Technical Report Documentation Page

1. Report No. SWUTC/07/473700-00092-1	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle
Development and Evaluation of a Multi-Agent Based Neuro-Fuzzy Arterial Traffic Signal Control System

5. Report Date
September 2007

6. Performing Organization Code

7. Author(s)
Yunlong Zhang, Yuanchang Xie, and Zhirui Ye

8. Performing Organization Report No.
473700-00092

9. Performing Organization Name and Address
Texas Transportation Institute
The Texas A&M University System
College Station, Texas 77843-3135

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTRS99-G-0006

12. Sponsoring Agency Name and Address
Southwest Region University Transportation Center
Texas Transportation Institute
Texas A&M University System
College Station, Texas 77843-3135

13. Type of Report and Period Covered
Technical Report
Sept. 2006- Sept. 2007

14. Sponsoring Agency Code

15. Supplementary Notes
Supported by a rant from the U.S. Department of Transportation, University Transportation Centers Program.

16. Abstract

Arterial traffic signal control is a very important aspect of traffic management system. Efficient arterial traffic signal control strategy can reduce delay, stops, congestion, and pollution and save travel time. Commonly used pre-timed or traffic actuated signal control do not have the capability to fully respond to real-time traffic demand and pattern changes. Although some of the well-known adaptive control systems have shown advantageous over the traditional per-timed and actuated control strategies, their centralized architecture makes the maintenance, expansion, and upgrade difficult and costly.

Distributed artificial intelligence technologies such as multi-agent systems are well suited for arterial signal control and they have the ability to decompose and accomplish complicated control problems by cooperatively simple agents such that flexibility, efficiency, robustness, and cost effectiveness can be achieved. An in-depth investigation of applying the multi-agent technology in arterial signal control is conducted in this research, and two multi-agent arterial adaptive signal control systems based on neuro-fuzzy reinforcement learning are developed and evaluated using VISSIM simulation and real world traffic data collected in College Station, Texas. The two multi-agent arterial adaptive control systems are compared with optimized coordinated pre-timed and actuated controls. Encouraging results are obtained from both multi-agent control systems.

17. Key Words
Multi-Agent; Neuro-Fuzzy; Arterial Traffic Signal Control; Reinforcement Learning

18. Distribution Statement
No restrictions. This document is available to the public through NTIS:
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

19. Security Classification.(of this report) Unclassified	20. Security Classification.(of this page) Unclassified	21. No. of Pages 122	22. Price

DEVELOPMENT AND EVALUATION OF A MULTI-AGENT BASED NEURO-FUZZY ARTERIAL TRAFFIC SIGNAL CONTROL SYSTEM

by

Yunlong Zhang, Ph.D., P.E.
Assistant Professor
Texas A&M University

Yuanchang Xie
Graduate Research Assistant
Texas A&M University

and
Zhirui Ye
Graduate Research Assistant
Texas A&M University

Research Report 473700-00092-1

Sponsored by the
Southwest Region University Transportation Center
Texas Transportation Institute

September 2007

Texas Transportation Institute
Texas A&M University System
College Station, Texas 77843-3135

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ABSTRACT

Arterial traffic signal control is a very important aspect of traffic management system. Efficient arterial traffic signal control strategy can reduce delay, stops, congestion, and pollution and save travel time. Commonly used pre-timed or traffic actuated signal controls does not have the capability to fully respond to real-time traffic demand and pattern changes. Although some of the well-known adaptive control systems have shown advantageous over the traditional per-timed and actuated control strategies, their centralized architecture makes the maintenance, expansion, and upgrade difficult and costly.

Distributed artificial technologies such as multi-agent system is well suited for arterial signal control, and it has the ability to decompose complicated control problems and accomplish them by cooperatively simple agents such that flexibility, efficiency, robustness, and cost effectiveness can be achieved. An in-depth investigation of applying multi-agent technology in arterial signal control is conducted in this research, and two multi-agent arterial adaptive signal control systems based on neuro-fuzzy reinforcement learning are developed and evaluated using VISSIM simulation and real world traffic data collected in College Station, Texas. The two multi-agent arterial adaptive control systems are compared with optimized coordinated pre-timed and actuated control, and encouraging results are obtained from both multi-agent control systems.

EXECUTIVE SUMMARY

This research aims at developing and evaluating a multi-agent adaptive traffic signal control system for arterials. This multi-agent control system is based on reinforcement learning, which is an important research area in distributed artificial intelligence and has been extensively used in many applications including real-time control.

In this research, a systematic comparison between reinforcement learning control and existing adaptive traffic control methods is first presented from the theoretical perspective. This comparison shows both the connections between them and the benefits of using reinforcement learning. A Neuro-Fuzzy Actor-Critic Reinforcement Learning (NFACRL) method is then introduced for traffic signal control. NFACRL integrates fuzzy logic and neural networks into reinforcement learning and can better handle the problems associated with ordinary reinforcement learning and existing adaptive traffic control methods.

This NFACRL method is first applied to isolated intersection control and two implementation schemes are considered. The first scheme uses a fixed phase sequence and variable cycle length, while the second one considers a variable phase sequence and is not constrained to the concept of cycle. Testing results on isolated intersections show that both implementation schemes outperform optimized pre-timed and actuated controls in most cases.

The two implementation schemes are further extended for arterial control, with each intersection controlled by a NFACRL control agent and the arterial modeled as a multi-agent system. Different multi-agent coordination strategies are reviewed. A simple but robust method is adopted for coordinating traffic signal control agents along the arterial. Based on the two NFACRL implementation schemes, two multi-agent arterial adaptive traffic signal control systems are developed and evaluated using VISSIM simulation under different traffic volume scenarios. The two multi-agent control systems are compared with optimized coordinated pre-timed and actuated controls. Testing results show that both multi-agent control systems have a very encouraging performance and the multi-agent control system with variable phase sequence performs the best in most cases. Finally, issues on how to further improve the NFACRL method and implement it in the real world are discussed.

ACKNOWLEDGMENTS

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers Program to the Southwest Region University Transportation Center.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 PROBLEM STATEMENT

Many urban areas have been experiencing explosive vehicular traffic growth on arterials, causing large amount of delay at arterial intersections. Optimal isolated intersection control and signal coordination along an arterial have been identified as efficient and low cost methods for reducing delay and congestion (Meyer, 1979; Schrank and Lomax, 2005). It was estimated that traffic signal coordination alone reduced delay by 11 million hours and saved $187 million from congestion cost for 85 urban areas in the United States in 2003 (Schrank and Lomax, 2005). Considering there are more than 300,000 traffic signals in North America (FHWA, 1995), the potential saving from improving traffic signal timing is very significant.

Most current traffic signal control systems used in the real world are either pre-timed or actuated control. One major problem with the pre-timed signal control is that it does not have the capability to respond to short-term traffic demand and pattern changes (Li, 2002). Traffic-actuated control can partially solve this problem by extending green phases in response to real-time traffic arrivals. However, this green phase extension strategy makes decision primarily based on traffic arrivals of the movements being served. Even very long queues on other movements may not stop the extension of the current green phase (Bingham, 2001; Zhang et al., 2005). When traffic demand is heavy, actuated control may have unsatisfying control results (Zhang et al., 2005).

Adaptive signal control, which adjusts signal timing parameters in response to real-time traffic flow fluctuations, has a great potential to outperform both pre-timed and actuated controls and has been researched for a few decades. Several adaptive signal control systems such as RHODES and OPAC have been developed and better performance compared with pre-timed and actuated control was reported (Mirchandani and Head, 2001; Gartner, 1983). However, many existing adaptive traffic signal control systems are based on dynamic programming and these systems' applicability may be limited due to restrictions from their problem formulations as well as solution procedures. In addition, for some of the adaptive control systems using centralized architecture, the maintenance and expansion are difficult and costly. Therefore, it is very important to develop new and more flexible distributed adaptive control strategies.

1.2 OVERVIEW OF THE PROPOSED METHODOLOGY

In this research, a multi-agent based neuro-fuzzy arterial traffic signal control method is proposed, which is based on reinforcement learning. As one of the key elements of artificial intelligence, reinforcement learning has been successfully applied to control problems such as elevator operation (Crites and Barto, 1998) and robot soccer games (Park et al., 2001). It has also been used for supply chain modeling (Swaminathan et al., 1998), activity-travel pattern analysis (Janssens et al., 2007), and dynamic resource allocation (Vengerov, 2007). In the field of reinforcement learning control, the controller is often referred to as agent, which is formally defined as anything that can observe the environment and act upon it. The environment is the subject to be controlled. A system consists of a group of agents that interact with each other is called a multi-agent system (MAS) (Vlassis, 2003). At each decision step, the agent applies an action to the environment in response to the environment's current state. Under the effect of this action, the environment may change accordingly and result in a new state and a feedback signal called reward (or penalty). Based on the new state and the reward, the agent can adjust its policy and learn how to achieve a certain goal from the interactions with the environment (Sutton and Barto, 1998). This learning approach is called reinforcement learning. One advantage of using the reinforcement learning for control applications is that it can learn the optimal control policy directly from interactions between the controller and the environment without knowing the underlying model of the subject to be controlled. In addition, the reinforcement learning method can well circumvent the problems associated with dynamic programming algorithms used in some of the existing adaptive traffic signal control systems. Also, it is conceptually desirable to model arterial traffic signal control problems using reinforcement learning and the MAS framework.

In the case of isolated intersection traffic control, the agent is the traffic signal controller and the environment consists of all other traffic and geometry factors related to the intersection. Queue length or total delay can be used as the penalty. The concept of using reinforcement learning for isolated intersection traffic control is shown in Figure 1.

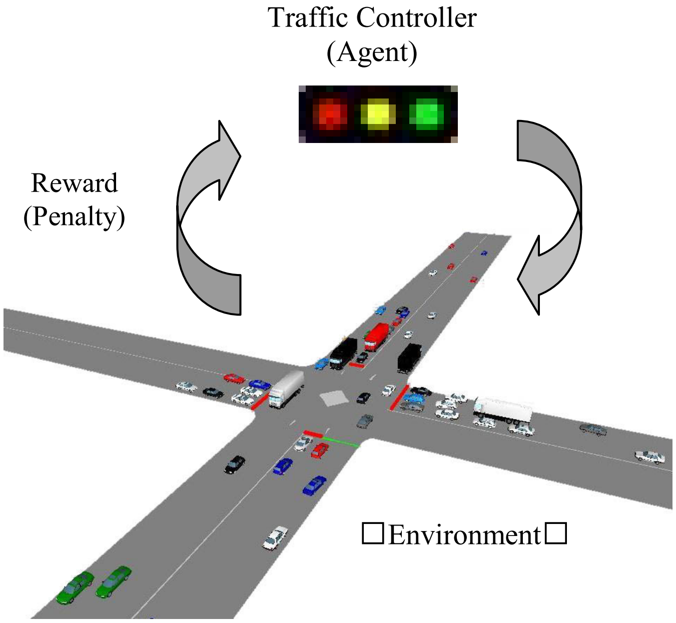


Figure 1 Modeling intersection traffic signal control as agent and environment system.

Arterial traffic signal control can be modeled as a MAS and solved by the reinforcement learning method. For a signalized arterial, the signal controller of each intersection is an individually-motivated agent. The agents at different intersections interact with each other and try to optimally control traffic along the arterial. Under the framework of MAS, it is possible to decompose a complicated control system by coordinating agents such that flexibility, efficiency, robustness, and cost effectiveness can be achieved.

Despite the many potential benefits of using MAS for arterial traffic signal control, few thorough investigations have been found. The potential of applying reinforcement learning and agent technology to traffic signal control, especially arterial traffic signal control, has not been explored fully. Thus, it is imperative to conduct an in-depth research on this topic.

1.3 RESEARCH OBJECTIVES

The following objectives are identified in this study.

1. To develop an isolated intersection control method using reinforcement learning. The new control method should be truly demand-responsive and has the ability to better solve the curse of dimensionality and generalization problems.

2. To develop a reinforcement learning control method for arterials based on the proposed isolated intersection reinforcement learning control method.
3. To perform a comprehensive evaluation of the proposed reinforcement learning arterial traffic control method based on a widely-accepted microscopic traffic simulation platform.
4. To provide directions for further studies and field implementation of the proposed reinforcement learning arterial traffic control method.

1.4 RESEARCH OVERVIEW

This report consists of six chapters. In the next chapter, various traffic signal control types and strategies are reviewed. The review covers pre-timed control, actuated control, and adaptive control, with a focus on adaptive traffic signal control methods.

In Chapter 3, a systematic introduction of reinforcement learning is presented. The introduction of reinforcement learning starts with the discussion of Markov property and Markov Decision Processes (MDP), and then proceeds to review various commonly-used reinforcement learning methods such as SARSA, Q-Learning, and Actor-Critic method. Following the introduction is a review of studies that applied reinforcement learning to traffic signal control. Problems with the existing applications are also discussed.

In Chapter 4, a new reinforcement learning method based on fuzzy logic control and neural networks is discussed in details. This neuro-fuzzy reinforcement learning method is then applied to isolated intersection and arterial traffic control and two application schemes are proposed. Both schemes are further extended for arterial traffic control based on a simple but robust coordination strategy.

Chapter 5 first describes the data and the microscopic traffic simulation platform used for evaluating the proposed neuro-fuzzy reinforcement learning control method. A test design is then presented that describes how the proposed control method is evaluated at both isolated intersection and arterial levels as well as under different traffic demand conditions. Finally, the evaluation results from the proposed neuro-fuzzy reinforcement learning control, pre-timed, and actuated control methods are presented, compared, and discussed.

Chapter 6 summarizes findings and highlights contributions of this research. Possible future extensions on this research topic are also provided.

CHAPTER 2. TRAFFIC SIGNAL CONTROL BACKGROUND AND LITERAURE REVIEW

2.1 INTRODUCTION

Intersection traffic control first emerged in the form of manually turned semaphores in London in 1868 (Abbas, 2001). As an important method to resolve traffic conflicts, improve operational efficiency, and enhance safety at intersections, this idea was soon adapted by other nations and eventually evolved into three major types of traffic signal control strategies: pre-timed, actuated, and adaptive control. Each control type can be applied at an isolated intersection in its simplest form. By properly considering coordination, they can also be used for arterial and network traffic control. This research focuses on the development of an adaptive control method that can be used for both isolated intersections and signalized arterials. Conceptually, the new algorithm introduced in this study can be expanded for network traffic control.

In the rest of this chapter, pros and cons of pre-timed, actuated, and existing adaptive traffic control methods are reviewed in details. In addition, several rule-based control methods are also discussed.

2.2 PRE-TIMED TRAFFIC SIGNAL CONTROL

2.2.1 Pre-Timed Isolated Intersection Traffic Signal Control

Figure 2 is a typical four-approach isolated intersection with eight movements (each through movement and its associated right-turn movement are combined as one movement). Each movement is usually labeled by a number between 1 and 8 in NEMA convention (NEMA, 1992). Pre-timed signal control operates in a cyclic manner. In each cycle there are several signal phases. For each signal phase, one or more non-conflicting movements are allowed. For pre-timed control, the phase sequence and phase duration are fixed. Thus, the cycle length is also fixed. Figure 3 shows a typical example of protected left-leading pre-timed control for the isolated intersection shown in Figure 2.

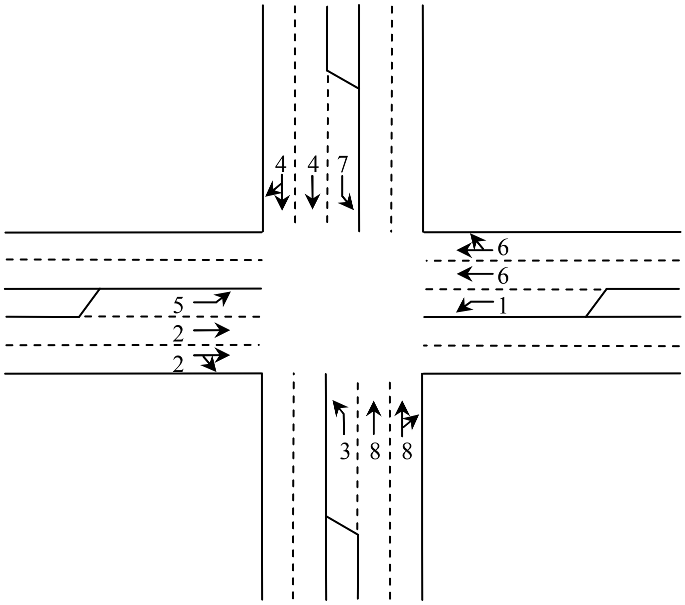


Figure 2 A typical four-approach intersection.

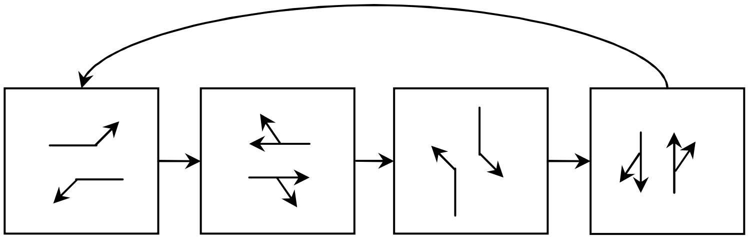


Figure 3 An example of protected left-leading pre-timed control.

For isolated intersections, the control parameters are usually optimized based on either the Webster (Webster, 1958) method or the procedure in Highway Capacity Manual (HCM) (TRB, 2000), and are determined based on average traffic volume data. In the real world, traffic volume may change considerably throughout the day and also in short intervals. Obviously, a control method based on average traffic volume data cannot effectively consider traffic flow fluctuations and may result in suboptimal control. Therefore, in practice, the applications of pre-timed control are often limited to locations with less variable traffic flows. Besides, the control parameters of pre-timed control need calibrations from time to time to reflect mid-term or long-term traffic flow pattern changes.

2.2.2 Pre-Timed Arterial Traffic Signal Control

For closely spaced intersections on an arterial, vehicle arrivals to a downstream intersection are often affected by the control strategies of the upstream intersections. Vehicles also travel in platoons. Thus, it is desirable to coordinate the pre-timed traffic signals of adjacent intersections such that platoons of vehicles can get through a number of intersections without being stopped. For this purpose, offset is used in addition to cycle length, phase sequence, and phase duration to coordinate adjacent traffic signals (Morgan and Little, 1964).

A commonly-used signal coordination strategy is to maximize the bandwidth of through movements along the arterial, which maximizes the number of vehicles that go through the arterial without being stopped. However, this strategy may not be effective due to the following reasons:

1. Coordinated pre-timed method usually requires all traffic signal controllers being coordinated to have the same cycle length. For different intersections on an arterial, their optimal cycle lengths are most likely different. Requiring a common cycle length for all intersections may cause additional delay to vehicles at some of the intersections.
2. Offsets are calculated based on the distances between two intersections and the average speed. In many cases, travel time between two adjacent intersections may vary depending on flow and queuing situations.
3. Large percentage of turning traffic can make the control strategy less efficient.
4. For two-way traffic, the offset in one direction also determines the offset in the other direction. It is difficult to give both directions equally good coordination.
5. Coordinated pre-timed control gives higher priority to traffic on main streets. This may cause cross street traffic to experience unreasonably large delays.

2.3 ACTUATED TRAFFIC SIGNAL CONTROL

2.3.1 Actuated Signal Control at Isolated Intersection

Actuated control provides an intermediate solution between pre-timed control and adaptive control (Abbas, 2001). It can be further classified into semi- and fully-actuated control (Roess et al., 2004). Both semi- and fully-actuated control methods are based on the same fundamental principle. The length of green phase falls between the preset minimum and maximum green times. After the minimum green time is served, as long as a vehicular actuation occurs before the preceding vehicle extension ends and the total green extension has not exceeded the preset maximum green time, another green extension will be given to the current green phase. This actuated control strategy can partially solve the criticism attributed to the pre-timed control strategy in a sense that it can respond to the real-time traffic arrivals of the current green phase. However, actuated control strategy does not take into consideration of the queue lengths on other conflicting movements, and may result in suboptimal control, especially under heavy traffic conditions.

2.3.2 Actuated Traffic Signal Control on Arterial

When applying actuated signal control to isolated intersections, the cycle length may vary from cycle to cycle, as the phase durations are variable depending on actual traffic arrivals. When applying actuated control to arterials, the coordinated actuated control must have a constant cycle length and a coordinated phase should be defined for each intersection. Actuated control is considered to be more suitable for arterial traffic signal control than pre-timed control (Gartner, 1983; Abbas, 2001). However, it still has unsolved problems such as “the-early-return-to-green” (Abbas, 2001), which may cause unnecessary stops of vehicles.

2.4 ADAPTIVE SIGNAL CONTROL

In the following sections, a number of well known or recently developed adaptive traffic signal control systems are reviewed. Adaptive traffic signal control systems are usually complex and include prediction and estimation modules, it is difficult to cover every detailed aspect of each system. Therefore, this review only focuses on the system designs and architectures, problem formulations, solution procedures, and optimization algorithms of the existing systems. The existing systems that are reviewed include UTCS (Gartner, 1983), SCOOT (Hunt, 1981), SCAT (Sims and Dobinson, 1980; Lowrie, 1982), DYPIC (Robertson and Bretherton, 1974), OPAC (Gartner, 1983), RHODES (Mirchandani and Head, 2001), ALLONS-D (Porche, 1997), and MDP&DP (Yu and Recker, 2006).

2.4.1 Urban Traffic Control System (UTCS)

Starting from the 1970s, the U.S. Department of Transportation (USDOT) conducted several research projects on urban traffic control system (UTCS) (Gartner, 1983). The intersection control strategies proposed and evaluated in these projects can largely be classified into three categories: first-generation control (first-GC), second-generation control (second-GC), and third-generation control (third-GC). First-GC strategy generates traffic control plans based on historically averaged traffic volume data. Depending on the time-of-day (TOD), different pre-timed control plans are selected and implemented. The updating frequency for the control plans is usually 15 minutes. Second-GC strategy optimizes traffic signal control plans every 5 minutes based on predicted traffic volume data instead of historical data. The updating frequency for traffic signal control plans is restricted to be no less than 10 minutes in belief that this can avoid transition disturbances. Third-GC is similar to the second-GC, but updates signal timing plans using a shorter interval of 3-5 minutes (Gartner, 1982).

2.4.2 Split, Cycle and Offset Optimization Technique (SCOOT)

Hunt et al. (1981) developed the SCOOT system, which is considered to be equivalent to a second-GC (Gartner et al., 1995) or third-GC (Abbas, 2001) method. In SCOOT, intersections are grouped into many sub-areas and signal controllers in each sub-area operate at a common cycle length. SCOOT makes frequent and small changes to signal control parameters such as cycle length, phase duration, and offset of a pre-timed plan based on actual traffic flow variations (Abbas, 2001; Shelby, 2004*)*. The adjustment of signal control parameters is based on a traffic model that predicts delay and stops resulted from different signal timing plans. The plans that can best reduce delay and stops are then selected and implemented (Hunt, 1981). SCOOT has been widely used in the United Kingdom. There are also a few implementations of it in other countries. The latest version of SCOOT is SCOOT MC3 (Peek, 2007), which has some new features such as the ability to skip phases for bus priority purpose.

2.4.3 Sydney Coordinated Adaptive Traffic System (SCATS)

SCATS was developed by Australian researchers (Sims and Dobinson, 1980; Lowrie, 1982). It is similar to SCOOT and is considered to be an adaptive control method between first-GC and second-GC (Abbas, 2001). The major difference between SCATS and SCOOT is that SCATS does not have a traffic model or a traffic signal control plan optimizer. SCATS selects the best phase durations and offsets from some predefined plans (Abbas, 2001) based on real time traffic flow conditions.

SCATS has a hierarchical system structure with three levels. The lowest level consists of the local controllers at each signalized intersection. They perform tasks such as data collection, data preprocessing, and assessment of detector malfunctions. The middle level includes the regional masters, which form the core of SCATS. Each regional master controls up to several hundred local controllers, and these controllers are further grouped into systems and sub-systems. Sub-systems usually consist of several intersections and are the smallest control element on the multi-intersection level. The highest level is the control center, which does not really perform any specific control operations. The main purpose of the control center is to monitor the entire system.

2.4.4 Dynamic Programmed Intersection Control (DYPIC)

Robertson and Bretherton (1974) developed an optimal control method called DYPIC based on dynamic programming for an isolated intersection. A simple intersection with only two conflicting movements was used by Robertson and Bretherton to illustrate their method. Since there were only two conflicting movements, the control decisions were to either extend or terminate current green signal. In their study, Robertson and Bretherton assumed that exact traffic arrival information in the next few minutes (over the decision horizon) was known. However, this is impossible in real world applications.

In the DYPIC method, the entire decision horizon was divided into N intervals. Each interval was 5 seconds long. At the end of each small interval (decision point), the control logic made a decision to either extend the current green phase or terminate it and give green to the other movement. There were no constraints such as minimum and maximum green times. Robertson and Bretherton (1974) formulated this intersection control as a dynamic programming problem. Specifically, the decision point corresponded to the concept of stage in dynamic programming; states at each stage were characterized by the signal state (green or red) and the queue lengths on each approach. As the exact traffic arrival information was assumed to be known for the entire decision horizon, queue lengths of each approach at any stage can be estimated using some traffic models. The optimization goal was to find an optimal control strategy consisting of a sequence of actions $A = \{a_1, ..., a_N\}$ that minimize the total delay. Based on the initial signal states, queue lengths of each approach, and future traffic arrival information, the entire decision process can be illustrated by a decision tree as shown in Figure 4.

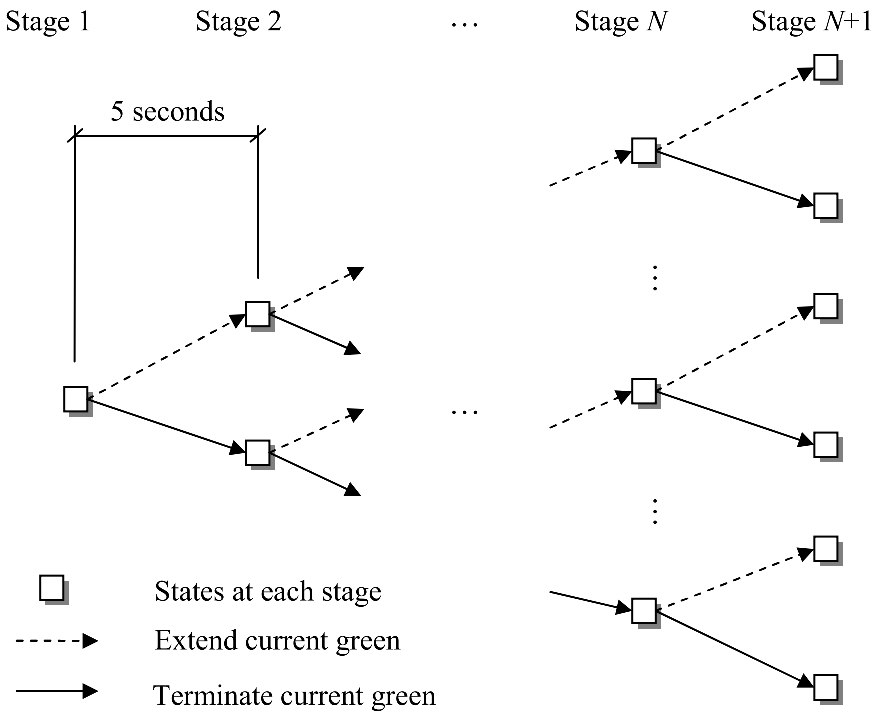


Figure 4 Illustration of the DYPIC method.

The following formula was used in DYPIC to find the optimal control strategy for the decision problem shown in Figure 4.

$$f_i(j) = \min_{a_i}\{C_{jk} + f_{i+1}(k)\}, \quad i = 1,...,N, \quad j \in S_i, \quad k \in S_{i+1} \tag{1}$$

where

$S_i =$ all possible states at stage *i*;

$S_{i+1} =$ all possible states at stage *i*+1;

$C_{jk} =$ total delay associated with transition from state *j* at stage *i* to state *k* at stage *i*+1;

$f_i(j) =$ value function for state *j* at stage *i*;

$a_i =$ action taken at stage *i* (either extension or termination); and

N = number of stages minus 1.

