The Report Committee for Hao Liu
Certifies that this is the approved version of the following Report:

Traffic Signal Control Using Queueing Theory

APPROVED BY
SUPERVISING COMMITTEE:

____
John Hasenbein, Supervisor

____
Randy Machemehl, Co-Supervisor

Traffic Signal Control Using Queueing Theory

by

Hao Liu

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2018

Dedication

I would like to dedicate this work to my parents, Sitai and Xiaoli, and my wife Yifei.

Acknowledgements

First of all, I would like to acknowledge my supervisor, Dr. Hasenbein, for his guidance, knowledge sharing and help throughout this work. I would also like to thank Dr. Hanasusanto for providing important suggestions on how to solve the convex optimization models. Moreover, I am grateful to my co-supervisor, Dr. Machemehl, for his support and constant help. Last but not least, I acknowledge Dr. Strigler for helping me improve my academic writing significantly from the writing class.

Abstract

Traffic Signal Control Using Queueing Theory

Hao Liu, M.S.Stat.

The University of Texas at Austin, 2018

Supervisor: John Hasenbein

Co-Supervisor: Randy Machemehl

Traffic signal control has drawn considerable attention in the literatures thanks to its ability to improve the mobility of urban networks. Queueing models are capable of capturing performance or effectiveness of a queueing system. In this report, SOCPs (second order cone program) are proposed based on different queueing models as pre-timed signal control techniques to minimize total travel delay. Stochastic programs are developed in order to handle the uncertainties in the arrival rates. In addition, the superiority of the proposed model over Webster's model has been validated in a microscopic traffic simulation software named CORSIM.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Congestion existing in urban traffic networks is becoming increasingly significant, especially in some metropolises, because of the increase in car ownership. It has given rise to significantly negative impacts, such as reducing mobility, wasting time and fuel energy and causing air pollution. A variety of techniques used to mitigate the congestion have been proposed, such as road pricing, access restriction and dedicated bus lanes. In addition to these methods, traffic signal control may be the most cost-effective method to alleviate congestion thanks to its ability to respond to a myriad of traffic scenarios. Based on its working principle, traffic signal control techniques can be divided into two groups: pre-timed control and traffic-actuated control, also referred as adaptive signal control or real-time control. In pre-timed control, the optimal solution is a set of optimal signal timing plans for a certain number of predetermined time periods in one day. The optimal signal timing, including cycle length, splits and phase order, is derived based on historical traffic data. On the other hand, traffic-actuated signal control utilizes the prevailing traffic conditions, which can be captured by combining advanced sensors and prediction methodology, to provide the optimal signal timing in real-time. Thanks to its flexibility to accommodate varying traffic demands, adaptive signal control is considered superior to pre-timed control.

In traffic signal control research, minimization of the total vehicle control delay, an indicator representing the delay caused by the signal, is one of the most commonly used objective functions. In addition, other objective functions, such as minimization of person delay [1] and balancing queue lengths [2], have also been investigated. Different traffic delay models have been proposed in past decades. Dion et al. [3] compared the delay estimates from four classes of models: deterministic queueing models [4], [5], shock wave delay models [6]–[8], steady-state stochastic delay models [9]–[12] and time-dependent stochastic delay models [13]–[16]. They showed that at a pre-timed intersection, the estimates from these models are similar when the traffic demand is low, and the differences increase when the traffic demand approaches saturation flow.

Many traffic signal control methodologies have been published, some of which have been put into practice such as SCOOT [17] and SCATS [18]. Although these systems have been regarded as adaptive signal control, they do not execute optimization programs based on the current and forecasted traffic conditions. Instead, the idea was to choose the best timing plan from a preset timing plan set. Real adaptive signal control programs usually employ the information provided by sensors such as inductive-loop detectors and cameras to predict the traffic demand in a time horizon, then utilize optimization models to acquire the optimal signal plan in the same time horizon, which is called a rolling horizon scheme. To reduce the computational complexity, dynamic programming was used in the methods of DYPIC [19], OPAC [20], PRODYN [21]. An exhaustive comparison between these programs can be found in [22]. Head [23] proposed a model to predict the traffic flow at an intersection based on the triggering time of the upstream node detectors and upstream signal settings. The distance between the intersections and the upstream detectors determines the prediction time horizon, which is highly related to practicability. Taking the advantage of this prediction model, Mirchandani and Head [24] developed a real-time traffic signal control system, RHODES, in which a dynamic program model [25] was used to perform the optimization. Unlike most of the control methodology in which the minimum and the maximum are fixed parameters, Zheng and Recker [26] proposed a rolling horizon algorithm with a time-variant horizon length to optimize four basic parameters: phase sequence, minimum green, unit extension and maximum green time. In their model, a feedback mechanism for the optimization parameters exists to adjust signal timing to up-to-date traffic volume.

Queueing models have been used in the field of traffic signal design in past decades. Webster [9] proposed analytical expressions for queues and delays at traffic intersections assuming the arrivals are Poisson processes. After that, Miller [27] and Newell [10] proposed models to calculate the queue length at the end of green phases at an intersection with fixed signal timing. Mirchandani and Zou [28] modeled the intersection as a M/G/1 system and proposed a numerical algorithm to compute the long-run average delays and optimal phase splits under a exhaustive service policy. They claimed it is adaptive signal control while they only considered a general arrival rate instead of the prevailing traffic condition. Osorio and Bierlaire [29] proposed an analytical M/M/c/k queueing network model that can capture the evolution and the effect of congestion in this network. By employing this model in a simulation-based optimization framework, Osorio et al. [30], [31] proposed algorithms to optimize the signals at a network level.

