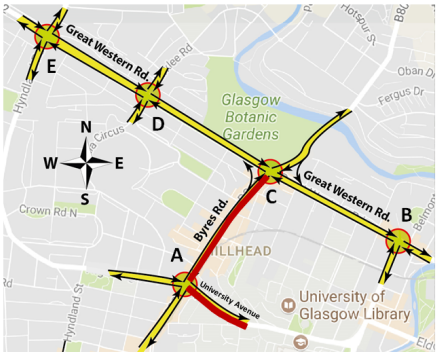
The signal control performances are shown in the right column of Figure 7.5. Overall, the methods [ARL], [NDR-1] and [NDR-2] outperform other benchmark models in all three scenarios. In particular, [NDR-1] has the best performance in the case (c-2). Figure 7.6 compares the performances of different signal controllers before (non-missing) and after removal of information. Similar to Figure 7.5, ARL and NDR perform consistently well compared to the benchmarks. It is noted that the three proposed models (ARL, NDR-1, NDR-2) are most affected in the "Byres Road + University Avenue" case, while the impact of missing data on the Great Western Road is relatively minor. These phenomena are caused by the different flows on these links and are manifestation of highly complex traffic dynamics and decision processes. However, such findings provide valuable insights on sensor location problem when the traffic is managed in real time by AI-based signal controls.



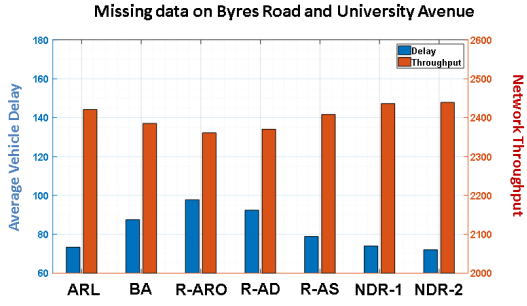
(a-1)



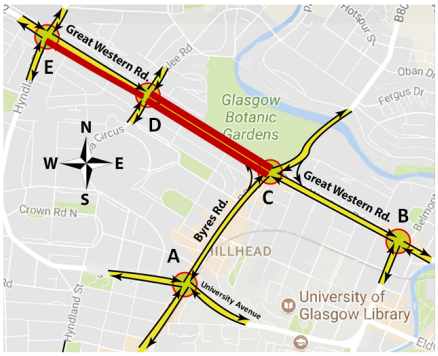
(a-2)



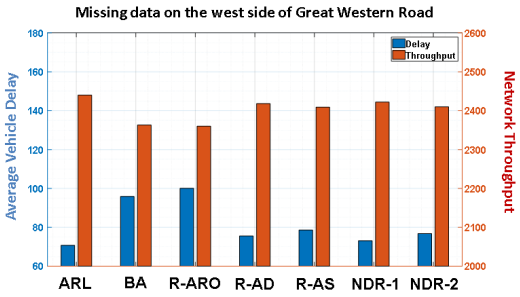
(b-1)



(b-2)



(c-1)



(c-2)

Figure 7.5: Performances of different signal control models in the presence of missing data. The left column illustrates the links with missing data in red bold curves. The right column indicates the corresponding performances of various signal control models in terms of vehicle delay (unit: sec) and network throughput (unit: veh)