Starting from stage N and working backwards, the values of each state at stages 1 through N can be obtained using Equation (1). The value for the initial state actually is the minimum delay resulted from the optimal control strategy. By tracking the path that leads to the value of the initial state, one can find the best control strategy. This method is often referred to as the backward dynamic programming.

There are four major problems with the DYPIC method. Firstly, as shown in Figure 4, since each action may result in two states in the next stage, if there are N+1 stages, the maximum possible number of states at the final stage is 2^N. If the decision horizon is 2 minutes and the interval is 5 seconds long, then the maximum possible number of states at the final stage could be $2^{120/5} = 16777216$. Although dynamic programming theoretically can give this problem a globally optimal solution, so many states will definitely make the computation time a serious problem for real time traffic signal control. Secondly, only an isolated intersection with two movements was considered in this simple example. For practical traffic signal control problems, usually there are eight movements and at each stage there could be multiple different actions. Thirdly, this example assumed all traffic arrivals during the decision horizon were known. This is impossible in reality. Finally, the DYPIC method assumed deterministic state transitions. Given current queue lengths, signal states, traffic arrival information, and the action to be applied, the resulted new state was determined. This assumption in fact may not always be true, as driver behaviors are very complicated and no traffic models can perfectly predict future traffic states.

Robertson and Bretherton (1974) compared the DYPIC control method with pre-timed control at an isolated intersection. Two different traffic arrival conditions were tested, which were random arrival and cyclic arrival. For random arrival condition, the results showed that the DYPIC method reduced delay by at least 50 percent. While for the cyclic arrival condition, limited tests showed that the DYPIC method reduced average delay by 3 seconds per vehicle.

2.4.5 Optimized Policies for Adaptive Control (OPAC)

The second-GC and third-GC strategies were expected to perform better than the first-GC strategy, as they seemed to be able to provide better responsiveness by using detected and predicted dada and shorter updating intervals. However, some field tests showed that the first-GC strategy in general outperformed the other two strategies (Gartner, 1982; Henry et al., 1976; Holroyd and Robertson, 1973). Due to the unsatisfactory results of the second-GC and third-GC strategies, Gartner (1982; 1983) suggested a truly demand-responsive control strategy that is not restricted to the conventional concepts of cycle length and phase durations.

In his research, Gartner first presented an isolated intersection traffic control example using a dynamic programming approach, later named as OPAC-1 (Gartner et al., 2001), which was similar to DYPIC (Robertson and Bretherton, 1974). Gartner discussed that though this dynamic programming approach can guarantee global optimality, it is not suitable for real time applications due to the excessive computation time and the requirement of exact traffic arrival data. Based on OPAC-1, Gartner proposed a simplified control algorithm using Optimal Sequential Constrained Search (OSCS) algorithm instead of dynamic programming (Gartner, 1983). The resulted new control method was referred to as OPAC-2.

In the same study (Gartner, 1983), Gartner also proposed a rolling horizon approach to predict traffic arrivals such that the constraint of knowing exact traffic arrivals was removed. By adding the rolling horizon prediction method, OPAC-2 evolved into OPAC-3 (Gartner et al., 2001). Gartner evaluated the OPAC-3 control strategy using a special version of NETSIM. The evaluation was based on data collected from an intersection in Tucson, Arizona. The results showed that compared to the existing control method deployed at that intersection, OPAC-3 reduced average delay by 30-50 percent and increased average speed by 10-20 percent. Although the improvement from OPAC-3 was significant, Gartner did not specify if the original control strategy was optimized or not.

In a subsequent study, Garter et al. (2001) summarized the development of OPAC and the application results of its latest version OPAC-4. OPAC-4 was developed to extend the application of OPAC from single intersections to arterials and networks. OPAC-4 uses a Virtual-Fixed-Cycle (VFC) technique and is often referred to as VFC-OPAC. The VFC-OPAC has a hierarchical structure with three layers as shown in Figure 5. The synchronization layer calculates the VFC every few minutes. Based on this VFC, the coordination layer optimizes the offset of each intersection. The local control layer optimizes signal changes subject to VFC and offset constraints from the synchronization and coordination layers. Although it is conceptually clear how the VFC-OPAC works for arterial and network traffic control, details about this control process are not available in Gartner et al. (2001) and any other literatures.

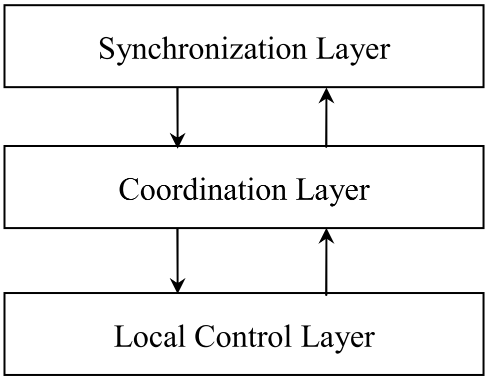


Figure 5 The hierarchical control structure of VFC-OPAC.

The OPAC-4 system was tested on an arterial in Reston, Virginia. The field test was carried out in two steps. In step one, the existing coordinated pre-timed control system was retimed and performance data were collected. In step two, the OPAC-4 system was implemented and its performance data were also collected. Comparison of travel time data showed that the performance of the existing coordinated pre-timed control and OPAC-4 control were not significantly different from each other. Gartner et al. (2001) explained that this might be caused by the traffic flow pattern changes between the two data collection periods.

2.4.6 Real-Time Hierarchical Optimized Distributed Effective System (RHODES)

RHODES was developed at the University of Arizona (Mirchandani and Head, 2001). RHODES has two core modules: prediction and control. The prediction module predicts future traffic arrival information such as when and how many vehicles will arrive, while the control module is used to control intersection and network traffic flows. The intersection control logic uses an algorithm developed by Sen and Head (1997). This algorithm is called Controlled Optimization of Phases (COP) and is also based on dynamic programming. The network control logic used in RHODES is based on both COP and REALBAND (Dell'Olmo and Mirchandani, 1995). REALBAND algorithm is used to produce progression bands in terms of observed platoons in the network, and these progression bands are then used as constraints for COP to develop optimal control strategies for individual intersections.

Although the COP algorithm is also based on dynamic programming, it uses definitions for stages, states, and actions that are quite different from DYPIC and OPAC methods. In COP, stages are defined as a sequence of phases; states at each stage are defined as the number of time steps that could be assigned to the current stage; the optimization goal is to find an optimal plan to allocate time steps to each stage (phase) such that the overall vehicle delay/number of stops/queue lengths could be minimized. This modeling approach is similar to applying dynamic programming to resource allocation problems (Butenko, 2005). The success of both the OPAC and RHODES models relies on an accurate prediction of traffic arrivals over the entire decision horizon.

2.4.7 Adaptive Limited Look-ahead Optimization of Network Signals – Decentralized (ALLONS-D)

Porche (1997) proposed a decentralized adaptive traffic signal control method called ALLONS-D in his dissertation. ALLONS-D is based on a depth-first branch and bound algorithm and uses a decision tree to help find the best control sequence (Porche, 1997). The decision tree used in ALLONS-D is similar to the one used in DYPIC as shown in Figure 4, in which each node represents a decision point and has a cost value associated with it while each arc is a control action. Figure 4 only shows the decision tree for an isolated intersection with two-phase control. For intersections with four or more phases, the size of the decision tree will make exhaustive search methods infeasible for real time applications. To improve searching efficiency, ALLONS-D uses the branch and bound algorithm and a special technique called "Serve the Largest Cost" (STLC) to find the best control sequence. The entire optimization process of ALLONS-D can be divided into two parts: 1) initial decision path (sequence) building, and 2) backtracking and exploration.

In the decision path building part, a feasible decision path is constructed. The construction of the decision path is based on the STLC technique. In terms of the STLC technique, at each decision point, the control phase incurring the highest delay in a most recent time period should be turned green. Following this STLC policy, a sequence of decisions is made until the initial queues and predicted traffic arrivals are cleared. This process can be better illustrated in Figure 6.

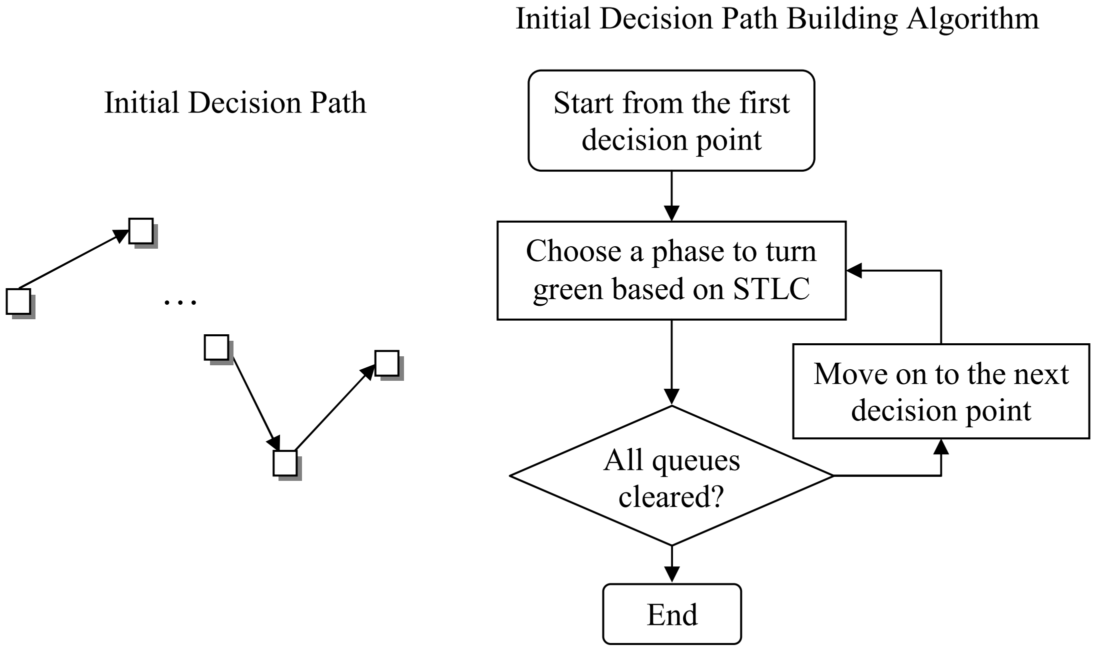

Initial Decision Path Building Algorithm
Initial Decision Path
…
Start from the first decision point
Choose a phase to turn green based on STLC
Move on to the next decision point
All queues cleared?
End

Figure 6 Initial decision path building of ALLONS-D.

The initial decision path in most cases is not optimal. Therefore, a backtracking and exploration process is needed to further improve the initial decision path. The backtracking process is similar to the backward recursive method for solving the dynamic programming problem shown in Equation (1), while the addition of an exploration process distinguishes it from the backward recursive method. The backtracking and exploration is a recursive process that starts from the end node of the initial decision path as shown in Figure 6. The corresponding cost value for the end node is zero, as all queues are assumed to be cleared at this point. Set the initial decision path as the Current Best Decision Path (CBDP). The process goes back one interval for each iteration and calculates the cost value of the current node. For every node except for the end one, all branches growing up from it will be evaluated and compared with the CBDP. A cost value is defined for both arcs and paths. The delay cumulated during each interval is defined as the arc cost, and the path cost is the summation of the costs of all arcs in the path. If any of the branches have a smaller cost than the branch in the CBDP, then the branch in the CBDP will be replaced by the new branch. Otherwise, the exploration from this node will be terminated, and the process will go back one interval and set the parent node as the current node.

The ALLONS-D algorithm introduced so far is for isolated intersection control. For arterial traffic control, Porche (1997) considered two coordination methods. The first coordination method assigns different weights to each direction. Porche tested this control method on an arterial. However, it seemed that this method did not perform very well. Another coordination method Porche proposed is game theory. Porche only conceptually showed that game theory may be used for coordinating traffic signal controllers. No experiments were conducted to show if this method can really be applied to coordinate traffic signal controllers.

2.4.8 Markov Decision Process and Dynamic Programming (MDP&DP)

More recently, Yu and Recker (2006) developed a stochastic adaptive traffic signal control model. The authors formulated traffic signal control as a Markov Decision Process (MDP) and solved it by dynamic programming. MDP is a discrete time stochastic process characterized by a set of states (*S*), actions (*A*), reward function (*r*), and state-transition function (*p*). In the context of intersection traffic signal control, the state variables are the queue lengths of all approaches; the action variables are the control actions that can be taken for each state; the reward function tells the immediate reward of each action under specific state; and the state-transition probability function is time-varying and dependent on actual traffic arrivals. To solve control problems modeled as MDPs, the first step is to find the optimal value function $V^*(s)$ based on Equation (2) (Sutton and Barto, 1998).

$$V^*(s) = \max_{a \in A(s)} \left\{ \sum_{k \in S} p_{sk}^a r_{sk}^a + \gamma \sum_{k \in S} p_{sk}^a V^*(k) \right\}, \ \forall s \in S \tag{2}$$

where $a \in A(s)$; $s \in S$ is the current state and $k \in S$ is the next state after action a is taken; p_{sk}^a and r_{sk}^a are the transition probability and reward, respectively, from state *s* to state *k* after action *a* is taken; and $\gamma \in [0,1)$ is a discount factor. Equation (2) is often referred to as the Bellman optimality equation (Sutton and Barto, 1998). Based on this Bellman equation, Yu and Recker (2006) used a dynamic programming method to solve the optimal value function. $V^*(s)$. After the $V^*(s)$ is found, the control problems simply become identifying the current system state *s* and applying the control action $a \in A(s)$ that lead to the optimal value function $V^*(s)$. This mapping from system state to an action is called policy, which is a very important concept that will be used frequently in this study.

Although dynamic programming algorithm can be used to solve this MDP problem and is guaranteed to find the optimal policy (Gosavi, 2003), it needs a well-defined state-transition probability function. In practice, this state-transition probability function is difficult and cumbersome to define. In the case of intersection traffic control, the state-transition probability function is affected by actual traffic arrivals and is often time-varying. Thus, it is even more difficult to give an accurate estimation. In addition, for intersection traffic signal control applications, the number of states is usually very large. This may make the computation time of dynamic programming algorithms a serious problem (Crites and Barto, 1998; Gosavi, 2003; Yu and Recker, 2006; Barto and Mahadevan, 2003). Nevertheless, it is a legitimate attempt to use MDP to model intersection traffic control problems. Unlike DYPIC and OPAC methods that assume a deterministic state transition, MDP implicitly acknowledges the uncertainty in state transition and reflects this uncertainty by a state-transition probability function.

2.5 TRAFFIC CONTROL USING FUZZY LOGIC AND RULES

Several studies have applied fuzzy logic to traffic signal control (Zhang et al., 2005; Pappis and Mamdani, 1977; Trabia et al., 1999; Murat and Gedizlioglu, 2005; Chiu and Chand, 1993; Niittymaki and Pursula, 2000). These fuzzy logic methods use queue lengths and traffic arrivals on all approaches as inputs, and the control action is usually determined based on a number of fuzzy rules. The followings are two simple examples of fuzzy rules that are used to determine the extension of the current green phase (Zhang et al., 2005).

1. **IF** current queue length is {Short} **AND** arrival is {Low} **AND** conflicting queue length is {Medium}, **THEN** extension is {Short}
2. **IF** current queue length is {Medium} **AND** arrival is {High} **AND** conflicting queue length is {Short}, **THEN** extension is {Long}

More discussions on fuzzy logic signal control will be provided later in Chapter 4. Interested readers can also refer to (Jiang et al., 1997; MathWorks, 2007) for detailed information on fuzzy logic. An obvious advantage of using fuzzy logic for traffic signal control is that it needs minimal computation resources. Similar to pre-timed and actuated control, it is much more computationally efficient than other adaptive methods.

Some researchers also proposed rule-based and knowledge-based adaptive traffic signal control systems (Lin, 1988; Elahi et al., 1987; Owen and Stallard, 1995). For instance, Owen and Stallard (1995) developed an adaptive traffic signal control method called Generalized Adaptive Signal Control Algorithm Project (GASCAP). GASCAP has three key components: queue estimation model, a set of rules for controlling uncongested traffic, and an algorithm for producing pre-timed plans for congested traffic. The major differences distinguishing this rule-based method from other aforementioned adaptive traffic control methods are the rules for controlling uncongested traffic. GASCAP has five sets of rules as shown below.

1. **Demand Rules:** This set of rules tends to give green time to movements with the largest queue lengths. Phase sequence is not considered in making decisions using the demand rules.
2. **Progression Rules:** The purpose of progression rules is to coordinate signal timings of adjacent intersections. Progression rules give suggestions on signal states of each intersection in terms of projected traffic arrivals.
3. **Urgency Rules:** Urgency rules are used to detect saturation conditions on any of the approaches to an intersection. If any upstream detectors are on consecutively for at least 15 seconds, urgency rules will recommend the corresponding movements to be given green signal.
4. **Cooperative Rules:** Cooperative rules are employed mainly to address problems such as spillback. For two adjacent intersections, if one movement of the downstream intersection is experiencing spillback, movements of the upstream intersection aggravating the spillback will not be given green signals.
5. **Safety Rules:** Safety rules are used to ensure proper minimum green times, prevent conflicting movements from being given green signals at the same time, and so forth.

2.6 SUMMARY

This chapter reviews pre-timed, actuated, and adaptive traffic signal control. The focus of the review is adaptive signal control systems or research prototypes including UTCS, SCOOT, SCAT, DYPIC, OPAC, RHODES, ALLONS-D, and MDP&DP.

Pre-timed traffic signal control has fixed cycle length, phase sequence, and phase duration. It cannot adapt to short-term traffic flow dynamics and is only suitable for stable flow conditions. Actuated control can partially solve the problem with pre-timed control by introducing the concept of vehicle extension based on vehicle actuation information. However, actuated control still has many preset constraints and is not flexible enough.

Adaptive traffic control conceptually can better handle real time traffic flow fluctuations and significantly reduce control delay. There are two major types of adaptive traffic control systems. UTCS, SCOOT, and SCATS are typical examples of the first type of adaptive traffic control systems. The rest of the adaptive traffic signal control systems reviewed in this chapter can be generally classified as the second type. The first type of adaptive traffic signal control systems still has fixed cycle length, phase duration, phase sequence, and offset within short time periods. The control systems adaptively adjust these parameters based on real time or projected traffic conditions.

The second type of adaptive traffic control systems often model traffic control as a multi-stage problem or a MDP and solve it by using dynamic programming or branch and bound. Fuzzy logic, rule-based methods, and knowledge-based methods have also been used. The second type of adaptive traffic control systems may not have the restrictions of cycle length, fixed phase sequence, phase duration, and offset, and has attracted considerable attention in recent years. However, this type of methods still has the following problems:

1. Under certain circumstances, the excessive computation requirement makes some systems based on dynamic programming not suitable for real time applications.
2. Both the multi-stage and MDP&DP modeling approaches require accurate traffic arrival information for the next one or two minutes to determine the best control plans. This information is very difficult to obtain.
3. Although using fuzzy logic, rule-based, or knowledge-based methods has minimum computation time requirement, it is difficult to determine the optimal rules.

CHAPTER 3. REINFORCEMENT LEARNING – THEORETIC BACKGROUND

3.1 INTRODUCTION

Adaptive traffic signal control can better respond to short-term traffic fluctuations and has been the focus of recent traffic control studies. Fuzzy logic, rule-based, and knowledge-based methods have been adopted for adaptive traffic signal control. These methods are generally referred to as rule-based adaptive traffic control methods. In contrast, adaptive control methods based on dynamic programming and branch and bound are called optimization-based adaptive traffic signal control. As summarized in Chapter 2, both the rule-based and the optimization-based adaptive traffic signal control methods have their limitations.

To overcome these limitations, a multi-agent method based on reinforcement learning and neuro-fuzzy logic is proposed in this research. The new method is named as Neuro-Fuzzy Actor-Critic Reinforcement Learning (NFACRL). The intersection traffic control is still formulated as a MDP as did by Yu and Recker (2006), but the NFACRL method is used in lieu of dynamic programming to solve for the optimal value function $V^*(s)$ in Equation (2) and to find the best control policy. There are two major advantages of using the NFACRL to solve MDP problems over using dynamic programming. Firstly, the NFACRL does not require state-transition probabilities and traffic arrival predictions as inputs. It can learn the state-transition probabilities interactively from the system operations, and it can also learn the state-transition probabilities from simulations (Barto and Mahadevan, 2003). Secondly, after the NFACRL is trained, it has the same low computational requirement as rule-based methods. Thus, it is more suitable for real-time applications.

To better present the NFACRL method, a systematic introduction of the MDP and reinforcement learning is presented in this chapter, and the NFACRL method will be introduced in Chapter 4. The rest of this chapter is organized as follows: in the subsequent section, the MDP and various methods that can be used for solving MDP problems are discussed. These methods include dynamic programming, SARSA, Q-Learning, and Actor-Critic learning; following this discussion is a comprehensive review of existing applications of reinforcement learning to traffic control; and the final section summarizes this chapter.

3.2 REINFORCEMENT LEARNING

3.2.1 Reinforcement Learning Problems

In reinforcement learning, the learner is often referred to as an agent. Everything except for the agent is called environment. For different applications of reinforcement learning, the contents of agent and environment can be quite different. For traffic signal control, the agent corresponds to the traffic signal controller and the environment includes many factors such as the queue length of each approach, traffic arrivals, and current signal state. The interaction process between agent and environment is shown in Figure 7.

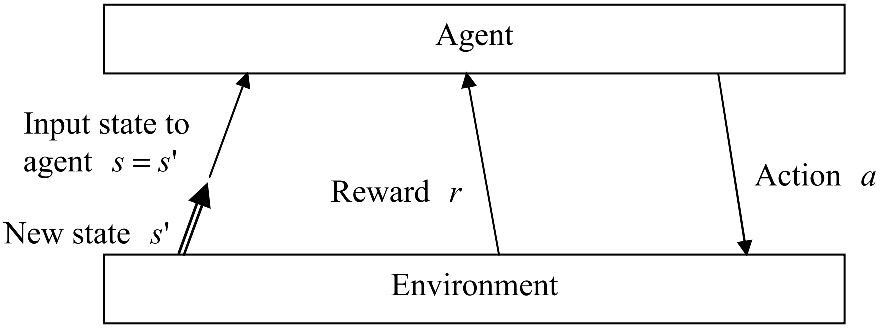


Figure 7 Agent and environment in reinforcement learning.

The interaction between agent and environment happens at any continuous time point, and theoretically the agent can make decisions at any time. For practical considerations, discrete time steps are often used. The following is a simple sample procedure to further explain how the interaction works at discrete time steps.

1. At time step *t*=0, observe the state of the environment $s_t \in S$. *S* is the collection of all possible states of the environment;
2. Based on $s_t \in S$, the agent chooses decision $a_t \in A(s_t)$. $A(s_t)$ is the collection of all available decision choices for state s_t ;
3. Apply $a_t \in A(s_t)$ to the environment at time step *t* and observe new environment state $s_{t+1} \in S$ and reward r_{t+1} at time step $t+1$;
4. Use s_t , a_t , s_{t+1} , and r_{t+1} to update the agent; and

5. Let $t = t + 1$ and go back to step 2.

At each time step, the agent chooses an action a_t based on the current environment state s_t. This mapping from states to actions is usually referred to as policy and is represented by π. In the following subsections, why this procedure works and how the agent is updated will be explained.

3.2.2 Markov Property and Markov Decision Processes

Reinforcement learning method is built based on Markov property and MDP. A stochastic process is said to have the Markov property if it satisfy the following condition:

$$\Pr\{s_{t+1} = s' \mid s_h, \forall h \le t\} = \Pr\{s_{t+1} = s' \mid s_t\} \tag{3}$$

This equation suggests that the state of the stochastic process at time step $t+1$ only depends on the state of the process at time step *t*, not on any of the states of the process at time steps $h < t$. If a stochastic process satisfies the Markov property, then it can be modeled as a MDP. A MDP is formally defined as a tuple (S, A, r, p) (Hu and Wellman, 1998; Puterman, 1994), where

1. S is the state space;
2. A is the action space;
3. r is a reward function, where $r_{ss'}^{a}$ represents the expected reward when the environment transfers from state s to state s' under the effect of action a at state s; and
4. p is a transition function, where $p_{ss'}^{a}$ represents the probability the environment will transfer from state s to state s' under the effect of action a at state s.

In addition to state, action, reward function, and state-transition probability function, another important concept of MDP is value function, which includes state value function and action value function (Sutton and Barto, 1998). State value function is a function representing how close each state is to the final (goal) state by following certain policy. In other words, it shows how good it is for the environment to be in each state under certain policy (Sutton and Barto, 1998). The goal state is generally the control objective. For traffic signal control problems, the goal state is reached when all queues are minimized. The state value function following policy π is defined in Equation (4).

$$V_\pi(s) = \sum_{s'} p_{ss'}^a \left[r_{ss'}^a + \gamma V_\pi(s') \right] \qquad \textbf{(4)}$$

where γ is a discount factor; *a* is the action decided by policy π when the environment is in state s; and $s' \in S$ are the resulted states after action *a* is taken when the environment is in state *s*. The action value function can be defined in a similarly way. If the current policy is π, the value of taking action *a* at state *s* is defined in Equation (5) (Sutton and Barto, 1998).

$$V_\pi(s,a) = \sum_{s'} p_{ss'}^a \left[r_{ss'}^a + \gamma V_\pi(s',a') \right] \qquad \textbf{(5)}$$

where *a* and a' represent the actions determined by policy π for state *s* and s', respectively. In fact, Equations (4) and (5) are equivalent.

As the state function values of each state represent how close they are to the control goal (final state), solving a control problem modeled as MDP is equivalent to finding an optimal policy $\pi *$ (a mapping from states to actions) to minimize (or maximize, depending on the problem under study) the state function values for each state. With the optimal policy $\pi *$, the following two equations hold (Sutton and Barto, 1998).

$$V^*(s) = \max_{a \in A(s)} \sum_{s'} p_{ss'}^a \left[r_{ss'}^a + \gamma V^*(s') \right] \qquad \textbf{(6)}$$

$$V^*(s,a) = \sum_{s'} p_{ss'}^a \left[r_{ss'}^a + \gamma \max_{a' \in A(s')} V^*(s',a') \right] \qquad \textbf{(7)}$$

Equations (6) and (7) are two different forms of the Bellman optimality equation. They are often used in combination with dynamic programming to solve for the optimal state or action value function. Once the optimal state or action value function is obtained, the optimal control policy

$\pi *$ can be readily determined by using Equation (8). For each state, one just needs to find the action that leads to the largest state value (Sutton and Barto, 1998).

$$\pi^*(s) = \arg\max_{a \in A(s)} \sum_{s'} p_{ss'}^{a} \left[r_{ss'}^{a} + \gamma V^*(s') \right], \ \forall s \qquad \textbf{(8)}$$

where argmax means the argument of the maximum. It returns the action that maximizes the state value of s.

It can also be shown that Equation (4) is equivalent to Equation (9), which is the summation of discounted rewards (Sutton and Barto, 1998).

$$V_{\pi}(s) = E_{\pi}\{R_t \mid s_t = s\} = E_{\pi}\left\{ \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \mid s_t = s \right\} \qquad \textbf{(9)}$$

where

R_t = summation of discounted rewards; and

r_{t+k} = reward at the (t+k)th time step.

Thus, finding the state value function $V^*(s)$ and optimal control policy $\pi *$ actually is to maximize the summation of discounted rewards shown in Equation (9).

3.2.3 Dynamic Programming for MDP

Both dynamic programming and reinforcement learning can be used to solve MDP problems. In this section, dynamic programming methods for MDP will be briefly discussed. The discussion serves as a basis for introducing reinforcement learning methods. Two dynamic programming methods have been used to solve MDP problems: policy iteration and value iteration. Policy iteration has two components, which are policy evaluation and policy improvement. Given certain policy π, policy evaluation tries to approximate the values of each state under this policy using Equation (4). The values of each state are the inputs to the policy improvement process. The purpose of policy improvement process is to adjust the policy according to the new state values, and the output of policy improvement is a new policy. Figure 8 shows how policy iteration is used to find the optimal policy for MDP problems (Sutton and Barto, 1998), where $\pi(s)$ is the action decided by policy π for state *s*.

