Design and Operation of a Last Mile Transportation System

By

Hai Wang

B.S., Civil Engineering, Tsinghua University (2009)

S.M., Transportation, Massachusetts Institute of Technology (2012)

S.M., Operations Research, Massachusetts Institute of Technology (2012)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Signature of Author.....
Sloan School of Management
May 15th, 2015

Certified by .....
Amedeo R. Odoni
Professor of Aeronautics and Astronautics and of Civil and Environmental Engineering
Thesis Supervisor

Certified by .....
Cynthia Barnhart
Chancellor, Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Professor of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

Design and Operation of a Last Mile Transportation System

By

Hai Wang

Submitted to the Sloan School of Management

on May 15, 2015, in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in Operations Research

Abstract

The Last Mile Problem refers to the provision of travel service from the nearest public transportation node to a home or office. Last Mile Transportation Systems (LMTS) are critical extensions to traditional public transit systems. We study the LMTS from three perspectives.

The first part of this thesis focuses on the design of a LMTS. We study the supply side of LMTS in a stochastic setting, with batch demands resulting from the arrival of groups of passengers at rail stations or bus stops who request last-mile service. Closed-form bounds and approximations are derived for the performance of LMTS as a function of the fundamental design parameters of such systems. It is shown that a particular strict upper bound and an approximate upper bound perform consistently and remarkably well. These expressions can therefore be used for the preliminary planning and design of Last Mile Transportation Systems.

The second part of the thesis studies operating strategies for LMTS. Routes and schedules are determined for a multi-vehicle fleet of delivery vehicles with the objective of minimizing the waiting time and riding time of passengers. A myopic operating strategy is introduced first. Two more advanced operating strategies are then described, one based on a metaheuristic using tabu search and the other using an exact Mixed Integer Programming model, which is solved approximately in two stages. It is shown that all three operating strategies greatly outperform the naïve strategy of fixed routes and fixed vehicle dispatching schedules.

The third part presents a new perspective to the study of passenger utility functions in a LMTS. The unknown parameters of a passenger utility function are treated as unobserved events, and the characteristics of the transportation trips made by the passengers are treated as observed outcomes. We propose a method to identify the probability measures of the events given observations of the frequencies of outcomes by introducing the concept and assumptions of the Core Determining Class. We introduce a

combinatorial algorithm in which the noise in the observations data is ignored and a general procedure in which data noise is taken into consideration.

Thesis Supervisor: Amedeo R. Odoni

Title: Professor of Aeronautics and Astronautics and of Civil and Environmental Engineering

Thesis Supervisor: Cynthia Barnhart

Title: Chancellor, Professor of Civil and Environmental Engineering

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Amedeo Odoni for everything he has done for me. In the past six years, he has been a constant source of knowledge, guidance, support and encouragement. I have benefited significantly from his immense knowledge, inspiration and patience. He always provides strong support, sound advice and tremendous guidance. I would not have come this far without him. He is the greatest mentor I have ever met or even heard of. His kindness and caring have made my journey at MIT a very pleasant one. He is an exceptional role model for me and it is my great honor to be one of the last group of PhD students working with him.

I would like to express my sincere appreciation to my advisor Prof. Cynthia Barnhart. Throughout the past three years, she has provided me with sound advice, insightful suggestions and lots of new ideas. Her broad perspective, high professional standards, and emphasis on the balance and combination of theory and practice have had a huge influence on me. This thesis would not have been possible without her. I am extremely fortunate to have Cindy as my advisor and I always feel I am very lucky to be working with her.

I would like to thank Prof. Patrick Jaillet, who also served on my thesis committee, for his valuable feedback, insightful suggestions and comments at various stages of my dissertation research. More importantly, he has been providing kind help ever since my early days as an MST student in MIT.

Additionally, this thesis has benefited from discussions with a number of members of MIT's faculty: Prof. Dimitris Bertsimas, Prof. Andreas Schulz, Prof. Juan Pablo, and Prof. Jinhua Zhao. Chapter 4 of this thesis is a result of collaborating with Mr. Ye Luo from the Department of Economics of MIT. I also thank Prof. Richard Larson and Prof. Arnold Barnett. It was an amazing experience to work with them as a teaching assistant for the course of Urban Operations Research. I learnt much from their professionalism.

I am very grateful to Maria Marangiello, Laura Rose, and Andrew Carvalho for their administrative assistance. This dissertation has been supported by the MIT Chyn Duog

Shish Memorial Fellowship and the Future Urban Mobility program of the Singapore-MIT Alliance for Research and Technology (SMART).

Many thanks go to my friends at ORC and MIT. I have been extremely lucky to be surrounded by a group of peers who are individually brilliant, yet exceptionally supportive of each other. The friends here made my stay at MIT much more enjoyable and memorable. I would also like to thank all my Chinese friends in the Boston and Cambridge area, who created a home-like atmosphere for me in my best age.

Last, but not least, I would like to dedicate this thesis to my parents – my father Zhengqing Wang and my mother Suqing Xu. They provide unconditional love and endless support all the time. They are my everything and I owe my deepest gratitude to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Literature Review

The Last Mile Problem (LMP) refers to the provision of travel service from a public transportation node to home or workplace ("last mile") or vice versa ("first mile"). This public transportation node could be the nearest rapid transit rail station or a stop of a scheduled bus line. The unavailability of this type of service is one of the main deterrents to the use of public transport in urban areas, especially for certain demographic groups, such as schoolchildren, seniors and people with certain physical handicaps. Currently, the default solutions to the LMP are walking, riding a bike, taking a taxi, or driving a private vehicle.

A conceptual Last Mile Transportation System (LMTS) is described schematically in Figure 1.1, which shows an urban area surrounding a public-transit rail station, where trains arrive and discharge passengers. The passengers' final destinations (homes, apartments, offices and workplaces) are distributed in the area. A fleet of vehicles transports these passengers to their eventual destinations (or locations which are very close to their eventual destinations) and empty vehicles return to the station to pick up waiting passengers or newly arriving ones. We describe the specific setting of LMTS in more detail in Chapter 2 and Chapter 3, in the context of system design and system operation, respectively.

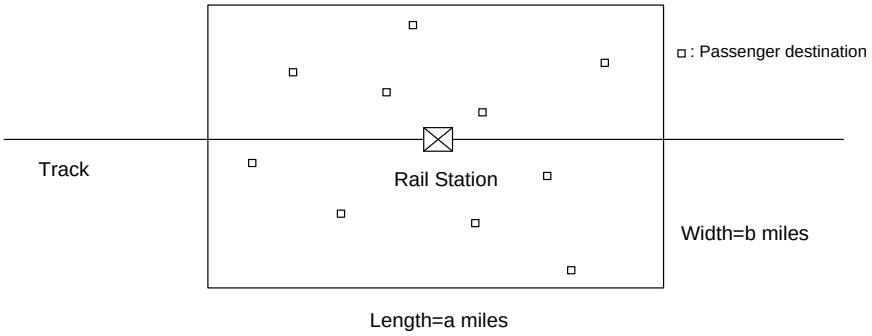


Figure 1.1 Schematic of a Last Mile Transportation System (LMTS)

Many issues must be addressed when designing and operating a LMTS. On the supply side, it is essential to deal with difficult questions concerning the stochastic aspects of the system; additionally, operating strategies addressing vehicle routing and scheduling, and passenger service assignment are needed to support system operations. The demand side requires an understanding and estimation of the potential LMTS loads as a function of demographic characteristics, nature of trip, level of service, time-of-day, cost, etc. The LMTS loads can be partially estimated through a study of passenger utility functions.

An extensive literature in this general area has generated various models for a number of application contexts related to the LMP with early work dating back to the 1960s. We mention here only a few that are among the most influential in the field or especially relevant to the approach we have adopted. Many references specific to the individual topics studied are provided, respectively, in Chapters 2, 3, and 4.

Several influential papers in the Operations Research literature have addressed problems with significant similarities to the LMP. The Dynamic Traveling Repairman Problem (DTRP) was introduced in two papers by Bertsimas and Van Ryzin. They consider the DTRP in the cases of a single-vehicle "fleet" (1991) and of multiple vehicles (1993). The Dynamic Pick-up and Delivery Problem (DPDP) was studied by Swihart and Papastavrou (1999), who derived bounds on the performance of several DPDP variants for light and heavy traffic. The Car Pooling Problem (CPP), introduced by Baldacci et al. (2004), also has features similar to the LMP – or, more exactly, to the First

Mile Problem. The paper presents both exact and heuristic methods for solving the CPP based on integer programming formulations.

Many more recent papers deal with last mile supply chains and freight last mile systems in the age of booming e-commerce. The related literature has been growing rapidly in the last 15 years, including Lee and Whang (2001), Punakivi et al. (2001), Esper et al. (2003), Balcik et al. (2008), Boyer et al. (2009), and Song et al. (2009). For example, Balcik et al. (2008) consider a vehicle-based last mile distribution system geared to the needs of humanitarian relief chains. They propose a mixed integer programming model to determine delivery schedules for vehicles and to allocate resources equitably in the face of certain types of constraints. Boyer et al. (2009) examine the effects of two factors, customer density and delivery window length, on the overall efficiency of "last mile routes" for package deliveries. As for passenger last mile systems, some case studies provide analysis of LMTS in different contexts, such as the study of a bicycle-sharing program for an LMTS in Beijing by Liu et al. (2012).

Personal rapid transit (PRT), which refers to a variety of transportation systems with characteristics similar in some ways to the last mile transportation system studied in this thesis, has also attracted significant attention in recent years. Papers considering PRT systems from a range of perspectives (all different from those presented here) include those by Anderson (1998), Bly and Teychenne (2005), Lees-Miller et al. (2009, 2010), Berger et al. (2011), and Mueller and Sgouridis (2011).

Finally, a large number of papers have dealt with the Dial-a-Ride Problem (DARP) and related variations – see, e.g., Jaw et al. (1986), Lei et al. (2012). A good critical review of the DARP literature by Cordeau and Laporte (2007) underlines, among other points, the fact that this body of work does not address well some of the queueing aspects of the subject systems – a deficiency that Chapter 2 tries to remedy.

It should also be noted that similarities exist between the LMP and various queuing, dispatching, routing, scheduling, service assignment, and resource allocation problems arising in entirely different contexts such as the design of manufacturing systems, the operation of elevator banks, and the scheduling of school-bus systems. The major

difference between the LMP and the more "traditional" problems is that, in the LMP, passengers (service requests) arrive in (possibly large) batches, not singly. This difference, however, makes the analysis of LMP much more difficult analytically than the analysis of these other problems.

1.2 Thesis Outline and Contributions

The main body of this thesis is organized as follows.

In Chapter 2, we study the supply side of the last mile transportation system in a stochastic setting, with stochastic batch demands resulting from the arrival of groups of passengers who request last-mile service at urban rail stations or bus stops. The main contribution of this chapter is the derivation of several closed-form expressions that approximate the principal performance characteristics of last mile transportations systems as a function of the fundamental design parameters of such systems. An initial set of results is obtained for the case in which a fleet of vehicles of unit capacity provides the Last Mile service and each delivery route consists of a simple round-trip between the rail station or bus stop and a single passenger's destination. These results are then extended to the general case in which the capacity of a vehicle is a small number (up to 20). It is shown through comparisons with simulation results that the approximations perform consistently well for a broad and realistic range of input values and conditions. These expressions can therefore be used for the preliminary planning and design of last mile transportation systems, especially for determining approximately resource requirements, such as the number of vehicles/servers needed to achieve some pre-specified level of service, as measured by the expected waiting time until a passenger is picked up from the station or delivered to her destination.

In Chapter 3, we consider the operation of a last mile transportation system with batch demands. The main contribution of this chapter is the development of operating strategies and algorithms for the design of passenger delivery schedules and vehicle routes for a multi-vehicle fleet of delivery vehicles with the objective of minimizing the waiting time and riding time of passengers. A myopic operating strategy is introduced first, for the

case in which the last mile demand from each group of arriving passengers is revealed sequentially. Two more advanced operating strategies are then described in detail, one based on a metaheuristic using tabu search and the other using an exact Mixed Integer Programming (MIP) model, which is solved approximately in two stages. These operating strategies are implemented in a number of computational experiments with a broad and realistic range of inputs values and conditions. It is shown that: the myopic strategy performs well for certain ranges of the input values and poorly for others; the tabu search metaheuristic provides solutions of good quality in a reasonably short computational time; and the MIP model provides the best solutions, but has greatly increased computational requirements. Thus the best approach to the routing and scheduling of the LMTS fleet depends on the context and the user's needs.

In Chapter 4, we present a novel approach to the study of the passenger utility function in a last mile transportation system. The passenger utility function provides critical information to LMTS service providers when it comes to understanding and estimating passenger demand and designing and operating their systems. From our new perspective, which is significantly different from existing ones, the unknown parameters in the passenger utility function are treated as unobserved events (defined in detail in Chapter 4), and the specific characteristics of transportation trips, such as passenger waiting time, vehicle travel time and monetary travel cost that can be collected in the real data, are considered as observed outcomes (defined in detail in Chapter 4). In this chapter, given a bipartite graph representing the relations between events and outcomes, we develop a combinatorial algorithm to identify irredundant linear inequalities to bound the probability measures of events using observations of the frequencies of the outcomes. We then extend the irredundant linear inequality identification problem and propose a general inequality selection procedure in which we take the data noise of the outcome observations into consideration. The primary model, which is similar to the Dantzig Selector in the l_1-regularization problem, is a linear programming formulation motivated by Farkas' lemma. It measures the importance of each linear inequality among an entire set of numerous inequalities under sparse assumptions. The novel approach presented in Chapter 4 calibrates the possible set of the unknown parameters in the passenger utility

function and is helpful in understanding and estimating passenger demand for last mile transportation systems.

Finally, in Chapter 5 we conclude with a summary of the thesis and discuss several directions for future research.

Chapter 2

Design of a Last Mile Transportation System

The focus of this chapter is on the stochastic analysis of the supply side of LMTS: given a probabilistic description of demand, design a LMTS that operates under dynamic and stochastic conditions according to certain guidelines and satisfies a set of Level of Service (LOS) requirements. This implies specifying such system characteristics as vehicle fleet size, service frequency, vehicle dispatching strategies, vehicle routing strategies, monitoring and control of operations, etc. We propose several closed-form expressions as functions of system parameters to bound and estimate system performance, such as the average passenger waiting time and the average passenger riding time. The analytical expressions derived herein can be very useful in designing LMTS, specifically in determining resource requirements for these systems, such as how many vehicles would be necessary to achieve a specified level of service (as measured by expected time until a passenger boards a vehicle or is delivered to his/her destination) and how many kilometers per day these vehicles would travel.

2.1 Background

Bounding and approximating the performance of a last mile transportation system is difficult analytically, as the planning and management of a LMTS generally involves such complications as: stochastic travel times; batch arrivals of prospective passengers; partitioning of demands among vehicles; routing of the vehicles; queueing issues; and,

obviously, numerous considerations concerning staffing and economic sustainability. With the exception of staffing and economic issues, these complications are addressed in this chapter in a static setting.

In a general LMTS, passengers arrive in batches, not singly. Moreover, the size of these batches is a random variable. Queueing systems with batch arrivals are notoriously difficult analytically. A further complication is that the "service times" of passengers are determined by the length (or the duration) of the routes traveled by each vehicle. Thus, in designing a LMTS, it is necessary to consider simultaneously the problems of: allocating passengers among vehicles; routing the vehicles and estimating the lengths of the routes; and computing the queueing performance characteristics of the system.

The main body of this chapter is organized as follows. In Section 2.2, we describe in detail the version of the LMP that we are studying and discuss the associated fundamental assumptions. Section 2.3 outlines our overall approach: we begin by deriving a set of queueing results by considering a fleet of vehicles with capacity for a single passenger ($c = 1$) and then extend the analysis by allowing the vehicle capacity to be arbitrary and by incorporating the resulting travel time estimates into the queueing expressions derived for the $c = 1$ case. Section 2.4 presents our analysis and results for the unit-capacity case. We derive an upper bound and an approximate expression for the performance of a LMTS as a function of its design parameters and then show through a set of simulation experiments that the resulting estimates approximate well the observed waiting times. Section 2.5 examines the general capacity case ($c > 1$) by, first, proposing approximate analytical expressions for the expected value and the variance of the travel times associated with fleets consisting of vehicles with general capacity, and then applying these expressions to the queueing approximation derived in Section 2.4. The results again compare well with those obtained from simulations. We also show that the relaxation time of the queue in the LMTS is significantly shorter than the duration of the time intervals during which the respective demand rates for an LMTS system can be approximated as being roughly constant. It is therefore reasonable to use the steady state approximations derived in this chapter. Section 2.6 contains a summary and concluding remarks.

2.2 Problem Description and Assumptions

We now describe in more detail the LMP scenario of Figure 1.1. The LMTS operates as follows: let STA be the transit rail station served by the LMTS. Consider a passenger, PAX, who boards a train at any station ("ORIGIN") for the purpose of traveling to STA and will then board a LMTS vehicle for transport to her home. PAX is required to provide advance notice to LMTS of her impending arrival at STA. The time interval between the advance notice and the actual arrival of PAX at STA is of the order of several minutes (e.g., at least 5 or 10 minutes) to give the LMTS system sufficient time to plan the service of PAX. In practical terms, the advance notice could be generated by PAX in a number of alternative ways. For example, PAX could use a smart-phone when she arrives at ORIGIN or when she enters her train to STA; or, she could tap a smart card on a special-purpose screen, as she is entering ORIGIN or while aboard the train. The resulting message to the LMTS includes the expected time of arrival of PAX at STA (easy to predict, once the passenger is at the ORIGIN station or aboard a train) and her ultimate destination, e.g., her home address. If the great majority of LMTS users are subscribers whose home addresses are pre-registered, then the only information that PAX will have to provide will be an identification number or code.

Once the information about PAX is received the LMTS will assign PAX to one of the vehicles of the LMTS fleet, plan the route of that vehicle so it includes a visit to the ultimate destination of PAX, estimate the departure time of the vehicle from STA, and notify PAX accordingly. PAX will receive a message (on her smart-phone or by tapping her card on a screen when she arrives at STA) that indicates the vehicle she has been assigned to and the planned departure time of the vehicle from STA (e.g., "please board Vehicle 123 which will depart from STA at 4:26 PM"). Once all the passengers assigned to a vehicle are on board, the vehicle will execute a delivery route, visiting the destination of each of the passengers and will then return to STA to pick up the passengers for its next delivery tour.

The LMTS described above possesses the generic system features that we are most interested in: arrivals of passengers in "batches" (groups) at STA; clustering of passengers into subgroups for assignment to a fleet of vehicles; routing of the vehicles to deliver the passengers on board; and a requirement for fast computation of waiting times and other performance parameters so that, for example, passengers can be notified in a timely way of the departure time of the vehicle they have been assigned to. Actual implementations would probably involve some simpler variants of the above features.

Given the service region's geometry, passenger demand, the spatial distribution of the passenger destinations, and the number, capacity and travel speed of the LMTS vehicles, examples of performance metrics that we wish to compute include: the average waiting time until boarding a vehicle, the average riding time of passengers, the average waiting time until delivery, the minimum number of vehicles we need to reach stable operation, vehicle productivity and workload, and eventually (but not in this thesis) the general cost of operating the system and the service vs. cost trade-offs involved.

We now identify briefly the specifics of the model considered. With reference to Figure 2.1, we make the following assumptions: (i) headways, h, between arrivals of trains at the station (and discharges of passengers) are constant; (ii) passengers are discharged in batches after each train's arrival; (iii) the batch size is a general random variable, N, with known expectation $E(N) = n$ and variance $Var(N) = \sigma_N^2$; (iv) all passengers arriving in a single batch request service essentially simultaneously; (v) given the size of any particular batch, $N = N_0$, the destinations of the N_0 passengers in the batch are distributed identically, uniformly and independently in a service region; (vi) the service region is convex and compact with known dimensions; (vii) the delivery fleet (or pick-up fleet, in the case of "First Mile" service) consists of m vehicles, each with integer capacity, c.

We believe that these assumptions are sufficiently general for approximating, to a first order, the characteristics of many potential variations of LMTS. Note that our model includes the most difficult, from the analytical point of view, features that one might encounter in an LMTS: batch arrivals, stochastic demand, stochastic service times, and the presence of queueing phenomena interfaced with clustering and routing problems.

2.3 Description of Overall Approach

Sections 2.4 and 2.5 of the chapter describe in detail our analysis and results. In this section we provide a brief description of the overall approach we have followed to provide perspective for these detailed sections. We have adopted a viewpoint under which the LMTS is regarded as a spatially distributed queueing system. In line, with typical queueing terminology, we shall refer henceforth to passengers as "customers". The m parallel servers (the vehicle fleet) serve customers in groups of c or smaller, where c is the capacity of each vehicle. The service time for each group is equal to the travel time associated with a vehicle tour that begins at the station/depot, visits each of the c (or fewer) customer destinations and returns to the station/depot to pick up a new group.

Because queueing systems with batch arrivals (like the arrivals of customers at STA) and batch services (like the service of groups of customers by each vehicle) are difficult to analyze, we resort to a two-step approach. In Step 1, we assume that $c = 1$, i.e., that the delivery vehicles have unit capacity. Thus, service times consist simply of the duration of a round-trip between STA and one customer's destination (Figure 2.1), with the destination being randomly and uniformly distributed within the service area per our assumption (v) in Section 2.2. In this way, we obtain a $D^N/G/m/\infty$ queueing system in queueing theory notation, where D^N indicates batch arrivals at constant ("Deterministic") intervals with the number of arriving customers in each batch described by random variable N; G denotes the fact that the distribution of service times (i.e., the duration of the round trips between STA and customer destinations) is "general"; and m and ∞ indicate, respectively, the number of service vehicles and the fact that no *a priori* limit is placed on the number of customers waiting for pickup at STA.

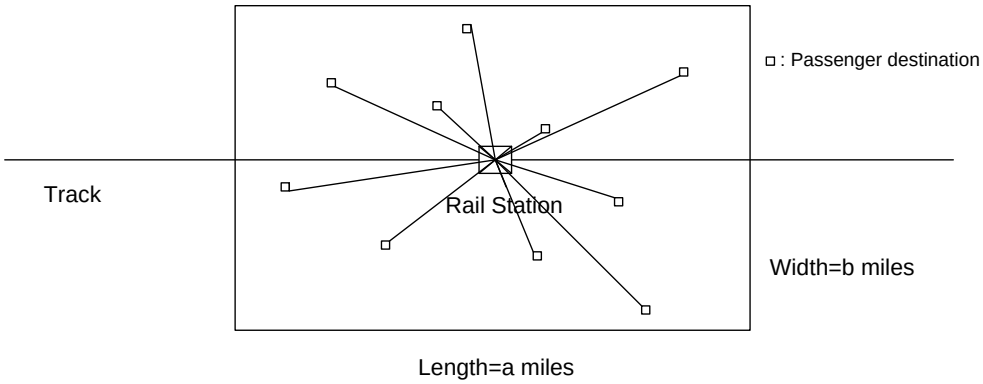


Figure 2.1 Customer destinations and vehicles routes of the Unit-Capacity, Multi-Vehicle LMP

As no closed-form expressions are available for the fundamental quantities that describe the performance of a $D^N/G/m/\infty$ system, we then attempt to obtain expressions for similar queueing systems, which are more tractable mathematically. Through a series of simplifications, we derive (i) an upper bound and (ii) an approximate expression for the mean waiting time associated with $D^N/G/m/\infty$ queues. We then carry out an extensive set of simple simulation experiments and conclude that these expressions provide good estimates of the performance of the system (with $c = 1$) under a broad range of system design parameters.

Step 2 examines the general case, in which service times are equal to the duration of delivery tours consisting of $c(> 1)$ or fewer delivery stops, as shown in Figure 2.2. To apply to the general capacity case the queueing expressions that were derived in Step 1, we need to partition the customers, design near-optimal tours/routes for the vehicles, and compute the approximate expectation and variance of the vehicle tour length. We accomplish this by using arguments from geometrical probability and from the literature on the Traveling Salesman Problem and Vehicle Routing Problem. We then use these expressions, along with the queueing-based approximation derived in Step 1, to complete the process of estimating the performance of the LMTS for the general case of arbitrary fleet size and arbitrary vehicle capacity. Finally, we compare again our approximate estimates to the results of a series of simulations over a broad range of input values.

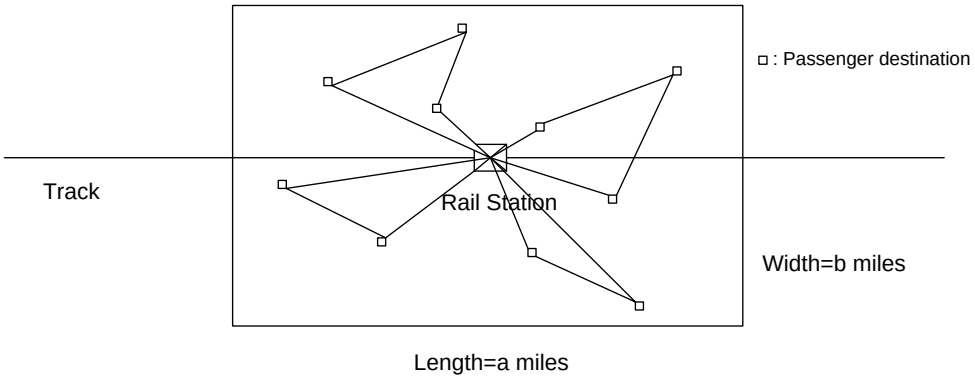


Figure 2.2 Customer destinations and vehicles routes of the General-Capacity, Multi-Vehicle LMP

2.4 The Unit-Capacity, Multi-Vehicle LMP

In this section we present the analysis of the case described in Section 2.3 as Step 1, in which $c = 1$, and m is an arbitrary positive integer. As already indicated (Figure 2.1), the length of the vehicle trips in this case is equal to two times the distance between the rail station and a customer's destination. If we postulate constant and unit travel speed, the expressions for travel times are identical with those derived for travel distances.

The basic notation is summarized as follows:

h = the constant headway between arrivals of trains (and discharges of customers) at the station STA;

N = a random variable denoting the number of LMTS customers ("batch size") discharged after the arrival of a train at STA, with the sizes of successive batches being mutually independent, and $E(N) = n$ and $Var(N) = \sigma_N^2$ denoting, respectively, the expectation and variance of N;

λ_a = the arrival rate of customer batches ($= 1/h$);

σ_a^2 = the variance of the inter-arrival time of customer batches ($\sigma_a^2 = 0$ for constant headways);

S = a random variable denoting the service time of a random LMTS customer with $E(S) = s$ and variance $Var(S) = \sigma_S^2$.

Note that the successive service times of any given vehicle in the fleet are independent and identically distributed. The traffic load (or utilization ratio) is given by $\rho = ns/hm$, since m/s is the service rate of the LMTS, while n/h is the rate of customer arrivals per unit of time.

We are particularly interested in the expected waiting time, W_q, of LMTS customers until they board one of the m vehicles to be transported to their eventual destination. Determining this expected waiting time as a function of the LMTS design parameters is a critical step toward developing the means to design LMTS satisfying certain level-of-service requirements.

2.4.1 General Upper Bound and Approximation

We begin by obtaining a general upper bound and approximate expression for W_q in the original Unit-Capacity, Multi-Vehicle $D^N/G/m/\infty$ model. To do this, for each train's arrival, we pre-assign the discharged customers to different vehicles and then construct a corresponding single-server queueing model $D^{N_S}/G/1/\infty$ for each vehicle, where N_S is the random variable indicating the number of customers from any single train assigned to the same vehicle. Each customer can be served only by the vehicle to which she has been pre-assigned.

With such an assignment policy, service inefficiencies will exist since a customer is required to wait for his or her assigned vehicle, even when other vehicles may be available. Thus, the average waiting time in this case will be larger than the average waiting time in the original model. The customer flow is shown schematically in Figure 2.3.

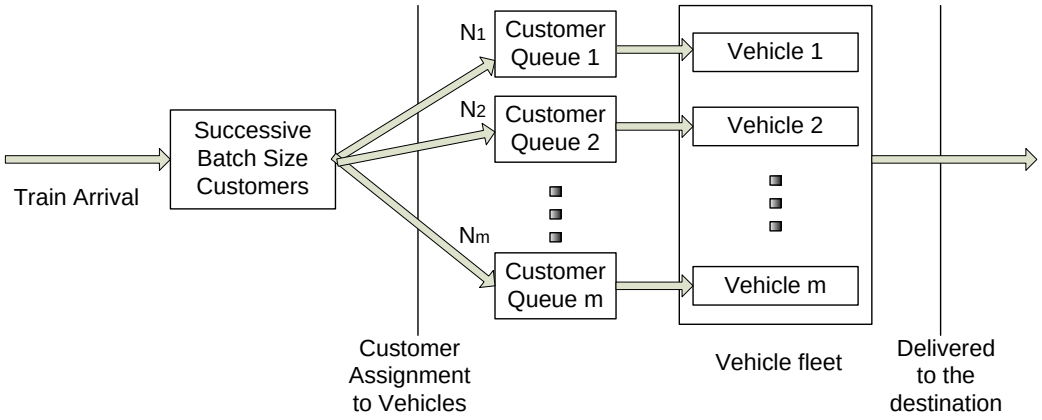


Figure 2.3 Customer flow in the pre-assignment policy

The $D^{N_S}/G/1/\infty$ model is still difficult to work with. To obtain approximate expressions for W_q, we decompose the problem into two parts. First, the N_S customers in a batch who are assigned to the same vehicle are treated as a single "macro-customer" P. If we only consider the "macro-customer", this reduces the $D^{N_S}/G/1/\infty$ model to the more tractable $D/G/1/\infty$ model and allows us to obtain an approximation for W_{q1}, the expected waiting time until the first customer in P receives service.

Let T be the service time of the "macro-customer", $T = \sum_{i=1}^{N_S} S_i$, where N_S depends on the assignment policy and $S_1, S_2, \ldots, S_N$ are the service times of the real customers, which are mutually independent and identically distributed. Note that N_S is a random variable. Therefore,

$$E(T) = \sum_{i=1}^{N_S} E(S_i) = E(N_S)s, \qquad Var(T) = E(N_S)\sigma_S^2 + s^2 Var(N_S),$$

$$C_T^2 = \frac{E(N_S)\sigma_S^2 + s^2 Var(N_S)}{E^2(N_S)s^2}$$

Additionally, $\sigma_a^2 = 0$ because of constant "macro-customer" inter-arrival times, $\lambda_a = 1/h,\ \rho = E(T)/h = E(N_S)s/h$. According to Kingman (1961), Kingman (1962), and Ott (1987), an upper bound for W_{q1}, the expected waiting time of $D/G/1/\infty$ queue is:

$$W_{q1} = W_q(D/G/1/\infty) \leq \frac{\lambda_a(\sigma_a^2 + \sigma_T^2)}{2(1-\rho)} = \frac{\frac{1}{h}(0 + Var(T))}{2(1 - \frac{E(T)}{h})} = \frac{E(N_S)\sigma_S^2 + s^2 Var(N_S)}{2(h - E(N_S)s)} \quad (2.1)$$

According to Kraemer et al. (1976), an approximation of W_{q1} is provided by:

$$W_{q1} = W_q(D/G/1/\infty) \approx \frac{Var(T)}{2\big(h - E(T)\big)} \cdot \exp\left[-\frac{2\big(h - E(T)\big)E(T)}{3Var(T)}\right]$$

$$= \frac{E(N_S)\sigma_S^2 + s^2 Var(N_S)}{2(h - E(N_S)s)} \cdot \exp\left[-\frac{2(h - E(N_S)s)E(N_S)s}{3E(N_S)\sigma_S^2 + 3s^2 Var(N_S)}\right] \quad (2.2)$$

In a second step, we then compute the additional expected waiting time, W_{q2}, until each of the individual customers in macro-customer *P* receives service, following the service to the "macro-customer". For the $i\,th$ customer in *P*, we consider the additional expected waiting time due to being preceded by $i-1$ other customers in *P*. If the macro-customer consists of k customers and $k \geq 1$, the customer in the $i\,th$ position suffers the expected additional total waiting time $W_{q,i\,th} = \sum_{j=1}^{i-1} s_j = (i-1)s$, where s_j is the expected service time of the $j\,th$ customer served before the $i\,th$ customer. Let $W_{q,k\ customers}$ denote the expected total additional waiting time of the k customers:

$$W_{q,k\ customers} = \frac{\sum_{i=1}^{k} W_{q,i\,th}}{k} = \frac{\sum_{i=1}^{k}(i-1)s}{k} = \frac{(k-1)s}{2}, \qquad k \geq 1$$

If $k = 0$, no customers are served, so that $W_{q,0\ customer} = 0$. According to the Law of Total Expectation, the expected additional waiting time of a customer is then given by:

$$W_{q2} = \frac{\sum_{k=0}^{\infty} P(k) W_{q,k\ customers} k}{\sum_{k=0}^{\infty} P(k)k} = \frac{sVar(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \quad (2.3)$$

Thus the upper bound we seek is:

$$W_q = W_{q1} + W_{q2} \leq \frac{E(N_S)\sigma_S^2 + s^2 Var(N_S)}{2(h - E(N_S)s)} + \frac{sVar(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \quad (2.4)$$

The approximation, using (2.2) and W_{q2}, is:

$$W_q = W_{q1} + W_{q2}$$

$$\approx \frac{E(N_S)\sigma_S^2 + s^2 Var(N_S)}{2(h - E(N_S)s)} \cdot \exp\left[-\frac{2(h - E(N_S)s)E(N_S)s}{3E(N_S)\sigma_S^2 + 3s^2 Var(N_S)}\right]$$

$$+ \frac{sVar(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \qquad (2.5)$$

Expression (2.4) and (2.5) are valid under general assumptions about the probability density functions of the batch size, N, and the service times, S. Moreover, (2.4) and (2.5) have been derived without considering how exactly customers are assigned to vehicles. We next analyze one particular reasonable policy for customer assignment to vehicles. The policy will provide a modified $D^{N_S}/G/1/\infty$ model with $E(N_S)$ and $Var(N_S)$, leading to corresponding expressions for W_{q1} and W_{q2}, and, ultimately, to an upper bound and an approximation for W_q.