Although traffic signal control has drawn considerable attention and queueing theory has been used to model traffic flow, to the best of our knowledge, relatively little research utilizing analytical queueing models in signal optimal research has been conducted. In this report, different queueing models based on corresponding assumptions will be used to develop signal optimization models and the solution will be tested in the traffic micro simulator called CORSIM. In addition, uncertainty in the arrival process will be considered in the model.

The rest of this report is organized as follows. In the second chapter, basic knowledge of traffic flow theory and queueing models is introduced. In the third chapter, traffic flow by queueing models, in the base of which we propose optimization models, is modeled. Then, the solutions of these models are tested in a micro traffic simulator in the fourth chapter. Finally, conclusions are drawn, and future research is discussed.

Chapter 2: Background

TRAFFIC SIGNAL DESIGN

The objective of traffic signal design is to ensure safe and efficient service for the traffic at an intersection. Phasing is the fundamental method by which a traffic signal serves the vehicles. The definition of phase varies across the literatures. In this report, the definition in [32] is used: a phase is defined as a controller timing unit associated with the control of one or more movements. Figure 1 shows a signal design pattern called the standard ring-and-barrier signal design. In each cycle, eight phases are served at this intersection. The horizontal axis indicates time, and two non-conflict phases activated at the same time are served simultaneously. For example, phase 1 and phase 2 are served at the same time, and the corresponding phase time is represented by $g1$. Phase 3 allows two non-conflict movements while phase 1 allows one. It is noteworthy that these two phases do not have to be terminated at the same time. In this report, we only consider the case in which any two non-conflict phases in the same stage start and end at the same time, and we call this period a stage. To ensure the safety of the signal switch, a new stage has to wait a period composed of a yellow time and an all-red time for clearing the vehicles in the middle of the intersection before it begins to be served. More general knowledge about traffic signal design can be found in [32].

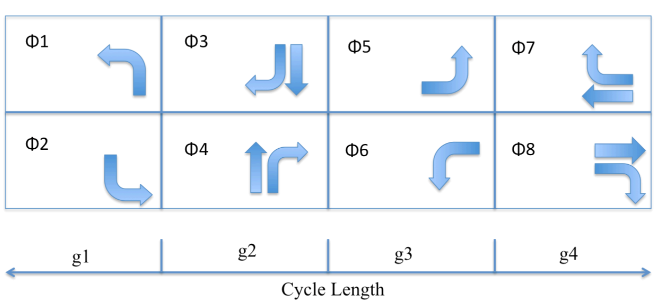

Figure 1: Standard ring-and-barrier signal design.

Queue length, number of stops and traffic delay are three essential measures of effectiveness when a traffic signal design is evaluated, and traffic delay is the most commonly used. Specifically, traffic delay can be classified as: stopped time delay, time-in-queue delay and control delay. Stopped time delay indicates the interval between the time a vehicle stops to wait for the green signal and time it begins to accelerate after the signal turns green. The time-in-queue represents the time from a vehicle joining a queue to its departure across the stop line. In general, the control delay illustrates the delay caused by a control device, such as a traffic signal and a stop sign. It is approximately equal to the sum of the time-queue-delay and delays resulting from deceleration and acceleration. The measure of effectiveness used in this report is system time rate in queueing theory, which is similar as stopped time delay, and it will be explained in the Chapter 3.

All of the components of a traffic signal, such as cycle length, phase pattern, phase sequence, phase time, lost time, even minimum and maximum green time, can be optimized in a traffic signal design program. For simplicity, only the phase time is optimized in this report.

QUEUEING THEORY MODELS

Queues are an unavoidable phenomenon in reality, and queueing theory plays a critical role in managing systems, such as factories and call centers. The customer arrival process and service time is almost stochastic, and the queueing system is modeled as a stochastic process in queueing theory. The key aspects used to describe a queue are as follows:

1. The arrival process. In queueing models, instead of giving the probability of arrival number, the distribution of interarrival time is always given. The common symbols to denote the interarrival time distribution are as follows: *M* for exponential (the arrival

process is a Poisson process), G for general, D for deterministic, E_k for Erlang with k phases.

2. The service times. The service times of customers are assumed to be i.i.d. (independent and identical distribution). The notation for the service times is the same as the arrival process.

3. The number of servers. In queueing theory, the number of servers determines the number of customers that can be served simultaneously. In traffic theory, because it represents the number of vehicles which can go through the intersection at the same time, it is equal to the number of lanes designed for the movements. The notation c is used to denote the server number.

4. The holding capacity. This is the maximal number of vehicles that a queue can accommodate. For a traffic queue, it can be approximated as the division of street length by the saturation vehicle space headway. The notation K is used to denote the queue capacity.

5. The service policy. The service policy represents the sequence in which the customers are served given their arrival order. For example, FIFO (first in first out) means the first arrived customer is served firstly when the server is available while LIFO (last in first out) policy gives the priority to the customer who is the last one joining the queue.

A queue can be fully described by these five parameters. For example, an $M/M/c/K$ represents a queue with Poisson arrival process, exponential service time, c servers and a capacity of K. The last letter is usually omitted if the capacity is assumed to be infinity. In most cases, given these parameters above, the number of customers in a queue can be modeled as stochastic process. The common performance measures are long-run average number in the system, system time, waiting time and queue length. In the following, the expression of these measures of different queue models will be derived.

M/M/1 queue

In the *M/M/1* model, we assume that the arrival process is Poisson with an arrival rate of λ, the service time is exponentially distributed with a rate of μ, and there is one server in the system and the capacity is infinity. Then, the number of the customers in the system can be modeled as a CTMC (continuous time Markov chain). The corresponding rate diagram is shown in Figure 2. This CTMC is a special case of birth-death process with uniform birth rates and uniform death rates.