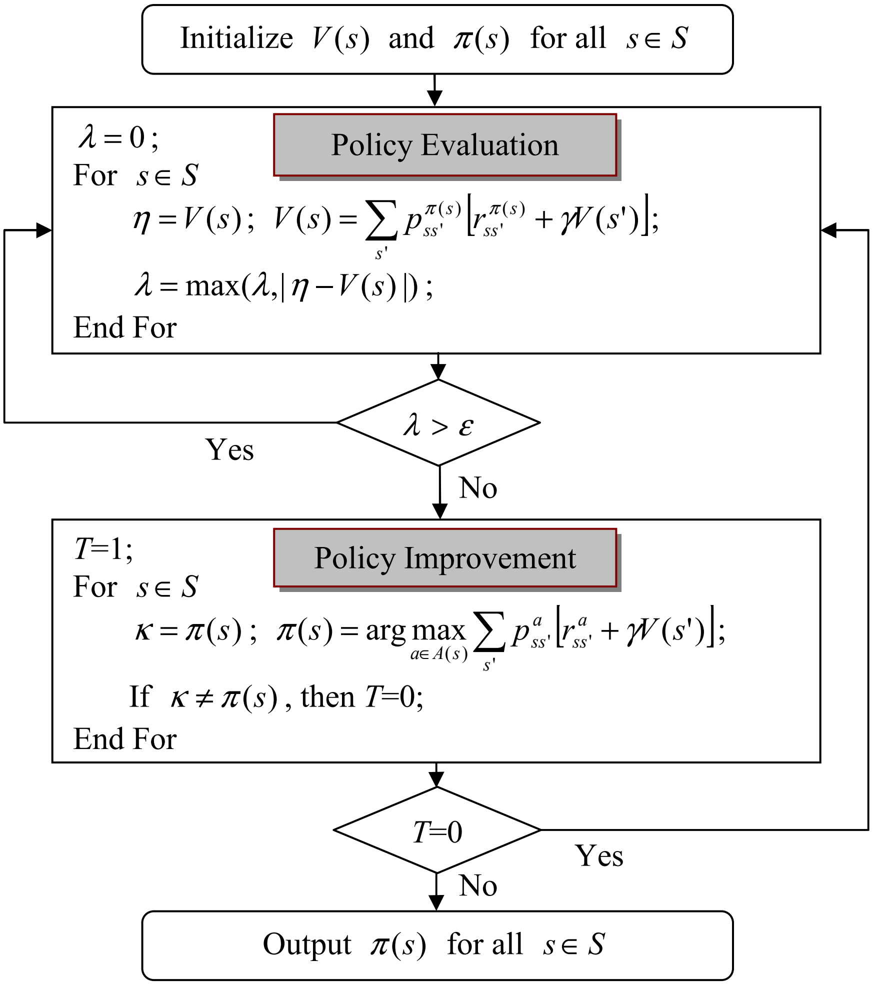

Initialize $V(s)$ and $\pi(s)$ for all $s \in S$
Policy Evaluation
$\lambda = 0$;
For $s \in S$
$\eta = V(s)$; $V(s) = \sum_{s'} p_{ss'}^{\pi(s)} \left[r_{ss'}^{\pi(s)} + \gamma V(s') \right]$;
$\lambda = \max(\lambda, |\eta - V(s)|)$;
End For
$\lambda > \varepsilon$
Yes
No
Policy Improvement
T=1;
For $s \in S$
$\kappa = \pi(s)$; $\pi(s) = \arg\max_{a \in A(s)} \sum_{s'} p_{ss'}^{a} \left[r_{ss'}^{a} + \gamma V(s') \right]$;
If $\kappa \neq \pi(s)$, then *T*=0;
End For
T=0
Yes
No
Output $\pi(s)$ for all $s \in S$

Figure 8 Policy iteration of dynamic programming.

Both policy evaluation and policy improvement need to visit each state multiple times and are computationally inefficient. Compared to policy iteration method, the value iteration method effectively integrates policy evaluation and policy improvement and has better computational efficiency. The value iteration method is illustrated in Figure 9.

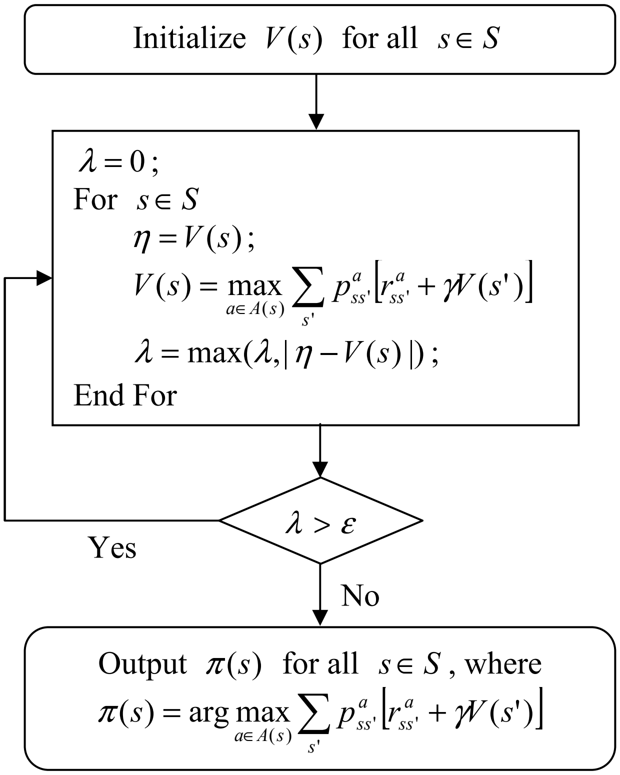

Initialize $V(s)$ for all $s \in S$
$\lambda = 0$;
For $s \in S$
$\eta = V(s)$;
$V(s) = \max_{a \in A(s)} \sum_{s'} p_{ss'}^{a} \left[r_{ss'}^{a} + \gamma V(s') \right]$
$\lambda = \max(\lambda, |\eta - V(s)|)$;
End For
$\lambda > \varepsilon$
Yes
No
Output $\pi(s)$ for all $s \in S$, where
$\pi(s) = \arg\max_{a \in A(s)} \sum_{s'} p_{ss'}^{a} \left[r_{ss'}^{a} + \gamma V(s') \right]$

Figure 9 Value iteration of dynamic programming.

Although the policy iteration and value iteration methods are different, both of them can guarantee the optimal solutions if accurate knowledge of the probability $p_{ss'}^{a}$ is provided (Sutton and Barto, 1998). For many practical problems such as adaptive traffic signal control (Yu and Recker, 2006), it is difficult to obtain accurate estimation of state transition probabilities. In addition, the dynamic programming method may have considerably high computational requirements if the state space is large. It would be great if some methods can solve the MDP problems without relying on the state transition probabilities and also have a low computational requirement. Fortunately, reinforcement learning can meet both requirements and will be introduced in the following subsections.

3.2.4 SARSA for MDP

3.2.4.1 SARSA Reinforcement Learning

SARSA is one of the three major reinforcement learning methods. The other two reinforcement learning methods are Q-Learning and Actor-Critic reinforcement learning. All these three reinforcement learning methods are based on a Temporal-Difference (TD) error (Sutton and Barto, 1998). The TD error is calculated in terms of observed changes from the environment, and is used to update the state value function and the action value function.

The following equation is used in SARSA to update the action value function (Sutton and Barto, 1998).

$$V_\pi(s_t, a_t) = V_\pi(s_t, a_t) + \phi\left[r_{t+1} + \gamma V_\pi(s_{t+1}, a_{t+1}) - V_\pi(s_t, a_t)\right] \tag{10}$$

where

s_t, s_{t+1} = observed states of the environment at time steps *t* and *t*+1, respectively;

ϕ = learning rate;

$r_{t+1} + \gamma V_\pi(s_{t+1}, a_{t+1}) - V_\pi(s_t, a_t)$ = TD error;

a_t, a_{t+1} = actions for state s_t and s_{t+1}, respectively;

r_{t+1} = observed reward at time step *t*+1; and

γ = discount factor.

By comparing Equations (10) and (6), one can see that both the dynamic programming and the SARSA method use the one-step reward and the state or action value of the resulted state to update the state or action value of the current state. The major difference is that dynamic programming method requires predefined state transition probabilities $p_{ss'}^{a}$, while the SARSA method does not. The SARSA method introduces a learning rate ϕ and updates the action value by a linear combination of its current action value and the TD error. By using the SARSA method, $V_\pi(s,a)$ can converge to the optimal value $V^*(s,a)$ asymptotically (Sutton and Barto, 1998). After the action value function has converged, the following Equation (11) is used to extract the optimal policy $\pi*$ from the action value function.

$$\pi^*(s) = \arg\max_{a \in A(s)} V^*(s,a) \tag{11}$$

Before using Equation (10), a reward function r_{t+1} has to be properly defined. The calculation of the reward function involves direct interactions between the control agent and the environment. This means that finding the optimal control policy requires implementation of the control system in real world or more likely through simulation. Using simulation as an example, the SARSA method is illustrated in Figure 10. This method can be better understood by taking a look at Figure 7, which shows the interaction between agent and environment.

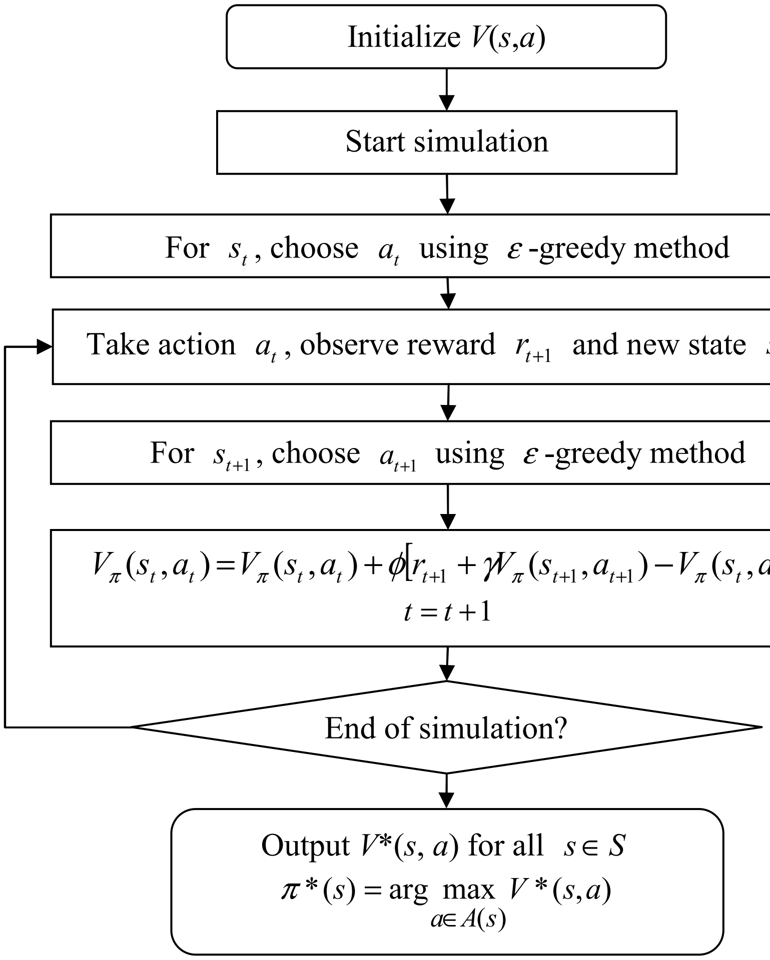

Initialize $V(s,a)$
Start simulation
For s_t, choose a_t using ε-greedy method
Take action a_t, observe reward r_{t+1} and new state s_{t+1}
For s_{t+1}, choose a_{t+1} using ε-greedy method
$V_\pi(s_t,a_t)=V_\pi(s_t,a_t)+\phi[r_{t+1}+\gamma V_\pi(s_{t+1},a_{t+1})-V_\pi(s_t,a_t)]$
$t=t+1$
End of simulation?
Output $V^*(s, a)$ for all $s \in S$
$\pi^*(s) = \arg\max_{a \in A(s)} V^*(s,a)$

Figure 10 SARSA for MDP.

3.2.4.2 Action Selection Methods

There are several methods that can be used for action selection given the current state of the environment. These methods include greedy, ε-greedy and softmax action selection methods

(Sutton and Barto, 1998). In this research, the ε-greedy action selection method is chosen due to its simplicity and effectiveness. The ε-greedy method is based on the greedy action selection method. For given state *s*, the greedy method always chooses an action with the largest action value *V*(*s*,*a*). However, sometimes two actions a_1 and a_2 may have approximately the same action value, and $V(s,a_1)$ is just slightly larger than $V(s,a_2)$. By using the greedy method, action a_1 will always be chosen. In fact, a_2 may be better than a_1, and $V(s,a_2)$ will be larger than $V(s,a_1)$ after one more value updating. To address this problem, an exploration strategy is incorporated into the greedy method and results in the ε-greedy selection method. For the ε-greedy selection method, actions with the largest action values are selected for most of the time. The remaining actions are selected with a small probability $\frac{\varepsilon}{|A(s)|}$. This method is described in Equation (12) more clearly (Sutton and Barto, 1998).

$$\pi(s,a_i) = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{|A(s)|} & \text{if } a_i = \arg\max\limits_{a\in A(s)} V(s,a) \\ \dfrac{\varepsilon}{|A(s)|} & \text{otherwise} \end{cases} \tag{12}$$

where

$\pi(s,a_i) =$ the probability that action a_i will be chosen for state *s*;

$\varepsilon =$ a small value; and

$|A(s)| =$ total number of possible actions for state *s*.

3.2.5 Q-Learning for MDP

Q-Learning is similar to SARSA. It uses Equation (13) to update action values.

$$V(s_t,a_t) = V(s_t,a_t) + \phi\left[r_{t+1} + \gamma \max_{a_{t+1}\in A(s_{t+1})} V(s_{t+1},a_{t+1}) - V(s_t,a_t)\right] \tag{13}$$

Equation (13) is slightly different from Equation (10), which is used by SARSA to update action values. SARSA is considered as an on-policy method while Q-Learning is an off-policy method.

A formal expression of the difference between on-policy and off-policy methods is that "*the distinguishing feature of on-policy methods is that they estimate the value of a policy while using it for control. In off-policy methods these two functions are separated. The policy used to generate behavior, called the behavior policy, may in fact be unrelated to the policy that is evaluated and improved, called the estimation policy. An advantage of this separation is that the estimation policy may be deterministic (e.g., greedy), while the behavior policy can continue to sample all possible actions*" (Sutton and Barto, 1998).

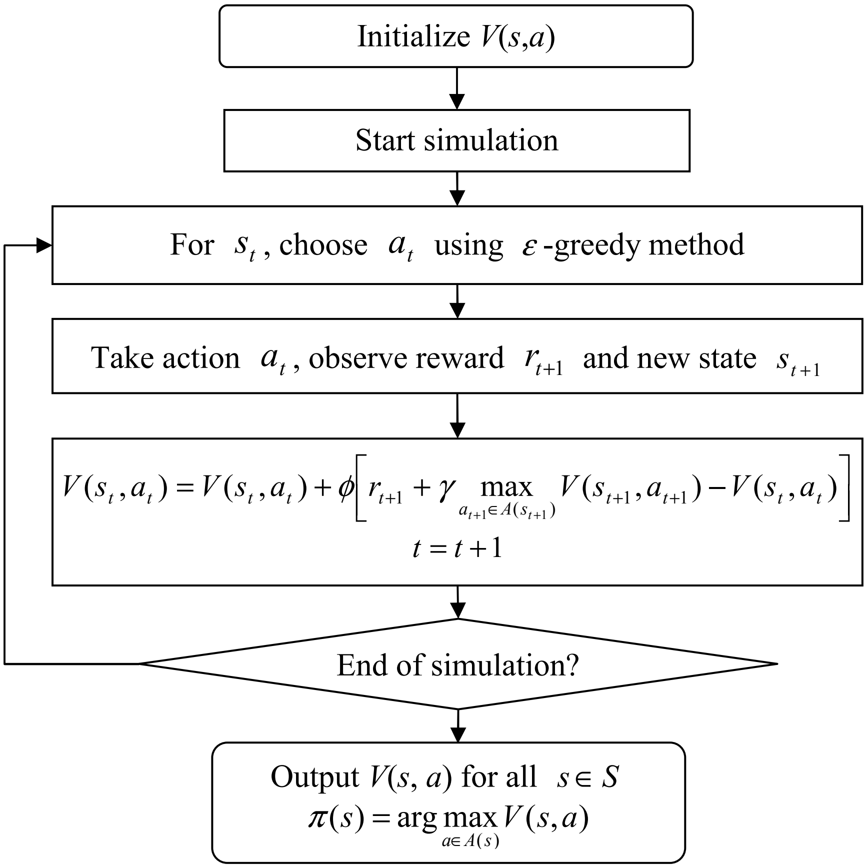

Initialize $V(s,a)$
Start simulation
For s_t, choose a_t using ε-greedy method
Take action a_t, observe reward r_{t+1} and new state s_{t+1}
$V(s_t,a_t)=V(s_t,a_t)+\phi\left[r_{t+1}+\gamma \max_{a_{t+1}\in A(s_{t+1})} V(s_{t+1},a_{t+1})-V(s_t,a_t)\right]$
$t=t+1$
End of simulation?
Output $V(s, a)$ for all $s\in S$
$\pi(s)=\arg\max_{a\in A(s)} V(s,a)$

Figure 11 Illustration of the Q-Learning algorithm.

The Q-Learning results are stored in the action value function $V(s,a)$, which is often in a table form as shown in Table 1. This table is called Q-Table. Note that the number of actions for different states could be different. When the environment is in certain state, in terms of Equation (11), the best action is determined by finding the corresponding row in Table 1 for the current state and then locating the action with the highest action value in that row.

Table 1 Learning results of Q-Learning method

State #	Action #		
	1	2	…
1	8	9	…
2	11	6	…
…	…	…	…

For each cell in Table 1, its action value is updated by Equation (13) using the iteration process shown in Figure 11. To approximate the true action value, the corresponding cell needs to be visited as often as possible. However, when the state or action space is large, visiting each cell many times requires considerable computation time. This is often referred to as the curse of dimensionality problem. Thus, the traditional Q-Learning may not be directly applicable for problems with large state or action space. Another relevant problem with the traditional Q-Learning based on Q-Table is generalization. During the learning process some cells in Table 1 may only be visited one or two times even though their neighboring cells are visited many times. This may produce inaccurate action values for those less visited cells. When the environment happens to be in the corresponding states during actual application, it is possible that suboptimal actions will be chosen that may lead to poor control performance. In fact, it is reasonable to expect neighboring states to have similar actions values. However, by using this traditional Q-Learning method, action values of neighboring cells cannot be used to update the action values of those less visited cells.

3.2.6 Actor-Critic Reinforcement Learning for MDP

Another well known reinforcement learning method is Actor-Critic Reinforcement Learning (ACRL) (Barto et al., 1983; Gajjar et al., 2003; Bhatnagar and Panigrahi, 2006; Borkar, 2005). ACRL has a more complex structure than SARSA and Q-Learning. For SARSA and Q-Learning, optimal policies are stored in action value functions. After the optimal action value functions are obtained, Equation (11) is used to extract the optimal policies from the optimal action value functions. Storing optimal policies in action value functions is straightforward and easy to understand. For ACRL, the policy and state value function are stored separately. Although this increases the complexity of the method and makes it difficult to analyze, the ACRL method does have two major advantages as pointed out by Sutton and Barto (1998).

For the ACRL method, the unit used to store policy is called *Actor*, and the unit used to store state value function is referred to as *Critic*. *Actor* and *Critic* can use different techniques such as neural networks and fuzzy logic (Berenji and Khedkar, 1992; Lin and Lee, 1994) to store policy and state value function. To simplify the introduction of ACRL, a generic description of this method is provided here. The following figure has been used by several researchers to illustrate the architecture of ACRL (Sutton and Barto, 1998; Gajjar et al., 2003).

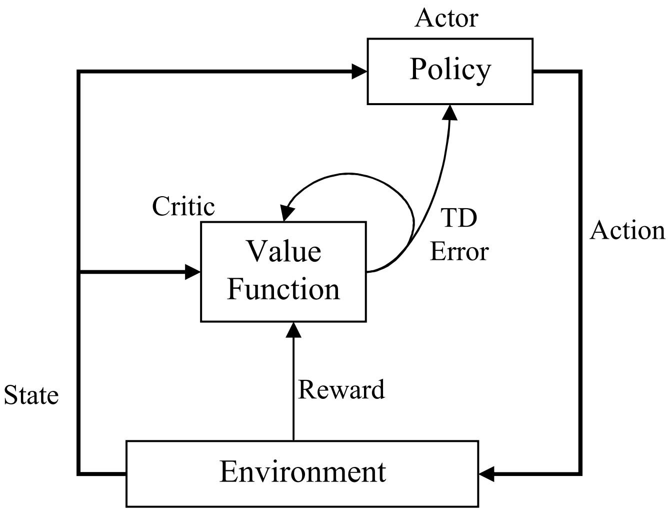


Figure 12 Architecture of ACRL method (Sutton and Barto, 1998; Gajjar et al., 2003).

At any decision point *t*, the *Actor* generates an action a_t based on the current environment state s_t. This action is then applied to the environment. Under the effect of action a_t, the environment may change accordingly. A reward value r_{t+1} and a new state s_{t+1} can be obtained. Also, a TD error is calculated using Equation (14).

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{14}$$

For SARSA and Q-Learning, the TD error is defined based on action values and used for updating action value functions, since for them policies are stored in action value functions. For

ACRL, the TD error is used to update both state value function and policy using Equations (15) and (16), respectively.

$$V(s_t) = V(s_t) + \alpha\delta_t \tag{15}$$

$$V(s_t, a_t) = V(s_t, a_t) + \beta\delta_t \tag{16}$$

where

$\alpha, \beta =$ step-size parameters; and

$V(s_t, a_t) =$ action value representing the preference to choose action a_t when the environment is in state s_t.

For an action, if its corresponding TD error is positive, then the preference of choosing this action should be reinforced. Otherwise, the preference of choosing this action should be decreased.

3.2.7 Comparison between Dynamic Programming and Reinforcement Learning

Both dynamic programming and reinforcement learning can be used for solving MDP problems. A major difference is that dynamic programming has to have an accurate model of the MDP problems. The reward and state transition probability functions need to be exactly known, while reinforcement learning methods such as SARSA, Q-Learning, and ACRL do not require perfect models of the MDP problems under study. They can implicitly learn the state transition probability functions and observe rewards from interactions between the agent and the environment. This property of reinforcement learning is very important and useful. For many practical problems that can be modeled as MDPs, it is usually very difficult to estimate the state transition probability and reward functions accurately. For instance, if an intersection has four approaches and eight movements as in Figure 2, assuming the queue lengths of each movement can be categorized into 5 classes, then there would be $5^8 \approx 39 \times 10^4$ possible states if one uses queue lengths as state variables. Finding the state transition probability function for this problem would be very computationally intensive. In practice, reinforcement learning would be a better choice for such problems.

3.3 REVIEW OF EXISTING INTERSECTION TRAFFIC CONTROL STUDIES USING REINFORCEMENT LEARNING

3.3.1 Traffic Control Using SARSA

Thorpe (1997) conducted one of the pioneering studies on traffic signal control using reinforcement learning. In his study, Thorpe tested a SARSA control method on a simple 4×4 grid traffic network with 16 intersections. Each intersection had four approaches and each approach had exactly one lane. The distance between any two intersections was 440 feet. Left-turn phase was not considered. Thus, each intersection only had two through (right-turn movements were combined with the corresponding through movements) phases. Each of the 16 intersections was controlled by one agent, and there was no coordination explicitly considered. The test was carried out based on a self-developed simulation program. A 2-second yellow time and a 1-second all red time were used between phase switches for safety consideration. The control decision was made at a 1-second interval.

One key issue for reinforcement learning application is how to define the state of the environment. In Thorpe's study, four different methods were used to define the environment state, which were:

1. **Vehicle count representation:** vehicle count representation first summed up vehicle counts in each direction (east-west and north-south bounds) and then categorized them into 10 states. Since there were two control actions, the total number of states using the vehicle count representation method was 200.
2. **Fixed-distance representation:** in Thorpe's study, each approach of an intersection was 440 feet long, which was divided evenly into four segments. Each segment had two states: with and without vehicles on it. This representation method finally resulted in 512 states.
3. **Variable-distance representation:** this representation was almost the same as the fixed-distance representation except for how each approach was divided into segments. For the variable-distance method, each approach was divided into four segments at distances 50, 110, and 220 feet starting from the stop line. The total number of states was also 512.

4. **Count/duration representation:** the count/duration representation was based on the vehicle count representation. In this case, the vehicle counts were classified into 8 groups. Since there were two directions and two signal states (green or not green), the total number of states was 128. The action space in this representation was expanded to 16, which consisted of different minimum green times for each direction. Thus, the total number of state action pairs became 2048.

Another very important issue that affects the performance of reinforcement learning is the definition of reward. Thorpe used two different definitions of reward. For the first definition, if at each decision point the environment state was not the goal state (goal state: all vehicles were cleared), then the value for the action taken at the previous decision point was updated by subtracting 1. For the second definition, the reward was defined in Equation (17).

$$r = \text{constant} + moved - stopped \tag{17}$$

where

constant = a constant value that was set to -3 in Thorpe's study;

moved = number of vehicles that have passed the intersection from approaches being given green signal; and

stopped = number of vehicles that have been stopped due to a red signal in the last interval.

Thorpe tested the four state representation methods and two reward definitions based on computer simulation, and compared the SARSA control method with a number of other strategies such as greatest-volume strategy and pre-timed control. A greedy action selection method was used to choose actions for each state. The test was conducted under different traffic demand levels. The results showed that the SARSA method with count/duration state representation performed the best in terms of average travel time. For average stopped time, the SARSA method with fixed-distance and variable-distance representations performed better than the other methods. Thorpe also showed that for count/duration representation, the best reward definition was the first one, while for fixed-distance and variable-distance representations, the best reward definition was the second one.

There were a few problems not well addressed in Thorpe's study. Firstly, Thorpe used the greatest-volume and pre-timed control strategies as benchmarks for comparison with the SARSA control method. However, he did not describe clearly how the greatest-volume and pre-timed strategies were designed. Secondly, two-phase control without considering left-turn movements is very uncommon in practice unless maybe for urban grids with one-way streets and left-turn restriction. Finally, Thorpe did not use any commonly used simulation tools such as CORSIM (FHWA, 1997) or VISSIM (PTV, 2007b) for the comparison of different control strategies. These commonly used microscopic traffic simulation packages should provide a more accurate traffic environment and more rigorous performance measure calculations, consequently a more convincing results comparison. In spite of all these problems, Thorpe's study provided useful information for conducting further research on this topic.

3.3.2 Adaptive Traffic Signal Control Using Q-Learning

Abdulhai et al. (2003) proposed a truly adaptive traffic signal control strategy based on Q-Learning. In their study, they discussed how to apply Q-Learning to both isolated intersection and arterial traffic control, and provided testing results for isolated intersection control. However, the authors did not provide testing results for arterial control, which are of primary interests to many traffic engineering researchers and practitioners.

For the application of Q-Learning to isolated intersection control, Abdulhai et al. considered an intersection without turning movements. Therefore, there were only two phases. Different from most of the adaptive traffic signal control methods reviewed in Chapter 2, Abdulhai et al. considered a fixed cycle length for the isolated intersection control. Since the isolated intersection was controlled by a fixed cycle length strategy and there were just two phases, in each cycle there was only one decision to make, and the action set was whether to make the phase switch or not. In their study, Abdulhai et al. used total delay accumulated between two consecutive phase switch points as the reward. As for state variables, they used queue lengths on each approach and the elapsed time since last phase switch. However, they did not make it clear how the states were defined in terms of queue lengths. If an approach can store up to 20 vehicles, then for this approach alone there could be 21 states in terms of the number of vehicles in the storage bay. When the state space is large, there could be a generalization problem. In their study, Abdulhai et al. used a technique called Cerebellar Model Articulation Controller (CMAC) for storing and generalizing the learned action value function.

The Q-Learning control was tested and compared with pre-timed control under different traffic flow patterns. The results showed that under uniform and constant-ratio flow conditions, Q-Learning control performed approximately the same as pre-timed control. While for variable traffic flow conditions, Q-Learning control reduced average delay by more than 50% compared to pre-timed control. Although the results were very promising under variable flow conditions, the authors did not mention if the pre-timed control was optimized or not. In addition, this comparison was only for two-phase control. In practice, most intersections have four-phase signal operation.

