

UC Berkeley

Earlier Faculty Research

Title

Dynamic and Stochastic Routing Optimization: Algorithm Development Analysis

Permalink

<https://escholarship.org/uc/item/0q34937d>

Author

Lu, Xiangwen

Publication Date

2001

UNIVERSITY OF CALIFORNIA,

IRVINE

Dynamic and Stochastic Routing Optimization:

Algorithm Development and Analysis

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Civil Engineering

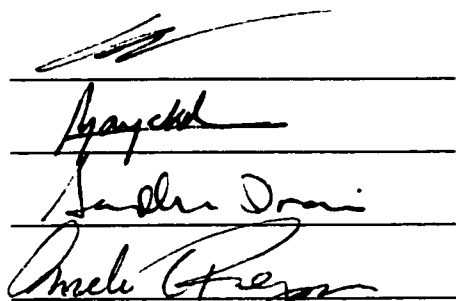
by

Xiangwen Lu

Dissertation Committee:
Professor Amelia C. Regan, Chair
Professor Sandra S. Irani
Professor Wilfred R. Recker
Professor R. Jayakrishnan

2001

The dissertation of Xiangwen Lu is
approved and is acceptable in quality
and form for publication on microfilm:



Three handwritten signatures are stacked vertically, each on a horizontal line. The top signature is a stylized, cursive name. The middle signature is 'Sudhakar D. D. D. D.'. The bottom signature is 'Aneli G. G. G. G.'.

Committee Chair

University of California, Irvine
2001

DEDICATION

To

my wife, Guangyu Zhang, my Parents, Daocai Lu and Xiuying Chang, and to my wife's
parents, Gishun Zhang and Shuhua Yang

DEDICATION

To

my wife, Guangyu Zhang, my Parents, Daocai Lu and Xiuying Chang, and to my wife's
parents, Gishun Zhang and Shuhua Yang

TABLE OF CONTENTS

LIST OF FIGURES	VII
ACKNOWLEDGEMENT	VIII
CURRICULUM VITA.....	X
ABSTRACT.....	XII
CHAPTER 1 RESEARCH OVERVIEW.....	1
1.1 Introduction and Motivation	1
1.2 Problems Examined	3
1.2.1 The Probabilistic Traveling Salesman Problem (PTSP).....	3
1.2.2 The Dynamic Traveling Repairman (DTRP) and Dynamic Traveling Salesman Problems (DTSP).....	4
1.2.3 The M/G/1 Queueing System with Switchover Costs	6
1.3 Literature Review.....	7
1.3.1 Vehicle Routing and Scheduling	7
1.3.2 Problems Examined in this Dissertation	11
1.4 Research Approach	13
1.5 Fundamental Insights.....	15
1.6 Organization of the Dissertation	16
CHAPTER 2 A HEURISTIC FOR THE EUCLIDEAN PROBABILISTIC TRAVELING SALESMAN PROBLEM.....	18
2.1 Introduction.....	18
2.2 Notation.....	20
2.3 A New Heuristic	21
2.3.1 The Related k – Median Problem (capacitated and un-capacitated)	21
2.3.2 Returning to the PTSP: Analysis of the Optimal A Priori Tour	23
2.3.2.1 Clustering According to An Optimal A Priori Tour	23

2.3.2.2 The Radii of the Groups.....	25
2.3.3 A New Algorithm for the PTSP.....	27
2.3.3.1 Sweep Algorithms.....	27
2.3.3.2 The k – Median Problem Φ	28
2.3.3.3 A New Heuristic Algorithm H	28
2.3.4 Properties of Algorithm H	30
2.3.5 Proof of the Lemmas.....	33
2.4 Conclusion	47
CHAPTER 3 THE M/G/1 QUEUE WITH SWITCHOVER COSTS:AN EXAMINATION OF ALTERNATIVE HEURISTICS	48
3.1 Introduction.....	48
3.2 Problem and Notation	51
3.3 Lower and Upper Bounds on the Average Waiting Time for the Optimal Algorithm	53
3.3.1 A Lower Bound for Algorithms Obeying the Continuous Condition	53
3.3.2 A Lower Bound for Algorithms which do not (necessarily) Obey the Continuous Condition	60
3.3.3 General Lower and Upper Bounds for the Optimal Algorithm	65
3.4 Lower and Upper Bound for the Optimal Algorithm Under Light Traffic.....	67
3.5 Simulation.....	70
3.5.1 The Simulation Model	70
3.5.2 Simulation Results	71
3.6 Conclusion	75
CHAPTER 4 THE DYNAMIC TRAVELING SALESMAN PROBLEM: AN EXAMINATION OF ALTERNATIVE HEURISTICS	76
4.1 Introduction.....	76
4.2 The DTSP on Networks in which the Optimal TSP Tour and Minimum Spanning Tree Involve Only Links of Equal Length.....	79
4.2.1 Notation.....	79
4.2.2 The DTSP on the Special Graph.....	80
4.3 The DTSP on a General Graph	81
4.3.1 A Heuristic Algorithm	81
4.3.2 Properties of the Heuristic Algorithm.....	81
4.4 The DTSP on the General Graph under Light Traffic Intensity	82

4.5 Simulation Results	85
4.6 Conclusion	89
CHAPTER 5 AN ASYMPTOTICALLY OPTIMAL ALGORITHM FOR THE DYNAMIC TRAVELING REPAIRMAN PROBLEM	91
5.1 Introduction.....	91
5.2 Algorithms for Single Server Case	93
5.2.1 Definition and Notation	93
5.2.2 Policies of Interest.....	94
5.2.3 General Partition Algorithms.....	95
5.2.4 Properties of General Partition Algorithms	96
5.2.5 The Small Partition Case	98
5.2.6 The Fixed Partition Case.....	102
5.2.7 An Asymptotically Optimal General Partition Algorithm.....	103
5.2.8 The Optimality of $BvR(n, k)^*$ Among the General Partition Class	104
5.3 The m-Server Case.....	105
5.4 Proof of the Theorems	106
5.4.1 Proof of the Propositions and Lemmas.....	106
5.4.1.1 The Smoothing Technique.....	107
5.4.1.2 Lemmas and Theorem BHM	108
5.4.1.3 Proof of Propositions	109
5.4.1.4 Proof of Lemmas.....	112
5.4.2 Proof of the Small Partition Case.....	124
5.4.3 Proof of the Fixed Partition Case.....	125
5.5 Conclusion	129
CHAPTER 6 CONCLUSION AND FUTURE RESERACH.....	131
6.1 Conclusion	131
6.2 Future Research	133
REFERENCES	135

LIST OF FIGURES

Figure 3.1 Closed Form Solution vs. Simulation Results.....	71
Figure 3.2 A Comparison of Three Heuristics when the Switching Costs are Constant...	73
Figure 3.3 The relative performance of cyclic polling and longest queue first for ρ approaching 1.....	73
Figure 3.4 The benefit of allowing the server to be patient when no customers are in the system.....	74
Figure 4.1. Example networks.....	78
Figure 4.2. Randomly generated six node networks with length of optimal TSP tour of 6 units.....	87
Figure 4.3. A Comparison of Three Heuristics when the Switching Costs are Proportional to Distance (randomly generated 6 node networks).....	88
Figure 4.4. An examination of the point at which Cyclic Polling outperforms Longest Queue First when travel time is proportional to Distance (randomly generated 6 node networks)	88

ACKNOWLEDGEMENT

I would like to express my deepest appreciation and gratitude to Professor Amelia Regan for her immeasurable amount of inspiration, guidance, direction and help in the development of my research and studies and, for her constant encouragement. I also would like to express my deepest appreciation to Professor Sandra Irani for her constant guidance, inspiration and help in development of my research.

I would like to thank my committee members, Professors Wilfred Recker and R. Jayakrishnan for their interest in this work and for providing valuable comments on previous drafts.

I would like to thank Jiri Herrman for providing support for the development of simulation models used in this research.

I would like to thank Dr. Xiubin Wang for the lively research discussions we had while he was a student at ITS. I also thank him for introducing me to this topic area and encouraging me to apply to UCI.

This research was partially supported by the University of California Transportation Center and the National Science Foundation. This support is gratefully acknowledged.

I would like to thank to all the students and faculty at the Institute of Transportation Studies, University of California at Irvine.

I would also like to take this opportunity to acknowledge the support I received during my earlier studies at Beijing University and my time as an instructor at Northern Jiaotong University in China. Special thanks are due to Professors Yuanpei Liu, Zhongguo Zheng, Yuke Meng.

Finally, I want to express my deepest feeling of love to my wife, my parents and to her parents for their support and love, they are my center of life.

CURRICULUM VITA

Xiangwen Lu

Education

- 1991 B. Sc. In Probability and Mathematical Statistics, Beijing University,
Beijing, P.R.China
- 1998-2001 1994 M. Sc. In Probability and Mathematical Statistics, Beijing
University, Beijing, P.R.China
- 2001 Ph.D. in Civil Engineering, University of California, Irvine

Employment

- 1994-1998 Instructor, Department of Mathematics, Northern Jiaotong
University, Beijing, P.R. China

Graduate Research Assistant, University of California, Irvine

Research Interests

Transportation, Logistics, Supply Chain Management and Inventory Control, Revenue
Management, Stochastic Process and Statistics, Queueing Theory

Awards

Doctoral Dissertation Fellowship, awarded by the University of California Transportation
Center, 2000-2001.

Publications and Working papers

Irani, S. X. Lu and A.C. Regan, The Online Algorithms for the Dynamic Traveling Repairman Problem, Proceedings of the 2002 Symposium on Discrete Algorithms SODA, under review, 2001.

Lu, X., A.C. Regan, S. Irani, The $M/G/1$ Queue with Switchover Costs: An Examination of Alternative Heuristics, Queueing Systems, under review, 2001.

Lu, X., S. Irani, A.C. Regan, The Dynamic Traveling Salesman Problem: An Examination of Alternative Heuristics, Transportation Science, under review, 2001.

Lu, X., A.C. Regan, S. Irani, An Asymptotically Optimal Algorithm for the Dynamic Traveling Repair Problem, Networks, under review, 2001.

Lu, X., S. Irani, A.C. Regan, A Heuristic Algorithm for the Probabilistic Traveling Salesman Problem, Mathematics of Operations Research, under review, 2001.

Regan, A., J. Herrmann and X. Lu, The Relative Performance of Heuristics for the Dynamic Traveling Salesman Problem, Transportation Research Record, under review, 2001.

ABSTRACT

Dynamic and Stochastic Routing Optimization: Algorithm Development and Analysis

By

Xiangwen Lu

Philosophy of Doctor in Civil Engineering

University of California, Irvine, 2001

Professor Amelia C. Regan, Chair

The last several years has witnessed a sharp increase in interest in stochastic and dynamic routing and scheduling. Because many systems contain inherently stochastic factors, decisions must often be made before all necessary information is available. To a certain degree, algorithm development has lagged behind implementation. In order to fully leverage advances in information technologies, algorithms which explicitly consider dynamic and stochastic factors should be examined. Or, if static algorithms are to be applied in these dynamic environments, proper attention should be given to examining the conditions under which these perform well. This is the primary theme of this research.

This dissertation examines several key dynamic and stochastic routing and scheduling problems: the probabilistic traveling salesman problem, the dynamic traveling salesman

problem and the dynamic traveling repair problem. In addition, as part of our research on the dynamic traveling salesman problem, we examine a related $M/G/1$ queueing problem with switching costs. These problems arise in pickup and delivery operations, repair fleet operations, and emergency vehicle and police operations in addition to many computing, telecommunications and manufacturing applications.

As part of our research, we demonstrate that heuristics which rely on partitioning the service region into smaller regions can be very effective for dynamic routing problems. Using a partitioning scheme we show that if a constant guarantee algorithm exists for the k -capacitated median problem, then a constant guarantee algorithm exists for the probabilistic traveling salesman problem. For the DTRP, we show that a partitioning algorithm is asymptotically optimal when the traffic intensity is high.

We show that robust a priori algorithms can be developed for dynamic routing problems. For the $M/G/1$ with switchover cost, we show that an a priori cyclic polling algorithm works very well using both theoretical and simulation analysis. Cyclic polling algorithm also works well for dynamic traveling salesman problem. For these both problems, we identify certain conditions under which the a priori (cyclic polling) solution is close to optimal. We demonstrate that the existence of the connection between the static and dynamic vehicle routing and scheduling problem that have been observed by earlier researchers.

CHAPTER 1 RESEARCH OVERVIEW

1.1 Introduction and Motivation

The last several years has witnessed a sharp increase in interest in stochastic and dynamic routing and scheduling. Several factors are driving this trend. The first is increasing opportunities to use on line decision techniques to take full advantage advances in the computer and communication technologies in the area of transportation, telecommunications and logistics. Steadily increasing computational power and the availability of information technologies capable of providing dispatchers with real-time updates on the status and location of vehicles and customers as well as traffic network conditions make real time decision making feasible. A second factor is an increase in customer service expectations, driven in part, by the promises made by a few aggressive companies (FedEx, Cox Cable, for example).

Because many systems contain inherently stochastic factors, decisions must often be made before all necessary information is available. Stochasticity in these systems takes many forms. In freight and fleet management systems the timing, location and level of demand may vary; the availability of resources available to meet demand can also vary due to traffic conditions, breakdowns, service times and dockside waiting times; and, operations in one zone, regions or area of operation can impact those in other zones. Once made, these decisions must be modified as new information is known. Increases in customer service expectations, the availability of real-time information on traffic network

conditions and the emergence of on-line freight marketplaces have all led to a dramatic rise in the number of freight and fleet management systems operating under explicitly dynamic conditions. To a certain degree, algorithm development has lagged behind implementation. Typical real-time fleet management systems employ static algorithms developed in the 1980's and early 1990's. In order to fully leverage advances in information technologies, algorithms which explicitly consider dynamic and stochastic factors should be examined. Or, if static or a priori algorithms are to be applied in these dynamic environments, proper attention should be given to examining the conditions under which these perform well on dynamic problems. Such an examination is the key theme of this research.

There are two main classes of dynamic and stochastic routing and scheduling models. The first class relies on a priori optimization in which solutions are generated for stochastic problems prior to the receipt of information regarding the realization of their random elements. The general approach is to generate an a priori solution that has the least cost in the expected sense. The second class of models involves making decisions and observing outcomes on a continuous, rolling horizon basis.

This dissertation examines several key dynamic and stochastic routing and scheduling problems. We examine both classes of models. The focus of this study is more theoretical than applied, though theoretical analyses of this sort can often lend significant insight to the development of practical applications. The main problems examined in this research are the probabilistic traveling salesman problem, the dynamic traveling salesman

problem and the dynamic traveling repair problem. In addition, as part of our research on the dynamic traveling salesman problem, we examine a related dynamic network problem, the M/G/1 queueing problem with switching costs. These problems arise in pickup and delivery operations, repair fleet operations, and emergency vehicle and police operations in addition to many computing, telecommunications and manufacturing applications. These problems are network optimization problems and are at the very heart of research on dynamic and stochastic logistics systems analysis.

1.2 Problems Examined

1.2.1 The Probabilistic Traveling Salesman Problem (PTSP)

We consider the following situation: A company (UPS, FedEx, for example) has a large number of delivery customers. Each driver is responsible for a given delivery region. The driver's knowledge of the route is an important factor affecting his or her efficiency and the level of customer service provided. In each service area there are many potential customers. However, in each day, only a subset of these potential customers require a delivery. The probabilistic traveling salesman problem is to design an a priori route for each driver, in which the route is followed exactly, simply skipping customers not requiring a visit. The goal is to find a priori tour with the least expected cost.

The PTSP is a very important problem. It represents a strategic planning model in which stochastic factors are considered explicitly. These types of problems are of central importance in many logistics and transportation planning applications in which heuristics

which perform well over a wide range of demand realizations are required or where re-optimization may be infeasible either for computational or operational reasons. In addition, of key interest is the examination of the robustness of optimal solutions for deterministic problems when the instances upon which these problems have been solved are modified.

The PTSP is well known to be NP-hard. In this dissertation we develop and examine a new class of heuristics for this problem. Depending upon the methods used to solve necessary sub-problems, these heuristics can have polynomial or quasi-polynomial performance. We show that one of the quasi-polynomial heuristics is a c -approximation heuristic. We demonstrate that it has good average case performance and that its worst case performance is bounded.

1.2.2 The Dynamic Traveling Repairman (DTRP) and Dynamic Traveling Salesman Problems (DTSP)

Assume that a utility firm (electric, gas, cable TV etc.) is responsible for maintenance over a large, geographical area. Failures occur over time throughout the geographic region. The firm operates a fleet of repair vehicles and technicians. Routing decisions are made over time based on real time information about the location of known failures and some prediction about future outages. The vehicles travel to customer locations to provide on site service. On site repair times are stochastic. The objective of the firm is to minimize the average waiting time experienced by customers.

A closely related application is real time product delivery (furniture, industrial gasses, home heating oil, etc.) in which delivery vehicles are dispatched with the objective of minimizing some combination of the delivery cost and customer waiting times. Another important application in the transportation area is the dispatch of highway patrol vehicles, where the highway patrol is assumed to be responsible for the control of a large freeway network. Estimates of the proportion of freeway delay in the U.S. attributable to the non-recurring (incident related) congestion range up to 60% and it is believed that this proportion is increasing (Lindley, 1987). Incidents occur at random over both time and space. Response time is critical for the clearance of incidents. The dispatcher's decision concerns the optimal location of the patrol vehicles if there is no incident, and, which patrol vehicles to dispatch if there is an incident. The objective is to minimize response time.

Additional important examples can be found in goods distribution and freight consolidation. Consider, for example, the freight consolidation problem of less-than-truckload carriers and package delivery services. The dispatch centers receive parcel loads designed for specific locations in service region. These parcel loads are queued and consolidated into full truckloads for delivery. Operators face the challenge of balancing operating costs (travel cost plus vehicle and driver cost) and service levels (the waiting time for service). In order to reduce customers' waiting time, vehicles should be dispatched more frequently at a higher cost; to reduce operational costs loads should be allowed to accumulate in the dispatch center so that denser delivery routes can be developed, resulting in higher customer waiting times.

The dynamic traveling salesman problem, is similar to the dynamic traveling repair problem, except that customer locations are limited to a set of known nodes, rather than any point in the Euclidean space. Please note that these are the historical definitions of these problems, which are generally accepted by the transportation research community. Some researchers have made different distinctions between these problems, for example, arguing that the dynamic traveling repair problem involves a mobile service fleet with non-zero service times over any metric space and that the dynamic traveling salesman problem involves a mobile fleet with instantaneous services times over any metric space. In our research, we adopt the historical (albeit imperfect) definitions which separate these two problems based on their discrete (DTSP) or continuous (DTRP) metric spaces.

1.2.3 The M/G/1 Queueing System with Switchover Costs

Consider the following situation: a single server provides service to n locations. At any point in time, each location may or may not have one or more waiting customers. The server can provide service to only one location at a time and must spend some time traveling between locations, thereby incurring a switchover cost.

Traffic control provides one example of an application of M/G/1 queueing systems with switchover costs. In order to minimize the average waiting time of the customers in the system, the controller must decide when to switch and where to switch to. Switching here means the signal switches from red to green or green to red. The problem is how long to allow the green and red time in each direction. Each time the light switches from red to green or green to red, there is a cost incurred during the time in which all directions are

temporarily set to red. More frequent switching decreases the service level of all drivers. On the other hand, less frequent switching can increase the waiting time of drivers in one direction or the other. More obvious examples of applications arise in telecommunications and computing systems. For example, a time-shared computer system consists of n terminals connected to the center computer. Data must be transferred between the terminals and the computer. The objective is to provide good service to all the terminals. In many manufacturing environments, a single facility may produce several different products. At any given time, the facility can produce only one product. A set-up cost is incurred each time it changes between products. Decisions about when to switch between products and in what order to switch affect the performance of the system. To model all these kind of situations, we consider the $M/G/1$ queueing system with switchover costs.

1.3 Literature Review

1.3.1 Vehicle Routing and Scheduling

Vehicle routing and scheduling involves finding a set of one or more routes to various demand or activity locations, in order to minimize a cost function (minimizing routing costs or a combination of fixed costs and routing costs). Vehicles may have capacity constraints. If the problem has unit demands, a single vehicle without capacity constraints, and the objective function is to minimize the total travel cost, then the vehicle routing problem reduces to the well known Traveling Salesman Problem (TSP). Fisher (1995) presents a recent review of the routing and scheduling literature while Desrosiers,

Dumas, Solomon and Soumis (1995) examine the special case in which time windows are considered.

The TSP is the probably the most extensively studied combinatorial optimization problem. For detailed comprehensive reviews of research, results and applications of the TSP, please refer to Lawler, Lenstra, Rinnooy Kan and Shmoys (1985) or to Laporte (1992). Two different research approaches are typical for the TSP and its variants.

On the side of exact methods, there has been a vast literature aimed at the development of various integer programming-based approaches. Typical research begins by presenting an integer programming formulation for the problem. The various formulations differ with respect to the sub-tour elimination methods used. The next step in solving these formulations is to develop appropriate relaxation and branch and bound schemes.

Heuristics for the TSP are typically classified as tour construction procedures or tour improvement procedures. Tour construction procedures involve building a solution by adding nodes to a partial tour. The best known tour construction heuristics are nearest neighbor and insertion heuristics. Tour improvement procedures involve improving a feasible solution by performing various exchanges. The best known of these are the 2-opt and 3-opt and 1-shift heuristics. In recent years, effective composite heuristics have been developed, see for example, Golden and Steward (1985).

For the general vehicle routing problem, typical solution approaches is to modify math programming based methods developed for the TSP or to use cluster first and route second heuristics, route first and cluster second heuristics.

One class of methods used to analyze the performance of heuristic algorithms is worst-case analysis. Worst case performance analysis is known as competitive analysis in the on-line algorithm literature (see for example Irani and Karlin, 1997). These techniques consider every possible input to the algorithm. The goal is to compare the ratio between the cost of the output of the heuristic algorithm and the cost of the output of the optimal algorithm. If the ratio of these costs can be bounded from above by some constant number c , we call the algorithm a c -constant guarantee algorithm or c -approximation algorithm. For example, the Christofides' TSP algorithm is $\frac{3}{2}$ -approximation algorithm. Recently, Arora used a very complex partitioning scheme to obtain a $(1 + \varepsilon)$ -approximation algorithm for the Euclidean TSP. This is a fully polynomial time approximation scheme. For any ε , the running time is a polynomial function of $1/\varepsilon$ and the cost of the output of the algorithm is within $(1 + \varepsilon)$ of the optimal. This is the best worst case bound to date for heuristics for the TSP. Worst case analysis provides a guarantee on the maximum relative difference between the solution of the heuristic algorithm and the optimal solution for any possible problem instance, even those instances that are not likely to occur in practice. It does not always accurately reflect the typical performance of algorithms. To overcome this drawback average case analysis is performed. In average case analysis, problem instances are generated based on some

probabilistic distributions and comparisons of the performance of the heuristics relative to that of the optimal methods or to other heuristics are performed. For the literature on this type of analysis for the TSP please refer to Karp and Steele (1985).

In practice, a third kind of performance analysis is also important, namely empirical analysis. Empirical analysis relies on examining the performance of heuristics on a set of classical or typical problems. One surprising result of empirical analysis of heuristics for the TSP is that one of the best algorithms is very simple - simply use a nearest neighbor tour construction procedure and then perform various local tour improvement procedures.

In the real world, many factors have a random component. Therefore, there is a need to analyze the influence of these stochastic factors in routing and scheduling solution methods. Routing and scheduling vehicles in the face of the uncertainty of demand is not new. The earliest explicit treatment of stochastic demands in the design of the vehicle tours is found in Tillman (1969) which presented a modified Clarke-Wright savings heuristic for Poisson-distributed demands. Later Stewart and Golden (1983) formulated the stochastic vehicle routing problem using both a chance-constrained and penalty function approach. Both formulations rely on stochastic programming. Stochastic programming has some inherent difficulties. Chance constrained formulations have some serious limitations. These are formulations in which the parameters of the problems are random variables and for which a solution must satisfy the constraints in the probabilistic sense. In our opinion, this leads to solutions generated from restrictive (pessimistic) assumptions. Penalty function formulations suffer from the size limitations and also

require the selection of a problem dependent penalty function. An excellent review of both chance constrained and penalty function formulations can be found in Dror, Laporte and Trudeau (1989).

1.3.2 Problems Examined in this Dissertation

Our research involves explicit consideration of the uncertainty of demand. The PTSP, DTSP and DTRP are fundamental dynamic and stochastic network problems. There are other well known dynamic and stochastic network problems. For example, the dynamic shortest path problem (Cooke and Halsey, 1966); the stochastic shortest path problem (Psarafits and Tsitsiklis, 1993); the dynamic vehicle allocation problem (Powell, 1986); and the dynamic traffic assignment problem (Friesz, Luque, Tobin and Wie, 1989).