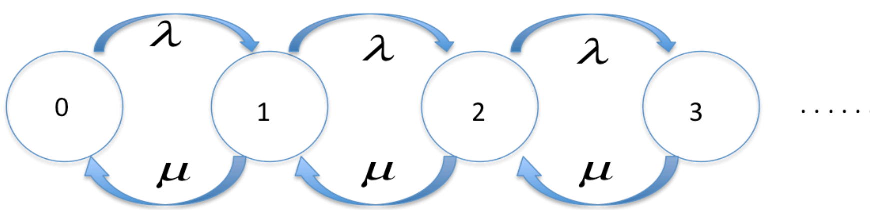

Figure 2: Rate diagram of *M/M/1* queue.

The detailed balance equations are,

$$\begin{cases} \mu p_1 = \lambda p_0 \\ \mu p_2 = \lambda p_1 \\ \dots \\ \sum_{i=0}^{\infty} p_i = 1 \end{cases} \tag{1}$$

where p_i is the stationary probability that *i* customers are in the system.

The stationary probability distribution is a geometric distribution,

$$p_i = (1-\rho)\rho^i \tag{2}$$

where $\rho = \frac{\lambda}{\mu}$ is called utilization of the server, which indicates the fraction of time the server is busy.

Performance metrics including the average number of customers in the system *L,* average system time *W,* waiting time *d,* queue length *Q* can be obtained based on (2). All of these metrics are average values in the long run.

$$L=\sum_{i=0}^{\infty} ip_i=\sum_{i=0}^{\infty} i(1-\rho)\rho^i=\frac{\rho}{1-\rho} \tag{3}$$

Little's Law: The long run average number of customers in a stationary system is equal to the long-term average effective arrival rate λ multiplied by the average time W that a customer spend in the system,

$$L=\lambda W. \tag{4}$$

According to *Little's Law,*

$$W=\frac{L}{\lambda}=\frac{1}{\mu-\lambda}. \tag{5}$$

The waiting time,

$$d=W-\frac{1}{\mu}=\frac{\rho}{\mu-\lambda}. \tag{6}$$

Again, according to *Little's Law,* the queue length, the number of customers waiting in the line, is expressed as,

$$Q=\lambda d=\frac{\rho^2}{1-\rho}. \tag{7}$$

M/G/1 queue

M/G/1 models assume the service time has a general distribution and the average service time $E(s)=\frac{1}{\mu}$. In this case, the number of customers in the system is not a CTMC. However, an embedded DTMC (discrete time Markov chain) can be modeled as,

$$Q_i=Q(t_i), i=1,2,3... \tag{8}$$

where t_i represents the departure time of the *i*th customer and $Q(t_i)$ indicates the number of customers in the system at t_i.

Following the definition (8),

$$Q_{n+1} = Q_n - 1 + \delta_n + A_{n+1} \tag{9}$$

where $\delta_n = I\{Q_n = 0\}$ and A_i 's are i.i.d random variables representing the number of arrival customers in service time S_{n+1}.

For simplicity, the process of deriving L is omitted in this report. The performance metrics of the *M/G/1* queueing model are given by the Pollaczek-Khinchine formula,

$$L = \rho + \frac{\rho^2 + \lambda^2 Var(s)}{2(1-\rho)} \tag{10}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu} + \frac{\rho + \lambda\mu Var(s)}{2(\mu - \lambda)} \tag{11}$$

$$d = W - \frac{1}{\mu} = \frac{\rho + \lambda\mu Var(s)}{2(\mu - \lambda)} \tag{12}$$

$$Q = \lambda d = \frac{\lambda\rho + \lambda^2 \mu Var(s)}{2(\mu - \lambda)} \tag{13}$$

where $Var(s)$ is the variance of service time.

M/D/1 queue

The *M/D/1* queue is a special case of *M/G/1* queue in which the service time is assumed to be constant, which implies $Var(s) = 0$. Therefore, substituting $Var(s) = 0$ into (10)-(13), the performance metrics of the M/D/1 queueing model can be obtained.

Chapter 3: Methodology

In this chapter, the traffic flow at a signalized intersection is modeled based on three queueing models introduced in chapter 2. Then, optimization models, considering both deterministic arrival rates and random arrival rates, are proposed to optimize the signal timing. In this report, the focus is only put on one isolated intersection with a phase design shown in Figure 1, which means coordination among intersections is not considered.

MODELS WITH DETERMINISTIC ARRIVAL RATE

M/M/1 queueing model

In this section, each phase in Figure 1 is modeled as an *M/M/1* queue, which implies the following assumption are made:

1. Arrival processes of these eight queues are independent Poisson processes with known rates;

2. Service times of these queues, which are queue discharge headways, are described by exponential distributions;

3. Each queue can only occupy one lane, i.e. the number of servers for each queue is equal to 1;

4. The street is long enough to accommodate infinitely long queues;

5. The moving time of a vehicle before it joins the queue is not considered;

6. The arrival rate is less than the service rate, which implies the no-spillover condition.

The notations used in this section are defined as follows:

S_i : saturation flow of queue i (vehicles per hour);

C : cycle length (s);

g_i : green time for queue i (s);

l : lost time, which is the sum of start-up lost time and clearance lost time, the sum of all-red and yellow time;

λ_i : arrival rate of queue i (vehicles per hour);

μ_i : service rate of queue i (vehicles per hour), which is equal to the reciprocal of time headway;

I : set of indices of queues (phases);

N: cardinality of set I.

The service rate of queue i is approximated as the product of the saturation flow of queue i and the fraction of its green time,

$$\mu_i = S_i \frac{g_i}{C} \tag{14}$$

The total system time rate, which is equal to the rate of total time spent in the queue, is the objective function,

$$f = \lim_{t \to \infty} \sum_{i \in I} \frac{\lambda_i t W_i}{t} = \sum_{i \in I} \lambda_i W_i \tag{15}$$

It turns out that this objective function is equivalent to the sum of average number of vehicles in the long run over all of the queues.