In their study, Abdulhai et al. also proposed a general framework for arterial and network traffic control using Q-Learning. They suggested including queue information from adjacent intersections as the state variables for the current intersection control agent, to facilitate the coordination among these intersections. However, they acknowledged that this may considerably increase the state space and make the training time of the Q-Learning method intractable.

3.3.3 Signal Control Using Actor-Critic Reinforcement Learning

Bingham (2001) proposed an isolated intersection traffic control strategy based on a Generalized Approximate Reasoning-based Intelligent Control (GARIC) algorithm developed by Berenji and Khedkar (1992). The GARIC algorithm was essentially an Actor-Critic Reinforcement Learning (ACRL) method. Bingham considered a very simple isolated intersection as the test bed. This intersection had two one-way streets. The author used two state variables. The first state variable *APP* was the number of vehicles in the movement being given green signal. The second state variable *QUE* was the number of vehicles in the movement being given red signal. The *APP* and *QUE* were the inputs to both the *Actor* and *Critic*. The action output of the ACRL was a continuous value, which represented the amount of extension that should be given to the current green signal.

A TD error defined in Equation (14) is used to update the *Actor* and *Critic* at each learning step. In Bingham's study, the *Critic* is a fully connected feed-forward neural network and the *Actor* the is a fuzzy logic controller; r_{t+1} was defined as minus total vehicle delay between two consecutive decision points; and $V(s_t)$ and $V(s_{t+1})$ were the outputs of the *Critic* when the environment was in state s_t and s_{t+1}, respectively (Berenji and Khedkar, 1992; Bingham, 1998).

Bingham compared the control performance of the original and the updated fuzzy logic controllers (*Actor* in her study) using a simulation program called HUTSIM. However, she did not compare the proposed method with any other pre-timed or actuated control.

3.3.4 Other Signal Control Using Reinforcement Learning

Choy et al. (2003a; 2003b) and Srinivasan and Choy (2006) modeled a regional traffic signal control problem using reinforcement learning. In both studies, each intersection was controlled by a pre-timed controller. Reinforcement learning was mainly used to dynamically update cycle lengths and other parameters of the pre-timed controllers in response to varying traffic flow conditions. The methods they proposed are similar to those investigated in the UTCS projects, and are strictly not demand-responsive adaptive control recommended by Gartner (1982; 1983).

3.4 PROBLEMS WITH THE EXISTING STUDIES

Existing studies applying reinforcement learning to intersection traffic control provide useful information for future research in this area. However, there are still several important issues that need to be further investigated.

First of all, reinforcement learning is based on the MDP framework. When the state space dimension is large, some reinforcement learning methods will suffer from the curse of dimensionality problem (Yu and Recker, 2006; Sen and Head, 1997).

Secondly, the coordination of different control agents has not been adequately investigated in previous studies. Bingham (2001) and Abdulhai et al. (2003) only reported results for isolated intersections. Although Thorpe (1997) did apply his proposed method to a 4×4 network, coordination was not explicitly considered or discussed in his study.

Thirdly, most previous studies used isolated intersections and networks with very simple structures for testing. Thorpe (1997) tested his reinforcement learning control method on a network without considering left-turn phases. Abdulhai et al. (2003) evaluated their adaptive reinforcement learning traffic control method on an isolated intersection without turning vehicles. Bingham (2001) evaluated an ACRL traffic controller on an isolated intersection consisting of two one-way streets. In all these studies, there were only two phases considered for each intersection. In reality, most intersections have eight movements and are typically controlled by three- or four-phase signal operation.

Finally, most of the previous studies did not use a widely accepted traffic simulation platform for algorithm evaluations. Thorpe (1997) used a simulation program developed by himself. Bingham (2001) used the HUTSIM developed by the Helsinki University of Technology. In the study by Abdulhai et al. (2003), they did not mention which simulation program was used.

3.5 SUMMARY

This chapter focuses on introducing reinforcement learning methods and their recent applications to intersection traffic control. Markov property and MDP are first discussed, which are the modeling bases of reinforcement learning methods. After that, dynamic programming and three reinforcement learning methods are introduced and compared. The three reinforcement learning methods discussed are SARSA, Q-Learning, and ACRL. Comparison shows that reinforcement learning has certain advantages over dynamic programming for intersection traffic control problems modeled as MDPs. This is mainly because reinforcement learning does not need to have perfect models of the systems to be controlled, and can implicitly learn the state transition probability function from the interactions between environments and agents.

Some recent applications of reinforcement learning to traffic signal control are reviewed. Several problems with these existing applications are identified and discussed. Despite of these limitations, the existing studies provide much useful information for this study and future research in this area. In the next chapter, a new reinforcement learning signal control method based on neural networks and fuzzy logic will be developed, and details about how to apply this new signal control method to both intersection and arterial control are also presented.

CHAPTER 4. DEVELOPMENT OF A MULTI-AGENT BASED NEURO-FUZZY ARTERIAL TRAFFIC SIGNAL CONTROL SYSTEM

4.1 INTRODUCTION

In Chapters 2 and 3, a comprehensive review of intersection traffic signal control, reinforcement learning, and reinforcement learning for adaptive traffic signal control was presented. The review showed that adaptive traffic signal control is conceptually more efficient than pre-timed and actuated controls. Many adaptive traffic signal control methods have been developed in the past. Compared to traditional adaptive traffic control methods such as OPAC and RHODES, modeling adaptive traffic control as a MDP problem has two major benefits. It does not require accurate traffic arrival predictions and can better account for the uncertainty in state transition by introducing a state transition probability matrix. The review also showed the advantages of using reinforcement learning over dynamic programming for adaptive intersection traffic control modeled as a MDP problem. In the meantime, problems with reinforcement learning and its applications to adaptive traffic control were also discussed in details. To address these problems, a Neuro-Fuzzy Actor-Critic Reinforcement Learning (NFACRL) method will be developed for both intersection and arterial traffic controls in this chapter. The NFACRL method is designed to consider more practical traffic signal control problems that have more than two phases with the consideration of left-turn movements. Compared to the traditional reinforcement learning methods such as Q-Learning, the NFACRL method can better handle dimensionality and generalization problems. Coordination of intersection traffic control agents will be also taken into account. In addition, the NFACRL method will be compared with optimized pre-timed and actuated control strategies using a commonly-accepted microscopic traffic simulation tool.

In the following sections, a concise description of fuzzy logic control and neural networks is first presented. The NFACRL method is then introduced and two implementation schemes for isolated intersection traffic control using the NFACRL are proposed. Following that are the discussions of coordination strategies and the development of a multi-agent arterial adaptive traffic control system using the NFACRL.

4.2 FUZZY LOGIC CONTROL AND NEURAL NETWORKS

4.2.1 Fuzzy Logic Control

4.2.1.1 Fuzzy Sets and Fuzzy State Representation

Before introducing fuzzy sets and fuzzy state representation, an example of discrete state representation is presented. Discrete state representation has been used in several previous studies (Yu and Recker, 2006; Thorpe, 1997). It uses crisp boundaries to partition observed state values into different categories. For example, if a set of boundary values shown in Table 2 is used for partitioning state values, then a queue of 6 vehicles or less will be classified as "Uncongested", while a queue of 7 vehicles or more will be classified as "Congested". Although the difference between queues of 6 and 7 vehicles is almost negligible, these two queues belong to distinctly different states according to the discrete state representation. Also for queues of 1 vehicle and 6 vehicles, although a queue of 6 vehicles is six times as long as a queue of only 1 vehicle, they all belong to the state "Uncongested" and are treated in the same way. Obviously, it is problematic to use such partition method for categorizing input state values. One way to address this problem is to use smaller partition intervals, but this will considerably increase the number of states and make the reinforcement learning problem intractable.

Table 2 Threshold values for each category

	Uncongested	Congested
Threshold values	<=6 vehicles	>=7 vehicles

This problem can be better solved by using fuzzy sets and fuzzy set representation. In the fuzzy set representation, each category in Table 2 will have a membership function associated with it. For a given queue length, there are two membership function values representing the degrees that the given queue belongs to each category. Using membership function values can avoid classifying a queue into a category absolutely. To explain how this works, the concept of fuzzy sets is formally defined below (Jiang et al., 1997).

$$A = \{(x_i, \mu_A(x_i)) \mid x_i \in X\} \tag{18}$$

where

A $=$ fuzzy set;

X $=$ a collection of values, which can be discrete or continuous and is often referred to as universe of discourse;

x_i $=$ input value to fuzzy set *A*; and

$\mu_A(x_i)$ $=$ membership function for fuzzy set *A*. Its values are always between 0 and 1 and represent the degrees that each x_i belongs to the current fuzzy set.

There are many types of membership functions, including Triangular, Trapezoidal, and Gaussian membership functions as defined in Equations (19) through (21).

Triangular membership function (Jiang et al., 1997)

$$\mu_A(x) = \begin{cases} (x-a)/(b-a) & x \in [a,b] \\ (c-x)/(c-b) & x \in [b,c] \\ 0 & \text{else} \end{cases}, \text{ where } a<b<c \tag{19}$$

Trapezoidal membership function (Jiang et al., 1997)

$$\mu_A(x) = \begin{cases} (x-a)/(b-a) & x \in [a,b] \\ 1 & x \in [b,c] \\ (d-x)/(d-c) & x \in [c,d] \\ 0 & \text{else} \end{cases}, \text{ where } a<b\le c<d \tag{20}$$

Gaussian membership function (Jiang et al., 1997)

$$\mu_A(x) = \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\} \tag{21}$$

A number of operations are defined for fuzzy sets, including union and intersection. The union of fuzzy sets *A* and *B* is denoted as $A \cup B$, and the membership function for the resulted new fuzzy set is defined as (Jiang et al., 1997; Ross, 2004)

$$\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x) = \max\{\mu_A(x), \mu_B(x)\} \quad (22)$$

The intersection of fuzzy sets *A* and *B* is denoted as $A \cap B$, for which the new membership function is defined as (Jiang et al., 1997; Ross, 2004)

$$\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x) = \min\{\mu_A(x), \mu_B(x)\} \quad (23)$$

If the fuzzy sets and fuzzy set representation is used to classify a queue into two categories as shown in Table 2, then x_i represents the queue length; *X* denotes all possible discrete queue length values; and there are two fuzzy sets *U* and *C*, which stands for "Uncongested" and "Congested" conditions, respectively. For fuzzy set *C*, if the membership function is a Triangular function with parameters *a*=5, *b*=7, and *c*=9, then given queue lengths 6 and 7, their corresponding membership function values are 0.5 and 1, respectively. Compared with the results from the discrete state representation presented at the beginning of this section, the results from the fuzzy set representation are more rational. Moreover, the number of states can be kept within a reasonable range. The process of applying the fuzzy set representation and calculating the membership function values is often called fuzzification.

4.2.1.2 Fuzzy Rules and Reasoning

Using fuzzy sets, state variables can be written in the following linguistic term

- Current Queue Length is {*A*}

where "Current Queue Length" is a state variable and also called a linguistic variable in this case. *A* is a linguistic value corresponding to a fuzzy set that could denote "Uncongested" or "Congested" condition. For each observed value of "Current Queue Length", there is a fuzzy membership function value associated with the linguistic term "Current Queue Length is {*A*}", and this membership function value is also called degree of compatibility. Action variables can also be expressed in the same way by using linguistic term. For instance,

- Green Time Extension is $\{G\}$

where "Green Time Extension" is an action variable (also a linguistic variable) and G is a linguistic value corresponding to a fuzzy set that could denote "Short" or "Long".

Based on linguistic terms, traffic control can be realized using fuzzy rules that consist of linguistic terms as in the following examples:

- **IF** Current Queue Length (q) is {Short} **AND** Arrival (a) is {Low} **AND** Conflicting Queue Length (c) is {Medium}, **THEN** Extension (e) is {Short}
- **IF** Current Queue Length (q) is {Medium} **AND** Arrival (a) is {High} **AND** Conflicting Queue Length (c) is {Short}, **THEN** Extension (e) is {Long}

A fuzzy rule usually has two components: antecedent and consequence. In the first fuzzy rule presented above, linguistic terms "Current Queue Length (q) is {Short}", "Arrival (a) is {Low}", and "Conflicting Queue Length (c) is {Medium}" are antecedents, while the last linguistic term "Extension (e) is {Short}" is a consequence (Jiang et al., 1997). Each antecedent or consequence has a degree of compatibility, which in fact is the fuzzy membership function value for the corresponding linguistic term.

Each fuzzy rule has a numerical value associated with it. This value is called firing strength. Firing strength is calculated based on the degrees of compatibility of antecedents. For the first fuzzy rule in the previous paragraph, the degrees of compatibility are $\mu_{Short}(q)$, $\mu_{Low}(a)$, and $\mu_{Medium}(c)$. There are basically two methods to calculate the firing strength (Jiang et al., 1997). The first one is to calculate it as the intersection of the degrees of compatibility of all antecedents, which is in Equation (24).

$$FS_{Rule\,1} = \mu_{Short}(q) \wedge \mu_{Low}(a) \wedge \mu_{Medium}(c) \tag{24}$$

The other method is to calculate it as the product of the degrees of compatibility of all antecedents (Jiang et al., 1997) as shown in Equation (25)

$$FS_{Rule\,1} = \mu_{Short}(q) \times \mu_{Low}(a) \times \mu_{Medium}(c) \qquad (25)$$

Firing strength is used for calculating the output of a fuzzy rule, and the output is an induced consequent fuzzy set. The entire process from fuzzy rules to the induced consequent fuzzy set is called fuzzy reasoning and is illustrated in Figure 13, in which there are two fuzzy rules. The first step of fuzzy reasoning is to calculate the degrees of compatibility of each antecedent. This step is also called fuzzification. Based on these degrees of compatibility, the second step is to calculate the firing strength of each fuzzy rule. As discussed before, there are mainly two different methods for calculating the firing strength. For the example in Figure 13, Equation (24) is used to calculate firing strengths. Each fuzzy rule has a consequence. The consequences in this example are two fuzzy sets: $\mu_{Short}(e)$ and $\mu_{Long}(e)$. The calculated firing strengths are then applied to these two consequences to obtain induced consequent fuzzy sets. The two induced consequent fuzzy sets are represented by the shaded areas in Figure 13. For fuzzy reasoning problems with two or more fuzzy rules, a union operation in Equation (22) is usually used to merge all induced consequent fuzzy sets to obtain a combined induced consequent fuzzy set. For the example shown in Figure 13 with two fuzzy rules, the combined induced consequent fuzzy set is $\mu_{Extension}(e)$.

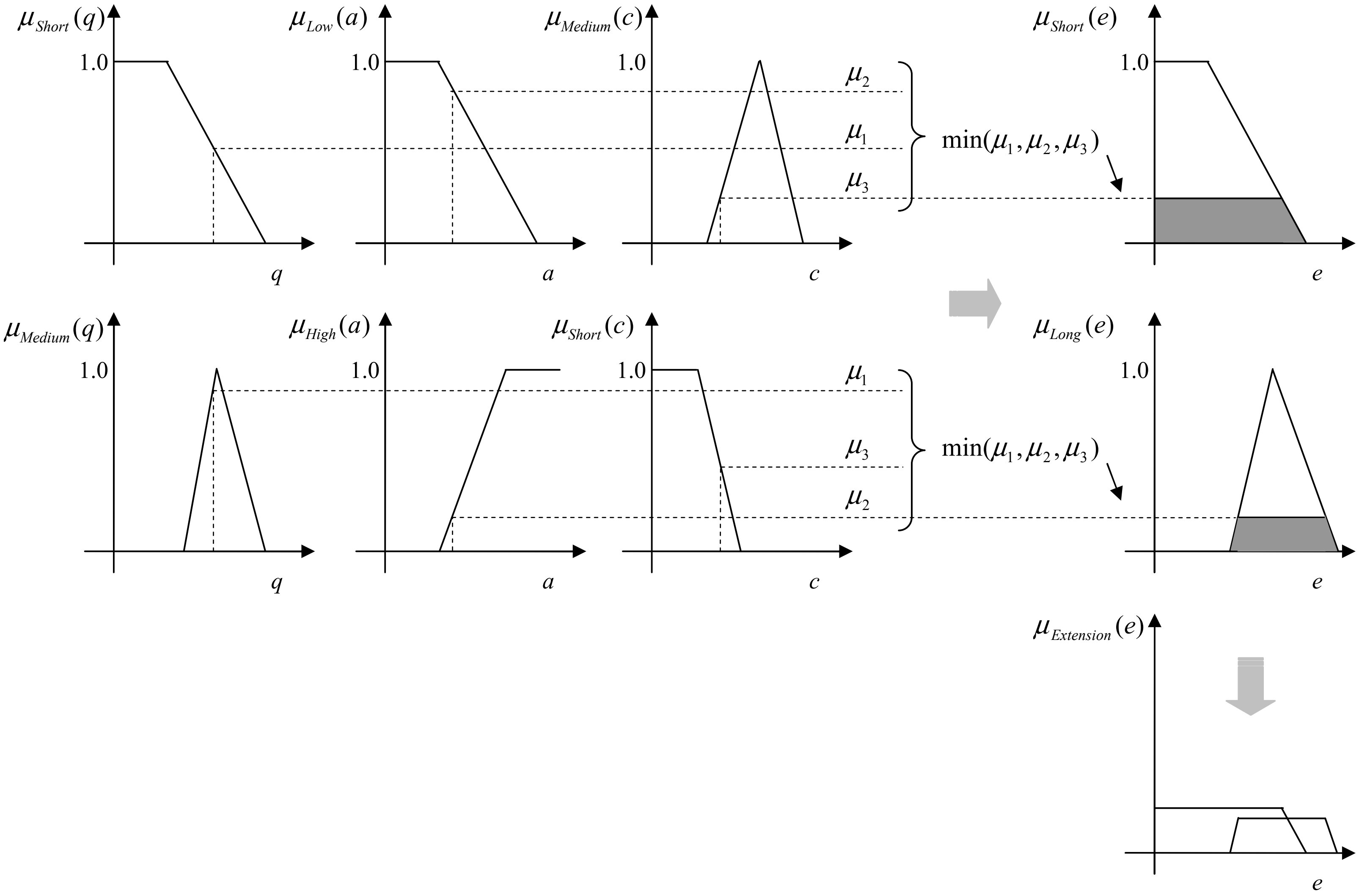


Figure 13 Example of fuzzy reasoning.

4.2.1.3 Fuzzy Logic Controller

A typical fuzzy logic controller has five major components, which are shown in Figure 14. The fuzzification process is to obtain the degrees of compatibility of each antecedent in fuzzy rules. The fuzzy inference component includes fuzzy rules and fuzzy reasoning, which are discussed in the previous section. As shown in Figure 13, the result from fuzzy inference is a combined induced consequent fuzzy set. To apply fuzzy logic controllers to practical control problems, a meaningful and crisp value usually needs to be obtained from the combined induced consequent fuzzy set.

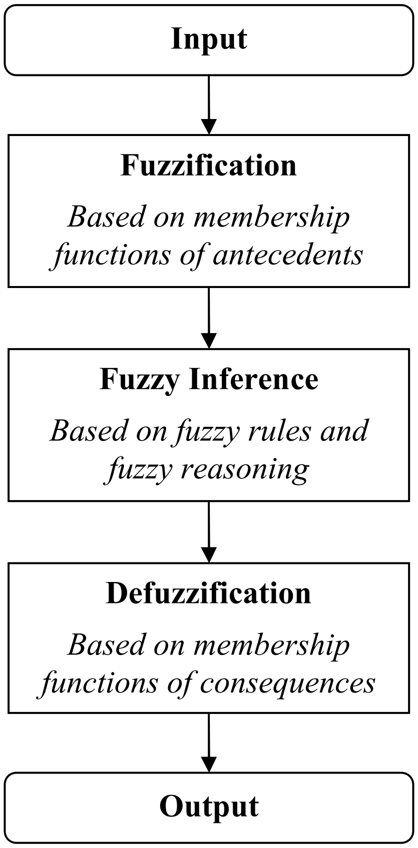

Input
Fuzzification
Based on membership functions of antecedents
Fuzzy Inference
Based on fuzzy rules and fuzzy reasoning
Defuzzification
Based on membership functions of consequences
Output

Figure 14 Structure of a typical fuzzy logic controller.

The process of obtaining a crisp value from the output of fuzzy inference, a combined induced consequent fuzzy set, is called defuzzification. A number of methods are available for this purpose, including Centroid of Area (COA), Bisector of Area (BOA), Mean of Max (MOM), Smallest of Max (SOM), and Largest of Max (LOM) (Jiang et al., 1997). After defuzzification, a crisp value can be obtained and used for practical applications. For the example in Figure 13, the output crisp value represents the amount of green time extension that should be given to the current green phase.

There are several different types of fuzzy logic controllers. The one mentioned above is called Mamdani fuzzy logic controller. Other well-known fuzzy logic controllers include Sugeno and Tsukamoto fuzzy logic controllers. Detailed information about them can be found in Jiang et al. (1997).

4.2.2 Neural Networks

In traditional reinforcement learning methods such as Q-Learning, a Q-Table is usually used for storing the learned control policy in the form of action values. As discussed in Chapter 3, this Q-Table method has certain limitations when the state or action space is large. In recent studies, neural networks are often used instead of the Q-Table in reinforcement learning for storing learned policies (Sutton and Barto, 1998; Berenji and Khedkar, 1992; Bingham, 2001), to improve generalization ability and better handle the curse of dimensionality problem. In the proposed NFACRL method, neural networks are combined with fuzzy logic control to approximate the best control policy. For better understanding of the proposed NFACRL method, a feed-forward back-propagation neural network is briefly described here.

Figure 15 shows the structure of a typical feed-forward back-propagation neural network. This network has three layers. The first layer is the input layer that takes inputs and sends them to the second layer. Each node in the first layer represents an input variable. The second layer is the hidden layer that consists of a number of hidden neurons, and each hidden neuron has a transfer function. The input to each transfer function is the summation of the weighted outputs from the first layer. The third layer is the output layer. In this example, it consists of only one neuron. In fact, there could be more than one neuron in the output layer depending on the problems to be solved. Similar to the hidden layer, the neuron in the output layer also has a transfer function. Its input is the summation of the weighted outputs from the hidden layer. The output of this transfer function is also the output of this neural network. In addition to these neurons, there are a number of weights and biases in the network. Before a neural network can be used to solve problems, these weights and biases have to be calibrated through a process called training. More details about neural networks can be found in (Jiang et al., 1997; Rumelhart et al., 1986).

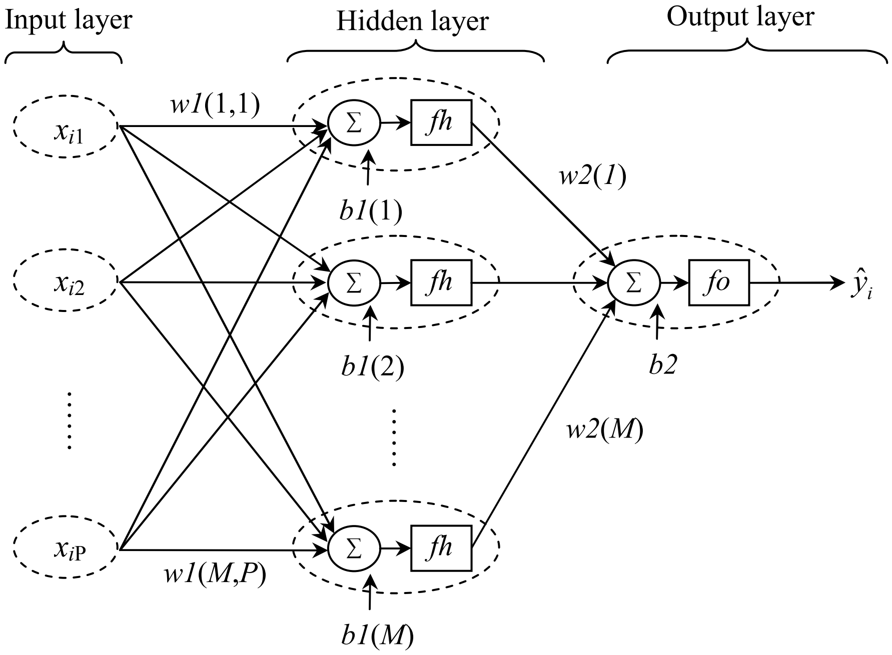


Figure 15 A typical feed-forward back-propagation neural network.

When applying neural networks to store the learned policy of a reinforcement learning traffic signal controller, the input to the network shown in Figure 15 could be the queue lengths and the output could be the amount of green time extension. Depending on the problems under study, neural networks can have multiple output units and each of them stands for a specific control action. The value of each output unit represents the preference that the corresponding action should be chosen.

4.3 NEURO-FUZZY ACTOR-CRITIC REINFORCEMENT LEARNING (NFACRL)

4.3.1 NFACRL Structure

The NFACRL method used in this research was developed by Jouffe (1996; 1998). NFACRL is also an actor-critic type of reinforcement learning, but it is different from the GARIC method used by Bingham (2001). The NFACRL method takes the form of neural networks and also incorporates fuzzy logic control into it. The structure of the NFACRL method is shown in Figure 16. The symbols used in Figure 16 are described below.

$S_i =$ the i^{th} input variable;

$K =$ the total number of input variables;

$NM_i =$ the number of fuzzy sets or membership functions for the i^{th} input variable;

$M_i^{a(i)} =$ the $a(i)^{th}$ fuzzy set or membership function for the i^{th} input variable;

$R_j =$ the j^{th} fuzzy rule;

$N =$ the total number of nodes in the third layer;

$\lambda^j =$ the weight connecting the j^{th} fuzzy rule and the critic output;

$w_q^j =$ the weight connecting the j^{th} fuzzy rule and the q^{th} action output;

$V =$ the critic output;

$A_q =$ the q^{th} action output;

$P =$ the total number of actions;

$a(i) \in \{1,...,NM_i\}$

$i = 1,...,K$;

$j = 1,...,N$; and

$q = 1,...,P$.

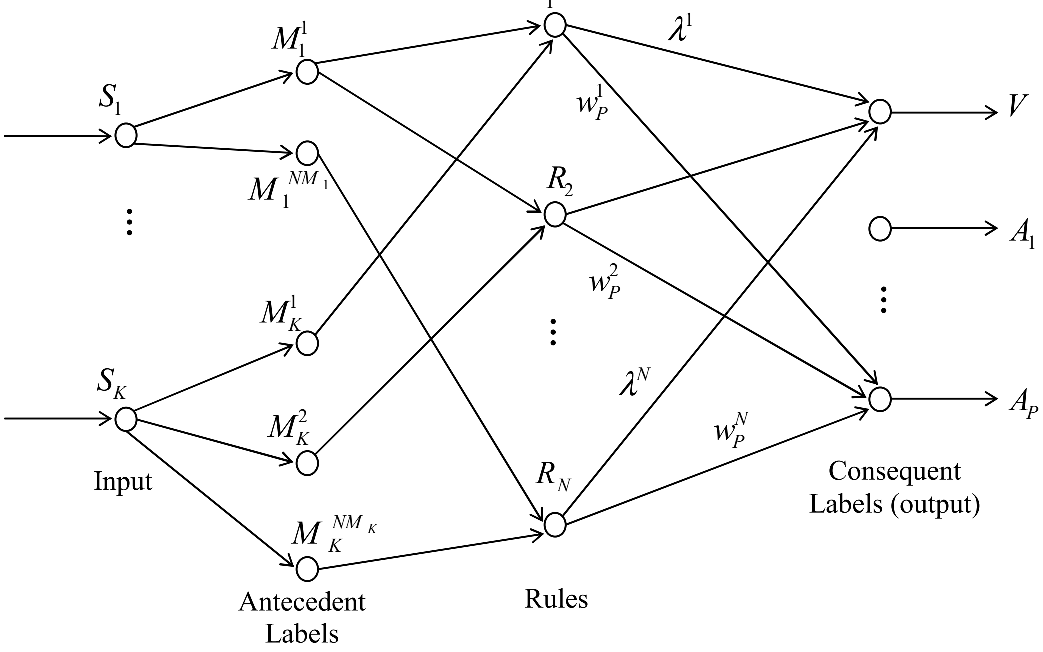


Figure 16 Example of the NFACRL (Jouffe, 1996).