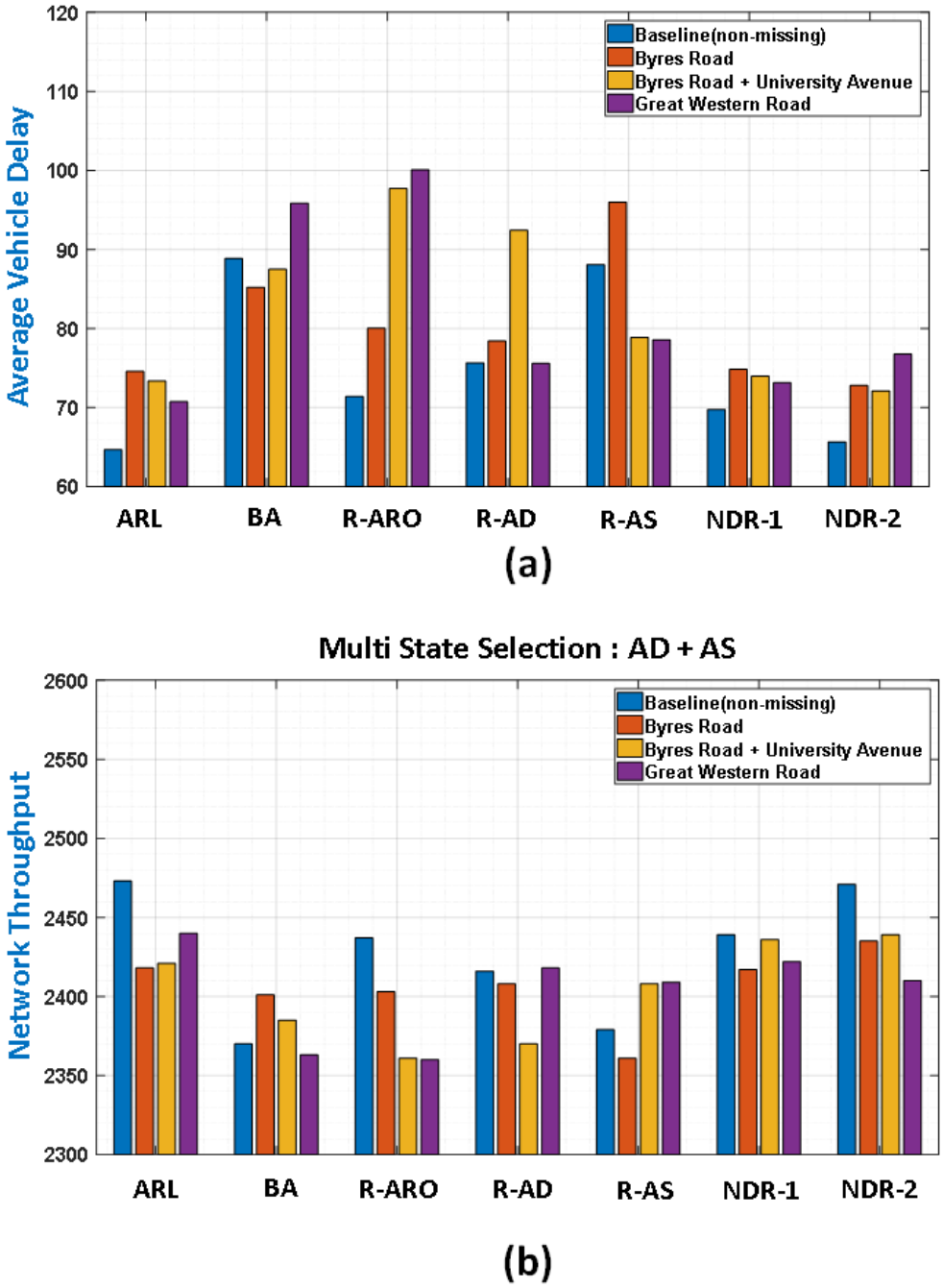


Figure 7.6: Performances in terms of vehicle delay (a) and network throughput (b) of different signal control models before (non-missing) and after (others) data removal on certain links.

7.3 Impact of different types of data quality issues

As a summary of the previously presented instances of data quality issues, namely selection of state variables, data noise and unknown states on parts of the network,

we compare their impacts pertaining to the signal control frameworks mentioned in this chapter.

When different combinations of state variables are used, the ranges of deterioration of the proposed methods are

- ARL: Delay: 7s - 25s; Throughput: 110 - 170 veh
- NDR-1: Delay: 4s - 22s; Throughput: 70 - 130 veh
- NDR-2: Delay: 6s - 23s; Throughput: 40 - 120 veh

When different levels of noises are applied to the state variables (AD or ARO), the ranges of deterioration of the proposed methods are:

- ARL: Delay: 10s-12s; Throughput: 120 - 130 veh
- NDR-1: Delay: 13s - 14s; Throughput: 100 -135 veh
- NDR-2: Delay: 6s - 8s; Throughput: 60 veh

When three scenarios are considered concerning the missing data on several links of the network (see Figure 7.5), the ranges of deterioration of the proposed methods are:

- ARL: Delay: 6s - 8s; Throughput: 100 - 110 veh
- NDR-1: Delay: 0.5s - 4s; Throughput: 10 - 90 veh
- NDR-2: Delay: 2s - 4s; Throughput: 0 - 60 veh

It can be seen that the different types of data quality issues have relatively minor impact on the performance of the proposed signal control measures. All things considered, NDR-2 (NDR with recurrent neural network) is the least impacted method among the three.

7.4 Summary

This chapter investigates the impact of data incompleteness and imperfectness on the performance of a range of real-time signal control methods. This is driven by the observation that most responsive signal control models in the literature, derived

either analytically/theoretically or via simulations, tend to make idealized assumptions on the quality of data, which gives rise to accurate and perfect information on the traffic states. These assumptions are challenged in this chapter, and we show the varying levels of performance deterioration of different signal controllers including the proposed ARL, NDR-1 and NDR-2, as well as benchmarks models such as the baseline reinforcement learning and RL with different reward shaping functions.

Three types of scenarios are considered, namely different selections of the state variables, data uncertainty/noise, and unknown states on parts of the network. It is found that the proposed ARL and NDR models perform consistently well compared to the benchmarks, despite the deterioration of data quality in the decision-making process.

The ranges of performance deterioration for the proposed methods (ARL, NDR-1, NDR-2) are summarized in Section 7.3. In terms of state variables, average delay (AD) or average relative occupancy (ARO) are shown to be suitable ones to represent traffic states when integrated with the ARL or NDR frameworks. When noises are applied to the state variables, the performance of the three methods remain stable with relatively minor deterioration. When the traffic state information is removed from certain links of the network, different methods are impacted and their effects vary, which is likely due to the different neural network structure inherent in these responsive signal control frameworks.

Chapter 8

Comparison, Recommendation, Conclusions and future research

This thesis addresses the four main challenges in responsive traffic signal control, namely

1. Uncertainty in the traffic network;
2. Multi-objective optimization;
3. Computational efficiency; and
4. Data availability and quality

by leveraging machine learning techniques integrated with optimization-based training procedures to yield timely decisions that optimizes traffic network performance in a centralized and reliable way. In addition, with strength points to the proposed frameworks in this thesis, this section first address the key difference from NDR/ARL with existing systems like SCOOT/SCAT. Additionally, this thesis provide a discussion on the pros and cons of Nonlinear Decision Rule(NDR) and Advanced Reinforcement Learning(ARL), and offer some suggestions when it comes to their deployment.

8.1 Comparison with Existing traffic system

NDR and ARL can be efficiently used to mitigate traffic congestion and reduce exhausted emission. As the responsive traffic signal control framework, the both

frameworks efficiently handle uncertainty in the traffic network, keep the balance between traffic and environmental objectives, achieve computational efficiency, and have a strong guarantee related to the issue of the data availability and quality. In order to recommend our frameworks, we need to address the key difference from NDR/ARL with existing systems like SCOOT.

First, unlike SCOOT or SCAT using only numerical data(especially, vehicle count and queue length), our framework can be applied with different type of data provided by various traffic sensors(like image-based sensors). That is very important because, with advanced technology, there are many state-of-the-art sensors to more accurately and efficiently detect events/vehicle's behaviours on the roads. The existing traffic signal system using SCOOT has a limitation to use the different type of traffic data since it requires only specific data(like vehicle counts and queue length). Our frameworks has a strong flexibility of using various type of traffic data(like imagery traffic data). That is, the proposed frameworks can be applied to varoius traffic networks in which the SCOOT/SCAT cannot be applied.