2.4.2 Cyclic Assignment Policy

One possible policy for allocating customers to vehicles is to assign customers in cyclic order to the vehicles: the first customer in the batch is assigned to Vehicle 1, the second to Vehicle 2, …, the $(m+1)\,th$ to Vehicle 1 again, and so forth. No jockeying of customers, after being assigned to vehicles, is allowed. Figure 2.4 illustrates this policy, which requires assigning an "identification number" to each vehicle to distinguish among them.

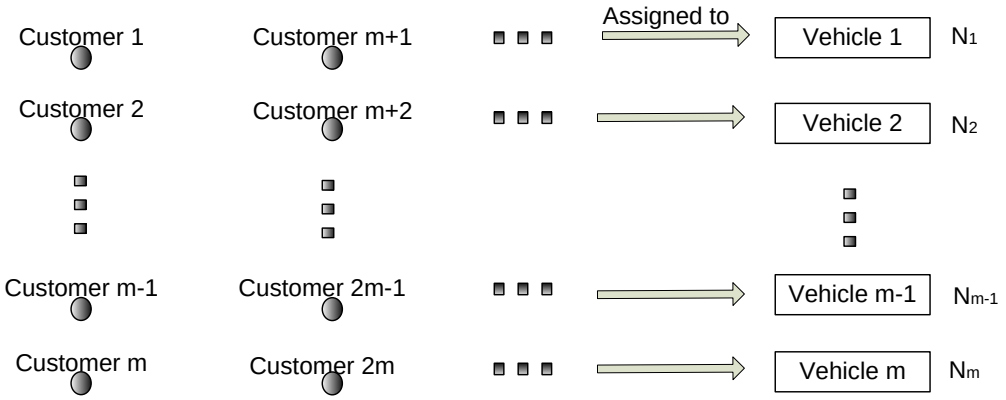


Figure 2.4 Cyclic assignment policy

We have a total of m vehicles, labeled as "Vehicle 1", "Vehicle 2",…, "Vehicle m". Let N_i be the random variable indicating the number of customers assigned to "Vehicle i" after the arrival of a particular train, with the assignment process upon arrival of each train being independent of the arrival process upon arrival of any other train. When one train arrives, we order the m vehicles in sequence: the vehicle that receives customers first is called "1st server", the vehicle that receives customers second is called "2nd server", etc. Let X_i be the random variable indicating the number of customers assigned to the "$i\ th$ server" after the arrival of a particular train. Then,

$$X_1 = \left\lfloor \frac{N+m-1}{m} \right\rfloor, X_2 = \left\lfloor \frac{N+m-2}{m} \right\rfloor, \dots, X_{m-1} = \left\lfloor \frac{N+1}{m} \right\rfloor, X_m = \left\lfloor \frac{N}{m} \right\rfloor$$

$$N = X_1 + X_2 + \cdots + X_{m-1} + X_m$$

If we order the vehicles randomly, the probability that Vehicle i will become the $j\ th$ server for some train is $1/m$. The modified model can be considered as $D^{N_i}/G/1/\infty$. Since $N_1, N_2, \dots, N_m$ are identically distributed, all $D^{N_i}/G/1/\infty$ models can be viewed as identical $D^{N_S}/G/1/\infty$ models although $N_1, N_2, \dots, N_m$ are not necessarily independent.

Recalling that N is the random variable indicating the total number of customers coming from one train, let $N = Km + R$, where $K = \lfloor N/m \rfloor$, and R is the remainder after division of N by m. We can therefore express N as a 2-dimensional random vector, (K, R).

$$X_i = \begin{cases} K+1, & 1 \le i \le R; \\ K, & R+1 \le i \le m; \end{cases}$$

$$E(N_S|(K,R)) = \frac{1}{m}[E(X_1|(K,R)) + E(X_2|(K,R)) + \dots + E(X_m|(K,R))] = \frac{N}{m};$$

$$Var(N_S|(K,R)) = P(N_S = K+1)(K+1-E(N_S|(K,R)))^2 + P(N_S = K)(K-E(N_S|(K,R)))^2$$
$$= \frac{R}{m} \cdot (K+1-\frac{Km+R}{m})^2 + \frac{m-R}{m} \cdot (K-\frac{Km+R}{m})^2 = \frac{Rm-R^2}{m^2}$$

$$E(N_S) = E(E(N_S|(K,R))) = E(\frac{N}{m}) = \frac{n}{m}$$

$$Var(N_S) = E(Var(N_S|(K,R))) + Var(E(N_S|(K,R))) = E(\frac{Rm - R^2}{m^2}) + Var(\frac{N}{m})$$
$$= \frac{E(Rm - R^2)}{m^2} + \frac{\sigma_N^2}{m^2}$$

Since $R < m$, it is also true that $Rm - R^2 \leq m^2/4$, and

$$Var(N_S) \leq \frac{1}{4} + \frac{\sigma_N^2}{m^2} = \frac{4\sigma_N^2 + m^2}{4m^2}$$

In practice, the number of customers N from each batch will typically be much larger than the number of vehicles m, and the remainder R will tend to be uniformly distributed in $\{0, 1, \dots, m-1\}$. Then,

$$E(Rm - R^2) \approx \frac{m^2 - 1}{6m^2}, \qquad Var(N_S) \approx \frac{6\sigma_N^2 + m^2 - 1}{6m^2}$$

By substituting the bound and approximation of $E(N_S)$ and $Var(N_S)$ into (2.4) and (2.5), respectively, the model corresponding to the cyclic assignment policy finally leads to the following upper bound and approximation for the case of a General service time distribution:

$$W_q \leq \frac{4mn^2(\sigma_S^2 + s^2) - 4n^3s^2 + 4hms(\sigma_N^2 + n^2) + hm^3s - 4hm^2ns}{8mn(hm - ns)} \tag{2.6}$$

$$W_q \approx \frac{6mn\sigma_S^2 + 6s^2\sigma_N^2 + m^2s^2 - s^2}{12m(hm - ns)} \cdot \exp\left[-\frac{4(hm - ns)ns}{6mn\sigma_S^2 + 6s^2\sigma_N^2 + m^2s^2 - s^2}\right] + \frac{(6\sigma_N^2 + m^2 + 6n^2 - 6mn - 1)s}{12mn} \tag{2.7}$$

Assuming the service area is a $b \times b$ square with the train station located at the square's center, the travel metric is right angle, and the travel speed is constant throughout the service region and equal to 1, and for Poisson batch sizes, the bound (2.6) becomes:

$$W_q \leq \frac{14b^2mn^2 + 12bhmn^2 - 12b^2n^3 + 12bhmn - 12bhm^2n + 3bhm^3}{24mn(hm - bn)} \tag{2.8}$$

The approximation for this special case is:

$$W_q \approx \frac{(m+6)b^2n + b^2m^2 - b^2}{12m(hm-bn)} \cdot \exp\left[-\frac{4(hm-bn)n}{bmn+6bn+bm^2-b}\right] + \frac{(m^2+6n^2+6n-6mn-1)b}{12mn} \quad (2.9)$$

2.4.3 Another Approximation

In addition to the approach described above, we have developed an alternative way to simplify and approximate the $D^N/G/m$ queue of the Unit-Capacity, Multi-Vehicle LMP. As shown in Figure 2.5, in this alternative approximation, the waiting time is decomposed into two parts: W_{q3}, the waiting time until the first passenger in a batch receives service; and W_{q4}, the waiting time until the following individual customers in that batch receive service. We treat all the customers from each arrival batch as a single "macro-customer" P' and do not pre-assign them to vehicles. This reduces the $D^N/G/m/\infty$ model to the $D/G/m/\infty$ model and allows us to obtain an approximation for W_{q3} using approximations of the $G/G/m/\infty$ model, such as those of Köllerström (1974) and Whitt (1993).

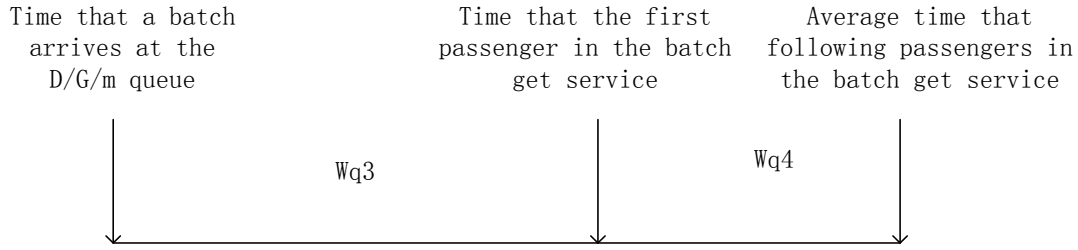


Figure 2.5 Waiting time component

The detailed derivation is described as follows.

Let T' be the service time of the "macro-customer" P', $T' = \sum_{i=1}^{N} S_i$, then:

$$E(T') = \sum_{i=1}^{N} E(S_i) = ns, \qquad Var(T') = n\sigma_s^2 + s^2\sigma_N^2, \qquad C_{T'}^2 = \frac{n\sigma_s^2 + s^2\sigma_N^2}{n^2s^2}$$

In addition, $\sigma_a^2 = 0$, $\lambda_a = 1/\,h,\ \rho = E(T')/mh = ns/mh$. According to Whitt (1994), a good approximation for $D/G/m/\infty$ queue can be written as:

$$W_{q3} \approx \phi(m,\rho)\frac{C_{T'}^2}{2}W_q(M/M/m)$$

where

$$\phi(m,\rho) = (1 - 4\cdot min\{0.24, \frac{(1-\rho)(m-1)\left((4+5m)^{\frac{1}{2}}-2\right)}{16m\rho}\})\cdot exp(\frac{-2(1-\rho)}{3\rho})$$

and

$$W_q(M/M/m) \approx \frac{E(T')(\rho^{(\sqrt{2(m+1)}-1)})}{m(1-\rho)}$$

Therefore, substituting $E(T'), Var(T'), C_{T'}^2$ and ρ, we obtain:

$$W_{q3} \approx (1 - 4\cdot min\{0.24, \frac{(mh-ns)(m-1)\left((4+5m)^{\frac{1}{2}}-2\right)}{16mns}\})\cdot exp(\frac{-2(mh-ns)}{3ns})\cdot \cdot\frac{h\cdot(ns)^{\left(\sqrt{2(m+1)}-2\right)}\cdot(n\sigma_s^2 + s^2\sigma_N^2)}{2\cdot(mh)^{(\sqrt{2(m+1)}-1)}\cdot(mh-ns)} \quad (2.10)$$

Next we study W_{q4}. At the point in time when the first passenger in the batch gets access to service, one server becomes available, while the other servers are still busy. Hence, the following passengers in the batch should wait for more available servers. We can approximate the expected waiting time until the next server becomes available as s/m. Therefore, assuming k passengers in the batch:

For the first passenger in the batch: $W_{q4,1st} = 0$;

For the second passenger in the batch: $W_{q4,2nd} = s/m$;

For the third passenger in the batch: $W_{q4,3rd} = 2s/m$;

…

For the $k\ th$ passenger in the batch: $W_{q4,k\ th} = (k-1)s/m$;

Let $W_{q4,k\ customers}$ denote the average additional waiting time of the k customers:

$$W_{q4,k\ customers} = \frac{\sum_{i=1}^{k} W_{q4,i\ th}}{k} = \frac{\sum_{i=1}^{k}(i-1)s/m}{k} = \frac{(k-1)s}{2m}, \qquad k \geq 1$$

If $k = 0$, no customers are served, so that $W_{q4,0\ customer} = 0$. According to the Law of Total Expectation, the expected additional waiting time of a customer is given by:

$$W_{q4} = \frac{\sum_{k=0}^{\infty} P(k) W_{q4,k\ customers} k}{\sum_{k=0}^{\infty} P(k) k} = \frac{sVar(N) + sE^2(N) - sE(N)}{2mE(N)} = \frac{s\sigma_N^2 + sn^2 - sn}{2mn}$$

(2.11)

Therefore, the approximation of the expected waiting time of passengers is obtained:

$$\begin{aligned} W_q &\approx W_{q3} + W_{q4} \\ &\approx (1 - 4 \cdot min\{0.24, \frac{(mh-ns)(m-1)\left((4+5m)^{\frac{1}{2}} - 2\right)}{16m\lambda s}\}) \\ &\cdot exp(\frac{-2(mh-ns)}{3ns}) \cdot \frac{h \cdot (ns)^{\left(\sqrt{2(m+1)}-2\right)} \cdot (n\sigma_s^2 + s^2\sigma_N^2)}{2(mh-ns) \cdot (mh)^{(\sqrt{2(m+1)}-1)}} + \frac{s\sigma_N^2 + sn^2 - sn}{2mn} \end{aligned}$$

(2.12)

Expression (2.12) is valid under general assumptions about the probability density functions of the batch size N, and the service times S. For Poisson batch sizes, a square service region with a right-angle distance metric, and vehicles with unit capacity and constant speed 1, the approximation (2.12) becomes:

$$\begin{aligned} W_q \approx &\left(1 - 4 \cdot min\left\{0.24, \frac{(mh-nb)(m-1)\left((4+5m)^{\frac{1}{2}} - 2\right)}{16mnb}\right\}\right) \cdot exp\left(\frac{-2(mh-nb)}{3nb}\right) \\ &\cdot \frac{7bh}{12(mh-nb)} \cdot \left(\frac{nb}{mh}\right)^{\left(\sqrt{2(m+1)}-1\right)} + \frac{sn}{2m} \end{aligned}$$

(2.13)

According to numerical experiments, these approximations (2.12) and (2.13) perform worse than the approximations (2.7) and (2.9) in most cases, except under extremely high utilization ratios when the system is unstable and both approximations perform poorly, anyway. Additionally, the expressions (2.7) and (2.9) are simple closed-form expressions, much simpler than expressions (2.12) and (2.13).

2.4.3 Numerical Experiments for the Unit-Capacity, Multi-Vehicle LMP

To assess the performance of the expressions obtained in Sections 2.4.1 and 2.4.2 under a broad range of conditions, a simple simulation of the Unit-Capacity, Multi-Vehicle LMP was carried out with a program written in java. We consider a square service region with geometry $b/v = 2.5\,min = 150\,sec$, headway $h = 10\,min = 600\,sec$, and Poisson-distributed batch sizes of $n = 20, 40, 60, 80$. We selected these parameters so that the system would make sense physically.

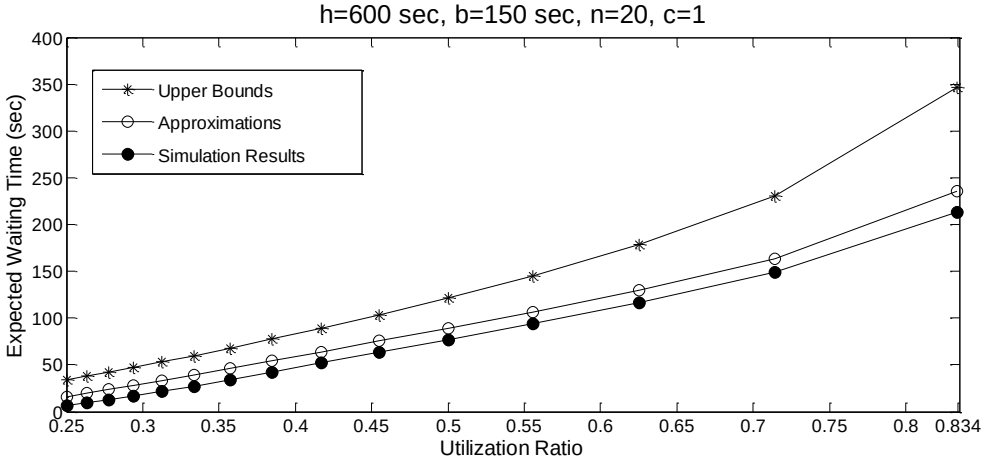


Figure 2.6 Simulation results, bounds and approximations of average waiting time when $n = 20$

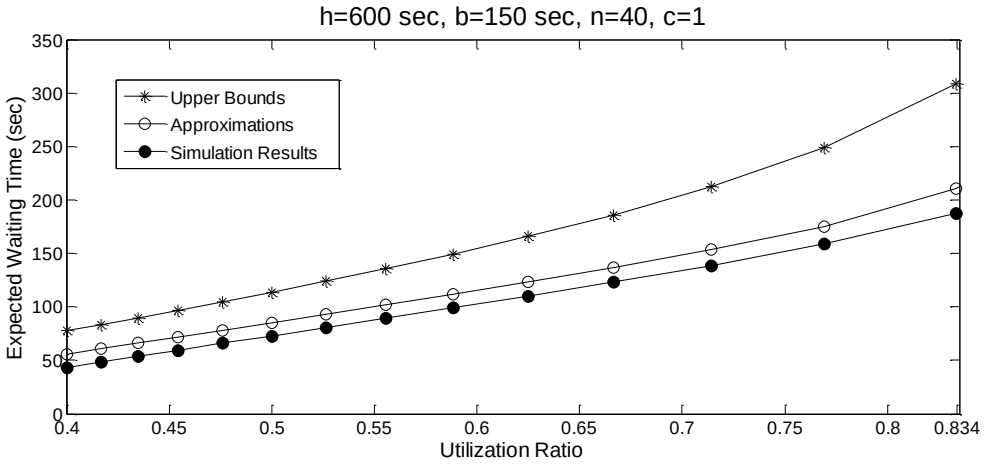


Figure 2.7 Simulation results, bounds and approximations of average waiting time when $n = 40$

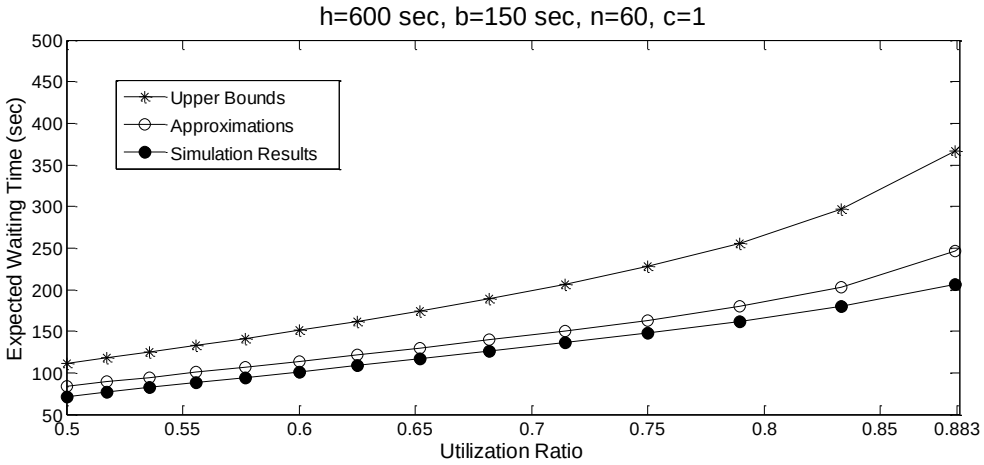


Figure 2.8 Simulation results, bounds and approximations of average waiting time when $n = 60$

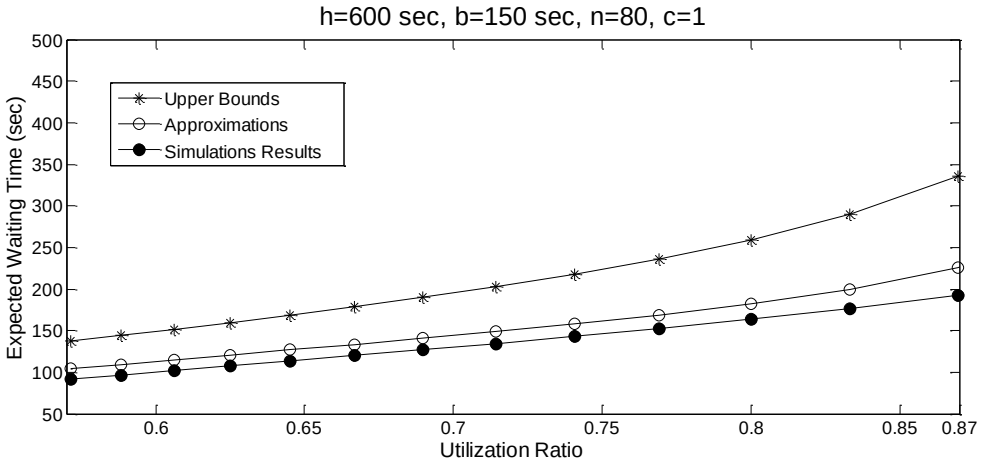


Figure 2.9 Simulation results, bounds and approximations of average waiting time when $n = 80$

Figures 2.6 – 2.9 plot the simulation results and our estimates for the average waiting time per customer W_q (in seconds) against the utilization ratio $\rho = sn/hm$. Since the simulated system has Poisson customer batch size and a square service region, the upper bound (expression (2.8)), and the approximation (expression (2.9)) from Sections 2.4.2 are applicable and considered here. For each demand intensity n, the utilization ratio ρ takes on a set of discrete values because the number of vehicles, m, is integer. We have plotted the points with utilization ratio less than 0.9, above which the system is highly unstable and the average waiting time is too long to be accepted practically.

It can be seen that (2.8) is a consistently reliable upper bound for W_q, while (2.9) provides a very good approximation for the entire range of parameter values for which the LMTS remains stable. In a practical system, it would be desirable to achieve values of 1 to 5 minutes, for the average waiting time until customers board a vehicle. Note from Figures 2.6 – 2.9 that for this range of values (60 to 300 seconds) the difference between the approximation and the simulation results stays small in absolute or percentage terms. For example, when $n = 20$ (Figure 2.6), this difference never exceeds the greater of 15 seconds or 12% for values of W_q between 1 and 4 minutes.

We have also performed simulation experiments with rectangular and diamond-shaped service regions and with discontinuities in the travel medium, such as an

impenetrable barrier to travel. For these environments we have derived expressions for W_q, analogous to (2.8) and (2.9), based on (2.6) and (2.7) – see Wang (2012). These experiments led to the conclusion that the analytical upper bound and approximation continue to perform well under a wide range of conditions.

2.5 General-Capacity, Multi-Vehicle LMP: Approximations

In this section we shall generalize the results of Section 2.4 by considering the General-Capacity, Multi-Vehicle LMP, in which the vehicle capacity, c, and the number of vehicles, m, are arbitrary positive integers. The vehicles will now travel along more complicated routes than in the $c = 1$ case to deliver customers to their destinations. In practice, one would expect the vehicle capacity to be smaller than that of a regular bus – typically a number between 3, for service provided by taxi-like vehicles, and 20, for large vans.

As explained in Section 2.3, the General-Capacity, Multi-Vehicle LMTS will be viewed as a spatially distributed queueing system in which the service times are equal to the amount of time it takes to complete a customer delivery tour and return to the train station (Figure 2.2). After each batch of arrivals, the customers must be partitioned into clusters and assigned to vehicles according to their destinations and the vehicles must then be routed with the objective of obtaining a shortest total travel distance – which translates into shortest service times and smallest overall queueing. This means that the estimation of model parameters, such as the expected value and the variance of service times, is now far more complicated than when $c = 1$.

2.5.1 Adjustment of the Queueing Model

We first need to make some adjustments to the principal expression (2.7) that we have derived from our queueing model. For the General $(c > 1)$ Capacity case, General distribution of customer batch size and General service times, the approximation for the waiting time until boarding a vehicle is given by:

$$W_{q,Board} \approx \frac{6mE(N_E)Var(S_E) + 6E^2(S_E)Var(N_E) + E^2(S_E)m^2 - E^2(S_E)}{12m\big(hm - E(N_E)E(S_E)\big)} \cdot exp\left[-\frac{4\big(hm - E(N_E)E(S_E)\big)E(N_E)E(S_E)}{6mE(N_E)Var(S_E) + 6E^2(S_E)Var(N_E) + E^2(S_E)m^2 - E^2(S_E)}\right] + \frac{(6Var(N_E) + m^2 + 6E^2(N_E) - 6mE(N_E) - 1)E(S_E)}{12mE(N_E)} \quad (2.14)$$

The expression (2.14) is exactly the same as (2.7), except S is substituted by S_E, the travel time to serve c customers, and N by N_E, the random variable indicating the number of tours formed following the arrival of a batch of customers.

Note that in (2.14) we have used the notation $W_{q,Board}$ for the expected waiting time until a customer will board a vehicle, while in (2.7) we used the notation W_q for the same quantity. This is because we also want to introduce here another quantity, W_{Riding}, which is defined as the expected time a customer will spend riding on the vehicle before being delivered to her destination. Considering the riding component of the trip, the total expected time from the instant a customer arrives at the rail station until she is delivered at her destination is given by

$$W_{Delivered} = W_{q,Board} + W_{Riding} \quad (2.15)$$

The expected riding time of the $i\,th$ delivered customer in a tour with c customer deliveries is approximated by $i \times E(S_E)/(c+1)$ and the expected riding time of a random customer is $E(S_E)/2$.

2.5.2 Approximating the Expected Value of Customer Service Times

We turn next to the task of evaluating the performance of the general expressions (2.14) and (2.15). To do this, expressions must be developed for all terms involving S_E and N_E. In this subsection and the next two, we propose a set of such approximate expressions for the case in which the destinations of the customers are uniformly and independently

distributed within a square $b \times b$ district, assuming Euclidean travel. An entirely different operating environment is examined in Section 2.5.6.

We start with the critical quantity $E(S_E)$, the expected travel time to deliver c customers. Consider the situation in which j customers are to be delivered by vehicles with capacity c each within the district of interest in a minimum total amount of travel time. Vehicles must return to their origin (the train station). This is a classical Vehicle Routing Problem (VRP).

Eilon et al. (1971) proposed an empirical formula for $E(TVRT_{j,c})$, the total length of vehicle routing tours when a total of j customers are delivered using vehicles of capacity c, but tested it for only up to $j = 70$ and $c = 10$. Daganzo (1984) provided another simple and intuitive analytical approximation:

$$E\left(TVRT_{j,c}\right) \approx \frac{2rj}{c} + 0.57\sqrt{jA} \tag{2.16}$$

where r is the average distance between the customers and the depot and A is the area of service region. For a $b \times b$ square region, uniformly distributed customer destinations, and the depot located at the center of the region, $r = 0.382b$ and expression (2.16) becomes:

$$E\left(TVRT_{j,c}\right) \approx 0.764\frac{j}{c}b + 0.57\sqrt{j}b \tag{2.17}$$

The expectation of a single route length of the vehicle routing tours $VRT_{j,c}$ can then be approximated as:

$$E\left(VRT_{j,c}\right) \approx \frac{E\left(TVRT_{j,c}\right)}{j/c} \approx 0.764b + 0.57\frac{c}{\sqrt{j}}b \tag{2.18}$$

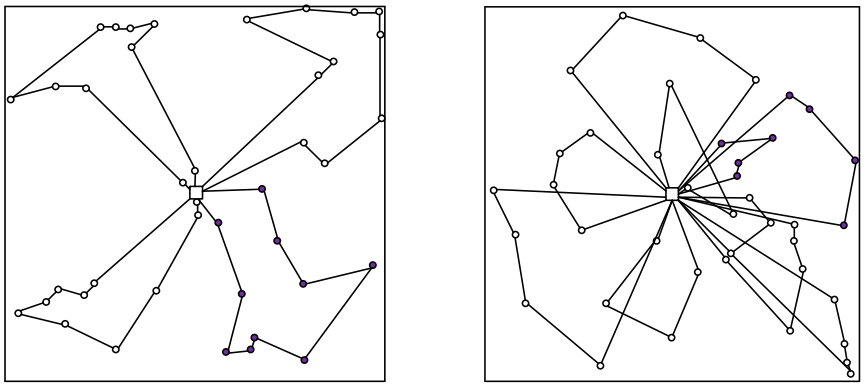


Figure 2.10 Best routes for a $j = 40, c = 10$ instance (left), a $j = 40, c = 4$ instance (right)

To assess the accuracy of (2.18), we took advantage of the fact that good heuristics exist for the VRP. Specifically, we simulated hundreds of thousands of instances of LMTS train arrivals and associated customer destinations. To create the clusters and routes we applied two widely used VRP heuristics, the Sweep algorithm (coupled with a TSP heuristic) and the Clark-Wright algorithm. According to Cordeau et al. (2002), these fast and simple heuristics provided an average deviation of 6.71% and 7.09%, respectively, from the best solutions obtained on CMT benchmark instances. We solved all the simulated instances we generated using each of the two heuristics separately and, for each instance, we chose the better of the two solutions. Figures 2.10 shows the best solutions obtained for two examples, both with $j = 40$ customers but in one case with $c = 10$ and in the other with $c = 4$. As might be expected, the Sweep algorithm generated the solution shown on the left and Clark-Wright the one shown on the right. For a broad range of vehicle capacities, c (2 to 20), and number of routes j/c (2 to 20), the average error of (18), in absolute value terms, was of the order of 2%. Table 2.1 shows part of this assessment.

		c							
		3	4	5	6	7	8	9	10
j/c	4	5.90%	2.58%	1.48%	0.94%	0.52%	0.29%	-0.12%	-0.49%
	5	6.02%	3.24%	2.40%	1.56%	0.79%	0.77%	0.68%	0.54%
	6	5.68%	2.39%	1.97%	1.75%	1.40%	1.06%	0.82%	0.97%
	7	5.04%	2.99%	2.14%	1.38%	1.31%	1.24%	1.24%	1.47%
	8	4.29%	2.88%	1.86%	1.41%	1.25%	1.24%	1.41%	1.30%
	9	5.20%	2.94%	1.79%	1.33%	1.29%	1.20%	1.34%	1.28%
	10	4.39%	2.44%	1.58%	0.97%	0.92%	0.77%	0.87%	1.03%
	11	4.36%	2.54%	1.73%	0.92%	0.90%	0.84%	0.90%	1.27%
	12	4.09%	2.27%	1.48%	1.11%	0.84%	0.64%	0.64%	0.92%
	13	4.18%	2.20%	1.55%	1.08%	0.64%	0.43%	0.53%	0.51%
	14	4.05%	2.30%	1.29%	0.85%	0.75%	0.25%	0.38%	0.49%
	15	4.12%	2.38%	1.58%	0.81%	0.65%	0.28%	0.27%	0.30%

Table 2.1 Error of expression (2.18) compared to results of simulation

In the queueing model, the service time, S_E, is the time that a vehicle takes to traverse a tour and deliver a group of c customers. Estimates of the length of a VRT can be converted into time units, using information about the speed of travel in the region of interest. To simplify this conversion, we shall continue to assume here that travel speed is constant and equal to 1 throughout the region. Considering that the size, j, of a customer batch is sampled from the General distribution of a random variable N, we finally have:

$$E(S_E) \approx \frac{\sum_{j=0}^{\infty} P(N=j)\cdot\frac{j}{c}\cdot E\left(VRT_{j,c}\right)}{\sum_{j=0}^{\infty} P(N=j)\cdot\frac{j}{c}} = \frac{\sum_{j=0}^{\infty} P(N=j)\cdot j\cdot E\left(VRT_{j,c}\right)}{E(N)}$$

$$\approx \frac{\sum_{j=0}^{\infty} P(N=j)\cdot j\cdot\left(0.764b + 0.57\frac{c}{\sqrt{j}}b\right)}{E(N)} = \frac{0.57cE\left(\sqrt{N}\right)b + 0.764E(N)b}{E(N)}$$

$$= \frac{0.57cE\left(\sqrt{N}\right)}{E(N)}b + 0.764b \qquad (2.19)$$

2.5.3 Approximating the Variance of Customer Service Times

Our simulation experiments indicated that the coefficient of variation of $VRT_{j,c}$ is small, typically of the order of 0.1 – 0.25 for most combinations of c and j/c values in the range of $c = 3 - 20$ and $j/c = 4 - 20$. Thus, the standard deviation of $VRT_{j,c}$ is small compared to its expected value. In addition, it is well known from queueing theory (and has been confirmed by the simulation experiments in the specific context of this chapter) that the variance of the service times has only a secondary impact on expected waiting times because it does not affect the utilization ratio, ρ, of a queueing system. For these reasons, we can use simple approximations for the variance (or for the coefficient of variation) of $VRT_{j,c}$ without affecting by much the quality of the approximations obtained for the expected waiting time and expected time to delivery of (2.14) and (2.15).