In the PTSP we assume that n nodes are spread over a bounded area. Each node has a given probability of requiring a visit. We refer to the probability of requiring a visit as the coverage probability. We assume service requirements are independent across nodes. The goal is to find an a priori tour with the least expected length. The a priori tour is one in which the nodes are visited in the order given by the tour, and those not requiring a visit are simply skipped. This problem was first defined and examined by Jaillet (1985, 1988). Later, it was examined by Bertsimas, Jaillet and Odoni (1990), Bertsimas (1992), Bertsimas and Howell (1993), Jaillet (1993), Bertsimas, Chervi and Peterson (1995). Bertsimas (1989) also examines the related probabilistic traveling salesman location problem. For a survey of research on these problems, please refer to Jaillet and Odoni (1988), Powell, Jaillet and Odoni (1995) or to Bertsimas and Simchi-Levi (1996).

Probabilistic analysis of the TSP, which is closely related to the PTSP has been extensive. In probabilistic analysis of the TSP problem, it is assumed that node locations are independently generated according to a common probability distribution. The objective is to analyze the length of the optimal TSP tour. The most important result of this kind was developed by Beardwood, Halton and Hammersley (1959). Later their method was generalized by Karp and Steele (1985). Both groups of researchers used partitioning schemes to examine the length of the optimal TSP tour.

The dynamic traveling salesman problem concerns the development of a routing policy for a single mobile server providing service to customers whose positions are known. Service requests are generated according to a Poisson process which is uniform across customer locations. The DTSP was first introduced by Psaraftis (1988). Bertsimas and van Ryzin (1991) studied a similar problem, the dynamic traveling repairman problem in which customer locations are either uniformly distributed in a bounded area in the Euclidean plane or distributed according to a distribution with probability density function $f(x)$. This problem was intensively examined by Bertsimas and van Ryzin (1991, 1992, 1993) and reviewed by Bertsimas and Simchi-Levi (1996) and Powel, Jaillet and Odoni (1995).

To facilitate the analysis of the DTSP, our research examines a closely related queueing system, the M/G/1 queue with switchover cost. Previous research on this problem has mainly focused on the examination cyclic polling algorithms and the characterization of the optimal algorithm for this problem. Under cyclic polling algorithms, nodes are visited

according in a predetermined cycle. Cooper, Niu and Srinivasan (1996) developed an explicit expression for the average waiting time under gated or exhaustive policies. Hofri and Ross (1987) studied the case of two queues and conjectured that the optimal algorithm will be of a threshold type. Liu, Nain and Towsley (1992) and Duenyas and van Oyen (1996) partially characterized the optimal algorithm.

Note that in the on line literature, Ausiello, E. Feuerstein, S. Leonardi, L. Stougie, M. Talamo (2001) examined the on line travelling salesman problem with arbitrary input to minimize the makespan (the time at which the tour is completed). They give a 2.5-constant guarantee algorithm for all continuous metric spaces. Irani, Lu and Regan (2001) perform an analysis of the DTRP with uniform time windows over arbitrary inputs in which the objective is to serve as many customers as possible and in which the on site service time is zero or near zero. That paper presents several heuristic algorithms and their worst case analysis. In addition they show that a simple algorithm for the DTRP with uniform time windows is a constant guarantee algorithm.

1.4 Research Approach

Our research begins with an examination of the probabilistic traveling salesman problem. We develop a quasi-polynomial c -approximation algorithm for this problem. We use probabilistic analysis to examine the performance of our heuristic algorithm and the relationship between the PTSP and the k -median problem. We show if there exists a c_1 -approximation algorithm for the k -capacitated median problem, then we can identify a c_2 -approximation algorithm for the PTSP where c_2 is a function of c_1 .

We then move to an examination of the dynamic traveling salesman and repairman problems. For these problems, we are concerned with when to visit each node and how long to remain during each visit, in addition to which node(s) to use as a depot when no demands are in the system. The goal of the decision maker in these problems is to strike a balance between the present (known demands) and the future (uncertain demands) to minimize the average waiting time for the customers.

We combine stochastic optimization, queuing theory and deterministic optimization to develop and analyze the performance of algorithms for the DTSP and DTRP. Our analysis provides bounds for the average waiting time for customers under various heuristics. We provide lower and upper bounds on the average waiting time in these systems for the DTSP, DTRP and the M/G/1 queueing model with switchover costs, and on the average distance traveled per demand served for the DTRP.

Asymptotic analysis of algorithms for vehicle routing and scheduling has become a commonly accepted approach. It has been demonstrated in many cases that algorithms which perform well in the limit also perform well under typical conditions with respect to congestion and problem size. However, these methods are not without their critics (see for example Psaraftis, 1984). For example, the rate of convergence to the optimal solution can be very slow. For this reason, we compliment our analytic results, which are the key contribution of this dissertation, with simulation based analysis of algorithms for the dynamic traveling salesman problem and the M/G/1 queueing system with

switchover costs. We demonstrate for these problems that algorithms with good asymptotic properties also have good empirical performance.

1.5 Fundamental Insights

In this dissertation research, we demonstrate that partitioning algorithms are very useful for solving complex problems. These methods reduce large problems to a series of smaller problems, which can often be solved optimally. When the size of the problem is very large, the loss due to partitioning, relative to the globally optimal solution, can be quite small. In fact, Karp (1985) showed that for the TSP, by partitioning the region and then using dynamic programming to obtain the optimal TSP solution for each small problem and combining these solutions leads to a near optimal solution. Later Arora (1997) developed a $(1 + \varepsilon)$ -approximation partitioning based algorithm for the Euclidean TSP problem. Spaccamela, Rinnooy Kan and Stougie (1984) demonstrated that a partition based heuristic is asymptotically optimal for a class of hierarchical vehicle routing problems. In these problems the first level involves the decision about the number of vehicles needed based on probabilistic information about the locations of future customers. The second level decision is to route the vehicles to provide service after customers have materialized. The goal is minimize a combination of the vehicle acquisition costs and the length of the longest route assigned to any vehicle. In our research we show that if a constant guarantee algorithm exists for the k -capacitated median problem, then a constant guarantee algorithm also exists for the PTSP. We base this claim on a partitioning algorithm. In addition, for the DTRP, we show that a partitioning algorithm is asymptotically optimal.

Another important finding is that robust a priori algorithms can be developed for dynamic routing problems. We show that for the M/G/1 with switchover cost, a cyclic polling algorithm works very well using both theoretical (mathematical) analysis and simulation modeling. We also show that cyclic polling works well for the dynamic TSP. We identify certain conditions under which the a priori (cyclic polling) solution is close to optimal. The fact that dynamic problems, under certain condition(s) the a priori solution is close to optimal, provides an indication of the connection between the static and dynamic solutions to dynamic or on-line problems.

Our research indicates that if we use static vehicle routing methods properly, we can obtain good or even near optimal solutions for dynamic problems. For the dynamic traveling repairman problem, we show that partition based heuristic that lets the customer accumulate in a region and uses a optimal TSP tour to serve the customers within the region is asymptotic optimal algorithm. This is very encouraging since it implies that the exact and heuristic algorithms and insights that have been developed over years for static vehicle routing and scheduling problems can be used to solve and analyze dynamic routing and scheduling problems. The problem becomes how to identify the best methods for each dynamic problem.

1.6 Organization of the Dissertation

In chapter 2, we examine the probabilistic traveling salesman problem. We introduce the problem, present previous results and the relevant literature and then present our heuristic and its properties. The dynamic traveling salesman problem is the subject of chapter 4.

To facilitate our examination, we first examine the related $M/G/1$ queueing system with switchover costs. Because these results are of independent interest, we present these separately in chapter 3. In that research we develop a lower bound on the average waiting time for the optimal algorithm. We also get an upper bound on the average waiting time for the optimal algorithm from Cooper, Niu and Srinivasan (1996). Then, we identify circumstances in which our lower bound is tight under heavy traffic intensity, implying that the cyclic polling algorithm is near optimal under some circumstances. We also provide an algorithm for low traffic intensity and show that the proposed algorithm is near optimal under those conditions. In chapter 4, we apply results obtained in the analysis of the $M/G/1$ queueing system with switchover cost to the DTSP. We examine special networks in which the optimal TSP tour and the minimum spanning tree involve only links of equal length. For these special networks, we obtain lower and upper bounds for the average waiting time for optimal algorithms. The ratio of the upper bound and lower bound is bounded by approximately 2. There are situations in which the ratio is near 1. Then, for general networks, we provide different lower and upper bounds on the average waiting time for service. For light traffic conditions, we identify an alternative algorithm and demonstrate its asymptotic optimality. This result has implications for the optimal location of emergency vehicles. In chapter 5, we focus on the dynamic traveling repairman problem. We prove the asymptotic optimality of a specific algorithm for this problem. Chapter 6 provides a conclusion and discusses potential future research on dynamic and stochastic network optimization problems.

CHAPTER 2 A HEURISTIC FOR THE EUCLIDEAN PROBABILISTIC TRAVELING SALESMAN PROBLEM

2.1 Introduction

The Probabilistic Traveling Salesman Problem is a fundamental stochastic network problem. Assume that n nodes are spread over a bounded area. Each node has a given probability of requiring a visit. We refer to the probability of requiring a visit as the *coverage probability*. We assume service requirements are independent across nodes. The goal in this problem is to find an priori tour with the least expected length. The a priori tour is one in which the nodes are visited in the order given by the tour, and nodes not requiring a visit are simply skipped. This problem was first defined and examined by Jaillet (1985, 1988). It is well known to be NP-hard.

In our research, we examine good heuristic algorithms for this problem. First we concentrate on a special case in which all nodes have the same coverage probability for a while. This is only for the reason of easy explanation. Our research on this topic focuses on the general case. Many good heuristic algorithms for the Traveling Salesman Problem (TSP) have been developed over the years. It is well known that when the coverage probability is high, that these can produce a good PTSP tour. However, when the coverage probability is low, heuristics for the TSP produce poor PTSP tours. Bertsimas, Jaillet and Odoni (1990) show that for some problem instances, the ratio of the expected length of the PTSP tour produced using the optimal TSP tour and the expected length of

the optimal PTSP solution approaches infinity as np approaches infinity (n here is the number of nodes and p is the coverage probability). The rate at which this ratio approaches infinity is $O(n)$. Later Bertsimas and Grigni (1989) studied the performance of spacefilling algorithms for the TSP and show that the ratio between the expected length of the tours produced using these heuristic algorithms and the expected length of the PTSP solution will approach infinity at the rate of $O(\log(n))$. For the worst case, the bound of $O(\log(n))$ can be achieved.

As before, if for any input, the cost of the output of a heuristic algorithm can be bounded by a constant c times the cost of the output of the optimal algorithm, we say that the heuristic algorithm is a constant guarantee algorithm or c -approximation algorithm. The terminology algorithm also applies to the non-constant bounds, i.e. $O(\log(n))$ -approximation algorithm. The existence of a constant guarantee algorithm for the PTSP remains an open question. In this paper, we address the existence of c -approximation algorithms for the PTSP. We show that if there exists a c -approximation algorithm for the capacitated k -median problem, we can find a c -approximation algorithm for PTSP. Our primary contribution is to show that there exists a reduction from the PTSP problem to the k -median problem. Garey and Johnson (1979) provides a thorough discussion of the development of reductions from one NP-hard problem to another.

The PTSP problem is a very important issue for both theoretical research and practical practice. It provides one way to explicitly explore the structure of the problem and the

effect of the stochastic factors. As stated by Jaillet (1993) and Bertsimas et al. (1993), there are two main motivations for the examining the PTSP problem. The first is the desire to formulate and analyze models that are appropriate for real world problems, in which randomness is not only present but a major concern. The second is an interest in investigating the robustness of optimal deterministic solutions to applied to stochastic problems. There are also many other probabilistic combinatorial optimization problems in which the PTSP plays a fundamental role.

This chapter is structured in the following way: first we present the formal definition of the problem followed by the heuristic for the PTSP and its properties. Later we give the proof of the results and the conclusion.

2.2 Notation

Let S be the set of all the nodes representing the potential customers. In each problem instance, only the nodes that belong to a subset s need to be visited. The probability that the set s requiring a visit is exactly given by $p(s)$. Assume we have an priori tour τ , let $L_\tau(s)$ be the length of the tour in which we visit all the nodes in s according to the order of the priori tour, skipping nodes that do not require a visit. $E[L_\tau]$ represents the expected length of the a priori tour τ , $E[L_\tau(S)] = \sum_{s \in S} p(s)L_\tau(s)$. $E[L_{PTSP}]$ represents the expected length of optimal a priori tour, by definition, $E[L_{PTSP}] = \min_\tau \{EL_\tau\}$.

Letting $E[L_H]$ represent the expected length of the tour produced using a heuristic algorithm H , $L_{TSP}(s)$ represent the length of the optimal TSP tour over nodes in s , $E[\Sigma]$ represent the expected length of the tour produced using a re-optimization technique in which the optimal TSP tour is produced after the problem instance is known, $E[\Sigma] = \sum_s p(s)L_{TSP}(s)$.

For a given heuristic algorithm H , if there exists a constant c such that for any set S , the following holds, $\frac{E[L_H]}{E[L_{PTSP}]} \leq c$, we call the heuristic H a constant guarantee heuristic.

2.3 A New Heuristic

Before we introduce the new heuristic, we first introduce the k – (Capacitated) median problem. Then we present the heuristic and analyze its properties.

2.3.1 The Related k – Median Problem (capacitated and un-capacitated)

The k – median problem

There exist n nodes each with demand d_i that must be served by one or more service facilities. The problem of finding the optimal locations for the k service facilities, when the set-up cost for the facilities is zero, is known as the k – median problem.

If we select medians $\{1, 2, \dots, k\}$, the total cost to serve all the demands is given by

$$\left\{ \sum_j d_{i,j} \min_{\{i, \text{median } i\}} c_{ij} \right\}, \text{ where } d_{i,j} \text{ is the size of the demand from node } j \text{ served from}$$

median i ($\sum_i d_{i,j} = d_j$) and $c_{i,j}$ is the distance from median i to node j . To find the best

k medians to minimize the cost to serve all the nodes is called k -median problem.

k – Capacitated Median Problem

If each median (facility) i has a capacity constraint C_i (i.e. $\sum_j d_{i,j} \leq C_i$, where $d_{i,j}$ is the

size of the demand from node j served from median i). The problem of finding the best

k such medians is called k -Capacitated Median Problem. The objective is to select k

locations to build the facilities and serve all the demands at the minimum cost.

Both of these two problems are NP-hard problems. As mentioned earlier, if we let

$C^*(S)$ be the optimal cost for the optimal algorithm on the problem instance S and

$C_H(S)$ be the cost under heuristic algorithm H for the problem S . If there exists c ,

such that for problem instance S , if $\frac{C_H(S)}{C^*(S)} \leq c$ always holds, then we call the heuristic

algorithm H a c constant guarantee heuristic algorithm or c -approximation algorithm.

Arora et al. (1998) obtain a $(1 + \epsilon)$ -approximate algorithm for the k -median problem.

Charikar et al. (1999) present a 4-approximation for k -median problem. For the

k -capacitated median problem, there are no known polynomial time c constant

approximation algorithms. Arora et al. (1998) obtain a quasi-polynomial time $(1 + \varepsilon)$ -approximate algorithm for the k -capacitated median problem.

2.3.2 Returning to the PTSP: Analysis of the Optimal A Priori Tour

2.3.2.1 Clustering According to An Optimal A Priori Tour

First, suppose we know the optimal PTSP solution over n nodes which is an a priori tour

$\tau^* = (X_1, X_2, \dots, X_n, X_1)$ with corresponding coverage probabilities p_1, p_2, \dots, p_n .

We select a parameter β which is bigger than one. We cluster the nodes according to their order in the a priori tour and their coverage probabilities in the following way:

Step 1. Let $m = \max \left\{ \left\lfloor \frac{\sum_i p_i}{\beta} \right\rfloor, 1 \right\}$, where $\lfloor x \rfloor$ represents the largest integer not

exceeding x (commonly called the floor function).

Step 2. If $\sum_i p_i \leq \beta$, all the nodes are clustered in one group, otherwise we follow step

three.

Step 3. Select k nodes until $\sum_{i=1}^{i=j-1} p_i < \frac{\sum p_i}{m}$ and $\sum_{i=1}^{i=j} p_i \geq \frac{\sum p_i}{m}$, if $\sum_{i=1}^{i=j} p_i > \frac{\sum p_i}{m}$, we split

X_j into two nodes X_j' and X_j'' . The coverage probability for X_k' is $p_k' = \frac{\sum p_i}{m} - \sum_{i=1}^{i=j-1} p_i$

while the coverage probability for X_k'' is $p_k'' = \sum_{i=1}^{i=j} p_i - \frac{\sum p_i}{m}$. Repeat this procedure for the

rest nodes of X_j', \dots, X_n until we have m groups of nodes. We use G_i to represent the i -th group we get, $i = 1, 2, \dots, m$.

For convenience, for the purposes of developing our proof, we arbitrarily assign this parameter β to be equal to 4. This is without loss of generality. Results presented from now on make this assumption.

Note that the sum of coverage probability over all the nodes within the group is $\frac{\sum p_i}{m}$.

If $\sum_i p_i \geq 4$, we have $4 \leq \frac{\sum p_i}{m} \leq 8$ or if $\sum_i p_i \leq 4$ then $\frac{\sum p_i}{m} = \sum_i p_i$.

Using the procedure outlined above, the demand at a node may be split between two clusters. For a split node X_j we do the following. Create two new nodes X_j' and X_j'' . The coverage probabilities for these nodes are p_j' and p_j'' . However, only one of these nodes can actually require a visit. We call the new tour with split nodes $\tilde{\tau}$.

The expected length of the optimal a priori tour τ^* is equal to the expected length of tour $\tilde{\tau}^*$. We make the following observation:

Observation 2.1. $E[L_{\tilde{\tau}^*}] = E[L_{\tau^*}]$.

2.3.2.2 The Radii of the Groups

For tour $\tilde{\tau}^*$, we continue to use G_i to represent the groups (clusters) ($i = 1, 2, \dots, m$). Let $E[L_{\tilde{\tau}^*}(i)]$ represent the expected length of the a priori tour among members in the group G_i .

From now on, we try to build the connection between the expected length of the optimal the a priori tour τ^* and the sum of the length of the radius of the group G_i , $i = 1, 2, \dots, m$.

For the radius of the group G_i , we will give it later.

We build the connection by analyzing the $E[L_{\tilde{\tau}^*}(i)]$, the radius and the expected length of the optimal the a priori tour τ^* . We express this idea in three steps, where the first two steps are used to give the definition of the radius of the group and step three to build the connection.

Step 1. Let A_i define the event in which $\{Z_i \geq 2\}$, where Z_i is the number of nodes from group G_i that have requests in a specific problem instance.

Step 2. We define $\bar{r}_i = \min_{\{X\}} \left\{ \overline{d(X, G_i)} \right\}$ as the radius of group G_i , where X is a point in the Euclidean plane and $\overline{d(X, G_i)} = \sum_{q \in G(i)_p} \frac{P_q}{\sum_{\alpha \in G(i)_p} P_\alpha} d(X_q, X)$ which represents the average distance between a point X in the Euclidean plane and the nodes in the group G_i . In other words, \bar{r}_i is the minimum average distance from any point in the Euclidean space to the nodes in the group G_i .

Step 3. If more than two nodes in group G_i require a visit, we select two nodes from group G_i at random where the likelihood of selection is proportional to their coverage probability. This is equivalent to randomly selecting two nodes from group G_i according to their coverage probability if there are at least two nodes having requirements of visits.

With this in mind, we obtain the following:

$$E[L_r] \geq \sum_{i=1}^{i=m} \left(E[L_r(i)] \right) \geq \sum_{i=1}^{i=m} P\{A_i\} E[L_r(i) | A_i] \geq \sum_{i=1}^{i=m} \left(P\{A_i\} \bar{r}_i \right). \quad (2.1)$$

In lemma 2.1, we provide a lower bound for $P(A_i)$ when $\sum_i p_i \geq 4$. Remember, as stated earlier that we could provide such a bound for any $\beta \geq 1$ where β a parameter. We use the value 4 for convenience of developing our proof.

Combining lemma 2.1 and (2.1), leads us to observation 2.2 which is the connection between the sum of the length of the radius of all the groups and the expected length of the optimal a priori tour τ^* .

Lemma 2.1 If $\sum_i p_i \geq 4$, $P(A_i) \geq \alpha$, where $\alpha = \frac{3}{5}(1 - e^{-4})$.

Observation 2.2 $\sum_i \bar{r}_i \leq \frac{E[L_r]}{\alpha}$, where $\alpha = \frac{3}{5}(1 - e^{-4})$.

2.3.3 A New Algorithm for the PTSP

First we introduce a sweep algorithm then define a k -median problem that is sub problem in our algorithm. Next we introduce our algorithm and its properties. In fact, the heuristic algorithm is a cluster-first and route-second heuristic. These are very common for solving vehicle routing problems. We use a heuristic algorithm for the k -(capacitated) median problem to generate the clusters and then use a sweep algorithm to provide the routes.

2.3.3.1 Sweep Algorithms

Assume that customers are located at points in the plane and that C_{ij} is the Euclidean distance between points i and j . A customer is chosen at random in a polar coordinate system with the origin at the center of the region and the ray from a customer to the center is "swept" either clockwise or counter clockwise (see for example Fisher,1995).

2.3.3.2 The k – Median Problem Φ

Let $m = \max \left\{ \left\lceil \frac{\sum_i p_i}{4} \right\rceil, 1 \right\}$. Assume that there are n customer locations (nodes) which are

represented by $\{X_1, X_2, \dots, X_n\}$. For node X_i , the size of its demand is equal to the node coverage probability p_i . The distance between any two points is the Euclidean distance.

The capacity for any median is $\frac{\sum_i p_i}{m}$ for the capacitated version of the problem.

2.3.3.3 A New Heuristic Algorithm H

Step 0. Pre-Processing

(I) If $\frac{9}{10} \leq \sum_i p_i \leq \frac{11}{10}$ and $\max_i (p_i) \geq \frac{2}{3} \left(\sum_i p_i \right)$, we select the node with highest

coverage probability as the center and connect all the other nodes directly to the center.

The expected length of the a priori tour generated in this way can be bounded by 2 times the expected length of optimal priori tour, please see the proof of lemma 2.3, case 3 for the detailed explanation.

(II) If neither $\frac{9}{10} \leq \sum_i p_i \leq \frac{11}{10}$ nor $\max_i (p_i) \geq \frac{2}{3} \sum_i p_i$ holds, we use a (capacitated)

k – median heuristic to solve problem Φ and let Y_1, Y_2, \dots, Y_m be the m medians selected by the heuristic algorithm.

Step 1. Clustering

Cluster the nodes according to the median from which it is serviced. Referring to the i -th median as Y_i , we call the group served by Y_i group $G(i)_H$.

Step 2. Routing the Groups

We select the median locations Y_1, Y_2, \dots, Y_m as the representatives of the groups $\{G(i)_H, i = 1, 2, \dots, m\}$ and then we use the well-known Christofides heuristic to construct a tour over all the representatives Y_1, Y_2, \dots, Y_m . The Christofides heuristic relies on the development of a minimum spanning tree and then solves a matching problem. Its performance is known to be guaranteed to be within $3/2$ of the optimal TSP solution. We can also use other good heuristic algorithms for the TSP. One good example is provided by Arora (1997).

Step 3. Routing the Nodes within the Groups

Connect each node to its representative and form a loop or use a sweep algorithm to construct an a priori tour within each group.

Note that in step 2, we get a tour through the representatives of the groups and in step three, we get the tour through the nodes within the group. Thus if we first start from a representative and visit all the nodes within group and follow the tours of the representatives to visit the next representative and the nodes within the same group and so on, we will get an a priori tour through all the nodes. We can further use the local improvement technique to the obtained an improved a priori tour.

2.3.4 Properties of Algorithm H

Now we will show that if we have a c_1 – approximation algorithm for the k – capacitated median problem, then the above algorithm is a c_2 – approximation algorithm for the PTSP for some constant c_2 . In practice, we can use any good heuristic algorithms for the k – median (capacitated or un-capacitated) problem to solve Φ .

We use $E[L_H]$ to represent the expected length for the tour developed using the specified steps.

Let $\hat{p}_i = P\{G(i)_H \text{ requires a visit}\}$ where $G(i)_H$ requires a visit means there are at least one node from group $G(i)_H$ requiring a visit.

We call \hat{p}_i the coverage probability for group $G(i)_H$. When the coverage probability is very high, it is very easy to find a heuristic algorithm with a constant guarantee. When the coverage probability is low, we transform the original problem into a new problem such that \hat{p}_i is very high by increasing the number of nodes in the groups.

We define $E[L_H(R)]$ is the expected length of the tour generated by the heuristic algorithm over the representatives with the covering probability \hat{p}_i and $E[L_H(\text{internal})]$ is the expected length of the tour within each group, $\{G(i)_H, i = 1, 2, \dots, m\}$. We have the following fact,

$$E[L_H] \leq E[L_H(\text{int})] + E[L_H(R)].$$

When the total sum of the coverage probabilities over all the nodes within each group is at least four and if we have a c -approximation algorithm for the capacitated version of problem Φ , we prove that both the $E[L_H(\text{internal})]$ and $E[L_H(R)]$ can be bounded by a constant times $E[L_{PTSP}]$. This is the lemma 2.2.