Second, existing systems(such as SCOOT and SCAT) only focus on the optimization of the vehicle delay and/or network capacity and show the efficiency of controlling traffic flow in real-life. But, these systems do not consider environmental perspective. As you can see the section 4.1.4.3, vehicle delay and Carbon Dioxide(CO_2) emission have the high correlation, but low correlation between vehicle delay and Black Carbon(BC). Cho (2019) mentioned that BC directly affects to global warming as a main factor of air pollutants. Many governments and researchers have recently paid attention to environmental pollution so that environmental perspectives might be required to traffic systems. In this thesis, our frameworks can consider two exhausted emissions(CO_2 and BC), but the two frameworks has a flexibility of considering other emissions. Therefore, our framework can efficiently solve air pollution.

Third, SCOOT is based on fixed traffic signal timing which can efficiently control the traffic in common situation. But, the fixed signal plan does not cover specific traffic condition or events(like festival or road closure). However, our frameworks based on machine learning techniques(including Neural Networks(NN) and Reinforcement Learning(RL)) can responsively control traffic flows in urban traffic network so that these frameworks can efficiently handle the uncertainties occurred from the roads.

In addition, the fixed timing cannot harmonize the traffic conditions of other intersections. As a result, although traffic flow can be optimized in a certain intersection, but it can cause serious conflicts or accidents. On the other hand, our frameworks can perform the network-level traffic signal control. So, these frameworks can consider traffic conditions among each intersections. As a result, they can avoid traffic conflicts effectively.

8.1.1 Recommendation

Two proposed frameworks in this thesis can efficiently be employed according to different conditions on traffic network, due to the main characteristic of the frameworks.

First, NDR is the rule-based framework. So, the NDR framework provides optimal traffic signal control strategy to the specific traffic network, which has a key capability to effectively find the optimal real-time traffic plan. If NDR is applied however, off-line optimisation should be performed to find the optimal traffic signal strategies corresponding to the different traffic network. But, the on-line implementation of the trained NDR framework is quite efficient and can accommodate real-time decision requirements. In addition, due to the nature of built-in neural network in the NDR framework, the more experience the NDR framework has with other traffic networks, the faster and easier it is to find the optimal traffic signal strategy.

Second, RL is the learning-based framework to control traffic signal. Although the RL framework can find the "near-optimal" traffic signal plan, it can more effectively and quickly react the unexpected events(including accidents, temporal road closure and any events). In addition, when it comes to different traffic networks, without any off-line implementation, the framework can find the efficient traffic signals based on the framework's experience and knowledge. Therefore, the framework has an excellent applicability to different traffic networks.

Therefore, according to urban planning of the government, if the government wants to optimize the traffic signal strategy only in a specific city, NDR is more useful. If each traffic signal control system of multiple cities need to be developed at the same time, RL can intensely reduce the development period because of the above-mentioned characteristic of the RL.

8.2 Main contributions

The four main challenges have been addressed in this thesis, as follows.

- Uncertainty in the traffic network

  - Unlike many stochastic optimization approaches where a priori distributions are known, the uncertainty of traffic network dynamics is difficult to characterize and calibrate. This has been addressed in the proposed NDR framework via a Monte-Carlo type simulation-based optimization approach, which is shown to effectively handle within-day and day-to-day variations of traffic quantities. Similarly, the ARL approach can account for the stochasticities in the state variables by deriving the optimal control policy. As a result, the resulting real-time controls are robust against traffic uncertainties, as shown in relevant simulation tests.

- Multi-objective problem

  - This thesis addresses traffic related objectives including vehicle delay, network throughput, and exhaust emissions such as CO_2 and black carbon. These objectives are incorporated in the training procedure of the machine learning models via a feedback loop based on traffic simulations. The test results of the proposed models demonstrate the capability of the machine learning based signal controllers to effectively balance different objectives. In addition, further analysis reveal the potential alignment and conflict of traffic and environmental objectives, which offer insights into the management of dynamic traffic networks for sustainability.
  - When the level of network demand increases (i.e. the network becomes more congested), the key performance indicators of signal controls are affected in different ways. Specific findings can be found in relevant sections.

- Computational efficiency

  - Both NDR and ARL frameworks tackle the challenge of excessive computational demands in an on-line optimization environment. Through

off-line training procedure that fully takes into account the possible realization of uncertain traffic states, the resulting real-time controls are guaranteed to maintain a satisfactory level of performance with little computational burden, which allows real-time decisions to be made.

- To reduce the size of the traffic state space and allow timely decisions to be made without consuming a vast amount of traffic information, this thesis investigate performances of NDR and ARL according to different selection and combination of traffic state variables. This allows the machine learning model to conducted supervised learning, as informed by realized traffic objectives through simulation, in order to optimize the configuration of the neural networks.

- Information availability and quality

  - the impact of data incompleteness and imperfectness on the performance of a range of real-time signal control methods is investigated. This is driven by the observation that most responsive signal control models in the literature, derived either analytically/theoretically or via simulations, tend to make idealized assumptions on the quality of data, which gives rise to accurate and perfect information on the traffic states. These assumptions are challenged in this research, and it is shown that the varying levels of performance deterioration of different signal controllers including the proposed NDR and ARL methods, as well as several benchmarks models.

  - In terms of state variables, average delay (AD) or average relative occupancy (ARO) are shown to be suitable ones to represent traffic states when integrated with the ARL, NDR-1, or NDR-2 frameworks. When noises are applied to the state variables, the performance of the three methods remain stable with relatively minor deterioration. When the traffic state information is removed from certain links of the network, different methods are impacted and their effects vary, which is likely due to the different neural network structure inherent in these responsive signal control frameworks.

8.3 Research limitation and future research

The research methodologies and/or findings may be extended in the following ways.

1. application of the proposed signal control frameworks to different types of traffic networks (with varying size, configuration and flow characteristics), in order to demonstrate the transferability and generality of the findings, despite the widely-held belief that machine learning based control models should work properly in different environments given sufficient data and training resources.