From (5), the signal control problem at an isolated intersection can be modeled as,

$$\min \quad f = \sum_{i \in I} \frac{\lambda_i}{\mu_i - \lambda_i} \tag{16}$$

$$s.t. \quad \mu_i = S_i \frac{g_i}{C}, \qquad \forall i \in I \tag{17.a}$$

$$\sum_{i \in I} g_i = C - l \tag{17.b}$$

$$\mu_i \geq \lambda_i, \qquad \forall i \in I \tag{17.c}$$

The decision variables are green times and the service rates of the queues. The second constraint ensures time conservation, and the third constraint represents the stability condition.

Although the objective is not a linear function of μ_i's, the Hessian matrix (18) of the objective function is a positive semidefinite diagonal matrix, which implies this function is convex in μ_i's. In addition, the constraints are linear, implying that this optimization problem is a convex problem.

$$H=\begin{bmatrix} \frac{\partial^2 f}{\partial \mu_1^2} & & & \\ & \frac{\partial^2 f}{\partial \mu_2^2} & & \\ & & \ddots & \\ & & & \frac{\partial^2 f}{\partial \mu_N^2} \end{bmatrix} = \begin{bmatrix} \frac{2\lambda_1}{(\mu_1-\lambda_1)^3} & & & \\ & \frac{2\lambda_2}{(\mu_2-\lambda_2)^3} & & \\ & & \ddots & \\ & & & \frac{2\lambda_N}{(\mu_N-\lambda_N)^3} \end{bmatrix}. \tag{18}$$

This convex program can be solved analytically through the method of Lagrange multipliers. Let the new objective function be,

$$z=\sum_{i\in I}-\frac{\lambda_i}{\mu_i-\lambda_i}+\sum_{i\in I}\alpha_i(\mu_i-S_i\frac{g_i}{C})+\beta(\sum_{i\in I}g_i-C+l) . \tag{19}$$

Set the partial derivatives to zero,

$$\begin{cases} \frac{\partial z}{\partial \mu_i}=-\frac{\lambda_i}{(\mu_i-\lambda_i)^2}+\alpha_i=0 & \forall i\in I \\ \frac{\partial z}{\partial g_i}=-\alpha_i\frac{S_i}{C}+\beta=0 & \forall i\in I \\ \frac{\partial z}{\partial \alpha_i}=-g_i\frac{S_i}{C}+\mu_i=0 & \forall i\in I \\ \frac{\partial z}{\partial \beta}=\sum_{i\in I}g_i-C+l=0 & \end{cases} . \tag{20}$$

By solving the equation set (20), the optimal solution can be obtained as,

$$\begin{cases} \beta^* = -\left(\dfrac{\sqrt{C}\sum_{i\in I}\sqrt{\dfrac{\lambda_i}{S_i}}}{C - l - C\sum_{i\in I}\dfrac{\lambda_i}{S_i}} \right)^2 \\ g_i^* = \dfrac{C}{s_i}\left(\sqrt{\dfrac{-\lambda_i S_i}{C\beta^*}} + \lambda_i \right) \\ \mu_i^* = \sqrt{\dfrac{-\lambda_i S_i}{C\beta^*}} + \lambda_i \\ \alpha_i^* = \dfrac{\lambda_i}{(\mu_i^* - \lambda_i)^2} \end{cases} \tag{21}$$

This solution has a similar form as the optimal solution to the problem of minimizing the total queue length in an open Jackson network under budget constraint.

M/G/1 queueing model

Although it is reasonable to assume that the arrival processes at an intersection are Poisson processes, the service time is not very likely to be exponentially distributed. In fact, a constant service time of 2 seconds is commonly used in traffic theory. In addition, because each queue is served under the alternation of red and green lights, the service rate changes between a deterministic value and infinity. Therefore, a general distribution for the service time, leading to M/G/1 models, is more appropriate in this case. By replacing M/M/1 models with M/G/1 models in (16), while the constraints stay the same, the optimization program is converted to,

$$\min \quad f = \sum_i \rho_i + \frac{\rho_i^2 + \lambda_i^2 Var(s_i)}{2(1-\rho_i)} \tag{22}$$

$$s.t. \quad (17)$$

where $\rho_i = \dfrac{\lambda_i}{\mu_i}$ is the server utilization of queue i, and $Var(s_i)$ is the variance of service time of queue i.

Like the objective function (16), the Hessian matrix of (22) is a diagonal matrix with elements of,

$$\frac{\partial^2 f}{\partial \mu_i^2} = \frac{1}{2}\left[\frac{2(1-\rho_i)(1+\lambda^2 \operatorname{var}(s_i))}{(1-\rho_i)^4}\left(\frac{d\rho_i}{d\mu_i}\right)^2 + \frac{\rho_i^{\,2} - 2\rho_i + 2 + \lambda^2 \operatorname{var}(s_i)}{(1-\rho_i)^2}\frac{d^2\rho_i}{d\mu_i^2}\right] \geq 0 \quad (23)$$

Therefore, this optimization problem is also convex. However, it is challenging to derive an analytical solution. By inserting additional variables τ_i's and β_i's, (22) can be transformed to,

$$\begin{aligned}
\min \quad & \sum_{i \in I} \rho_i + \tau_i + \beta_i \\
s.t. \quad & (17) \\
& 1 \geq \rho_i \geq \frac{\lambda_i}{\mu_i}, && \forall i \in I \\
& \tau_i \geq \frac{\rho_i^{\,2}}{2(1-\rho_i)}, && \forall i \in I \\
& \beta_i \geq \frac{\lambda_i^2 Var(s_i)}{2(1-\rho_i)}, && \forall i \in I \\
& \tau_i \geq 0, \;\; \beta_i \geq 0, && \forall i \in I
\end{aligned} \qquad (24)$$

Because $\dfrac{\rho_i^{\,2}}{2(1-\rho_i)}$ and $\dfrac{\lambda_i^2 Var(s_i)}{2(1-\rho_i)}$ are increasing with ρ_i, and the objective function is minimized, the equalities $\tau_i = \dfrac{\rho_i^{\,2}}{2(1-\rho_i)}$ and $\beta_i = \dfrac{\lambda_i^2 Var(s_i)}{2(1-\rho_i)}$ hold in the optimal solution, implying the equivalence between models (22) and (24).