When the sum of the probabilities over the all the nodes are less than four, we analyze the special step for this case and show directly that $E[L_H]$ can be bounded by $E[L_{PTSP}]$. This leads us to lemma 2.3. Combining the lemma 2.2 and 2.3, we get theorem 2.1.

Please note that the number four here is selected for ease of presentation of our proof. Any number greater than one would be fine.

Lemma 2.2. When $\sum_i p_i \geq 4$, if we have c_1 -approximation algorithm for Φ , then

$$E[L_H] \leq \left[12 + \frac{c_1}{\alpha} \left(24 + \frac{2 \sum_i p_i}{m} \right) \right] E[L_{PTSP}].$$

Lemma 2.3. When $\sum_i p_i \leq 4$, $E[L_H] \leq 134E[L_{PTSP}]$.

Theorem 2.1. If we have c_1 – approximation algorithm for the k – capacitated median problem, then for our heuristic algorithm H for the PTSP which is based on the solution to the k – capacitated median problem Φ , $E[L_H] \leq c_2 E[L_{PTSP}]$ where

$$c_2 = \max \left\{ 134, 12 + \frac{c_1}{\alpha} \left(24 + \frac{2 \sum_i p_i}{m} \right) \right\}.$$

Note that Arora (1998) provides us with a quasi-polynomial algorithm for the k – capacitated median problem which is bounded by $(1 + \varepsilon)$ times the optimal solution.

Therefore we obtain corollary 2.1.

Corollary 2.1. We have a quasi-polynomial time c – approximation algorithm for PTSP,

$$\text{where } c = \max \left\{ 134, 12 + \frac{1 + \varepsilon}{\alpha} \left(24 + \frac{2 \sum_i p_i}{m} \right) \right\}.$$

From step 3, we can see that our algorithm will generate a solution which is like a star in each group. Interestingly, Bertsimas et al. (1993) obtained near optimal solution for several hundred nodes with equal coverage probability and found that the near optimal solution is star like.

2.3.5 Proof of the Lemmas

Proof of lemma 2.1

We will consider only one particular group. Let k be the number of nodes in this group

and $y = \sum_{i=1}^{i=k} p_i$. Remember that y is at least four ($y \geq 4$).

First we show that $P\{X = 0\} \leq e^{-y}$.

Let $\omega = \max\{\prod_i (1 - p_i)\}$, s.t. $\sum_i p_i = y$ ($y \geq 4$).

We know that $P\{X = 0\} \leq \omega$.

After applying the standard optimization technique, we know that when all the p_i 's are equal, the maximum is obtained.

Therefore $\omega = \left(1 - \frac{y}{k}\right)^k$.

It can be shown that $f(k) = \left(1 - \frac{y}{k}\right)^k$, ($k > y$) increases over k and $f(k) \rightarrow e^{-y}$ as

$$k \rightarrow \infty, \text{ therefore } \omega = f(k) \leq e^{-y}. \quad (2.2)$$

From (2.2), we know that $\omega \leq e^{-y}$. Finally, we have an upper bound for $P\{X = 0\}$,

$$P\{X = 0\} \leq \omega \leq e^{-y}. \quad (2.3)$$

Now we compare $P\{X = 2\}$ and $P\{X = 1\}$.

$$\begin{aligned}
P(X = 2) &= \frac{1}{2} \sum_{i \neq j} \left(\frac{p_i}{1-p_i} \frac{p_j}{1-p_j} \Pi_k (1-p_k) \right) \\
&= \frac{1}{2} [\Pi_k (1-p_k)] \sum_i \left(\frac{p_i}{1-p_i} \sum_{j \neq i} \left(\frac{p_j}{1-p_j} \right) \right) \\
&\geq \frac{1}{2} [\Pi_k (1-p_k)] \sum_i \left(\frac{p_i}{1-p_i} \sum_{j \neq i} p_j \right) \\
&\geq \frac{3}{2} [\Pi_k (1-p_k)] \sum_i \left(\frac{p_i}{1-p_i} \right) = \frac{3}{2} P(X = 1). \tag{2.4}
\end{aligned}$$

Combining (2.3) (2.4), we know that

$$\begin{aligned}
P(X = 1) + \frac{3}{2} P(X = 1) &\leq P(X = 1) + P(X = 2) \\
&\leq 1 - P(X = 0) \\
\Rightarrow P(X = 1) &\leq \frac{2}{5} (1 - P(X = 0)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(A_i) = P(X \geq 2) &\geq 1 - P(X = 0) - \frac{2}{5} (1 - P(X = 0)) \\
&\geq \frac{3}{5} (1 - \omega) \geq \frac{3}{5} (1 - e^{-\gamma}) \geq \frac{3}{5} (1 - e^{-4}) = \alpha.
\end{aligned}$$

Proof of Lemma 2.2

After using a c_1 -approximation algorithm for the capacitated k -median problem Φ , we obtain the median locations such that the cost can be bounded by the cost of the optimal median locations times c_1 .

For the PTSP solution, we obtain k groups, each with the same average demand. This is one possible solution for the capacitated k -median problem Φ and its total cost is no less than the optimal solution.

First, we will prove two claims separately and then we use the two claims to prove lemma 2.2.

Claim 2.1. $E[L_H(\text{internal})] \leq 2c_1 \left(\frac{\sum_i p_i}{m} \right) \frac{E[L_i]}{\alpha}$, where $E[L_H(\text{internal})]$ is the

expected length of the total distance traveled within each group under the tour provided by the heuristic algorithm given by steps 0 to 3.

Claim 2.2. $E[L_H(R)] \leq 12E[\Sigma(R)]$, where $E[L_H(R)]$ is the expected length of the a priori tour over the representatives and $E[\Sigma(R)]$ is the expected length of the tour generated by the re-optimization over the representatives with coverage probability \hat{p}_i .

We prove these two claims and then use them to prove lemma 2.2. From now on, we focus the proof of claim 2.1.

Let $Y_1^*, Y_2^*, \dots, Y_m^*$ be the optimal solution for problem Φ and $G(i)_{H^*}$ be the corresponding groups obtained from $Y_1^*, Y_2^*, \dots, Y_m^*$. Let Y_1, Y_2, \dots, Y_m be the solution obtained from the heuristic algorithm.

Let $\bar{r}_i = \sum_{X_j \in G(i)_H} \frac{p_j d(X_j, Y_i)}{\sum_{X_k \in G(i)_H} p_k}$ and $\bar{\bar{r}}_i = \sum_{X_j \in G(i)_{H^*}} \frac{p_j d(X_j, Y_i^*)}{\sum_{X_k \in G(i)_{H^*}} p_k}$. Note we use \bar{r}_i to represent

the radius of the group we get from the optimal a priori tour τ^* .

The cost associated with selecting Y_1, Y_2, \dots, Y_m as the medians is

$$\sum_{i=1}^{i=m} \sum_{X_j \in G(i)_H} p_j d(Y_i, X_j) = \left(\frac{\sum_i p_i}{m} \right) \sum_{i=1}^{i=m} \bar{r}_i \leq c_1 \left(\frac{\sum_i p_i}{m} \right) \left(\sum_{i=1}^{i=m} \bar{\bar{r}}_i \right).$$

From observation 2.2, we know $\sum_{i=1}^{i=m} \bar{\bar{r}}_i \leq \sum_{i=1}^{i=m} \bar{r}_i \leq \frac{E[L_r]}{\alpha}$.

For our heuristic algorithm, the expected length within the groups can be bounded by

$$2 \left(\frac{\sum_i p_i}{m} \right) \sum_{i=1}^{i=m} \bar{r}_i. \text{ This leads us to claim 2.1.}$$

Now we provide the proof of claim 2.2. There are two main steps as follows.

Step 1. First, we can see for any given problem instance, after we obtain the TSP tour over the nodes that require a visit, we select one node at random, proportional to its coverage probability from the node(s) within the same group that require a visit. Then we begin a round-trip from the selected node to its representative. This lead to a tour through all the present representatives. From this, we can see that,

$$E[\Sigma] + 2 \sum_i \bar{r}_i \geq E[\Sigma(R)] \quad (2.5)$$

Step 2. We compare the length of the minimum spanning tree over all the representatives to $E[\Sigma(R)]$, where $E[\Sigma(R)]$ is the expected length of the re-optimization over the representatives with coverage probability of \hat{p}_i . For the definition of \hat{p}_i , please refer to the section 2.3.4.

Let Ψ be the set of representatives, $\{Y_1, Y_2, \dots, Y_m\}$, where m is the number of the total representatives. If there are at least one nodes from group $G(i)_H$ that requires a visit, the representative will appear. Thus by this way, we can see that the coverage probability for the representatives are $\hat{p}_i, i = 1, 2, \dots, m$.

Selecting $\left(\left\lceil\frac{m}{2}\right\rceil+1\right)$ representatives randomly (uniformly) from Ψ . We call the set of selected nodes Ψ_1 . We focus the set of Ψ_2 by selecting one node randomly from Ψ_1 adding it to Ψ/Ψ_1 . Thus Ψ_1 is the set of random selected $\left(\left\lceil\frac{m}{2}\right\rceil+1\right)$ nodes from Ψ and Ψ_2 is the set of random selected $\left(m-\left\lceil\frac{m}{2}\right\rceil\right)$ nodes from the set Ψ . Note the set of Ψ_1 and Ψ_2 don't dependent on the group coverage probability

Let $L_{TSP}(\Psi)$, $L_{TSP}(\Psi_1)$, $L_{TSP}(\Psi_2)$ be the length of TSP tours over Ψ, Ψ_1, Ψ_2 respectively.

Letting $L_{MST}(A)$ be the length of minimum spanning tree over the nodes in A , we use $E[L_{MST}(\Psi_1)]$ and $E[L_{MST}(\Psi_2)]$ to represent the expected length of the minimum spanning tree over randomly selected Ψ_1, Ψ_2 (this doesn't related to the coverage probability).

We know the following:

Observation 2.4. $L_{TSP}(\Psi) \geq \max\{L_{TSP}(\Psi_1), L_{TSP}(\Psi_2)\}$ and $L_{MST}(\Psi_1) \leq L_{TSP}(\Psi_1)$. (2.6)

The fact that the minimum spanning tree over Ψ_1 and Ψ_2 will form a connected graph over Ψ leads us to observation 2.5.

Observation 2.5. $L_{MST}(\Psi) \leq L_{MST}(\Psi_1) + L_{MST}(\Psi_2)$. (2.7)

From (2.7), we know that $\max\{E[L_{MST}(\Psi_1)], E[L_{MST}(\Psi_2)]\} \geq \frac{1}{2}L_{MST}(\Psi)$.

Assume that $E[L_{MST}(\Psi_1)] \geq E[L_{MST}(\Psi_2)] \geq \frac{1}{2}L_{MST}(\Psi)$ without losing generality.

Let Λ be the event in which at least $\left(\left\lceil \frac{m}{2} \right\rceil + 1\right)$ representatives are present. Let

$E[\Sigma(R) | \Lambda]$ represent the optimal TSP tour over the nodes appearing when at least

$m_1 = \left(\left\lceil \frac{m}{2} \right\rceil + 1\right)$ nodes require visits.

Let $L_{TSP}(A)$ be the length of the optimal TSP tour over node set A . If $B \subseteq A$, $L_{TSP}(A) \geq L_{TSP}(B)$. We call the property of the optimal TSP tour as the monotone property of the optimal TSP tour.

Now assume Λ happens, if more than m_1 nodes appear, we randomly select m_1 nodes and drop the rest nodes. By this way, when Λ happens, we have two facts.

Fact 1. When Λ happens, we always can obtain the node set Ψ_1 .

Fact 2. The length of the optimal TSP tour over the selected nodes is less or equal to the length of the original optimal TSP tour by the monotone property of the optimal TSP tour.

Combing these two facts, we have

$$E[\Sigma(R) | \Lambda] \geq E[L_{\text{TSP}}(\Psi_1)] \geq E[L_{\text{MST}}(\Psi_1)] \geq \frac{1}{2} L_{\text{MST}}(\Psi).$$

We will show later that $P(\Lambda) \geq \frac{1}{2}$. Therefore:

$$E[\Sigma(R)] \geq E[\Sigma(R) | \Lambda] P(\Lambda) \geq \frac{1}{4} L_{\text{MST}}(\Psi). \quad (2.8)$$

This leads to observation 2.6.

Observation 2.6. For any set Y_1, Y_2, \dots, Y_m , if $m \geq 2$, let $E[\Sigma]$ is the expected length of the re-optimization over Y_1, Y_2, \dots, Y_m with the coverage probability \hat{p}_i , $i = 1, 2, \dots, m$, we have

$$E[\Sigma] \geq \frac{1}{4} L_{\text{MST}}(Y_1, Y_2, \dots, Y_m).$$

Now, we finish the proof of claim 2.2.

Two cases are possible:

Case 1. $m = 1$. We know that $E[L_H(R)] = E[\Sigma(R)] = 0$.

Case 2. $m \geq 2$. Using Christofides' heuristic algorithm in step 3 of our algorithm, we know the following:

$$\mathbb{E}[L_H(\mathbf{R})] \leq \frac{3}{2} L_{TSP}(Y_1, Y_2, \dots, Y_m) \leq 3L_{MST}(Y_1, Y_2, \dots, Y_m).$$

From observation 2.6, we know $\mathbb{E}[L_H(\mathbf{R})] \leq 12\mathbb{E}[\Sigma(\mathbf{R})]$.

Combining case 1 and case 2 together, we have $\mathbb{E}[L_H(\mathbf{R})] \leq 12\mathbb{E}[\Sigma(\mathbf{R})]$. This completes the proof of claim 2.

Now we use the two claims to prove the lemma 2.2.

Using (2.5), Claim 2.1 and Claim 2.2 and Observation 2.2, we have the following,

$$\begin{aligned} \mathbb{E}[L_H] &\leq \mathbb{E}[L_H(\text{internal})] + \mathbb{E}[L_H(\mathbf{R})] \\ &\leq \mathbb{E}[L_H(\text{internal})] + 12\mathbb{E}[\Sigma(\mathbf{R})] \\ &\leq \mathbb{E}[L_H(\text{internal})] + 12\left(\mathbb{E}[\Sigma] + 2\sum_i \bar{r}_i\right) \\ &\leq 12\mathbb{E}[\Sigma] + c_1 \left(24 + \frac{2\sum_i p_i}{m}\right) \frac{\mathbb{E}[L_{r_i}]}{\alpha} \\ &\leq \left[12 + \frac{c_1}{\alpha} \left(24 + \frac{2\sum_i p_i}{m}\right)\right] \mathbb{E}[L_{PTSP}]. \end{aligned}$$

Now we show that $P(\Lambda) \geq \frac{1}{2}$.

From (2.3), we know $\hat{p}_i \geq 1 - e^{-1} = \frac{15}{16}$, $i = 1, 2, \dots, m$.

We know that X is the number of nodes having requirement of m nodes with coverage probability of $\widehat{p}_i \geq \frac{15}{16}$ ($i=1,2,\dots,m$).

Let Y be the number of the number of nodes having requirement of m nodes with coverage probability of $\frac{15}{16}$ ($i=1,2,\dots,m$) and Y be independent with X .

It is easy to see that, $P(X \geq j) \geq P(Y \geq j)$.

The above relationship between X and Y is known as stochastic bigger, please refer to Ross (1983) for detail.

Because $P(\Delta) = P(X \geq m_1) \geq P(Y \geq m_1)$, to prove $P(\Delta) \geq \frac{1}{2}$, we only need to show

$$P(Y \geq m_1) \geq \frac{1}{2}.$$

From now on, we show $P(Y \geq m_1) \geq \frac{1}{2}$.

$$\text{We know that } P(Y=i) = \binom{m}{i} \left(\frac{15}{16}\right)^i \left(\frac{1}{16}\right)^{m-i} = \binom{m}{i} 15^i \left(\frac{1}{16}\right)^m.$$

$$P(Y=m-i) = \binom{m}{i} \left(\frac{15}{16}\right)^{m-i} \left(\frac{1}{16}\right)^i = \binom{m}{i} 15^{m-i} \left(\frac{1}{16}\right)^m.$$

So we have $P(Y=0) \leq P(Y=m)$, $P(Y=1) \leq P(Y=m-1)$ and

$$P(Y=i) \leq P(Y=m-i), \quad i=1,2,\dots,m.$$

When m is odd, assume $m=2q+1$, then $m_1 = \left\lfloor \frac{m}{2} \right\rfloor + 1 = q+1$.

It is easy to see $P(Y \geq m_1) = P(Y \geq q+1) \geq P(Y \leq q) \Rightarrow P(Y \geq m_1) \geq \frac{1}{2}$.

If m is even and assume $m=2q$, then $m_1 = \left\lfloor \frac{m}{2} \right\rfloor + 1 = q+1$.

We compare the difference between $P\left(Y = \frac{m}{2} - 1\right)$ and $P\left(Y = \frac{m}{2} + 1\right)$ with $P\left(Y = \frac{m}{2}\right)$ as

follows.

$$\begin{aligned} & P\left(Y = \frac{m}{2} + 1\right) - P\left(Y = \frac{m}{2} - 1\right) \\ & \geq \frac{(2q)!}{(q-1)!(q+1)!} \left(\frac{1}{16}\right)^{2q} (15^{q+1} - 15^{q-1}) \\ & \geq \frac{(2q)!}{(q-1)!(q+1)!} \left(\frac{1}{16}\right)^{2q} 14(15^q) \\ & \geq \frac{(2q)!}{(q)!(q)!} \left(\frac{1}{16}\right)^{2q} (15^q) = P(Y=q). \end{aligned}$$

Now we have $P(Y \geq m_1) = P(Y \geq q+1) \geq P(Y \leq q) \Rightarrow P(Y \geq m_1) \geq \frac{1}{2}$.

Thus $P(\Lambda) = P(X \geq m_1) \geq P(Y \geq m_1) \geq \frac{1}{2}$.

This concludes the proof of lemma 2.2.

Proof of Lemma 2.3

Let X be the number of nodes present and let $y = \sum_i p_i$. When there are 1, 2 or 3 nodes

present, we can see that $\frac{E[L_H | X = i]}{E[L_{PTSP} | X = i]} = 1$. We consider four cases.

Case 1. $\frac{11}{10} \leq \sum_i p_i = y \leq 4$.

First, we show that $p(X \geq 2) \geq \frac{1}{21} \left(1 - e^{-\frac{10}{11}}\right)$.

$$\begin{aligned} P(X = 2) &= \frac{1}{2} \sum_{i \neq j} (p_i p_j \prod_{k \neq i, k \neq j} (1 - p_k)) \\ &\geq \frac{1}{2} \left[\prod_i (1 - p_i) \right] \sum_i \left(\frac{p_i}{1 - p_i} \left(\sum_{j \neq i} \frac{p_j}{1 - p_j} \right) \right) \\ &\geq \frac{1}{2} \sum_i \left(\frac{p_i (y - p_i)}{1 - p_i} \left(\prod_j (1 - p_j) \right) \right). \end{aligned}$$

Because $\frac{11}{10} \leq \sum_i p_i = y$, we have $y - p_i \geq \frac{1}{10}$. So we have

$$P(X = 2) \geq \frac{1}{20} \sum_i \left(\frac{p_i}{1 - p_i} \left(\prod_j (1 - p_j) \right) \right) = \frac{1}{20} P(X = 1).$$

From (2.3), we know that $P(X=0) \leq e^{-y}$ and $\frac{11}{10} \leq \sum_i p_i = y \leq 4$, we have,

$$P(X=0) \leq e^{-\frac{10}{11}}.$$

$$P(X \geq 2) = 1 - P(X=0) - P(X=1) \geq 1 - P(X=0) - 20P(X=2)$$

$$\Rightarrow 21P(X \geq 2) \geq 1 - P(X=0) \geq 1 - e^{-\frac{10}{11}}.$$

Finally we have the following lower bound for the $P(X \geq 2)$,

$$P(X \geq 2) \geq \frac{1}{21} \left(1 - e^{-\frac{10}{11}} \right).$$

$EL_{PTSP} \geq P\{X \geq 2\} E[L_{PTSP} | X \geq 2] \geq 2P(X \geq 2)\bar{r}$ and $EL_H \leq 2 \left(\sum_i p_i \right) \bar{r}$, so we have

$$\frac{EL_H}{EL_{PTSP}} \leq \frac{2 \left(\sum_i p_i \right)}{2P(A)} \leq \frac{y}{P(X \geq 2)} \leq \frac{y}{P(X \geq 2)} \leq \frac{84}{1 - e^{-\frac{10}{11}}} \leq 134.$$

Case 2. $\frac{9}{10} \leq \sum_i p_i \leq \frac{11}{10}$ and $\max_i (p_i) \leq \frac{2}{3}y$.

$$P(X=2) = \frac{1}{2} \sum_{i \neq j} (p_i p_j \prod_{k \neq i, k \neq j} (1 - p_k))$$

$$\geq \frac{y}{6} \sum_i \left(\frac{p_i}{1 - p_i} \left(\prod_{j \neq i} (1 - p_j) \right) \right)$$

$$= \frac{y}{6} P(X=1).$$

From (2.3), we have $P(X=0) \leq e^{-y} \leq e^{-\frac{9}{10}}$.

In a similar way as we have done in case 1, we can show that

$$P(X \geq 2) \geq \frac{1}{7} \left(1 - e^{-\frac{9}{10}} \right).$$

For the same reason as in case 1, $\frac{EL_H}{EL_{PTSP}} \leq \frac{2 \left(\sum_i p_i \right)}{2P(A)} \leq \frac{y}{P(X \geq 2)} \leq \frac{77}{10 \left(1 - e^{-\frac{9}{10}} \right)} \leq 134$.

Case 3. $\frac{9}{10} \leq \sum_i p_i \leq \frac{11}{10}$ and $\max_i (p_i) \geq \frac{2}{3} y$.

By definition of the algorithm for this specific case (step 0), we know the following:

$$\begin{aligned} EL_H &\leq 2 \sum_i p_i d(X_i, X_1) \text{ and } EL_{PTSP} \geq \frac{4y}{3} \sum_i p_i d(X_i, X_1) \\ \Rightarrow \frac{EL_H}{EL_{PTSP}} &\leq \frac{3}{2y} \leq \frac{5}{3}. \end{aligned}$$

Case 4. $\sum_i p_i \leq \frac{9}{10}$.

First, we know $EL_H \leq 2y\bar{r} - 2P(X=1)\bar{r}$ and $EL_{PTSP} \geq 2P(X \geq 2)\bar{r} \Rightarrow$

$$\frac{EL_H}{EL_{PTSP}} \leq \frac{y - P(X=1)}{P(X \geq 2)} \leq \frac{2 \sum_i \left(p_i \left(1 - \sum_{j \neq i} p_j \right) \right)}{\sum_{i \neq j} \left(p_i p_j \left(\prod_{k \neq i, k \neq j} (1 - p_k) \right) \right)}$$

$$\leq \frac{2 \sum_i (p_i (y - p_i))}{\sum_{i \neq j} (p_i p_j (1 - y))} = \frac{20 \sum_i (p_i (y - p_i))}{\sum_i (p_i (y - p_i))} = 20.$$

Combining these cases we demonstrate that lemma 3 holds.

Proof of Theorem 2.1

$EL_H \leq c_2 EL_{PTSP}$ follows from lemma 2.2 and lemma 2.3.

2.4 Conclusion

Stochastic network optimization plays an important role in the real problems and in theoretical research. The probabilistic traveling salesman problem is an important problem in this class. In this chapter we have shown that, for some constant c_1 , if there exists a c_1 -approximation algorithms for k -capacitated median problem, for some constant c , we can find a c -approximation algorithm for PTSP. As a corollary, we developed a quasi-polynomial algorithm which is a c -approximation algorithm for PTSP. We can also use the standard route improvement scheme to improve the tour.

CHAPTER 3 THE M/G/1 QUEUE WITH SWITCHOVER COSTS:AN EXAMINATION OF ALTERNATIVE HEURISTICS

3.1 Introduction

Many systems can be modeled as M/G/1 queues with switchover costs. For example, in many production systems, it is common that a facility is responsible for replenishing inventories of several items and that a switchover time is incurred whenever a switch is made. If only one server is responsible for providing service to customers at different locations, a travel cost will be incurred when the server switches from one location to another. We need an explicit way to consider the influence of switchover costs on the performance of these systems.

In this chapter, we consider the M/G/1 queue with switchover costs. This model involves n identical independent queues, each fed by Poisson arrivals with identical parameter λ . A single server provides service to all customers. The server can provide service to only one queue at a time. A constant switchover cost is incurred each time the server switches nodes. A decision strategy or policy for this problem specifies the action to be taken at each decision instance. The server may remain working at the present queue, remain idle at the present queue or switch to another queue. Decision instances include the arrival of a customer, the completion of service for a customer, and may also occur any time the server is idle. The objective is to minimize the overall average waiting time for each customer.