2. integration with localized (decentralized) coordination (such as offset optimization) and actuation (such as a transit priority signal) to further reduce emissions at local junctions;

3. a systematic and quantitative approach to optimal sensor location that is compatible with the proposed neural network configuration; and

4. an in-depth (theoretically or experimentally) investigation of the matching between the structure and depth of the neural networks, and the complexity of the underlying control system. Such a priori knowledge could offer valueable insights into the construction of nonlinear decision rules and tuning of their parameters for better efficiency and less redundancy.

References

Abdoos, M., Mozayani, N., & Bazzan, A. L. (2011). Traffic light control in non-stationary environments based on multi agent q-learning. In *2011 14th International IEEE conference on intelligent transportation systems (ITSC)*, (pp. 1580–1585). IEEE. 20

Ahmed, A., Naqvi, S. A. A., Watling, D., & Ngoduy, D. (2019). Real-time dynamic traffic control based on traffic state estimation. *Transportation Research Record*. 31, 36

Araghi, S., Khosravi, A., Johnstone, M., & Creighton, D. (2013). Intelligent traffic light control of isolated intersections using machine learning methods. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, (pp. 3621–3626). IEEE. 83

Arel, I., Liu, C., Urbanik, T., & Kohls, A. (2010). Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems*, *4*(2), 128–135. 3, 8, 17, 19, 73

Aslani, M., Mesgari, M. S., Seipel, S., & Wiering, M. (2018a). Developing adaptive traffic signal control by actor–critic and direct exploration methods. In *Proceedings of the Institution of Civil Engineers-Transport*, (pp. 1–10). Thomas Telford Ltd. 19, 38, 73, 83

Aslani, M., Mesgari, M. S., & Wiering, M. (2017). Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. *Transportation Research Part C: Emerging Technologies*, *85*, 732–752. 19

Aslani, M., Seipel, S., & Wiering, M. (2018b). Continuous residual reinforcement

learning for traffic signal control optimization. *Canadian Journal of Civil Engineering*, *45*(8), 690–702. 19, 38

Aziz, H. A., Zhu, F., & Ukkusuri, S. V. (2018). Learning-based traffic signal control algorithms with neighborhood information sharing: An application for sustainable mobility. *Journal of Intelligent Transportation Systems*, *22*(1), 40–52. 3, 26, 38, 83, 103

Balaji, P., German, X., & Srinivasan, D. (2010). Urban traffic signal control using reinforcement learning agents. *IET Intelligent Transport Systems*, *4*(3), 177–188. 17, 19, 38, 73, 83

Banks, A., Vincent, J., & Anyakoha, C. (2007). A review of particle swarm optimization. part i: background and development. *Natural Computing*, *6*(4), 467–484. 47

Barisone, A., Giglio, D., Minciardi, R., & Poggi, R. (2002). A macroscopic traffic model for real-time optimization of signalized urban areas. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, vol. 1, (pp. 900–903). IEEE. 29

Barth, M., & Boriboonsomsin, K. (2008). Real-world carbon dioxide impacts of traffic congestion. *Transportation Research Record*, *2058*(1), 163–171. 2

Barth, M., & Boriboonsomsin, K. (2009). *Traffic congestion and greenhouse gases*. ACCESS Magazine. 32

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, *53*(3), 464–501. 46

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. 27

Bullinaria, J. A. (2004). Self organizing maps: fundamentals. *Introduction to Neural*. 15

Cambridge Systematics, I. (2005). Cambridge systematics, inc. In *Traffic Congestion and Reliability, Trends and Advanced Strategies for Congestion Mitigation.*, 7091.720. Federal Highway Administration. 1, 2

Casas, N. (2017). Deep deterministic policy gradient for urban traffic light control. *arXiv preprint arXiv:1703.09035*. 105

Castro, G. B., Hirakawa, A. R., & Martini, J. S. (2017). Adaptive traffic signal control based on bio-neural network. *Procedia Computer Science*, *109*, 1182–1187. 17

Chakraborty, P. S., Tiwari, A., & Sinha, P. R. (2015). Adaptive and optimized emergency vehicle dispatching algorithm for intelligent traffic management system. *Procedia Computer Science*, *57*, 1384–1393. 1, 2

Chang, H. S. (2006). Reinforcement learning with supervision by combining multiple learnings and expert advices. In *2006 American Control Conference*, (pp. 4159–4164). IEEE. 38, 74, 86

Chang, L., & Hui, W. (2016). Traffic emission control based on emission pricing and signal timing. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, (pp. 467–472). IEEE. 33

Chang, T.-H., & Lin, J.-T. (2000). Optimal signal timing for an oversaturated intersection. *Transportation Research Part B: Methodological*, *34*(6), 471–491. 27

Chang, T.-H., & Sun, G.-Y. (2004). Modeling and optimization of an oversaturated signalized network. *Transportation Research Part B: Methodological*, *38*(8), 687–707. 7, 28, 52

Chen, H., Bai, R., Ma, J., & Wang, D. (2012). Research on intersection signal timing model considering emissions effects. In *Twelfth COTA International Conference of Transportation ProfessionalsAmerican Society of Civil EngineersTransportation Research Board*. 32, 34, 53

Chen, X., Osorio, C., Marsico, M., Talas, M., Gao, J., & Zhang, S. (2015). Simulation-based adaptive traffic signal control algorithm. Tech. rep., Transportation Research Board. 34, 35, 37

Chin, Y. K., Bolong, N., Kiring, A., Yang, S. S., & Teo, K. T. K. (2011). Q-learning based traffic optimization in management of signal timing plan. *International Journal of Simulation, Systems, Science & Technology*, *12*(3), 29–35. 8, 19, 73

Chin, Y. K., Kow, W. Y., Khong, W. L., Tan, M. K., & Teo, K. T. K. (2012). Q-learning traffic signal optimization within multiple intersections traffic network. In *2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation*, (pp. 343–348). IEEE. 79

Chiu, S. (1992). Adaptive traffic signal control using fuzzy logic. In *Proceedings of the Intelligent Vehicles92 Symposium*, (pp. 98–107). IEEE. 17

Cho, R. (2019). The damaging effects of black carbon.
URL `https://blogs.ei.columbia.edu/2016/03/22/the-damaging-effects-of-black-carbon/` 118