Similar to neural networks, the NFACRL structure has four layers. The first layer is the input layer. It receives state variable values and sends them to different fuzzy membership functions in the second layer. Each node in the first layer represents an input (state) variable. Each node in the second layer is a fuzzy set with a fuzzy membership function associated with it. The inputs to the second layer are the state variable values, and the outputs of the second layer are fuzzy membership function values. The inputs and fuzzy sets of the second layer constitute many linguistic terms such as "Queue is {Short}" and "Queue is {Long}". Thus, the outputs of the second layer can also be considered as degrees of compatibility. The third layer corresponds to fuzzy rules in a fuzzy logic controller, and the outputs of the third layer can be considered as firing strengths. The fourth layer is a collection of nodes representing consequences. The first node stands for the *Critic* (see Figure 12), and its output value shows how good the current state is. The remaining nodes correspond to the available actions that can be taken, and their output values are the preferences to choose each action given the current state inputs.

There are three major differences between the architectures of the NFACRL method and the GARIC method used by Bingham (2001).

1. The NFACRL method has multiple outputs that are crisp values representing *Critic* and actions, while the GARIC method only has one continuous output. Multiple outputs can be more useful for modeling phase sequence optimization than the single continuous output. Since GARIC only has a continuous output, it can only be used to decide weather and how to extend the current green phase. While for the NFACRL method, the multiple action outputs can be used to decide which control phase should be chosen for the next step.
2. Bingham (2001) mainly used GARIC for fine tuning the parameters of fuzzy membership functions. The fuzzy rules in her study needed to be prespecified. If the control problem has many input variables, specifying the fuzzy rules could be cumbersome and prone to error. It will be shown later that the NFACRL dos not need to specify the fuzzy rules.
3. GARIC uses a neural network as the *Critic* and a fuzzy logic controller as the *Actor*, and *Critic* and *Actor* are relatively independent of each other. For the NFACRL,

Critic and *Actor* are closely related. Both of them use the same fuzzy membership function values as the inputs.

For the GARIC method, fuzzy rules need to be prespecified based on users' experience. If a control problem has many state variables, the fuzzy rules will become very complicated as the example shown below, and are difficult to specify even for very experienced experts.

- ♦ **IF** S_1 is {1} **AND** S_2 is {2} **AND** S_3 is {1} **AND** S_4 is {1} **AND** … **AND** S_K is {2}, **THEN** Action Output is { A_t }

Specifying fuzzy rules in the GARIC method is similar to determining how to connect the nodes in the second, third, and fourth layers in Figure 16. The maximum possible number of fuzzy rules is

$$Nmax = P\prod_{i=1}^{K} NM_i \tag{26}$$

For control problems with many state variables, there could be several hundreds of complicated fuzzy rules need to be specified manually if the GARIC method is used. In the NFACRL method, by introducing the weights between the third and fourth layers, one can simply use $Nmax$ fuzzy rules. Through fine tuning the weights between the third and fourth layers, the best fuzzy rules can be found automatically even though the number of fuzzy rules can still potentially be large.

4.3.2 Calculation Procedure of the NFACRL

For NFACRL control, the input and fuzzification parts are the same as the typical fuzzy logic control. Given the input and fuzzy membership functions, fuzzy membership function values are generated and fed into the third layer of NFACRL. The fuzzy inference method used in NFACRL is a little different from what is shown in Figure 13. Assuming the j^{th} fuzzy rule has the following *K* antecedents

$$S_1 \in M_1^{a(1)}, S_2 \in M_2^{a(2)}, \ldots, S_K \in M_K^{a(K)} \tag{27}$$

then the firing strength of the j^{th} fuzzy rule is calculated as

$$FS_{R_j} = \prod_{i=1}^{K} \mu_{a(i)}^{j}\left(S_i\right) \tag{28}$$

where

$a(i) \in \{1,\ldots,NM_i\} =$ one of the fuzzy sets for the i^{th} input variable; and

$\mu_{a(i)}^{j}\left(S_i\right) =$ the membership function value of the $a(i)^{th}$ fuzzy set for the i^{th} input variable, and this value is used in the j^{th} fuzzy rule.

Some of the firing strengths may be zeroes, which means that the corresponding fuzzy rules will not affect the final control output. After the firing strengths of each fuzzy rule are obtained, the next step is to calculate the preference of choosing each action using Equation (29).

$$\mathrm{Pref}(A_q) = \sum_{j=1}^{N} FS_{R_j} w_q^j \tag{29}$$

where

$\mathrm{Pref}(A_q) =$ preference of choosing the q^{th} action; and

$q = 1,\ldots,P.$

w_q^j in Equation (29) is also referred to as action weight. If the following two row vectors are used to represent firing strengths and action weights,

$$FS = \{FS_{R_1}, \ldots, FS_{R_N}\} \tag{30}$$

$$w_q = \{w_q^1, \ldots, w_q^N\} \tag{31}$$

then Equation (29) can be rewritten as

$$\text{Pref}(A_q) = FS(w_q)^T \tag{32}$$

In Equation (32), *T* means transpose. Similarly, the critic output of NFACRL is defined in Equation (33).

$$V = \sum_{j=1}^{N} FS_{R_j} \lambda^j = FS(\lambda)^T \tag{33}$$

where

$\lambda = \{\lambda^1, ..., \lambda^N\}$; and

$\lambda^j =$ the i^{th} critic weight connecting the i^{th} fuzzy rule and the critic output.

4.3.3 Learning Procedure of the NFACRL

The previous subsection describes how to calculate the outputs of NFACRL for given action and critic weights. In this subsection, the process of fine tuning the action and critic weights will be introduced (Jouffe, 1996; Jouffe, 1998).

Let $\lambda(t) = \{\lambda^1(t), \lambda^2(t), ..., \lambda^N(t)\}$ represent the critic weights at time step *t*, and $w_q(t) = \{w_q^1(t), w_q^2(t), ..., w_q^N(t)\}$ denote the action weights at time step *t* for the q^{th} action output. If the state variables at time step t are $S(t)=\{S_1(t), ..., S_K(t)\}$, then the critic and action outputs for state *S*(*t*) using weights at time step *t* are

$$V_t(S(t)) = FS(S(t))[\lambda(t)]^T \tag{34}$$

$$\text{Pref}_t(A_q, S(t)) = FS(S(t))[w_q(t)]^T \tag{35}$$

where

$FS\ (S(t)) =$ firing strengths calculated based on state variables at time step *t*;

$V_t(S(t)) =$ critic output calculated based on state variables at time step *t* and weights at time step *t*; and

$\mathrm{Pref}_t(A_q, S(t)) =$ preference for the q^{th} action calculated based on state variables at time step *t* and weights at time step *t*.

After all the action outputs have been calculated, an ε-greedy algorithm is used to choose an action based on the calculated preferences of each action. If action *j* is selected and executed, and the resulted new state at time step *t*+1 is $S(t+1)$, then the critic output for state $S(t+1)$ using weights at time step *t* is

$$V_t(S(t+1)) = FS(S(t+1))[\lambda(t)]^T \tag{36}$$

The transition from states *S*(*t*) to *S*(*t*+1) also results in a reward r_{t+1} at time step *t*+1. Based on $V_t(S(t))$, $V_t(S(t+1))$, and r_{t+1}, A TD error is calculated using Equation (37).

$$\delta_t = r_{t+1} + \gamma V_t(S(t+1)) - V_t(S(t)) \tag{37}$$

This TD error is used to update both the critic and action weights using Equations (38) and (39), respectively.

$$\lambda(t+1) = \lambda(t) + \beta\delta_t FS(S(t)) \tag{38}$$

$$w_j(t+1) = w_j(t) + \beta\delta_t FS(S(t)) \tag{39}$$

where β is a learning rate to be specified. Note that at each step only the action weights connecting to the chosen (j^{th}) action are updated.

If the changes of critic and action weights are less than a prespecified small value after certain updating steps, or the control performance tends to be stable, the learning process is terminated and the trained NFACRL is then used for real world control applications. After the learning process is terminated, a greedy action selection strategy should be used in lieu of the

ε-greedy action selection method, such that the NFACRL method will not give irrational instructions during implementation. The entire learning process of the NFACRL method is summarized in Figure 17.

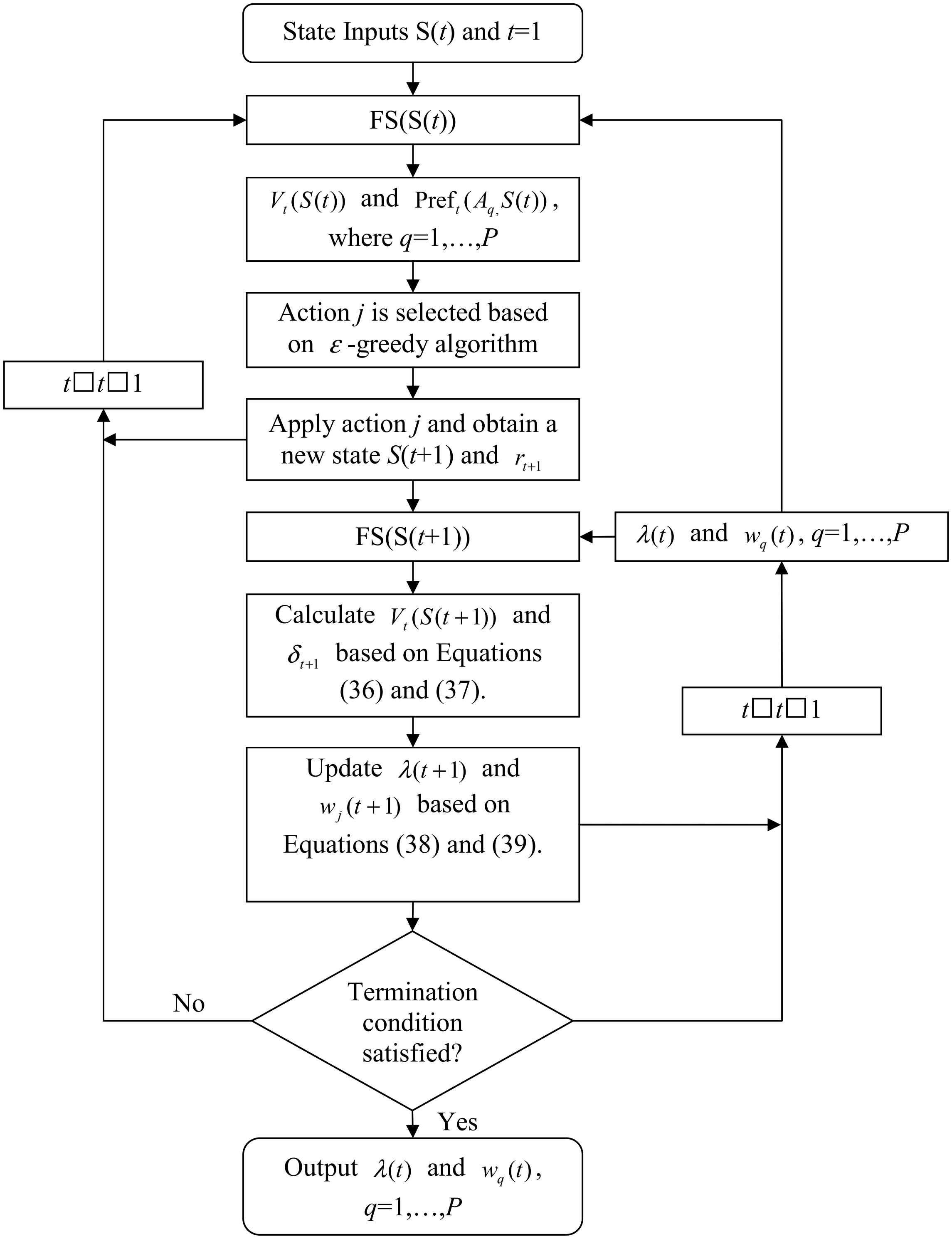

State Inputs S(t) and t=1
FS(S(t))
$V_t(S(t))$ and $\text{Pref}_t(A_{q,}S(t))$, where q=1,…,P
Action j is selected based on ε-greedy algorithm
t□t□1
Apply action j and obtain a new state S(t+1) and r_{t+1}
FS(S(t+1))
$\lambda(t)$ and $w_q(t)$, q=1,…,P
Calculate $V_t(S(t+1))$ and δ_{t+1} based on Equations (36) and (37).
t□t□1
Update $\lambda(t+1)$ and $w_j(t+1)$ based on Equations (38) and (39).
No
Termination condition satisfied?
Yes
Output $\lambda(t)$ and $w_q(t)$, q=1,…,P

Figure 17 Training process of the NFACRL method.

4.3.4 Summary of NFACRL

The NFACRL method is a combination of neural networks, fuzzy logic control, and actor-critic reinforcement learning and is different from the GARIC method used by Bingham (2001). It has the abilities to handle phase sequence optimization of traffic signal control, large state space, generalization ability, and complicated fuzzy rules.

The following three problems can have significant effects on the performance of NFACRL. Before applying the NFACRL method to traffic signal control, the following three problems need to be properly addressed.

1. Choices of state variables and actions;
2. Definition of reward; and
3. Coordination of control agents.

In the following two sections, these three problems are addressed and two intersection and arterial control methods based on NFACRL are proposed.

4.4 ISOLATED INTERSECTION ADAPTIVE TRAFFIC CONTROL BASED ON NFACRL

4.4.1 Fixed Phase Sequence Control Based on NFACRL

There could be many different ways of applying the NFACRL method to intersection traffic control. One option is to consider a fixed phase sequence. In this case, the action space is to either extend the current green phase or terminate it, which is similar to what has been used in previous studies (Thorpe, 1997; Abdulhai et al., 2003).

In this research, only three- and four-approach intersections are considered, as they are the most common types of intersections in the real world. For a typical four-approach intersection in Figure 2, the following phase sequence shown in Figure 18 may be used. The control logic starts with the first phase (ϕ_1) and then visits the remaining five phases one by one in order. After the last phase (ϕ_6) in the sequence has been visited, the control logic goes back to the first phase and repeats the entire process. Similar phase sequences are also used in the pre-timed and actuated control strategies that are to be compared with the NFACRL control. These pre-timed and actuated control strategies are optimized by Synchro (Husch and Albeck, 2001).

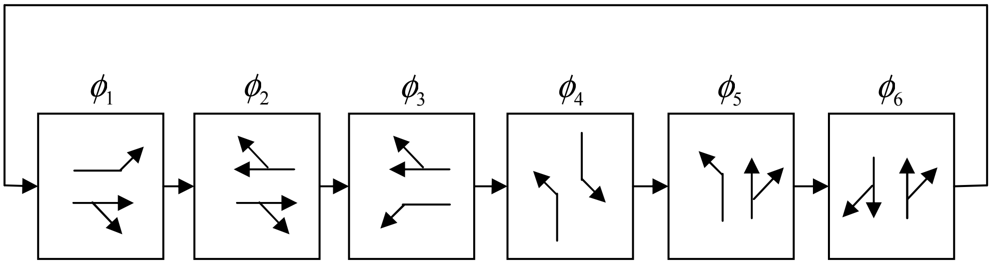


Figure 18 Phase plan for a four-approach isolated intersection.

For an isolated three-approach intersection with five movements as shown in Figure 19, the fixed phase sequence shown in Figure 20 may be used.

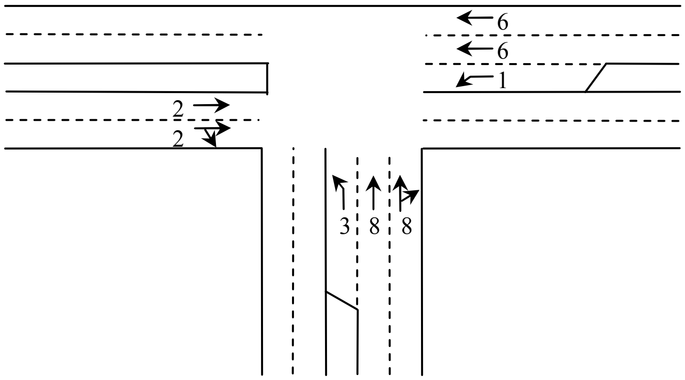


Figure 19 Layout of a typical three-approach intersection.

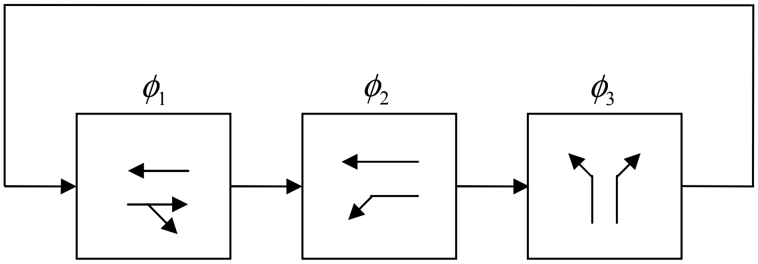


Figure 20 Phase plan for a three-approach isolated intersection.

4.4.1.1 Choices of State Variables

In most previous reinforcement learning traffic control studies, queue lengths were used as state variables (Bingham, 2001; Thorpe, 1997; Abdulhai et al., 2003). For the fixed phase sequence control based on NFACRL, queue lengths are also used as state variables. In addition to queue lengths, another state variable representing the current signal status is included.

For isolated four-approach intersections with eight movements (each through movement and its associated right-turn movement are combined as one movement) as in Figure 2, totally nine state variables are used, which means *K* in Figure 16 is equal to nine. The first eight state variables are used to represent the queue lengths and the last state variable is used to indicate the current signal state. More specifically, the first eight state variables are defined in Equation (40).

$$S_i = Q_i \tag{40}$$

where

$S_i =$ the i^{th} state variables;

$Q_i =$ queue length of the i^{th} movement (see Figure 2); and

$i = 1,...,8$.

The last state variable is defined as

$$S_9 = \begin{cases} 1 & \phi_1 = \text{Green, and } \phi_{i\neq 1} = \text{Red} \\ 2 & \phi_2 = \text{Green, and } \phi_{i\neq 2} = \text{Red} \\ 3 & \phi_3 = \text{Green, and } \phi_{i\neq 3} = \text{Red} \\ 4 & \phi_4 = \text{Green, and } \phi_{i\neq 4} = \text{Red} \\ 5 & \phi_5 = \text{Green, and } \phi_{i\neq 5} = \text{Red} \\ 6 & \phi_6 = \text{Green, and } \phi_{i\neq 6} = \text{Red} \end{cases}, \ i=1,\ldots,6 \tag{41}$$

For isolated three-approach intersections as the one shown in Figure 19, six state variables are used. Consequently, the parameter *K* in Figure 16 is equal to six. Among the six state variables, the first five ones are used to represent the queue lengths and are defined as

$$S_1 = Q_1 \tag{42}$$

$$S_2 = Q_2 \tag{43}$$

$$S_3 = Q_3 \tag{44}$$

$$S_4 = Q_6 \tag{45}$$

$$S_5 = Q_8 \tag{46}$$

The last state variable is used to indicate the current signal state and is defined as

$$S_6 = \begin{cases} 1 & \phi_1 = \text{Green, and } \phi_2, \phi_3 = \text{Red} \\ 2 & \phi_2 = \text{Green, and } \phi_1, \phi_3 = \text{Red} \\ 3 & \phi_3 = \text{Green, and } \phi_1, \phi_2 = \text{Red} \end{cases} \tag{47}$$

4.4.1.2 Fuzzy Membership Functions

To apply the NFACRL method, a set of fuzzy membership functions needs to be defined for the state variables. For each queue length state variable, two fuzzy sets are defined, which are {Short, Long}. The membership function for fuzzy set {Short} is defined in Equation (48).

$$\mu_{Short}(x) = \begin{cases} 1 & x \leq 0 \\ (10-x)/10 & x \in (0,10) \\ 0 & x \geq 10 \end{cases} \tag{48}$$

The membership function for fuzzy set {Long} is defined in Equation (49).

$$\mu_{Long}(x) = \begin{cases} 0 & x \leq 0 \\ x/10 & x \in (0,10) \\ 1 & x \geq 10 \end{cases} \tag{49}$$

The value 10 in both Equations (48) and (49) is a subjective number selected for this study. These fuzzy membership functions are illustrated in Figure 21.

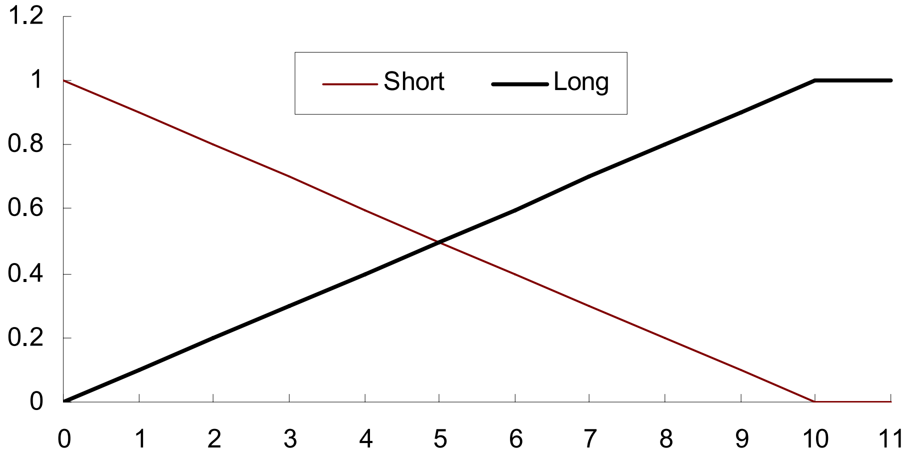


Figure 21 Fuzzy membership functions for queue length state variables.

For the state variable representing signal status, the definition of its fuzzy membership function is a little different. Using the three-approach intersection shown in Figure 19 as an example, the state variable S_6 has three fuzzy sets, which are $\{\phi_1, \phi_2, \phi_3\}$. The corresponding fuzzy membership functions are defined in Equations (50) through (52).

$$\mu_{\phi_1}(S_6) = \begin{cases} 1 & S_6 = 1 \\ 0 & \text{else} \end{cases} \tag{50}$$

$$\mu_{\phi_2}(S_6) = \begin{cases} 1 & S_6 = 2 \\ 0 & \text{else} \end{cases} \tag{51}$$

$$\mu_{\phi_3}(S_6) = \begin{cases} 1 & S_6 = 3 \\ 0 & \text{else} \end{cases} \tag{52}$$

Using the same principle, a set of fuzzy membership functions is defined for state variable S_9 for four-approach intersections.

4.4.1.3 Fuzzy Rules

For this fixed phase sequence control scheme, the third and fourth layers of the NFACRL (Figure 16) are assumed to be fully connected. Each node in the third layer has *K* connections with the second layer, and one for each input state variable. Taking the three-approach intersection control as an example, a sample fuzzy rule is presented below

♦ **IF** S_1 is {Long} **AND** S_2 is {Short} **AND** S_3 is {Long} **AND** S_4 is {Short} **AND** S_5 is {Short} **AND** S_6 is { ϕ_1 }, **THEN** Next Action is {Extension}

Since each of the five queue length state variables has two categories and the signal state variable has three values, totally there are 96 nodes in the third layer of the NFACRL (see Figure 16).

Similarly, for the four-approach intersection control, each of the eight queue length state variables has two fuzzy sets associated with it, and the signal state variable has six possible states. Therefore, for the four-approach intersection control there are a total of 1536 nodes in the third layer of the NFACRL (see Figure 16).

4.4.1.4 Definition of Reward

As shown in Equation (6), the objective of reinforcement learning is to find an optimal policy π^* (a mapping from states to actions) to maximize the reward of each state, and it is equivalent to maximize the summation of discounted rewards shown in Equation (53) (Sutton and Barto, 1998).

$$\begin{aligned} V^*(s) &= \max_{a\in A(s)} E\left\{ \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \mid s_t = s \right\} \\ &= \max_{a\in A(s)} E\{r_{t+1} + \gamma V^*(s') \mid s_t = s\} \end{aligned} \tag{53}$$

This is similar to the DYPIC method based on dynamic programming, whose optimization goal is in Equation (54).

$$f_i(j) = \min_{a_i}\{C_{jk} + f_{i+1}(k)\}, \quad i = 1,\dots,N, \quad j \in S_i, \quad k \in S_{i+1} \tag{54}$$

where C_{jk} is the total delay associated with transition from state *j* at stage *i* to state *k* at stage *i*+1. Comparing Equations (53) and (54) suggests that the minus delay between two decision intervals can be used as the reward.

Thorpe (1997) used a linear combination of discharged and stopped vehicles as the reward. Bingham (2001) used minus delay as the reward. Abdulhai et al. (2003) also used minus total delay between two decision points as the reward, and the total delay was calculated by counting queue lengths every 1 second. It makes sense to use minus total delay as the reward, as minimizing delay is often used as the objective of traffic signal control. However, simply using queue length to represent delay in the reward function may not be enough, as queue length can not accurately reflect the delay caused by acceleration and deceleration maneuvers. Also, sometimes it is desirable to consider minimizing the number of stops. Thus, in this research, the following reward definition is used.

$$r = \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 + \beta_4 x_4 - \beta_5 x_5 \tag{55}$$

where

x_1 = number of vehicles that have passed the intersection from approaches being given green signal;

x_2 = number of vehicles in queues;

x_3 = number of vehicles newly added to queues;

x_4 = number of vehicles in approaches being given green signal;

x_5 = number of vehicles being stopped when signal is switched from green to red; and

β_i = nonnegative coefficients for each variable. $i = 1,\dots,5$

x_1 encourages moving more vehicles through the intersection during two decision points; x_2 represents stopped delay; x_3 is used to account for deceleration delay; x_4 is to have more vehicles in the current green phase; and x_5 is used to penalize switching green signal to red while there are many vehicles being served by this green signal.

4.4.2 Variable Phase Sequence Control Based on NFACRL

Fixed phase sequence control based on NFACRL can significantly reduce the dimension of state and action spaces, consequently reducing the number of action and critic weights. However, the fixed sequence NFACRL control may lack the flexibility to fully adapt to traffic flow fluctuations due to the fixed phase sequence constraint. In this section, a variable phase sequence control method based on NFACRL is proposed. The variable phase sequence NFACRL control also use queue lengths and signal states as inputs. But the decision output is not extension or termination. Instead, the decision output is any of the available control actions. For the three-approach intersection in Figure 19, the decision output could be ϕ_1, ϕ_2, or ϕ_3 shown in Figure 20. In this case, a sample fuzzy rule is

- ♦ **IF** S_1 is {Long} **AND** S_2 is {Short} **AND** S_3 is {Long} **AND** S_4 is {Short} **AND** S_5 is {Short} **AND** S_6 is {ϕ_1}, **THEN** Next Action is {ϕ_3}

For the four-approach intersection in Figure 2, the decision output could be any of the eight phases in Figure 22.

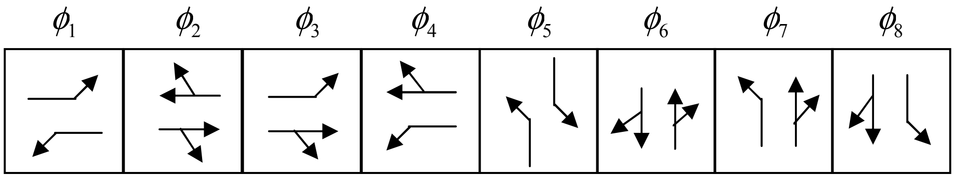


Figure 22 Possible control actions for a four-approach intersection.

For the three-approach intersection, five queue length state variable and one signal state variable are used in the variable phase sequence NFACRL control. These variables are defined exactly the same as in the fixed phase sequence NFACRL control. Namely, each queue length state variable has two fuzzy sets and each signal state variable has three fuzzy sets. Therefore, the variable phase sequence NFACRL has 96 nodes in the third layer (see Figure 16).

For the four-approach intersection in Figure 2, eight queue length state variable and one signal state variable are used in the variable phase sequence NFACRL control. The eight queue length state variables are defined the same as in the fixed phase sequence NFACRL control. The signal state variable is defined a little differently, which is shown in Equation (56).

$$S_9 = j, \text{ if } \phi_j = \text{Green and } \phi_{i \neq j} = \text{Red}, \ (i,j = 1,...,8) \tag{56}$$

Therefore, for four-approach intersection control with variable phase sequence, the NFACRL has 2048 nodes in the third layer (see Figure 16).