Previous research in this area has had as its focus the characterization of the best service algorithms or examination of such systems under various kinds of cyclic polling algorithms. Much of the research on cyclic polling systems has focus on the analysis of the waiting time or queue length under various service policies. For example, Cooper, Niu and Srinivasan (1996) developed an explicit expression for the average waiting time under gated or exhaustive policies. Srinivasan, Niu and Cooper (1995) extended those results to describe the relationship between the waiting time distributions in systems with zero and non-zero switchover costs when a gated or exhaustive service discipline is enforced. Recently, Borst and Boxma (1997) extend that research to the general case in which algorithms satisfy a more general discipline referred to as a branching property. Eisenberg (1994) analyzed the polling system in which the server comes to a stop when the system is empty rather than continuing to cycle. That work examines a variety of stopping and starting rules. Later, Srinivasan and Gupta (1996) consider the circumstances under which the server should be patient (which means to stop when the system is empty). They show that while the patient server mechanism is generally better than the roving server mechanism, there do exist cases where roving is better. In all of these studies, the assumption is that if customers exist in the system that the server will not be idle.

To date, neither the optimal algorithm nor the explicit expression of average waiting time for the optimal algorithm is known for the M/G/1 queue with switchover cost. Hofri and Ross (1987) studied the case of two queues and conjectured that the optimal algorithm will be of a threshold type. Liu, Nain and Towsley (1992) and Duenyas, and van Oyen

(1996) have partially characterized the optimal algorithm. We discuss this characterization in detail in the first section of this chapter.

Our research provides a lower bound for the waiting time in an M/G/1 queue with switchover cost. From Cooper, Niu and Srinivasan (1996), we know the average waiting time for the M/G/1 queue under cyclic polling. This average waiting time provides an upper bound for the optimal algorithm. Comparing our lower bound to this upper bound, we can see that in general it is not very tight. However, under certain conditions, worst case analysis of the cyclic polling algorithm shows that its competitive ratio (the ratio of the average waiting time under cyclic polling to the optimal) is close to 2. We also identify circumstances under which our lower bound is very close to the cyclic polling upper bound which implies that under certain conditions the cyclic polling algorithm is close to optimal. In this way, we partially identify the optimal algorithm and provide additional justification for the popular cyclic polling algorithm. We also compare the average waiting time for the cyclic polling algorithm and the longest queue first algorithm (here the longest queue first means when the server switches to the other queue, it selects the queue with longest queue).

This chapter is organized as follows. In section 3.2, we discuss the definition of the problem and the properties of the optimal algorithm, we also provide the notation needed later. In section 3.3 we provide a lower bound for the optimal algorithm. In section 3.4 we provide a new lower bound for the light traffic intensity case. In section 3.5, we provide some simulation analysis. Finally we end with some conclusions.

3.2 Problem and Notation

The M/G/1 queue with switchover costs involves n identical independent queues, each fed by Poisson arrivals with identical parameter λ . A single server provides service to all customers. The service time is a random variable with first and second moments \bar{s} and \bar{s}^2 , respectively. The server can provide service to only one queue at a time. A constant switchover cost, $\frac{d_i}{v}$, is incurred each time the server switches to node i . Neither the service of a job nor the process of switching to a different queue can be interrupted prior to its completion.

A decision strategy or policy for this problem specifies the action to be taken at each decision instance. The server may remain working at the present queue, remain idle at the present queue or switch to another queue. Decision instances include the arrival of a customer, the completion of service, and may also occur any time the server is idle.

For the case of uniform switchover costs over all the nodes, from Liu, Nain and Towsley (1992) we know that the best algorithm will remain at its current location as long as there are unserved customers at the current location. An algorithm satisfying this property is said to be *exhaustive*. When it transfers to another queue, it will choose the queue with the largest number of waiting customers, this property is known as *longest queue criteria*. For the general case, from Duenyas, and van Oyen (1996), we know that the best algorithm will remain working instead of idling in the current location if jobs remain at the current location. Algorithms satisfying these conditions are differentiated only by the

switching rule employed. That is, the server must decide whether to wait at a node after it has completed all the jobs waiting at that node or whether to switch to another node and also when to switch to another node. We refer to the following condition as the *continuous* condition: if there are customers in the system when the server completes service at a single arrival process, the server will depart its current location for a location where there are unserved customers. Without providing proof, we conjecture that as ρ approaches 1.0, the optimal algorithm will have this property (In that case the probability that the system is idle will be very small). If this conjecture is true, the optimal algorithm must be *exhaustive*, *continuous* and must also satisfy the *longest* queue criterion. Taken together, these completely specify the optimal algorithm. In addition, we note that under the special case of zero switchover costs, our system reduces to an M/G/1 queueing model.

Notation

Let

n represent the number of sub-queues in the system,

\bar{s} and $\overline{s^2}$ represent the first and second moments of the on-site service time respectively,

λ the parameter for the Poisson process at each sub-queue,

$\rho = n\lambda\bar{s}$ and $\rho_1 = \lambda\bar{s}$ the fraction of time the server spends providing on on-site service to all sub-queues and the fraction of time spent in on-site service for a single sub-queue, respectively,

$\bar{W}_1 = \frac{n\lambda\bar{s}^2}{2(1-\rho)}$, the average waiting time in an M/G/1 queue with arrival rate $n\lambda$, service rate $\mu = \frac{1}{s}$ and no switchover costs. This is known as Pollaczek-Khinchin(P-K) formula.

Please see Bertsekas and Gallager (1992) for a discussion of the P-K formula.

3.3 Lower and Upper Bounds on the Average Waiting Time for the Optimal Algorithm

In this section we develop separate lower bounds for algorithms that obey and do not necessary obey the continuous condition. We combine these to obtain a lower bound for the optimal algorithm. There are two reasons that we address the two cases separately. The first is that the examination of the lower bound for algorithms obeying the continuous condition is of independent interest. The second is that proof of the lower bound for the first case leads to the proof of the lower bound for the second case.

3.3.1 A Lower Bound for Algorithms Obeying the Continuous Condition

To develop our lower bound we begin by selecting an arbitrary algorithm obeying the continuous condition. Further we require the algorithm to perform in such a way as to ensure that all customers eventually receive service. We use a simple, but important concept in the development of this lower bound. The total waiting time for all customers served during a single visit to a node must be greater than the total waiting time experienced by customers already at the location when the server arrives. In addition, the total waiting time experienced at the moment when the server arrives must be greater than the total waiting time experienced at the moment of the arrival of the last customer

that arrives prior to the arrival of the server. This last quantity can be estimated and is used to provide bounds for the other two.

Remember that \bar{W}_1 is the average waiting time for M/G/1 queue model and without

switchover costs,
$$\bar{W}_1 = \frac{n\lambda s^2}{2(1-\rho)}.$$

Let \bar{W}_c represent the average waiting time for an arbitrary algorithm satisfying the continuous condition.

Let
$$\omega_1 = \frac{(1-\rho_1)^2}{2nv(1-\rho)} \left(\sum_{i=1}^{in} \sqrt{d_i} \right)^2 - \frac{(1-\rho_1)}{2\lambda}.$$

Lemma 3.1: $\bar{W}_c \geq \max \{ \bar{W}_1, \omega_1 \}.$

Proof of lemma 3.1

Select an arbitrary algorithm satisfying the continuous condition. Under steady-state conditions, let Z_i be the number of customers served during one visit to node i and z_i be the number of customers waiting at the node i when the server arrives.

In an M/G/1 queue in which the arrival rate is λ and the service rate is $\frac{1}{s}$, the expected

length of a busy period is $\frac{\rho_1}{\lambda(1-\rho_1)}$. We define the customers that arrive during the

service time of one demand α to be the children of α . We define the grandchildren of α

to be those customers that arrive during service to α 's children. We refer to the children, grandchildren and later generations of α as the descendants of α .

Now we let the server serve the first customer in the queue and all of its descendants in the order of their arrival; the second customer in the queue and all of its descendants in the order of their arrival; the third customer in the queue and all of its descendants in the order of their arrival; and so on until the last customer in the queue and all its descendants have been served. We observe that the service time for a customer and all of its descendants is an i.i.d. random variable as is the number of descendants. Furthermore, the number of descendants of a single customer is distributed in the same way as the number of new customers served during a single busy period in a classical M/G/1 queue.

Observation 3.1: (A lower bound for $E[z_i]$)

The expected number of descendants of a single customer is less than or equal to

$$\lambda \frac{\rho_1}{\lambda(1-\rho_1)} = \frac{\rho_1}{1-\rho_1}.$$

If there are z_i customers present at the beginning of service, the expected number of

descendants for all z_i customers is equal to $\frac{z_i \rho_1}{1-\rho_1}$. This implies that the following,

$$E[Z_i | z_i] \leq z_i + \frac{z_i \rho_1}{1-\rho_1} = \frac{z_i}{1-\rho_1}.$$

Because $E[E[Z_i | z_i]] = E[Z_i]$, we know that,

$$\frac{E[z_i]}{1-\rho_1} \geq E[Z_i] \text{ i.e. } E[z_i] \geq (1-\rho_1)E[Z_i] \tag{3.1}$$

Observation 3.2: (A lower bound for the total waiting time)

The total waiting time for all customers served is greater than or equal to the total waiting time experienced by the customers that arrived prior to the arrival of the server. At stable state, Z_i is the number of customers served during one visit to node i and z_i is the number of customers waiting at node i when the server arrives. Let s_i be the set of customers waiting at node i when the server arrives. The number of elements in set s_i is z_i . Assume that the inter-arrival times between successive customers arriving at location i belonging to s_i are given by $Y_1^{z_i}, Y_2^{z_i}, \dots, Y_{z_i}^{z_i}$. A lower bound for the total waiting time experienced by customers in the set s_i , is given by $\sum_{j=1}^{z_i} (j-1)Y_j^{z_i}$.

Because our customers arrive according to a Poisson process, the expected length of the interarrival period, $E[Y_k^{z_i}] = \frac{1}{\lambda}$. Therefore $E\left[\sum_{j=1}^{z_i} (j-1)Y_j^{z_i}\right] = \frac{z_i(z_i-1)}{2\lambda}$.

Define W^i to be the waiting time of a random selected demand from the customers served at node i . Now, first we randomly select a served customer. We know that the customers are uniformly distributed over the n nodes, so the probability that the selected customer is at any specific node is $\frac{1}{n}$. Then, we select a customer randomly from the customers served at the selected node. Let W_r be the waiting time of such a selected customer. Therefore:

$$\begin{aligned} E[W_r] &= \frac{1}{n} \sum_{i=1}^{i=n} E[W^i] = \frac{1}{n} \sum_{i=1}^{i=n} E[E[W^i | z_i = k]] \\ &\geq \frac{1}{n} \sum_{i=1}^{i=n} \sum_{k=2}^{k=\infty} E\left[\frac{1}{2\lambda Z_i} \left(\sum_{j=1}^{j=z_i} (j-1) Y_j^i\right) \mid z_i = k\right] P\{z_i = k\} \end{aligned}$$

Using these observations:

i) From (3.1), we know that $\frac{1}{E[Z_i]} \geq \frac{1-\rho_1}{z_i}$.

ii) Let $Z_i = z_i + x_i$, where x_i is the number of descendants of the z_i customers in set s_i .

Note that $\{Y_1^i, Y_2^i, \dots, Y_{z_i}^i\}$ are dependent only on the arrival process before time t (the moment of the arrival of server to node i) and the x_i 's are dependent only on the arrival process after t and the process of serving after time t . This implies that the variables $\{Y_1^i, Y_2^i, \dots, Y_{z_i}^i\}$ are independent of x_i , given z_i . This gives us the following equation:

$$\begin{aligned} E\left[\frac{1}{2\lambda(z_i + x_i)} \left(\sum_{j=1}^{j=z_i} (j-1) Y_j^i\right) \mid z_i = k\right] &= E\left[\frac{1}{2\lambda(k + x_i)} \left(\sum_{j=1}^{j=k} (j-1) Y_j^i\right) \mid z_i = k\right] \\ &= \frac{1}{2\lambda E[k + x_i | z_i = k]} E\left[\sum_{j=1}^{j=k} (j-1) Y_j^i\right] = \frac{k(k-1)}{2\lambda E[Z_i | z_i = k]}. \end{aligned}$$

Therefore we obtain the following:

$$\begin{aligned} E[W_r] &\geq \frac{1}{n} \sum_{i=1}^{i=n} \sum_{k=2}^{k=\infty} \left[\frac{k(k-1)}{2\lambda E[Z_i | z_i = k]} P\{z_i = k\} \right] \geq \frac{1}{n} \sum_{i=1}^{i=n} \sum_{k=2}^{k=\infty} \frac{1-\rho_1}{k} \frac{k(k-1)}{2\lambda} P\{z_i = k\} \\ &= \frac{1}{n} \sum_{i=1}^{i=n} \frac{(1-\rho_1)E[z_i - 1]}{2\lambda} = \frac{1}{n} \sum_{i=1}^{i=n} \frac{(1-\rho_1)E[z_i]}{2\lambda} - \frac{1-\rho_1}{2\lambda}. \end{aligned}$$

Which leads to our lower bound for $E[W_r]$, which is,

$$E[W_r] \geq \frac{1}{n} \sum_{i=1}^{i=n} \frac{(1-\rho_1)E[z_i]}{2\lambda} - \frac{1-\rho_1}{2\lambda}. \quad (3.2)$$

Observation 3.3: (A lower bound for $E[z_i]$)

The average switchover cost per customer served is given by $\frac{1}{n} \sum_{i=1}^{i=n} E\left[\frac{d_i}{vZ_i}\right]$ (3.3)

For $X \geq 0$, from Schwarz's Inequality, we know that $E\left[\frac{1}{X}\right] \geq \frac{1}{E[X]}$.

From (3.1), we know that,

$$E\left[\frac{1}{Z_i}\right] \geq \frac{1}{E[Z_i]} \geq \frac{1-\rho_1}{E[z_i]} \quad (3.4)$$

From (3.3) and (3.4), we can see that the average switchover cost is bounded from below

$$\text{by } \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)d_i}{vnE[z_i]} \right).$$

Remember here that we assume that our system is stable. A necessary condition to reach

steady state is that $\hat{\rho} < 1$. In our system $\hat{\rho} = n\lambda \left(\bar{s} + \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)d_i}{vnE[z_i]} \right) \right)$.

Therefore $n\lambda \left(\bar{s} + \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)d_i}{vnE[z_i]} \right) \right)$ must be less than 1.

Remembering that $\rho = n\lambda\bar{s}$ and $\rho_1 = \lambda\bar{s}$, respectively.

The following algebraic manipulation gives us

$$n\lambda\bar{s} + n\lambda \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)d_i}{v n E[z_i]} \right) < 1 \Rightarrow n\lambda \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)d_i}{v n E[z_i]} \right) < 1 - \rho \text{ which implies that}$$

$$\sum_{i=1}^{i=n} \frac{d_i \lambda}{v E[z_i]} < \frac{1 - \rho}{1 - \rho_1} \quad (3.5)$$

From (3.2) and (3.5), we know that

$$\text{minimizing } \sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)E[z_i]}{2n\lambda} \right) - \frac{1-\rho_1}{2\lambda}$$

$$\text{subject to: } \sum_{i=1}^{i=n} \frac{d_i \lambda}{v E[z_i]} < \frac{1-\rho}{1-\rho_1}$$

leads us to a lower bound for \bar{W}_c .

Using classical methods to solve the above problem, we obtain, $\bar{W}_c \geq \omega_1$.

$$\text{Where, } \omega_1 = \frac{(1-\rho_1)^2}{2nv(1-\rho)} \left(\sum_{i=1}^{i=n} \sqrt{d_i} \right)^2 - \frac{(1-\rho_1)}{2\lambda}.$$

Another obvious lower bound can be obtained from an M/G/1 queueing model with zero switchover cost, \bar{W}_1 . Combining this lower bound with the previous one, we prove lemma 1.

3.3.2 A Lower Bound for Algorithms which do not (necessarily) Obey the Continuous Condition

For algorithms that do not (necessarily) obey the continuous condition we must make some observations and define two new variables. First we observe that in these systems, a server may arrive at a node, provide continuous service to customers until the sub-queue is empty. We call this the initial busy period. The server may then remain idle at the current location until a new customer arrives at that location. At that time it enters into what we refer to as a subsequent busy period. In principle, a server may have many of these subsequent busy periods (later we show that only poor algorithms would allow this). Let p represent the fraction of customers that arrive during the initial busy periods and $(1-p)$ represent the fraction that arrive during either the idle periods or the subsequent busy periods.

Let A represent the set of $2n$ -tuples $(\alpha_1, \alpha_2, \dots, \alpha_n, p_1, p_2, \dots, p_n)$ which satisfy the following constraints:

$$\text{i) } \sum_{i=1}^{i=n} \frac{\lambda d_i p_i (1 - \rho_i)}{v \alpha_i} < 1 - \rho - (n - \rho)(1 - p), \text{ ii) } \frac{1}{n} \sum_{i=1}^{i=n} p_i = p, \text{ and iii) } p_i \leq 1 \text{ for any } i \text{ (3.6)}$$

Now let $\omega_2 = \min_A \left\{ \sum_{i=1}^{i=n} \left[\frac{(1-\rho_1)p_i\alpha_i}{2n\lambda} \right] \right\} - \frac{p(1-\rho_1)}{2\lambda}$,

$$\omega_3 = \left(\frac{np^2(1-\rho_1)^2}{2v(1-\rho-(n-\rho)(1-p))} \right) \left(\frac{n}{\sum_{i=1}^{i=n} \frac{1}{d_i}} \right) - \frac{p(1-\rho_1)}{2\lambda},$$

$$\text{and } \omega_4 = \left(\frac{n(1-\rho_1)^2}{2v(1-\rho)} \right) \left(\frac{n}{\sum_{i=1}^{i=n} \frac{1}{d_i}} \right) - \frac{1-\rho_1}{2\lambda}.$$

Now let \bar{W}_{nc} represent the average waiting time for an algorithm that does not satisfy the continuous condition.

Lemma 3.2: $\bar{W}_{nc} \geq \max \{W_1, \min_p \{\omega_2\}\} \geq \max \{W_1, \min_p \{\omega_3\}\} \geq \max \{W_1, \omega_4\}$.

Proof of lemma 3.2

Observation 3.4: (A lower bound for \bar{W}_{nc} based on $E[z_i]$)

$E[Z_i]$ is the average number of customers served during the initial busy period per visit to node i . Let p_i represent the fraction of customers served at node i during an initial busy period and let $(1-p_i)$ represent the fraction of customers served at node i during subsequent busy periods. As we have shown in the proof of lemma 3.1, if we only consider the customers served during the initial busy periods, the average waiting time

for these customers is bounded from below by $\sum_{i=1}^{i=n} \left(\frac{(1-\rho_1)(E[z_i]-1)}{2n\lambda} \right)$ and a lower bound

for the average waiting time for the customers at the node i is $\frac{(1-\rho_1)(E[z_i]-1)}{2\lambda}$.

Therefore, a lower bound for the average waiting time in the system is given by

$$\sum_{i=1}^{i=n} \left(\frac{p_i(1-\rho_1)(E[z_i]-1)}{2n\lambda} \right) = \sum_{i=1}^{i=n} \left(\frac{p_i(1-\rho_1)E[z_i]}{2n\lambda} \right) - \frac{p(1-\rho_1)}{2\lambda}$$

$$\text{subject to } \frac{1}{n} \sum_{i=1}^{i=n} p_i = p. \quad (3.7)$$

Observation 3.5: (A constraint for $E[z_i]$ based on p_i)

The average switchover cost per demand served is $E \left[\frac{d_i}{v(Z_i + X_i)} \right]$ if the demand is from

node i , where X_i is the number of customers served during the subsequent busy periods.

Because at steady-state, on average, a fraction of the total customers equal to p_i comes

from Z_i and $(1-p_i)$ comes from X_i , $\frac{E[X_i]}{E[Z_i]} = \frac{1-p_i}{p_i} \Rightarrow E[X_i] = \frac{1-p_i}{p_i} E[Z_i]$.

$$E \left[\frac{d_i}{v(Z_i + X_i)} \right] \geq \frac{d_i}{E[v(Z_i + X_i)]}$$

$$\frac{d_i}{E[v(Z_i + X_i)]} = \frac{d_i}{vE[Z_i + X_i]} = \frac{p_i d_i}{vE[Z_i]}$$

Combining with (3.1), we have : $E\left[\frac{d_i}{v(Z_i + X_i)}\right] \geq \frac{p_i(1-\rho_1)d_i}{vE[z_i]}$.

Now we have a lower bound for the average switchover cost per demand served:

$$\frac{1}{n} \sum_{i=1}^{i=n} \frac{p_i(1-\rho_1)d_i}{vE[z_i]} \quad (3.8)$$

Now we make the following observation: the average idle time per demand served during each subsequent busy period is bounded from below by $\frac{1}{\lambda}(1-\rho_1)$. This comes from dividing the average interarrival time by the average number of customers served during a single busy period in an M/G/1 queue. So a lower bound for the average extra-cost due to idle periods per customer served for customers at node i is $\frac{1}{\lambda}(1-\rho_1)(1-p_i)$ and the average extra-cost due to idle periods per overall demand served is bounded by

$$\frac{1}{n} \sum_{i=1}^{i=n} \left[(1-p_i) \frac{1}{\lambda} (1-\rho_1) \right] = \frac{1-p}{\lambda} (1-\rho_1). \quad (3.9)$$

From (3.8) and (3.9), we know that the average extra-cost due to switching and idling is

$$\text{at least } \sum_{i=1}^{i=n} \frac{p_i(1-\rho_1)d_i}{nvE[z_i]} + \frac{1-p}{\lambda} (1-\rho_1). \quad (3.10)$$

Because our system is in steady-state $n\lambda \left[\frac{1}{s} + \sum_{i=1}^{i=n} \frac{p_i(1-\rho_1)d_i}{nvE[z_i]} + \frac{1-p}{\lambda} (1-\rho_1) \right] < 1$.

$$\text{Algebraic manipulation leads to } \sum_{i=1}^{i=n} \left[\frac{\lambda d_i p_i (1-\rho_1)}{vE[z_i]} \right] < 1 - \rho - (n-\rho)(1-p). \quad (3.11)$$

Observation 6: (A lower bound for \overline{W}_{nc})

In order to find a bound on the average waiting time in the system we solve a minimization problem where $E[z_i]$ and p_i are variables, the objective function is given by (3.7) and the constraints are implied by (3.11). We show this problem below.

$$\begin{aligned} &\text{Minimizing } \sum_{i=1}^{i=n} \left(\frac{p_i(1-\rho_1)\alpha_i}{2n\lambda} \right) - \frac{p(1-\rho_1)}{2\lambda} \\ &\text{subject to } \sum_{i=1}^{i=n} \left[\frac{\lambda d_i p_i (1-\rho_1)}{v\alpha_i} \right] < 1 - \rho - (n-\rho)(1-p) \end{aligned}$$

leads to us a lower bound for \overline{W}_{nc} based on p .

If $p=1$, this leads to the continuous case and we obtain the same result as before; for $p < 1$, we cannot obtain an explicit expression. Therefore, we relax some constraints, which means we obtain a looser but explicit lower bound.

$$\text{The lower bound is } \omega_2, \text{ defined earlier to be } \omega_2 = \min_i \left\{ \sum_{i=1}^{i=n} \left[\frac{(1-\rho_1)p_i\alpha_i}{2n\lambda} \right] \right\} - \frac{p(1-\rho_1)}{2\lambda}.$$

Relaxing the constraints that $p_i \leq 1, \forall i$ and minimizing the $\sum_{i=1}^{i=n} \left[\frac{p_i(1-\rho_1)\alpha_i}{2n\lambda} \right]$ leads us to the following lower bound based on p . We observe that this is equal to ω_3 , defined

$$\text{earlier as } \left(\frac{np^2(1-\rho_1)^2}{2v(1-\rho-(n-\rho)(1-p))} \right) \left(\frac{n}{\sum_{i=1}^{i=n} \frac{1}{d_i}} \right) - \frac{p(1-\rho_1)}{2\lambda}.$$

If we minimize ω_2 and ω_3 over the range of p , we get $\bar{W}_{nc} \geq \min_p \{\omega_2\} \geq \min_p \{\omega_3\}$.

When the number of nodes $n \geq 2$, we can show that $\frac{np^2}{2v(1-\rho-(n-\rho)(1-p))}$
 $\geq \frac{n}{2v(1-\rho)}$ and because $\frac{1-\rho_1}{2\lambda} \geq \frac{p(1-\rho_1)}{2\lambda}$, we have $\bar{W}_{nc} \geq \omega_4$.

Another obvious lower bound is $\bar{W}_{nc} \geq \bar{W}_1$. Combining these, we have proven lemma 3.2.

3.3.3 General Lower and Upper Bounds for the Optimal Algorithm

Let \bar{W}^* be the average waiting time for optimal algorithm.

Theorem 3.1 $\bar{W}^* \geq \max \{W_1, \min [\omega_1, \min_p \{\omega_2\}]\} \geq \max \{W_1, \omega_4\}$.

Proof of theorem 3.1

$\bar{W}^* \geq \max \{W_1, \min [\bar{W}_c, \bar{W}_{nc}]\}$, from lemmas 3.1 and 3.2, we prove theorem 3.1.