Chow, A. H., Sha, R., & Li, S. (2019). Centralised and decentralised signal timing optimisation approaches for network traffic control. *Transportation Research Part C: Emerging Technologies*. 3, 7, 8

Christofa, E., Ampountolas, K., & Skabardonis, A. (2016). Arterial traffic signal optimization: A person-based approach. *Transportation Research Part C: Emerging Technologies*, *66*, 27–47. 7, 28, 30

Christofa, E., & Skabardonis, A. (2011). Traffic signal optimization with application of transit signal priority to an isolated intersection. *Transportation Research Record*, *2259*(1), 192–201. 6, 28, 52

Chu, T., Wang, J., Codecà, L., & Li, Z. (2019). Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*. 25

Chunxiao, L., & Shimamoto, S. (2011). A real time traffic light control scheme for reducing vehicles co2 emissions. In *The 8th Annual IEEE Consumer Communications and Networking Conference-Emerging and Innovative Consumer Technologies and Applications*. 34

D'Acierno, L., Gallo, M., & Montella, B. (2012). An ant colony optimisation algorithm for solving the asymmetric traffic assignment problem. *European Journal of Operational Research*, *217*(2), 459–469. 1, 28, 36

de Palma, A., & Lindsey, R. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1377–1399. 2

Dong, L., & Chen, W. (2010). Real-time traffic signal timing for urban road multi-intersection. *Intelligent Information Management*, *2*(08), 483. 6, 28

Dotoli, M., Fanti, M. P., & Meloni, C. (2003). Real time optimization of traffic signal control: application to coordinated intersections. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 4, (pp. 3288–3295). IEEE. 29

Dotoli, M., Fanti, M. P., & Meloni, C. (2004). Coordination and real time optimization of signal timing plans for urban traffic control. In *IEEE International Conference on Networking, Sensing and Control, 2004*, vol. 2, (pp. 1069–1074). IEEE. 29

El-Tantawy, S., Abdulhai, B., & Abdelgawad, H. (2013). Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): methodology and large-scale application on downtown toronto. *IEEE Transactions on Intelligent Transportation Systems*, *14*(3), 1140–1150. 3, 17, 19, 22, 25, 38, 73, 83, 103

El-Tantawy, S., Abdulhai, B., & Abdelgawad, H. (2014). Design of reinforcement learning parameters for seamless application of adaptive traffic signal control. *Journal of Intelligent Transportation Systems*, *18*(3), 227–245. 8

Ellis, D. (2009). Cost per hour and value of time calculations for passenger vehicles and commercial trucks for use in the urban mobility report. *Texas Transportation Institute*, (p. 7). 2

Evans, R. (2007). Central london congestion charging scheme-ex-post evaluation of the quantified impacts of the original scheme. 2

Feng, Y., Head, K. L., Khoshmagham, S., & Zamanipour, M. (2015). A real-time adaptive signal control in a connected vehicle environment. *Transportation Research Part C: Emerging Technologies*, *55*, 460–473. 29

Frejo, J. R. D., Papamichail, I., Papageorgiou, M., & De Schutter, B. (2019). Macroscopic modeling of variable speed limits on freeways. *Transportation research part C: emerging technologies*, *100*, 15–33. 2

Friesz, T. L. (2010). *Dynamic optimization and differential games*, vol. 135. Springer Science & Business Media. 27, 45

Gao, J., Shen, Y., Liu, J., Ito, M., & Shiratori, N. (2017). Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network. *arXiv preprint arXiv:1705.02755*. 8, 19, 23, 38, 73, 83, 103

Garavello, M., Han, K., & Piccoli, B. (2016). *Models for vehicular traffic on networks*, vol. 9. American Institute of Mathematical Sciences (AIMS), Springfield, MO. 27

Genders, W., & Razavi, S. (2016). Using a deep reinforcement learning agent for traffic signal control. *arXiv preprint arXiv:1611.01142*. 8, 23, 24

Genders, W., & Razavi, S. (2018). Evaluating reinforcement learning state representations for adaptive traffic signal control. *Procedia computer science*, *130*, 26–33. 25, 105

Genders, W., & Razavi, S. (2019). Asynchronous n-step q-learning adaptive traffic signal control. *Journal of Intelligent Transportation Systems*, *23*(4), 319–331. 19

Gkatzoflias, D., Kouridis, C., Ntziachristos, L., & Samaras, Z. (2007). Copert 4 user manual (version 5.0). 51

Grześ, M., & Kudenko, D. (2010). Online learning of shaping rewards in reinforcement learning. *Neural Networks*, *23*(4), 541–550. 85, 86

Guardiola, I. G., Leon, T., & Mallor, F. (2014). A functional approach to monitor and recognize patterns of daily traffic profiles. *Transportation Research Part B: Methodological*, *65*, 119–136. 13

Hájek, P. (2011). Municipal credit rating modelling by neural networks. *Decision Support Systems*, *51*(1), 108–118. 15

Han, K. (2017). Framework for real-time traffic management with case studies. *Transportation Research Record*, *2658*(1), 35–43. 3, 42, 46

Han, K., & Gayah, V. V. (2015). Continuum signalized junction model for dynamic traffic networks: Offset, spillback, and multiple signal phases. *Transportation Research Part B: Methodological*, *77*, 213–239. 44, 81, 91, 93

Han, K., Gayah, V. V., Piccoli, B., Friesz, T. L., & Yao, T. (2014). On the continuum approximation of the on-and-off signal control on dynamic traffic networks. *Transportation Research Part B: Methodological*, *61*, 73–97. 7, 28, 52, 91, 93

Han, K., Liu, H., Gayah, V. V., Friesz, T. L., & Yao, T. (2016). A robust optimization approach for dynamic traffic signal control with emission considerations. *Transportation Research Part C: Emerging Technologies*, *70*, 3–26. 33, 52

Han, K., Sun, Y., Liu, H., Friesz, T. L., & Yao, T. (2015). A bi-level model of dynamic traffic signal control with continuum approximation. *Transportation Research Part C: Emerging Technologies*, *55*, 409–431. 61