Beardwood et al. (1959) proposed a famous asymptotic expression for the expectation of the length of a Traveling Salesman Tour (TST) with k independent, uniformly distributed points in a square of area A with Euclidean travel,

$$E(TST)_k \approx \beta_{1,k}\sqrt{kA} \tag{2.20}$$

For very large values of k, the best available estimate seems to be $\beta_{1,\infty} \approx 0.7124$ (Johnson 1996). Gremlich et al. (2004) have suggested that, the variance of the length of a TST with a large number of points can be approximated as,

$$Var(TST)_\infty \approx \beta_{2,\infty}A \tag{2.21}$$

where $\beta_{2,\infty} \approx 0.1385$. Let C denote a coefficient of variation. Then, using (2.20) and (2.21), we shall use

$$C_{TST,k} = \frac{\sqrt{Var(TST)_k}}{E(TST)_k} \approx \frac{\sqrt{Var(TST)_\infty}}{E(TST)_\infty} \approx \sqrt{\frac{\beta_{2,\infty}}{k{\beta_{1,\infty}}^2}} \tag{2.22}$$

as an approximate expression for the coefficient of variation of the length of a TST.

For the case at hand, in which the number of points visited by vehicles with capacity c is $c+1$, we shall use the further approximation

$$C_{VRT,j,c} \approx C_{TST,c+1} \approx \sqrt{\frac{\beta_{2,\infty}}{(c+1)\beta_{1,\infty}{}^2}} \tag{2.23}$$

on the premise that the variability of a VRT and a TST that visit the same number of points should be similar. This leads to

$$Var\left(VRT_{j,c}\right) = E^2(VRT_{j,c}) \cdot C^2_{VRT,j,c} \approx \frac{\beta_{2,\infty}}{(c+1)\beta_{1,\infty}{}^2} E^2(VRT_{j,c}) \tag{2.24}$$

and finally, for speed of travel equal to 1, to

$$Var(S_E) \approx \frac{\beta_{2,\infty}}{(c+1)\beta_{1,\infty}{}^2} E^2(S_E) \tag{2.25}$$

We tested the accuracy of (2.25) against the results of our simulations. Despite the rough nature of approximations (2.22) – (2.23) on which (2.25) is based, the observed average errors were of the order of only 30% for a broad range of values of the vehicle capacity c and the number of routes j/c. As the next subsection indicates this is very adequate due to the limited impact of $Var(S_E)$ on the value of the expected waiting time.

2.5.4 Simulation and Comparisons for the General-Capacity, Multi-Vehicle LMP

Expressions (2.14) and (2.15) will now be tested for the case in which the size of customer batches has a Poisson distribution with intensity n. All the other assumptions (independent and uniform locations of customer destinations, Euclidean travel, square district with size b, travel speed equal to 1) are the same as above.

Under the Poisson assumption $E\left(\sqrt{N}\right) \approx \sqrt{n}$ in (2.15) and therefore,

$$E(S_E) \approx \frac{0.57c\sqrt{n}}{n} b + 0.764b = \frac{0.57c}{\sqrt{n}} b + 0.764b \tag{2.26}$$

The variance of S_E is given in (2.25), while the various other terms of (2.14) and (2.15) take on the following values:

$$E(N_E) \approx \frac{E(N)}{c} = \frac{n}{c} \tag{2.27}$$

$$VAR(N_E) \approx Sd^2\left(\left\lceil\frac{N}{c}\right\rceil\right) \approx \left\lceil\sqrt{Var\left(\frac{N}{c}\right)}\right\rceil^2 \approx \left\lceil\frac{\sqrt{n}}{c}\right\rceil^2 \tag{2.28}$$

$$E\left({N_E}^2\right) \approx \left\lceil\frac{\sqrt{n}}{c}\right\rceil^2 + \left(\frac{n}{c}\right)^2 \tag{2.29}$$

A simulation of a General-Capacity, Multi-Vehicle LMTS was performed with a program written in java. We consider a square service district with geometry $a/v_x = b/v_y = 2.5\,min = 150\,sec$, headways between train arrivals of $h = 10\,min = 600\,sec$, vehicle capacity $c = 3 - 20$ and customer arrivals with batch sizes described by a Poisson distribution with n = 40, 80 and 120. These parameters were selected so that the system would make sense physically. As before, vehicle tours were generated by using the two well-known vehicle routing heuristics, the Sweep algorithm and the Clark-Wright algorithm. Specifically, the simulation generated sets of points, uniformly and independently distributed in a $b \times b$ square, and vehicle tours through these points were drawn using the better of the two solutions (shortest total length of the delivery tours).

Figures 2.11 through 2.18 present a sample of comparisons between the simulation results and the analytical approximations of Section 2.5.1 for the following respective cases: $c = 5, n = 40, 80, 120$; $c = 10, n = 40, 80, 120$; and $c = 15, n = 120$; $c = 20, n = 120$.

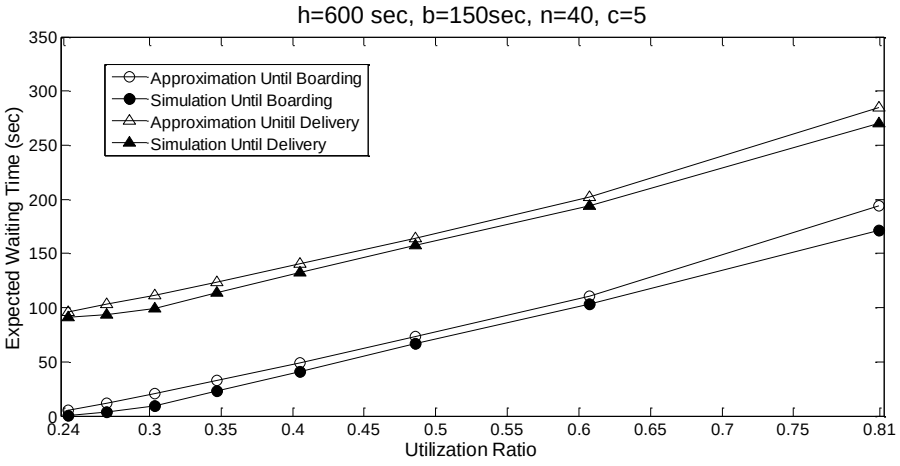


Figure 2.11 Simulation and analytical results when $c = 5$ and $n = 40$

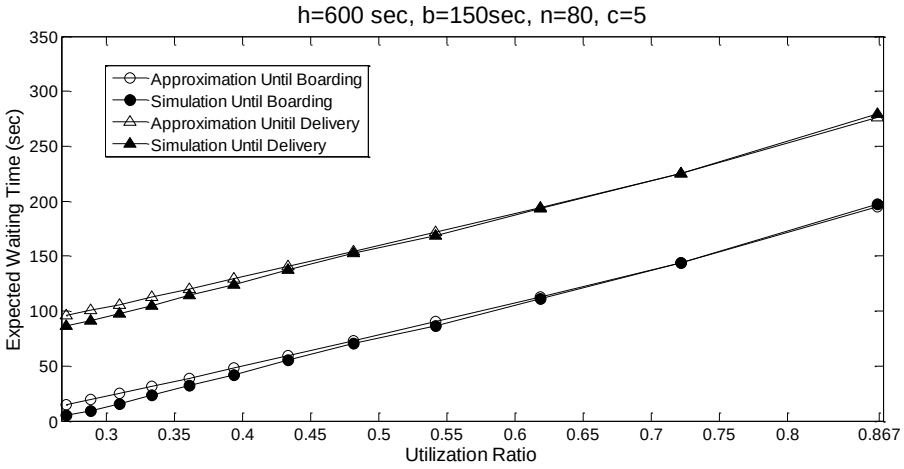


Figure 2.12 Simulation and analytical results when $c = 5$ and $n = 80$

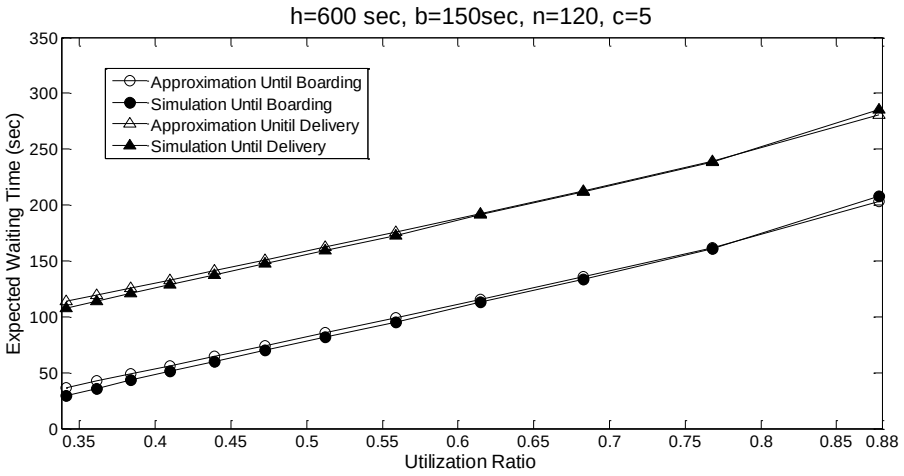


Figure 2.13 Simulation and analytical results when $c = 5$ and $n = 120$

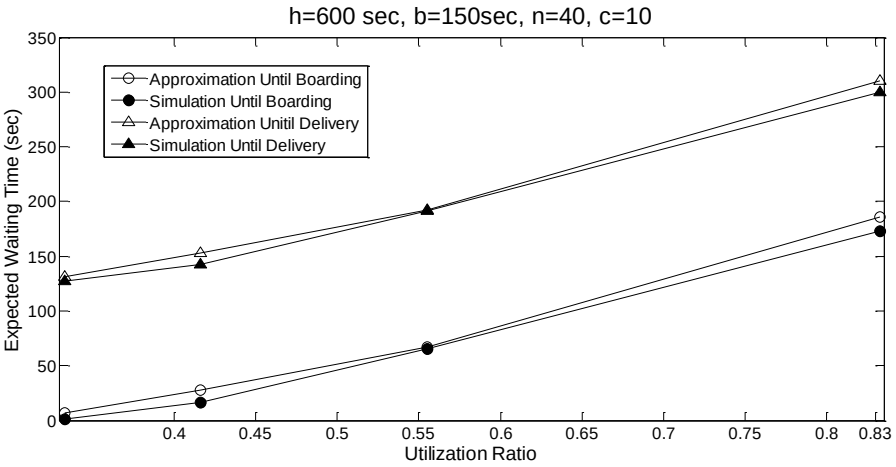


Figure 2.14 Simulation and analytical results when $c = 10$ and $n = 40$

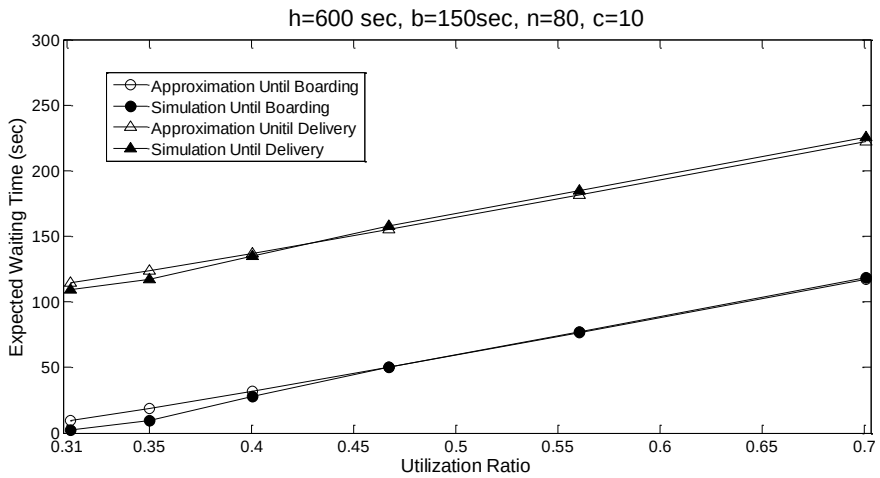


Figure 2.15 Simulation and analytical results when $c = 10$ and $n = 80$

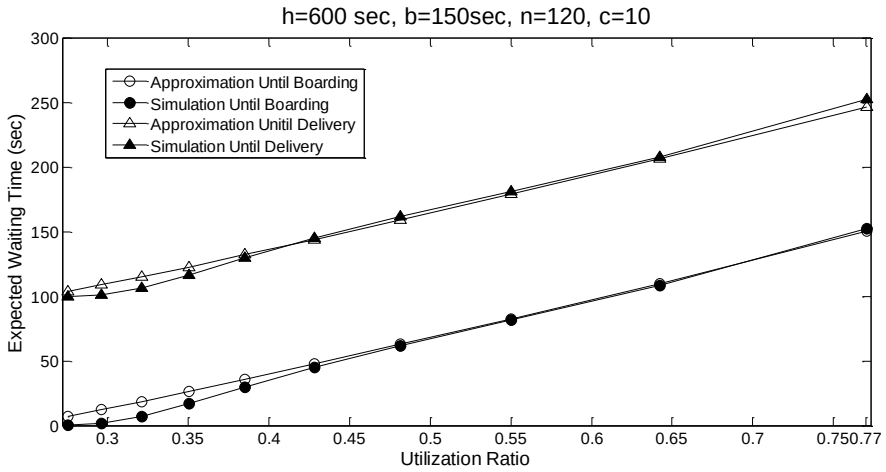


Figure 2.16 Simulation and analytical results when $c = 10$ and $n = 120$

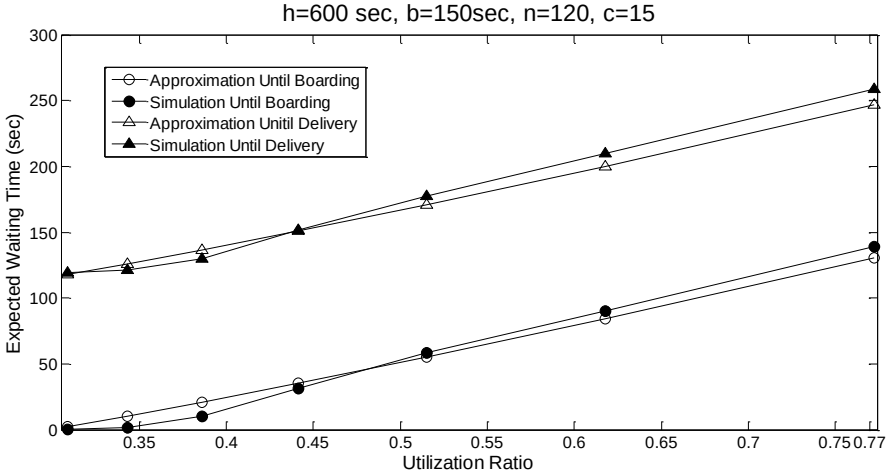


Figure 2.17 Simulation and analytical results when $c = 15$ and $n = 120$

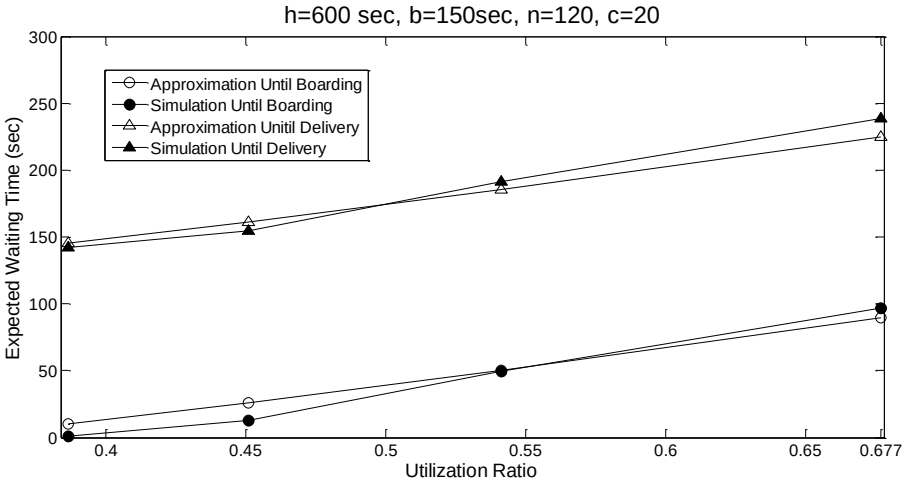


Figure 2.18 Simulation and analytical results when $c = 20$ and $n = 120$

The horizontal axis in Figures 2.11 – 2.18 shows the utilization ratio $\rho = E(S_E)E(N_E)/hm$, while the vertical axis shows the expected waiting time until boarding a vehicle and the expected total waiting time spent between arrival at the station and delivery at the customer's destination. "Approximation Until Boarding" is obtained from (2.14) and "Approximation Until Delivery" from (2.15). For each combination of vehicle capacity c and demand intensity n, the utilization ratio ρ takes on a set of discrete values because the number of vehicles, m, is integer.

We plot in Figures 2.11 - 2.18 the discrete points corresponding to utilization ratios less than 0.9. At values of ρ higher than 0.9 the system is highly unstable and the average waiting time is too long to be acceptable in practice. As shown in Figures, the approximate expression (2.14) for the expected passenger waiting time until boarding a vehicle performs very well for both small and large vehicles and for the broad range of customer arrival intensities (n = 40, 80, and 120) examined. The difference between the simulated average time until boarding and the analytical expression (2.14) is of the order of the greater of 5% or 10 seconds. Turning to the estimation of passenger expected total time until delivery, the analytical expression (2.15) also works well. The difference between the analytical and simulation results is now of the order of the greater of 5% or 15 seconds.

For similar LMTS with irregular service region and non-uniform demand, the VRT length (and time, with assumption of speed) expectation can be approximated following the method described by Daganzo (1984), and the variance can be approximated following the method described in Section 2.5.3. Expression (2.14) and (2.15) can be then used to estimate the system performance.

2.5.5 Relaxation Time

In order to assess whether the steady state expressions derived in Section 2.5.1 can be used as reasonable approximations of expected queuing performance during a typical time interval in which we can assume a roughly steady demand rate, we look at the "relaxation time" of the LMTS $D^N/G/m$ queue. If the relaxation time is small compared to a typical time interval with steady demand rate, such as 2 to 3 hours, it is reasonable to use the steady state approximations.

We note that the queuing literature provides very few exact expressions regarding relaxation times. Most of the (few) available results provide expressions for upper bounds on these relaxation times.

We define t_1 as the relaxation time of the $D^N/G/m$ queue in the original model, t_2 as the relaxation time of the $D^{Ns}/G/1$ queue with customer pre-assignment, and t_3 as the relaxation time of the $D/G/1$ queue with macro-customers.

There exists mutual help (collaboration) between servers (vehicles) in the original $D^N/G/m$ queue, while customers are pre-assigned to different vehicles in the $D^{Ns}/G/1$ queue. This implies that the $D^N/G/m$ queue has more flexibility than the $D^{Ns}/G/1$ queue. Therefore, the $D^N/G/m$ queue will reach steady state faster than the $D^{Ns}/G/1$ queue under the same demand and service conditions, i.e., $t_1 \leq t_2$.

t_3, the relaxation time for macro-customers waiting in the $D/G/1$ queue, is exactly the same as the relaxation time for the first individual customer waiting in the $D^{Ns}/G/1$ queue.

For a random individual customer in a macro-customer group, the waiting time is given by $W = W_{q1} + W_{q2}$, where W_{q1} is the waiting time until the first individual customer in the macro-customer group receives service, and W_{q2} is the additional waiting time that this customer will suffer after the first individual customer receives service. The transient behavior of W depends only on the transient behavior of W_{q1}. W_{q2} does not affect in any way the transient behavior of W, because W_{q2} is determined solely by the number of individual customers in the macro-customer group.

In other words, the relaxation time for the first individual customer in the macro-customer group is also the relaxation time for all individual customers in the macro-customer, i.e., $t_2 = t_3$.

Combining the two relations above, we obtain $t_1 \leq t_3$.

According to Odoni and Roth (1983), the relaxation time for a single-server queue is upper-bounded by:

$$u_r = (C_A^2 + C_S^2)/(2.8\mu(1 - \sqrt{\rho})^2) \tag{2.30}$$

where C_A^2 is the square of the coefficient of variation for inter-arrival times, C_S^2 is the square of coefficient of variation for the service times, μ is the average service rate, and ρ is the system utilization ratio.

According to Newell (1971), the relaxation time can be approximated as:

$$a_r = (\rho C_A^2 + C_S^2)/(\mu(1 - \rho)^2) \tag{2.31}$$

We evaluate u_r and a_r for the relaxation time for some of the typical sets of parameters used in this Chapter (Figures 2.11 – 2.18). Parts of the results are shown in Table 2.2. For each set of parameters, we indicate the estimates of the relaxation time (shown in bold in minutes with u_r first and a_r second) for the highest possible utilization ratio (shown as the first of the three numbers in each box), which corresponds to the longest possible relaxation time. We find that, in all cases with utilization ratios of 0.85 or lower, the upper bound for the relaxation time, u_r, is less than or in the order of 25 minutes. The estimated approximate relaxation time, a_r, is even smaller. Only when the

utilization ratio is extremely large (such as 0.93) does the highest possible relaxation time become roughly equal to a 2-3 hour interval. However, as noted in the Chapter, the LMTS is highly unstable at these very high utilization ratios and the average waiting time is too long to be acceptable.

$h = 600\ sec, b = 150\ sec$				
$(\rho, \boldsymbol{u_r}, \boldsymbol{a_r}$ $\boldsymbol{in\ minutes})$	$n = 40$		$n = 80$	
$c = 5$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
	$(0.81, \mathbf{22.2}, \mathbf{17.2})$	$(0.61, \mathbf{4.9}, \mathbf{4.3})$	$(0.87, \mathbf{32.8}, \mathbf{24.7})$	$(0.72, \mathbf{7.3}, \mathbf{6.0})$
$c = 10$	$m = 2$	$m = 3$	$m = 3$	$m = 4$
	$(0.83, \mathbf{38.7}, \mathbf{29.6})$	$(0.56, \mathbf{5.6}, \mathbf{5.1})$	$(0.93, \mathbf{169.6}, \mathbf{122.8})$	$(0.70, \mathbf{8.2}, \mathbf{6.8})$

Table 2.2 Upper bound and approximation of relaxation time t_3

Note, in addition, that, in a real LMTS, the queueing system does not begin from rest (i.e., demand does not start from 0). The time to reach steady state should therefore be smaller than the computed upper limits for the relaxation time in Table 2.2.

Thus, the time for the queue to reach steady state should be significantly shorter than the duration of the time intervals (e.g., morning rush period, or evening rush period, or midday period) during which the respective demand rates for an LMTS system can be approximated as being roughly constant. It is therefore reasonable to use the steady state approximations for all but the extremely high (i.e., greater than 0.9) utilization cases, which are unrealistic in practice, anyway.

2.5.5 Another Test

As a second test of the performance of (2.14) and (2.15), we study a last mile transportation system that operates only along two main streets, which intersect at the location of a subway station, as shown in Figure 2.19. The destinations of passengers alighting from the subway are uniformly distributed along the two streets up to a distance

of $b = 150$ from the station. LMTS shuttles, operating in each of the four directions emanating from the station, deliver arriving passengers to their eventual destinations.

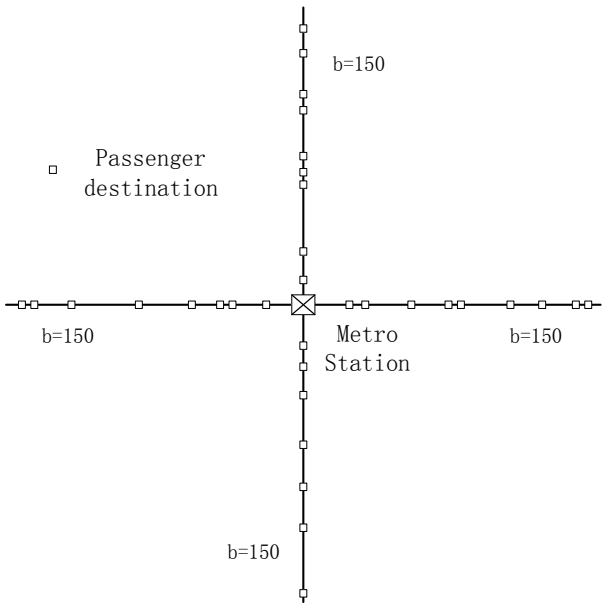


Figure 2.19 Schematic LMTS around crossroad

We have obtained approximate expressions for the expectation and variance of S_E in this case as follows:

$$E(S_E) \approx \frac{2n + c + 1}{2(n+1)} b \qquad Var(S_E) \approx 4 \times \frac{1}{12} b^2 = \frac{1}{3} b^2$$

We have applied these expressions to (2.14) and (2.15) and compared the resulting analytical approximations with results from simulations. For practical waiting times (1 – 5 min), the difference between the simulated average time until boarding and the analytical approximation (2.14) is less than the greater of 15% and 15 seconds, while the difference between the simulated average time until delivery and the analytical approximation (2.15) is less than 15%. The approximations thus also work well for an environment that is very different from that of Sections 2.5.2 – 2.5.4.

2.6 Conclusion

This chapter has developed a set of fully analytical expressions to support the approximate estimation of the performance of a quite general version of a Last-Mile Transportation System (LMTS). Given a lengthy list of inputs about the system's characteristics (headways between arrivals of trains at the station, passenger batch size from each train, number of vehicles in the service fleet, capacity of vehicles, dimensions and travel-related properties of the urban district served), the expressions we have developed estimate the expected waiting time until a passenger can board a vehicle, and the expected time between arrival at the station and delivery to the passenger's destination. A number of simple simulation experiments suggest that these expressions approximate well the expected performance of LMTS under a broad range of conditions typical of what one may encounter in practice.

On the methodological side, the principal contribution of this research is the development of approaches for bounding and approximating the performance of a very difficult type of queueing system involving batch arrivals and requiring the simultaneous consideration of vehicle routing, queueing issues and the use of geometrical probability arguments. The analytical expressions can be very useful in designing LMTS, specifically in determining resource requirements for these systems, such as how many vehicles would be necessary to achieve a specified level of service (as measured by expected time until one boards a vehicle or is delivered to one's destination) and how many kilometers per day these vehicles would travel.

Chapter 3

Operation of a Last Mile Transportation System

3.1 Introduction

In this chapter, we study the operation of a last mile transportation system. The setting of the LMTS is slightly different from that in Chapter 2. As illustrated in Figure 3.1, while a passenger's final destination can be any point in the service region, the LM stops of the vehicles are limited to a finite number of locations which are convenient for the vehicles to load/unload passengers, such as existing public transit stops, entrances of hotels, crossroads near office buildings, and points located close to residential buildings or complexes. The routes and schedules of the vehicles in the LMTS are flexible. They may change over time based on the specific last mile service requests. Essentially, LMTS is an on-demand urban transportation system with batch demands. We describe the setting in more detail in Section 3.2.

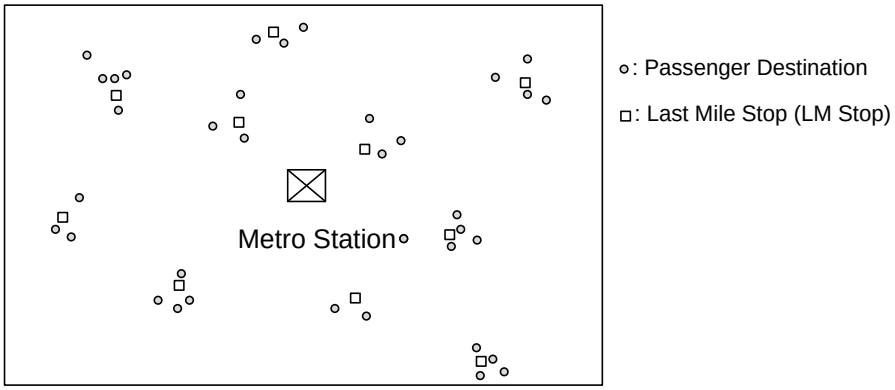


Figure 3.1 Schematic of a Last Mile Transportation System with LM stops

The focus of this chapter is on the system's operations: given the geometric configuration of the service region, the number and location of demand destinations, and the number and capacity of the vehicles, we provide efficient strategies of passenger assignment, vehicle routing and scheduling for operating the LMTS. The objective of the operation is to maximize the LOS, as measured by the average waiting time until a passenger is picked up from the metro station or delivered to her destination.

Addressing these questions is difficult, as the operational decisions generally involve complicated combinatorial optimization problems, e.g., multi-server scheduling of batch demands and numerous service options. As shown in Section 3.5, the exact formulation of the problem is a large-scale Mixed Integer Programming (MIP) model, which is hard to solve optimally or even near-optimally in acceptable computational times.

Routing and scheduling problems have been studied for a long time in Operations Research and a very extensive literature exists. We mention here only a few papers that are among the most influential in the field and especially relevant to our problem. The Vehicle Routing Problem with Time Windows (VRPTW) has been the subject of intensive research using both heuristic and exact optimization approaches. VRPTW considers temporal demand information as in the case of LMTS operations, but with a different objective, namely to minimize the number of vehicles used and/or the total travel distance of vehicles. A good review of the VRPTW literature can be found in Bräysy and Gendreau (2005a, 2005b). Scheduling for multi-server vehicle systems is an important problem which has been studied in diverse contexts. Examples include Liu and Liu (1998), Zee et al. (2001), and Lee et al. (2006). The problems of lot-sizing and scheduling also have similarities with the LMTS operation problem. A survey on these intensely studied problems can be found in Drexl and Kimms (1997).

Our goal is to propose methodologies to efficiently identify feasible solutions of good quality for realistically dimensioned instances of the problem.

The main body of this chapter is organized as follows: In Section 3.2, we describe in detail the operation problem of LMTS we are studying and discuss the associated fundamental assumptions. In Section 3.3, we describe a myopic operating strategy which

can be easily implementable and approximate approaches could be utilized in practical application, especially when limited information about future demand is available. Section 3.4 presents a quick tabu search metaheuristic, which can often improve significantly the solution provided by the myopic operating strategy. Section 3.5 proposes an exact MIP model for the LMTS routing and scheduling problem and a simple two-stage heuristic method to solve it. Section 3.6 defines a set of test instances and performs computational experiments. Section 3.7 contains a summary and concluding remarks.

3.2 Problem Description

We now describe in more detail the problem with reference to Figure 3.1. The setting is slightly different from that in Chapter 2. The LMTS operates as follows: STA denotes the transit metro station served by the LMTS. Any passenger, PAX, who needs last mile service is required to register in a service reservation system (either through a smart-phone application or on a website), indicating the LM stop which is closest to her final destination. As described in Chapter 2, PAX is required to provide advance notice to LMTS of her arrival time at STA, i.e., of the time she will need last mile service. In practical terms, the advance notice could be generated in a number of alternative ways. Each alternative may be associated with a different length of advance notice and a different service horizon. For example, at one extreme, consider the case in which all passengers are regular subscribers, each passenger follows an exact known schedule every day ("request last mile service from STA to LM Stop 1 at 6:00 pm from Monday to Friday") and the metro service is punctual and reliable, then, the LMTS operator has a long advance notice of the service requirements of each passenger and a service horizon that may span a sequence of many metro arrivals ("we shall serve every afternoon about 65 passengers with known destinations who will arrive in the sequence of six metro trains that reach STA between 5 and 6 PM"). In this environment, the LMTS operator will wish to optimize service to the entire (known) set of passengers over the entire service horizon (5 – 6 PM, in this instance). At the opposite extreme, if the service subscribers have a variable schedule from day to day (or if the metro system is crowded and unreliable), service requests may be known only a short time before passengers arrive at STA.

Specifically, PAX could use a smart-phone or tap a smart card on a special-purpose screen to send the service request when she arrives at any station ("ORIGIN") for the purpose of traveling to STA or when she enters her train to STA. The resulting message to the LMTS includes the time of arrival of PAX at STA (easy to predict through the train schedules, once the passenger is at the ORIGIN station or aboard a train) and her ultimate LM stop. Thus, the advance notice is of the order of 10 – 20 minutes and the LMTS operator can plan service ("service horizon") for passengers arriving on only the next very few (perhaps 1 to 3) metro trains.