Theorem 3.2 When $d_i = d_1, \forall i$, $\bar{W}^* \geq \max \left\{ \bar{W}_1, \frac{nd_1(1-\rho_1)^2}{2v(1-\rho)} - \frac{1-\rho_1}{2\lambda} \right\}$.

Proof of theorem 3.2

In this case, $\omega_1 = \omega_4$, from theorem 3.1, we know $\bar{W}^* \geq \max \left\{ W_1, \frac{nd_1(1-\rho_1)^2}{2v(1-\rho)} - \frac{1-\rho_1}{2\lambda} \right\}$.

From Cooper, Niu and Srinivasan (1996), we know that the average waiting time for the M/G/1 queue with switching costs under cyclic polling, \overline{W}_{cyclic} , is the following:

Theorem 3.3 (Due to Cooper, Niu and Srinivasan, 1996)
$$\overline{W}_{cyclic} = W_1 + \frac{(1-\rho_1) \sum_{i=1}^{i=n} d_i}{2v(1-\rho)}$$

Further $\overline{W}^* \leq \overline{W}_{cyclic}$.

Consider the case where $d_i = d$. Let $x = \frac{v\lambda s^2}{d}$ (the ratio of λs^2 to the switching cost for a single switchover). We compare the average waiting time under cyclic polling with that of the optimal algorithm in the corollaries below. Corollary 3.2 explains corollary 3.1 in words.

Corollary 3.1: If $d_i = d$, for all i and if $x \leq 1$, then
$$\lim_{\rho \rightarrow 1} \frac{\overline{W}_{cyclic} - \overline{W}^*}{\overline{W}^*} \leq \frac{\rho_1}{1-\rho_1} + x;$$

If $x > 1$,
$$\lim_{\rho \rightarrow 1} \frac{\overline{W}_{cyclic} - \overline{W}^*}{\overline{W}^*} \approx \frac{1-\rho_1}{x}.$$

Corollary 3.2: For the special case in which switchover costs are uniform, as ρ goes to 1, if x is very large, the cyclic polling algorithm is asymptotically the same as the optimal; if x is very small, when n is very large, the cyclic polling algorithm is also asymptotically the same as the optimal.

In fact, we can interpret $\frac{n(1-\rho)^2}{2(1-\rho)v}$ as the contribution of the switchover to the total cost and \bar{W}_1 as the contribution of the randomness of the arrival process and service time.

3.4 Lower and Upper Bound for the Optimal Algorithm Under Light Traffic

If the arrival rate λ is very low, we develop an alternative lower bound. We introduce a heuristic algorithm that locates the server at the node with the longest set-up time. In this way, some of the set-up time is absorbed into the system idle time. When $\lambda \rightarrow 0$, this algorithm achieves the alternative lower bound. Under light traffic this algorithm is approximately optimal. First we introduce the heuristic algorithm and then provide the result in Theorem 3.4.

A New Heuristic

Let d_i represent the switching cost incurred in switching to node i . Note that this is independent of node that the server switches from. Sort d_i in non-decreasing order as $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}$, and locate the server at the node with the longest set-up time $d_{(n)}$. When a customer arrives at a node, the server switches to that node and provides the service to that customer. Immediately after completing service server switches back to its original location.

Let the average waiting time for this algorithm be \bar{W}_H .

Theorem 3.4 $\bar{W} \geq \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$ where $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}$ and $\lim_{\lambda \rightarrow 0} \bar{W}_H = \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$.

Proof of theorem 3.4

We consider the waiting time for customer i . We divide this waiting time into two components: the waiting time due to the server's travel prior to serving customer i , and the waiting time due to the on-site service time of customers served prior to customer i .

\bar{W}_i^d and \bar{W}_i^s represent the two components, respectively. Because we use ideas very close to those of Bertsimas and van Ryzin (1991) to obtain our first bound we use notation identical to theirs in this section. We have the following relationship,

$\bar{W}_i = \bar{W}_i^d + \bar{W}_i^s$. Taking expectations and letting i approach infinity

($\bar{W}^d = \lim_{i \rightarrow \infty} E[\bar{W}_i^d]$ and $\bar{W}^s = \lim_{i \rightarrow \infty} E[\bar{W}_i^s]$) we obtain $\bar{W} = \bar{W}^d + \bar{W}^s$.

A lower bound for \bar{W}^d is the expected travel time between the optimal location of the

server and the location of a random demand, $\bar{W}^d \geq \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$, therefore $\bar{W} \geq \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$.

In order to complete our proof we need to calculate the average waiting time for the heuristic algorithm. Because the server returns to the same location after every service completion, we can use an M/G/1 queueing model to calculate the average waiting time \bar{W}^s .

Because the on-site service time is independent of the travel time, we know that the first

and second moments of the service time are bounded from above by $\bar{s} + \frac{2d_{(n)}}{v}$ and

$\bar{s}^2 + \frac{4d_{(n)}^2}{v^2}$, respectively. Using the classical Pollaczek-Khinchin (P-K) formula,

$$\bar{W} = \frac{\lambda s_c^2}{2(1-\rho_c)},$$

where the subscript c represents the classical definitions for the second

moment of the service time and the utilization factor, we obtain

$$\bar{W}_H^v \leq \frac{n\lambda \bar{s}^2}{2(1-\hat{\rho})} + \frac{4n\lambda d_{(n)}^2}{v(1-\hat{\rho})} \text{ where } \hat{\rho} = \rho + \frac{2n\lambda d_{(n)}}{v}.$$

Letting e be the event that when a new customer arrives in the system the server is busy, we obtain the following

$$\bar{W}_d^H \leq P\{e\} E[W_i^d | e] + P\{\bar{e}\} E[W_i^d | \bar{e}] \leq (1-\rho) \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv} + 2\hat{\rho} \frac{d_{(n)}}{v}.$$

$$\rho \rightarrow 0, \hat{\rho} \rightarrow 0 \quad \bar{W}_H^v \leq \frac{n\lambda \bar{s}^2}{2(1-\hat{\rho})} + \frac{4n\lambda d_{(n)}^2}{v(1-\hat{\rho})} \rightarrow 0 \text{ as } \lambda \rightarrow 0.$$

$$\text{Therefore } \lim_{\lambda \rightarrow 0} \bar{W}_H \leq \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}.$$

Because $\lim_{\lambda \rightarrow 0} \bar{W} \geq \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$, our upper and lower bounds for $\lim_{\lambda \rightarrow 0} \bar{W}_H$ are identical.

Therefore we obtain $\bar{W}_H \rightarrow \frac{\sum_{i=1}^{i=n-1} d_{(i)}}{nv}$ as $\lambda \rightarrow 0$.

3.5 Simulation

3.5.1 The Simulation Model

The simulation model was developed in C++. The run lengths for the simulation are very long. It is well known that even simple queuing systems require very long periods to reach steady state behavior. Empirical analysis of our system showed that this was much more true than we could have imagined. However, we found that including a warm up period, however long, had no impact on the long run averages of the performance measures of the system. This is because even for rather high values of ρ , the fraction of time the server is busy, the system returns periodically to an empty state. We present results obtained from serving one million consecutive customers. Because our simulation run times are so long, we do not present confidence intervals for the performance measures of interest. Instead, we present these performance measures directly. The two performance measures of interest are the average wait time for service and the average number of customers in the system. We compare the following policies: cyclic polling with exhaustive service, cyclic polling with gated service and longest queue first with exhaustive service. These results assume that the service times are exponential.

3.5.2 Simulation Results

We begin by presenting a comparison of closed form results for cyclic polling systems in which switchover costs are constant versus our simulation results. We show this in figure 3.1. The cyclic polling systems examined here assume that the servers are “impatient”. This means that they keep moving when the system is empty. One can see that the results of our simulation model are virtually identical to those predicted by the closed form solution for the number of customers in the system under exhaustive cyclic polling.

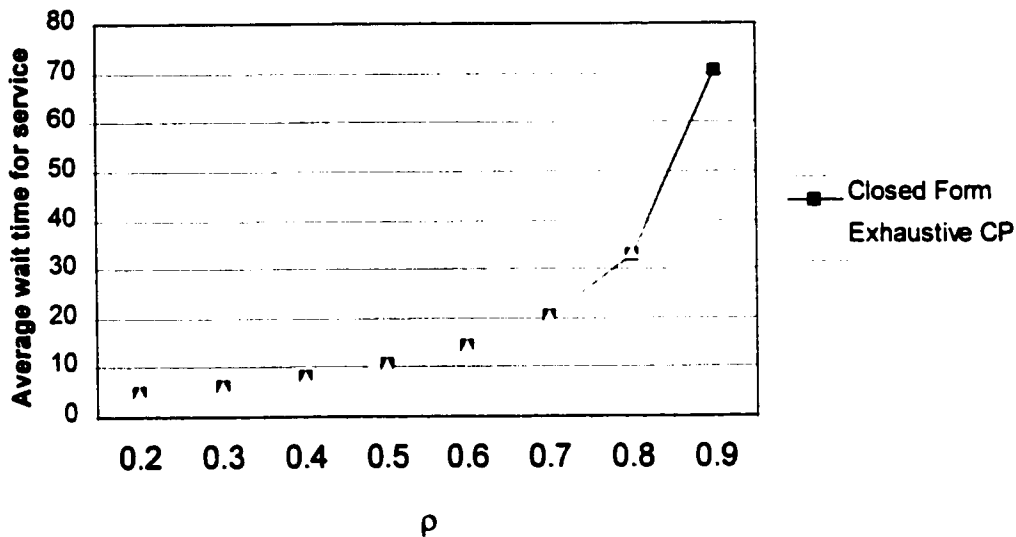


Figure 3.1 Closed Form Solution vs. Simulation Results

We next examine the relative performance of three heuristics for the M/G/1 queue with switchover cost. As predicted, longest queue first out-performs the others when the switching costs are constant. Figure 3.2 presents those results. However, we find that for high values of ρ , the performance of cyclic polling is very close to that of longest queue

first. Figure 3.3 presents their relative performance for values of ρ between 0.95 and 0.99.

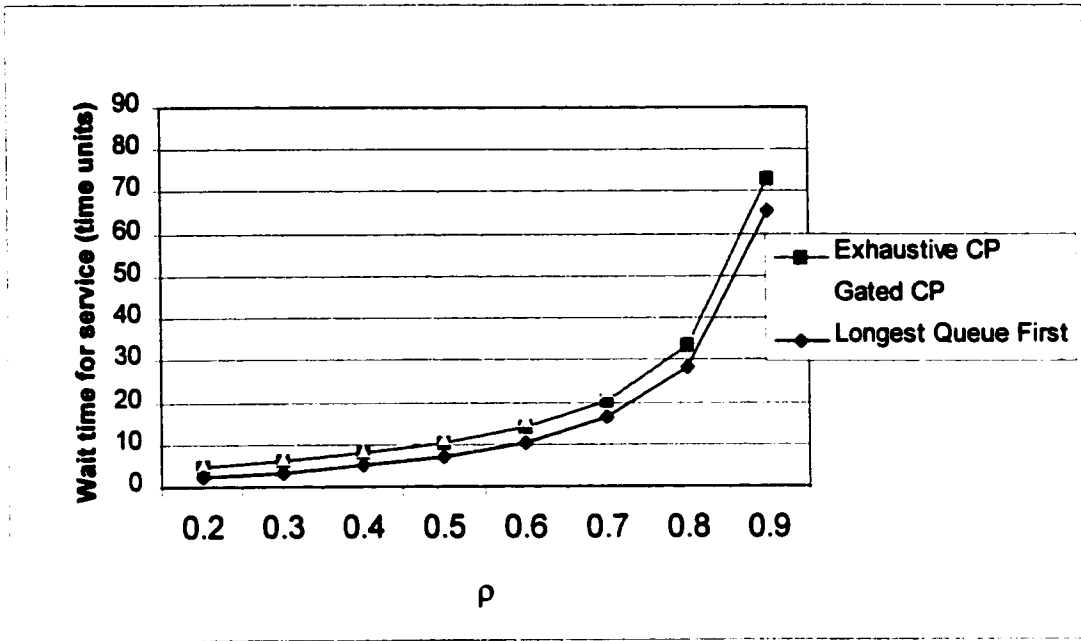


Figure 3.2 A Comparison of Three Heuristics when the Switching Costs are Constant

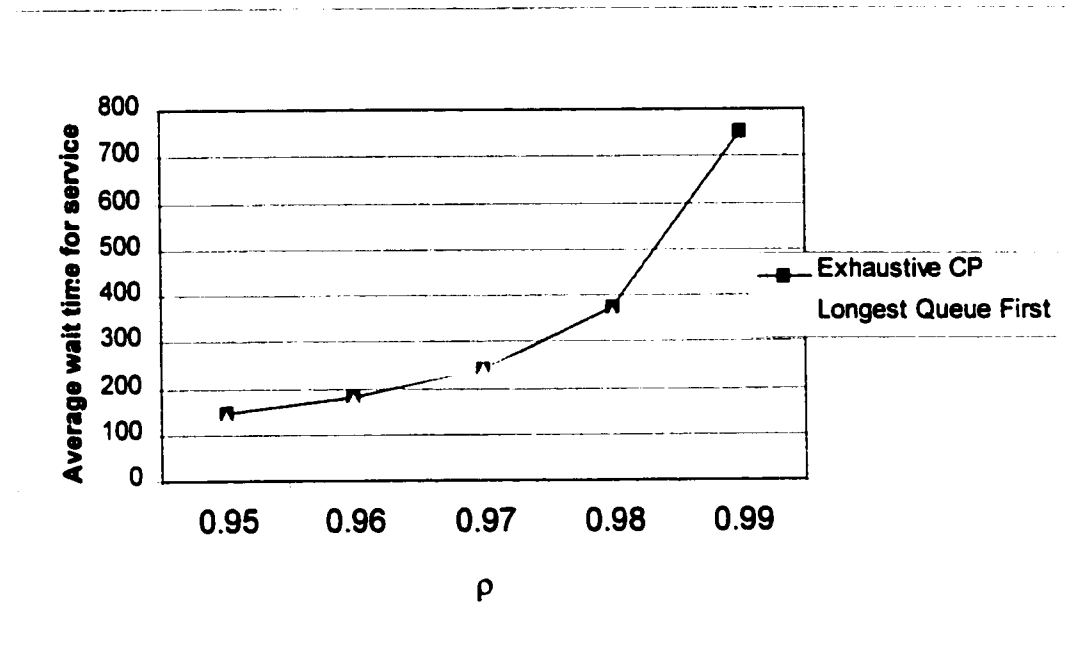


Figure 3.3 The relative performance of cyclic polling and longest queue first for ρ Approaching 1

Now we examine the benefits of allowing a “patient” server to wait if there are no customers in the system. We assume that the server waits at its current location until a service request arrives in the system. Then it begins to move again. While the gains over the case with the impatient server are small, they are non-zero. The intuition here is that since arrivals occur according to the same Poisson process at each node, the next arrival is equally likely to occur at any node. It will definitely not occur between nodes, which is where an impatient server will be when the next customer arrives.

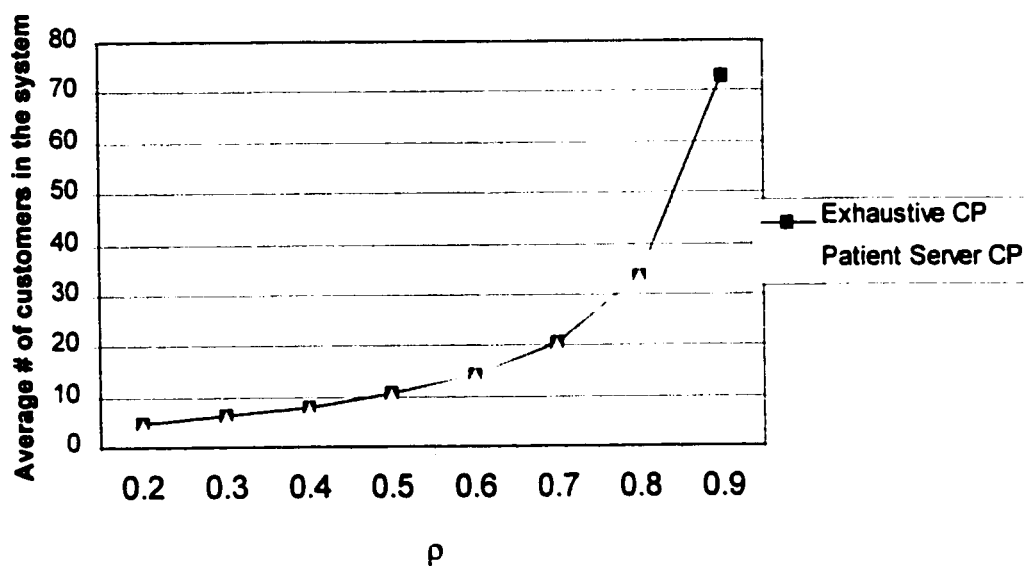


Figure 3.4 The benefit of allowing the server to be patient when no customers are in the system.

3.6 Conclusion

We examine the M/G/1 queueing model with switchover costs. We develop a lower bound for the waiting time in these systems under any arbitrary algorithm, including those that are optimal. We point out that in general this is not a very tight bound. Next we examine systems in which service is provided according to a cyclic polling algorithm. We show that for the special case where the switchover costs are identical and traffic intensity is high, the average waiting time of cyclic algorithm is bounded by approximately 2 times the average waiting time of optimal algorithm. We also show that for this special case when $\rho \rightarrow 1$ and x (the ratio of $\overline{\lambda s^2}$ to the switching cost for a single switchover) is very small or when $\rho \rightarrow 1$ and x and n , the number of individual queues in our system, are both very large, cyclic polling is also close to optimal. Under the special case of very low demand intensity, cyclic polling performs poorly. In this case we provide an alternative heuristic and a lower bound for the average waiting time of optimal algorithm. When $\lambda \rightarrow 0$, our heuristic algorithm is approximately optimal. The simulation results indicates that the exhaustive longest queue first policy is better than the exhaustive cyclic and gated cyclic. In addition, letting the server be patient is slighter better continuing to cycle when no customers are present in the system.

CHAPTER 4 THE DYNAMIC TRAVELING SALESMAN

PROBLEM: AN EXAMINATION OF ALTERNATIVE HEURISTICS

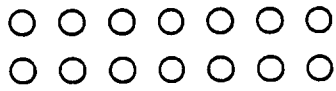
4.1 Introduction

The dynamic traveling salesman problem concerns the development of a routing policy for a single mobile server providing service to customers whose positions are known. Service requests are generated according to a Poisson process which is uniform across customer locations. We assume that the mean service time is known and its variance is bounded. Service time is independent of customer location. This problem, called the Dynamic Traveling Salesman Problem (DTSP), was first introduced by Psaraftis (1985). Bertsimas and van Ryzin (1991) studied a similar problem, the Dynamic Repairman Problem (DTRP), in which customer locations are either uniformly distributed in a bounded area in the Euclidean plane or distributed according to a distribution with probability density function $f(x)$.

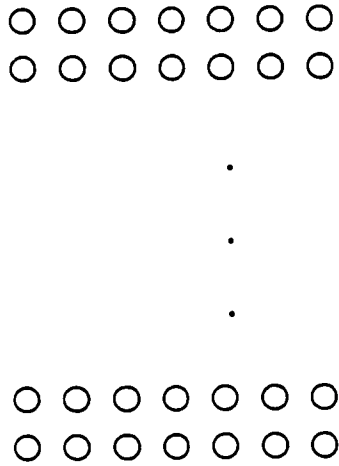
In this chapter we begin by examining a special case of the DTSP. The special case involves networks in which the optimal TSP tour and minimum spanning tree across customer locations involve only links of equal length (see Figure 4.1 for some examples). For this special case, through analysis of a related queueing system, we show that the average waiting time when the server follows the a priori tour generated by the well known “cyclic polling” algorithm is approximately bounded by $\frac{2-\rho_1}{1-\rho_1}$ times the average

waiting time of the optimal algorithm where ρ_1 , the fraction of time spent in on-site service time for a single node, is less than $\frac{1}{n}$. Note that when n is large $\frac{2-\rho_1}{1-\rho_1}$ is close to 2, and $\frac{2-\rho_1}{1-\rho_1}$ is always bounded by 3. We also identify circumstances under which our bound is very tight. This implies that under certain conditions the cyclic polling algorithm is close to optimal.

Next, we introduce a heuristic algorithm for the DTSP on a general graph. We provide a lower bound on the waiting time for the optimal algorithm. From Cooper, Niu and Srinivasan (1996), we know the average waiting time of the cyclic polling algorithm. This provides us with an upper bound for the average waiting time under the optimal algorithm. Finally, when the arrival rate is very low, we provide an alternative heuristic and show it is approximately optimal as the arrival rate approaches zero. We also present some simulation result for randomly generated six nodes network which shows that the cyclic polling algorithm is robust for these networks.



Example network with two rows



Example network with $2 \cdot K$ rows ($K = 1, 2, 3, \dots$)

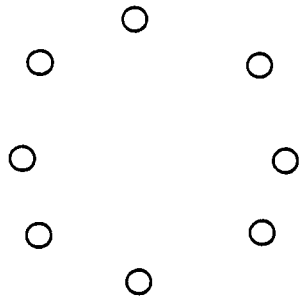


Figure 4.1. Example networks

4.2 The DTSP on Networks in which the Optimal TSP Tour and Minimum Spanning Tree Involve Only Links of Equal Length

We assume here that the length of each link involved in the minimum spanning tree is 1 and the travel speed is v , so $\frac{1}{v}$ is the minimum switchover time between any two nodes.

4.2.1 Notation

Let

n represent the number of nodes in the network,

\bar{s} and \bar{s}^2 represent the first and second moments of the on-site service time for each demand served, respectively,

λ the parameter for the Poisson process at each node which is uniform over all nodes,

$\rho = n\lambda \bar{s}$ and $\rho_1 = \lambda \bar{s}$ the fraction of time the server spends providing on-site service to all nodes and the fraction of time spent in on-site service for a single node, respectively,

$\bar{W}_1 = \frac{n\lambda \bar{s}^2}{2(1-\rho)}$, the average waiting time for the classical M/G/1 without switchover cost,

\bar{W}^* = the average waiting time for the optimal algorithm,

\bar{W}_{cyclic} = the average waiting time for cyclic algorithm,

In the next sub-section, we examine the DTSP on graphs in which the TSP tours and minimum spanning trees include only links with equal length.

4.2.2 The DTSP on the Special Graph

Theorem 4.1: $\bar{W}_{cyclic} = \bar{W}_1 + \frac{n(1-\rho_1)}{2v(1-\rho)}$ and $\bar{W}^* \geq \max \left\{ \bar{W}_1, \frac{n(1-\rho_1)^2}{2v(1-\rho)} - \frac{1-\rho_1}{2\lambda} \right\}$.

Proof of theorem 4.1

If we regard the switchover time as the time between the time when the server leaves a node until it reaches another node then the switchover time must be at least $\frac{1}{v}$. From Theorem 3.2, we have a lower bound for the average waiting time for the optimal algorithm. From Theorem 3.3, we know the average waiting time for the cyclic algorithm. Together these give us theorem 4.1.

For the case in which $d_i = 1$, let $x = v\lambda \bar{s}^2$ (the ratio of $\lambda \bar{s}^2$ to the switching cost for a single switchover), we compare the average waiting time under cyclic polling with that of the optimal algorithm in the corollaries below. Corollary 4.2 explains corollary 4.1 in words.

Corollary 4.1: If $d_i = 1$, for all i and if $x \leq 1$, then $\lim_{\rho \rightarrow 1} \frac{\bar{W}_{cyclic} - \bar{W}^*}{\bar{W}^*} \leq \frac{\rho_1}{1-\rho_1} + x$; If

$$x > 1, \lim_{\rho \rightarrow 1} \frac{\bar{W}_{cyclic} - \bar{W}^*}{\bar{W}^*} \approx \frac{1-\rho_1}{x}.$$

Corollary 4.2: For the special case in which switchover costs are uniform, as ρ goes to 1, if x is very large, the cyclic polling algorithm is asymptotically the same as the optimal algorithm; if x is very small, when n is very large, the cyclic polling algorithm is also asymptotically same as the approximately optimal.

4.3 The DTSP on a General Graph

4.3.1 A Heuristic Algorithm

First, find a TSP solution for all nodes in the network and then visit the nodes along the TSP tour, providing exhaustive service at each node and skipping nodes with no demands. This is essentially a cyclic polling algorithm, we use \overline{W}_{cyclic} to represent the average waiting time for this algorithm.

4.3.2 Properties of the Heuristic Algorithm

We develop a lower bound for the average waiting time for the optimal algorithm from the Theorem 3.1. We also obtain the average waiting time in our system from the work of Cooper, Niu and Srinivasan (1996).

Let $\omega_1 = \frac{n(1-\rho_1)^2}{2\nu(1-\rho)} \left(\frac{n}{\sum_{i=1}^{i=n} \frac{1}{\widehat{d}_i}} \right) - \frac{1-\rho_1}{2\lambda}$ where $\widehat{d}_i = \min_{j \neq i} \{d_{ji}\}$ and d_{ji} is the distance

between nodes j and i .