(24) is equivalent to a SDP (semidefinite program) as,

$$
\begin{aligned}
\min \quad & \sum_{i\in I} \rho_i + \tau_i + \beta_i \\
s.t. \quad & (17) \\
& \begin{bmatrix} \rho_i & \sqrt{\lambda_i} \\ \sqrt{\lambda_i} & \mu_i \end{bmatrix} \succ 0, \qquad \forall i \in I \qquad (25) \\
& \begin{bmatrix} \tau_i & \rho_i \\ \rho_i & 2(1-\rho_i) \end{bmatrix} \succ 0, \qquad \forall i \in I \\
& \begin{bmatrix} \beta_i & \lambda_i \sqrt{Var(S_i)} \\ \lambda_i \sqrt{Var(S_i)} & 2(1-\rho_i) \end{bmatrix} \succ 0, \quad \forall i \in I
\end{aligned}
$$

where "$A \succ B$" means $A - B$ is a semidefinite positive matrix. Therefore,

$$
\begin{bmatrix} \rho_i & \sqrt{\lambda_i} \\ \sqrt{\lambda_i} & \mu_i \end{bmatrix} \succ 0 \Leftrightarrow \begin{vmatrix} \rho_i & \sqrt{\lambda_i} \\ \sqrt{\lambda_i} & \mu_i \end{vmatrix} \Leftrightarrow \rho_i \geq \frac{\lambda_i}{\mu_i} \qquad (26)
$$

.

Because the product of the eigenvalues of this 2×2 matrix is equal to the determinant of the matrix, and the sum of the eigenvalues is equal to the trace $\rho_i + \lambda_i \geq 0$, the semidefinite condition is equivalent to the determinant being positive, which indicates the first equivalence in (26). Therefore, (24) is equivalent to (26).

Further, (24) can be converted to an SOCP (second-order cone program) by transforming the hyperbolic constraints in (24) to hyperbolic constraints given the following equivalence, which will not be proved in this report for simplicity,

$$
x^T x \leq yz, \quad y,z \geq 0 \quad \Leftrightarrow \quad \left\| \begin{bmatrix} 2x \\ y-z \end{bmatrix} \right\|_2 \leq y+z, \quad y,z \geq 0 \qquad (27)
$$

.

Based on (27), the model (24) is equivalent to,

$$
\begin{aligned}
\min \quad & \sum_{i\in I} \rho_i + \tau_i + \beta_i \\
s.t. \quad & (17) \\
& \left\| \begin{matrix} 2\sqrt{\lambda_i} \\ \rho_i - \mu_i \end{matrix} \right\|_2 \le \rho_i + \mu_i, \qquad \forall i \in I \\
& \left\| \begin{matrix} 2\rho_i \\ 2\tau_i - 1 + \rho_i \end{matrix} \right\|_2 \le 2\tau_i + 1 - \rho_i, \quad \forall i \in I \\
& \left\| \begin{matrix} 2\lambda\sigma(S_i) \\ 2\beta_i - 1 + \rho_i \end{matrix} \right\|_2 \le 2\beta_i + 1 - \rho_i, \quad \forall i \in I \\
& \tau_i \ge 0, \quad \beta_i \ge 0, \qquad \forall i \in I
\end{aligned} \tag{28}
$$

where $\sigma(S_i)$ is the standard deviation of the service time. The solver called MOSEK will be employed to solve this SOCP in Matlab.

M/D/1 queueing model

In traffic flow theory, it is reasonable to assume the discharge headway is a constant, which requires an M/D/1 queueing model instead of an M/G/1 model. The only change needed to be made in the M/D/1 queueing model from the M/G/1 model is to remove the variation term $Var(S_i)$'s. The solution between these models will be compared in the next chapter, but the model itself is not shown for brevity.

MODELS WITH RANDOM ARRIVAL RATES

The previous section shows that the optimal signal timing is highly dependent on the arrival rates of the upstream queues. However, in most cases, the forecast has uncertainty, and this uncertainty may have a significant impact on the traffic. In this section, stochastic programming models are proposed to investigate the effect of uncertainty in the arrival rates. All of the models in this section are based on the M/M/1 model.

In this section, the expected value of total queue lengths in the long term at an intersection is chosen as the risk measure and the distribution of arrival rates are assumed known.