Since the number of LM stops served by the LMTS is finite, the number of possible vehicle routes (sequences of LM stops in a delivery trip, also referred as route types) is also finite. Based on the service region's geometry, we pre-select a set of feasible routes that are practical in the sense of satisfying some typical constraints, such as the maximum number of LM stops in a single route and the maximum travel distance (or travel time) in a route. For each feasible route, the optimal sequence in which its LM stops are visited (with shortest total travel distance/time as the criterion) and the corresponding travel distance/time to each LM stop are obtained from TSP heuristics. For example, if the maximum number of LM stops in a single route is set to 4 in the LMTS illustrated in Figure 3.2, the blue route is feasible, while the red route is not.

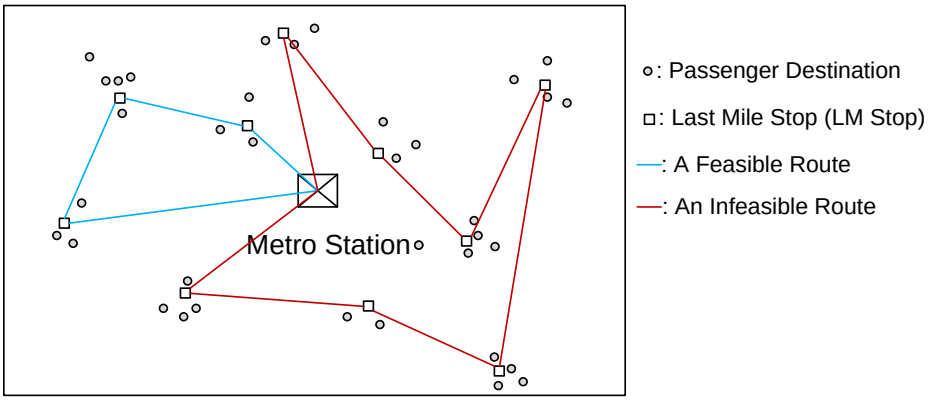


Figure 3.2 Examples of feasible route and infeasible route

In the LMTS operation problem, we determine the routes and schedules of vehicles for delivering all the PAX to their LM stops. The operational decisions include the assignment of each PAX to a vehicle as well as the selection of the route option and

schedule of each vehicle. Drivers of vehicles are provided with a detailed plan that indicates the route, schedule and the number of passengers to each LM stop in every service trip (e.g., "depart STA at 6:36 PM and follow the route Stop 3 - Stop 1 - Stop 2 that delivers 4 passengers to Stop 3, 2 passengers to Stop 1 and 3 passengers to Stop 2"). PAX will receive a message (on her smart-phone or by tapping her smart-card on a screen when she arrives at STA) that indicates the vehicle she has been assigned to, the planned departure time from STA, the planned route, and the planned arrival time at her LM stop (e.g., "board Vehicle #123 which will depart from STA at 6:36 PM; the route of the trip is Stop 3 - Stop 1- Stop 2; you will arrive at your destination LM Stop 1 at 6:41 PM"). The vehicle executes the service trip, visits the planned LM stops in the sequence specified and returns to STA to pick up the passengers for its next planned service trip.

Given the service region's geometry (location of LM stops, feasible routes travel distances, etc.), passenger demand (arrival time at STA and destination LM stops), and the number, capacity and travel speed of the LMTS vehicles, we wish to provide the detailed plan of vehicle operations with the objective of minimizing the passenger waiting time until boarding a vehicle and the passenger riding time.

With reference to Figure 3.2, we make the following assumptions: (i) the LM stops are pre-specified; (ii) the set of feasible routes for the LMTS vehicles are pre-selected; (iii) the train schedules are fixed and known for a specified period of time; (iv) the arrival time and destination LM stop of every passenger (i.e., demand information) are known in advance for a pre-specified period of time; and (v) the delivery fleet consists of m vehicles, each with integer capacity, c.

As noted earlier, the length of time for which the demand is known in advance depends on the practical implementation and service reservation requirements of the LMTS at hand. In this chapter, it is assumed that, once a demand becomes known (whether only 10 minutes in advance of arrival at STA or hours in advance), that demand will materialize exactly as expected. This "deterministic" version of the LMTS operation problem is a reasonable approximation to reality in many contexts. Its solution can also serve as a benchmark for contexts in which large stochastic variability exists.

3.3 Myopic Operation

In this section, we describe a myopic approach for solving the LMTS operation problem. In the myopic approach, we make decisions on the vehicle routes assuming that the demand information is revealed sequentially, one train at a time. This is equivalent to assuming that the last mile service request of a passenger becomes known only at the instant when must then she actually arrives at STA (i.e., the extreme case without passenger advance notice). When a train (batch of passengers) arrives at STA, we consider (i) the new LMTS passengers that arrive on that train, and (ii) any previously unserved passengers who are already waiting for LM service at the station. Under the myopic approach, the LMTS operator specifies delivery routes upon arrival of each train at STA based on the revealed demand information, i.e., the passengers in classes (i) and (ii) at STA. The myopic method is easy to implement in practical terms. It may also be the default solution to the LMTS operation problem in the case of systems where no advance demand information is available to the LMTS operator.

3.3.1 Procedure of Myopic Operation

Let J denote the number of pre-specified LM stops. Let u_j be the number of unserved passengers with LM stop j as their destination, and denote by $U = \{u_1, u_2, \dots, u_J\}$ the set of numbers of unserved passengers to all destinations. Let K be the total number of pre-selected feasible vehicle routes. Let $S = \{k_1, k_2, \dots\}$ denote the set of suggested route types (see Section 3.3.2), where $k_i \in \{1, \dots, K\}$. The procedure of the myopic operation is illustrated in Figure 3.3 and described in Table 3.1.

(0) Whenever a new train arrives at STA:

(1) Update the set of unserved passengers U using the information of the newly arrived passengers.

(2) Empty the set of suggested route types S.

(3) Use a Mixed Integer Programming model (Myopic Formulation (MF), Section 3.3.2) to suggest a set of route types S.

(4) Use ranking criteria (Section 3.3.3) to determine the selection priorities of the route types in S; rank the route types in S in the order of selection priority.

(5) Whenever there are idle vehicles:

(5.1) Dispatch an idle vehicle to provide a service trip of the route type with the highest priority in S.

(5.2) Update the status of passengers.

(5.3) Update the status of vehicles.

(5.4) Delete the selected one from the set of suggested route types S.

Table 3.1 Procedure of myopic operation

The process is repeated every time a train (batch of passengers) arrives until the end of the time horizon for the LMTS operation problem. The waiting time and riding time for each passenger is calculated and reported.

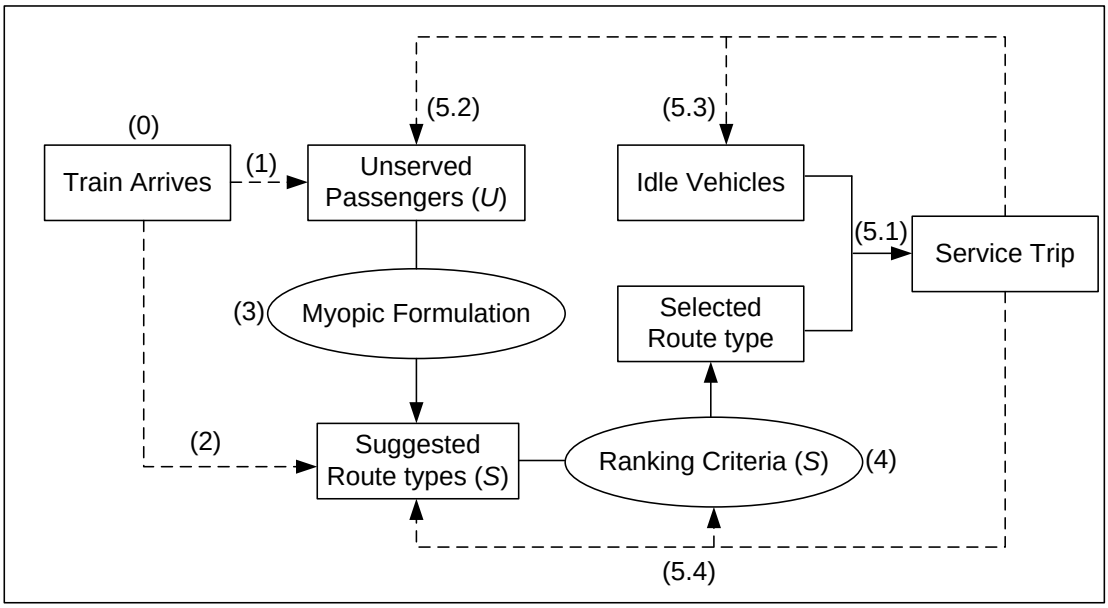


Figure 3.3 Procedure of myopic operation

We use an LMTS operation problem with just 2 train arrivals as an example to demonstrate the myopic operation approach. The headway between the two trains is 5 minutes. We assume the following system settings:

4 LM stops: j_1, j_2, j_3, j_4.

10 feasible route types:

- k_1: serve stop j_1 with total service time of 5 minutes
- k_2: serve stop j_2 with total service time of 3 minutes
- k_3: serve stop j_3 with total service time of 6 minutes
- k_4: serve stop j_4 with total service time of 5 minutes
- k_5: serve stops j_1 and j_2 with total service time of 7 minutes
- k_6: serve stops j_1 and j_3 with total service time of 10 minutes
- k_7: serve stops j_2 and j_3 with total service time of 8 minutes
- k_8: serve stops j_1 and j_4 with total service time of 8 minutes
- k_9: serve stops j_3 and j_4 with total service time of 9 minutes
- k_{10}: serve stops j_1, j_2 and j_3 with total service time of 12 minutes

The number of vehicles $m = 2$ and we assume that both vehicles are available at the time when the first of the two trains arrives.

Vehicle capacity: $c = 5$.

The myopic operation of this LMTS is illustrated in Table 3.2.

(0.1) Train 1 arrives at $t = 0$:

(1) Train 1 brings 3 passengers to stop j_1, 2 passengers to stop j_2, 4 passengers to stop j_3, and 4 passengers to stop j_4:

- $U = \{3,2,4,4\}$

(2) Empty the set of suggested route types:

- $S = \emptyset$

(3) Use the Myopic Formulation (described in Section 3.3.2) to suggest a set of route types:

- $S = \{k_3, k_4, k_5\}$

(4) Use a ranking criterion (described in Section 3.3.3) to determine the selection priority:

- $S = \{k_4, k_3, k_5\}$

(5) Vehicle 1 is idle at $t = 0$:

(5.1) Provide a trip of route type k_4, serving 4 passengers to stop j_4;

(5.2) $U = \{3,2,4,0\}$;

(5.3) Vehicle 1 returns to STA at $t = 5$;

(5.4) $S = \{k_3, k_5\}$;

Vehicle 2 is idle at $t = 0$:

(5.1) Provide a trip of route type k_3, serving 4 passengers to stop j_3;

(5.2) $U = \{3,2,0,0\}$;

(5.3) Vehicle 2 returns to STA at $t = 6$;

(5.4) $S = \{k_5\}$;

(0.2) Train 2 arrives at $t = 5$:

(1) Train 2 brings 2 passengers to stop j_1, 1 passenger to stop j_2, 1 passengers to stop j_3, and 4 passengers to stop j_4:

- $U = \{5,3,1,4\}$

(2) Empty the set of suggested route types:

- $S = \emptyset$

(3) Use the Myopic Formulation (described in Section 3.3.2) to suggest a set of route types:

- $S = \{k_1, k_2, k_9\}$

(4) Use a ranking criterion (described in Section 3.3.3) to determine the selection priority:

- $S = \{k_1, k_2, k_9\}$

(5) Vehicle 1 becomes idle at $t = 5$:

(5.1) Provide a trip of route type k_1, serving 5 passengers to stop j_1;

(5.2) $U = \{0,3,1,4\}$;

(5.3) Vehicle 1 returns to STA at $t = 10$;

(5.4) $S = \{k_2, k_9\}$;

Vehicle 2 becomes idle at $t = 6$:

(5.1) Provide a trip of route type k_2, serving 3 passengers to stop j_2;

(5.2) $U = \{0,0,1,4\}$;

(5.3) Vehicle 2 returns to STA at $t = 9$;

(5.4) $S = \{k_9\}$;

Vehicle 2 becomes idle at $t = 9$:

(5.1) Provide a trip of route type k_9, serving 1 passenger to stop j_3 and 4 passengers to stop j_4;

(5.2) $U = \{0,0,0,0\}$;

(5.3) Vehicle 2 returns to STA at $t = 18$;

(5.4) $S = \emptyset$;

Table 3.2 Example of myopic operation

3.3.2 Myopic Formulation

In Step 3 of the myopic operation approach, a set of route types S is suggested for the service trips that could be provided before next train arrives. The suggestions are based on the details of the passenger demand, i.e., the number of unserved passengers with destination at each LM stop. A Mixed Integer Programming model (Myopic Formulation) is proposed to make the suggestions S.

The notation for the Myopic Formulation is introduced in Table 3.3.

Parameters:

$n_j^{i,U}$: number of unserved passengers with destination at LM stop j before the arrival of train i;

n_j^i : number of passengers with destination at LM stop j brought by train i;

ϕ_{jk} : 1 if LM stop j is served by route type k; 0 otherwise;

t_k : total service time of route type k;

t_{jk} : travel time to LM stop j on route type k;

c : maximum number of passengers served by a vehicle (vehicle capacity);

β_1 : coefficient in the objective function;

β_2 : coefficient in the objective function;

β_3 : coefficient in the objective function;

Decision variables:

z_{jk} : number of passengers with destination at LM stop j assigned to a trip of route type k;

w_k : number of trips of route type k;

m : total number of trips to serve all passengers in the decision epoch.

Table 3.3 Notation for myopic formulation

The Myopic Formulation for the decision epoch of arrival of train i (MF_i) is defined as follows:

$$\min \quad \beta_1 \cdot m + \beta_2 \cdot \sum_k t_k \cdot w_k + \beta_3 \cdot \sum_j \sum_k t_{jk} \cdot z_{jk} \tag{3.1}$$

$$\sum_k z_{jk} \cdot \phi_{jk} = n_j^{i,U} + n_j^i, \quad \forall j \tag{3.2}$$

$$\sum_j z_{jk} \cdot \phi_{jk} \leq c \cdot w_k, \quad \forall k \tag{3.3}$$

$$\sum_k w_k = m, \tag{3.4}$$

$$m, w_k \in \boldsymbol{Z}^*, z_{jk} \in \boldsymbol{R}^*, \quad \forall j, k \tag{3.5}$$

With coefficients $\beta_1 \gg \beta_2 \gg \beta_3$, the objective (3.1) has three hierarchies: (i) first, minimize the number of trips to serve all passengers, (ii) second, minimize the total travel distance/time of vehicles, and (iii) finally, minimize the total travel distance/time of passengers. In practice, the values of β_1, β_2 and β_3 can be adjusted to incorporate different decision preferences.

In the formulation MF_i, any passengers and last mile service requests that appear after train i has arrived are assumed to have no influence on the MF_i decisions. MF_i is essentially suggesting route types and the number for each route type, without considering the passengers who will arrive in the future. This is indeed a "myopic" decision. Constraint (3.2) makes sure that every passenger is assigned to a route type; (3.3) guarantees the vehicle capacity is not exceeded; (3.4) captures the total number of trips needed; (3.5) defines the domains of the decision variables.

When m is set to a fixed value, the mathematical structure of the formulation MF_i is exactly the same as the traditional Capacitated Facility Location Problem (CFLP) if ϕ_{jk} is not taken into consideration in MF_i: w_k is analogous to the location choice of facility and z_{jk} is analogous to the assignment of demand to the chosen facility. CFLP is a well-studied problem. One can refer to a review in Sridharan (1995) for the various heuristic

and exact methods for CFLP. When solving MF_i, we begin by setting the decision variable $m = \lceil (\sum_j (n_j^{i,U} + n_j^i)) / c \rceil$. Because of the limitations of the pre-selected route types and topological relations between the route types and LM stops, MF_i may be infeasible with this initial value of m. We increase m by 1 whenever MF_i is infeasible. When the number of LM stops is not very large (e.g., J does not exceed 20), the corresponding MF_i with fixed value of m (which is similar to the CFLP) can be solved directly and quickly with common commercial optimization software, such as ILOG CPLEX.

3.3.3 Ranking Criterion

The Myopic Formulation MF_i suggests a set of route types that could be carried out in the inter-arrival time between train i and train $i+1$. Before dispatching idle vehicles to provide service trips, we need a ranking criterion to determine the selection priorities of the suggested route types in S.

Generally, a trip with shorter travel time serving more passengers should be given higher priority than a trip with longer travel time serving fewer passengers. Therefore, as a simple criterion, for the route type with $w_k = 1$, we use the value $\sum_j z_{jk} \cdot \phi_{jk} / t_k$ to decide the selection priority, where $\sum_j z_{jk} \cdot \phi_{jk}$ is the total number of passengers that trips of route type k will serve, and t_k is the total service time of the trip of route type k. The suggested route types in S are then ranked and selected in the descending order of $\sum_j z_{jk} \cdot \phi_{jk} / t_k$. When $w_k > 1$, we dispatch passengers to fill up one trip before starting a new one.

3.4 Tabu Search

In this section, we describe a method based on tabu search, a local search metaheuristic that explores the solution space by moving, at each iteration, from the current solution to the best solution in its neighborhood. Tabu search (TS) was proposed and developed by

Glover (see Glover, 1989, 1990a, 1990b), inspired by the principles of artificial intelligence.

The main concepts used in tabu search are: attributes, neighborhood and neighborhood size, moves and evaluation of the moves, tabu list, tabu list size, aspiration criteria, and termination conditions. Tabu search has been applied intensively to various types of routing and scheduling problems, with very good results. Examples of applications include vehicle routing (Gendreau et al, 1994, Cordeau and Maischberger, 2012), multi-purpose machine job scheduling (Hurink et al, 1994), nurse scheduling (Dowsland, 1998), real-time vehicle routing and dispatching (Gendreau et al, 1999), vehicle routing with time windows (Cordeau et al, 2001), split delivery vehicle routing (Archetti et al, 2006), vehicle routing with simultaneous pick-up and delivery service (Montané and Galvao, 2006), and dynamic dial-a-ride (Berbeglia et al, 2012).

In this section, we assume that the demand information for a certain period of time is known before the LMTS operator make operational decisions. The LMTS operator has an advance notice of the service requirements of each passenger and a service horizon that span a sequence of several metro arrivals. In this environment, the LMTS operator will wish to optimize service to the entire (known) set of passengers over the entire service horizon.

In what follows, we first introduce the notation and attributes used in the method. We then provide detailed descriptions of the tabu search concepts for the LMTS operation problem.

3.4.1 Notation and Attributes

In this tabu search metaheuristic, the solution attributes are the route types of the trips initiated during each inter-arrival time of trains (batches of passengers). Let T_i denote the arrival time of train i, and let $h_i = [T_i, T_{i+1})$ denote the inter-arrival time between train i and train $i+1$. Solution s is then represented by $(R_1, R_2, \dots R_I)$, where R_i is the set of route types of the trips initiated during inter-arrival time h_i (R_I is the set of route types of the trips initiated after the arrival of the last train, i.e., train I).

For example, in an LMTS operation problem with 3 trains, a solution s denoted by ($R_1 = \{k_1, k_3\}$, $R_2 = \{k_2\}$, $R_3 = \{k_1, k_4\}$) represents the operation plan in which: vehicle fleet initiates two service trips in the inter-arrival time h_1, and one trip is of route type k_1 and the other is of route type k_3; one service trip of route type k_2 in the inter-arrival time h_2; and two service trips in the time period h_3, one of route type k_1 and the other of route type k_4.

Note that, the solution $s = (R_1, R_2, \dots, R_I)$ represents only the route types of the trips initiated during each inter-arrival time, while the sequence/priorities of route types within each inter-arrival time should be determined by some ranking criteria. We can use, for example, the same ranking criterion used in the myopic operation method described in Section 3.3.3.

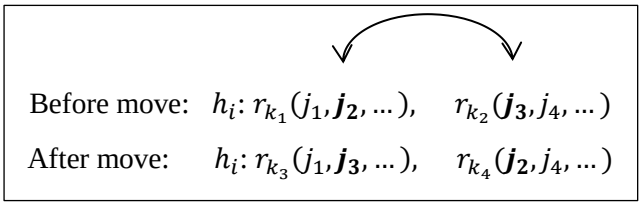
3.4.2 Neighborhood Exploration

Before deciding on the search mechanism to be used, it is important to consider the space (neighborhood) over which the search will be conducted. The LMTS operation problem has two natural neighborhoods. The first, and simplest, involves changing the route types of trips within a single inter-arrival time h_i: the possible changes of route types include swaps of LM stops, shifts, elimination and addition. The second neighborhood involves changing the route types of trips interactively in two consecutive inter-arrival times, h_i and h_{i+1}(or h_{i-1}): the possible changes of route types include swaps and shifts of LM stops.

Let $r_k(j_1, j_2, \dots, j_L)$ denote the service trip of route type k visiting the LM stops j_1, j_2,…, and j_L in that order. Then the details of the possible routing changes, referred to henceforth as “moves” can be described as follows:

Move within a Single Inter-arrival Time h_i

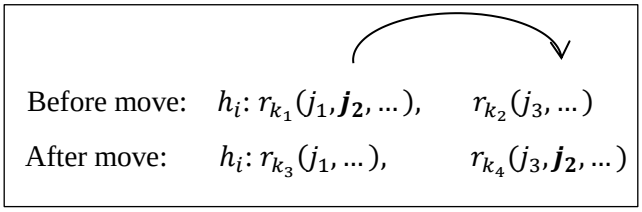
(1) Swap LM stops between two trips.



The move is valid if and only if both the route type k_3 and k_4 are feasible according to the route pre-selection requirements.

We use (i, k, A or E) to denote the route type change of the service trips, where i is the index of inter-arrival times (arrival train i), k is the route type of the changed route, A means route type k is added and E means route type k is eliminated. Therefore, the move (Mv) above is denoted as $Mv = (i, k_1, E) \cap (i, k_2, E) \cap (i, k_3, A) \cap (i, k_4, A)$.

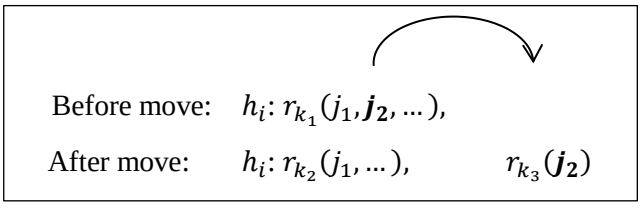
(2) Shift an LM stop from one trip to another.



The move is valid if and only if both the route type k_3 and k_4 are feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i, k_2, E) \cap (i, k_3, A) \cap (i, k_4, A)$.

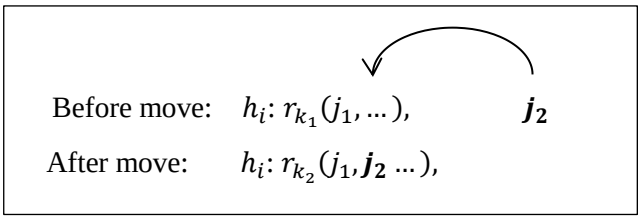
(3) Split a trip into two trips: one serves a single LM stop and the other serves the remaining LM stop(s).



The move is valid if and only if both the route types k_2 and k_3 are feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i, k_2, A) \cap (i, k_3, A)$.

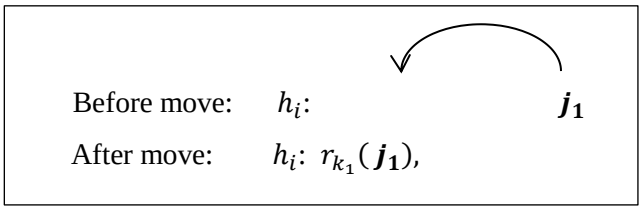
(4) Add an LM stop to a trip.



The move is valid if and only if the route type k_2 is feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i, k_2, A)$.

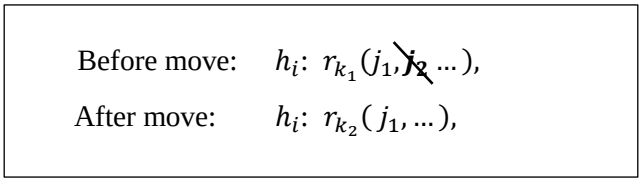
(5) Add a trip serving a single LM stop.



The move is valid if and only if the route type k_1 is feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, A)$.

(6) Eliminate an LM stop from a trip.

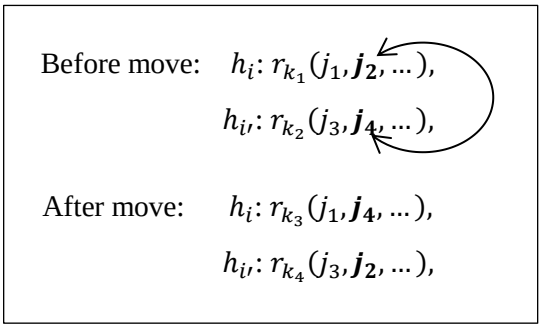


The move is valid if and only if the route type k_2 is feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i, k_2, A)$.

Move involving Two Consecutive Inter-arrival Times, h_i and $h_{i'}$ ($i' = i - 1$ or $i + 1$)

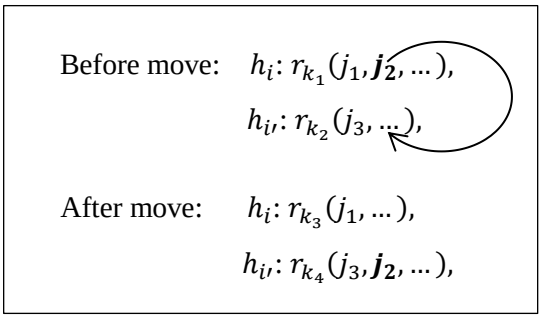
(1) Swap LM stops between a trip in h_i and a trip in $h_{i'}$.



The move is valid if and only if both the route type k_3 and k_4 are feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i', k_2, E) \cap (i, k_3, A) \cap (i', k_4, A)$.

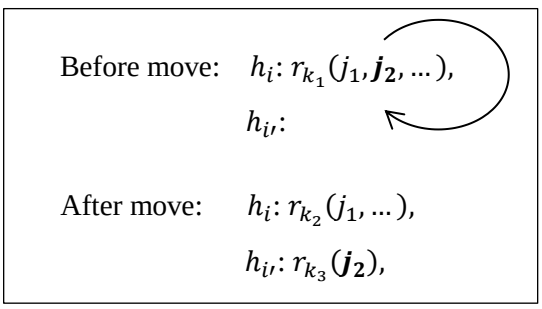
(2) Shift an LM stop from a trip in h_i to an existing trip in $h_{i'}$.



The move is valid if and only if both the route type k_3 and k_4 are feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i', k_2, E) \cap (i, k_3, A) \cap (i', k_4, A)$.

(3) Shift an LM stop from a trip in h_i to a new single LM stop trip in $h_{i'}$.



The move is valid if and only if both the route type k_2 and k_3 are feasible according to the route pre-selection requirements.

The move is denoted as $Mv = (i, k_1, E) \cap (i, k_2, A) \cap (i', k_3, A)$.

With the solution format $s = (R_1, R_2, \dots, R_I)$ in which the sequence of route types in each inter-arrival time is not taken into consideration, it is obvious that any solution (including the optimal solution) can be obtained by imposing a limited number of moves described above on any other solution.

3.4.3 Tabu List

Each move described in Section 3.4.2 contains one or several route type changes, which are denoted as(i, k, A) or (i, k, E). In this chapter, a move is tabu if it contains a change to (i, k, A) (or (i, k, E)) while any move in the tabu list contains a change to (i, k, E) (or (i, k, A)). Essentially, this means that any move that reverses a change of a route type in recent iterations (recorded in the tabu list) is forbidden.

The best size of the tabu list for each kind of problem has to be found empirically. Computational tests need to be implemented for different problems. Some work on similar problems provides observations on good tabu list sizes. For example, Cordeau et al. (1997) observes that the best tabu list size for solving the Periodic Vehicle Routing Problem (PVRP) and Multi-Depot Vehicle Routing Problem (MDVRP) is $7.5\, log_{10}\, n$, where n is the number of customers. Other work has shown experimentally that a tabu list of variable size tends to give better results than a fixed one. For example, Taillard (1991) sets the size of the tabu list to a random number in a specified interval.

For our problem, we have tested several simple tabu list sizes and chosen a fixed size $1 + J/2$ for our algorithm, where J is the number of pre-specified LM stops in the LMTS. A variable size of the tabu list can be easily implemented anyway.

3.4.4 Evaluation of Moves and Aspiration Criteria

The objective value (passenger waiting time + riding time) of the solution $s = (R_1, R_2, \dots, R_I)$ can be obtained if we have a criterion to decide the selection priorities of the route types in each R_i . The evaluation procedure is similar to that in the myopic operation method, while the set of suggested route types S during each inter-arrival time is replaced by R_i in the solution s.

The evaluation procedure is illustrated in Figure 3.4 and described in Table 3.4.

(0) When considering train i:

(1) Update the set of unserved passengers U using the information on the passengers arriving on train i.

(2) Empty the set of suggested route types S.

(3) R_i in solution s is the new set of suggested route types.

(4) Use some ranking criteria (e.g., the same ranking criterion described in the myopic method) to determine the selection priorities of the route types in R_i; rank the route types in R_i in the order of selection priority.

(5) Whenever there are idle vehicles:

(5.1) Dispatch an idle vehicle to provide a service trip of the route type with the highest priority in R_i.

(5.2) Update the status of passengers.

(5.3) Update the status of vehicles.

(5.4) Delete the selected one from the set of suggested route types R_i.

Table 3.4 Evaluation procedure for solution s in tabu search

The process is repeated every time a train (batch of passengers) arrives until the end of the time horizon for the LMTS operation problem. The waiting time and riding time for each passenger is calculated and reported.

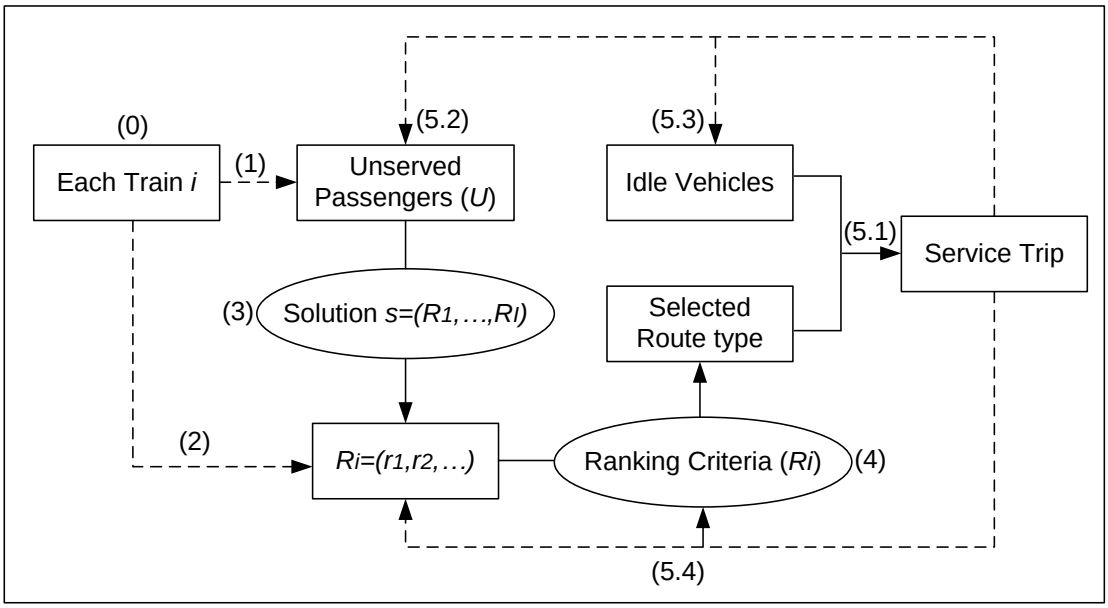


Figure 3.4 Evaluation procedure for solution s in tabu search

Aspiration criteria are the set of criteria that, if satisfied, allow moves that override tabus. In our problem, we allow a move that overrides a tabu, if that move results in an objective value (passenger waiting time and riding time) that is better than the best-known objective value so far.

3.4.5 Termination Conditions

Our termination rule is based on a limit to the maximum number of total iterations and a limit to the maximum number of iterations without solution improvement. The search terminates if a maximum number of total iterations (N_1) is reached or if the best solution so far has not been improved for a certain number of iterations (N_2). After tests with a number of computational experiments, we have set $N_1 = 500$ and $N_2 = 50$ when we use the solution from the myopic approach as the initial solution in the tabu search.