Theorem 4.3: $\overline{W}_{cyclic} \leq \overline{W}_1 + \frac{(1-\rho_1)L_{TSP}}{2\nu(1-\rho)}$ where L_{TSP} is the length of the optimal TSP

tour over all the nodes and $\overline{W} \geq \max\{\overline{W}_1, \omega_1\}$.

Proof of theorem 4.3

We can see the switchover time for a whole cycle is $\frac{L_{TSP}}{\nu}$, from lemma 3, we know the

first part of the theorem. Because \widehat{d}_i is the minimum distance traveled to reach node i ,

it therefore provides a lower bound on the switchover cost to node i . From theorem 3.1 we know the result.

4.4 The DTSP on the General Graph under Light Traffic Intensity

First we define a median of a graph as a location such that the average distance to all nodes is minimized, let B represent the bounded region:

$$\frac{1}{n} \sum_{i=1}^{i=n} \|X_0 - X_i\| = \min_{X \in B} \left\{ \frac{1}{n} \sum_{i=1}^{i=n} \|X - X_i\| \right\} \text{ where } \|X_i - X_0\| \text{ is the Euclidean distance}$$

between X_i and X_0 .

If the arrival rate λ is very low, we introduce a heuristic algorithm that locates the server on the median of the graph, whenever there is demand, the server leaves the median and goes directly to the node to provide service. After the demand is served, it returns immediately to the median. Let the average waiting time for this algorithm be \bar{W}_H .

Theorem 4.2: $\bar{W} \geq \min_{\{X_0 \in B\}} \left\{ \frac{1}{n\nu} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}$ and $\bar{W}_H \rightarrow \min_{\{X_0 \in B\}} \left\{ \frac{1}{n\nu} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}$

as $\lambda \rightarrow 0$.

Proof of theorem 4.2

We consider the waiting time for demand i and we divide the waiting time into two components: the first is the waiting time due to the server's travel prior to serving i , and the second is the waiting time due to the on-site service times of demands served prior to demand i . \bar{W}_i^d, \bar{W}_i^s represent the two components, respectively. We have the following relationship, $\bar{W}_i = \bar{W}_i^d + \bar{W}_i^s$. Taking expectations and letting i approach infinity, $\bar{W}^d = \lim_{i \rightarrow \infty} E[\bar{W}_i^d]$ and $\bar{W}^s = \lim_{i \rightarrow \infty} E[\bar{W}_i^s]$. We have $\bar{W} = \bar{W}^d + \bar{W}^s$.

Essentially, the proof is saying that we have chosen X_0 to minimize travel distance. Since demand is low, there will be no demands waiting at the queue when a new request arrives.

A lower bound for \overline{W}^d is the expected travel time between the optimal location of the server and the location of a random demand. Since $\overline{W}^d \geq \min_{\{X_0 \in B\}} \left\{ \frac{1}{nv} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}$,

$$\overline{W}^d \geq \min_{\{X_0 \in B\}} \left\{ \frac{1}{nv} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}.$$

In order to complete our proof we need to calculate the average waiting time for the heuristic algorithm. Because the server returns to the median every time it finishes a demand and it begins service from the same location every time, we can use an M/G/1 queue model to calculate the average waiting time due to the on-site service time of demands served prior to demand i (This idea is due to Berman, Larson and Chiu, 1985).

Because the on-site service time is independent of the travel time, the first and second moments of the service time are bounded from above by $\overline{s} + \frac{2 \sum_{i=1}^{i=n} \|X_i - X_0\|}{v}$,

$$\overline{s}^2 + 4 \left(\frac{\sum_{i=1}^{i=n} \|X_i - X_0\|}{v} \right)^2 + 2 \frac{\overline{s} \left(\sum_{i=1}^{i=n} \|X_i - X_0\| \right)}{v} \text{ respectively. This is obtained from the}$$

classical result of the P-K formula: $\overline{W} = \frac{\lambda \overline{s_c^2}}{2(1 - \rho_c)}$, where the subscript c represents the

classical definitions for the second moment of the service time and the utilization factor.

Let \overline{W}_H^s represent the average waiting time due to the on-site service time of customers served prior to the selected customer.

$$\bar{W}_H^s \leq \frac{n\lambda \bar{s}^2}{2(1-\hat{\rho})} + \frac{2n\lambda \left(\sum_{i=1}^{i=n} \|X_i - X_0\| \right)^2}{v^2(1-\hat{\rho})} + \frac{n\lambda \bar{s} \left(\sum_{i=1}^{i=n} \|X_i - X_0\| \right)}{v(1-\hat{\rho})}$$

$$\text{where } \hat{\rho} = \rho + \frac{2n\lambda \sum_{i=1}^{i=n} \|X_i - X_0\|}{v}.$$

We observe that as $\lambda \rightarrow 0$, $\bar{W}_H^s \rightarrow 0$.

Next we examine the average waiting time due to the server's travel prior to serving the selected customer.

Let \bar{W}_H^d represent this average waiting time. $\bar{W}_H^d = \min_{\{X_0 \in B\}} \left\{ \frac{1}{nv} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}$.

Because $\bar{W}_H = \bar{W}_H^d + \bar{W}_H^s$ as $\lambda \rightarrow 0$, therefore, $\bar{W}_H \rightarrow \min_{\{X_0 \in B\}} \left\{ \frac{1}{nv} \sum_{i=1}^{i=n} \|X_0 - X_i\| \right\}$.

4.5 Simulation Results

The same simulation framework discussed in chapter 3 is used for this simulation. The model was developed in C++. The run lengths for the simulation are very long to ensure steady state results.

We compare the performance of the three heuristics for randomly generated networks where the locations of nodes are generated uniformly in a unit square. Then, the relative

positions of the nodes are maintained but their locations are uniformly scaled so that the length of the optimal TSP tour on these nodes is equal to exactly six time (space) units. The simulations are then run on twenty of these random networks and the values obtained for the twenty (one million customer) runs are averaged.

In this case, the cyclic polling tour follows the optimal TSP tour across the customers. The travel time in this case is proportional to distance. Figure 4.2 shows some examples of these networks. The cyclic polling solution, which corresponds to following the a priori generated TSP tour outperforms longest queue first for large values of ρ . Figure 4.3 presents those results. Of particular interest is the value of ρ after which cyclic polling outperforms longest queue first. Figure 4.4 presents more detailed simulation (smaller step size for ρ) for the region from $\rho = 0.75$ to 0.94 . We can observe that cyclic polling consistently outperforms longest queue first for $\rho > 0.85$.

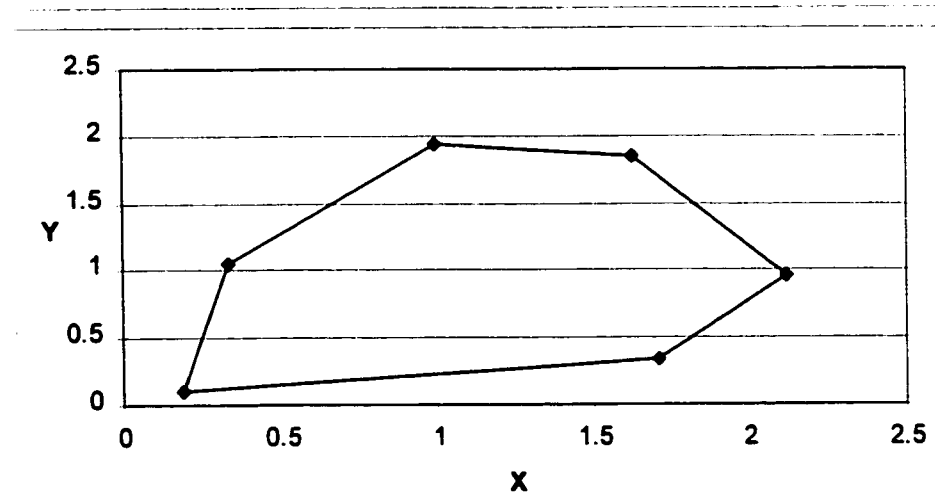
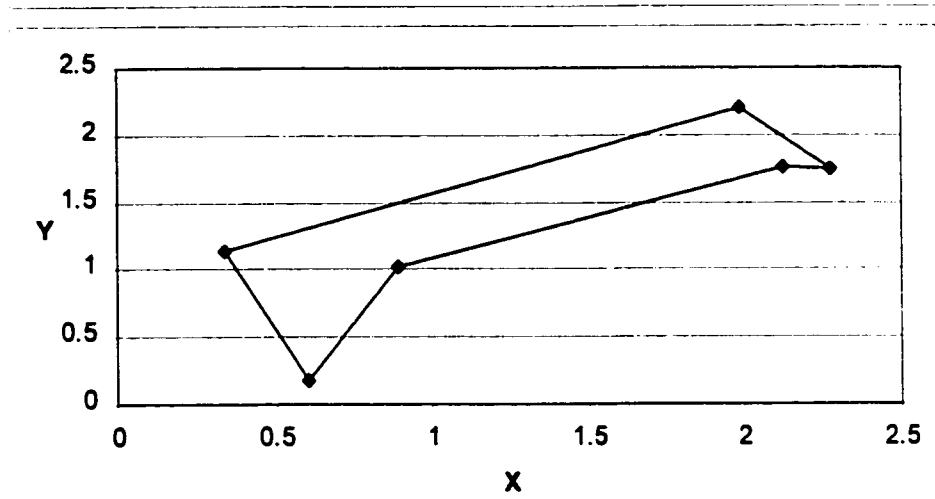
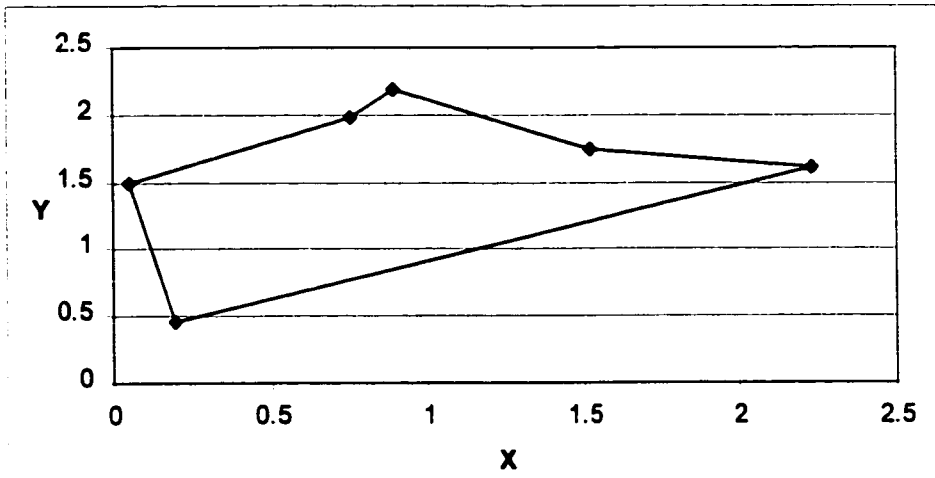


Figure 4.2. Randomly generated six node networks with length of optimal TSP tour of 6 units

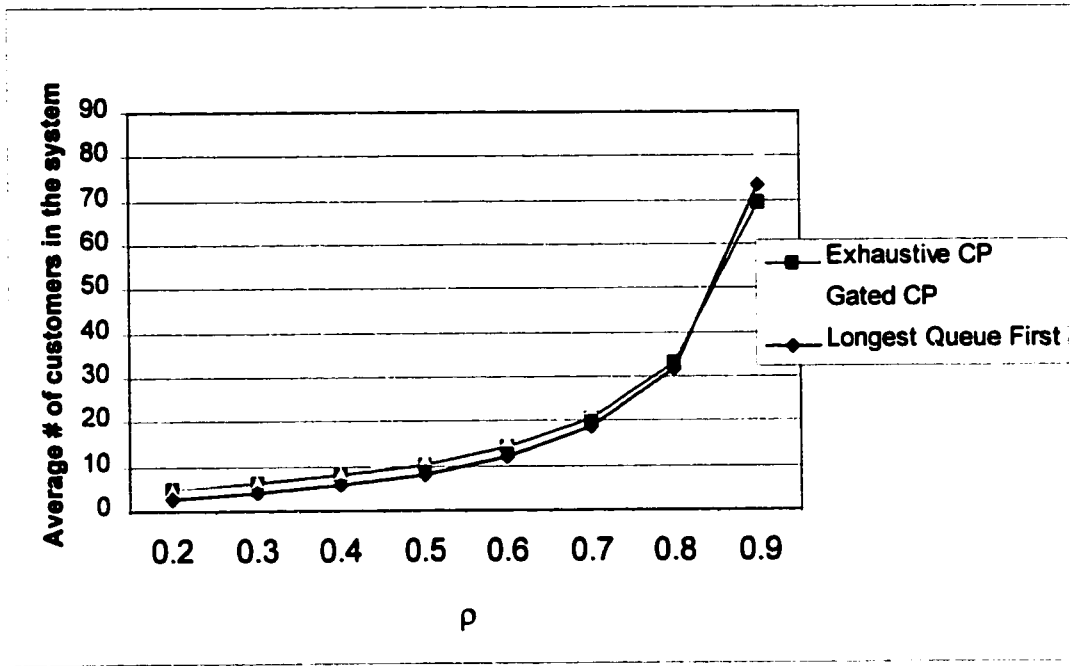


Figure 4.3. A Comparison of Three Heuristics when the Switching Costs are Proportional to Distance (randomly generated 6 node networks)

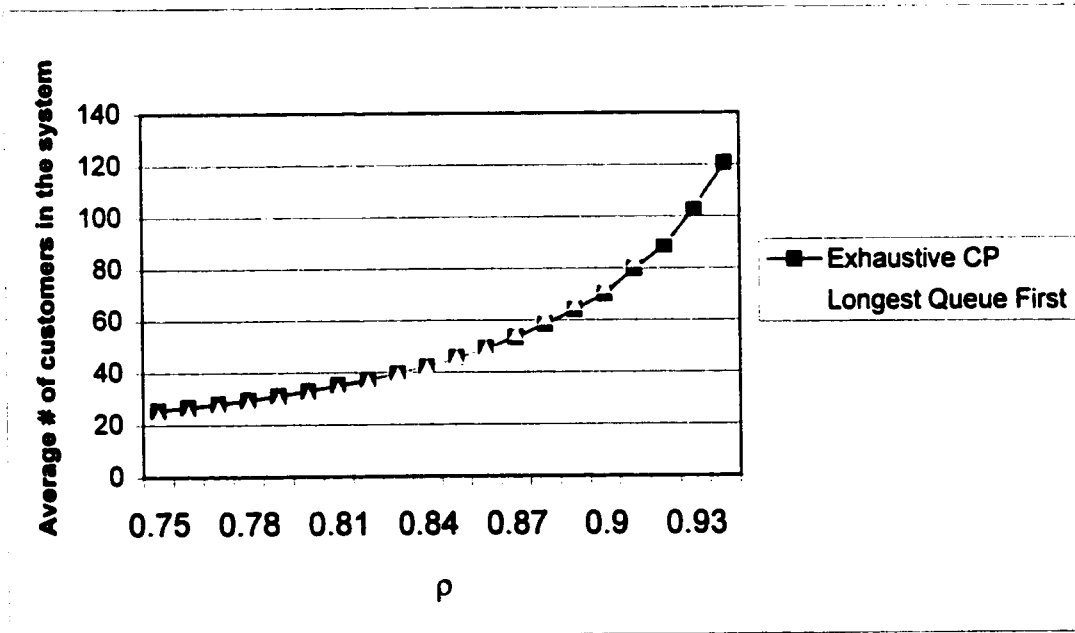


Figure 4.4. An examination of the point at which Cyclic Polling outperforms Longest Queue First when travel time is proportional to Distance (randomly generated 6 node networks)

4.6 Conclusion

In this chapter we examine a very difficult problem, the dynamic traveling salesman problem. We examine both a special case and the general case. For the special case we examine the performance of a specific algorithm. For the general case we provide bounds on the performance of the best on-line algorithms.

Psaraftis (1985) discusses some basic unanswered questions about the DTSP. We repeat these here and then partially address each of these. (1) If the on-site service time is zero, what is the best algorithm? (2) Under what circumstances is the myopic policy, which optimizes over known demands only, best? (3) Does it make sense to let demands accumulate before the vehicle departs?

We address question (1) for the special case of networks in which both the optimal TSP tour and the minimum spanning tree contain only links of equal length. We show that when the demand arrival rate is relatively high and when both the on-site service time and its variance are very small and when the number of locations n is large, the cyclic polling algorithm will be very close to optimal. That is, the server travels along the optimal TSP tour and provides service to each node as it passes the node in its tour. This is a direct result of corollary two.

We address question (2), in part by showing that for these graphs, when the arrival rate is very low, a myopic algorithm is close to optimal. We show this by providing proof for Psaraftis' conjecture that for the case of low demand, the best algorithm is to locate the server on the median and to provide service by traveling from the median to the customer

locations. The case of moderately heavy or moderately light traffic remains an open question.

We address question (3) also in part. Though we do not prove conclusively that algorithms obeying the continuous condition perform better than those which do not (hence, allowing demands to accumulate), we do provide lower bounds on the waiting time for service for both types of algorithms and show that the lower bound for algorithms obeying this condition dominates the lower bound for those that do not. Therefore, we conjecture that these perform better.

In a dynamic or on-line system, decisions are made over time, based on current information. In a static or a priori system, decisions are made before demand is realized. The fact that under certain condition(s) the a priori solution is optimal for the dynamic problem provides another indication of the connection between the static and dynamic solutions to dynamic or on-line problems. This connection has been mentioned by many researchers during the last two decades.

We examine the performance of cyclic polling algorithm and the longest queue first algorithm for a set of 20 randomly generated networks and found that when ρ is big enough, the average performance of the cyclic polling algorithm is better than the longest queue first algorithm.

CHAPTER 5 AN ASYMPTOTICALLY OPTIMAL ALGORITHM FOR THE DYNAMIC TRAVELING REPAIRMAN PROBLEM

5.1 Introduction

The dynamic traveling repair problem (DTRP) is the following: m mobile servers are positioned within a bounded region A in the Euclidean plane. The servers travel at a fixed, constant speed v per time unit. Service requests arrive over time according to a Poisson process with arrival rate λ . When requests arrive, they are distributed to the bounded region A independently according to a uniform distribution. The server must spend some time traveling to the customer locations and it must spend some time providing on-site service. The on-site service time for each customer is independently and identically distributed according to a probability distribution with mean \bar{s} and variance σ_s^2 . The goal is to minimize the average waiting time of all customers. In this chapter we examine this problem and address a conjecture about the asymptotic optimality of a partitioning algorithm for sequencing service to customers made by earlier researchers.

In a paper titled “Stochastic and Dynamic Vehicle Routing with General Demand and Interarrival Time Distributions”, Bertsimas and van Ryzin (1993, p. 962) examined the following $GI/G/m$ service policy, which we refer as the $BvR(n, k)$ algorithm:

For a fixed integer k , the service area is partitioned arbitrarily into k areas of equal size. Service requests are distributed uniformly over the whole area. When batches of $\frac{n}{k}$ customer requests accumulate in a partition, these are deposited into a queue in a first-come first-served manner as in a $GI/G/m$ queue. In each partition, customer requests are served according to the optimal TSP tour across their locations.

We define the expected fraction of time the vehicle spends providing on-site service as follows: $\rho = \lambda \bar{s} / m$. The researchers conjectured that there exists a function $g(k, \rho)$ which determines n , ($n = g(k, \rho)$) such that as $\rho \rightarrow 1$ and $k \rightarrow \infty$, the $BvR(g(k, \rho), k)$ algorithm is asymptotically optimal. We refer to $BvR(g(k, \rho), k)$ as the optimal $BvR(n, k)$ algorithm and use $BvR(n, k)^*$ to denote it.

We consider a class of *General Partition Algorithms* which include the $BvR(n, k)$ algorithms as sub-class and develop a lower bound on the average wait time under this class of algorithms. We develop the lower bound for two different systems configurations. In the first, the size of the partitions depends upon the number of customers in the system. In the other, the partitions are fixed a priori. We refer to these as small partition and fixed partition cases. We obtain exactly the same bound for general partition algorithms applied to these systems. This lower bound matches the upper bound on the average waiting time provided by $BvR(n, k)^*$. Therefore, we show

that $BvR(n, k)^*$ is optimal among algorithms in this class. Finally, we identify an algorithm falling into this class whose asymptotic (heavy-traffic) performance is close to optimal. Therefore, we also prove the asymptotic optimality of $BvR(n, k)^*$.

5.2 Algorithms for Single Server Case

5.2.1 Definition and Notation

Bertsimas and van Ryzin showed that when ρ is less than one, there exists a function of $g(k, \rho)$ which determines n , ($n = g(k, \rho)$) such that such that the $BvR(g(k, \rho), k)$ algorithm can satisfy all service requests. This implies that N , the expected number of requests in queue is finite.

Assume for now that there is just a single server and that the service region is a unit square. We number the demands according to the order in which they are served. Let d_i be the distance traveled from $(i-1)^{th}$ demand to i^{th} demand. Let s_i be the on-site service time for demand i . The total service time includes the travel time $\frac{d_i}{v}$ and the on-site service time s_i . If, for all times t , the number of waiting requests in the system is bounded almost surely under a specific policy, we call this a stable policy. Using the definitions and notation presented by Bertsimas and van Ryzin (1991, 1993), for a stable policy, we let W_i denote the waiting time for demand i . The waiting time is the time between the arrival of demand i and the arrival of the server at the location of demand i .

The steady state expected waiting time is defined as $\bar{W} = \lim_{i \rightarrow \infty} E[W_i]$, and the steady state expected value of d_i is defined as $\bar{d} = \lim_{i \rightarrow \infty} E[d_i]$. The steady state expected number of requests in the queue, is defined to be $N = \lambda \bar{W}$.

5.2.2 Policies of Interest

Let $W(x)$ be the expected waiting time for a randomly selected customer located at point x . And, let W be the average waiting time under the algorithm of interest. We only consider algorithms that satisfy the following condition: there exists $\bar{\omega}$ and $\underline{\omega}$, such that

$$0 < \underline{\omega} \leq \frac{W(x)}{W} \leq \bar{\omega} \quad (5.1)$$

This is a technical requirement for our proof. If $\bar{\omega}$ grows large and $\underline{\omega}$ grows small, constraint (5.1) becomes progressively less tight and the class of algorithms satisfying the constraint increases.

Another requirement is that when the server finishes the service for a customer, it will go to the next customer or go back to the depot.

This is also a technical requirement to ensure us to estimate the average waiting time over the customers in a region based on the number of customers in that region.

5.2.3 General Partition Algorithms

We now define a class of algorithms for the single vehicle DTRP which we refer to as *General Partition Algorithms*. These work as follows: using a grid, divide the area A into k partitions of equal size (Note that Bertsimas and van Ryzin used a sweep algorithm rather than a grid. Any method to develop equal partitions will do. We use a grid to facilitate the development of our proof). A general partition algorithm will, for each region, partition time into periods. Let $t_{i,j}$ denote the end of region i 's j^{th} period. $t_{i,0}$ is defined to be 0 for each region i . The server then serves the requests in a sequence of *visits*. In each visit, the server selects a region to serve based on the accumulated demand in each region. This is the only information considered in the sequencing decision. In the j^{th} visit to region i , the server will serve exactly those requests which arrive in region i in the interval from $t_{i,j-1}$ through $t_{i,j}$. The order in which the regions are visited and the way that time is partitioned for each region will depend on the particular partition algorithm applied. If the server travels across regions in which there are no waiting customers, en route to a region in which there are waiting customers, we say that the empty region has also been visited.

$BvR(n, k)$ is a general partition algorithm in which a period ends for a region when $\frac{n}{k}$ new requests have accumulated in that region. The regions are visited in first-come-first-serve order according to when each period ends.

Note that it is not required that the current period for a region be completed when the server arrives in that region. For example, the class of *Exhaustive Partition Algorithms* will visit a region, will serve all waiting customers, and will also serve those customers that arrive while service is being provided to waiting customers. Thus, the stopping criterion for a period is when there are no outstanding requests in the region.

5.2.4 Properties of General Partition Algorithms

We will now show how a general partition algorithm gives rise to a distribution function f which describes consecutive served requests are distributed over the area. Fix a general partition algorithm and positive integers M and i . We are interested in the i^{th} request through the $(i + M - 1)^{\text{th}}$ request. However, for convenience, we would like to focus on a sequence of requests which start and end at the boundaries of visits. Suppose that the i^{th} request is served in the middle of the l^{th} visit. Let r_b be the first request served in visit l . (We use b for 'begin'). Now suppose that the $(i + M - 1)^{\text{th}}$ request is served in the middle of the p^{th} visit. Let r_e be the last request served in visit p . (We use e for 'end'). We will focus on requests r_b, \dots, r_e .