Hauser, T. A., & Scherer, W. T. (2001). Data mining tools for real-time traffic signal decision support & maintenance. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, vol. 3, (pp. 1471–1477). IEEE. 17, 18

He, Q., Head, K. L., & Ding, J. (2014). Multi-modal traffic signal control with priority, signal actuation and coordination. *Transportation Research Part C: Emerging Technologies*, *46*, 65–82. 6, 27, 29, 51

He, W., Lu, T., & Wang, E. (2013). A new method for traffic forecasting based on the data mining technology with artificial intelligent algorithms. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(12), 3417–3422. 1, 15, 36

Hirschmann, K., Zallinger, M., Fellendorf, M., & Hausberger, S. (2010). A new method to calculate emissions with simulated traffic conditions. In *13th International IEEE Conference on Intelligent Transportation Systems*, (pp. 33–38). IEEE. 34, 35, 36

Hoplaros, D., Tari, Z., & Khalil, I. (2014). Data summarization for network traffic monitoring. *Journal of network and computer applications*, *37*, 194–205. 13

Jamshidnejad, A., Papamichail, I., Papageorgiou, M., & De Schutter, B. (2018). Sustainable model-predictive control in urban traffic networks: Efficient solution based on general smoothening methods. *IEEE Transactions on Control Systems Technology*, *26*(3), 813–827. 34

Ji, Y., Hu, B., Han, J., & Tang, D. (2014). An improved algebraic method for transit signal priority scheme and its impact on traffic emission. *Mathematical Problems in Engineering*, *2014*. 32, 52

Jiao, P., Li, Z., Liu, M., Li, D., & Li, Y. (2015). Real-time traffic signal optimization model based on average delay time per person. *Advances in Mechanical Engineering*, *7*(10), 1687814015613500. 29

Jin, J., & Ma, X. (2015). Adaptive group-based signal control by reinforcement learning. *Transportation Research Procedia*, *10*, 207–216. 19, 38, 73, 83

Jin, J., & Ma, X. (2018). Hierarchical multi-agent control of traffic lights based on collective learning. *Engineering applications of artificial intelligence*, *68*, 236–248. 8

Junchen, J., & Xiaoliang, M. (2016). A learning-based adaptive signal control system with function approximation. *IFAC-PapersOnLine*, *49*(3), 5–10. 19

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences*, (pp. 995–1004). IEEE. 8, 9, 109

Khamis, M. A., & Gomaa, W. (2014). Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework. *Engineering Applications of Artificial Intelligence*, *29*, 134–151. 8, 21

Kianfar, J., & Edara, P. (2013). A data mining approach to creating fundamental traffic flow diagram. *Procedia-Social and Behavioral Sciences*, *104*, 430–439. 16

Kuyer, L., Whiteson, S., Bakker, B., & Vlassis, N. (2008). Multiagent reinforcement learning for urban traffic control using coordination graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (pp. 656–671). Springer. 21

Lee, J., Abdulhai, B., Shalaby, A., & Chung, E.-H. (2005). Real-time optimization for adaptive traffic signal control using genetic algorithms. *Journal of Intelligent Transportation Systems*, *9*(3), 111–122. 30, 36

Lefebvre, W., Fierens, F., Trimpeneers, E., Janssen, S., Van de Vel, K., Deutsch, F., Viaene, P., Vankerkom, J., Dumont, G., Vanpoucke, C., et al. (2011). Modeling the effects of a speed limit reduction on traffic-related elemental carbon (ec) concentrations and population exposure to ec. *Atmospheric Environment*, *45*(1), 197–207. 32, 34, 62

Li, L., Lv, Y., & Wang, F.-Y. (2016a). Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, *3*(3), 247–254. 17, 20, 38, 73, 74, 75

Li, S.-b., Li, Y., Fu, B.-b., & Dang, W.-x. (2016b). Study on simulation optimization of dynamic traffic signal based on complex networks. *Procedia Engineering*, *137*, 1–10. 6, 28

Li, X., Li, G., Pang, S.-S., Yang, X., & Tian, J. (2004). Signal timing of intersections using integrated optimization of traffic quality, emissions and fuel consumption: a note. *Transportation Research Part D: Transport and Environment*, *9*(5), 401–407. 31

Liang, X., Du, X., Wang, G., & Han, Z. (2018). Deep reinforcement learning for traffic light control in vehicular networks. *arXiv preprint arXiv:1803.11115*. 2, 8, 16, 21, 72, 75, 88

Lin, S., De Schutter, B., Xi, Y., & Hellendoorn, H. (2012). Efficient network-wide model-based predictive control for urban traffic networks. *Transportation Research Part C: Emerging Technologies*, *24*, 122–140. 34

Lin, S., De Schutter, B., Xi, Y., & Hellendoorn, H. (2013). Integrated urban traffic control for the reduction of travel delays and emissions. *IEEE Transactions on Intelligent Transportation Systems*, *14*(4), 1609–1619. 33

Lin, Y., Dai, X., Li, L., & Wang, F.-Y. (2018). An efficient deep reinforcement learning model for urban traffic control. *arXiv preprint arXiv:1808.01876*. 8, 23, 38, 72, 73, 83, 103

Liu, B., Han, K., & Hu, J. (2016). Global optimization framework for real-time route guidance via variable message sign. *arXiv preprint arXiv:1611.08343*. 2

Liu, H., Han, K., Gayah, V., Friesz, T., & Yao, T. (2015). Data-driven linear decision rule approach for distributionally robust optimization of on-line signal control. *Transportation Research Procedia*, *7*, 536–555. 2, 8, 30, 40, 46, 47

LIU, M., DENG, J., XU, M., ZHANG, X., & WANG, W. (2017). Cooperative deep reinforcement learning for traffic signal control. In *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Halifax 2017*. 3, 8, 22

Lucas, D. E., Mirchandani, P. B., & Larry Head, K. (2000). Remote simulation to evaluate real-time traffic control strategies. *Transportation Research Record*, *1727*(1), 95–100. 7, 28, 52