3.4.6 Tabu Search Algorithm

Based on the concepts discussed in Section 3.4.1 – Section 3.4.5, the tabu search algorithm is described in Table 3.5

---

0. Obtain an initial solution s_0 from another method, e.g., myopic operation method
   Set best solution $s^* = s_0$
   Set current solution $s^c = s_0$
   Tabu list $TL = \emptyset$
1. REPEAT:
   IF termination condition is satisfied
   THEN
   STOP;
   ELSE
   1.1 For each neighbor in the neighborhood of s^c, DO
   Calculate the objective value using the evaluation procedure
   1.2 Select and move to the best neighbor, which is not tabu or satisfies the aspiration criteria
   1.3 Update s^*, s^c and TL

Table 3.5 Tabu search algorithm

3.5 Mixed Integer Programming Formulation

In this section, we present an exact Mixed Integer Programming (MIP) model for the LMTS operation problem described in Section 3.5.1. In this section, we also assume the LMTS operator has an advance notice of the service requirements of each passenger and a service horizon that span a sequence of several metro arrivals. The LMTS operator will wish to optimize service to the entire (known) set of passengers over the entire service horizon. The large scale of the formulation makes it difficult to solve the MIP quickly. In Section 3.5.2 and Section 3.5.3 we propose a two-stage heuristic to solve the MIP model approximately.

3.5.1 Exact MIP model

First, we introduce the following some additional notation in Table 3.6:

Parameters:	
n_j^t	: number of passengers with destination at LM stop *j* arriving at station at time t; obtained from train schedules and the LM service reservation system;
β_w	: weight of passenger waiting time until boarding in the objective function;
β_r	: weight of passenger in-vehicle riding time in the objective function;
Decision variables:	
z_{jk}^t	: number of passengers with destination at LM stop *j* that board a vehicle initiating a trip of route type k at time t;
w_k^t	: number of trips of route type k initiated at time t;
Intermediate variables:	
r_j^t	: number of unserved passengers with destination at LM stop *j* waiting at the train station at the end of time t;
v^t	: number of available vehicles at the end of time t;

Table 3.6 (Additional) Notation for the exact MIP model

In this formulation, we discretize the time into intervals of one minute, so we can approximate what will happen in practice.

The objective function (3.6) is defined as minimizing the weighted summation of the time spent by all passengers in the LMTS: (i) waiting time until boarding a vehicle and (ii) in-vehicle riding time:

$$\min \quad \beta_w \cdot \sum_t \sum_j r_j^t + \beta_r \cdot \sum_t \sum_j \sum_k t_{jk} \cdot z_{jk}^t \tag{3.6}$$

For example, since we discretize the time into one minute intervals and r_j^t counts the unserved passengers at the end of time period t, if $r_j^t = 20$ for some t, this means that 20 minutes of waiting were added during the minute t for passengers going to stop j.

If $\beta_w = \beta_r$, the objective is to minimize the summation of passenger waiting time and riding time, i.e., the time from the arrival of passengers at STA to delivery at their destination LM stops.

The formulation has the following constraints:

a) Passenger flow constraints: expressions (3.7), (3.8) and (3.9) define and constrain the number of unserved passengers with destination at each LM stop waiting in the station at the end of each time t.

$$r_j^0 = n_j^0 - \sum_k z_{jk}^0 \cdot \phi_{jk}, \quad \forall j \tag{3.7}$$

$$r_j^t = r_j^{t-1} + n_j^t - \sum_k z_{jk}^t \cdot \phi_{jk}, \quad \forall j, t \geq 1 \tag{3.8}$$

$$r_j^t \geq 0, \quad \forall j, t \tag{3.9}$$

b) Vehicle flow constraints: expressions (3.10), (3.11) and (3.12) define and constrain the number of available vehicles waiting at the station at the end of each time t.

$$v^0 = m - \sum_k w_k^0, \tag{3.10}$$

$$v^t = v^{t-1} + \sum_k w_k^{t-t_k} - \sum_k w_k^t, \quad \forall t \geq 1 \tag{3.11}$$

$$v^t \geq 0, \quad \forall t \tag{3.12}$$

c) Service capacity constraints: expression (3.13) guarantees the vehicle service capacity restriction.

$$\sum_j z_{jk}^t \cdot \phi_{jk} \leq c \cdot w_k^t, \quad \forall k, t \tag{3.13}$$

d) Domains of decision variables.

$$w_k^t \in \boldsymbol{Z}^*, z_{jk}^t \in \boldsymbol{R}^* \quad \forall j,k,t \tag{3.14}$$

The model described above deviates from a large class of traditional vehicle routing and scheduling problems. Note that, (i) the model considers the routes and schedules of a multi-vehicle fleet, (ii) the temporal demand for a set of LM stops is known, (iii) a set of feasible routes are pre-selected beforehand, and (iv) the performance evaluation metric is the waiting time and riding time of passengers, instead of the travel distance/time of vehicles. These characteristics make the model important for other applications. Similar models can be applied very usefully in many other contexts, particularly in problems involving multi-server systems that provide flexible service options.

It is hard to obtain optimal or even near-optimal solutions for the MIP model described above. In Section 3.5.2 and Section 3.5.3, we provide a two-stage heuristic to solve the formulation approximately, aiming at solutions of good quality.

3.5.2 First Stage: Solve MIP to the Level of Inter-arrival Time

In the first stage, we modify the original exact MIP model by replacing the time dimension t in the decision variables z_{jk}^t and w_k^t with the train's ID i. In other words, instead of making detailed decisions for every minute t in the original formulation, we shift our focus to making more aggregate decisions for every inter-arrival time h_i. The modified model can reduce the problem scale and the required computational time.

The notation is modified as in Table 3.7.

Parameters:

n_j^i : number of passengers with destination at LM stop j arriving on train i;

Decision variables:

z_{jk}^i : number of passengers with destination at LM stop j that board a vehicle initiating a trip of route type k during the inter-arrival time h_i;

w_k^i : number of trips of route type k initiated during the inter-arrival time h_i;

Intermediate variables:

r_j^i	: number of unserved passengers with destination at LM stop *j* waiting at the station at the end of inter-arrival time h_i;

Table 3.7 (Additional) Notation for the first stage model

The objective function (3.15) captures (i) part of passenger waiting time until boarding vehicle and (ii) all the passenger in-vehicle riding time:

$$\min \quad \beta_w \cdot h_i \cdot \sum_i \sum_j r_j^i + \beta_r \cdot \sum_i \sum_j \sum_k t_{jk} \cdot z_{jk}^i \tag{3.15}$$

The modified formulation has the following constraints:

a) Passenger flow constraints: expressions (3.16), (3.17), and (3.18) are directly modified from expressions (3.7), (3.8), and (3.9), respectively. The dimension of time t is replaced with train i.

$$r_j^1 = n_j^1 - \sum_k z_{jk}^1 \cdot \phi_{jk}, \quad \forall j \tag{3.16}$$

$$r_j^i = r_j^{i-1} + n_j^i - \sum_k z_{jk}^i \cdot \phi_{jk}, \quad \forall j, i \geq 2 \tag{3.17}$$

$$r_j^i \geq 0, \quad \forall j, i \tag{3.18}$$

b) Vehicle flow constraints: we cannot capture the detailed vehicle schedules without decision variables w defined at every time t; instead, we use some heuristic constraints to make the vehicle schedules roughly feasible. Constraints (3.19) and (3.20) limit the total number of trips initiated within one and two consecutive inter-arrival times, respectively; constraints (3.21) and (3.22) limit the total service time of trips initiated within one and two consecutive inter-arrival times, respectively. The values of the upper limits $m_{\max 1}$, $m_{\max 2}$, $t_{\max 1}$ and $t_{\max 2}$ could be set as their realized values in solution to other approaches, for example, the myopic operation approach.

$$\sum_k w_k^i \leq m_{\max 1}, \quad \forall i = 1, 2, \ldots, I \tag{3.19}$$

$$\sum_k (w_k^i + w_k^{i+1}) \leq m_{\max 2}, \quad \forall i = 1, 2, \dots, I-1 \tag{3.20}$$

$$\sum_k t_k \cdot w_k^i \leq t_{\max 1}, \quad \forall i = 1, 2, \dots, I \tag{3.21}$$

$$\sum_k t_k \cdot (w_k^i + w_k^{i+1}) \leq t_{\max 2}, \quad \forall i = 1, 2, \dots, I-1 \tag{3.22}$$

c) Service capacity constraints: expression (3.23) is modified from expression (3.13).

$$\sum_j z_{jk}^i \cdot \phi_{jk} \leq c \cdot w_k^i, \quad \forall k, i \tag{3.23}$$

d) Domains of decision variables.

$$w_k^i \in \boldsymbol{Z}^*, z_{jk}^i \in \boldsymbol{R}^* \quad \forall j, k, i \tag{3.24}$$

Essentially, the modified formulation in the first stage uses the inter-arrival time between trains as the smallest time unit to make decisions. However, unlike the myopic operation approach, the modified formulation does take into consideration the mutual interactions that exist among demands from all the trains.

3.5.3 Second Stage: Column Generation in the Original Formulation

The solution to the modified formulation in the first stage suggests the service route types that could be provided in each inter-arrival time h_i. In the second stage, we implement the original exact MIP model proposed in Section 3.5.1 with the decision variables (columns) generated using the information revealed in the first stage solution. Specifically, if $w_k^i > 0$ in the optimal solution of the first stage problem, we generate vehicle decision variables $w_{k'}^t$, for: (i) every time $t \in h_i$; and (ii) every route type k' which is a sub-tour of route type k, including route type k itself.

For example, if a route type k serving three LM stops is selected for the inter-arrival time h_i in the first stage solution, as shown in Figure 3.5, we generate decision variables $w_{k'}^t$, for: (i) every time t in $h_i = [T_i, T_{i+1})$; and (ii) the $2^3 - 1 = 7$ specific route types, where each route type serves a subset of the three LM stops, as shown in Figure 3.6.

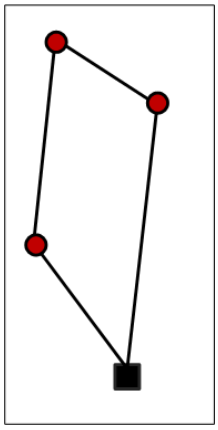


Figure 3.5 Route type selected in the first stage

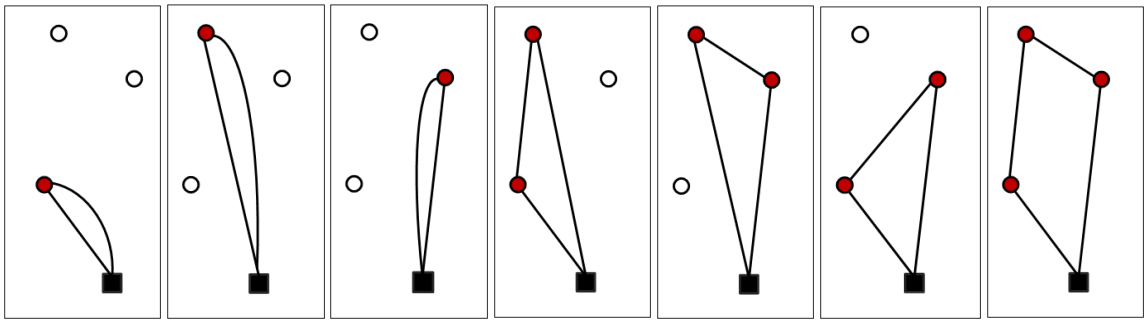


Figure 3.6 Route types of decision variables generated in the second stage

With the decision variables generated as described above, the exact MIP model in Section 3.5.1 can be solved in Stage 2 in much less computational time.

3.6 Computational Study

We present a computational study based on the approaches described in Section 3.3 to Section 3.5. We compare the results of the myopic operating strategy, the tabu search metaheuristic, and the MIP model, which is solved approximately in two stages. A conventional service with fixed routes and schedules is taken as a benchmark to evaluate the performance of the LMTS with each of these three different operating strategies. The computational experiments are coded in Java and run on 64-bit computers with 2.9 GHZ

processors and 4GB RAM. All the corresponding MIP problems are solved using ILOG CPLEX (version 12) with a time limit of 5 minutes for each instance.

We first discuss the settings of test instances for the computational experiments in Section 3.6.1. We then describe in Section 3.6.2 a common multi-vehicle conventional transportation system with fixed routes and schedules that will serve as our benchmark of comparisons. Finally, Section 3.6.3 presents our computational results, followed by a brief discussion.

3.6.1 Settings of Test Instances

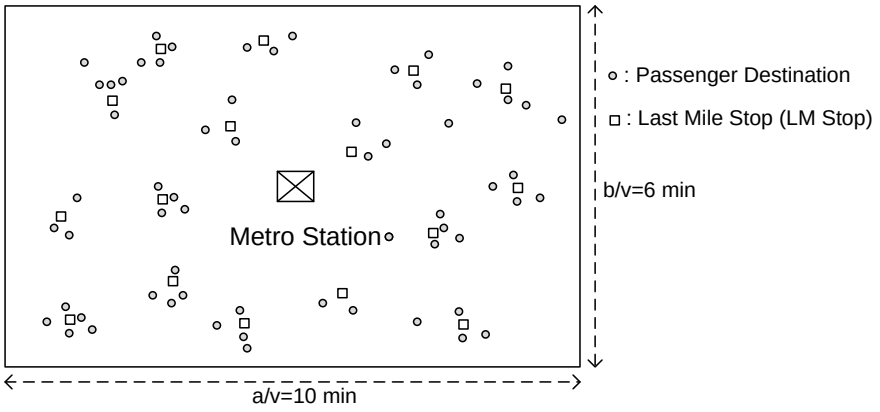


Figure 3.7 Schematic of the LMTS in the computational experiments

We consider an LMTS in a rectangular service region illustrated in Figure 3.7. To travel through the region's length and width, LMTS vehicles require 10 minutes and 6 minutes, respectively. The number of pre-specified LM stops (assumed to be uniformly distributed in the region) is denoted as J and takes values ranging from 8 to 12 in the experiments. Assuming the vehicles travel according to the Euclidean metric and the service time at each LM stop (e.g., time for vehicle deceleration, loading/unloading of passengers and vehicle acceleration) is set to 1 minute. The feasible routes are then selected to satisfy the requirements that (i) the maximum number of LM stops in a route is 3, and (ii) the maximum total service time (travel time + service time at stops) in a route is 14 minutes. The number of feasible routes, K, under such conditions will be in the region of 100 to 300.

Trains with passengers arrive at the metro station every 10 minutes. The size of a passenger batch from each train is assumed to be Poisson-distributed with intensity N taking values ranging from 10 to 30. Passenger destinations are assumed to be distributed among the LM stops either (a) uniformly, or (b) slightly heterogeneously or (c) extremely heterogeneously. An example of demand intensity at LM stops in a case with $J = 8$ and $N = 16$ is shown in Table 3.8.

$J = 8, N = 16$	Ratio of highest demand over lowest demand	Demand intensity at LM stops
Uniform (UN)	2.0/2.0 = **1**	2.0 for all
Slightly heterogeneous (SH)	3.0/0.5 = **6**	0.5, 1.5, 1.5, 2.0, 2.0, 2.5, 3.0, 3.0
Extremely heterogeneous (EH)	4.0/0.2= **20**	0.2, 0.6, 1.0, 1.8, 2.2, 2.8, 3.4, 4.0

Table 3.8 Demand intensity at LM stops

A fleet of vehicles with capacity taking values ranging from 3 to 12 serves the LMTS. Experiments with different fleet sizes are used to evaluate the performance of LMTS in situations of high and low vehicle utilization.

We have selected the above parameters so that the system would make sense physically. For each combination of parameter settings, we carry out 10 test instances.

3.6.2 Conventional Service with Fixed Routes and Schedules

In order to study the potential advantages of on-demand service in LMTS, we introduce a commonly used multi-vehicle conventional transportation system with fixed routes and schedules as a benchmark of comparisons. In the conventional system, each vehicle in the fleet follows a pre-designed fixed bus route. We apply a simple integer programming model to design the bus routes and service. As general design guidance, we require that each LM stop be served by at least one bus route, the service capacity provided to each LM stop is roughly proportional to its passenger demand rate, and we aim to minimize the total service time of all bus routes while keeping these service times roughly similar. Table 3.9 introduces the relevant notation.

Parameters:

t_d : upper limit on the difference between the service times of different bus lines;

c_d : upper limit on the difference between general service capacities provided to each LM stop;

λ_j : passenger demand rate to LM stop j;

M : a large positive number;

Decision variables:

x_k : binary integer variable to indicate whether route type k is selected as a bus line in the conventional service;

Table 3.9 (Additional) Notation for bus line design for conventional service

The integer programming model for the bus lines is defined as follows:

$$\min \quad \sum_k t_k \cdot x_k \tag{3.24}$$

$$\sum_k \phi_{jk} \cdot x_k \geq 1, \quad \forall j \tag{3.25}$$

$$\sum_k x_k = m, \tag{3.26}$$

$$\frac{\sum_k c \cdot \phi_{j_i k} \cdot x_k / (d_k \cdot t_k)}{\lambda_{j_i}} - \frac{\sum_k c \cdot \phi_{j_s k} \cdot x_k / (d_k \cdot t_k)}{\lambda_{j_s}} \leq c_d, \tag{3.27}$$

$$\forall j_i, j_s \in \{1, \dots, J\}$$

$$t_{k_i} - t_{k_j} \leq M \cdot \left(2 - x_{k_i} - x_{k_j}\right) + t_d, \quad \forall k_i, k_j \in \{1, \dots, K\} \tag{3.28}$$

$$x_k \in \{0,1\}, \qquad \forall k \tag{3.29}$$

Objective function (3.24) is to minimize the total service time of the selected bus lines; constraint (3.25) makes sure that every LM stop is served by at least one bus line;

constraint (3.26) makes sure that the number of bus lines equals the number of available vehicles in the fleet; constraints (27) ensures the general service capacity difference between any pair of LM stops does not exceed the upper limit; constraints (3.28) ensures the service time difference between any pair of selected bus lines does not exceed the upper limit; constraint (3.29) defines the domains of the binary decision variables.

3.6.3 Results and Discussion

Tables 3.10 – 3.14 display the objective values (passenger waiting time + riding time) and the computational time associated with different operating strategies with diverse parameter settings. UN denotes uniform demand among LM stops, SH denotes slightly heterogeneous demand (the ratio of the highest demand over the lowest demand is 6), and EH denotes extremely heterogeneous demand (the ratio is 20). The objective value is in minutes and the running time is in seconds for every instance.

$J = 8$, $N = 16$, $c = 6$, $m = 3$	UN		SH		EH	
	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)
Conventional	13.08		14.67		18.78	
Myopic	7.53	0.6	7.16	0.4	6.38	0.5
Myopic + Tabu	5.36	0.6 + 4	5.35	0.4 + 7	4.77	0.5 + 6
MIP (2-stage)	**5.09**	244	5.13	172	**4.43**	167
MIP (2-stage) + Tabu	**5.09**	244 + 11	**5.10**	172 + 7	4.49	167 + 6

Table 3.10 Results for $J = 8, N = 16, c = 6, m = 3$

$J = 8$, $N = 16$, $c = 6$, $m = 7$	UN		SH		EH	
	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)
Conventional	9.18		10.49		6.82	
Myopic	4.16	0.4	3.65	0.4	3.82	0.4
Myopic + Tabu	3.62	0.4 + 8	2.93	0.4 + 9	3.32	0.4 + 15
MIP (2-stage)	**3.56**	126	2.81	167	**3.24**	115
MIP (2-stage) + Tabu	**3.56**	126 + 9	**2.80**	167 + 11	**3.24**	115 + 6

Table 3.11 Results for $J = 8, N = 16, c = 6, m = 7$

$J = 12,$ $N = 30,$ $c = 6,$ $m = 5$	UN		SH		EH	
	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)
Conventional	18.47		26.19		32.19	
Myopic	9.09	1	8.62	1	7.12	0.7
Myopic + Tabu	6.93	1 + 33	6.58	1 + 42	5.51	0.7 + 31
MIP (2-stage)	**5.95**	448	**5.95**	443	**4.94**	334
MIP (2-stage) + Tabu	6.25	448 + 29	6.23	443+30	5.20	334 + 32

Table 3.12 Results for $J = 12, N = 30, c = 6, m = 5$

$J = 12,$ $N = 30,$ $c = 6,$ $m = 7$	UN		SH		EH	
	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)
Conventional	9.51		11.79		18.82	
Myopic	5.00	1	4.29	0.8	4.07	0.7
Myopic + Tabu	4.17	1 + 63	3.81	0.8 + 49	3.57	0.7 + 50
MIP (2-stage)	4.17	275	3.81	296	**3.50**	301
MIP (2-stage) + Tabu	**4.07**	275 + 60	**3.77**	296+48	3.55	301+74

Table 3.13 Results for $J = 12, N = 30, c = 6, m = 7$

$J = 8,$ $N = 20,$ $c \times m = 24,$ UN	$c = 12, m = 2$		$c = 8, m = 3$		$c = 4, m = 6$	
	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)	Objective value (min)	Running time (sec)
Conventional	12.55		12.83		14.05	
Myopic	8.48	0.7	8.02	0.4	4.36	0.3
Myopic + Tabu	5.73	0.7 + 2	5.65	0.4 + 8	3.82	0.3 + 7
MIP (2-stage)	5.51	311	5.28	211	**3.71**	232
MIP (2-stage) + Tabu	**5.26**	311 + 2	**5.27**	211 + 6	3.75	232 + 6

Table 3.14 Results for $J = 8, N = 20, c \times m = 24$, demand is UN

The conventional service is taken as a benchmark. The myopic operating strategy, which can be implemented easily in LMTS, could reduce the passenger waiting time and riding time significantly compared to the conventional service. For example, in the uniform demand case (UN) in Table 3.8, the passenger waiting time and riding time in the LMTS myopic strategy is only 57.6% of that in the conventional service. In addition, it can be seen that the advanced operating strategies, i.e., the tabu search metaheuristic

and the MIP model provide even better routing and scheduling solutions. In the same instances in Table 3.10, the tabu search, using the myopic solution as its initial solution, provides an objective value which is only 41.0% of that for the conventional service; the MIP model solved approximately in two stages provides an objective value which is only 38.9% of the conventional service. However, the tabu search that uses the MIP solution as its initial solution does not achieve much improvement, which can be seen as evidence of the high-quality of the MIP solution itself.

In terms of computational time, the myopic operating strategy can provide solutions in seconds; the computational time of the tabu search metaheuristic depends on the parameters in the search termination conditions: if the maximum total number of iterations (N_1) is 500 and the maximum number of iterations without improvement (N_2) is 50, the tabu search takes a computational time ranging from several seconds in small cases (small J and K) to 1 minutes in large cases (large J and K); the MIP method requires the longest computational times.

The advantage of flexible LMTS over conventional bus service is greater when vehicle capacity is small than when vehicle capacity is large. Table 3.14 displays results for systems with same geometric configuration, same passenger demand, and equal total vehicle capacity ($c \times m = 24$) in the fleet. The improvements of the best solutions achieved by LMTS are 58%, 59%, 73%, for $= 12$, $c = 8$, and $c = 4$ respectively, compared to conventional service.

In the case of low demand and low utilization of vehicles, it is highly probable that the batch of passengers from any particular train can be served before the next train arrives. In these conditions, the performance of the myopic operating strategy is quite close to that of the tabu search or of the MIP method. Stated differently, information about passengers in the "next" train does not help much in improving system performance, if most passengers on the "current" train can be served before the "next" train arrives. By contrast, when demand and vehicle utilization are high, it is more beneficial to apply advanced methods such as the tabu search metaheuristic and the MIP method as these methods consider simultaneously demands from all trains, i.e., take advantage of information about demand from "future trains". For example, Table 3.10

and Table 3.11 have the same system parameters, with the exception of the number of vehicles. Table 3.10 shows a high utilization situation with 3 vehicles, while Table 3.11 a low utilization one with 7 vehicles. In the UN case, the tabu search metaheuristic (or the MIP method) reduces the passenger waiting time and riding time compared to myopic operating strategy by 13.0% (or 14.4%) in the low utilization situation, while the improvement is 28.8% (or 32.4%) in the high utilization situation. The same trend can be found in Tables 3.12 and 3.13: the improvement is 16.6% (or 18.6%) in the low utilization situation and 23.8% (or 34.5%) in the high utilization situation.

Figures 3.8 and 3.9 compare some of the detailed characteristics of the solutions shown in Table 3.12 for the UN case. Taking conventional service with fixed routes and schedules as the benchmark, it is obvious (Figure 3.8) that all the non-naïve methods for operating LMTS provide better service, i.e., of reduced passenger waiting time and riding time. Additionally, the LMTS achieves reduced total vehicle service time (Figure 3.9). Specifically, the tabu search metaheuristic starting from the myopic solutions tends to provide operating plans with good service quality and the least vehicle service time. The MIP method provides operating plans with the best service quality, which result from customized routes that typically use a larger number of trips than the other methods (Figure 3.8).

The results of Tables 3.10 – 3.14 also suggest that LMTS may provide significant cost savings, for both the service users (passengers) and the service providers (e.g., government, private company). First, less waiting time and riding time in LMTS actually means monetary savings for its users. According to Gómez-Ibañez et al. (1999), for work trips in San Francisco, the monetary value of a unit transfer waiting time is 195% of the users' after-tax wages, and the monetary value of a unit in-vehicle riding time is 76% of the users' after-tax wages. The equivalent economic savings are very large when we consider these monetary values of time. Second, the reduced vehicle service time achieved by LMTS also means monetary savings for its operators. Given a fixed size of vehicle fleet, the operation cost of the transportation system is largely proportional to the vehicle service time, e.g., the labor cost of the drivers is positively correlated with the vehicle service time, and the fuel cost of the vehicles directly depends on the vehicle travel time/distance.

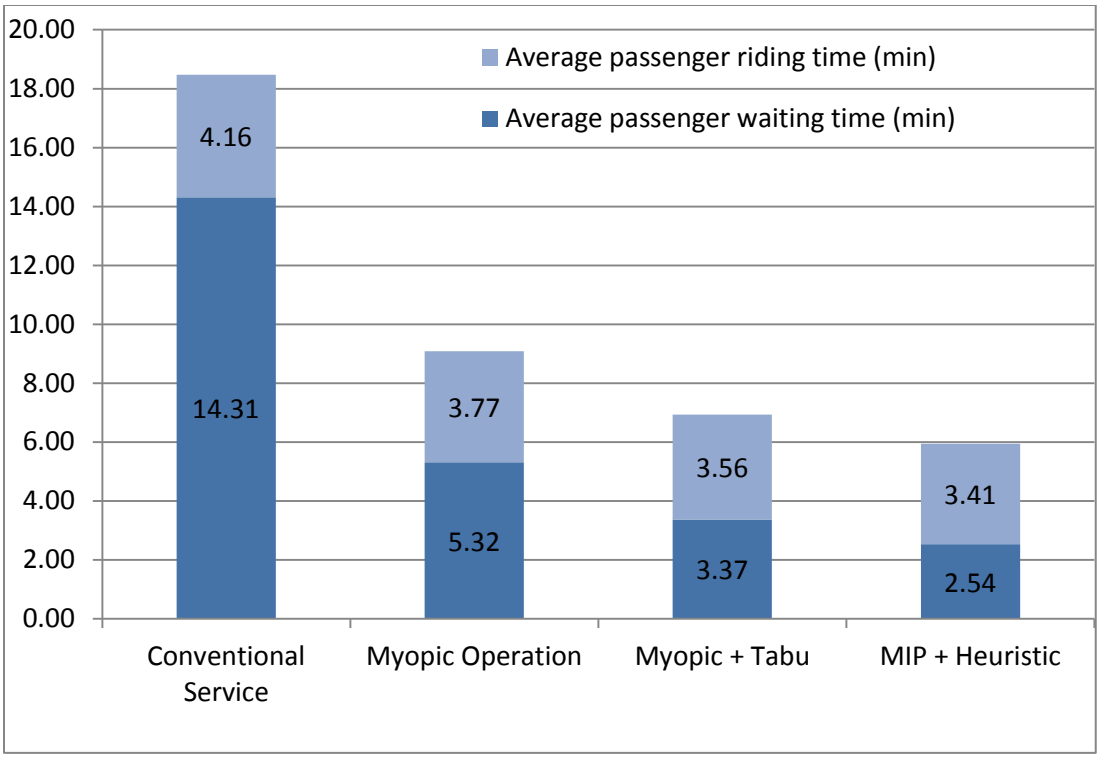


Figure 3.8 Passenger waiting time and riding time for $J = 12, N = 30, c = 6, m = 5$

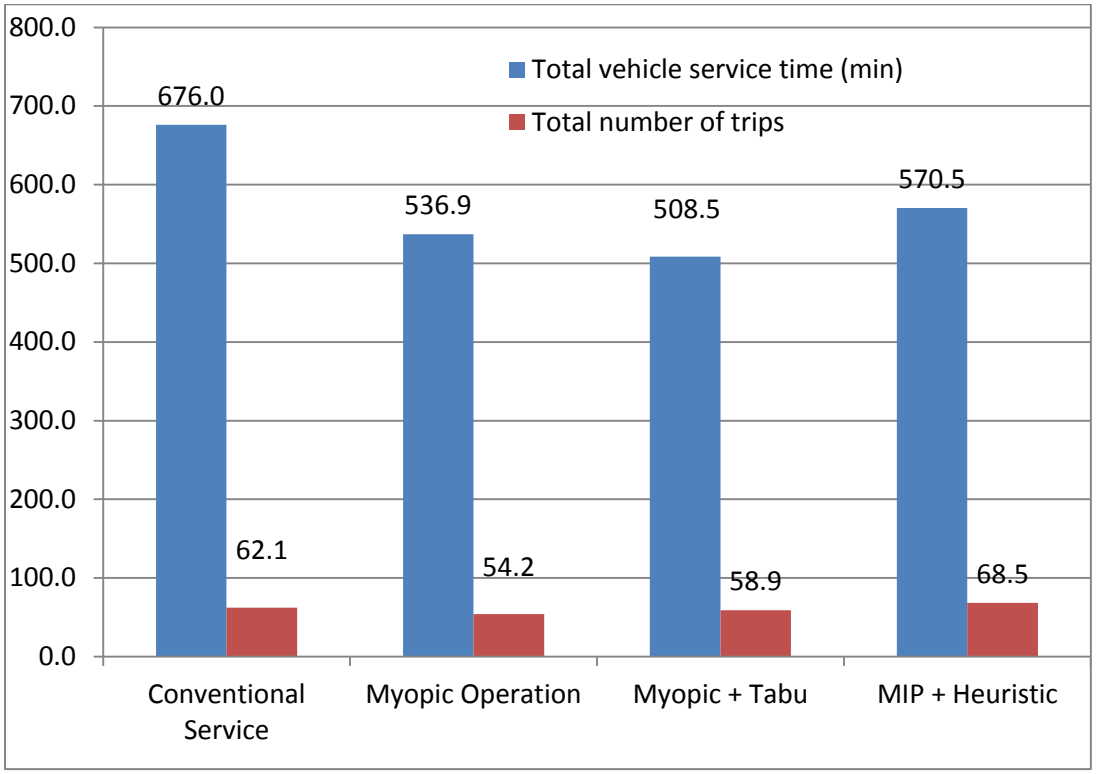


Figure 3.9 Vehicle service time and number of trips for $J = 12, N = 30, c = 6, m = 5$

3.7 Conclusion

This chapter has proposed several methods for operating the vehicle fleet of a Last Mile Transportation System (LMTS). Given the geometry of the service region (pre-specified LM stops, feasible routes), the number of vehicles in the service fleet, the capacity of the vehicles, and a set of known LM service requests (arrival time and destination of passengers), the operation methods we have developed provide the detailed routing and scheduling plan for the vehicle fleet, with the objective of minimizing passenger waiting time until boarding a vehicle and passenger riding time on a vehicle. Computational experiments suggest that, compared to a conventional service system with fixed routes and schedules, an LMTS operating with any non-naive method, such as the myopic operating strategy, the tabu search metaheuristic, and the MIP model, performs better under a broad range of conditions. The myopic operation method can be implemented easily and quickly; the tabu search metaheuristic provides solutions of good quality in a short computational time; the MIP method provides the best solution, but with the greatest computational requirements.

On the methodological side, the principal contribution of this research is the development of approaches for the routing and scheduling operation of a very difficult type of problem involving batch arrivals, batch service and multiple servers. On the practical side, we believe that the operation methods we have proposed can be very useful for LMTS, providing good operating plans for these complex systems.

Chapter 4

A New Perspective to Study Passenger Utility Functions – Core Determining Class: Construction, Approximation and Inference

Fast expanding data availability in a variety of industries is bringing unprecedented opportunities to the Operations Research community. Huge volumes of sales data, service records, or even instant messages, are generated in this modern era of "big data" through the wide utilization of information and communications technologies. The data allows us to infer critical aspects of service systems and markets, such as service-specific demand distributions and hidden customer preferences. This also makes it possible to pursue promising improvements in many areas, such as better planning strategies, more efficient operations, customized products, and accurate marketing.