Suppose that requests r_b, \dots, r_e comprise q consecutive visits. Suppose that r distinct regions are visited during these q visits. We call these R_1, \dots, R_r . Fix these q visits, let v_1, v_2, \dots, v_r be the first visit to region R_1, \dots, R_r and v'_1, v'_2, \dots, v'_r be the last visit to region R_1, \dots, R_r belonging to these q visits. Remembering that each visit consists of an arrival

period and the each visit only clear the customers arrived during that period. For each $j \in \{1, \dots, r\}$, Let $t_{j,begin}$ be the beginning of v_j . Let $t_{j,end}$ be the end of v_j . So $t_{j,end} - t_{j,begin}$ is therefore the time interval associated with the Poisson arrival process for region R_j . Let Z_j denote the number of requests served in region R_j during the interval $t_{j,begin}$ to $t_{j,end}$. Z_j is exactly the number of requests that arrive in R_j during the interval $t_{j,begin}$ to $t_{j,end}$. We use $|R_j|$ to denote the area of partition R_j . Z_j is therefore a Poisson random variable with mean $\lambda |R_j| (t_{j,end} - t_{j,begin})$. Since the partitions are of equal size, $|R_j|$ is equal to $\frac{1}{k}$, where k is the number of partitions. Thus, Z_j is a Poisson random variable with mean $\frac{\lambda (t_{j,end} - t_{j,begin})}{k}$. Further, $\{Z_j, j = 1, \dots, r\}$ are independent Poisson random variables. The independence results from the fact that the regions are disjoint.

Next we generate a random permutation x_1, \dots, x_n of the set of consecutively served requests r_1, \dots, r_n where n is the number of requests. We observe the following: we have r independent Poisson random variables, each representing the number of customers in region R_j , with mean $\frac{\lambda (t_{j,end} - t_{j,begin})}{k}$. Further, the locations of these requests are uniformly distributed in region R_j . Let $I_{R_j}(x) = \begin{cases} 1 & x \in R_j \\ 0 & \text{otherwise} \end{cases}$ define an index function over R_j . The locations of the served requests, x_1, \dots, x_n , are a realization of i.i.d. random variables with probability density function of the form

$$f(x) = \frac{(t_{j,end} - t_{j,begin}) I_{R_j}(x)}{\sum_{j=1}^{j=r} (t_{j,end} - t_{j,begin}) |R_j|}. \quad (5.2)$$

5.2.5 The Small Partition Case

Let N represent the expected number of customers in queue. As before, let R be the size of the partitions. For any fixed $\bar{\theta}_\varepsilon$ and $\underline{\theta}_\varepsilon$, we consider only those general partition algorithms under which $\bar{\theta}_\varepsilon \geq NR \geq \underline{\theta}_\varepsilon$. For any $\varepsilon > 0$, let $M = N\varepsilon$ and assume that M is an integer. For a fixed general partition algorithm Γ , when the system is in steady-state, for any randomly selected request i , we are interested in requests from request i through the next M consecutively served requests. We expand the sequence of requests as described in the previous sections so that this sequence begins and ends at the boundaries of visits. Let Q denote the random variable that indicates the number of requests in this expanded sequence. Note that if the area of each partition is proportional to $\frac{1}{(1-\rho)^2}$, this ensures that the probability that the number of request in a randomly selected region is more than $N\varepsilon$ is negligible. With very high probability, $N\varepsilon \leq Q \leq 3N\varepsilon$, as ρ approaches one. Let x_1, x_2, \dots, x_Q represent the locations of Q consecutively served customers. Let the index represent the order in which service is performed. Now let y_1, y_2, \dots, y_Q be a random permutation of x_1, x_2, \dots, x_M . Therefore, y_1, y_2, \dots, y_M are distributed independently according to a distribution $f_\Gamma(x)$ where $f_\Gamma(x)$ is a piece-wise (discontinuous) function. Let $W(x)$ be the expected waiting time for a random selected

customer i that is located at point x . We present three propositions about $f_\Gamma(x)$ and $W(x)$.

Let W be the average waiting time for fixed algorithm Γ , we show that that the expected waiting time for any M demands can be bounded by W . This leads to a constraint (5.5) on $f_\Gamma(x)$, the distribution of customer locations.

Proposition 5.1 If $\bar{\theta}_i \geq NR \geq \underline{\theta}_i$, the expected waiting time for a customer located in x

$$\text{is } W(x). \text{ When } \rho \text{ is big enough } W(x) \geq \frac{Qf_\Gamma(x)}{4\lambda} - \frac{1}{2\lambda R_i}. \quad (5.3)$$

Proposition 5.2 If we focus on policies under which $0 < \underline{\omega} \leq \frac{W(x)}{W} \leq \bar{\omega}$, when ρ is

$$\text{large enough, then } f_\Gamma(x) \leq \frac{4\bar{\omega}}{\varepsilon} + \frac{2}{\varepsilon \underline{\theta}_i}. \quad (5.4)$$

Proposition 5.3 For any $\varepsilon_1 > 0$, when time is sufficiently large, $f_\Gamma(x)$ satisfies the following

$$\text{constraint } \int_A f_\Gamma^2(x) dx \leq \frac{2N(1+\varepsilon_1)}{Q\omega} \quad (5.5)$$

$$\text{where } \omega = \frac{1}{1 - 2\lambda R_i (\bar{s} + \sqrt{2}/v)}.$$

To obtain our lower bound for W , we analyze the average distance between consecutive demands served. First, we obtain a lower bound for the distance traveled to serve all of these selected M demands expressed in $f_{\Gamma}(x)$ which leads a lower bound on the average distance traveled per customer served. This is shown in lemma 5.1. We obtain lemma 5.1 by generalizing the classical TSP result of Beardwood et al (1959) and applying a smoothing technique to the distribution function $f_{\Gamma}(x)$. Next, minimizing the lower bound obtained under constraint (55) leads to our lower bound for the average distance traveled to serve each customer (5.6.a). We use lemma 3 to obtain the lower bound (5.6.a). We then provide a lower bound on the average waiting time for service (5.6.b) based on (5.6.a). For the proof of theorem 51 and the related lemmas, please see section 5.5.

Lemma 5.1 Let \bar{d} be the expected average distance traveled per demand served and let N be the average number of customers awaiting service, R be the size of the partition and Q represent the number of consecutively served customers. We consider algorithms for which for any $\varepsilon > 0$, we fix $\bar{\theta}_{\varepsilon}$ and $\underline{\theta}_{\varepsilon}$. For any general partition algorithms under which $\bar{\theta}_{\varepsilon} \geq NR \geq \underline{\theta}_{\varepsilon}$, for any $\varepsilon_1 > 0 \exists \rho_0$, such that when $\rho > \rho_0$, $\sqrt{Qd} \geq (\beta - \varepsilon_1) \int_A \sqrt{f_{\Gamma}(x)} dx$, where β is the TSP constant defined by Beardwood et al (1959).

Lemma 5.2 Letting $Z = \min \left\{ \sum_{i=1}^n \sqrt{x_i A_i} \right\}$ subject to $\sum_{i=1}^n x_i^2 A_i \leq 1 + \varepsilon$ and $\sum_{i=1}^n x_i A_i = 1$,

$$Z \geq \frac{1}{\sqrt{1 + \varepsilon}}.$$

Theorem 5.1 Let \bar{d} be the expected distance traveled per demand served, N be the expected number of customers awaiting service and R be the size of each partition. For any fixed $\bar{\theta}_\varepsilon$ and $\underline{\theta}_\varepsilon$, for any general partition algorithm under which $\bar{\theta}_\varepsilon \geq NR \geq \underline{\theta}_\varepsilon$, the following result holds:

$$\lim_{\rho \rightarrow 1} \sqrt{2N} \bar{d} \geq \beta \tag{5.6.a}$$

$$\lim_{\rho \rightarrow 1} \{(1 - \rho)^2 W\} \geq \frac{\lambda \beta^2 A}{2v^2} \tag{5.6.b}$$

Note that when $\bar{\theta}_\varepsilon$ is larger and $\underline{\theta}_\varepsilon$ is smaller, the class of general partition algorithms considered becomes broader. At some point, all general partition algorithms obey the required conditions.

Finally, as $\rho \rightarrow 1$ we find that if we want to minimize the average waiting time, under the optimal algorithm among the class of algorithms considered, the distribution function $f(x)$ that describes the spatial distribution of M consecutively served customers will be almost uniform in a small area and zero over all other areas. In fact, the lower bounds (equations 5.6.a and 5.6.b) hold for a class of algorithms that satisfying the following condition.

For any $\varepsilon > 0$, we consider $M = N\varepsilon$ consecutively served customers and let y_1, y_2, \dots, y_M be a random permutation of the locations. Consider algorithms under which y_1, y_2, \dots, y_M are distributed independently according to a distribution $f(x)$, where $f(x)$ is required to satisfy the following α -Lipschitz condition:

There exists α such that $|f(x) - f(y)| \leq \alpha \|x - y\|$, where $\|x - y\|$ is the distance between x and y . Note that α does not depend on ρ .

5.2.6 The Fixed Partition Case

Up to this point, we have examined algorithms for which the number and size of the partitions depends upon ρ . Now we consider algorithms for which the partitions are pre-determined. We state the result for m servers case and the proof of theorem 5.2 is also for m server case.

First use a grid to divide the unit square into fixed small partitions. Let Δ be the area of each partition. When the server selects a partition to serve it will either finish all customers waiting in the queue at the selected partition or finish all the customers arriving during some arbitrarily selected period. We show that as Δ approaches zero, the

average waiting time for service will be bounded from below by $\frac{\lambda\beta^2 A}{2m^2v^2(1-\rho)^2}$, where m

is the number of servers.

Theorem 5.2 For any algorithm falling into the category mentioned above, when Δ is small enough, $\lim_{\rho \rightarrow 1} W(1-\rho)^2 > \frac{\lambda\beta^2}{2v^2m^2}$. Note that this exactly matches equation (5.6.b), the bound for the small partition case.

5.2.7 An Asymptotically Optimal General Partition Algorithm

In this section we show the asymptotic optimality of a specific general partition algorithm. Let P_j be the partition which divides the area into j^*j squares and OPT denote the optimal algorithm. For a specific arrival sequence, let $\sigma = r_1, r_2, \dots$ denote the order in which the requests are satisfied by OPT . Given a partition P of the area, we will devise a General Partition Algorithm called A_p based on the behavior of OPT . We take the sequence σ and remove some of the requests to obtain another sequence σ' as follows: Examine each r_i in turn. Suppose that request r_i is located in region R_i . Suppose that at the time that r_i is reached by OPT 's server, there are other outstanding requests in R_i . Remove these additional outstanding requests from σ and continue. The requests that were removed from σ will be called *extra* requests. We denote the sequence $\sigma' = r_{i_1}, r_{i_2}, \dots$. The algorithm A_p will work as follows: for each request r_{i_j} in σ' , visit region R_{i_j} and satisfy r_{i_j} and any extra requests which are waiting in R_{i_j} at the time OPT serves r_{i_j} . In other words, the periods are chosen so that when OPT serves a request r_{i_j} from σ' , the current period for R_{i_j} ends.

Proposition 5.4 (Bertsimas and van Ryzin, 1991) $\bar{d}_\lambda^* \geq \frac{\gamma\sqrt{A}}{\sqrt{N+m/2}}$ where γ is a

constant, $\gamma \geq \frac{2}{3\sqrt{2\pi}}$, m is the number of servers.

Proposition 5.5 For a fixed arrival rate λ , let \bar{d}_λ^* denote the average distance traveled per customer for the optimal algorithm. For any ε , for the partition P_j , when j is large enough, the average distance traveled per customer served for the algorithm defined above is at most $\bar{d}_\lambda^* + \frac{\varepsilon}{\sqrt{N}}$, which implies the lower bound in theorem 1 applies to the optimal algorithm.

If we let $\bar{d}_{\lambda,\rho}$ be the average distance traveled per customer served, combining propositions 5.4 and 5.5, we show that $\bar{d}_{\lambda,\rho} = \bar{d}_\lambda^* + o(\bar{d}_\lambda^*)$.

5.2.8 The Optimality of $BvR(n,k)$ Among the General Partition Class

Lemma 5.3 (Bertsimas and van Ryzin) $W^* \leq \frac{\lambda\beta^2 A}{2m^2v^2(1-\rho)^2} + O\left(\frac{1}{(1-\rho)^{3/2}}\right)$, where m is

the number of servers, $\rho = \frac{\lambda\bar{s}}{m}$.

Lemma 5.3 provides an upper bound for the average waiting time under the $BvR(n, k)$ algorithm. Theorem 5.1 and proposition 5.5 provides a lower bound for the optimal algorithm. As $\rho \rightarrow 1$ these bounds converge. Therefore, we show that under high traffic intensity, $BvR(n, k)$ is asymptotically optimal among all algorithms.

5.3 The m-Server Case

Now we assume that instead of a single server that there are m mobile servers in the Euclidean service region. The definition is for the W, W, d, d and N is similar as one server case. If we only focus on the customers consecutively served by the same server, we have the following theorem.

Theorem 5.3 Let \bar{d} be the expected distance traveled per demand served, N be the expected number of customers awaiting service and R be the size of each partition. For any fixed $\bar{\theta}_\epsilon$ and $\underline{\theta}_\epsilon$, for any general partition algorithm under which $\bar{\theta}_\epsilon \geq NR \geq \underline{\theta}_\epsilon$, the following result holds:

$$\lim_{\rho \rightarrow 1} \sqrt{2N} \bar{d} \geq \beta$$

$$\lim_{\rho \rightarrow 1} \{(1 - \rho)^2 W\} \geq \frac{\lambda \beta^2 A}{2m^2 v^2}$$

Now we define the following algorithm, which we refer to as $BvR(n, k) - m$,

Divide the area into m sub-regions of equal size and assign each server to a single sub-region, let each server works independently in their own assigned area according to the $BvR(n, k)$.

In the similar way as what we have done for one server case, we can show that $BvR(n, k) - m$ is asymptotic optimal. Comparing the single server and multiple server cases we find the following interesting result. If we the algorithm that is asymptotically optimal for the single server case is also asymptotically optimal for m server case.

5.4 Proof of the Theorems

5.4.1 Proof of the Propositions and Lemmas

First, in section 5.4.1.1, we introduce a smoothing technique to prove lemma 5.1. In section 5.4.1.2, we introduce the lemmas we need to prove lemma 5.1. We use lemma 5.1.1 and 5.1.2 to proof lemma 5.1. Using lemma 5.1.3 to proof lemma 5.1.1. In section 5.4.1.3, we provide the proof for the propositions and in the last section, we provide proof for the lemmas. To prove lemma 5.1.1 and lemma 5.1.3, we borrow the method from the classical paper by Beardwood et al (1959). To prove lemma 5.2, we rely on optimization methods and algebraic techniques.

5.4.1.1 The Smoothing Technique

To prove lemma 5.1, we require a smoothed version of f . A variety of smoothing techniques will work. We choose one to make our discussion precise.

We select a parameter η , for each area A_i on which $f(x)$ is not zero, we define a new area \tilde{A}_i which contains in the original one with the same center, the ratio of the parameter of new area to the original one is $1 - 2\eta$.

First we define a $g(x)$ based on $f(x)$ and η : $g(x) = \frac{f(x)}{\eta} [\eta - d(x, \tilde{A}_i)]$ when $x \in A_i$.

We define $f_\eta(x)$ to be a smoothed version of $f(x)$ as following: $f_\eta(x) = \frac{g(x)}{\int_A g(x) dx}$.

Note that as η gets small, f_η approaches f in the limit. Furthermore, for any fixed value of η , we can find a constant α such that for any two points x and y in the area A ,

$$|f_\eta(x) - f_\eta(y)| \leq \alpha \|x - y\|,$$

where $\|x - y\|$ is the Euclidean distance from x to y and $\alpha = \frac{\max_x \{f(x)\}}{\eta}$.

Note that $f_\eta(x)$ is bounded and so after applying this smoothing technique, the smoothed version of $f_\eta(x)$ will satisfy the α -Lipschitz condition.

5.4.1.2 Lemmas and Theorem BHM

Lemma 5.1.1 Let $y_{1,\mu}, y_{2,\mu}, \dots, y_{n,\mu}$ be the i.i.d. random variables with distribution

$f_\mu \in \Sigma_\alpha$, where $\Sigma_\alpha = \{f \mid |f(x) - f(y)| \leq \alpha \|x - y\|\}$. Let $L_{TSP}(y_{1,\mu}, y_{2,\mu}, \dots, y_{n,\mu})$ be the

length of optimal TSP tour over $y_{1,\mu}, y_{2,\mu}, \dots, y_{n,\mu}$, for any $\varepsilon > 0$, we can find N_0 , when

$n > N_0$, we have for any μ , $E \frac{L_{TSP}(y_{1,\mu}, y_{2,\mu}, \dots, y_{n,\mu})}{\sqrt{n}} \geq \beta \int_A \sqrt{f_\mu(x)} dx - \varepsilon$, where β is

the TSP constant defined in Beardwood *et al.* (1959).

Lemma 5.1.2 Assume y_1, \dots, y_M are i.i.d. random variables with common distribution f

and z_1, \dots, z_M are i.i.d. random variables with distribution f_η . Let $L_{TSP}(y_1, \dots, y_M)$ and

$L_{TSP}(z_1, \dots, z_M)$ be the length of the optimal TSP tour over y_1, \dots, y_M and z_1, \dots, z_M

respectively. For any ε_1 , when η is sufficient small and M is sufficient large, we have,

$$\frac{|E[L_{TSP}(y_1, \dots, y_M) - L_{TSP}(z_1, \dots, z_M)]|}{\sqrt{M}} < \varepsilon_1.$$

Lemma 5.1.3 Let $\{B_i\}_{i=1}^{i=n}$ be a grid partition over A such that each B_i has the same area.

Let $f(x)$ be constant within B_i and X_1, X_2, \dots, X_n be i.i.d. random variables distributed

according to $f(x)$. For any $\varepsilon_3 > 0$, we can find $N_1(\varepsilon_3)$, when $n > N_1(\varepsilon_3)$, we have

$$E \left[\frac{L_{TSP}(X_1, X_2, \dots, X_n)}{\sqrt{n}} \right] \geq \beta \int_A \sqrt{f(X)} dX - \varepsilon_3 \text{ holds for any } f(x).$$

Theorem BHM: Assume that X_1, X_2, \dots, X_n are i.i.d. random variables with distribution $f(x)$ and let $L_{TSP}(X_1, X_2, \dots, X_n)$ be the length of TSP tour over X_1, X_2, \dots, X_n .

$$\lim_{n \rightarrow \infty} \left\{ \frac{L_{TSP}(X_1, X_2, \dots, X_n)}{\sqrt{n}} \right\} = \beta \int_A \sqrt{f(X)} dX \quad \text{a.s.} \quad (5.7)$$

5.4.1.3 Proof of Propositions

Proof of proposition 5.1

We have Q customers, the locations of these customers are i.i.d. random variables with distribution $f(x)$. We fix a region first, let it be R_i , Assume $f(x) = c$, when $x \in R_i$.

Let n_i be the number of customers located at R_i . It is easy to know that the distribution of n_i is Binomial with parameter c, R_i . Let $W(x)$ be the expected waiting time for a random selected customer located at x .

Assume from the moment that the server enters a region and begins to provide the service and keeps working until there is no other customer in the current region, there are r customers served totally. We estimate the average waiting time in the following way: when the server finishes the last one, there are r customers in the region, the average waiting time is bigger than $\frac{1}{2}$ of the length of the total sum of $(r-1)$ interarrival time minus the length of the total stay period, which is less than the sum of r on site service time and travel time (which is less than $\frac{\sqrt{2}}{v}$) to provide service.

We have,

$$W(x) = E[E[W_i | n_i = r]] \geq E\left[\frac{n_i - 1}{2\lambda R_i} - n_i \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right)\right] = \frac{Qf(x)R_i - 1}{2\lambda R_i} - Qf(x)R_i \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right) =$$

$$Qf(x) \left(\frac{1}{2\lambda} - R_i \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right)\right) - \frac{1}{2\lambda R_i}.$$

Note that when $\rho \rightarrow 1$, $N \rightarrow \infty$ implies $R_i \rightarrow 0$ as $\rho \rightarrow 1$.

For any NR such that $\frac{1}{4\lambda} > R_i \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right)$ and $NR \geq \underline{\theta} > 0$, we know, when ρ is big

enough, we have $W(x) \geq \frac{Qf(x)}{4\lambda} - \frac{1}{2\lambda R_i}$.

Proof of proposition 5.2

From proposition 1, we know when ρ is big enough,

$$W(x) \geq \frac{Qf(x)}{4\lambda} - \frac{1}{2\lambda R_i} \Rightarrow \frac{W(x)}{W} \geq \frac{Qf(x)}{4N} - \frac{1}{2NR_i} \Rightarrow f_r(x) \leq \frac{4NW(x)}{WQ} + \frac{4N}{2NRQ}.$$

If $\bar{\theta}_s \geq NR \geq \underline{\theta}_s$ and $0 < \underline{\omega} \leq \frac{W(x)}{W} \leq \bar{\omega}$, we know $f_r(x) \leq \frac{4\bar{\omega}}{\varepsilon} + \frac{2}{\varepsilon \underline{\theta}_s}$.

Proof of proposition 5.3

First we try to give a lower bound on the expected total waiting time.

Let Z_j denote the number of requests in R_i . Let T_j be the total waiting time for the customers in R_i . Let n_j be the total number of customers served in R_i .

Using the fact that the distribution of the Q requests is $f(x)$, the distribution of these random variables of Z_j 's is given by the following multinomial distribution:

$$P\{Z_1 = z_1, \dots, Z_r = z_r\} = \frac{Q!}{\prod_j (z_j)} \prod_j (c_j | R_i)^{z_j}.$$

$$\begin{aligned} E[T_j] &= E[E[T_j | n_j]] \geq E\left[\frac{n_j(n_j-1)}{2\lambda R_i} - n_j(n_j-1)\left(\bar{s} + \frac{\sqrt{2}}{\nu}\right)\right] = \frac{Q(Q-1)c_j^2 R_i}{2\lambda} \\ &\quad - Q(Q-1)c_j^2 R_i^2 \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right). \end{aligned}$$

$$\text{We have } \frac{\sum_j E[T_j]}{Q} = \frac{(Q-1)}{2\lambda} \int_A f^2 dx \left[1 - 2\lambda R_i \left(\bar{s} + \frac{\sqrt{2}}{\nu}\right)\right].$$

Because $E[W_i] \rightarrow W \Rightarrow E\left[\sum_{j=i}^{i+Q-1} W_j\right] \rightarrow W$ as $i \rightarrow +\infty$. So for any $\varepsilon_1 > 0$, there exists

$$T > 0, \text{ when } t > T, \text{ we have } E\left[\frac{\sum_{j=i}^{i+Q-1} W_j}{W}\right] \leq 1 + \varepsilon_1.$$

$$\text{So when } t > T, \text{ we have the } \int_A f_r^2(x) dx \leq \frac{2N(1+\varepsilon_1)}{Q\omega}.$$

Proof of Proposition 5.5

The maximum diameter of a region under a given partition P_j is $\frac{\sqrt{2}}{\sqrt{j}}$, then each extra

request introduces a distance of at most $\frac{2\sqrt{2}}{\sqrt{j}}$ to A_p . When $j \gg N$, we know

$\frac{2\sqrt{2}}{\sqrt{j}} \ll \frac{1}{\sqrt{N}}$. Thus, as the partitions become more fine, the additional distance

introduced by an extra request decreases. Furthermore, as the partitions become more

fine, the probability that any given request is an extra request also becomes smaller.

Using these two facts, we conclude that for any arrival rate λ , when j big enough, the average distance traveled per customer served for the algorithm defined above is at most

$$\bar{d}_\lambda + \frac{\varepsilon}{\sqrt{N}}.$$

5.4.1.4 Proof of Lemmas

Proof of lemma 5.1.1

Observation 5.1. There exists $C_1 > 0$, such that $\sup_{f(x) \in \Sigma_\alpha} \int_A \sqrt{f(x)} dx \leq C_1$.

For any $f(x) \in \Sigma_\alpha$, let $f(y_0) = \min_{x \in A} \{f(x)\}$.

Because $\int_A f(x) dx = 1$, we know: $f(y_0) D(A) \leq 1 \Rightarrow f(y_0) \leq \frac{1}{D(A)} \Rightarrow$

$\max_{x \in A} \{f(x)\} \leq f(y_0) + \alpha D(A) \leq \frac{1}{D(A)} + \alpha D(A)$, Where $D(A)$ is the diameter of A .

Let $\|A\|$ be the area of A and $C_1 = \|A\| \sqrt{\frac{1}{D(A)} + \alpha D(A)}$, we have $\sup_{f(x) \in \Sigma_\alpha} \int_A \sqrt{f(x)} dx \leq C_1$.

Observation 5.2. For any given $1 > \varepsilon_3 > 0$, let $\delta = \frac{\varepsilon_3^2}{\alpha}$, for any for any x and y satisfying $\|x - y\| \leq \delta$, we have, $|\sqrt{f(x)} - \sqrt{f(y)}| \leq \varepsilon_3$.