For the variable phase sequence NFACRL control, the same fuzzy membership function in Figure 21 is used for all queue length state variables. The fuzzy membership functions for the signal state variables are defined in the same way as in the fixed phase sequence NFACRL control. The reward definition used in the fixed phase sequence NFACRL control is also used in the variable phase sequence NFACRL control, but different coefficients are chosen.

4.5 MULTI-AGENT ARTERIAL ADAPTIVE TRAFFIC CONTROL BASED ON NFACRL

4.5.1 Multi-Agent Reinforcement Learning

Isolated intersection control is a single agent decision problem. For a system that has more than one intersection, multiple control agents should be used. A system consists of several agents is usually referred to as multi-agent system (MAS). As many practical control problems, such as arterial traffic control, can be modeled as MASs, multi-agent reinforcement learning (MARL) has attracted considerable attention over the past two decades (Hu and Wellman, 1998; Panait and Luke, 2005; Chalkiadakis and Boutilier, 2003; Tan, 1993). In the following subsections, three major MARL methods are briefly reviewed.

4.5.1.1 MARL Based on Independent-Agent

Independent-agent is the simplest MARL method. It directly applies single-agent reinforcement learning to MAS. Each agent treats all other agents as part of the environment (Hu and Wellman, 1998). One potential problem of this method is that the existence of other agents may affect the environment and invalidate the Markov property assumption (Hu and Wellman, 2003).

4.5.1.2 MARL Based on SG

Many MARL studies have focused on using stochastic game (SG) or Markov game (MG). SG is a natural extension of MDP to handle problems with multiple agents. Recall that in Chapter 3 a MDP is defined by a tuple (S, A, r, p). Similarly, a SG is defined by a more complicated tuple as $(n, S, A_1, ..., A_n, r_1, ..., r_n, p)$ (Hu and Wellman, 1998; Bowling and Veloso, 2002; Littman, 1994), where

1. n is the total number of agents in the MAS;
2. S is a set of discrete states;
3. $A_i \;\; (i = 1, ..., n)$ is the action space for the i^{th} agent;
4. $r_i \;\; (i = 1, ..., n)$ is the reward function for the i^{th} agent, which is affected by the current system state and all actions that will be taken; and
5. p is a transition function, which gives the probability that the system will be in each state provided with the current system state and actions to be taken.

Under the framework of SG, the state transition is still assumed to satisfy the Markov Property.

Littman (1994) appears to be the first researcher to use SG as the framework for solving MARL problems. He studied a two-agent zero-sum SG problem, and proposed a minimax-Q algorithm similar to Q-Learning for solving this problem. For the two-agent zero-sum SG problem, there are two competing agents. The gain of one agent always leads to the loss of another, and the summation of gains from both agents is equal to zero. For arterial traffic signal control, the gain of one control agent does not necessarily mean the loss of other agents. Therefore, the zero-sum SG framework is not suitable for modeling arterial traffic control problems.

Hu and Wellman (1998) further researched the MARL problem under the framework of general-sum SG, in which different agents can increase their gains simultaneously. They developed a multi-agent Q-Learning algorithm to solve n-agent general-sum SG problems. For ease of description, the following discussions only consider a two-agent general-sum SG problem. Different from the Q-Learning for MDP, the multi-agent Q-Learning proposed by Hu and Wellman (1998) requires each agent to keep two Q-Tables, one for itself and one for the other agent in the system. Using agent one as an example, during the learning process, it updates it own Q-Table using Equation (57).

$$V_{t+1}^{1}(s_t,a_t^1,a_t^2)=V_t^1(s_t,a_t^1,a_t^2)+\phi\left[r_{t+1}^{1}+\gamma\pi^1(s_{t+1})V_t^1(s_t,a_t^1,a_t^2)\pi^2(s_{t+1})-V_t^1(s_t,a_t^1,a_t^2)\right] \qquad \textbf{(57)}$$

where

$V_{t+1}^{1}(s_t,a_t^1,a_t^2)=$ action function value for agent 1 at time step *t*+1;

$a_t^1=$ action taken by agent 1 at time step *t*;

$a_t^2=$ action taken by agent 2 at time step *t*;

$\pi^1(\cdot)=$ policy function of agent 1 at time step *t*;

$\pi^2(\cdot)=$ policy function of agent 2 at time step *t*;

$r_{t+1}^{1}=$ reward for agent 1 at time step *t*+1; and

$\pi^1(s_{t+1})V_t^1(s_t,a_t^1,a_t^2)\pi^2(s_{t+1})=$ is the expected reward of agent 1 under the mixed strategy Nash Equilibrium (Hu and Wellman, 1998);

Note that updating agent 1's state action function (Q-Table) needs the policy function information of agent 2. This can be done by keeping track of agent 2's Q-Table using the following Equation (58). Detailed updating procedure can be found in (Hu and Wellman, 1998).

$$V_{t+1}^{2}(s_t,a_t^1,a_t^2)=V_t^2(s_t,a_t^1,a_t^2)+\phi\left[r_{t+1}^{2}+\gamma\pi^1(s_{t+1})V_t^2(s_t,a_t^1,a_t^2)\pi^2(s_{t+1})-V_t^2(s_t,a_t^1,a_t^2)\right] \qquad \textbf{(58)}$$

There are major difficulties in applying this multi-agent Q-Learning method to arterial traffic control. Mostly, with multiple intersections, the number of state variables will become very large and make the learning process extremely slow. Based on previous discussions on a four-approach intersection, there could be 9 state variables. If an arterial has four such intersections, then the total number of state variables is 36. Assuming each state variable has 2 categories, the total number of possible states is $2^{36} \approx 6.9 \times 10^{10}$. The huge number of possible states will not only considerably slow down the reinforcement learning process, but also give rise to the generalization problem.

4.5.1.3 MARL Based on Cooperative-Agent

Tan (1993) conducted a study to compare the performance of independent-agent and cooperative-agent in a MAS. For cooperative-agent method, agents share information with each other. Tan (1993) experimented with the following three cooperation strategies:

1. The first strategy shared real-time state information among all agents. Although testing results showed that sometimes cooperative-agent method using this strategy could moderately outperform the independent-agent method, this strategy significantly increased the state space of each agent in the system and might not be suitable for arterial traffic signal control.
2. The second strategy shared experiences among all agents. These experiences were different from the instant information shared in the first strategy. They were past state, action, and reward information experienced by each agent. Tan reported that the second strategy improved the learning speed. However, it produced approximately the same performance as the independent-agent method did.
3. The third strategy was similar to the first one. But the author applied it to a new problem, in which two agents were designed to accomplish a common task. In addition to the large state space problem with the first strategy, the third strategy required a lot of communications between the two agents.

4.5.2 Multi-agent Arterial Adaptive Traffic Control Using NFACRL

The review in the previous section shows that there are basically three MARL methods:

1. MARL based on independent-agent;
2. MARL under the framework of SG; and
3. MARL based on cooperative-agent that share experiences or information.

Due to the large state and action spaces problem, the second method under the framework of SG is ruled out for arterial traffic control in this research. In fact, this method so far is mainly used in theoretical studies. The cooperative method may also not be a good idea as each intersection is controlled by an agent in this research. Since different intersections may have different geometric settings, their environments are most likely different. Under this circumstance, sharing experience among different agents may not be useful. In addition, the previous study by Tan (1993) showed that sharing experience among agents only expedited the learning process and did not appear to improve the learning results.

For the independent-agent MARL method, agents are expected to learn how to coordinate implicitly. Although the first method is simple, it can be very useful in practice. Compared to the other two more complicated MARL methods, it has the following nice properties:

1. No communication devices need to be installed between adjacent intersections.
2. Simplicity sometimes means robustness. In this case, the malfunction of other controllers will not directly affect the function of the current controller.

With all the above considerations, the independent-agent method is chosen to coordinate different control agents in this research.

4.6 SUMMARY

In this chapter, a neuro-fuzzy actor-critic reinforcement learning (NFACRL) method is introduced for adaptive traffic signal control. NFACRL uses a neuro-fuzzy network to store the actor and critic values of each state, such that the curse of dimensionality and generalization problems can be properly handled. It also has the ability to model discrete action outputs and can be used to optimize phase sequence of traffic signal control.

After the NFACRL method is introduced, two implementation schemes are proposed to apply the NFACRL method to isolated intersection traffic control. The first scheme considers a fixed phase sequence and the second one does not. For both implementation schemes, the implementation details such as the choice of state and action variables, fuzzy membership functions, fuzzy rules, and reward functions are discussed in details.

The two NFACRL control methods are further extended for the traffic control of an arterial consisting of several intersections. Each intersection is controlled by an agent and the arterial traffic signal control is modeled as a multi-agent system. Various methods to coordinate different agents in this multi-agent system are reviewed. Based on the review, a simple but robust independent-agent method is adopted for arterial adaptive traffic signal control.

CHAPTER 5. EVALUATION OF THE NFACRL TRAFFIC CONTROL BASED ON MICROSCOPIC SIMULATION

5.1 INTRODUCTION

This chapter discusses in details the evaluation of the NFACRL traffic control using VISSIM microscopic traffic simulation. The evaluations are carried out at both isolated intersection and arterial levels based on simulation network created from real world data. The fixed and variable NFACRL control schemes for isolated intersection traffic control are first evaluated. Both NFACRL control schemes are then extended to arterial traffic control by using an independent-agent coordination method. For the isolated intersection evaluation, the two NFACRL control schemes are compared with optimized pre-timed and actuated controls. For the arterial evaluation, the two NFACRL control schemes are compared with optimized coordinated pre-timed and coordinated actuated control.

The rest of this chapter is organized as the following. Firstly, data used for setting up the simulation traffic network are described. Secondly, the VISSIM microscopic traffic simulation program used in this research is discussed. Details about how to code the simulation traffic network and various control algorithms are also presented. Thirdly, test design is described. Following that are the testing results at both intersection and arterial levels. The last section summarizes this chapter.

5.2 DATA DESCRIPTION

Data from a real world arterial network in College Station, Texas are used. The chosen arterial is a segment of FM 2818 (Harvey Mitchell Parkway), shown in Figure 23, which include three four-approach intersections and one three-approach intersection. The traffic volume data for each intersection in Figure 23 are listed in Tables 3 through 5. The morning peak period traffic data were collected on October 7, 2004 from 7:00 A.M. to 8:00 A.M.; the noon peak period traffic data were collected on October 12, 2004 from 11:45 A.M. to 12:45 P.M.; and the afternoon peak period traffic data were also collected on October 12, 2004 but from 4:45 P.M. to 5:45 P.M.

Table 3 Traffic volume data during morning peak hour

Intersection	Time	Southbound				Westbound				Northbound				Eastbound			
		R	T	L	All	R	T	L	All	R	T	L	All	R	T	L	All
Longmire & FM 2818	7:15:00 AM	4	4	6	14	2	204	29	235	56	10	61	127	24	150	4	178
	7:30:00 AM	4	5	6	15	7	277	33	317	63	19	81	163	55	237	8	300
	7:45:00 AM	11	7	4	22	7	162	29	198	43	12	27	82	56	184	2	242
	8:00:00 AM	4	7	3	14	4	73	25	102	39	9	29	77	32	128	2	162
Southwood & FM 2818	7:15:00 AM	16	5	13	34	2	205	1	208	9	8	26	43	14	151	27	192
	7:30:00 AM	28	10	7	45	3	315	0	318	13	32	53	98	16	231	60	307
	7:45:00 AM	30	9	25	64	3	327	3	333	15	24	51	90	11	235	38	284
	8:00:00 AM	13	6	15	34	3	104	6	113	4	15	17	36	16	163	24	203
Rio Grande & FM 2818	7:15:00 AM	-	-	-	-	-	217	8	225	41	-	54	95	8	148	-	156
	7:30:00 AM	-	-	-	-	-	353	12	365	86	-	113	199	16	179	-	195
	7:45:00 AM	-	-	-	-	-	404	14	418	112	-	109	221	24	217	-	241
	8:00:00 AM	-	-	-	-	-	191	10	201	50	-	52	102	18	174	-	192
Welsh & FM 2818	7:15:00 AM	16	26	21	63	22	145	12	179	39	53	61	153	6	93	15	114
	7:30:00 AM	17	39	45	101	61	206	16	283	26	101	82	209	1	102	30	133
	7:45:00 AM	30	47	58	135	74	208	10	292	10	99	96	205	8	124	32	164
	8:00:00 AM	29	50	49	128	23	142	24	189	14	78	49	141	8	99	11	118

Note: L – Left-Turn Movement; T – Through Movement; R – Right-Turn Movement

Table 4 Traffic volume data during noon peak hour

Intersection	Time	Southbound				Westbound				Northbound				Eastbound			
		R	T	L	All	R	T	L	All	R	T	L	All	R	T	L	All
Longmire & FM 2818	12:00:00 PM	10	11	12	33	11	118	64	193	82	9	42	133	38	106	6	150
	12:15:00 PM	3	13	11	27	5	131	69	205	88	19	58	165	34	106	8	148
	12:30:00 PM	3	12	8	23	6	92	49	147	74	6	61	141	46	114	8	168
	12:45:00 PM	8	18	6	32	7	80	41	128	57	20	39	116	48	149	8	205
Southwood & FM 2818	12:00:00 PM	12	10	13	35	4	144	6	154	10	8	17	35	32	147	18	197
	12:15:00 PM	29	17	8	54	7	168	3	178	8	13	24	45	24	138	24	186
	12:30:00 PM	15	10	13	38	6	143	6	155	7	12	17	36	17	133	15	165
	12:45:00 PM	18	8	5	31	6	121	5	132	11	11	20	42	24	187	26	237
Rio Grande & FM 2818	12:00:00 PM	-	-	-	-	-	161	19	180	29	-	21	50	18	165	-	183
	12:15:00 PM	-	-	-	-	-	184	22	206	23	-	20	43	17	170	-	187
	12:30:00 PM	-	-	-	-	-	168	24	192	21	-	14	35	7	148	-	155
	12:45:00 PM	-	-	-	-	-	145	14	159	31	-	11	42	14	202	-	216
Welsh & FM 2818	12:00:00 PM	5	17	23	45	10	115	39	164	21	12	21	54	15	146	1	162
	12:15:00 PM	4	14	16	34	8	150	25	183	22	25	28	75	23	118	3	144
	12:30:00 PM	1	20	20	41	27	108	26	161	21	25	29	75	14	105	4	123
	12:45:00 PM	3	26	20	49	13	104	23	140	46	20	29	95	26	128	8	162

Note: L – Left-Turn Movement; T – Through Movement; R – Right-Turn Movement

Table 5 Traffic volume data during afternoon peak hour

Intersection	Time	Southbound				Westbound				Northbound				Eastbound			
		R	T	L	All	R	T	L	All	R	T	L	All	R	T	L	All
Longmire & FM 2818	5:00:00 PM	9	10	12	31	12	197	47	256	64	22	59	145	71	185	12	268
	5:15:00 PM	10	28	11	49	16	204	49	269	75	21	68	164	81	170	9	260
	5:30:00 PM	10	16	11	37	13	161	61	235	62	12	56	130	68	184	8	260
	5:45:00 PM	4	14	12	30	19	196	55	270	73	17	43	133	72	200	11	283
Southwood & FM 2818	5:00:00 PM	23	13	17	53	10	249	6	265	19	17	27	63	20	219	34	273
	5:15:00 PM	49	23	20	92	7	231	15	253	14	13	22	49	27	248	17	292
	5:30:00 PM	39	19	20	78	13	237	10	260	10	12	13	35	30	248	34	312
	5:45:00 PM	31	17	12	60	8	193	7	208	7	12	12	31	17	288	32	337
Rio Grande & FM 2818	5:00:00 PM	-	-	-	-	-	261	48	309	40	-	18	58	19	229	-	248
	5:15:00 PM	-	-	-	-	-	231	54	285	34	-	25	59	41	256	-	297
	5:30:00 PM	-	-	-	-	-	241	56	297	34	-	32	66	40	287	-	327
	5:45:00 PM	-	-	-	-	-	205	49	254	48	-	15	63	42	284	-	326
Welsh & FM 2818	5:00:00 PM	6	48	37	91	29	170	51	250	51	51	27	129	50	143	13	206
	5:15:00 PM	13	53	42	108	40	163	52	255	53	50	30	133	61	197	12	270
	5:30:00 PM	16	83	27	126	22	144	49	215	55	46	29	130	54	190	20	264
	5:45:00 PM	14	76	41	131	26	105	59	190	43	54	12	109	40	147	14	201

Note: L – Left-Turn Movement; T – Through Movement; R – Right-Turn Movement

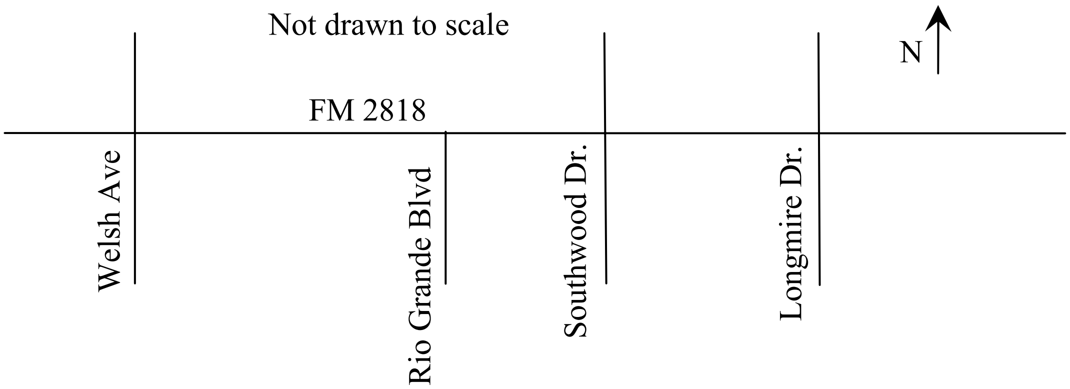


Figure 23 Diagram of testing arterial network.

5.3 MICROSCOPIC TRAFFIC SIMULATION

Microscopic traffic simulation has been used as a standard method for testing and comparing different traffic control strategies. Compared to evaluating traffic control strategies in the real world, using microscopic traffic simulation has the following advantages:

1. It is cost effective. Testing a traffic control system using microscopic simulation is much easier than doing it in the real world. This saves a lot of efforts including installment of communication hardware and deployment of detectors.
2. It is safe. For new traffic control systems that are still in the testing stage, evaluating it in the real world may cause unexpected results such as serious traffic accidents.
3. It is fast. Implementing a traffic control system in microscopic simulation can be done in a few days, and the testing can usually be accomplished with a desktop computer.
4. It is very flexible. Traffic analysts can modify parameters or traffic network settings conveniently to suit different analysis purposes. Doing the same in the real world would be cumbersome or even impossible.
5. It is controllable. By using the same random number, traffic analysts can test different traffic control strategies under exactly the same traffic condition. While it is usually impossible to replicate the exact same conditions in the real world. Since different traffic control strategies will have to be tested during different time periods, there is no way to expect the traffic conditions during those time periods to be exactly the

same. The difference in traffic conditions often makes the comparison results questionable, causing difficulties to draw valid and convincing conclusions from the results (Gartner et al. 2001).

There are many microscopic traffic simulation packages being used, including VISSIM (PTV, 2007b), CORSIM (FHWA, 1997), AIMSUN (TSS, 2003), and Paramics (Quadstone, 1999). There have been studies comparing different traffic simulation programs (Birst et al., 2007), however, no universal consensus has been reached as to which program is the best one. In this research, VISSM is chosen mainly for the following reasons:

1. VISSIM is one of the most popular microscopic traffic simulation software being widely used around the world, and has been trusted by many traffic engineering researchers and practitioners. Using VISSIM as the testing platform makes it easy for other researchers to compare their traffic control methods with the one proposed in this research.
2. VISSIM provides a NEMA editor that can code actuated traffic signal control. Actuated traffic signal control is considered to be better than pre-timed control and is used as one of the baselines in this study.
3. VISSIM has a signal control DLL (Dynamic-Link Library) interface that can be used to code and test the proposed NFACRL control method.

5.4 TESTING DESIGN

5.4.1 Testing Procedure

Testing of the proposed NFACRL control method is conducted at both isolated intersection and arterial levels. The intersection at Welsh Avenue and the intersection at Rio Grande Boulevard (three-approach intersection) in Figure 23 are chosen for isolated intersection control testing, and the entire arterial network in Figure 23 is used for arterial control testing.

For testing on the two isolated intersections, the fixed and variable phase sequence NFACRL control schemes are evaluated and compared with pre-timed and actuated control. The pre-timed and actuated control plans are optimized by Synchro (Husch and Albeck, 2001). The two NFACRL controllers are first trained using simulated traffic data and then applied to control the same simulated traffic. To make the evaluation and comparison results more convincing, each of the four control methods are tested 30 times using different random seeds.

The fixed and variable phase sequence NFACRL control schemes are extended to control the entire arterial using an independent-agent coordination method. Based on this coordination method, each intersection is controlled by one NFACRL controller. These NFACRL controllers treat each other as part of the environment and learn how to coordinate implicitly. The NFACRL controllers based on the two schemes are trained and evaluated. Their performances on arterial control are then compared with those of coordinated pre-timed and coordinated actuated control. Again, the coordinated pre-timed and coordinated actuated control plans are optimized by Synchro (Husch and Albeck, 2001). Each of the four control methods is tested 30 times independently using different random seeds.

5.4.2 Testing Under Different Flow Patterns

To better illustrate the difference among traffic flow patterns during morning, noon, and afternoon peak periods, the total entrance traffic volumes during each of these three peak periods are plotted in Figure 24. This figure shows that the total entrance traffic volumes during morning and afternoon peak periods are significantly larger than that during noon peak period. In order to give a thorough evaluation of the proposed two NFACRL control schemes, they are tested using these three sets of traffic volume data at both the isolated intersections and the arterial levels.

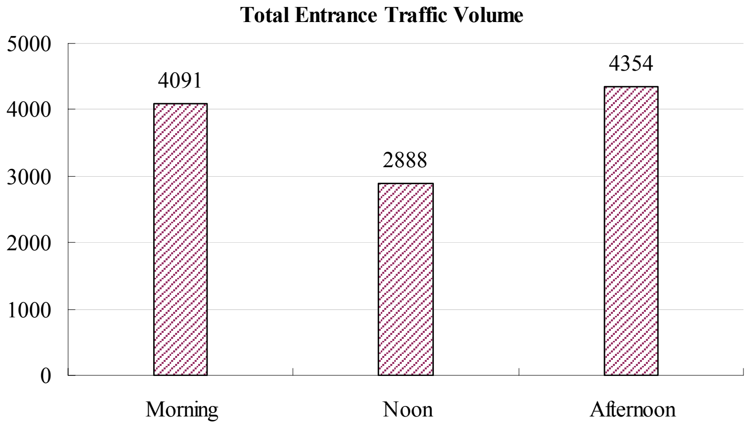


Figure 24 Total entrance traffic volumes.

5.4.3 Network Coding

GIS data from the website of the City of College Station (CCS, 2007) are used to code the arterial network. The coded arterial network in VISSIM is shown in Figure 25.

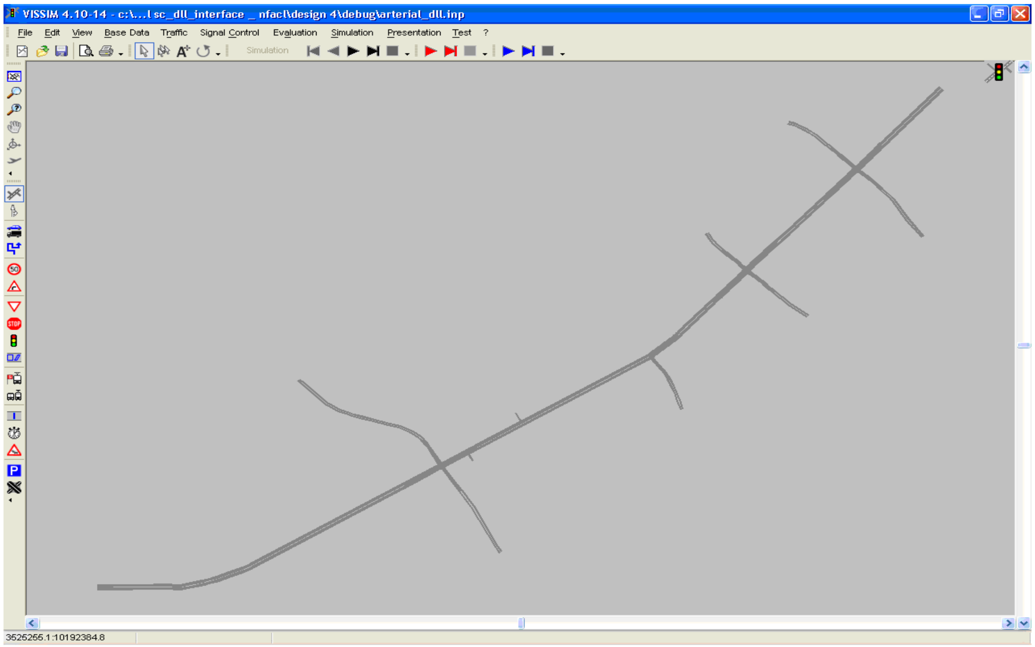


Figure 25 Coded arterial network.

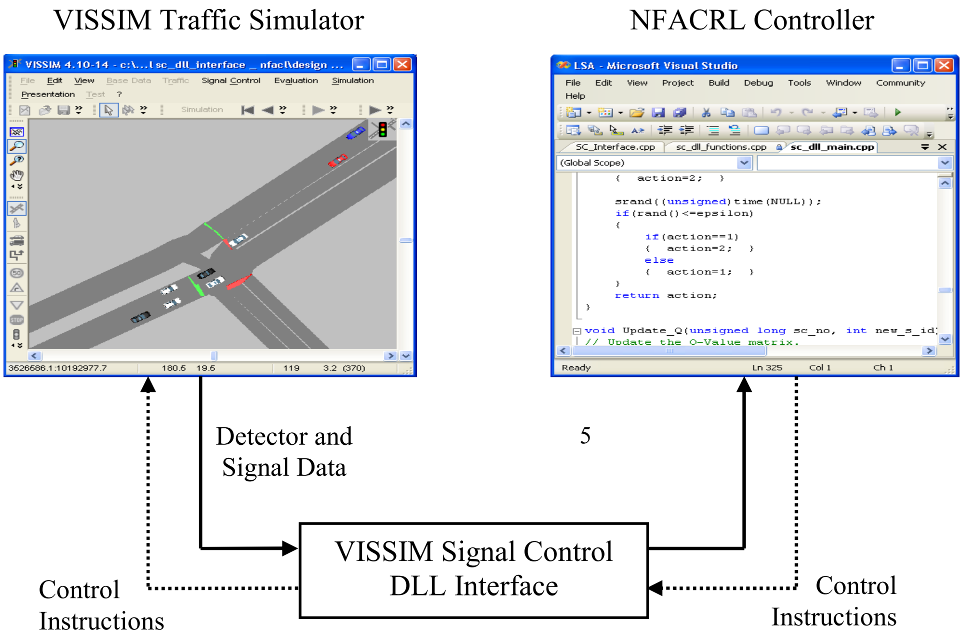
5.4.4 Algorithm Implementation

5.4.4.1 Pre-Timed and Actuated Control

Many software packages can be used to optimize pre-timed and actuated traffic control plans for both isolated intersections and arterials. Those packages include Synchro (Husch and Albeck, 2001), PASSER II (TTI, 2007) and V (TTI, 2007), and TRANSYT-7F (McTrans, 2007). Synchro is chosen for this research as it has a very friendly user interface and its performance is also comparable with or better than other packages (Zhang and Xie, 2006). Synchro is also more commonly used in practice than other packages. The optimized pre-timed and actuated control plans are coded in VISSIM using the provided fix timed controller and the NEMA controller (PTV, 2007a).

5.4.4.2 Reinforcement Learning Control

One reason for choosing VISSIM as the simulation platform is that VISSIM has a convenient DLL interface. With the help of this DLL interface, users can implement their own logics to control the simulated traffic. In this study, the two NFACRL control schemes are first coded as DLL files using the C++ language. The NFACRL control schemes in the form of DLL files communicate with the simulated traffic through the DLL interface. The entire idea of control flow using the DLL feature is illustrated in Figure 26.



```
    {   action=2;   }

    srand((unsigned)time(NULL));
    if(rand()<=epsilon)
    {
        if(action==1)
        {   action=2;   }
        else
        {   action=1;   }
    }
    return action;
}

void Update_Q(unsigned long sc_no, int new_s_id)
// Update the Q-Value matrix.
```

Figure 26 DLL interface and implementation of NFACRL control schemes.