The field of transportation and logistics provides an excellent example of this major development with fast expanding data availability in every segment and mode of the sector. With the wide application of information technologies and devices, such as sensors deployed along streets/roads, portable GPS devices, personal location and timing record systems, and electronic payment tools, the characteristics and information corresponding to many aspects of a transportation trip, including the average/maximum vehicle speed, passenger waiting time, in-vehicle riding time, distance from origin to final destination, and travel cost/fee, can be measured, collected and stored easily. The

possibility to collect and utilize the data creates opportunities to study transportation systems in very different ways and perspectives. In this Chapter, we provide a new perspective for studying passenger utility functions, taking advantage of increasing data availability. This also offers an example of how hidden information can be inferred from observed data in a context in which only some correlations are known between unobservable events and observed outcomes. This is a key problem in the era of "big data".

From this point on, this Chapter presents joint work with Mr. Ye Luo, a PhD candidate in the Department of Economics in MIT. The contributions of the thesis author is (1) a combinatorial algorithm to construct the Core Determining Class (defined in Section 4.3) in the case without observation data noise; (2) a linear optimization formulation to select an approximated Core Determining Class in the case with observation data noise; (3) and a set of numerical experiments to evaluate the statistical property of the formulation.

4.1 Overall Approach

Utility is the (perceived) ability of something to satisfy needs or wants. In general, it represents the satisfaction experienced by the consumer of a good. Not coincidentally, a good is something that satisfies human wants and provides utility. For a passenger, who is the consumer of a transportation service, the good is the transport service provided. The utility to passengers is derived through the benefits, satisfaction or happiness attained as a result of arrival at their final destinations. The passenger utility function, a function to measure and quantify the utility experienced by the passengers in a transportation service, plays an important role in understanding and estimating the distribution of the passenger demand. This is critical information that can be used by the service provider to design and operate a transportation system.

Utility is very difficult to observe and measure directly and utility functions are typically estimated and calibrated in various indirect ways. In this Chapter, we study the estimation of utility functions from a very new perspective, significantly different from

existing ones. We provide an alternative angle for calibrating the probability measures of the unknown parameters in passenger utility functions.

As an overall approach in this Chapter, (1) we treat the unknown parameters (which need calibration) in the passenger utility function as unobserved events, and the observed characteristics of transportation trips, such as passenger waiting time, in-vehicle riding time and monetary travel cost, as observed outcomes; then, (2) we construct a bipartite graph representing the relationships between the events and the outcomes; and finally, (3) we propose a general method for identifying the probability measures of the events given the observations of the frequencies of the outcomes, including a combinatorial algorithm in which the data noise of the observations is ignored and a general procedure in which data noise is taken into consideration.

The closest researches to our topic are Galichon and Henry (2006, 2011) and Chesher and Rosen (2012). Galichon and Henry (2011) propose the Core Determining Class problem, i.e., finding the minimum set of inequalities to describe the feasible region of probability measure on unobserved events. Chesher and Rosen (2012) provide an inequality selection algorithm, but may still contain some redundant inequalities in the selected set. Andrews and Soares (2010) propose moment inequality selection procedure using criterions such as BIC.

There are many studies on performing inference of sets. Chernozhukov et al. (2007, CHT) proposes general inference procedure with moment inequality constraints. Romano and Shaikh (2010) provide improvements for CHT (2007). Beresteanu et al. (2011) use random set theory to perform inference with convex inequality restrictions. Andrews and Shi (2013) construct inference based on conditional moment inequalities. For related empirical studies, see Manski and Tamer (2002), Bajari et al. (2007), Bajari et al. (2010) and etc..

We address the redundancy of linear constraints in the problem of identification of probability measures. There is also a wide literature on detection and elimination of redundant constraints when data noise is not taken into consideration. For example, Telgen (1983) develops two methods to identify redundant constraints and implicit

equalities. Caron et al. (1989) present a degenerate extreme point strategy which classifies linear constraints as either redundant or necessary. Paulraj et al. (2006) propose a heuristic approach using an intercept matrix to identify redundant constraints.

The main body of this Chapter is organized as follows. In Section 4.2, we construct in detail a bipartite graph relating the unobserved events (unknown parameters in the utility function) and the observed outcomes (observed characteristics of transportation trips), assuming a reasonably simple form of the passenger utility function. Section 4.3 introduces the general model and basic assumptions of the Core Determining Class. Section 4.4 studies the Core Determining Class from the structure of the bipartite graph and provides a combinatorial algorithm to construct the exact Core Determining Class when data noise of observations is not taken into consideration. Section 4.5 proposes a general linear inequality selection procedure under noisy data with the definition of sparse assumptions. Section 4.6 discusses the additional technical assumptions and proves the main theorems of the statistical properties of the selection procedure, with application in the Core Determining Class problem. Section 4.7 provides some concluding remarks.

4.2 Example of Bipartite Graph

In this Chapter, we construct a bipartite graph as an example to demonstrate the relationships between unobserved events and observed outcomes. We assume a linear form of the passenger utility function:

$$u = d - \beta \cdot (\alpha t_w + t_r) - \gamma \cdot p \tag{4.1}$$

where

d: distance from origin to final destination (observed);

t_r: passenger in-vehicle riding time (observed);

t_w: passenger waiting time before boarding (observed);

p: travel cost (observed);

α: weight to capture different perceptions of waiting time and riding time (obtained from other sources);

β: unknown non-negative parameter (target);

γ: unknown non-negative parameter (target);

A basic assumption in the example is that, a transportation trip will be chosen (by the passenger), realized (by the service provider), and then observed (by any observers), if and only if a passenger derives from the trip a utility with non-negative value:

$$u = d - \beta \cdot (\alpha t_w + t_r) - \gamma \cdot p \geq 0$$

which is equivalent to,

$$\gamma + \beta \cdot \frac{\alpha t_w + t_r}{p} \leq \frac{d}{p}$$

If we denote $(\alpha t_w + t_r)/p$ as c_1 and d/p as c_2, then a trip will be observed if and only if expression (4.2) is satisfied.

$$\gamma + \beta \cdot c_1 \leq c_2 \qquad (4.2)$$

For any combination of c_1 and c_2, there is a set of combinations of β and γ that satisfies expression (4.2). If (a) we discretize the continuous values of c_1, c_2, β and γ to discretized segments, (b) define the combination of a β and γ discretized segment as an unobserved event, and (c) define the combination of a c_1 and c_2 discretized segment as an observed outcome, then we can construct a correspondence mapping to represent the relationships between the events and the outcomes which may generate non-negative utility for passengers and realize the transportation trip. The correspondence mapping is a bipartite graph. A link between an event and an outcome means, the event (combination of β and γ) together with the outcome (combination of c_1 and c_2) satisfies expression (4.2) and makes the trip possible to be realized and observed.

As a simple example, we assume the range of β is [0.1,1] and discretize it into two segments: [0.1,0.5) and [0.5,1]; we assume the range of γ is $[0.1,1] \cup \{0\}$ and discretize it into three segments: $\{0\}, [0.1,0.5)$ and [0.5,1] (where $\gamma = 0$ means the passenger is not sensitive to the travel cost for some reason, e.g., the travel cost is fully reimbursed by his/her employer). Let u denote events, which are the Cartesian product of a β and a γ segment. Then there are 6 events:

u_1: $\beta \in [0.1,0.5) \cap \gamma \in \{0\}$;

u_2: $\beta \in [0.1,0.5) \cap \gamma \in [0.1,0.5)$;

u_3: $\beta \in [0.1,0.5) \cap \gamma \in [0.5,1]$;

u_4: $\beta \in [0.5,1] \cap \gamma \in \{0\}$;

u_5: $\beta \in [0.5,1] \cap \gamma \in [0.1,0.5)$;

u_6: $\beta \in [0.5,1] \cap \gamma \in [0.5,1]$;

As for the outcomes, we assume the range of c_1 is [1,5] and discretize it into two segments: [1,3) and [3,5]; and the range of c_2 is [0,2] and discretize it into three segments: $[0,0.5), [0.5,1)$ and $[1,2]$. Let y denote outcomes, which are the Cartesian product of a c_1 and a c_2 segments. Then there are 6 outcomes:

y_1: $c_1 \in [1,3) \cap c_2 \in [0,0.5)$;

y_2: $c_1 \in [1,3) \cap c_2 \in [0.5,1)$;

y_3: $c_1 \in [1,3) \cap c_2 \in [1,2)$;

y_4: $c_1 \in [3,5] \cap c_2 \in [0,0.5)$;

y_5: $c_1 \in [3,5] \cap c_2 \in [0.5,1)$;

y_6: $c_1 \in [3,5] \cap c_2 \in [1,2)$;

A link between an event u (combination of β and γ) and an outcome y (combination of c_1 and c_2) is added to the bipartite graph if there exist parameter values of this

combination of u and y such that expression (4.2) is satisfied. The corresponding bipartite graph is illustrated in Figure 4.1. With the bipartite graph and the observations of the frequencies of the outcomes (c_1 and c_2), we can apply the methodologies described in Section 4.3 to 4.6 to identify the probability measures of the hidden events, i.e., the unknown parameters β and γ in the passenger utility function in which we are interested.

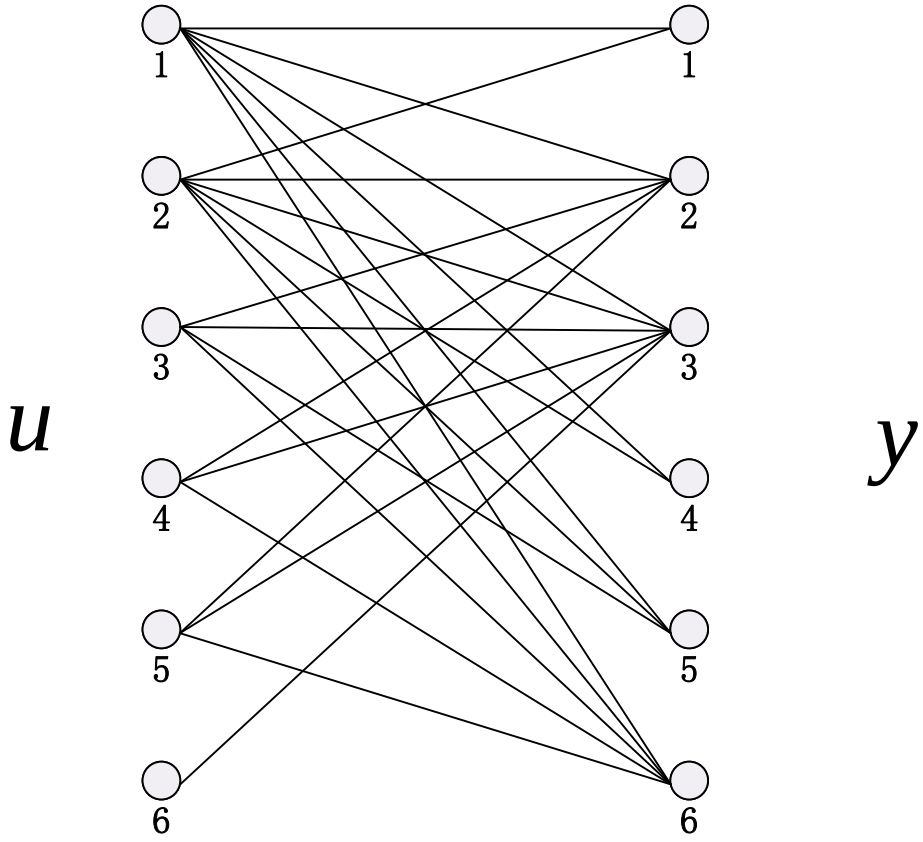


Figure 4.1 Example of a bipartite graph

4.3 Core Determining Class

Given a bipartite graph $G = (U, Y, \varphi)$, suppose U is a set of vertices representing events, and Y is a set of vertices representing outcomes. Suppose an event $u \in U$ leads to a set of possible outcomes $\varphi(u) \in Y$, where $\varphi(u)$ is a set of vertices in Y. For any set $A \subset U$, $\varphi(A) := \cup_{u \in A} \varphi(u)$. Therefore, $\varphi: 2^U \to 2^Y$ is a correspondence mapping between U and

Y. The inverse of φ, denoted as φ^{-1}, is defined as φ^{-1}: $2^Y \to 2^U, \varphi^{-1}(B) := \{u \in U | \varphi(u) \cap B \neq \emptyset\}, \ \forall B \subset Y$.

Let v be the probability measure on U. Let $\mu_{n,0}$ be the true measure on Y which could change with the model. Let $\hat{\mu}_n$ be the measure observed in a sample set of outcomes Y. Denote $d_1 = |U|$ and $d_2 = |Y|$. For a bipartite graph $G = (U, Y, \varphi)$, we say G is connected if $\forall A_1, A_2 \subset G$ and $A_1 \cup A_2 = G$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$.

Assumption: 4.1 [Non-Degeneracy of G, $\mu_{n,0}$ and $\hat{\mu}_n$] (1) Assume G is connected: we say G is non-degenerate if G is connected; (2) for the probability measure $\mu = \mu_{n,0}$ or $\hat{\mu}_n$, assume that for any $y \in Y$, $\mu(y) > 0$: we say that μ is non-degenerate if $\mu(y) > 0$ for any $y \in Y$.

We assume that Assumption 4.1 holds throughout this Chapter. The parameter of interest is the $d_1 \times 1$ vector v, which is the probability measure which generates the events $u \in U$. In general, we are unable to obtain a point estimation of v unless additional information is provided. Instead, we aim to obtain inequality bounds on v given the bipartite graph $G = (U, Y, \varphi)$ and the measure μ on Y. More specifically, for any set of events $A \subset U$, the probability measure on the corresponding outcomes should fall into the set $\varphi(A)$. Thus, for any $A \subset U$, we can obtain the inequality $v(A) := \sum_{u \in A} v(u) \leq \mu\big(\varphi(A)\big) := \sum_{y \in \varphi(A)} \mu(y)$.

The Artstein's theorem stated in Artstein (1983) presents that all information of v in the biparte graph model $G = (U, Y, \varphi)$ can be characterized by the set of linear inequality constraints described in Lemma 4.1.

Lemma 4.1 [Artstein's Theorem] The following set of inequalities/equalities contains sharp information of v:

(1) $\forall A \subset U$, $v(A) := \sum_{u \in A} v(u) \leq \mu(\varphi(A))$, where $\mu\big(\varphi(A)\big) := \sum_{y \in \varphi(A)} \mu(y)$;

(2) $\sum_{u \in U} v(u) = 1$.

Our model, denoted as P_G, is presented in Definition 4.1.

Definition 4.1 [Model P_G] Finding the set of all feasible probability measure v on U such that:

$$\forall A \subset U, v(A) \leq \mu(\varphi(A)) \tag{4.3}$$

$$\sum_{u \in U} v(u) = 1 \tag{4.4}$$

The non-degeneracy assumption prevents the model P_G from decomposition, i.e., we cannot decompose graph G into sub-graphs G_1 and G_2 and proceed with sub-problems P_{G_1} and P_{G_2}. Otherwise the problem can be simplified by looking at G_1 and G_2 separately.

In general, the set of inequality constraints stated in Definition 4.1 contains redundant inequalities. Define the minimum model T_0 of P_G as the set of linear inequality constraints stated in expression (4.3) such that T_0 together with the equality (4.4) has a minimum number of constraints which generate the same set of feasible measure as P_G. In other words, T_0 consists of all irredundant constraints in P_G. If the number of irredundant constraints in T_0 is much less than $2^{d_1} - 1$ stated in Definition 4.1, then it is more accurate and computational efficient to conduct inference on the Core Determining Class (defined in Definition 4.2) using T_0. Galichon and Henry (2011) propose the concept "Core Determining Class" as follows.

Definition 4.2 [Core Determining Class problem] The Core Determining Class problem is the problem of finding all binding constraints in model P_G. The Core Determining Class is any collection of subsets of U that contains the sharp information on v, i.e., the corresponding inequalities includes all binding inequalities. The exact Core Determining Class is defined as the set of subsets of U which corresponds to the irredundant inequalities in the model T_0.

The definition of Core Determining Class in Galichon and Henry (2006) is slightly different from the definition in this Chapter. Galichon and Henry (2006) define Core-Determining Class as any set that contains all the binding inequalities. In this Chapter, we refer "exact Core Determining Class" as the set of all binding inequalities, i.e., the smallest (in cardinality) set which characterizes the identified set of the parameters of interest.

In many cases, there may exist a parametric model for v, which is denoted as $v_i = F_i(\theta)$. The function F_i can be non-linear. The inference problem on θ is generally difficult if the number of inequalities about v is large. Therefore, if we can find the truly binding inequalities about v, we would perform estimation and inference on θ much faster.

In reality, the true probability measure $\mu_{n,0}$ on the outcome set Y is unobservable. Instead, given some data, we could observe the empirical measure $\hat{\mu}_n$ on Y. Due to uncertainty of the data, we would like to solve a relaxed problem P_G', the solution set of which covers the solution set of the true model P_G with probability approaching 1 as the sample size n of the data approaching infinity.

The relaxed problem P_G' provides conservative inference for the model P_G.

Definition 4.3 [Model P_G']. For a small λ, finding the set of all feasible probability measure v on U such that:

$$\forall A \subset U,\ v(A) \coloneqq \sum_{u \in A} v(u) \leq \hat{\mu}_n\big(\varphi(A)\big) + \lambda \tag{4.5}$$

$$\sum_{u \in U} v(u) = 1 \tag{4.6}$$

Ideally λ should converge to 0 when $n \to \infty$. The dimensionality of the problem, $|U|$, and the number of inequalities in P_G', should affect the tuning parameter λ. In fact, λ should be chosen properly such that: (1) the feasible set of v found in model P_G' covers the feasible set of v found in model P_G with probability approaching 1, so P_G' provides reliable inference on P_G; and (2) λ is not be too large to exaggerate the feasible set of v found in model P_G'. We will discuss the choice of λ in Section 4.6.

According to the Artstein's theorem, model P_G contains $2^{d_1} - 2$ inequalities, which is a very large number when d_1 is large and even grows with n in some contexts. The numerous inequalities lead to both computational difficulties and undesirable statistical properties. In fact, some or even most of the inequalities stated in the Artstein's theorem may be redundant. Galichon and Henry (2011) analyze the monotonic structure of the graph G and claim that there are at most $2d_1 - 2$ sets in the Core Determining Class

under a special structure. Chesher and Rosen (2012) provide an algorithm which could get rid of some, but not necessarily all, redundant inequalities. In Section 4.4, we fully characterize the Core Determining Class by exploring the combinatorial structure of the bipartite graph G. We prove that the Core Determining Class only rely on the structure of graph G under the non-degeneracy assumption of μ and G. The results are novel compared to existing studies. We also propose a fast algorithm in Section 4.4 to compute the exact Core Determining Class when data noise of observation is not taken into consideration.

In addition, besides the redundant inequalities, many of the binding inequalities could be "nearly" redundant, meaning that although they are informative in model P_G with empirical $\hat{\mu}$, they could be "implied" by other inequalities in model P_G' with a small relaxation λ. Therefore, it may be possible to use a smaller number of inequalities, i.e., a "small" model, to approximate the full and exact one. Such a smaller model will enjoy better statistical properties compared to the full model, i.e., it will be less sensitive to modeling errors. In Section 4.5, we propose a general inequality selection procedure similar to the Dantzig Selector in regression to obtain a smaller model.

4.4 Exact Core Determining Class

In this section, we present our discovery of the combinatorial structure of the Core Determining Class, along with a fast algorithm to generate the Core Determining Class. In Galichon and Henry (2011), whether an inequality $v(A) \leq \mu(\varphi(A))$ is in the Core Determining Class is examined by numerical computations using the probability measure μ.

In fact, given the correspondence mapping φ of the bipartite graph $G(U, Y, \varphi)$, we can identify the redundant inequalities without any observations of the outcomes in Y. For example, for $A_1 \in U$ and $A_2 \in U$, if $A_1 \cap A_2 = \emptyset$ and $\varphi(A_1) \cap \varphi(A_2) = \emptyset$, then the two inequalities, $v(A_1) \leq \mu(\varphi(A_1))$ and $v(A_2) \leq \mu(\varphi(A_2))$ can generate the inequality $v(A_1 \cup A_2) = v(A_1) + v(A_2) \leq \mu\big(\varphi(A_1)\big) + \mu\big(\varphi(A_2)\big) = \mu\big(\varphi(A_1) \cup \varphi(A_2)\big) = \mu\big(\varphi(A_1 \cup A_2)\big)$, which is exactly the inequality corresponding to the set $A = A_1 \cup A_2$. In

other words, the inequality $v(A) \leq \mu(\varphi(A))$ with $A = A_1 \cup A_2$ is redundant given $v(A_1) \leq \mu(\varphi(A_1))$ and $v(A_2) \leq \mu(\varphi(A_2))$. Also, if $u \notin A$ satisfies $\varphi(\{u\}) \subset \varphi(A)$, then the inequality $v(\{u\} \cup A) \leq \mu\big(\varphi(\{u\} \cup A)\big) = \mu(\varphi(A))$ will imply a redundant inequality $v(A) \leq \mu(\varphi(A))$.

In this section, we propose a combinatorial method to generate the exact Core Determining Class. We prove that, in theory, if the probability measure μ is non-degenerate, our method excludes all redundant inequalities in the model P_G regardless the values of μ. That is to say, the Core Determining Class can be exactly constructed through combinatorial method and it is independent from μ.

Definition 4.4 [Set S_U] $S \subset 2^U$ is the collection of all non-empty subsets $A \subset U$ and $A \neq U$, such that $v^M(A) > \mu\big(\varphi(A)\big)$,

where $v^M(A) \coloneqq \max\{v(A) | v(A') \leq \mu\big(\varphi(A')\big), \forall A' \subset U, A' \neq A\}$.

Set S_U is defined using probability measure μ. The inequality generated by any $A \in S_U$ is informative: it is irredundant given other inequalities described in expression (4.3). Essentially, S_U identifies the irreducible inequalities for model P_G when the critical equality $\sum_{u \in U} v(u) = 1$ is not taken into consideration.

Definition 4.5 [Set S'_U] $S' \subset 2^U$ is the collection of all non-empty subsets $A \subset U$ and $A \neq U$, such that:

(1) A is self-connected, i.e., $\forall A_1, A_2 \subset U$ such that $A_1, A_2 \neq \emptyset$ and $A_1 \cup A_2 = A$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$;

(2)There exists no $u \in U$, such that $u \notin A$ and $\varphi(u) \subset \varphi(A)$.

Lemma 4.2. If μ is non-degenerate, the collection of subsets defined in Definition 4.4 and Definition 4.5 are identical, i.e., $S_U = S'_U$.

Proof: see Appendix A.1

S_U and S'_U describe the irreducible inequalities in P_G if the equality $\sum_{u \in U} v(u) = 1$ is not taken into consideration. Theorem 5 of Chesher and Rosen (2012) proposes a subset

of inequalities with property (1) stated in Definition 4.5. This subset contains the set of all binding inequalities, which is Core Determining. Lemma 4.2 shows that with an additional property (2) in Definition 4.5, we can find all binding inequalities when the equality $\sum_{u\in U} v(u) = 1$ is ignored. In fact, adding this equality can further substantially reduce the number of inequalities and it is impossible to find the minimum set of inequalities in P_G without the key equality $\sum_{u\in U} v(u) = 1$.

To find the minimum binding set of inequalities, i.e., the exact Core Determining Class, we further look at the problem P_G from the opposite direction: consider the inequalities from Y to U. For any non-degenerate probability measure $\tilde{v}$ on U, we define S_Y and S_Y', which are collections of subsets of Y and similar to S_U and S_U'.

Definition 4.6 [Set S_Y] Given a non-degenerate probability measure $\tilde{v}$ on U, $S_Y \subset 2^Y$ is the collection of all subsets $B \subset Y$ and $B \neq Y$, such that $\mu^M(B) > \tilde{v}\left(\varphi^{-1}(B)\right)$,

where $\mu^M(B) \coloneqq \max\{\tilde{\mu}(B) | \tilde{\mu}(B') \leq \tilde{v}\left(\varphi^{-1}(B')\right), \forall B' \subset Y, B' \neq B\}$ and $\tilde{\mu}$ is a probability measure on Y.

Definition 4.7 [Set S_Y'] $S_Y' \subset 2^Y$ is the collection of all subsets $B \subset Y$ and $B \neq Y$, such that:

(1) B is self-connected, i.e., $\forall B_1, B_2 \subset B$, such that $B_1, B_2 \neq \emptyset$ and $B_1 \cup B_1 = B$, it holds that $\varphi^{-1}(B_1) \cap \varphi^{-1}(B_2) \neq \emptyset$;

(2) There exists no $y \in Y$, such that $y \notin B$ and $\varphi^{-1}(y) \subset \varphi^{-1}(B)$.

The Lemma below presents result similar to Lemma 4.2.

Lemma 4.3 $S_Y = S_Y'$

The proof is similar to the proof of Lemma 4.2.

Definition 4.8 [Set S_Y^{-1}] S_Y^{-1} is the collection of $A \subset U$ and $A \neq U$ such that there exists $B \subset S_Y'$ satisfying $A = \varphi^{-1}(B)^c$.

Below we give a numerical definition of the exact Core Determining Class using linear programming.

Definition 4.9 [Set S^*] The Core Determining Class S^* is the collection of all subsets $A \subset U$ and $A \neq U$, such that $v^{M^*}(A) > \mu(\varphi(A))$,

where $v^{M^*}(A) \coloneqq \max\{v(A) | v(A') \leq \mu\big(\varphi(A')\big), \forall A' \subset U, A' \neq A; v(U) = 1\}$.

In Definition 4.9, the equality $v(U) = 1$ is taken into consideration. S^* contains subsets in U corresponding to irreducible inequalities under $v(U) = 1$. The theorem below characterizes and constructs the Core Determining Class S^*.

Theorem 4.1. The Core Determining Class S^* is characterized by the following equation:

$$S^* = S_U \cap S_Y^{-1}$$

Proof: See Appendix A.2

Notice that both S_U' and S_Y' are defined via combinatorial rules, so S_U and S_Y^{-1} can be found via combinatorial rules and the Core Determining Class S^* is independent from μ if μ is non-degenerate. In an example followed, we show that considering only S_U may not be able to substantially reduce the number of inequalities, when $S_U \cap S_Y^{-1}$ can be a very small set in cardinality.

Consider the set $U = \{u_1, \dots, u_{d_1}\}$ and the set $Y = \{y_1, \dots, y_{d_1+1}\}$. φ is the correspondence mapping between U and Y such that $\varphi\big(u_j\big) = \{y_j, y_{j+1}\}$ for all $1 \leq j \leq d_1$. The correspondence mapping is illustrated in Figure 4.2.

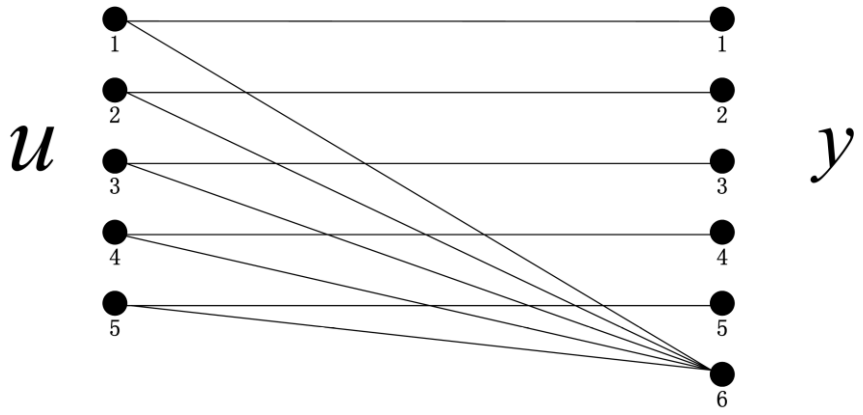


Figure 4.2 Correspondence mapping of an example

If we only consider S_U, we can obtain $S = 2^U - \{\emptyset, U\}$, which consists of $2^{d_1} - 2$ subsets and essentially makes no selection of inequalities. The Core Determining Class S^* constructed in Theorem 4.1 is $\{U - u_j | 1 \leq j \leq d_1\}$. It is obvious that it is the minimum number of subsets carrying full information on v for model P_G. The Core Determining Class S^* contains d_1 inequalities, which is much less than the $2^{d_1} - 2$ inequalities selected by Theorem 5 in Chesher and Rosen (2012).

Therefore, we utilize the combinatorial structure revealed in Definition 4.5 and Definition 4.7 to construct the sets S'_U and S'_Y: algorithm illustrated in Figure 4.3 generates set S'_U and a similar algorithm generates set S'_Y, and then set S_Y^{-1}. Then we can obtain the exact Core Determining Class $S^* = S_U \cap S_Y^{-1}$.

input : Bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$

output: Set $\mathcal{S}'_u$

Initiation: $\mathcal{S}'_u = \{\emptyset\}$

for $i \leftarrow 0$ **to** $|\mathcal{U}| - 1$ **do**

  Identify additional $A' \in \mathcal{S}'_u$ *as union of* $u \in \mathcal{U}$ *and* $A \in \mathcal{S}'_u$ *with* $|A| = i$,

  foreach $A \in \mathcal{S}'_u$ *with* $|A| = i$ **do**

    foreach $u \notin A$ *that* $\varphi(u) \cap \varphi(A) \neq \emptyset$ **do**

      $A' \leftarrow A \cup \{u\}$,

      if $\varphi(A') < 1$ **then**

        foreach $u' \notin A'$ **do**

          if $\varphi(u') \subset \varphi(A')$ **then** $A' \leftarrow A' \cup u'$;

        end

        if $A' \notin \mathcal{S}'_u$ **then** $\mathcal{S}'_u \leftarrow \mathcal{S}'_u \cup \{A'\}$;

      end

    end

  end

end

Termination: $\mathcal{S}'_u \leftarrow \mathcal{S}'_u - \{\emptyset\}$

Figure 4.3 Algorithm to generate set S'_U

The complexity of the algorithm is $O(2^{\max\{d_u, d_y\}} \cdot d_1^2 \cdot d_2^2)$, where

d_u is defined as

$$d_u \coloneqq \max_A |A|$$

$$s.t. A \subset U$$

$$\varphi(A) = Y$$

$$\varphi(A/u) \subsetneq Y, \forall u \in A$$

d_y is defined as

$$d_y \coloneqq \max_B |B|$$

$$s.t. B \subset Y$$

$$\varphi^{-1}(B) = U$$

$$\varphi^{-1}(B/y) \subsetneq U, \forall y \in B$$

Under the assumption of non-degenerate G and μ, in a bipartite graph with practical applications, d_u and d_y is much smaller than d_1 and d_2 respectively, so the algorithm is fast in practice.

4.5 A General Selection Procedure and Sparse Assumption

Essentially, the objective of model P_G is to obtain a feasible set of measure v given the observation $\hat{\mu}$, i.e., to obtain $\hat{Q} \coloneqq \{v|v(A) \leq \hat{\mu}_n(\varphi(A)), \forall A \subset U; \sum_{u \in U} v(u) = 1\}$. In Section 4.4 we explore the structure of the bipartite graph G to obtain the set of irreducible inequalities to define $\hat{Q}$. In this section, we propose a procedure for a general problem of linear inequality selection with data noise of $\hat{\mu}$. This procedure chooses the set of linear inequalities with sharp information on v as $n \to \infty$. It can identify the inequalities which are binding but "close" to redundant, so to further reduce the number of inequalities in $\hat{Q}$. The procedure can be applied to general linear inequality selection problems, including the Core Determining Class problem allowing mixed strategy as defined in Galichon and Henry (2011).

4.5.1 General Selection Procedure

Problem *P* can be interpreted as computing the feasible region of a collection of linear inequality constraints. It could be generalized as computing the feasible region of

$$Q \coloneqq \{v|Mv \leq b, v \geq 0\},$$

where M is a $m \times d_1$ matrix, v is a $d_1 \times 1$ vector, and b is a $m \times 1$ vector.

In many situations, the number of inequalities, m, is too large for us to effectively conduct any known estimation and inference procedures such as the CHT inference in Chernozhukov et al. (2007). For example, there are $m = 2^{d_1}$ inequalities in the Core Determining Class problem in Section 4.3 without implementation of any inequality selection procedures (we could view $v(U) = 1$ as two inequalities: $v(U) \leq 1$ and $v(U) \geq 1$).

There are two reasons that we do not use the entire set of the 2^{d_1} inequalities: first, there could be many redundant inequalities which are not informative at all; second, when m and d_1 are growing, there could be many inequalities which are nearly redundant, compared to the scale of noise in the data.

Notice that the random noise of b, which comes from $\hat{\mu}_n - \mu_{n,0}$, is ignored in Section 4.4 when data noise of observation is not taken into consideration. In this section, we develop a procedure to select informative inequalities in a general Q considering the random noise of b.