Discrete arrival rate

First, let us consider the case in which the arrival rates are discrete and the distribution is,

$$p_{ij} = P(\lambda_i = j), \quad \forall i \in I, j \in \Lambda_i \quad \text{and} \quad \sum_{j \in \Lambda_i} p_{ij} = 1, \quad \forall i \in I \tag{29}$$

where Λ_i is the support of λ_i. Then, the stochastic optimization problem can be modeled as,

$$\min \quad \sum_{i \in I} \sum_{j \in \Lambda_i} p_{ij} \left(\frac{\lambda_{ij}}{\mu_i - \lambda_{ij}} \right) \tag{30}$$

$$s.t. \quad \mu_i = S_i \frac{g_i}{C}, \qquad \forall i \in I \tag{31.a}$$

$$\sum_{i \in I} g_i = C - l \tag{31.b}$$

$$\mu_i \geq \lambda_{ij}, \qquad \forall i \in I, j \in \Lambda_i \tag{31.c}$$

This model can be transformed to an SOCP expressed as,

$$\min \quad \sum_{i \in I} \sum_{j \in \Lambda_i} p_{ij} \beta_{ij}$$

$$s.t. \quad \left\| \begin{matrix} 2\sqrt{\lambda_{ij}} \\ \beta_{ij} - \mu_i + \lambda_{ij} \end{matrix} \right\| \leq \beta_{ij} + \mu_i - \lambda_{ij} \quad \forall i \in I, j \in \Lambda_i \tag{32}$$

(31)

Continuous arrival rate

Assume the arrival rates are continuously distributed, and the joint distribution is given, then the signal control problem can be modeled as,

$$\begin{aligned} \min \quad & \sum_{i \in I} \int_{\Xi} f(\vec{\lambda})(\frac{\lambda_i}{\mu_i - \lambda_i}) d\vec{\lambda} \\ s.t. \quad & (31) \end{aligned} \tag{33}$$

where $\vec{\lambda}$ is the arrival rate vector, Ξ is the support set, and $f(\vec{\lambda})$ is the joint distribution of the arrival rates. In most cases, it is challenging to solve the stochastic program with a continuously distributed random coefficient. However, Monte Carlo methods can be utilized to obtain an approximate solution as follows. First, a large number, *N,* of arrival rates are sampled from the given distribution; then, (33) can be approximated as,

$$\begin{aligned} \min \quad & \frac{1}{N} \sum_{k \in [N]} \sum_{i \in I} \frac{\lambda_{ik}}{\mu_i - \lambda_{ik}} \\ s.t. \quad & \mu_i = S_i \frac{g_i}{C}, & \forall i \in I \\ & \sum_{i \in I} g_i = C - l \\ & \mu_i \geq \lambda_{ik}, & \forall i \in I, k \in [N] \end{aligned} \tag{34}$$

This model can also be solved by converting it to an SOCP.

Chapter 4 Results

CORSIM is a widely used microscopic traffic simulator, in which the behavior of every driver-vehicle unit is simulated, and it is an appropriate option to test the performance of the models proposed in Chapter 3.

Webster [9] proposed a model to determine the green splits,

$$g_i = \frac{\alpha_i}{\sum_{i \in I} \alpha_i}(C - l) \tag{35}$$

where α_i is the critical saturation level of stage i, which is the maximum traffic λ / S ratio among the phases belonging to stage i. For example, in Figure 1, $\alpha_1 = \max\{\lambda_1 / S_1,\ \lambda_2 / S_2\}$. The Highway Capacity Manual [4] proposed another model allocating the green time to ensure all stages have the same critical v/c ratio. Although these two methods have different interpretations, they result in the same solutions.

The performance, in terms of travel delay at an intersection, between Webster's model and the proposed models in Chapter 3 will be compared through CORSIM.

CORSIM MODEL CALIBRATION

Figure 3 shows the network used in this chapter. Node 2 is the signalized intersection, and it has four upstream links. Each of the links consists of one only-turn-left lane and another lane allowing right turns and through movements. The phase set shown in Figure 1 is used.

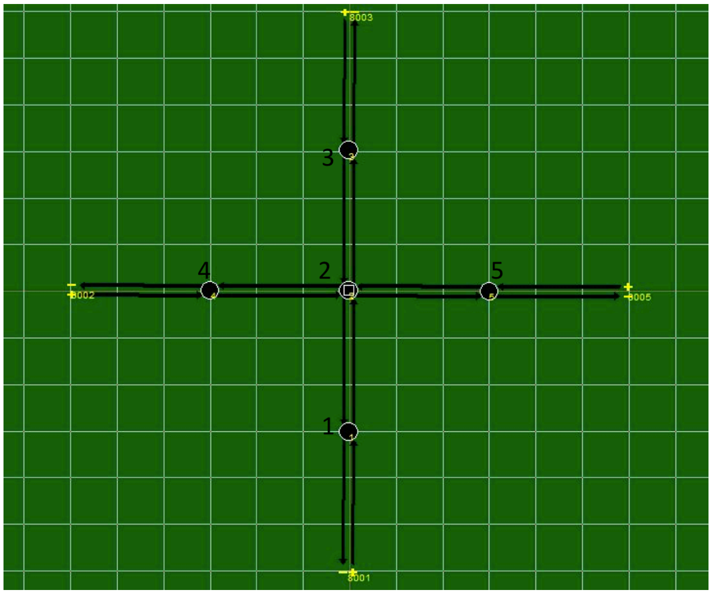

Figure 3: Network schematic.

A preset cycle length of 90 seconds is chosen, and the yellow time and all red time during each signal change are both equal to 1 second. The mean startup delay in each link is 0, so the effective green time in each cycle is 82 seconds. The discharge headway is 2 seconds. In addition, by running preliminary cases with oversaturated volume, the saturation flows of left-turn movement and right-turn and through movement are obtained as 1487 vph (number of vehicles per hour) and 1600 vph, separately. The total simulation time is 2 hours.

COMPARISON

In this section, the signal timing from both Webster's models and the proposed models corresponding to various arrival rate scenarios are acquired, and the total travel delay at node 2 is employed as the performance measure. The scenarios of arrival rates are listed in Table 1.

Scenario #	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase 7	Phase 8
1	100	100	200	200	300	300	400	400
2	100	450	100	450	250	250	250	250
3	50	250	50	250	250	250	250	250
4	50	250	50	250	200	200	200	200

Table 1: Arrival rate (vph) scenarios.