Mannion, P. (2017). Knowledge-based multi-objective multi-agent reinforcement learning. 38, 72, 75, 86, 97, 99

Mannion, P., Duggan, J., & Howley, E. (2015). Learning traffic signal control with advice. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2015)*. 8, 26, 103

Manolis, D., Pappa, T., Diakaki, C., Papamichail, I., & Papageorgiou, M. (2018). Centralised versus decentralised signal control of large-scale urban road networks in real time: a simulation study. *IET Intelligent Transport Systems*, *12*(8), 891–900. 3, 7, 8

Mascia, M., Hu, K., Han, K., & North, R. (2015). Simulation output for traffic scenarios for the city of glasgow. Tech. rep., Technical report, CARBOTRAF. 44, 56, 81

Mascia, M., Hu, S., Han, K., North, R., Van Poppel, M., Theunis, J., Beckx, C., & Litzenberger, M. (2017). Impact of traffic management on black carbon emissions: a microsimulation study. *Networks and Spatial Economics*, *17*(1), 269–291. xvi, 33, 37, 50, 51, 52, 53, 62

McHugh, D. (2015). Traffic prediction and analysis using a big data and visualisation approach. *Department of Computer Science, Institute of Technology Blanchardstown*. 15

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, *18*(3), 11–11. 86

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*. 18

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529. 18

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, *13*(1), 103–130. 72

Mousavi, S. S., Schukat, M., & Howley, E. (2017). Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intelligent Transport Systems*, *11*(7), 417–423. 8, 24

Muresan, M., Fu, L., & Pan, G. (2019). Adaptive traffic signal control with deep reinforcement learning an exploratory investigation. *arXiv preprint arXiv:1901.00960*. 25

Nagurney, A., & Zhang, W.-B. (2007). Mathematical models of transportation and networks. *Mathematical Models in Economics*, *2*, 346–384. 27

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, vol. 99, (pp. 278–287). ICML. 38, 74, 85, 86

Ntziachristos, L., & Boulter, P. (2013). Emep/eea air pollutant emissions inventory guidebook 2013: Road vehicle tyre and brake wear; road surface wear. copenhagen. *European Environment Agency*. 51

Osorio, C., Chen, X., & Santos, B. F. (2017). Simulation-based travel time reliable signal control. *Transportation Science. Forthcoming. Available at: http://web. mit. edu/osorioc/www/papers/osoCheSanReliableSO. pdf*. 34, 36

Osorio, C., Flötteröd, G., & Zhang, C. (2015). A metamodel simulation-based optimization approach for the efficient calibration of stochastic traffic simulators. *Transportation Research Procedia*, *6*, 213–223. 34, 35, 36

Osorio, C., & Selvam, K. K. (2015). Solving large-scale urban transportation problems by combining the use of multiple traffic simulation models. *Transportation Research Procedia*, *6*, 272–284. 34, 37

Panait, L., & Luke, S. (2005). Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, *11*(3), 387–434. 75

Papageorgiou, M. (1990). Dynamic modeling, assignment, and route guidance in traffic networks. *Transportation Research Part B: Methodological*, *24*(6), 471–495. 2

Papatzikou, E., & Stathopoulos, A. (2015). An optimization method for sustainable traffic control in urban areas. *Transportation Research Part C: Emerging Technologies*, *55*, 179–190. 30

Peeta, S., & Gedela, S. (2001). Real-time variable message sign–based route guidance consistent with driver behavior. *Transportation Research Record*, *1752*(1), 117–125. 2

Pengra, D. B., & Dillman, L. (2009). Notes on data analysis and experimental uncertainty. *Ohio Wesleyan University, University of Washington*. 6

Polson, N., & Sokolov, V. (2016). Deep learning predictors for traffic flows. *arXiv preprint arXiv:1604.04527*. 15

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons. 72

Rakha, H., Ahn, K., & Trani, A. (2004). Development of vt-micro model for estimating hot stabilized light duty vehicle and truck emissions. *Transportation Research Part D: Transport and Environment*, *9*(1), 49–74. 32, 34

Savsani, V., Rao, R., & Vakharia, D. (2010). Optimal weight design of a gear train using particle swarm optimization and simulated annealing algorithms. *Mechanism and machine theory*, *45*(3), 531–541. 46

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*. 87, 88, 90

Schrank, D., Lomax, T., & Turner, S. (2010). Tti's 2010 urban mobility report. Tech. rep., Texas A&M University: College Station, TX. 2

Schultz, L., & Sokolov, V. (2018). Deep reinforcement learning for dynamic urban transportation problems. *arXiv preprint arXiv:1806.05310*. 8

Scotland, T. (2011a). Aire (analysis of instantaneous road emissions) user guidance. Tech. rep., Transport Scotland. 34

Scotland, T. (2011b). Aire analysis of instantaneous road emissions user guidance. *Transport Scotland, Glasgow, UK*. 50

Sha, D., & Hsu, C.-Y. (2008). A new particle swarm optimization for the open shop scheduling problem. *Computers & Operations Research*, *35*(10), 3243–3261. 46

SIAS (2011). *S-Paramics 2011 reference manual*. SIAS. 34, 50, 54

Smith, B. L., Scherer, W. T., & Hauser, T. A. (2001). Data-mining tools for the support of signal-timing plan development. *Transportation research record*, *1768*(1), 141–147. 2, 14

So, J., Motamedidehkordi, N., Wu, Y., Busch, F., & Choi, K. (2018). Estimating emissions based on the integration of microscopic traffic simulation and vehicle dynamics model. *International Journal of Sustainable Transportation*, *12*(4), 286–298. 34, 35, 36

Sobrino, N., Monzon, A., & Hernandez, S. (2016). Reduced carbon and energy footprint in highway operations: the highway energy assessment (hera) methodology. *Networks and Spatial Economics*, *16*(1), 395–414. 58

Song, J., Hu, S., & Han, K. (2017). Real-time adaptive traffic signal control: trade-off between traffic and environment objectives. Tech. rep., Transportation Research Board. 6, 14, 36, 40, 53