For any subset I of set $\{1,2,\dots,m\}$, denote M_I as the matrix comprised of the rows indexed by I in matrix M. Similarly, denote b_I as the subvector of b comprised of the elements indexed by I. By the Farkas' Lemma, for a general matrix M and a vector b, if the set of constraints indexed by I can imply all other constraints, i.e., the set $Q_I \coloneqq \{v | M_I v \leq b_I, v \geq 0\}$ equals Q, then there must exist a non-negative $m \times m$ matrix Π such that:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b$$

$$\Pi_{lj} = 0$$

for any $1 \leq l \leq m$ and $j \notin I$.

For any $j \in \{1,2,\dots,m\}$, we denote M_j as the j^{th} row of M, Π_j as the j^{th} row of Π and Π^k as the k^{th} column of Π. The coefficient matrix Π can serve as a signal of the importance of each inequality. If all the coefficients of the k^{th} inequality, Π^k, are zero or close to 0, then this inequality is not very informative to v. Inspired by the Farkas' Lemma, we propose the following selection procedure which slightly relaxes the constraints on Π:

Problem $\hat{R}$

$$\min_{\Pi} \sum_{k=1}^{m} g(\Pi^k)$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi \hat{b} \leq \hat{b} + \Lambda_n$$

where observed $\hat{b}$ is a $m \times 1$ vector which converges to b as the data sample size n goes to ∞, and $\Lambda_{n,m} = (\lambda_{n,m}, \dots, \lambda_{n,m})'$ in which $\lambda_{n,m}$ is a relaxation parameter measuring the maximum error allowed for each inequality.

The Vector $\Lambda_{n,m}$ can also be chosen to be specific to each inequality. For the inequality which we believe to be too important to be ruled out, we could set the corresponding $\lambda_{n,m}$ in $\Lambda_{n,m}$ to 0.

We choose the objective function $g(\cdot)$ such that it measures the importance of the constraints. One choice is $g(\Pi^k) = \text{sign}(\sum_{1 \leq j \leq m} \Pi_{jk})$. With this objective function $g(\cdot)$, the selection procedure $\hat{R}$ is essentially a binary integer programming model to select a minimum number of inequalities. We call the procedure L^0 selector.

The L^0 selector is extremely difficult to implement when m is large. However, many studies on LASSO and the Dantzig Selector show that some L^1 objective functions could enjoy nice statistical properties in model selection and also low computational costs. Below we propose a feasible L^1 objective function $g(\cdot)$:

$$g(\Pi^k) = \max_{1 \le j \le m} \Pi_{jk}$$

where Π_{jk} is the $(j,k)^{th}$ entry of Π.

With the above choice of $g(\cdot)$ and homogenous $\lambda_{n,m}$(i.e., λ), the formulation of the problem $\hat{R}$ is rewritten as:

Problem $\hat{R}$:

$$\min_{\Pi} \sum_{k=1}^{m} \max_{1 \le j \le m} \Pi_{jk}$$

subject to:

$$\Pi M \ge M, \Pi \ge 0$$

$$||(\Pi \hat{b} - \hat{b})_+||_\infty \le \lambda$$

4.5.2 Sparse Assumptions

Sparse assumptions play an essential role in the analysis of some L^1 penalization procedures, such as LASSO and the Dantzig Selector. In this subsection, we define sparse assumptions.

For any $1 \le j \le m$, we define a separation of inequality j as:

$$c_j \coloneqq \max_{v \in Q_j} M_j v - b_j$$

where

$$Q_j \coloneqq \{v | M_i v \le b_i, \forall i \ne j; v \ge 0\}$$

c_j measures the maximal separation of the j^{th} inequality for all points in Q_j. If $c_j > 0$, the j^{th} inequality is irredundant, otherwise the j^{th} inequality is redundant. Let T_0 be the set of indices j with $c_j > 0$ denoting the set of irredundant inequalities. Since c_j

characterizes the information carried by the j^{th} inequality, we can define sparse assumptions using c_j.

Definition 4.10 [Exact Sparse] Recall that T_0 is the subset of $\{1,2,\dots,m\}$ denoting all the irredundant inequalities, let $\widetilde{\Pi}^*$ be the solution to the following problem:

Problem R:

$$\min_{\Pi} \sum_{k \in T_0}^{m} g(\Pi^k)$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b$$

$$\Pi^k = 0 \text{ if } k \notin T_0$$

For any $m \times m$ matrix Π, denote $g(\Pi) \coloneqq (\, g(\Pi^1), \dots, g(\Pi^m))'$, which is a $m \times 1$ vector. The exact sparse assumption is defined as follows.

There exists absolute positive constants K^u, r and K, and an absolute constant $c_{g,n}$ which may depend on n, such that:

(1) $s_0 \coloneqq |T_0| = o(n \wedge m)$, which may increase at a slow rate as m and n increases;

(2) The sum of coefficients needed to construct each inequality is bounded: $\max_{1 \leq j \leq m} ||\widetilde{\Pi}_j^*||_1 \leq K d_1^r$;

If $||M||_2$ is normalized to be 1, then in general $r = 1/2$. In the Core Determining Class problem, we can prove that $K = 1$ and $r = 1$.

(3) $\max_{1 \leq k \leq m} g(\widetilde{\Pi}^{*k}) \leq K^u$;

(4) $\min_{j \in T_0} c_j \geq c_{g,n}$.

The exact sparse assumption assumes that all the binding constraints are informative and we are able to distinguish these constraints when the data noise is small enough. Denote set I^* as the set of indices with non-zero components in $g(\widetilde{\Pi}^*)$. In general set I^* is not necessarily the same as the set of indices corresponding to the minimum number of constraints in T_0. We use T_0 to denote the set of indices corresponding to its constraints if there is no confusion. In the special case of the Core Determining Class problem with a bipartite graph, we can show that $I^* = T_0$. That is to say, the L^1 selector recovers the Core Determining Class when λ is set to be 0. We expect the set I^* should not be too large compared to T_0. We show that in the next section, similar to the results in Candes and Tao (2007), the number of non-zero components in I^* has an order of $O(s_0)$ with probability approaching 1 if we employ a cutoff value $0 < \eta < 1$ to the solution $g(\widetilde{\Pi}^*)$.

For a q dimensional vector $\tilde{b}$ and a scalar λ, we define $\tilde{b} + \lambda \coloneqq (\tilde{b}_1 + \lambda, \dots, \tilde{b}_q + \lambda)$. Throughout the Chapter, we assume that M is a matrix with fixed value. Define $F \coloneqq \{\tilde{Q}_I(\tilde{b}) | \tilde{Q}_I(\tilde{b}) = \{v | M_I v \leq \tilde{b}_I\}, \tilde{b} \in R^m\}$ as a collection of sets which takes the formulation $\{v | M_I v \leq \tilde{b}_I\}$ for some $\tilde{b} \in R^m$ and $I \subset \{1,2, \dots, m\}$. Define the operation $\oplus$ which maps a set $\tilde{Q}_I(\tilde{b}) \in F$ and a real number λ to another set $\tilde{Q}_I(\tilde{b}) \oplus \lambda = \{v | M_I v \leq b_I + \lambda\}$. In the rest of this Chapter, let $\tilde{Q}_I \oplus \lambda$ be the abbreviation of $\tilde{Q}_I(\tilde{b}) \oplus \lambda$ if there is no confusion.

By analogy with the exact sparse assumption, we propose a more feasible approximate sparse assumption.

Definition 4.11 [Approximate Sparse] Suppose we can order the separations $c_1, \dots, c_m$ into $c_{(1)} \geq c_{(2)} \geq \cdots \geq c_{(m)}$ and suppose there exists a positive integer s^* such that:

(1) $s^* = o(n \wedge m)$;

Let T^* be the set of indices of the inequalities with the first s^* largest separations. Suppose K and r are absolute positive constants. Let $\sigma^2 \coloneqq \max_{1 \leq j \leq m} Var(\hat{b}_j - b_j)$. Let $\widetilde{\Pi}^*$ be the solution to the following problem.

Problem R:

$$\min_{\Pi} \sum_{k \in T^*}^{m} g(\Pi^k)$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b + K d_1^r \sigma \sqrt{\frac{\log(s^*)}{n}}$$

$$\Pi^k = 0 \text{ if } k \notin T^*$$

Then, it holds that:

(2) $\max_{1 \leq j \leq m} ||\widetilde{\Pi}_j^*||_1 \leq K d_1^r$;

(3) There exists an absolute constant K^u such that $\max_{1 \leq k \leq m} g(\widetilde{\Pi}^{*k}) \leq K^u$;

(4) $Q_{T^*} \subset Q \oplus K d_1^r \sigma \sqrt{\frac{\log(s^*)}{n}}$.

In the approximate sparse assumption, we allow $c_j > 0$ for all $1 \leq j \leq m$. Therefore, in the worst case, $g(\widetilde{\Pi}^{*k}) > 0$ for all $k \in \{1,2, \dots, m\}$. Essentially, the approximate sparse assumption assumes that there is a small set T^* indicating a set of inequalities to describe a feasible region similar to Q while the size of T^* is much smaller than m.

4.6 Properties of the Selection Procedure with Application in the Core Determining Class Problem

4.6.1 General Properties

In this subsection, we analyze the property of the selection procedure $\hat{R}$ and the choice of the relaxation parameter $\lambda_{n,m}$. We impose high level assumptions on $\hat{b}$ and $\lambda_{n,m}$ and then discuss a set of sufficient conditions for the assumptions.

Assumption 4.2 [Dominance of λ] Suppose we have data $B_1, B_2, \dots, B_n$ with dimension $m \times 1$ such that $b = E(B_i), 1 \le i \le n$. Suppose in practice we use $\hat{b} \coloneqq E_n(B_i)$ to estimate b. Suppose that with probability at least $1 - \alpha$,

(1) $\max_{1 \le j \le m} |\hat{b}_j - b_j| \le \lambda_{n,m}$;

(2) $\lambda_{n,m} \to 0$.

In Assumption 4.2, we require that the choice of relaxation parameter $\lambda_{n,m}$ dominate the maximal discrepancy between $\hat{b}_j$ and b_j for all $j \in \{1,2, \dots, m\}$. In addition, $\lambda_{n,m}$ should converge to 0 as sample size n increases to guarantee consistency.

Given $\lambda_{n,m}$, suppose that the solution to $\hat{R}$ is $\hat{\Pi}$. Denote $\hat{g}_k \coloneqq \max_{1 \le j \le m} \hat{\Pi}_{jk}$ for all $1 \le k \le m$. Define $\hat{I} \coloneqq \{k | \hat{g}_k \ne 0\}$ as the set of indices selected by the procedure $\hat{R}$. We consider the post-selection estimator $\hat{Q}_{\hat{I}} \coloneqq \{v | M_{\hat{I}} v \le \hat{b}_{\hat{I}}\}$ as the feasible set defined by the inequalities with indices in $\hat{I}$.

Lemma 4.4. If Assumption 4.2 holds, then with probability $\ge 1 - \alpha$, $\hat{Q}_{\hat{I}}$ satisfies:

(1) $Q \subset \hat{Q}_{\hat{I}} \oplus \lambda_{n,m}$;

(2) $\hat{Q}_{\hat{I}} \subset Q \oplus 2\lambda_{n,m}$.

Proof: see Appendix A.3

Lemma 4.4 shows that Q and $\hat{Q}_{\hat{I}}$ are very close to each other. Therefore, $\lambda_{n,m}$ should be at least as large as the $(1 - \alpha)$ quantile of the random variable $r_{n,m}$, where

$$r_{n,m} \coloneqq \max_{1 \le j \le m} |\hat{b}_j - b_j|$$

Chernozhukov, Chetverikov and Kato (2012) (CCK later) show that the distribution of $\sqrt{n} r_{n,m}$ can be well approximated by the distribution of the maxima of a Gaussian

vector under certain conditions and $\frac{(\log m)^7}{n} \to 0$ along with other mild regularity conditions. The calculation can be easily performed via Gaussian Multiplier bootstrap. A weaker bound (but still relatively sharp in many cases) of the $(1-\alpha)$ quantile of $r_{n,m}$ could be obtained using modest deviation theory of self-normalized vectors described in De La Puna (2009), which requires $\frac{(\log m)^{(2+\delta)}}{n} \to 0$ where $\delta > 0$.

Assumption 4.3 [Regularity Conditions]

(1) The data B_i is i.i.d. (The i.i.d. assumption can be extended to the i.n.i.d. assumption as both Lemma 4.5 and Lemma 4.6 allow i.n.i.d data with small modifications in the statement).

(2) There exists an absolute positive constant $C > 0$ such that

$$\max_{1 \le i \le n, 1 \le j \le m} |B_{ij}| \le C$$

(3) There exist an absolute positive constant $c_1 > 0$ such that

$$\min_{1 \le i \le n, 1 \le j \le m} E(B_{ij}^2) \ge c_1$$

The statement (2) in Assumption 4.3 holds for the Core Determining Class problem with the constant $C = 1$. Statement (3) may not be true in the Core Determining Class problem when the dimension d_1 grows. However, the problem can be fixed by multiplying $\sqrt{d_1}$ to B_{ij} when we make the assumption that $\frac{c}{d_1} \le v(u_i) \le \frac{c'}{d_1}$ for some absolute positive constants c and c'. We use Assumption 4.3 to derive properties for general selection procedure $\hat{R}$. In Section 4.6.2, we apply $\hat{R}$ to the Core Determining Class problem without Assumption 4.3.

Under Assumption 4.3, we are able to obtain the following two Lemmas on the choice of relaxation parameter λ. These two Lemmas are based on the results stated in De La Puna et al. (2009).

Lemma 4.5 [Choosing λ using Multiplier Bootstrap] Let $r_{n,d}^G := \max_{1 \le j \le m} \frac{\sum_{1 \le i \le n} B_{ij} e_{ij}}{n}$, where e_{ij} are independent standard normal random

variables. Suppose Assumption 4.3 holds and $\frac{(\log m \vee n)^7}{n} \to 0$, then the $(1-\alpha)$ quantile of $\sqrt{n}r_{n,d}^G$ is a consistent estimator of the $(1-\alpha)$ quantile of $\sqrt{n}r_{n,d}$.

Proof: Theorem 3.1 of CCK (2012) shows that under Assumption 4.3, at quantile $(1-\alpha)$, the multiplier bootstrap estimator of $\sqrt{n}r_{n,d}^G$ is consistent with $\sqrt{n}r_{n,d}$.

Lemma 4.6 [Choosing λ using Modest Deviation Theory of Self-Normalized Vectors] Denote $\hat{\sigma}^2 \coloneqq \max_{1\le j\le m}\{E_n(B_{ij}^2) - E_n(B_{ij})^2\}$. Let $\lambda_{n,m} \coloneqq \frac{C\hat{\sigma}^2\Phi^{-1}(1-\frac{\alpha}{2m})}{\sqrt{n}}$ for some constant $C > 1$. Suppose Assumption 4.3 holds and $\frac{(\log m)^{(2+\delta)}}{n} \to 0$ for some $\delta > 0$, then as $n \to \infty$, with probability at least $1-\alpha$,

$$\max_{1\le j\le m} |\hat{b}_j - b_j| \le \lambda_{n,m}$$

Proof: We refer the proof in De La Puna et al. (2009).

Next we discuss the performance of the L^1 selector $\hat{g}$ under the sparse assumptions.

Theorem 4.2 [Recovery of Informative Inequalities under Exact Sparse Assumption] Suppose Assumptions 4.2, 4.3 and the exact sparse assumption hold. Recall that c_j is the maximal separation of the j^{th} inequality and $c_{g,n} \le c_j$ for all $j \in T_0$. Let $0 < \eta < 1$ be an absolute constant. Assume that m, n, s_0, d_1 and $c_{g,n}$ obey a key growing condition:

$$\frac{(d_1^{2r}\log(s_0)) \vee \log(m)}{nc_{g,n}^2} \to 0$$

Consider the following two-step procedure:

(a) Step 1: set $\lambda_S \coloneqq (1+\epsilon)Kd_1^r\hat{\sigma}\sqrt{\frac{\log\left(\frac{4s_0}{\alpha}\right)}{n}} + \lambda_{n,m}$, with $\epsilon > 0$ be an absolute constant and $\lambda_{n,m}$ to be chosen according to Lemma 4.5 or Lemma 4.6.

(b) Step 2: let $\hat{g}_S$ be the solution to $\hat{R}$ with $\lambda = \lambda_S$ and let $\hat{I}_S \coloneqq \{j|\hat{g}_{S,j} \ne 0\}$, then, construct set $\hat{I}_{S,\eta} \coloneqq \{j|\hat{g}_{S,j} \ge \eta\}$.

Then, with probability $\geq 1-\alpha$, $\hat{I}_{S,\eta}$ has the following properties:

(1) There exists an absolute constant C_T such that $||\hat{I}_{S,\eta}||_0 \leq \frac{C_T s_0}{\eta}$;

(2) $\hat{I}_{S,\eta} \supset T_0$;

(3) $Q \subset \hat{Q}_{\hat{I}_{S,\eta}} \oplus \lambda_{n,m}$;

(4) $\hat{Q}_{\hat{I}_{S,\eta}} \subset Q \oplus \lambda_{n,m}$.

Proof: see Appendix A.4

In Theorem 4.2, we consider a two-step procedure. First, we select the inequalities using a larger relaxation parameter λ_S. Such relaxation can significantly reduce the number of inequalities. However, as soon as λ_S converges to 0 fast enough, all the informative inequalities will be preserved. Second, the cutoff strategy additionally throws away some nearly redundant inequalities which were not detected in the first step selection. The set of those inequalities survive the two-step procedure has good properties: (1) it has the same size compared to the minimum set of inequalities, T_0, up to a constant multiplier; (2) it contains T_0 with probability approaching 1; (3) $\hat{Q}_{\hat{I}_{S,\eta}}$ is close to the true feasible region $Q = \{v|Mv \leq b\}$, with error up to $O(\lambda_{n,m})$.

The constant K can be computed via M. If $||M_{ij}||_2 = 1$ for all j and $M_{ij} > 0$ for all i and j, then $K \leq 1$ and $r = 1/2$. In practice, s_0 is unknown, so we recommend to use n for s_0 as a starting value and then iterate a few times. We also recommend to use $\epsilon = 0.1$ in practice.

Theorem 4.3 [Recovery of Informative Inequalities under Approximate Sparse Assumption] Suppose Assumptions 4.2, 4.3 and the approximate sparse assumption hold. Let $0 < \eta < 1$ be an absolute constant. Assume that m, n, s_0, d_1 and $c_{g,n}$ obey a key growing condition:

$$\frac{(d_1^{2r}\log(s^*)) \vee \log(m)}{nc_{g,n}^2} \to 0$$

Consider the following estimation procedure:

(a) Step 1: set $\lambda_S \coloneqq 2(1+\epsilon)Kd_1^r\hat{\sigma}\sqrt{\frac{\log\left(\frac{4s^*}{\alpha}\right)}{n}} + \lambda_{n,m}$, with $\epsilon > 0$ be an absolute constant and $\lambda_{n,m}$ to be chosen according to Lemma 4.5 or Lemma 4.6.

(b) Step 2: let $\hat{g}_S$ be the solution to $\hat{R}$ with $\lambda = \lambda_S$ and let $\hat{I}_S \coloneqq \{j|\hat{g}_{S,j} \neq 0\}$, then, construct set $\hat{I}_{S,\eta} \coloneqq \{j|\hat{g}_{S,j} \geq \eta\}$.

Then, with probability $\geq 1-\alpha$, $\hat{I}_{S,\eta}$ has the following properties:

(1) There exists an absolute constant C_T such that $||\hat{I}_{S,\eta}||_0 \leq \frac{C_T s^*}{\eta}$;

(2) $Q \subset \hat{Q}_{\hat{I}_{S,\eta}} \oplus \lambda_{n,m}$;

(3) $\hat{Q}_{\hat{I}_{S,\eta}} \subset Q \oplus \frac{\lambda_{n,m}+\lambda_S}{2}$.

Proof: see Appendix A.5

Again, in practice we can set $s^* = n$ as a starting value and then iterate for a few times. If the approximate sparse assumption holds instead of the exact sparse assumption, the estimation procedure suffers from additional estimation error with size λ_S, which depends on the unknown parameter s^*.

4.6.2 Application in Estimating Measure v in the Core Determining Class Problem

To find the Core Determining Class given a bipartite graph $G = (U, Y, \varphi)$, we can use the method proposed in Section 4.4 to eliminate all the redundant inequalities and find exact solution when data noise of observation is not taken into consideration. We can also use the L^1 selector proposed in Section 4.6.1 to find an approximate solution to the Core Determining Class problem. In addition, we can consider a hybrid method: first, we find the exact solution according to the method described in Section 4.4, and second, we apply

the selection procedure presented in Section 4.6.1 using the inequalities selected from the previous step.

The hybrid method may speed up the selection procedure significantly. In this subsection, we discuss the general selection procedure first, and then briefly discuss the hybrid method.

In the Core Determining Class problem, the equality $v(U) = 1$ is never redundant. Therefore, we let the $(m-1)^{th}$ and m^{th} inequalities be $v(U) \geq 1$ and $v(U) \leq 1$ among the total m inequalities. Since there is no reason to drop the last two inequalities, we define problems R^C and $\hat{R}^C$:

Problem R^C:

$$\min_{\Pi} \sum_{k=1}^{m-2} \max_{1 \leq j \leq m-2} \Pi_{jk}$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b$$

Problem $\hat{R}^C$:

$$\min_{\Pi} \sum_{k=1}^{m-2} \max_{1 \leq j \leq m-2} \Pi_{jk}$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi \hat{b} \leq \hat{b} + \Lambda$$

where $\Lambda = (\lambda_{n,m}, \dots, \lambda_{n,m}, 0, 0)$ with $\lambda_{n,m}$ left to be chosen.

Let $\widehat{\Pi}$ be the solution to $\hat{R}^C$. First, we prove an important result specific to the Core Determining Class problem.

Lemma 4.7 [Perfect Recovery of the Minimum Mode T_0] If $\hat{\mu}_n$ is non-degenerate and $\lambda_{n,m} = 0$, let $\hat{g}(\widehat{\Pi}^k) = \max_{1 \le j \le m-2} \widehat{\Pi}_{jk}$, then:

(1) The L^0 norm of $\hat{g}$, $||\hat{g}||_0$, satisfies $||\hat{g}||_0 = s_0$;

(2) $\max_{1 \le j \le m-2} ||\widehat{\Pi}_j||_1 \le d_1$;

(3) $\max_{1 \le k \le m-2} \hat{g}(\widehat{\Pi}^k) \le 1$;

(4) The set of indices with non-zero entries of $\hat{g}$ satisfies:

$$\hat{I} \coloneqq \{k | \hat{g}_k \neq 0\} = T_0$$

As a special case, $I^* = T_0$.

Proof: see Appendix A.6

Lemma 4.7 indicates that under the exact sparse assumption, the recovery of model T_0 could be done simply by looking at the non-zero entries of the solution to the problem $\hat{R}^C$. Due to the special property presented in Lemma 4.7, we show that the relaxation parameter λ_S in Theorem 4.3 can be much tighter.

Therefore, the selection procedure would require much less number of observations in order to achieve good performance.

Definition 4.12 [Approximate Sparse on Core Determining Class] Suppose we can order the separations $c_1, \dots, c_{m-2}$ into $c_{(1)} \ge c_{(2)} \ge \cdots \ge c_{(m-2)}$ and suppose there exists a positive integer s^* such that:

(1) $s^* = o(n \wedge m)$;

Let T^* be the set of indices of the inequalities with the first s^* largest separations. Suppose K and r are absolute positive constants. Let $\sigma^2 \coloneqq \max_{1 \le j \le m-2} Var(\hat{b}_j - b_j)$. Let $\widetilde{\Pi}^*$ be the solution to the following problem:

Problem R:

$$\min_{\Pi} \sum_{k \in T^*}^{m-2} g(\Pi^k)$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b + K d_1^r \sigma \sqrt{\frac{\log(s^*)}{n}}$$

$$\Pi^k = 0 \text{ if } k \notin T^*$$

Then, it holds that:

(2) $\max_{1 \leq j \leq m} ||\widetilde{\Pi}_j^*||_1 \leq K d_1^r$;

(3) There exists an absolute constant K^u such that $\max_{1 \leq k \leq m-2} g(\widetilde{\Pi}^{*k}) \leq K^u$;

(4) $Q_{T^*} \subset Q \oplus K d_1^r \sigma \sqrt{\frac{\log(s^*)}{n}}$.

Define $\hat{\sigma}^2 \coloneqq \max_{1 \leq j \leq m-2} (E_n(B_{ij}^2) - E_n(B_{ij})^2)$.

Lemma 4.8 [Recovery of Informative Inequalities under Core Determining Class] Suppose Assumptions 4.2, 4.3 and the exact sparse assumption hold. Suppose G and $\hat{\mu}_n$ are non-degenerate. Recall that c_j is the maximal separation of the j^{th} inequality and $c_{g,n} \leq c_j$ for all $j \in T^*$. Let $0 < \eta < 1$ be an absolute constant and set $\lambda_S^C \coloneqq (1 + \epsilon)\hat{\sigma} \sqrt{\frac{\log(\frac{4s^*}{\alpha})}{n}}$ where $\epsilon > 0$ is a constant. Assume that s^* and $c_{g,n}$ obey a key growing condition:

$$\frac{\log(s^*)}{n c_{g,n}^2} \to 0$$

Assume that with probability approaching 1, the empirical measure $\hat{\mu}_n$ obeys:

$$\max_{1\le l\le d_2} \frac{|\hat{\mu}_n(l)-\mu(l)|}{\mu(l)} \to 0$$

Let $\widehat{\Pi}$ be the solution to R^C with $\lambda_{n,m}=\lambda_S^C$. Let $\hat{g}_{S,k} \coloneqq \max_{1\le j\le m-2} \widehat{\Pi}_{jk}$.

Then, with probability $\ge 1-\alpha$, the set $\hat{I}_{S,\eta} \coloneqq \{j|\hat{g}_{S,k}\ge \eta\}$ has the following properties:

(1) There exists an absolute constant C_T such that $||\hat{I}_{S,\eta}||_0 \le \frac{C_T s^*}{\eta}$;

(2) $Q \subset \hat{Q}_{\hat{I}_{S,\eta}} \oplus \lambda_S^C$;

(3) $\hat{Q}_{\hat{I}_{S,\eta}} \subset Q \oplus 2\lambda_S^C$.

The value s^* can be obtained iteratively by setting $s^*=n$ as an initial value.

Proof: see Appendix A.7

The key assumption $\max_{1\le l\le d_2} \frac{|\hat{\mu}_n(l)-\mu(l)|}{\mu(l)} \to 0$ mainly relies on the growing rate of d_2. When $\mu(l)=O(\frac{1}{d_2})$ and $\frac{d_2^3}{n}\to 0$, the assumption $\max_{1\le l\le d_2} \frac{|\hat{\mu}_n(l)-\mu(l)|}{\mu(l)} \to 0$ holds. Lemma 4.8 obtains stronger results compare to Theorem 4.3 due to the structure of the bipartite graph.

It is natural to consider a hybrid estimation strategy combining the combinatorial method in Section 4.4 and the selection procedure in Section 4.6. There are a few points that we would like to make about the hybrid method:

(1) When s_0 is small, the hybrid method performs similarly to the combinatorial method only.

(2) When s_0 is large, there may be significant gains from the hybrid method in terms of computational speed compared to the selection procedure in Section 4.6 only, and

significant inequality reduction compared to the combinatorial method in Section 4.4 only.

4.7 Conclusion

In this Chapter, we present a novel approach to the study of the passenger utility functions in a last mile transportation system. The unknown parameters in the passenger utility function are treated as unobserved events and the specific characteristics of transportation trips are treated as observed outcomes. We consider estimating the probability measure on the unobservable events given observations of the frequencies of the outcomes.

We try to select the set of a minimum number of inequalities, which is called the Core Determining Class, to describe the feasible set of the target probability measure. We propose a procedure to construct the exact Core Determining Class when data noise of observation is not taken into consideration. We prove that, if there is no degeneracy, the Core Determining Class only depends on the structure of the bipartite Graph, not the probability measure μ on the outcomes.

For a general problem of linear inequality selection under data noise, we propose a selection procedure similar to the Dantzig selector in regression. A formulation is proposed to identify the importance of each inequality in a feasible set defined by many inequalities constraints. We describe the exact sparse assumptions and approximate sparse assumptions, which are similar to the traditional sparse assumptions in a linear regression environment. We prove that the selection procedure has good statistical properties under the sparse assumptions. We also apply the selection procedure to the Core Determining Class problem and develop a hybrid selection method combining a combinatorial method and a selection formulation.

Chapter 5

Concluding Remarks

Last mile transportation systems are critical extensions to traditional public transit systems. The unavailability of this type of service is one of the main deterrents to the use of public transport in urban areas. Good design and operations of LMTS can make the overall public transportation systems more efficient and attractive. In this thesis, we study the LMTS from three perspectives. We have presented queueing, optimization and inference approaches for design and operation of the LMTS.

In Chapter 2, we study the LMTS from a queueing perspective. We consider the supply side of the LMTS in a stochastic setting, with stochastic batch demands resulting from the arrival of groups of passengers who request last-mile service at urban rail stations or bus stops. We study a very difficult type of queueing system involving batch arrivals and requiring the simultaneous consideration of vehicle routing, queueing issues and the use of geometrical probability arguments. We derive several closed-from expressions for bounding and approximating the principal performance characteristics of systems as a function of the fundamental design parameters of such systems. The expressions perform consistently well for a broad and realistic range of input values and conditions. On the practical side, these expressions can therefore be used for the preliminary planning and design of last mile transportation systems, especially for approximately determining resource requirements, such as the number of vehicles/servers needed to achieve some pre-specified level of service.

In Chapter 3, we study the LMTS from an optimization perspective. We have developed several operating strategies and algorithms for the design of passenger delivery schedules and vehicle routes for a multi-vehicle fleet of delivery vehicles with the objective of minimizing the waiting time and riding time of passengers. A myopic operating strategy is introduced first for the case in which the last mile demand from each group of arriving passengers is revealed sequentially. Two more advanced operating strategies are then described in detail, one based on a metaheuristic using tabu search and the other using an exact Mixed Integer Programming model, which is solved approximately in two stages. These operating strategies are implemented in a number of computational experiments with a broad and realistic range of input values and conditions. We believe that the operating strategies we have proposed can be very useful for LMTS, providing good operating plans for these complex systems. The best approach to the passenger service assignment, vehicle routing, and scheduling of the LMTS depends on the context and the user's needs.

In Chapter 4, we present a novel approach to study the passenger utility function in a last mile transportation system. The passenger utility function provides critical information to LMTS service providers when it comes to understanding and estimating passenger demand and designing and operating their systems. In this chapter, we treat the unknown parameters in the passenger utility function as unobserved events, and the observed characteristics of transportation trips, such as passenger waiting time, in-vehicle riding time and monetary travel cost, as observed outcomes. We construct a bipartite graph representing the relationships between the events and the outcomes. We propose a general method for identifying the probability measures of the events given the observations of the frequencies of the outcomes, including a combinatorial algorithm in which the data noise of the observations is ignored and a general procedure in which data noise is taken into consideration. This chapter offers an example in the field of transportation and logistics of how hidden information can be inferred from observed data in a context in which only some correlations are known between unobservable events and observed outcomes. This is a key problem in the era of "big data" with fast expanding data availability in a variety of industries.

A natural extension of our analysis is to study the bounding and approximation of the performance of a transportation system combining the "last mile" service described here with a "first mile" service, i.e., have the vehicles pick up passengers from the serviced district and transport them to the rail station. Under certain conditions, this may lead to increased efficiencies since vehicles will not be returning to the rail station empty. However, the analysis of this type of combined first- and last-mile service is significantly more complicated, if it is to be carried out at a similar level of detail as the analysis presented in Chapter 2 for last-mile services alone. We can also study the stochastic versions of the LMTS operation problem described in Chapter 3, involving some combination of unreliable train schedules, probabilistic last mile service requests, and uncertainty about vehicle service times due to traffic congestion. As for the information inference topic in Chapter 4, an important future direction is to consider the possibility of discretization and segmentation of hidden events and observable outcomes with continuous values.

Appendix A

Proofs in Chapter 4

A.1 Proof of Lemma 4.2

For any $A \notin S'_U$, suppose (1) $\exists A_1, A_2 \subset A$, $A_1, A_2 \neq \emptyset$, and $A_1 \cup A_2 = A$, such that $\varphi(A_1) \cap \varphi(A_2) = \emptyset$; or (2) $\exists u \in U$, such that $u \notin A$ and $\varphi(u) \subset \varphi(A)$.