It can be seen from Figure 26 that the DLL interface functions as a relay. It obtains detector outputs and signal states from the VISSIM microscopic traffic simulator and sends them to the NFACRL controller. In return, the NFACRL controller gives control instructions back to the VISSIM simulator to control the simulated traffic.

5.4.5 Performance Evaluation Criteria

VISSIM provides various outputs, including average delay per vehicle, average stopped delay per vehicle, and average number of stops per vehicle. Similar outputs have been used by several researchers for evaluating traffic control methods (Li, 2002; Abbas, 2001; Zhang, 1996). In this research, all these three outputs are adopted as performance evaluation criteria for both isolated intersection and arterial controls. For ease of description, these three performance criteria hereinafter are referred to as delay, stopped delay, and number of stops per vehicle.

For arterial control, three additional criteria are used, which are overall average speed, throughput, and average arterial travel time. Overall average speed is the average speed of all vehicles in the arterial system. Throughput is defined as the number of vehicles that have passed through the arterial system. Average arterial travel time is defined as the average travel time for vehicles traveling from one end of the arterial to the other end. In the rest of the report, overall average speed and average arterial travel time are referred to as speed and arterial travel time, respectively.

5.5 PERFORMANCE EVALUATION ON ISOLATED INTERSECTIONS

Two isolated intersections are chosen for evaluating the proposed NFACRL methods. The first is the intersection of Welsh Avenue and FM 2818, which is a four-approach intersection. The second is the intersection of Rio Grande Boulevard and FM 2818, which is a three-approach intersection. For ease of description, hereinafter these two intersections are referred to as four-approach and three-approach intersections. Also, the two NFACRL methods with fixed and variable phase sequences are referred to as NFACRL-F and NFACRL-V, respectively.

NFACRL controllers are like neural networks. They need to be trained before they can be used. In this study, the NFACRL controllers were trained 90 runs based on the data from each selected intersection. The trained NFACRL controllers were then applied to the two intersections for performance evaluation. As described before, for each intersection under different traffic flow conditions, the performance evaluation process was repeated 30 runs with different random seeds.

5.5.1 Evaluation with Morning Data

5.5.1.1 Four-Approach Intersection

Table 6 shows the simulation results for the four-approach intersection based on the morning peak period traffic volume data (Table 3). It can be seen that for all performance criteria, the NFACRL-V control consistently outperforms the pre-timed control, actuated control, and NFACRL-F control. Also, the NFACRL-F control performs better than the pre-timed and actuated controls in terms of delay and stopped delay. Compared to the pre-timed control, the NFACRL-F control reduces delay by 6.6 seconds per vehicle, which is 14.7% of the delay resulted from the pre-timed control. Although the NFACRL-F control performs slightly worse in terms of number of stops per vehicle, this could be solved by fine tuning the weight β_5 in Equation (55). For this scenario, actuated control has better performance than pre-timed control. This may be explained by actuated control's ability to adjust green signal length according to real-time traffic flow conditions.

The fact that for most performance criteria both NFACRL methods perform better than the optimized pre-timed and actuated control is very encouraging. It shows that it is feasible to use reinforcement learning in practical intersection traffic control problems with more than two phases, and reinforcement learning can be used to decide when to make phase switch as well as how to choose phase sequence.

Table 6 Simulation results for four-approach intersection based on morning peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	38.0	25.8	0.90
	Stdev	4.2	3.7	0.04
NFACRL-V	Average	**32.7**	**21.0**	**0.85**
	Stdev	2.7	1.8	0.05
Pre-Timed	Average	44.6	31.4	0.92
	Stdev	3.8	2.7	0.06
Actuated	Average	41.7	29.2	0.89
	Stdev	3.5	2.5	0.05
NFACRL-F vs. Pre-Timed	Improvement	6.6	5.7	0.02
	Percent. Improve. (%)	14.7	18.0	2.7
NFACRL-F vs. Actuated	Improvement	3.7	3.4	-0.01
	Percent. Improve. (%)	8.9	11.8	-0.9
NFACRL-V vs. Pre-Timed	Improvement	11.8	10.5	0.07
	Percent. Improve. (%)	26.6	33.3	7.9
NFACRL-V vs. Actuated	Improvement	9.0	8.2	0.04
	Percent. Improve. (%)	21.6	28.2	4.4

Note: $\text{Percent. Improve. (\%)} = \dfrac{\text{Improvement}}{\text{Average value of Pre} - \text{Timed or Actuated control}}$

For each of the 30 evaluation runs, the four control methods (Pre-timed, actuated, NFACRL-F, and NFACRL-V) used the same random seeds, and were evaluated under exactly the same traffic conditions. Paired-*t* tests were conducted to further compare the performance of the four control methods. One-tail paired-*t* test and a significance level of 0.05 were used in this study. The results of paired-*t* tests are presented in Table 7. Data in Tables 6 and 7 show that NFACRL-V control significantly outperforms the pre-timed and actuated control in terms of all three performance criteria. NFACRL-F control significantly reduces delay and stopped delay compared to the optimized pre-timed and actuated control. Although NFACRL-F control slightly increases the number of stops per vehicle compared to the optimized actuated control, the results in Table 7 indicate that this increase is statistically insignificant. Thus, the overall performance of the NFACRL-F control is better than that of the optimized pre-timed and actuated control.

Table 7 Paired-*t* test for four-approach intersection based on morning peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.011
	Comparison	Better	Better	Better.
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.193
	Comparison	Better	Better	No Diff.
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.001
	Comparison	Better	Better	Better

5.5.1.2 Three-Approach Intersection

Table 8 presents the simulation results for the three-approach intersection based on the morning peak period traffic data (Table 3), and Table 9 shows the corresponding paired-*t* test results. The general trend shown in Tables 8 and 9 is similar to what has been suggested by the data in Tables 6 and 7. The only difference is that the NFACRL-F control performs even better in this case, and both NFACRL control methods significantly outperform the optimized pre-timed and actuated control for all three performance criteria, indicated by the improved averages and the paired *t*-test results.

Table 8 Simulation results for three-approach intersection based on morning peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	15.0	6.9	0.47
	Stdev	0.5	0.4	0.01
NFACRL-V	Average	**14.2**	**5.8**	**0.47**
	Stdev	0.8	0.7	0.02
Pre-Timed	Average	16.5	7.3	0.56
	Stdev	1.0	0.5	0.03
Actuated	Average	16.1	7.1	0.55
	Stdev	1.0	0.5	0.03
NFACRL-F vs. Pre-Timed	Improvement	1.6	0.4	0.08
	Percent. Improve. (%)	9.5	5.6	14.8
NFACRL-F vs. Actuated	Improvement	1.1	0.2	0.07
	Percent. Improve. (%)	7.1	2.6	13.3
NFACRL-V vs. Pre-Timed	Improvement	2.3	1.5	0.08
	Percent. Improve. (%)	14.2	20.8	14.9
NFACRL-V vs. Actuated	Improvement	1.9	1.3	0.07
	Percent. Improve. (%)	11.9	18.3	13.3

Table 9 Paired-*t* test for three-approach intersection based on morning peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.024	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better

5.5.2 Evaluation with Noon Data

5.5.2.1 Four-Approach Intersection

Table 10 lists the simulation results based on the noon peak period traffic data (Table 4) for the four-approach intersection. The NFACRL-V control produces the lowest delay and number of stops per vehicle. The NFACRL-F control also generates lower delay and number of stops per vehicle than the pre-timed and actuated control. In addition, paired-*t* test results in Table

11 show that these improvements on delay and number of stops per vehicle from the NFACRL methods are statistically significant.

While the delay and number of stops per vehicle results show that the two NFACRL methods outperform the optimized pre-time and actuated control, one inconsistency in Tables 10 and 11 indicates that the two NFACRL methods do not perform well for stopped delay. For isolated intersection control, the most important performance evaluation criteria are delay and number of stops per vehicle. Since in this example the delay and numbers of stops per vehicle for the two NFACRL control methods are significantly less than those for the optimized pre-timed and actuated control, the two NFACRL methods can still be considered to be better.

Table 10 Simulation results for four-approach intersection based on noon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	18.6	**11.1**	0.62
	Stdev	0.4	0.3	0.02
NFACRL-V	Average	**18.3**	11.9	**0.52**
	Stdev	0.7	0.6	0.02
Pre-Timed	Average	19.9	11.7	0.69
	Stdev	0.8	0.6	0.02
Actuated	Average	19.2	**11.1**	0.68
	Stdev	0.7	0.6	0.02
NFACRL-F vs. Pre-Timed	Improvement	1.3	0.6	0.06
	Percent. Improve. (%)	6.6	5.5	9.2
NFACRL-F vs. Actuated	Improvement	0.6	0.0	0.06
	Percent. Improve. (%)	3.2	0.4	8.6
NFACRL-V vs. Pre-Timed	Improvement	1.6	-0.2	0.17
	Percent. Improve. (%)	8.0	-1.4	24.4
NFACRL-V vs. Actuated	Improvement	0.9	-0.8	0.16
	Percent. Improve. (%)	4.6	-6.9	23.8

Table 11 Paired-*t* test for four-approach intersection based on noon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.314	0.000
	Comparison	Better	No Diff.	Better
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.116	0.000
	Comparison	Better	No Diff.	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Worse	Better

5.5.2.2 Three-Approach Intersection

Table 12 shows the simulation results based on the noon peak period traffic volume data (Table 4) for the three-approach intersection and Table 13 shows the corresponding paired-*t* test results. The results show that both NFACRL methods significantly outperform the optimized pre-timed and actuated control in terms of all performance criteria.

Table 12 Simulation results for three-approach intersection based on noon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	8.2	2.8	0.33
	Stdev	0.8	0.3	0.03
NFACRL-V	Average	**7.1**	**2.4**	**0.26**
	Stdev	0.6	0.3	0.03
Pre-Timed	Average	9.5	3.6	0.38
	Stdev	0.3	0.2	0.01
Actuated	Average	8.9	3.4	0.37
	Stdev	0.3	0.2	0.01
NFACRL-F vs. Pre-Timed	Improvement	1.3	0.9	0.06
	Percent. Improve. (%)	14.0	24.2	14.6
NFACRL-F vs. Actuated	Improvement	0.8	0.6	0.04
	Percent. Improve. (%)	8.7	18.3	10.5
NFACRL-V vs. Pre-Timed	Improvement	2.3	1.3	0.12
	Percent. Improve. (%)	24.7	35.5	31.0
NFACRL-V vs. Actuated	Improvement	1.8	1.0	0.10
	Percent. Improve. (%)	20.1	30.6	27.7

Table 13 Paired-*t* test for three-approach intersection based on noon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better

5.5.3 Evaluation with Afternoon Data

5.5.3.1 Four-Approach Intersection

Table 14 shows the simulation results based on the afternoon peak period traffic volume data (Table 5) for the four-approach intersection.

Table 14 Simulation results for four-approach intersection based on afternoon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	36.6	25.2	**0.85**
	Stdev	1.5	1.3	0.02
NFACRL-V	Average	**33.4**	**21.6**	0.86
	Stdev	2.5	2.0	0.04
Pre-Timed	Average	45.7	31.1	1.01
	Stdev	5.3	3.3	0.10
Actuated	Average	43.3	29.5	0.97
	Stdev	4.7	3.1	0.08
NFACRL-F vs. Pre-Timed	Improvement	9.0	5.9	0.15
	Percent. Improve. (%)	19.8	18.9	15.3
NFACRL-F vs. Actuated	Improvement	6.7	4.2	0.11
	Percent. Improve. (%)	15.4	14.3	11.8
NFACRL-V vs. Pre-Timed	Improvement	12.3	9.5	0.14
	Percent. Improve. (%)	26.9	30.5	14.0
NFACRL-V vs. Actuated	Improvement	9.9	7.8	0.10
	Percent. Improve. (%)	22.9	26.6	10.5

The data in Table 14 illustrate that both NFACRL methods perform better than the optimized pre-timed and actuated control for all three criteria. More specifically, compared to the pre-timed and actuated control, the NFACRL-F control reduces delay by 9.0 and 6.7 seconds per vehicle, respectively. The corresponding improvements from the NFACRL-V control are even greater, at 12.3 and 9.9 seconds, respectively. The paired-*t* test results in Table 15 indicate that the improvements from the two NFACRL methods relative to the optimized pre-timed and actuated control are statistically significant.

Table 15 Paired-*t* test for four-approach intersection based on afternoon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better

5.5.3.2 Three-Approach Intersection

Table 16 present the simulation results based on the afternoon peak period traffic volume data (Table 5) for the three-approach intersection. The results show that for all three evaluation criteria the NFACRL-F control outperforms the optimized pre-timed and actuated control, and the NFACRL-V control again performs better than the NFACRL-F control in terms of all three criteria. The paired-*t* test results in Table 17 show that compared to the optimized pre-timed and actuated control, all the improvements from the two NFACRL methods are statistically significant.

Table 16 Simulation results for three-approach intersection based on afternoon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F	Average	11.1	4.5	0.37
	Stdev	0.4	0.4	0.02
NFACRL-V	Average	**10.4**	**4.3**	**0.32**
	Stdev	0.6	0.7	0.01
Pre-Timed	Average	13.3	5.9	0.45
	Stdev	0.7	0.5	0.02
Actuated	Average	13.3	6.0	0.45
	Stdev	0.9	0.6	0.02
NFACRL-F vs. Pre-Timed	Improvement	2.2	1.4	0.07
	Percent. Improve. (%)	16.6	24.2	16.6
NFACRL-F vs. Actuated	Improvement	2.2	1.5	0.08
	Percent. Improve. (%)	16.3	24.6	16.9
NFACRL-V vs. Pre-Timed	Improvement	3.0	1.6	0.13
	Percent. Improve. (%)	22.2	27.3	29.0
NFACRL-V vs. Actuated	Improvement	2.9	1.7	0.13
	Percent. Improve. (%)	22.0	27.7	29.2

Table 17 Paired-*t* test for three-approach intersection based on afternoon peak period data

Model	Statistics	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000
	Comparison	Better	Better	Better

5.5.4 Summary and Comparison of Performance during Morning, Noon, and Afternoon Peak Periods

The evaluation results on isolated intersections show that in general the NFACRL-F and NFACRL-V control perform significantly better than the optimized pre-timed and actuated control, and the NFACRL-V control outperforms the NFACRL-F control in most cases. For isolated intersection control, delay and number of stops per vehicle are the two most critical performance criteria. For all six scenarios considered for isolated intersections in this study, the NFACRL-V control produces lower delay and lower number of stops per vehicle than the optimized pre-timed and actuated control. The NFACRL-F control generates lower delay and lower number of stops than the optimized pre-timed and actuated control in all cases except for the morning peak period for the four-approach intersection. In this case, the NFACRL-F control generates lower delay but approximately the same number of stops compared to the optimized actuated control.

5.6 PERFORMANCE EVALUATION ON ARTERIAL

The proposed NFACRL-F and NFACRL-V methods were also evaluated on the entire arterial network shown in Figure 23. Again, the NFACRL-F and NFACRL-V controllers were trained 90 runs based on the arterial traffic data during morning, noon, and afternoon peak periods. The trained NFACRL controllers were then applied to the arterial for performance evaluation. For each of the three different traffic flow conditions (morning, noon, and afternoon), the performance evaluation process was repeated 30 runs with different random seeds. In addition to delay, stopped delay, and number of stops per vehicle, and three new performance criteria are considered for performance evaluations on arterial. These three criteria are speed, throughput, and arterial travel time.

5.6.1 Evaluation with Morning Data

Table 18 lists the simulation results for the arterial based on the morning peak period traffic data. Paired *t*-test is also performed and the results are provided in Table 19. The results show that the NFACRL-V method performs the best. It significantly improves all performance criteria over the pre-timed control and significantly improves all performance criteria but throughput over the actuated control.

Results in Table 18 also suggest that the NFACRL-F method can effectively reduce delay and increase speed. However, it does not perform well on number of stops per vehicle and arterial travel time compared to the optimized coordinated pre-timed and actuated control. One reason for this phenomenon is that the coordinated pre-timed and actuated control strategies use fixed cycle length and offset to maximize the green signal bandwidth along the arterial. This is like giving vehicles traveling along the arterial higher priority. Thus the number of stops for vehicles traveling along the arterial and the arterial travel time can be significantly reduced. While for the NFACRL-F method, vehicles from arterial and cross streets are treated the same. Another possible reason is that the fixed phase sequence restricts NFACRL-F method’s ability to minimize number of stops per vehicle. It may often happen that a platoon of vehicles is coming from the arterial and there are only a few vehicles waiting in the cross streets, and the arterial directions cannot be given green signal immediately due to the phase sequence restriction and minimum green times that have to be served for the cross-street movements.

Table 18 Simulation results for arterial based on morning peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F	Average	25.1	55.9	**31.5**	1.52	4104	212.1
	Stdev	0.4	2.3	1.7	0.05	10	3.1
NFACRL-V	Average	**25.6**	**52.9**	32.1	**1.25**	4104	**186.1**
	Stdev	0.4	1.9	1.5	0.04	9	2.5
Pre-Timed	Average	23.8	63.6	42.2	1.46	4093	211.3
	Stdev	0.5	3.1	2.0	0.06	11	3.0
Actuated	Average	24.7	58.3	38.4	1.35	**4112**	194.4
	Stdev	0.5	2.8	1.9	0.05	11	2.5
NFACRL-F vs. Pre-Timed	Improvement	1.3	7.7	10.7	-0.06	11	-0.8
	Percent. Improve. (%)	5.4	12.1	25.4	-4.2	0.3	-0.4
NFACRL-F vs. Actuated	Improvement	0.4	2.4	6.9	-0.17	-8	-17.7
	Percent. Improve. (%)	1.6	4.1	18.0	-12.6	-0.2	-9.1
NFACRL-V vs. Pre-Timed	Improvement	1.8	10.7	10.1	0.21	11	25.2
	Percent. Improve. (%)	7.7	16.8	24.0	14.2	0.3	11.9
NFACRL-V vs. Actuated	Improvement	1.0	5.4	6.3	0.10	-8	8.3
	Percent. Improve. (%)	3.9	9.3	16.5	7.3	-0.2	4.3

Table 19 Paired-t test for arterial based on morning peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000	0.000	0.000	0.139
	Comparison	Better	Better	Better	Worse	Better	No Diff.
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Worse	Worse	Worse
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Better	Worse	Better

5.6.2 Evaluation with Noon Data

Table 20 presents the simulation results for the arterial based on the noon peak period traffic data. It is obvious that the NFACRL-V method outperforms the coordinated pre-timed and coordinated actuated controls for all performance criteria. The paired-*t* test results in Table 21 also indicate that compared to the coordinated pre-timed and coordinated actuated controls all improvements from the NFACRL-V method are statistically significant.

The NFACRL-F control in this case produces better speed, delay, stopped delay, and throughput results than the coordinated pre-timed and coordinated actuated controls, and these improvements are statistically significant. However, the NFACRL-F control generates larger number of stops per vehicle than the coordinated pre-timed control and longer arterial travel time than the coordinated actuated control. The result from the NFACRL-F control in this case shows the same trend as the result from the NFACRL-F control based on the morning data suggests. It seems that the optimized coordinated pre-timed and coordinated actuated control methods favor the arterial more than the cross streets. Thus, they generally produce good performance along the arterial such as less arterial travel time. However, the overall performance of the system may not be the best. The NFACRL methods give the arterial and the cross streets equal priority. They may not have the best performance for the arterial. However, they usually generate better performance for the entire system than the optimized coordinated pre-timed and coordinated actuated control.

Table 20 Simulation results for arterial based on noon peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F	Average	28.3	42.0	**23.0**	1.38	**2881**	196.2
	Stdev	0.2	1.0	0.6	0.04	10	2.6
NFACRL-V	Average	**28.8**	**39.9**	25.0	**1.06**	2877	**168.4**
	Stdev	0.2	1.0	0.9	0.03	10	1.3
Pre-Timed	Average	26.8	49.0	31.0	1.35	2859	200.0
	Stdev	0.3	1.4	0.9	0.02	10	1.7
Actuated	Average	26.3	52.0	28.4	1.67	2855	180.4
	Stdev	0.8	4.1	2.0	0.11	14	1.6
NFACRL-F vs. Pre-Timed	Improvement	1.5	7.0	8.0	-0.03	22	3.8
	Percent. Improve. (%)	5.4	14.3	25.7	-2.3	0.8	1.9
NFACRL-F vs. Actuated	Improvement	2.0	10.0	5.4	0.29	26	-15.8
	Percent. Improve. (%)	7.6	19.3	19.0	17.3	0.9	-8.7
NFACRL-V vs. Pre-Timed	Improvement	1.9	9.0	6.0	0.29	18	31.6
	Percent. Improve. (%)	7.2	18.5	19.3	21.6	0.6	15.8
NFACRL-V vs. Actuated	Improvement	2.5	12.1	3.4	0.61	22	12.0
	Percent. Improve. (%)	9.4	23.2	12.0	36.6	0.8	6.7

Table 21 Paired-t test for arterial based on noon peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F vs. Pre-Timed	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Worse	Better	Better
NFACRL-F vs. Actuated	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Better	Better	Worse
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	Comparison	Better	Better	Better	Better	Better	Better

5.6.3 Evaluation with Afternoon Data

Table 22 presents the simulation results for the arterial based on the afternoon peak period traffic data. In this scenario, the NFACRL-V method consistently performs the best in terms of all performance criteria. The paired-*t* test results in Table 23 also suggest that compared to the coordinated pre-timed and coordinated actuated controls all improvements from the NFACRL-V method are statistically significant.

The results in Table 23 also show that, compared to the coordinated pre-timed and coordinated actuated control, the NFACRL-F control produces better or approximately the same results in terms of speed, delay, stopped delay, and throughput. However, the NFACRL-F control generates larger number of stops per vehicle and longer arterial travel time than both the coordinated pre-timed and coordinated actuated control.

Table 22 Simulation results for arterial based on afternoon peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F	Average	24.1	62.3	37.5	1.66	4365	208.7
	Stdev	0.3	2.2	1.7	0.04	15	2.2
NFACRL-V	Average	**25.2**	**56.1**	**35.7**	**1.31**	**4368**	**179.6**
	Stdev	0.4	2.2	1.9	0.03	15	1.8
Pre-Timed	Average	23.9	63.9	40.4	1.45	4359	189.4
	Stdev	0.7	4.5	2.8	0.08	16	2.4
Actuated	Average	24.2	62.0	39.9	1.43	4359	181.5
	Stdev	0.7	4.3	2.8	0.08	15	2.1
NFACRL-F vs. Pre-Timed	Improvement	0.2	1.6	2.8	-0.20	7	-19.4
	Percent. Improve. (%)	1.0	2.5	7.1	-14.0	0.2	-10.2
NFACRL-F vs. Actuated	Improvement	-0.1	-0.3	2.4	-0.23	6	-27.2
	Percent. Improve. (%)	-0.3	-0.5	5.9	-16.0	0.1	-15.0
NFACRL-V vs. Pre-Timed	Improvement	1.3	7.8	4.7	0.14	10	9.8
	Percent. Improve. (%)	5.5	12.3	11.6	9.9	0.2	5.2
NFACRL-V vs. Actuated	Improvement	1.0	5.9	4.2	0.12	9	1.9
	Percent. Improve. (%)	4.1	9.6	10.6	8.4	0.2	1.1

Table 23 Paired-*t* test for arterial based on afternoon peak period data

Model	Statistics	Speed (mph)	Delay (s/veh)	Stopped Delay (s/veh)	Number of Stops per Veh	Throughput (# of Veh.)	Arterial Travel Time (s)
NFACRL-F vs. Pre-Timed	*p*-value	0.023	0.018	0.000	0.000	0.004	0.000
	Comparison	Better	Better	Better	Worse	Better	Worse
NFACRL-F vs. Actuated	*p*-value	0.268	0.328	0.000	0.000	0.008	0.000
	Comparison	No Diff.	No Diff.	Better	Worse	Better	Worse
NFACRL-V vs. Pre-Timed	*p*-value	0.000	0.000	0.000	0.000	0.001	0.000
	Comparison	Better	Better	Better	Better	Better	Better
NFACRL-V vs. Actuated	*p*-value	0.000	0.000	0.000	0.000	0.001	0.000
	Comparison	Better	Better	Better	Better	Better	Better

5.6.4 Summary and Comparison of Performance during Morning, Noon, and Afternoon Peak Periods

The paired-*t* test results in Tables 19, 21, and 23 show that, for all tests on the arterial, the NFACRL-V method produces better results than the coordinated pre-timed and coordinated actuated controls for almost all performance criteria. The only exception is the throughput value for the test based on morning peak period data, which is slightly less than that of the coordinated actuated control. The difference between the two throughputs is 8 vehicles, which is only 0.2% of the throughput resulted from the coordinated actuated control. In the meantime, the NFACRL-V control reduces delay by 5.4 seconds per vehicle, which is 9.3% of the delay resulted from the coordinated actuated control.

For the NFACRL-F control, the paired-*t* test results suggest that it generally performs better on delay, speed, stopped delay, and throughput compared to the coordinated pre-timed and coordinated actuated control. However, it does not perform well on number of stops per vehicle and arterial travel time.

To further compare the performance of the proposed NFACRL methods during morning, noon, and afternoon peak periods, the delay reductions from the two NFACRL methods relative to the coordinated pre-timed control during different peak periods are plotted in Figure 27. Also plotted are the total entrance traffic volume data during these three peak periods. From Figure 27, it seems that there is no direct connection between the total entrance traffic volume and the relative delay reductions from the two NFACRL methods.

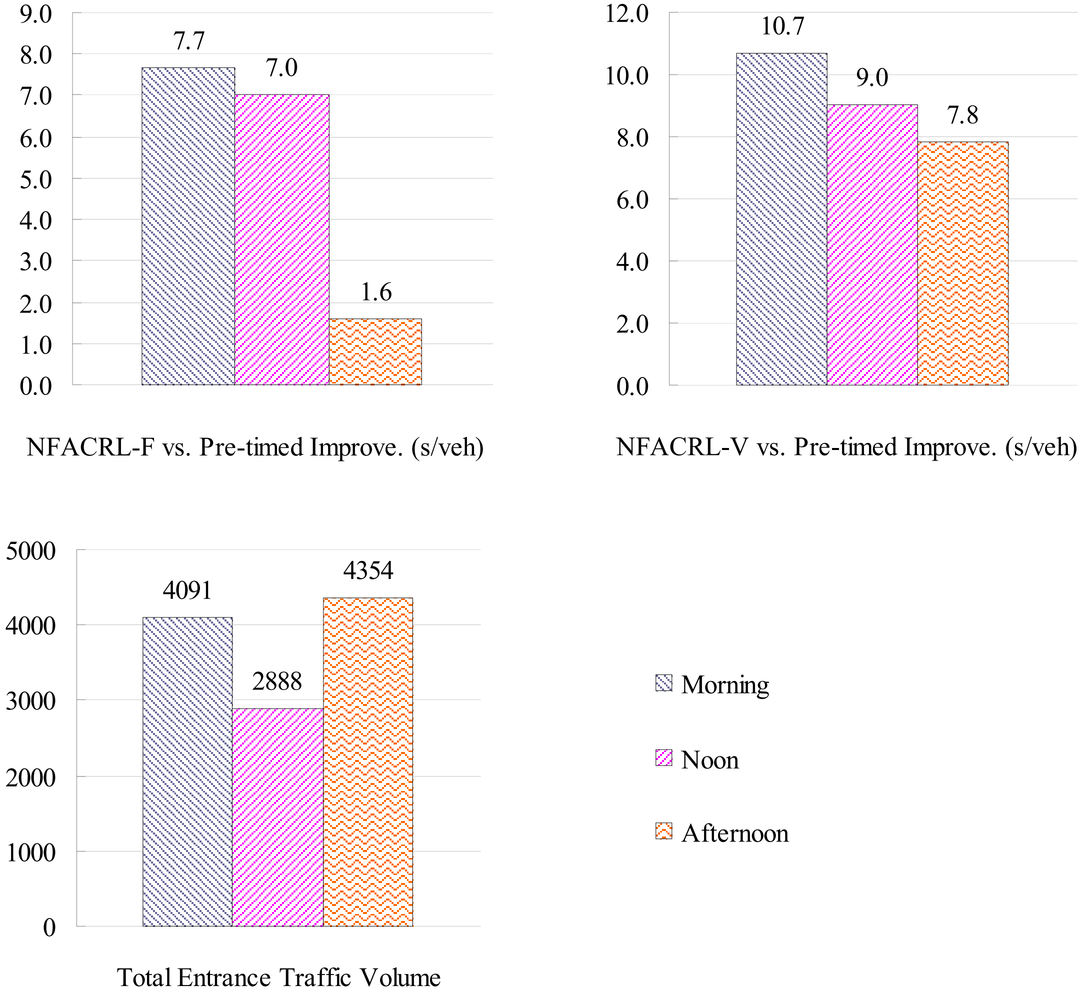


Figure 27 Delay improvements from the NFACRL methods relative to the coordinated pre-timed control and corresponding traffic volumes for the arterial.