Although 8 phases exist in the signal setting shown in Figure 1, Webster's model only considers one of the movements, which is the critical movement, for each stage. On the other hand, the models proposed in Chapter 3 involve all of the queues, which should be beneficial for signal control. It is noteworthy that, in this section, the proposed models with uncertainties in the arrival rates are not tested because it is challenging to do Monte Carlo simulation in CORSIM. However, it could be realized through an external application, known as CORSIM run-time extension (RTE), that can interface directly with CORSIM. Table 2 shows the comparison of the green time splits between these models. The M/G/1 queueing model is not tested because it is challenging to consider the uncertainty of service time in CORSIM. Table 2 indicates that the M/M/1 model and M/D/1 model lead to the same optimal signal timing under all of the scenarios. It is noteworthy that the optimal solutions of these two models are close to each other, but they are not exactly the same. However, because CORSIM only accepts integer number of green times, the values in Table 2 are rounded from the originally optimal solutions. Although the objective functions of these two models are different, the constraints are the same. The difference between these two objective functions may not have a significant effect on the optimal solution.

	Scenario 1				Scenario 2			
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 1	Stage 2	Stage 3	Stage 4
Webster	8	15	25	34	27	26	15	14
M/M/1	10	17	25	30	27	26	15	14
M/D/1	10	17	25	30	27	26	15	14
	Scenario 3				Scenario 4			
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 1	Stage 2	Stage 3	Stage 4
Webster	21	20	21	20	24	22	19	17
M/M/1	20	19	22	21	22	21	20	19
M/D/1	20	19	22	21	22	21	20	19

Table 2: Green time splits (seconds).

Table 3 shows the total travel delay corresponding to the signals in table 2. Because CORSIM assigns random seeds to each simulation, 10 runs are executed for each signal scenario, and the average value of total travel delay of each scenario are shown in Table 3. In addition, paired t-test is used to demonstrate if the difference between total travel delay from these models is statistically significant.

Because the M/M/1 model and M/D/1 model lead to the same signal timing, we do not differentiate between these two models, and they are referred to as the proposed model in the following. In the first scenario, the arrival rates of the two queues belonging to the same stage are balanced, which means the difference between them is small enough to be ignored. In this case, Webster's models and the proposed model obtain very similar result. Although the proposed model reduces the total travel delay by 0.72%, this improvement is not significant. In the second scenario, they result in the same signal timing. The overall arrival rate in this scenario is approximately equal to the overall capacity; the constraints

(17) force the proposed model assign green time based on the ratio of critical saturation level, which is the basic idea of Webster's model. Therefore, the comparison of scenario 2 is not displayed in Table 3. In the last two scenarios, the overall traffic at this intersection is under saturated, and the arrival rates are unbalanced. Because the proposed model considers all of the queues instead of only critical movements, its signal has better performance than Webster's. In addition, the superiority becomes more significant as the gap between arrival rates within the same stage becomes larger. The average total travel delays of 10 runs decrease by 1.27% and 2.21% in the third and fourth scenario, respectively, and the difference in both cases are statistically significant.

	Scenario 1		Scenario 3		Scenario 4	
Run number	Webster	M/M/1	Webster	M/M/1	Webster	M/M/1
1	2214.999	2228.444	1844.999	1807.761	1584.437	1545.703
2	2157.976	2139.926	1875.224	1851.501	1543.274	1508.049
3	2280.037	2299.473	1988.212	1967.244	1687.787	1632.241
4	2305.139	2311.303	1850.603	1829.366	1622.747	1602.856
5	2214.841	2135.963	1840.319	1805.057	1538.202	1505.074
6	2242.86	2215.965	1838.638	1836.232	1657.556	1595.244
7	2201.7	2177.81	1854.431	1813.702	1562.244	1532.208
8	2321.544	2278.658	1932.402	1883.756	1632.877	1582.416
9	2189.262	2224.607	1818.72	1821.469	1582.436	1551.334
10	2247.271	2202.265	1892.322	1882.412	1568.969	1571.831
Reduction(%)	0.72		1.27		2.21	
p-value	0.1779		0.0017		0.0002	

Table 3: Total travel delay (minutes) comparison by paired t-test.

Chapter 5: Conclusion

In this report, we proposed pre-timed optimization models using queueing theory to optimize the signal setting at an isolated intersection. All of these models can be converted to SOCPs (second-order cone program) which can be solved efficiently. Three optimization models are derived based on an M/M/1 model, M/G/1 model and M/D/1 model, respectively with deterministic arrival rates. In addition, the performance of these models is compared to Webster's method by simulation using CORSIM. It shows that the total travel delay from the proposed models is approximately equal to that from Webster's when the arrival phase rates within the same stage are balanced or the overall arrival rate reaches the overall capacity. However, the proposed models demonstrate their superiority over Webster's model when the balance is violated, and this superiority increases significantly with the gap of the arrival rates. In addition, stochastic programming models are proposed to consider the uncertainties in the arrival rates. Although it is challenging to simulate, the importance of taking the uncertainty into consideration cannot be underestimated.

We are interested in the following future research directions:

(1). Adaptive signal control has inspired considerable interest in transportation research because of its ability to adjust the signal timely based on the prevailing traffic conditions. Extending the proposed models to adaptive signal control models is necessary to improve their utility.

(2). On the other hand, intersections in a network are correlated with each other, and it is essential to develop a network level model to investigate the effect of their correlation on the signal control.

(3). Moreover, the capacities of queues are assumed to be infinity so that we do not need to worry about the blocking effect. This assumption limits the ability of the proposed models when the traffic is oversaturated. Therefore, queue capacities should be introduced into the future models.

References

[1] E. Christofa, K. Ampountolas, and A. Skabardonis, “Arterial traffic signal optimization: A person-based approach,” *Transp. Res. Part C Emerg. Technol.*, vol. 66, pp. 27–47, 2016.