If (1) is true, then $v^M(A) = v^M(A_1 \cup A_2) = v(A_1) + v(A_2) \leq \mu(\varphi(A_1)) + \mu(\varphi(A_2)) = \mu(\varphi(A_1) \cup \varphi(A_2)) = \mu(\varphi(A_1 \cup A_2)) = \mu(\varphi(A))$, so $A \notin S_U$.

If (2) is true, then $v^M(A) \leq v(A \cup \{u\}) \leq \mu\big(\varphi(A \cup \{u\})\big) = \mu(\varphi(A))$, so $A \notin S_U$.

Therefore, by definition of S'_U, we obtain $S_U \subset S'_U$.

For any $A \notin S_U$, assuming elements in S_U are denoted as A_i for $1 \leq i \leq |S_U|$. For simplicity of notations, we can consider A_i as a vector in $\{0,1\}^{d_1}$. By definition, $\exists \pi \geq 0$, s.t., (1) $\sum_{i=1}^{r} \pi_i A_i \geq A$, and (2) $\sum_{i=1}^{r} \pi_i \mu(\varphi(A_i)) \leq \mu(\varphi(A))$, where $r \coloneqq |S_U|$. Without loss of generality, assume $\pi_i > 0, \forall i = 1,2, \dots, r$, otherwise we would simply omit the A_i which corresponds to $\pi_i = 0$ in the sum above. Such an assumption does not affect our analysis below.

Since $\sum_{i=1}^{r} \pi_i A_i \geq A$, we have $\sum_{i=1}^{r} \pi_i 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geq 1(A \cap \varphi^{-1}(y) \neq \emptyset))$ for any $y \in Y$. By Galichon and Henry (2011), μ is sub-modular. Therefore,

$$\sum_{i=1}^{r} \pi_i \mu\big(\varphi(A_i)\big)$$
$$= \sum_{y \in Y} \sum_{i=1}^{r} \pi_i \mu(y) 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geq \sum_{y \in Y} \mu(y) 1(A \cap \varphi^{-1}(y) \neq \emptyset)$$
$$= \mu(\varphi(A))$$

However, by construction, we know that $\sum_{i=1}^{r} \pi_i \mu\big(\varphi(A_i)\big) \leq \mu(\varphi(A))$. Hence the inequality above holds as an equality, i.e., for any $y \in Y$,

$$\sum_{i=1}^{r} \pi_i 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) = 1(A \cap \varphi^{-1}(y) \neq \emptyset)$$

On the other hand, we know that $\sum_{i=1}^{r} \pi_i A_i \geq A$. Therefore, for any $y \in Y$, we have $\varphi^{-1}(y) \cap A \subsetneq A_i$ or $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ for all i.

We prove the above argument by contradiction. Assuming that there exists a $y \in Y$ and $1 \leq i \leq r$ such that $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ and $\varphi^{-1}(y) \cap A \subsetneq A_i$, then, there exists $u \neq u'$ such that $u, u' \in \varphi^{-1}(y)$, $u \in A \cap A_i$, $u' \in A$ but $u' \notin A_i$. Thus,

$$\sum_{i=1}^{r} \pi_i A_i 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) = \pi_i + \sum_{j \neq i} \pi_j A_j 1\big(A_j \cap \varphi^{-1}(y) \neq \emptyset\big)$$
$$\geq \pi_i + \sum_{j \neq i} \pi_j 1\big(u' \in A_j\big) = \pi_i + \sum_{i=1}^{r} \pi_j 1\big(u' \in A_j\big) \geq \pi_i + 1 > 1$$
$$= 1(A \cap \varphi^{-1}(y) \neq \emptyset)$$

It is a contradiction. Thus, for any $y \in Y$, we have $\varphi^{-1}(y) \cap A \subsetneq A_i$ or $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ for all i.

The above statement immediately implies the following conclusion:

If A is self-connected, then for any A_i, either $A_i \cap A = \emptyset$ or $A_i \cap A = A$. By the equality, for any A_i, there exists no $y \in \varphi(A_i)$ such that $y \notin \varphi(A)$. So we have $\varphi(A_i) = \varphi(A)$. Since $A = A_i$, there exists $u \in U$ such that $\varphi(u) \subset \varphi(A)$. Since $U \notin A$, we have $A \notin S'_U$.

Otherwise A is not self-connected and we have $A \notin S'_U$.

Therefore, in both cases, $A \notin S'_U$. This means that $S_U \supset S'_U$.

Combining with the result that $S'_U \supset S_U$, we have $S_U = S'_U$.

A.2 Proof of Theorem 4.1

Since S^* is the minimum set of inequalities which contains all information if condition $v(U) = 1$ holds, therefore, $S^* \subset S_U$ and $S^* \subset S_Y^{-1}$. We have $S^* \subset S_U \cap S_Y^{-1}$.

For any $A \in S_U \cap S_Y^{-1}$, by contradiction, suppose $A \notin S^*$. So there exists $\pi_i > 0$ and $A \in S^*$, $1 \leq i \leq r$ and $\pi_0 \geq 0$, such that:

(1) $\sum_{1 \leq i} \pi_i A_i - \pi_0 \geq A$;

(2) $\sum_{1 \leq i} \pi_i \mu(\varphi(A_i)) - \pi_0 \geq \mu(\varphi(A))$.

By the similar argument of Lemma 4.2, all the inequalities in (2) must hold as equalities. Again, for any $y \in Y$, either $\varphi^{-1}(y) \cap A$ is a subset of A_i, or it does not intersect with A_i. Since $A \in S_U$ is connected, we have either $A_i \supset A$ or $A_i \cap A = \emptyset$ for any A_i.

Since there exists B such that $\varphi^{-1}(B) = A^c$, then $\varphi^{-1}(\varphi(A)^c) = A^c$. Without loss of generality, let $B = \varphi(A)^c$. Since the graph is connected, it must hold that $\varphi(u) \cap \varphi(A) \neq \emptyset$ for some $u \in A^c$. Since $\pi_0 > 0$, then there must exist a set A_{i0} such that $u \in A_{i0}$. So $A_i \supset A$ due to $\varphi(u) \cap \varphi(A) \neq \emptyset$. Also, for any $y \in B$, it is also required that $\varphi^{-1}(b) \subset A_i$ or $\varphi^{-1}(b) \cap A_i = \emptyset$.

However, the set B is self-connected. Therefore, for any A_i, we have $B \subset A_i$ or $B \cap A_i = \emptyset$. Hence, $A_{i0} = U$, which contradicts with the definition of S^*.

Therefore, $S^* = S_U \cap S_Y^{-1}$.

A.3 Proof of Lemma 4.4

The selected set $\hat{I}$ implies all relaxed inequalities $M_j v \leq \hat{b}_j + \lambda_{n,m}$. Therefore, $\hat{Q}_{\hat{I}} \subset \hat{Q} \oplus \lambda_{n,m}$. According to Assumption 4.2, $\max_{1\leq j\leq m} |\hat{b}_j - b_j| \leq \lambda_{n,m}$ with probability $1-\alpha$, so $Q \subset \hat{Q} \oplus \lambda_{n,m}$ and $\hat{Q} \subset Q \oplus \lambda_{n,m}$with probability $1-\alpha$.

Therefore, $Q \subset \hat{Q}_{\hat{I}} \oplus \lambda_{n,m}$ and $\hat{Q}_{\hat{I}} \subset \hat{Q} \oplus \lambda_{n,m} \subset Q \oplus 2\lambda_{n,m}$ with probability $1-\alpha$.

A.4 Proof of Theorem 4.2

Consider $\widetilde{\Pi}^*$ defined in Definition 4.10, for every $1 \leq j \leq m$, $|\widetilde{\Pi}_j^*(\hat{b}-b)| \leq ||\widetilde{\Pi}_j^*||_1 \cdot \max_{j\in T_0} |\hat{b}_j - b_j| \leq K d_1^r \hat{\sigma}\sqrt{\frac{\log(\frac{4s_0}{\alpha})}{n}}$ with probability at least $1-\alpha$. Therefore, it is easy to see that $\widetilde{\Pi}^*$ is a feasible solution to the problem $\hat{R}$ with probability at least $1-\alpha$. Now we focus on the event when $\widetilde{\Pi}^*$ is a feasible solution to $\hat{R}$.

Let $\widehat{\Pi}$ be the solution to the problem $\hat{R}$, so

$$||g(\widehat{\Pi})||_1 \leq ||g(\widetilde{\Pi}^*)||_1 \leq s_0 K^u$$

and

$$\hat{I}_{S,\eta} \leq \frac{s_0 K^u}{\eta}$$

For any $j \in T_0$, let v_j be the point such that the maximal separation is realized while other inequalities hold for v. Therefore, by construction,

$$\widehat{\Pi}\left(Mv_j - \hat{b}\right) \geq Mv_j - \hat{b} - \lambda_S$$

We have $Mv_j \geq b + c_{g,n}$ and $Mv_{j'} - \hat{b} \leq 0$ for all $j' \neq j$. So the j^{th} inequality indicates that

$$\widehat{\Pi}_{jj}\left(c_{g,n} - \hat{b}_j + b_j\right) \geq c_{g,n} - \lambda_S - \hat{b}_j + b_j$$

Therefore,

$$\widehat{\Pi}_{jj} \geq \frac{c_{g,n} - \lambda_S - (\hat{b}_j - b_j)}{c_{g,n} - (\hat{b}_j - b_j)} \geq \frac{c_{g,n} - \lambda_S - \lambda_{n,m}}{c_{g,n} - \lambda_{n,m}}$$

The growing condition $\frac{d_1^{2r}\log(s_0)\vee\log(m)}{nc_{g,n}^2} \to 0$ guarantees that $\widehat{\Pi}_{jj} > \eta$ for any $\eta < 1$ as $n \to \infty$. Therefore, $j \in \hat{I}_S$, $j \in \hat{I}_{S,\eta}$ and $\hat{I}_{S,\eta} \supset T_0$.

Since we know that $T_0 \subset \hat{I}_{S,\eta}$, then $\hat{Q}_{\hat{I}_{S,\eta}} \subset \hat{Q}_{T_0} \subset Q \oplus \lambda_{n,m}$.

By construction, $Q \subset \hat{Q} \oplus \lambda_{n,m} \subset \hat{Q}_{\hat{I}_{S,\eta}} \oplus \lambda_{n,m}$.

A.5 Proof of Theorem 4.3

Consider $\widetilde{\Pi}^*$ defined in Definition 4.11, for every $1 \leq j \leq m$, $|\widetilde{\Pi}_j^*(\hat{b} - b)| \leq ||\widetilde{\Pi}_j^*||_1 \cdot \max_{j \in T_0} |\hat{b}_j - b_j| \leq K d_1^r \hat{\sigma} \sqrt{\frac{\log(\frac{4s^*}{\alpha})}{n}}$ with probability at least $1 - \alpha$. Therefore, it is easy to see that $\widetilde{\Pi}^*$ is a feasible solution to the problem $\hat{R}$ with probability at least $1 - \alpha$. Now we focus on the event when $\widetilde{\Pi}^*$ is a feasible solution to $\hat{R}$.

Let $\widehat{\Pi}$ be the solution to the problem $\hat{R}$, so

$$||g(\widehat{\Pi})||_1 \leq ||g(\widetilde{\Pi}^*)||_1 \leq s^* K^u$$

and

$$\hat{I}_{S,\eta} \leq \frac{s^* K^u}{\eta}$$

For any $j \in T^*$, let v_j be the point such that the maximal separation is realized while other inequalities hold for v. Therefore, by construction,

$$\widehat{\Pi}\left(M v_j - \hat{b}\right) \geq M v_j - \hat{b} - \lambda_S$$

We have $Mv_j \geq b + c_{g,n}$ and $Mv_{j'} - \hat{b} \leq 0$ for all $j' \neq j$. So the j^{th} inequality indicates that

$$\widehat{\Pi}_{jj}\left(c_{g,n} - \hat{b}_j + b_j\right) \geq c_{g,n} - \lambda_S - \hat{b}_j + b_j$$

Therefore,

$$\widehat{\Pi}_{jj} \geq \frac{c_{g,n} - \lambda_S - (\hat{b}_j - b_j)}{c_{g,n} - (\hat{b}_j - b_j)} \geq \frac{c_{g,n} - \lambda_S - \lambda_{n,m}}{c_{g,n} - \lambda_{n,m}}$$

The growing condition $\frac{d_1^{2r}\log(s^*)\vee\log(m)}{nc_{g,n}^2} \to 0$ guarantees that $\widehat{\Pi}_{jj} > \eta$ for any $\eta < 1$ as $n \to \infty$. Therefore, $j \in \hat{I}_S$, $j \in \hat{I}_{S,\eta}$, and $\hat{I}_{S,\eta} \supset T^*$.

Since we know that $T^* \subset \hat{I}_{S,\eta}$, so $\hat{Q}_{\hat{I}_{S,\eta}} \subset \hat{Q}_{T^*} \subset Q \oplus \frac{\lambda_S + \lambda_{n,m}}{2}$.

By construction, $Q \subset \hat{Q} \oplus \lambda_{n,m} \subset \hat{Q}_{\hat{I}_{S,\eta}} \oplus \lambda_{n,m}$.

A.6 Proof of Lemma 4.7

Let Π be a feasible solution to the problem R:

$$\min_{\Pi} \sum_{k=1}^{m-2} \max_{1 \leq j \leq m-2} \Pi_{jk}$$

subject to:

$$\Pi M \geq M, \Pi \geq 0$$

$$\Pi b \leq b$$

$$\Pi_{ij} = 0, \text{if } j \notin T_0$$

A feasible solution to this above problem is that $\Pi_{ii} = 1$ for all $i \in T_0$, and $\Pi_{ij} = 0$ for all $i \leq j$. Hence, the optimal value of the objective function is no worse than s_0.

In our case, if we denote $p = m$, except for the p^{th} row of M, every row satisfies $M_{ii} \in \{0,1\}^d$. Again, for the problem R, a feasible solution is $\Pi_{ii} = 1$ for any $i \in T_0$.

Therefore, the value of the objective function is no worse than s_0. Meanwhile, for any $i \notin T_0$, by definition, there exists $\alpha_j \geq 0$ for any $j \neq i, j \in T_0$ and $\alpha_p \geq 0$ such that:

$$\sum\nolimits_{j \in T_0} \alpha_j M_j - \alpha_p \geq M_i$$

and

$$\sum\nolimits_{j \in T_0} \alpha_j b_j - \alpha_p \leq b_i$$

Without loss of generality, we could assume that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r > 0 = \alpha_{r+1} = \cdots = \alpha_{p-1}$. Next we prove that there must be a feasible vector of α_i such that $\alpha_1 \leq 1$. Then we could conclude that the minimum value of the objective function in problem R is s_0, and the optimal solution exactly recovers the true model. Denote the set A correspond to M_j, and A_i correspond to M_i. Without loss of generality, by Galichon and Henry (2011), $\mu(\varphi(A))$ is a sub-modular, since $b_j = \mu(\varphi(A))$, then,

$$\begin{aligned}\sum\nolimits_{1 \leq i \leq r} \alpha_i b_j - \alpha_p & \\ &= \sum\nolimits_{1 \leq i \leq r} \alpha_i \mu(\varphi(A_i)) - \alpha_p \mu(\varphi(U)) \geq \mu(\varphi\left(\sum\nolimits_{1 \leq i \leq r} \alpha_i A_i - \alpha_p\right) \\ &\geq \mu(A) = b_j\end{aligned}$$

Therefore, the above equality holds as an equality. If $\alpha_1 > 1$, then $\alpha_p > 0$. So for any $u \notin A_i$, there must be $j \in T_0$ such that $u \in M_j$.

So for any $y \in \varphi(A)$, either $\varphi^{-1}(y) \cap A_i \cap A = \emptyset$ or $(\varphi^{-1}(y) \cap A) \subset A_i$. Similarly, for any $y \notin \varphi(A)$, we have $\varphi^{-1}(y) \cap A_i = \emptyset$ or $\varphi^{-1}(y) \subset A_i$.

(1) A is connected. Let A' be $\{u | \varphi(u) \subset A\}$. So A' implies A. We only need to prove that A' can be constructed via $\sum_{1 \leq i \leq r} \alpha_i A_i - \alpha_p U$.

(2) A is connected and there is no $u \notin A$ such that $\varphi(u) \subset \varphi(A)$, we have $A \subset S_U$. Hence $B \coloneqq \varphi(A)^c$ is not connected. Let $B_1, \dots, B_r$ denote all the disconnected branches of B. Let $C_k = \varphi(B_k)$ for any $1 \leq k \leq r$. Therefore, $\cup_{k=1}^{r} C_k = A^c$, $C_{k_1} \cap C_{k_2}$ for any $k_1 \neq k_2$, and each C_k is connected with A.

If we denote $C^k = \{u | u \in A^c, u \notin C_k\}$, then $A \cup C^1$, $A \cup C^2$,..., $A \cup C^r$ are sets in S_U. They are also sets in S_Y^{-1} since $C_k = (A \cup C^k)^c$ is connected. Therefore, All these sets are in S^*. Let $\alpha_i = 1$ and $\alpha_p = r - 1$, we could reconstruct the inequality indicated by A. Since $r \geq 2$, so all the coefficients $\alpha_k \leq 1$.

(3) A is not connected. Let $A_1, \dots, A_w$ be the connected branches. Let $B = \varphi(A^c)$. Without loss of generality, similar to (1), we could assume that $A_i \in S_U$ for $1 \leq i \leq w$. Assume $B_1, \dots, B_k$ is the connected branches of B. Let $C_i = \varphi^{-1}(B_i)$ for $1 \leq i \leq k$, then $C_{i_1} \cap C_{i_2} = \emptyset$ for any $i_1 \neq i_2$ and $C_i \cap A \neq \emptyset$ for any i. Therefore, C_i for $1 \leq i \leq k$ and A_j for $1 \leq j \leq w$ form a bipartite graph G_0. For every A_i, let $AC_1, \dots, AC_{i_r}$ be the connect branches of $G_0 - \{A_i\}$. Since the entire graph is connected, so AC_i is connected with A_i for $1 \leq i \leq i_r$. Let $AC^i \coloneqq \{u | u \notin AC_i\}$, then AC^i is a set in $S_U \cap S_Y^{-1} = S^*$. Therefore, the set A_i could be constructed by $\sum_{k=1}^{i_r} AC_k - (i_r - 1)U$.

If for some set AC_k appears in different i, let AC be such a set that it appears in $1 \leq i \leq J, J \geq 2$. Hence $A_1, A_2, \dots, A_J \subset AC = \emptyset$. Without loss of generality, we suppose $C_1, \dots, C_q \subset AC$, $q \geq 1$ and $C_{q+1}, \dots, C_k \cap AC = \emptyset$. Then, for any $1 \leq i \leq J$, $AC - A_i$ is a connected branch in $G_0 - A_i$, which means that $C_1, \dots, C_q$ does not connected with $A - AC$, and $C_{q+1}, \dots, C_k$ does not connect with $AC - A_i$. If $J \geq 2$, $C_{q+1}, \dots, C_k$ does not connect with $AC - A_1$ and $AC - A_2$. However, we have $(AC - A_1) \cup (AC - A_2) = A$. So $C_{q+1}, \dots, C_k$ does not connect with AC and $C_1, \dots, C_q$ does not connect with AC. Then it is known that AC and A are not connected. Each AC_k can appear twice in constructing A_i, $1 \leq i \leq k$. Therefore, there exists one way to construct A from S^* such that all the coefficients $\pi_{ij} \leq 1$, for $1 \leq j \leq p - 2$.

Therefore, the optimal solution to the problem R is s_0, and $I^* = T_0$.

A.7 Proof of Lemma 4.8

The proof is similar to the proof in Theorem 4.3. However, this Lemma achieves better rates because the structure of the Core Determining Class is special. For any $\Pi \geq 0$ such that $\Pi M \geq M$, as we show in the proof of Lemma 4.7, the residual $\Pi b \geq b$ can be rewritten as a sum $\sum_{1 \leq l \leq d_2} \alpha_l \mu(l)$ where $\alpha_l > 0$ for all $1 \leq l \leq d_2$. Therefore, when we replace μ with $\hat{\mu}_n$, the residual $\Pi b - b$ and $\Pi \hat{b} - \hat{b}$ are very close. According to the assumption that $\max_{1 \leq l \leq d_2} \frac{|\mu(l) - \hat{\mu}_n(l)|}{\mu} \to 0$, $\Pi b - b = \Pi \hat{b} - \hat{b}(1 + o_p(1))$. Therefore, with probability $\geq 1 - \alpha$, $\Pi^* \hat{b} - \hat{b} = (\Pi^* b - b)(1 + o_p(1)) \leq \lambda_S$. So Π^* is a feasible solution to $\hat{R}$ with probability $1 - \alpha$. The rest of the derivation follows the proof of Theorem 4.3.

Bibliography

[1]. Anderson, J. E. (1998). Control of personal rapid transit systems. Journal of advanced transportation, 32(1), 57-74.

[2]. Andrews, D. W., & Shi, X. (2013). Inference based on conditional moment inequalities. Econometrica, 81(2), 609-666.

[3]. Andrews, D. W., & Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica, 78(1), 119-157.

[4]. Archetti, C., Speranza, M. G., & Hertz, A. (2006). A tabu search algorithm for the split delivery vehicle routing problem. Transportation Science, 40(1), 64-73.

[5]. Artstein, Z. (1983). Distributions of random sets and random selections. Israel Journal of Mathematics, 46(4), 313-324.

[6]. Bajari, P., Benkard, C. L., & Levin, J. (2007). Estimating dynamic models of imperfect competition. Econometrica, 75(5), 1331-1370.

[7]. Bajari, P., Hong, H., & Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. Econometrica, 78(5), 1529-1568.

[8]. Balcik, B., Beamon, B. M., & Smilowitz, K. (2008). Last mile distribution in humanitarian relief. Journal of Intelligent Transportation Systems, 12(2), 51-63.

[9]. Baldacci, R., Maniezzo, V. and Mingozzi, A. (2004). An exact method for the car pooling problem based on Lagrangean column generation. Operations Research, 52(3), 422-439.

[10]. Beardwood, J., Halton, J. H. and Hammersley, J. M. (1959). The shortest path through many points. Mathematical Proceedings of the Cambridge Philosophical Society, 55(4), 299-327.

[11]. Berbeglia, G., Cordeau, J. F., & Laporte, G. (2012). A hybrid tabu search and constraint programming algorithm for the dynamic dial-a-ride problem. INFORMS Journal on Computing, 24(3), 343-355.

[12]. Beresteanu, A., Molchanov, I., & Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. Econometrica, 79(6), 1785-1821.

[13]. Berger, T., Sallez, Y., Raileanu, S., Tahon, C., Trentesaux, D., & Borangiu, T. (2011). Personal Rapid Transit in an open-control framework. Computers & Industrial Engineering, 61(2), 300-312.

[14]. Bertsimas, D. J. and Van Ryzin, G. (1991). A stochastic and dynamic vehicle routing problem in the Euclidean plane. Operations Research, 39(4), 601-615.

[15]. Bertsimas, D. J. and Van Ryzin, G. (1993). Stochastic and dynamic vehicle routing in the Euclidean plane with multiple capacitated vehicles. Operations Research, 41(1), 60-76.

[16]. Bly, P. H., & Teychenne, R. (2005, May). Three financial and socio-economic assessments of a personal rapid transit system. In Proceedings of the tenth international conference on automated people movers (pp. 1-4).

[17]. Boyer, K. K., Prud'homme, A. M., & Chung, W. (2009). The last mile challenge: evaluating the effects of customer density and delivery window patterns. Journal of Business Logistics, 30(1), 185-201.

[18]. Bräysy, O., & Gendreau, M. (2005a). Vehicle routing problem with time windows, Part I: Route construction and local search algorithms. Transportation science, 39(1), 104-118.

[19]. Bräysy, O., & Gendreau, M. (2005b). Vehicle routing problem with time windows, Part II: Metaheuristics. Transportation science, 39(1), 119-139.

[20]. Candes, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2313-2351.

[21]. Caron, R. J., McDonald, J. F., & Ponic, C. M. (1989). A degenerate extreme point strategy for the classification of linear constraints as redundant or necessary. Journal of Optimization Theory and Applications, 62(2), 225-237.

[22]. Chernozhukov, V., Chetverikov, D., & Kato, K. (2012). Central limit theorems and multiplier bootstrap when p is much larger than n (No. CWP45/12). cemmap working paper, Centre for Microdata Methods and Practice.

[23]. Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models1. Econometrica, 75(5), 1243-1284.

[24]. Chesher, A., Rosen, A. (2012). Simultaneous equations models for discrete outcomes: coherence, completeness, and identification, cemmap working paper CWP21/12.

[25]. Cordeau, J. F., Gendreau, M., & Laporte, G. (1997). A tabu search heuristic for periodic and multi-depot vehicle routing problems. Networks, 30(2), 105-119.

[26]. Cordeau, J. F., Gendreau, M., Laporte, G., Potvin, J. Y., and Semet, F. (2002). A guide to vehicle routing heuristics. Journal of the Operational Research society, 512-522.

[27]. Cordeau, J. F., and Laporte, G. (2007). The dial-a-ride problem: models and algorithms. Annals of Operations Research, 153(1), 29-46.

[28]. Cordeau, J. F., Laporte, G., & Mercier, A. (2001). A unified tabu search heuristic for vehicle routing problems with time windows. Journal of the Operational research society, 52(8), 928-936.

[29]. Cordeau, J. F., & Maischberger, M. (2012). A parallel iterated tabu search heuristic for vehicle routing problems. Computers & Operations Research, 39(9), 2033-2050.

[30]. Daganzo, C. F. (1984). The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application. Transportation Science, 18(4), 331-350.

[31]. Dowsland, K. A. (1998). Nurse scheduling with tabu search and strategic oscillation. European journal of operational research, 106(2), 393-407.

[32]. Drexl, A., & Kimms, A. (1997). Lot sizing and scheduling—survey and extensions. European Journal of operational research, 99(2), 221-235.

[33]. Eilon, S., Watson-Gandy, C. D. T., and Christofides, N. (1971). Distribution management: mathematical modelling and practical analysis (pp. 197-200). London: Griffin.

[34]. Esper, T. L., Jensen, T. D., Turnipseed, F. L., & Burton, S. (2003). The last mile: an examination of effects of online retail delivery strategies on consumers. Journal of Business Logistics, 24(2), 177-203.

[35]. Galichon, A., Henry, M. (2006). Inference in incomplete models. Available at SSRN 886907.

[36]. Galichon, A., Henry, M. (2011). Set identification in models with multiple equilibria. The Review of Economic Studies, 78(4), 1264-1298.

[37]. Gendreau, M., Guertin, F., Potvin, J. Y., & Taillard, E. (1999). Parallel tabu search for real-time vehicle routing and dispatching. Transportation science, 33(4), 381-390.

[38]. Gendreau, M., Hertz, A., & Laporte, G. (1994). A tabu search heuristic for the vehicle routing problem. Management science, 40(10), 1276-1290.

[39]. Gómez-Ibañez, J., Tye, W., & Winston, C. (1999). Essays in Transportation Economics and Policy, 42. Brookings Institution Press, Washington D.C..

[40]. Glover, F. (1989). Tabu search-part I. ORSA Journal on computing, 1(3), 190-206.

[41]. Glover, F. (1990a). Tabu search: A tutorial. Interfaces, 20(4), 74-94.

[42]. Glover, F. (1990b). Tabu search—part II. ORSA Journal on computing, 2(1), 4-32.

[43]. Gremlich, R., Hamfelt A., and Valkovsky, V. 2004. Prediction of the Optimal Decision Distribution for the Traveling Salesman Problem. Proceedings of IPSI International Conf., Sveti Stefan, Montenegro.

[44]. Hurink, J., Jurisch, B., & Thole, M. (1994). Tabu search for the job-shop scheduling problem with multi-purpose machines. Operations-Research-Spektrum, 15(4), 205-215.

[45]. Jaw, J. J., Odoni, A. R., Psaraftis, H. N., and Wilson, N. H. (1986). A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. Transportation Research Part B: Methodological, 20(3), 243-257.

[46]. Johnson, D. S., McGeoch, L. A., and Rothberg, E. E. (1996). Asymptotic experimental analysis for the Held-Karp traveling salesman bound. In Proceedings of the seventh annual ACM-SIAM symposium on Discrete algorithms (pp. 341-350). Society for Industrial and Applied Mathematics.

[47]. Kingman, J. F. C. (1961). The single server queue in heavy traffic. In Mathematical Proceedings of the Cambridge Philosophical Society, 57(4), 902-904).

[48]. Kingman, J. F. C. (1962). Some inequalities for the queue GI/G/1. Biometrika, 315-324.

[49]. Köllerström, J. (1974). Heavy traffic theory for queues with several servers. I. Journal of Applied Probability, 544-552.

[50]. Kraemer, W., and Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/1. In Congressbook of the Eight International Teletraffic Congress, pages 2351/8, Melbourne.

[51]. Lee, H. L., & Whang, S. (2001). Winning the last mile of e-commerce. MIT Sloan Management Review, 42(4), 54-62.

[52]. Lee, J. W., Mazumdar, R. R., & Shroff, N. B. (2006). Opportunistic power scheduling for dynamic multi-server wireless systems. Wireless Communications, IEEE Transactions on, 5(6), 1506-1515.

[53]. Lees-Miller, J. D., Hammersley, J. C., & Davenport, N. (2009). Ride sharing in personal rapid transit capacity planning. Automated People Movers 2009, 321-332.

[54]. Lees-Miller, J. D., Hammersley, J. C., & Wilson, R. E. (2010). Theoretical maximum capacity as benchmark for empty vehicle redistribution in personal rapid transit. Transportation Research Record: Journal of the Transportation Research Board, 2146(1), 76-83.

[55]. Lei, H., Laporte, G., and Guo, B. (2011). Districting for routing with stochastic customers. EURO Journal of Transportation and Logistics, 1-19.

[56]. Liu, L., & Liu, X. (1998). Dynamic and static job allocation for multi-server systems. IIE transactions, 30(9), 845-854.

[57]. Liu, Z., Jia, X., & Cheng, W. (2012). Solving the last mile problem: Ensure the success of public bicycle system in Beijing. Procedia-Social and Behavioral Sciences, 43, 73-78.

[58]. Manski, C. F., & Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. Econometrica, 70(2), 519-546.

[59]. Montané, F. A. T., & Galvao, R. D. (2006). A tabu search algorithm for the vehicle routing problem with simultaneous pick-up and delivery service. Computers & Operations Research, 33(3), 595-619.

[60]. Mueller, K., & Sgouridis, S. P. (2011). Simulation-based analysis of personal rapid transit systems: service and energy performance assessment of the Masdar City PRT case. Journal of Advanced Transportation, 45(4), 252-270.

[61]. Newell, G. F. (1971). Applications of queueing theory. Chapman & Hall, London.

[62]. Odoni, A. R., & Roth, E. (1983). An empirical investigation of the transient behavior of stationary queueing systems. Operations Research, 31(3), 432-455.

[63]. Ott, T. J. (1987). Simple inequalities for the D/G/1 queue. Operations research, 35(4), 589-597.

[64]. Paulraj, S., Chellappan, C., & Natesan, T. R. (2006). A heuristic approach for identification of redundant constraints in linear programming models. International Journal of Computer Mathematics, 83(8-9), 675-683.

[65]. Puna, V., T, Lai and Qi, Shao(2009). Self-Normalized Processes: Limit Theory and Statistical Applications. Springer.

[66]. Punakivi, M., Yrjölä, H., & Holmström, J. (2001). Solving the last mile issue: reception box or delivery box?. International Journal of Physical Distribution & Logistics Management, 31(6), 427-439.

[67]. Romano, J. P., & Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. Econometrica, 78(1), 169-211.

[68]. Song, L., Cherrett, T., McLeod, F., & Guan, W. (2009). Addressing the Last Mile Problem. Transportation Research Record: Journal of the Transportation Research Board, 2097(1), 9-18.

[69]. Sridharan, R. (1995). The capacitated plant location problem. European Journal of Operational Research, 87(2), 203-213.

[70]. Swihart, M. R. and Papastavrou, J. D. (1999). A stochastic and dynamic model for the single-vehicle pick-up and delivery problem. European Journal of Operational Research, 114(3), 447-464.

[71]. Taillard, E. (1991). Robust taboo search for the quadratic assignment problem. Parallel computing, 17(4), 443-455.

[72]. Telgen, J. (1983). Identifying redundant constraints and implicit equalities in systems of linear constraints. Management Science, 29(10), 1209-1222.

[73]. Wang, H. (2012). Approximating the Performance of a Last Mile Transportation System, S.M. Thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

[74]. Wang, H., & Odoni, A., (2014) Approximating the Performance of a "Last Mile" Transportation System. Transportation Science.

[75]. Whitt, W. (1983). Approximations for the GI/G/m queue. Production and Operations Management, 2(2), 114-161.

[76]. Zee, D. J. V. D., Harten, A. V., & Schuur, P. (2001). On-line scheduling of multi-server batch operations. IIE Transactions, 33(7), 569-586.