A close look at the traffic volume data in Tables 3 through 5 suggests that the performance of the NFACRL methods relative to the coordinated pre-timed control may depend on the proportion of cross-street traffic and cross-street turning traffic. Cross-street traffic and cross-street turning traffic are defined in Figure 28 and Figure 29, respectively. The cross-street and cross-street turning traffic data are listed in Table 24. It shows that during morning peak period 66.2% and 49.7% of the total entrance traffic are cross-street traffic and cross-street turning traffic, respectively. While for afternoon peak period only 54.7% and 38.0% of the total entrance traffic are cross-street traffic and cross-street turning traffic, respectively. Higher percentages of cross-street traffic and cross-street turning traffic may result in narrower green signal bandwidth along the arterial, thus the benefits of coordinating different traffic signal controllers become smaller. Also, the large amount of traffic turning into the arterial can considerably affect the queue lengths of the arterial direction at each intersection. The initial queues may have negative impact on the performance of the coordinated pre-timed and coordinate actuated control. This probably is the reason why during the morning peak period the relative improvements from the two NFACRL methods are more significant.

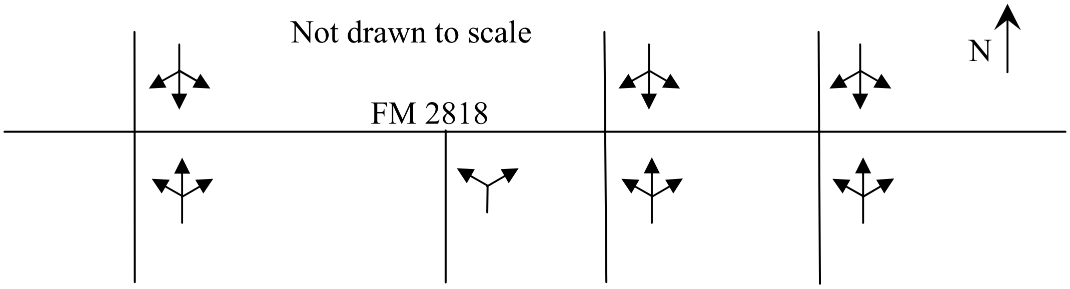


Figure 28 Cross-street traffic.

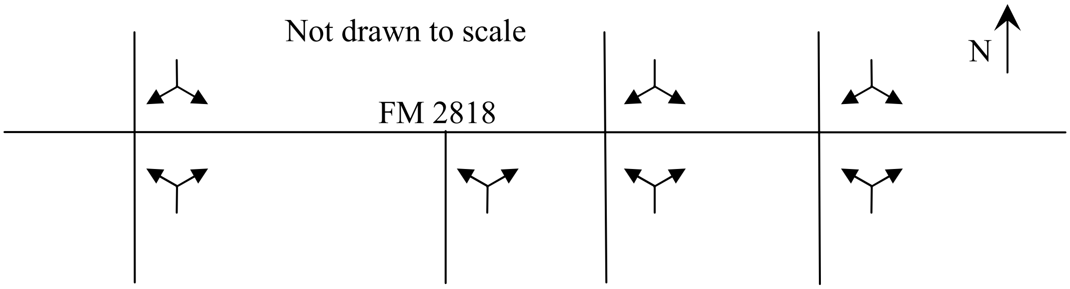


Figure 29 Cross-street turning traffic.

Table 24 Cross street traffic during morning, noon, and afternoon peak periods

	Morning	Noon	Afternoon
Total entrance traffic (A)	4091	2888	4354
Total cross-street traffic (B)	2710	1624	2383
(B/A)*100	**66.2**	**56.2**	**54.7**
Total cross-street turning traffic (C)	2035	1268	1656
(C/A)*100	**49.7**	**43.9**	**38.0**

5.7 SUMMARY

In this chapter, the proposed NFACRL-F and NFACRL-V control schemes are evaluated at two isolated intersections and on the entire arterial. The evaluations are performed using VISSIM simulation based on geometric and traffic data collected from a four-intersection arterial (FM 2818) in College Station, Texas. To better assess the performance of the proposed new methods under different traffic demand conditions, the evaluations are conducted using traffic data during morning, noon, and afternoon peak periods.

A four-approach intersection and a three-approach intersection in the arterial are selected for isolated intersection testing. The testing results show that in almost all cases, the two NFACRL control methods produce lower delay, stopped delay, and number of stops per vehicle compared to the optimized pre-timed and actuated control. Paired-*t* tests are also conducted and show that all the improvements from the two NFACRL methods are statistically significant. Although the two NFACRL methods produce slightly larger stopped delay for the four-approach intersection using the noon peak period data, they are still considered to perform better than the optimized pre-timed and actuated control due to the considerably reduced delay and number of stops per vehicle.

The NFACRL-F and NFACRL-V control schemes are also extended for arterial control. The results show that the NFACRL-V control significantly outperforms the coordinated pre-timed and coordinated actuated control for speed, delay, stopped delay, number of stops per vehicle, and arterial travel time. The NFACRL-V control also produces good throughput results in most cases. The NFACRL-F control in this research exhibits less flexibility than the NFACRL-V control. However, in most cases the NFACRL-F control still generates better delay, speed, stopped delay, and throughput results compared to the optimized coordinated pre-timed and coordinated actuated control. The results also show that the NFACRL-F control does not perform well for number of stops per vehicle and arterial travel time compared to the coordinated pre-timed and coordinated actuated control. Possible reasons for this are discussed, and future studies are needed to address this problem. The performance of the NFACRL methods during the morning, noon, and afternoon peak periods are also compared and analyzed. The result suggests that the benefits of using the NFACRL methods for arterial control may be even larger with higher proportions of cross-street traffic and cross-street turning traffic.

CHAPTER 6. CONCLUSIONS AND FUTURE RESEARCH

6.1 CONTRIBUTIONS

This research investigates the application of reinforcement learning to adaptive traffic signal control. An adaptive traffic signal control method based on neuro-fuzzy actor-critic reinforcement learning (NFACRL) is developed and evaluated at both isolated intersections and arterial. Compared to previous studies using reinforcement learning for traffic signal control, this research has the following contributions:

1. A comprehensive review of existing traffic signal control methods and reinforcement learning is presented in this study. This review systematically points out the connection between MDP, dynamic programming, and reinforcement learning. It also clearly explains the advantages of modeling traffic signal control as a MDP problem and using reinforcement learning methods to solve it.
2. By introducing the NFACRL, the curse of dimensionality and generalization problems associated with traditional reinforcement learning methods can be properly solved.
3. Bingham (2001) also combined fuzzy logic and reinforcement learning for traffic signal control. In that study, the fuzzy rules need to be prespecified explicitly. When the state space is large, there could be several hundreds of fuzzy rules. Specifying so many fuzzy rules is very cumbersome. In the proposed NFACRL method, action weights are introduced and there is no need to define each fuzzy rule.
4. Most previous studies considered traffic signal control with only two phases, and phase sequence optimization was not investigated. In practice, typical intersections are controlled by four or even more phases. To show the potential of applying reinforcement learning control methods in the real world, a fixed (NFACRL-F) phase sequence control strategy and a variable (NFACRL-V) phase sequence control strategy are proposed in this research. Four-phase control and three-phase control are used for the four-approach and three-approach intersections, respectively. This is the first time that complicated and realistic phase configurations are considered in truly adaptive traffic signal control based on reinforcement learning. Also, it is the first time that reinforcement learning is used for phase sequence selection.

5. Various strategies for coordinating agents in a multi-agent system are reviewed and their pros and cons are discussed in details. Finally, a simple but robust independent-agent strategy is adopted to coordinate different NFACRL-F and NFACRL-V controllers.
6. A new reward function is proposed in this research. The new reward function takes into account multiple factors such as delay and number of stops, and has been shown to perform well in this research. Theoretical reasons for choosing this new reward function are also provided.
7. A comprehensive comparison of the proposed NFACRL control methods with pre-timed and actuated controls is conducted based on VISSIM simulation. Most previous studies did not use a commonly used microscopic traffic simulation platform for performance evaluation and did not compare their methods with optimized pre-timed or actuated controls. Self-developed simulation platforms give users more control over the simulation process, but the simulated traffic environment may not be as close to the true traffic condition as those from commonly-used simulation tools such as VISSIM. Also, pre-timed and actuated control strategies are the two most widely used control methods in practice. It is necessary to show that the proposed new methods are better than them.

6.2 MAJOR FINDINGS

There are three major findings in this study. First of all, this study shows that it is feasible to apply reinforcement learning to adaptive isolated intersection control with more than two phases and complicated phase configurations. For all tests on isolated intersections, the proposed NFACRL-F and NFACRL-V methods produce considerably less delay than the optimized pre-timed and actuated control. In most cases, the two NFACRL methods generate significantly smaller stopped delay and number of stops per vehicle.

Secondly, it is found that reinforcement learning can be used for phase sequence selection. In fact, the NFACRL-V method with phase sequence selection ability consistently outperforms the NFACRL-F method with fixed phase sequence for most performance evaluation criteria in this research. As only two phases were considered in most previous studies, none of them investigated the phase sequence selection using reinforcement learning.

Lastly, this research shows that reinforcement learning has the potential to be used for realistic arterial adaptive traffic signal control. The proposed NFACRL-F and NFACRL-V control strategies are applied to the control of a four-intersection arterial network based on VISSIM simulations. A simple but robust independent-agent coordination strategy is considered, in which control agents learn to coordinate with each other implicitly. The evaluation results show that both NFACRL methods can effectively increase overall network speed and reduce delay and stopped delay compared to the optimized coordinated pre-timed and coordinated actuated control. In addition, the NFACRL-V control exhibits more flexibility and produces significantly better performance in terms of number of stops per vehicle and arterial travel time, compared to the optimized coordinated pre-timed and coordinated actuated control. It is also found that the NFACRL-F method does not perform well for criteria such as number of stops per vehicle and arterial travel time. This could be the problem of the reward function or the definition of state variables. Overall, the test results show that the proposed NFACRL-F and NFACRL-V methods are promising tools for arterial adaptive traffic signal control.

6.3 FUTURE RESEARCH

Although encouraging results are obtained in this research, further studies are still needed to address the following problems:

1. In this research, the decision interval is either 3 or 7 seconds. If a green extension decision is made, the next decision point is 3 seconds later. In other words, the green extension is 3 seconds; if a termination decision is made, the next decision point would be 7 seconds later. Because, in addition to 3-second minimum green time for the next phase, a 3-second yellow time and a 1-second all-red time need to be considered to ensure safety. In future studies, smaller green extension such as 2 seconds can be considered, this should further improve the performance of the NFACRL methods.
2. The coordination strategy considered in this research is fairly simple. Even with this simple coordination strategy, arterial control using the NFACRL-V method still performs better than the coordinated pre-timed and coordinated actuated control in all cases. The NFACRL-F method also outperforms the coordinated pre-timed and coordinated actuated controls in many cases in terms of speed, delay, and stopped delay. However, the NFACRL-F control does not perform well for number of stops per vehicle and arterial travel time. One possible solution to this problem is to add four new state variables

representing upstream traffic arrival information. With advanced traffic arrival information, the NFACRL-F control should be able to better adjust its control strategy such that the number of stops and arterial travel time can be reduced.

3. A number of factors can affect the performance of the NFACRL control methods, for instance, the fuzzy membership function, learning rate β (Equations (38) and (39)), choice of state variables, ε in the action selection method (Equation (12)), and β_i in the reward function (Equation (55)). Due to limited computation resources and considerable amount of time spent on algorithm developing and debugging, a comprehensive evaluation of the effects of various factors on the NFACRL control performance is not conducted in this research. In future studies, a comprehensive evaluation needs to be conducted.
4. In this research, the proposed NFACRL methods are only applied to isolated intersection and arterial control. By using the same coordination strategy adopted in this research, it would be interesting to see how the NFACRL methods perform on a signalized street network.
5. In this research, it is assumed that queue lengths can be observed accurately and there are no pedestrians. To apply the proposed NFACRL method to the real world, it is necessary to conduct future studies that take pedestrians into account. Also, an accurate and reliable queue detection method needs to be developed.

REFERENCES

Abbas, M. M. A Real Time Offset Transitioning Algorithm for Coordinating Traffic Signals. Ph.D. Dissertation. Department of Civil Engineering, Purdue University. 2001.

Abdulhai, B., R. Pringle, and G. J. Karakoulas. Reinforcement Learning for True Adaptive Traffic Signal Control. *Journal of Transportation Engineering*, Vol. 129, No. 3, 2003, pp. 278-285.

Barto, A. G., R. S. Sutton, and C. W. Anderson. Neuronlike Adaptive Elements that Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 13, No. 5, 1983, pp. 834-846.

Barto, A. G. and S. Mahadevan. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems: Theory and Applications*, Vol. 13, No. 4, 2003, pp. 343-379.

Berenji, H. R. and P. Khedkar. Learning and Tuning Fuzzy Controllers through Reinforcement. *IEEE Transactions on Neural Networks*, Vol. 3, No. 5, 1992, pp. 724-740.

Bhatnagar, S., and J. R. Panigrahi. Actor-Critic Algorithm for Hierarchical Markov Decision Processes. *Automatica*, Vol. 42, No. 4, 2006, pp. 637-644.

Bingham, E. Neurofuzzy Traffic Signal Control. Master Thesis. Department of Engineering Physics and Mathematics, Helsinki University of Technology. 1998.

Bingham, E. Reinforcement Learning in Neurofuzzy Traffic Signal Control. *European Journal of Operational Research*, Vol. 131, No. 2, 2001, pp. 232-241.

Birst, S., J. Baker, and E. Shouman. Comparison of Traffic Simulation Models to HCM 2000 Using Various Traffic Levels under Pre-timed Signal Control. In *Proceedings (CD-ROM) of the 86th Transportation Research Board Annual Meeting*. Washington D.C., 2007

Borkar, V. S. An Actor-Critic Algorithm for Constrained Markov Decision Processes. *System & Control Letters*, Vol. 54, No. 3, 2005, pp 207-213.

Bowling, M., and M. Veloso. Multiagent Learning Using a Variable Learning Rate. *Artificial Intelligence*. Vol. 136, No. 2, 2002, pp. 215-250.

Butenko, S. Class Notes for INEN 623 – Nonlinear and Dynamic Programming. Department of Industrial and Systems Engineering, Texas A&M University. College Station, TX, 2005.

City of College Station GIS file (CCS). Brazos County Streets. http://www2.cstx.gov/gisdownloads/downloads/County_Streets.zip. Accessed on May 20, 2007.

Chalkiadakis, G. and C. Boutilier. Coordination in Multiagent Reinforcement Learning: A Bayesian Approach. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS-03)*. Melbourne, Australia, 2003, pp. 709-716.

Chiu, S., and S. Chand. Adaptive Traffic Signal Control Using Fuzzy Logic. In *Proceedings of Second IEEE International Conferences on Fuzzy Systems*. San Francisco, CA, USA. 1993, pp. 1371-1376.

Choy, M. C., D. Srinivasan, and R. L. Cheu. Cooperative, Hybrid Agent Architecture for Real-Time Traffic Signal Control. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 33, No. 5, 2003a, pp. 597-607.

Choy M. C., R. L. Cheu, D. Srinivasan, and F.Logi. Real-Time Coordinated Signal Control Using Agents with Online Reinforcement Learning. In *Proceedings (CD-ROM) of the 82nd Transportation Research Board Annual Meeting*. Washington D.C., 2003b.

Crites, R. H. and A. G. Barto. Elevator Group Control Using Multiple Reinforcement Learning Agents. *Machine Learning*, Vol. 33, No. 2-3, 1998, pp. 235-262.

Dell'Olmo, P., and P. B. Mirchandani. REALBAND: An Approach for Real-Time Coordination of Traffic Flows on a Network. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1494*, TRB, National Research Council, Washington, D.C., 1995, pp. 106-116.

Elahi, S. M., A. E. Radwan, and K. M. Goul. Knowledge-Based System for Adaptive Traffic Signal Control. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1324*, TRB, National Research Council, Washington, D.C., 1987, pp. 115-122.

Federal Highway Administration (FHWA). Improving Traffic Signal Operations. http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/13466.pdf, 1995, Accessed on May 28, 2007.

Federal Highway Administration (FHWA). CORSIM User's Manual, Version 1.03, Federal Highway Administration, Mclean, VA, 1997.

The MathWorks, Inc. Fuzzy Logic Toolbox 2 User's Guide. Natick, MA., 2007.

Gajjar, G. R., S. A. Khaparde, P. Nagaraju, and S. A. Soman. Application of Actor-Critic Learning Algorithm for Optimal Bidding Problem of a Genco. *IEEE Transactions on Power Systems*, Vol. 18, No. 1, 2003, pp. 11-18.

Gartner, N. H. Prescription for Demand-Responsive Urban Traffic Control. In *Transportation Research Record: Journal of the Transportation Research Board, No. 881*, TRB, National Research Council, Washington, D.C., 1982, pp. 73-76.

Gartner, N. H. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record: Journal of the Transportation Research Board, No. 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 75-81.

Gartner, N.H., C. Stamatiadis, and F. J. Tarnoff. Development of Advanced Traffic Signal Control Strategies for Intelligent Transportation Systems: Multilevel Design. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1494*, TRB, National Research Council, Washington, D.C., 1995, pp. 98-105.

Gartner, N. H., F. J. Pooran, and C. M. Andrews. Implementation of the OPAC Adaptive Control Strategy in a Traffic Signal Network. In *Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference*. Oakland, CA, 2001, pp. 195-200.

Gosavi, A. Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning. Kluwer Academic Publishers. Norwell, Massachusetts, 2003.

Henry, R. D., R. A. Ferlis, and J.L. Kay. Evaluation of UTCS Control Strategies—Executive Summary. FHWA Report No. FHWA-RD-76-149, 1976.

Transportation Research Board (TRB). *Highway Capacity Manual 2000*, National Research Council, Washington, D. C., 2000.

Holroyd, J. and D.I. Robertson. Strategies for Area Traffic Control Systems Present and Future. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Report No. LR569, 1973.

Hu, J. and M. P. Wellman. Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Madison, WI, 1998, pp. 242-250.

Hu, J., and M. P. Wellman. Nash Q-Learning for General-Sum Stochastic Games. *Journal of Machine Learning Research*. Vol. 4, 2003. pp. 1039-1069

Husch, D, and J. Albeck. Synchro 5 User Guide. Trafficware, Albany, CA, 2001.

Hunt, P.B. A Traffic Responsive Method of Coordinating Signals. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Report No. LR1014, 1981.

Janssens, D., Y. Lan, G, Wets, and G. Chen. Allocating Time and Location Information to Activity–Travel Patterns through Reinforcement Learning. *Knowledge-Based Systems*, Vol. 20, No. 5, 2007, pp. 466-477.

Jiang, J. S. R, C. T. Sun, and E. Mizutani. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice Hall. Upper Saddle River, New Jersey, 1997.

Jouffe, L. Actor-Critic Learning Based on Fuzzy Inference System. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*. Beijing, China. Vol. 1, 1996, pp. 339-344.

Jouffe, L. Fuzzy Inference System Learning by Reinforcement Methods. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*. Vol. 28, No. 3, 1998, pp. 338-355.

Li, H. Traffic Adaptive Control for Isolated, Over-Saturated Intersections. Ph.D. Dissertation. Department of Civil Engineering, University of Hawaii, 2002.

Lin, F. B. Comparative Analysis of Two Logics for Adaptive Control of Isolated Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1194*, TRB, National Research Council, Washington, D.C., 1988, pp. 6-14.

Lin, C. T. and C. S. G. Lee. Reinforcement Structure/Parameter Learning for Neural-Network-Based Fuzzy Logic Control Systems. *IEEE Transactions on Fuzzy Systems*. Vol. 2, No. 1, 1994, pp. 46-63.

Littman, M. L. Markov Games as A Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the 11th International Conference on Machine Learning*. San Francisco, CA. 1994, pp. 157-163.

Lowrie, P. R. The Sydney Co-ordinated Adaptive Traffic System – Principles, □ Methodology, Algorithms. In *Proceedings of the International Conference on Road Traffic Signaling*. London, UK, 1982, pp. 67-70.

Meyer, M. D. A Toolbox for Alleviating Traffic Congestion and Enhancing Mobility. Institute of Transportation Engineers, 1997.

Mirchandani, P. and L. Head. A Real-Time Traffic Signal Control System: Architecture, Algorithms, and Analysis. *Transportation Research Part C*, Vol. 9, No. 6, 2001, pp. 415-432.

Morgan, J. T. and J. D. C. Little. Synchronizing Traffic Signals for Maximal Bandwidth. *Operations Research*, Vol. 12, No. 6 (Special Transportation Science Issue), 1964, pp. 896-912.

Murat, Y. S., and E. Gedizlioglu. A Fuzzy Logic Multi-Phased Signal Control Model for Isolated Junctions. *Transportation Research Part C,* Vol. 13, No. 1, 2005, pp. 19-36.

National Electrical Manufacturers Association (NEMA), Traffic Controller Assemblies. NEMA Standards, Publication N. TS 2-1992. Washington, D.C., 1992.

Niittymaki, J., and M. Pursula. Signal Control Using Fuzzy Logic. *Fuzzy Sets and Systems*, Vol. 116, No. 1, 2000, pp. 11-22.

Owen, L. E., and C. M. Stallard. Rule-Based Approach to Real-Time Distributed Adaptive Signal Control. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1683*, TRB, National Research Council, Washington, D.C., 1995, pp. 95-101.

Panait, L., and S. Luke. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems*. Vol. 11, 2005, pp. 387-434.

Pappis, C. P., and E. H. Mamdani. A Fuzzy Logic Controller for a Traffic Junction. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 7, No. 10, 1977, pp. 707-717.

Park, K. H., Y. J. Kim, and J. H. Kim. Modular Q-learning Based Multi-Agent Cooperation for Robot Soccer. *Robotics and Autonomous Systems*, Vol. 35, No. 2, 2001, pp. 109-122.

Texas Transportation Institute (TTI). PASSER II-02, Version 1.0, http://ttisoftware.tamu.edu/fraPasserII_02.htm. Accessed on May 20, 2007.

Texas Transportation Institute (TTI). PASSER V-03, Version 2.3, http://ttisoftware.tamu.edu/fraPasserV_03.htm. Accessed on May 20, 2007.

Porche, I. R. Dynamic Traffic Control: Decentralized and Coordinated Methods. Ph.D. Dissertation. Department of Electrical Engineering and Computer Science, The University of Michigan. 1997.

PTV America, Inc. NEMA Editor Manual Version 5. http://ptvamerica.com/nse/manual/Manual_Nema.pdf. Accessed on May 20, 2007a

PTV Planung Transport Verkehr AG, VISSIM User Manual, Version 4.30, Karlsruhe, Germany March 2007b.

Puterman, M. L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., New York, NY, 1994.

Quadstone, Ltd, Paramics Traffic Simulation Modeler Reference Manual, Version 3.0, Edinburgh, Quadstone, 1999.

Robertson, D. I. and R. D. Bretherton. Optimal Control of an Intersection for Any Known Sequence of Vehicle Arrivals. 2nd IFAC-IFIP-IFORS Symposium of Traffic Control and Transportation Systems., North-Holland, Amsterdam, 1974, pp. 3-17.

Roess, R. P., E. S. Prassas, W. R. McShane. Traffic Engineering (3rd Edition). Pearson Education, Inc. Upper Saddle River, New Jersey, 2004.

Ross, T. J. Fuzzy Logic with Engineering Applications (2nd Edition). John Wiley & Sons, Ltd., England, 2004.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. Learning Representations by Back-Propagating Errors. *Nature*, Vol. 323, 1986, pp. 533-536.

Schrank, D. and T. Lomax. The 2005 Urban Mobility Report. Texas Transportation Institute. College Station, TX, 2005.

Peek Traffic Limited, Siemens Traffic Controls and TRL Limited (Peek). SCOOT http://www.scoot-utc.com/WhatIsSCOOT.php?menu=Overview Accessed on April 27, 2007.

Sen, S. and K. L. Head. Controlled Optimization of Phases at An Intersection. *Transportation Science*. Vol. 31, No. 1, 1997, pp. 5-17.

Shelby, S. G. Single-Intersection Evaluation of Real-Time Adaptive Traffic Signal Control Algorithms. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1867*, TRB, National Research Council, Washington, D.C., 2004, pp. 183-192.

Sims, A. G. and K. W. Dobinson. The Sydney Coordinated Adaptive Traffic (SCAT) System Philosophy and Benefits. *IEEE Transactions on Vehicular Technology*, Vol. VT-29, No. 2, 1980, pp. 130-137.

Srinivasan, D. and M. C. Choy. Cooperative Multi-Agent System for Coordinated Traffic Signal Control. *IEE Proceedings - Intelligent Transport Systems*. Vol. 153, No. 1, 2006, pp. 41-50.

Swaminathan, J. M., S. F. Smith, and N. M. Sadeh. Modeling Supply Chain Dynamics: A Multiagent Approach. *Decision Sciences*. Vol. 29, No. 3, 1998, pp. 607-632.

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

Tan, M. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Proceedings of the 10th International Conference on Machine Learning*. Amherst, MA, 1993, pp. 330-337.

Thorpe, T. L. Vehicle Traffic Light Control Using SARSA. http://www.cs.colostate.edu/~anderson/pubs/thorpems.ps.gz, 1997, Accessed on December 30, 2006.

Trabia, M. B., M. S. Kaseko, and M. Ande. A Two-Stage Fuzzy Logic Controller for Traffic Signals. *Transportation Research Part C,* Vol. 7, No. 6, 1999, pp. 353-367.

McTrans Center, http://mctrans.ce.ufl.edu/featured/TRANSYT-7F/**.** Accessed on May 20, 2007.

Transportation Simulation Systems (TSS), AIMSUN User Manual, Version 4.1.4, 2003.

Vengerov, D. A Reinforcement Learning Approach to Dynamic Resource Allocation. *Engineering Applications of Artificial Intelligence*. Vol. 20, No. 3, 2007, pp. 383-390.

Vlassis, N. A Concise Introduction to Multiagent Systems and Distributed AI. http://staff.science.uva.nl/~vlassis/cimasdai/cimasdai.pdf, 2003, Accessed on December 30, 2006.

Webster, F. V. Traffic Signal Settings. Road Research Technical Paper. No. 39. Department of Scientific and Industrial Research, Road Research laboratory. London, U.K, 1958.

Yu, X. H. and W. W. Recker. Stochastic Adaptive Control Model for Traffic Signal Systems. *Transportation Research Part C*. Vol. 14, No. 4, 2006, pp. 263-282.

Zhang, L., H. Li, and P. D. Prevedouros. Signal Control for Oversaturated Intersections Using Fuzzy Logic. In *Proceedings (CD-ROM) of the 84th Transportation Research Board Annual Meeting*. Washington D.C., 2005.

Zhang, Y. Optimal Traffic Control for A Freeway Corridor under Incident Conditions. Ph.D. Dissertation. Department of Civil Engineering, Virginia Polytechnic Institute and State University, 1996.

Zhang, Y., and Y. Xie. Comparison of PASSER V, Synchro, and TRANSYT-7F for Arterial Signal Timing Based on CORSIM Simulation. In *Proceedings of the 9th International Conference on Applications of Advanced Technology in Transportation*. Chicago, Illinois. 2006, pp. 479-484.