Spall, J. C., & Chin, D. C. (1997). Traffic-responsive signal timing for system-wide traffic control. *Transportation Research Part C: Emerging Technologies*, *5*(3-4), 153–163. 34

Srinivasan, D., Choy, M. C., & Cheu, R. L. (2006). Neural networks for real-time traffic signal control. *IEEE Transactions on intelligent transportation systems*, *7*(3), 261–272. 6, 17, 18

Stevanovic, A., Stevanovic, J., So, J., & Ostojic, M. (2015). Multi-criteria optimization of traffic signals: Mobility, safety, and environment. *Transportation Research Part C: Emerging Technologies*, *55*, 46–68. 34, 35, 36, 52

Sun, D., Benekohal, R. F., & Waller, S. T. (2006). Bi-level programming formulation and heuristic solution approach for dynamic traffic signal optimization. *Computer-Aided Civil and Infrastructure Engineering*, *21*(5), 321–333. 6, 28, 52

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163. 72

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge. 78

Sweet, M. (2014). Traffic congestion's economic impacts: Evidence from us metropolitan regions. *Urban Studies*, *51*(10), 2088–2110. 2

Teo, K. T. K., Yeo, K. B., Chin, Y. K., Chuo, H. S. E., & Tan, M. K. (2014). Agent-based optimization for multiple signalized intersections using q-learning. *International Journal of Simulation: Systems, Science and Technology (IJSSST 2014)*, *15*(6), 90–96. 23, 38, 73, 83

Touhbi, S., Babram, M. A., Nguyen-Huu, T., Marilleau, N., Hbid, M. L., Cambier, C., & Stinckwich, S. (2017). Adaptive traffic signal control: Exploring reward definition for reinforcement learning. *Procedia Computer Science*, *109*, 513–520. 24

Ukkusuri, S. V., Ramadurai, G., & Patil, G. (2010). A robust transportation signal control problem accounting for traffic dynamics. *Computers & Operations Research*, *37*(5), 869–879. 28, 36

Uselton, S., Treinish, L., Ahrens, J., Bethel, W., et al. (1998). Multi-source data analysis challenges. In *Proceedings of the conference on Visualization'98*, (pp. 501–504). IEEE Computer Society Press. 8

Van der Pol, E., & Oliehoek, F. A. (2016). Coordinated deep reinforcement learners for traffic light control. *Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016)*. 8, 21, 38, 73, 83

Wang, C., Li, X., Zhou, X., Wang, A., & Nedjah, N. (2016). Soft computing in big data intelligent transportation systems. *Applied Soft Computing*, *38*, 1099–1108. 14, 36

Watling, D., & Van Vuren, T. (1993). The modelling of dynamic route guidance systems. *Transportation Research Part C: Emerging Technologies*, *1*(2), 159–182. 2

Webster, F. V. (1958). *Traffic signal settings*. Road Research Lab. 28

Wei, H., Zheng, G., Yao, H., & Li, Z. (2018). Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 2496–2505). ACM. 2, 8, 22, 38, 73, 83, 103

Wibisono, A., Jatmiko, W., Wisesa, H. A., Hardjono, B., & Mursanto, P. (2016). Traffic big data prediction and visualization using fast incremental model trees-drift detection (fimt-dd). *Knowledge-Based Systems*, *93*, 33–46. 13

Wiering, M. (2000). Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, (pp. 1151–1158). 8, 20, 21

Yan, S., Wu, C., Dai, W., Ghanem, M., & Guo, Y. (2012). Environmental monitoring via compressive sensing. In *Proceedings of the sixth international workshop on knowledge discovery from sensor data*, (pp. 61–68). ACM. 14

Yang, B., Yoon, J. W., & Monterola, C. (2016). Microscopic statistical characterisation of the congested traffic flow and some salient empirical features. *arXiv preprint arXiv:1603.04272*. 16

Yin, P.-Y. (2006). Particle swarm optimization for point pattern matching. *Journal of Visual Communication and Image Representation*, *17*(1), 143–162. 46

Yin, Y. (2008). Robust optimal traffic signal timing. *Transportation Research Part B: Methodological*, *42*(10), 911–924. 28

Yu, M., & Fan, W. D. (2019). Optimal variable speed limit control in connected autonomous vehicle environment for relieving freeway congestion. *Journal of Transportation Engineering, Part A: Systems*, *145*(4), 04019007. 2

Zhai, C., Luo, F., Liu, Y., & Xu, J. (2018). Adaptive control of isolated intersections based on sequential signal-stage optimisation. In *Proceedings of the Institution of Civil Engineers-Transport*, (pp. 1–12). Thomas Telford Ltd. 30

Zhang, K., Batterman, S., & Dion, F. (2011). Vehicle emissions in congestion: Comparison of work zone, rush hour and free-flow conditions. *Atmospheric Environment*, *45*(11), 1929–1939. 32

Zhang, L., Yin, Y., & Chen, S. (2013). Robust signal timing optimization with environmental concerns. *Transportation Research Part C: Emerging Technologies*, *29*, 55–71. 32, 36

Zhang, L., Yin, Y., & Lou, Y. (2010). Robust signal timing for arterials under day-to-day demand variations. *Transportation Research Record*, *2192*(1), 156–166. 6, 28, 51

Zheng, L., Xu, C., Jin, P. J., & Ran, B. (2019). Network-wide signal timing stochastic simulation optimization with environmental concerns. *Applied Soft Computing*, *77*, 678–687. 36

Zhou, Z., & Cai, M. (2014). Intersection signal control multi-objective optimization based on genetic algorithm. *Journal of Traffic and Transportation Engineering (English Edition)*, *1*(2), 153–158. 34, 35, 36

Zuurbier, F. S., van Zuylen, H. J., Hoogendoorn, S. P., & Chen, Y. (2006). Generating optimal controlled prescriptive route guidance in realistic traffic networks: a generic approach. *Transportation research record*, *1944*(1), 58–66. 2