[2] K. Aboudolas, M. Papageorgiou, A. Kouvelas, and E. Kosmatopoulos, “A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks,” *Transp. Res. Part C Emerg. Technol.*, vol. 18, no. 5, pp. 680–694, 2010.

[3] F. Dion, H. Rakha, and Y.-S. Kang, “Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections,” *Transp. Res. Part B Methodol.*, vol. 38, no. 2, pp. 99–122, Feb. 2004.

[4] TRB, “Highway Capacity Manual,” in *Special Report 209, third edtion*, 1994.

[5] TRB, “Highway Capacity Manual,” in *Special Report 209, third edition*, 1997.

[6] M. J. Lighthill and G. B. Whitham, “On kinematic waves. II. A theory of traffic flow on long crowded roads,” in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1955, vol. 229, no. 1178, pp. 317–345.

[7] P. I. Richards, “Shock waves on the highway,” *Oper. Res.*, vol. 4, no. 1, pp. 42–51, 1956.

[8] P. G. Michalopoulos and V. B. Pisharody, “Derivation of delays based on improved macroscopic traffic models,” *Transp. Res. Part B Methodol.*, vol. 15, no. 5, pp. 299–317, Oct. 1981.

[9] F. Webster, “Traffic signal settings, road research technical paper no. 39,” *Road Res. Lab.*, 1958.

[10] G. F. Newell, “Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light,” *SIAM Rev.*, vol. 7, no. 2, pp. 223–

240, Apr. 1965.

[11] G. F. Newell, "Queues for a Fixed-Cycle Traffic Light," *Ann. Math. Stat.*, vol. 31, no. 3, pp. 589–597, Sep. 1960.

[12] D. Heidemann, "Queue length and delay distributions at traffic signals," *Transp. Res. Part B Methodol.*, vol. 28, no. 5, pp. 377–389, Oct. 1994.

[13] R. Akcelik and Australian Road Research Board., *Traffic signals : capacity and timing analysis*. Australian Road Research Board, 1981.

[14] R. Akçelik and N. M. Rouphail, "Estimation of delays at traffic signals for variable demand conditions," *Transp. Res. Part B Methodol.*, vol. 27, no. 2, pp. 109–131, Apr. 1993.

[15] R. Akcelik, "The Highway Capacity Manual Delay Formula for Signalized Intersections."

[16] D. Fambro and N. Rouphail, "Generalized Delay Model for Signalized Intersections and Arterial Streets," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1572, pp. 112–121, Jan. 1997.

[17] P. B. Hunt, D. I. Robertson, R. D. Bretherton, and R. I. Winton, *SCOOT-A Traffic Responsive Method of Coordinating Signals*. Crowthorne, England: Transport and Road Research Laboratory, 1981.

[18] A. G. Sims and K. W. Dobinson, "The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits," *IEEE Trans. Veh. Technol.*, vol. 29, no. 2, pp. 130–137, May 1980.

[19] D. I. Robertson and R. D. Bretherton, "Optimum Control of an Intersection for Any Known Sequence of Vehicular Arrivals," *Proc. 2nd IFAC/IFIP/IFORS Symp. Traffic Control Transp. Syst.*, 1974.

[20] N. H. Gartner, "OPAC: A demand-responsive strategy for traffic signal control," *Transp. Res. Rec.*, no. 906, pp. 75–81, 1983.

[21] J. J. Henry, J. L. Farges, and J. Tuffal, "THE PRODYN REAL TIME TRAFFIC ALGORITHM," in *Control in Transportation Systems*,

Elsevier, 1984, pp. 305–310.

[22] C. Cai, C. K. Wong, and B. G. Heydecker, “Adaptive traffic signal control using approximate dynamic programming,” *Transp. Res. Part C Emerg. Technol.*, vol. 17, no. 5, pp. 456–474, 2009.

[23] K. L. Head, “Event-based short-term traffic flow prediction model,” *Transp. Res. Rec.*, no. 1510, pp. 45–52, 1995.

[24] P. Mirchandani, “A real-time traffic signal control system: architecture, algorithms, and analysis,” *Tramsportation Res. Part C Emerg. Technol.*, vol. 9, no. 6, pp. 415–432, 2001.

[25] S. Sen and K. L. Head, “Controlled Optimization of Phases at an Intersection,” *Transp. Sci.*, vol. 31, no. 1, pp. 5–17, 1997.

[26] X. Zheng and W. Recker, “An adaptive control algorithm for traffic-actuated signalized networks,” *Transp. Res. Part C Emerg. Technol.*, vol. 30, pp. 93–115, 2010.

[27] A. J. Miller, “Settings for Fixed-Cycle Traffic Signals,” *J. Oper. Res. Soc.*, vol. 14, no. 4, pp. 373–386, Dec. 1963.

[28] P. B. Mirchandani and N. Zou, “Queuing Models for Analysis of Traffic Adaptive Signal Control,” *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 50–59, Mar. 2007.

[29] C. Osorio and M. Bierlaire, “An analytic finite capacity queueing network model capturing the propagation of congestion and blocking,” *Eur. J. Oper. Res.*, vol. 196, no. 3, pp. 996–1007, 2009.

[30] C. Osorio and M. Bierlaire, “A Simulation-Based Optimization Framework for Urban Transportation Problems,” *Oper. Res.*, vol. 61, no. 6, pp. 1333–1345, 2013.

[31] O. Carolina and L. Chong, “A Computationally Efficient Simulation-Based Optimization Algorithm for Large-Scale Urban Transportation Problems,” *Transp. Sci.*, vol. 49, no. February, pp. 623–636, 2015.

[32] T. Urbanik *et al.*, *Signal Timing Manual - Second Edition*.

Washington, D.C.: Transportation Research Board, 2015.