The reason for it is as following: for any x, y satisfying $\|x - y\| \leq \delta$, assume $f(y) = \tau$ and $f(x) = \tau + \xi$ where $\tau \geq 0$, $\xi \geq 0$, note that $\sqrt{\xi} \leq \varepsilon_3$, we have,

$$|\sqrt{f(x)} - \sqrt{f(y)}| = \left| \frac{\xi}{\sqrt{\tau + \xi} + \sqrt{\tau}} \right| \leq \max_{\tau \geq 0} \left\{ \frac{\xi}{\sqrt{\tau + \xi} + \sqrt{\tau}} \right\} = \sqrt{\xi} \leq \varepsilon_3.$$

We divide A into grid partitions of identical size with diameter δ . As before, let B_j be the j^{th} partition in A .

$$\text{Let } f_\delta(x) = \min_{\{y, y \in B_j\}} \{f(y)\}, x \in B_j.$$

From observation 5.2, we have observation 5.3.

$$\text{Observation 5.3} \quad \int_A f_\delta(x) dx \geq 1 - \varepsilon_3 \quad \text{and} \quad \int_A \sqrt{f_\delta(x)} dx \geq \int_A \sqrt{f(x)} dx - \varepsilon_3.$$

We place Y_i into one of the two sets (Ω_1, Ω_2) as follows:

If $f_\delta(Y_i) = 0$, let $Y_i \in \Omega_1$ with probability one; otherwise, let $Y_i \in \Omega_1$ with probability

$$1 - \frac{f_\delta(Y_i)}{f(Y_i)} \text{ and } Y_i \in \Omega_2 \text{ with probability } \frac{f_\delta(Y_i)}{f(Y_i)}.$$

Let n_2 be the number of requests belonging to Ω_2 and $L_{TSP}(\Omega_2)$ be the length of optimal TSP tour over all the nodes belonging to Ω_2 .

We can show that

$$(I) \frac{n_2}{n} \rightarrow \int_A f_\delta(x) dx \text{ a.s. as } n \rightarrow \infty.$$

(II) The random variables in the set of Ω_2 are i.i.d. random variables with the probability

$$\text{density function of } \frac{f_\delta(x)}{\int_A f_\delta(x) dx}.$$

Using lemma 5.1.3 and (II), we know that for any $\varepsilon_4 > 0$, there exists $N_4(\varepsilon_4)$, such that

when $n_2 > N_4(\varepsilon_4)$, the following holds,

$$E \left[\frac{L_{TSP}(\Omega_2)}{\sqrt{n}} \right] \geq \beta \sqrt{\frac{n_2}{n}} \int_A \sqrt{\frac{f_\delta(x)}{\int_A f_\delta(x) dx}} dx - \varepsilon_4$$

From (I), we know there exists $n_0 > 0$, when $n > n_0$, we have

$$P \left(\frac{n_2}{n \int_A f_\delta(x) dx} > 1 - \varepsilon_3 \right) > 1 - \varepsilon_3.$$

When $n > \max \left\{ n_0, \frac{N_4(\varepsilon_4)}{1-2\varepsilon_3} \right\}$, we have

$$\mathbb{E} \left[\frac{L_{TSP}(\Omega_2)}{\sqrt{n}} \right] \geq \beta(1-\varepsilon_3) \int_A \sqrt{(1-\varepsilon_3)f_\delta(x)} dx - \varepsilon_4 \geq \beta \int_A \sqrt{f_\delta(x)} dx - 2\varepsilon_3 \int_A \sqrt{f_\delta(x)} dx - \varepsilon_4$$

From observation 5.1, observation 5.3, we know

$$\mathbb{E} \left[\frac{L_{TSP}(\Omega_2)}{\sqrt{n}} \right] \geq \beta \int_A \sqrt{f(x)} dx - \beta\varepsilon_3 - 2\varepsilon_3 C_1 - \varepsilon_4.$$

Let $\varepsilon_5 = \varepsilon_3 - 2\varepsilon_3 C_1 - \varepsilon_4$, because C_1 and β are constant and $\varepsilon_3, \varepsilon_4$ are arbitrary small number, ε_5 is arbitrary. At last, we have

$$\mathbb{E} \left[\frac{L_{TSP}(\Omega)}{\sqrt{n}} \right] \geq \mathbb{E} \left[\frac{L_{TSP}(\Omega_2)}{\sqrt{n}} \right] \geq \beta \int_A \sqrt{f(x)} dx - \varepsilon_5.$$

We finish the proof of lemma 5.1.1.

Proof of lemma 5.1.2

Step 1. We place Y_j into one of the two sets (Ω_1, Ω_2) as follows:

If $f_\eta(Y_j) \leq f(Y_j)$, let $Y_j \in \Omega_2$ with probability $\frac{f_\eta(Y_j)}{f(Y_j)}$ and $Y_j \in \Omega_1$ with probability

$$1 - \frac{f_\eta(Y_j)}{f(Y_j)}.$$

Step 2. Let $\tau = \int_{f(x) > f_\eta(x)} (f(x) - f_\eta(x)) dx$. Now we dispatch the elements in the set Ω_1 according to the $\frac{f_\eta(x) - f(x)}{\tau}$.

After these two steps, from y_1, \dots, y_M , we get z_1, \dots, z_M , where z_1, \dots, z_M are i.i.d. random variables with distribution f_η .

Now we examine the number of elements in set Ω_2 . Let n_2 be the number of requests belonging to Ω_2 and $L_{TSP}(\Omega_2)$ be the length of optimal TSP tour over all the nodes belonging to Ω_2 .

To calculate the difference between $E[L_{TSP}(Z_1, Z_2, \dots, Z_M)]$ and $E[L_{TSP}(Y_1, Y_2, \dots, Y_M)]$, we note the following fact,

$$(I) L_{TSP}(Z_1, Z_2, \dots, Z_M) \geq L_{TSP}(\Omega_2) \text{ and } L_{TSP}(Y_1, Y_2, \dots, Y_M) \geq L_{TSP}(\Omega_2).$$

(II) From lemma 2 in Karp and Steel (1985), we know there exists a tour whose length is less than $2\sqrt{(M - n_2)} + 2 + \sqrt{2}$ over $M - n_2$ nodes.

Combing these two facts together, we know

$$\begin{aligned} |E[L_{TSP}(Z_1, Z_2, \dots, Z_M)] - E[L_{TSP}(Y_1, Y_2, \dots, Y_M)]| &\leq E[2\sqrt{(M - n_2)} + 2 + \sqrt{2}] \\ &\leq 2\sqrt{E[M - n_2]} + 2 + \sqrt{2}. \end{aligned}$$

It is easy to see that when η goes to zero, $E[M - n_2]$ goes to zero too.

So for any ε_1 , when η is sufficient small, M is big enough, we have

$$\frac{|E[L_{TSP}(y_1, \dots, y_M) - L_{TSP}(z_1, \dots, z_M)]|}{\sqrt{M}} < \varepsilon_1.$$

Proof of lemma 5.1.3

We prove lemma 5.1.3 by induction based on the different values $f(x)$ can hold.

First we assume that $f(x)$ is equal to zero or any constant over all the B_i i.e.

$$f(x) = \sum_i c_i I_{B_i}(x), \quad I_{B_i}(x) = \begin{cases} 1 & \text{if } x \in B_i \\ 0 & \text{otherwise} \end{cases} \text{ and } c_i \text{ equals to zero or } c.$$

Let m be the number of the partitions on which $f(x)$ is c . The probability that any demand falls into any specific partition on which $f(x) \neq 0$ is $\frac{1}{m}$.

Let L_{TSP_i} be the length of optimal TSP subtour over all the demands belonging to the i^{th} piece of the partition on which $f(x)$ is c . Let L_{TSP} be the optimal TSP tour over all the demands. Let x is the length of the circumference of the partition B_i , For the optimal TSP tour over all the nodes, we are interested in the points that lies in the optimal TSP tour and the perimeter of the partition. For these points, we construct a tour through all the points and select each. After these steps, for each partition, we have a connected graph which each node has even degree. We know there exists one Euler tour that

traverse all the links exactly once. Remembering that n_0 is the total number of partitions,

The length of this tour is less than $L_{TSP} + 3n_0x$ and is at least the same length of the

$$\sum_i L_{TSP_i}.$$

Finally, we obtain $L_{TSP} \geq \sum_i L_{TSP_i} - 3n_0x$.

Now we try to estimate $\frac{L_{TSP}(X_1, X_2, \dots, X_n)}{\sqrt{n}}$. Because $\frac{n_0x}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$, we focus on

$$\frac{\sum_i L_{TSP_i}}{\sqrt{n}}.$$

To express the idea clearly, from now on, we focus only on the partitions that $f(x)$ is c ,

assume these partitions are $B_i, i \in \{1, 2, \dots, m\}$.

Let n_i be the number of nodes in B_i , let $D_i = \left\{ n_i > \frac{n}{m} - k_1 \sqrt{\frac{n}{m}} \right\}$.

Using Chebychev's Inequality, $P\{D_i\} \geq 1 - \frac{1}{k_1^2}$ (5.8)

By De Morgan's rule, $P\{\cap_i D_i\} = 1 - P\{\cup_i D_i^c\} \geq 1 - \sum_i [1 - P\{D_i\}]$

Finally, $P\{\cap_i D_i\} \geq 1 - \sum_i [1 - P\{D_i\}] \geq 1 - \frac{m}{k_1^2} \rightarrow 1$ as $k_1 = \left(\frac{n}{m}\right)^{\frac{1}{4}} \rightarrow \infty$ (5.9)

Let $|B|$ be the size of the area of any partition and $|A|$ be the size of the whole area, we use the result of Beardwood et al. (1959), please refer the (5.7) and combining with (5.8), (5.9), for any $\varepsilon_2 > 0$, there exists $n(\varepsilon_2) > 0$, when $n_i > n(\varepsilon_2)$, we have

$$P \left\{ \frac{L_{TSPi}}{\sqrt{n_i}} > (\beta - \varepsilon_2) \sqrt{|B|} \right\} > 1 - \varepsilon_2. \quad (5.10)$$

$$P \left\{ \frac{L_{TSPi}}{\sqrt{n}} > (\beta - \varepsilon_2) \sqrt{|B| \frac{n_i}{n}} \right\} > 1 - \varepsilon_2 \quad (5.11)$$

After some calculation, we know when $n > \max \left\{ 16m, m \left[n(\varepsilon_2) \right]^{4/3} \right\}$, we have

$$n_i > \frac{n}{m} - k_1 \sqrt{\frac{n}{m}} > n(\varepsilon_2). \text{ So when } n > \max \left\{ 16m, m \left[n(\varepsilon_2) \right]^{4/3} \right\}, \text{ with at least } P\{\cap_i D_i\}$$

probability that all n_i satisfies $n_i > n(\varepsilon_2)$.

$$\text{Because when } n_i > n(\varepsilon_2), \text{ we have } \frac{n_i}{n} > \frac{1}{m} - k_1 \sqrt{\frac{1}{nm}}. \quad (5.11)$$

From (5.11), (5.12) and $P(AB) \geq P(A) + P(B) - 1$, we know,

$$P \left\{ \sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} > (\beta - \varepsilon_2) \sum_{i=1}^{i=m} \sqrt{|B| \frac{n_i}{n}} \right\} > 1 - m\varepsilon_2 - 1 - \frac{m}{k_1^2}.$$

$$P \left\{ \sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} > (\beta - \varepsilon_2) m \sqrt{|B| \left(\frac{1}{m} - k_1 \sqrt{\frac{1}{nm}} \right)} \right\} > 1 - m\varepsilon_2 - 1 - \frac{m}{k_1^2}$$

$$P \left\{ \sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} > (\beta - \varepsilon_2) \sqrt{m|B| \left(1 - k_1 \sqrt{\frac{m}{n}} \right)} \right\} > 1 - m\varepsilon_2 - 1 - \frac{m}{k_1^2}.$$

Because $k_1 = \left(\frac{n}{m} \right)^{1/4}$, so at last we have

$$P \left\{ \sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} > (\beta - \varepsilon_2) \sqrt{m|B| \left(1 - \left(\frac{m}{n} \right)^{1/4} \right)} \right\} > 1 - m\varepsilon_2 - 1 - \frac{m^{3/2}}{\sqrt{n}}.$$

$$\Rightarrow E \left[\sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} \right] \geq \left(1 - m\varepsilon_2 - 1 - \frac{m^{3/2}}{\sqrt{n}} \right) (\beta - \varepsilon_2) \sqrt{m|B| \left(1 - \left(\frac{m}{n} \right)^{1/4} \right)}.$$

Note that $\int_A \sqrt{f} dx = \sqrt{m|B|}$, we have

$$E \left[\sum_{i=1}^{i=m} \frac{L_{TSPi}}{\sqrt{n}} \right] \geq \left(1 - m\varepsilon_2 - 1 - \frac{m^{3/2}}{\sqrt{n}} \right) (\beta - \varepsilon_2) \left[\int_A \sqrt{f} dx \right] \sqrt{\left(1 - \left(\frac{m}{n} \right)^{1/4} \right)}.$$

Because $\frac{m}{n} \rightarrow 0$, $\frac{m^{3/2}}{\sqrt{n}} \rightarrow 0$ and $\frac{n_0 x}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$, $L_{TSP} \geq \sum_i L_{TSPi} - 3n_0 x$ and ε_2 is

arbitrary, so for any $\varepsilon_3 > 0$, we can find $N_1(\varepsilon_3)$, when $n > N_1(\varepsilon_3)$, we have

$$E \left[\frac{L_{TSP}(X_1, X_2, \dots, X_n)}{\sqrt{n}} \right] \geq \beta \int_A \sqrt{f(X)} dX - \varepsilon_3.$$

The above argument holds for any fixed m , when m goes from 1 to n_0 , we fix

$n^* = \max_m n^*(m)$, we know when $n > n^*$, we have

$$E \left[\frac{L_{TSP}(X_1, X_2, \dots, X_n)}{\sqrt{n}} \right] \geq \beta \int_A \sqrt{f(X)} dX - \varepsilon_3.$$

The last step in this induction proof is to assume the lemma is true when $f(x)$ has k different values and to show that the lemma holds when $f(x)$ have $k+1$ different values. The proof is based on similar ideas and is rather tedious, so we omit the rest of proof here.

Proof of lemma 5.1

We know that y_1, y_2, \dots, y_M are i.i.d. random variables with distribution of $f_\Gamma(x)$. After applying the smoothing technique, we know that the for any $f_\Gamma(x)$ that satisfies the condition of $f_\Gamma(x) \leq \frac{4\bar{\omega}}{\varepsilon} + \frac{2}{\varepsilon\theta_\varepsilon} + \varepsilon$, we can find common α such that $f_\eta(x)$, the smoothing version of $f(x)$ has the following two properties:

Property I: they satisfy the α Lipnizs condition;

Property II: Let z_1, z_2, \dots, z_Q be i.i.d. r.v.s with distribution of $f_\eta(x)$. From lemma 5.1.2, we know for any $\varepsilon_2 > 0$, $\exists \rho_1(\varepsilon_2) > 0$, when $\rho > \rho_1(\varepsilon_2)$, we have

$$\frac{EL_{TSP}(y_1, y_2, \dots, y_Q)}{\sqrt{Q}} \geq \frac{EL_{TSP}(z_1, z_2, \dots, z_Q)}{\sqrt{Q}} - \varepsilon_2.$$

From lemma 5.1.1 and property I, we know for any $\varepsilon_3 > 0$, $\exists \rho_2$, when $\rho > \max\{\rho_1, \rho_2\}$,

$$\frac{EL_{TSP}(z_1, z_2, \dots, z_Q)}{\sqrt{Q}} \geq \int_A \sqrt{f_\eta(x)} dx - \varepsilon_3$$

Because $\int_A \sqrt{f_\eta(x)} dx \geq (1 - 4\eta + 4\eta^2) \int_A \sqrt{f_\Gamma(x)} dx$, and $\varepsilon_2, \varepsilon_3$ are arbitrary,

$$\sup_{f_{\Gamma}(x)} \int_A \sqrt{f_{\Gamma}(x)} dx \leq |A| \sqrt{\frac{4\overline{\omega}}{\varepsilon} + \frac{2}{\varepsilon\theta_{\varepsilon}}}.$$

Using an argument similar to the one used in the proof of lemma 5.1.1, we can show that

$$\text{for any } \varepsilon_1 > 0, \exists \rho_0, \text{ such that when } \rho > \rho_0, \sqrt{Qd} \geq (\beta - \varepsilon_1) \int_A \sqrt{f_{\Gamma}(x)} dx.$$

Proof of lemma 5.2

By adding some additional unbounded constraints we can show that the original problem

can be translated into the following problem: $\min \left\{ \sum_{i=1}^{i=n} \sqrt{x_i A_i^*} \right\}$ subject to: $\sum_{i=1}^{i=n} x_i A_i^* = 1,$

$$\sum_{i=1}^{i=n} x_i^2 A_i^* \leq 1 + \varepsilon, \quad x_i \geq \varepsilon_1, \quad \forall i \quad \text{for some small number of } \varepsilon_1. \quad \text{The reason that the}$$

constraints $\{x_i \geq \varepsilon_1\}_{i=1}^{i=n}$ do not affect the solution of the problem when ε_1 is small enough

because these constraints are not bounded.

For this new problem, we use a standard optimization technique as follows:

$$\text{Let } L(\overline{X}, \lambda, \mu, \overline{\gamma}) = \sum_{i=1}^{i=n} \sqrt{x_i A_i^*} + \left(\lambda \sum_{i=1}^{i=n} x_i A_i^* - 1 \right) + \mu \left(1 + \varepsilon - \sum_{i=1}^{i=n} x_i^2 A_i^* \right) + \sum_{i=1}^{i=n} \gamma_i (x_i - \varepsilon_1).$$

Considering the Kuhn-Tucker conditions we know that the optimal solution must satisfy

the following: $\frac{\partial L}{\partial x_i} = 0, \frac{\partial L}{\partial \lambda} = 0, \frac{\partial L}{\partial \mu} = 0, \gamma_i = 0 \quad \forall i$. Therefore, for the optimal solution

$$1 + 2\lambda x_i^{\frac{1}{2}} - 4\mu x_i^{\frac{3}{2}} = 0, \quad \forall i \text{ must hold.}$$

Considering the set of equations $1 + 2\lambda x_i^{\frac{1}{2}} - 4\mu x_i^{\frac{3}{2}} = 0, \forall i$. From the theory of algebraic equations, we know that one of the two cases listed below much apply.

Case One: The equations have the same nonnegative solution i.e. $x_i = x_j$ for any i, j .

Case Two: There are at most two different position solutions.

Assume that the positive solutions are a and b respectively and that $a \leq b$. Let

$$Z = \min \left\{ \sum_{i=1}^{i=n} \sqrt{x_i} A_i^* \right\} \text{ subject to:}$$

$$\sum_{i=1}^{i=n} x_i A_i^* = 1$$

$$\sum_{i=1}^{i=n} x_i^2 A_i^* \leq 1 + \varepsilon$$

$$x_i \in \{a, b\}$$

$$Z = \min \left\{ \sum_{\{i, x_i = a\}} \sqrt{a} A_i^* + \sum_{\{i, x_i = b\}} \sqrt{b} A_i^* \right\} \text{ subject to}$$

$$\sum_{\{i, x_i = a\}} a A_i^* + \sum_{\{i, x_i = b\}} b A_i^* = 1$$

$$\sum_{\{i, x_i = a\}} a^2 A_i^* + \sum_{\{i, x_i = b\}} b^2 A_i^* \leq 1 + \varepsilon .$$

So if we let $x = \sum_{\{i, x_i = a\}} A_i^*$ and $y = \sum_{\{i, x_i = b\}} A_i^*$, we have,

$Z \geq \min \{ \sqrt{ax} + \sqrt{by} \}$ subject to:

$$ax + by = 1$$

$$a^2x + b^2y \leq 1 + \varepsilon$$

$$a > 0, b > 0, b \geq a.$$

Therefore, $Z \geq \frac{1}{\sqrt{1+\varepsilon}}$. This completes the proof of lemma 5.2.

5.4.2 Proof of the Small Partition Case

Proof of Theorem 5.1

From lemma 5.1, for any $\varepsilon_1 > 0$, $\exists \rho_0$, when $\rho > \rho_0$,

$$\sqrt{Qd} \geq (\beta - \varepsilon_1) \int_A \sqrt{f_\Gamma(x)} dx \Rightarrow \sqrt{Nd} \geq \frac{(\beta - \varepsilon_1) \sqrt{N} \int_A \sqrt{f_\Gamma(x)} dx}{\sqrt{Q}}.$$

From proposition 5.3, for the above fixed $\varepsilon_1 > 0$, we have, $\int_A f_\Gamma^2(x) dx \leq \frac{2N(1+3\varepsilon+\varepsilon_1)}{Q\omega}$,

Where $\omega = \frac{1}{1-2\lambda R(\bar{s} + \sqrt{2}/v)} \rightarrow 1$ as $\rho \rightarrow 1$.

Using lemma 5.2, when $\rho > \rho_0$,

$$\sqrt{Nd} \geq (\beta - \varepsilon_1) \sqrt{\frac{\omega}{2(1+3\varepsilon+\varepsilon_1)}} = (\beta - \varepsilon_1) \sqrt{\frac{1}{2(1+3\varepsilon+\varepsilon_1)}}.$$

Letting $\varepsilon_1 \rightarrow 0$, $\lim_{\rho \rightarrow 1} \sqrt{Nd} \geq \beta \sqrt{\frac{1}{2(1+3\varepsilon)}}$.

Let $\varepsilon \rightarrow 0$, we obtain (5.6.a): $\lim_{\rho \rightarrow 1} \sqrt{2N\bar{d}} \geq \beta$.

Now we show (5.6.b) based on (5.6.a).

Recall $\bar{s} + \frac{\bar{d}}{v}$ is the actual average service time for each demand. In a stable system

$\lambda \left(\bar{s} + \frac{\bar{d}}{v} \right)$ must be less than 1.

We know that $\lambda \left\{ \bar{s} + \frac{\beta}{v\sqrt{2N}} + o\left(\frac{1}{\sqrt{2N}}\right) \right\} < 1$. Recalling that $N = \lambda W$ and that $\rho = \lambda \bar{s}$, we

obtain, $\lim_{\rho \rightarrow 1} \left\{ (1 - \rho)^2 W \right\} \geq \frac{\lambda \beta^2 A}{2v^2}$.

5.4.3 Proof of the Fixed Partition Case

Proof of theorem 5.3

We observe that in these systems, a server may arrive at a region, and provide continuous service to customers until there are no customers in the region. We call this the initial busy period. The server may then remain idle at the current region until a new customer arrives. At that time it enters into what we refer to as a subsequent busy period. In principle, a server may have many of these subsequent busy periods (later we show that only poor algorithms will allow the server to remain idle). Let p represent the fraction of customers that arrive during the initial busy periods and $(1 - p)$ represent the fraction that arrive during the either the idle periods or the subsequent busy periods.

Let Z_i be the number of customers served during the initial busy period for region R_i .

Define X_i to be the number of customers served during the subsequent busy periods for region R_i . Let p_i represent the fraction of customers served at region R_i during the initial busy period and $(1 - p_i)$ represent the fraction of customers served in region R_i during subsequent busy periods. At steady state we know that

$$\frac{E[X_i]}{E[Z_i]} = \frac{1 - p_i}{p_i} \Rightarrow E[X_i] = \frac{1 - p_i}{p_i} E[Z_i].$$

From now on, we only consider the case in which ρ is relatively large ($\rho > \frac{3}{4}$).

Observation 5.4: (The constraints on $E[Z_i]$ and $E[Z_i + X_i]$)

We show in the appendix that when we have at least two partitions ($\Lambda < \frac{1}{2}$), a necessary condition for the system to be stable is $p_i > \frac{1}{2}$.

If we only consider the algorithms which satisfy the constraints that $0 < \underline{w} \leq \frac{W(x)}{W} \leq \bar{w}$,

we know that when $\rho > \frac{3}{4}$, $E[Z_i] \geq \frac{2\underline{w}N\Lambda p_i - 1}{1 - p_i + p_i^2} > \frac{4}{3}(\underline{w}N\Lambda - 1)$.

Let $g(i)$ be the ratio of the average waiting time for partition R_i to W . $Z_i + X_i$ is the total number of customer served during one visit to region R_i .

$$E[X_i + Z_i] \leq \frac{2g(i)N\Lambda}{p_i(1-\rho_1)}.$$

Observation 5.5: (A Lower bound for the average distance traveled per customer served)

Given that the locations of demands in any given area is uniformly distributed, from

Beardwood, et al. (1959), we know, for any given $\varepsilon > 0$, there exists $n_0(\varepsilon)$, such that

when $n > n_0(\varepsilon)$,

$$E\left[\frac{L_{TSP}(X_1, \dots, X_n)}{\sqrt{n}}\right] \geq \beta - \varepsilon.$$

If the demand is from region R_i , The average travel distance per demand served is

$$E\left[\frac{L_{TSP}(Z_i + X_i)}{(Z_i + X_i)}\right].$$

$$\text{When } Z_i > n_0, \text{ we know } E\left[\frac{L_{TSP}(Z_i + X_i)}{(Z_i + X_i)} \mid Z_i + X_i\right] \geq \frac{\beta - \varepsilon}{\sqrt{Z_i + X_i}}.$$

$$\text{Let } \varpi = \sum_i \left[\sum_{n > n_0} \frac{\beta - \varepsilon}{\sqrt{n}} P\{Z_i + X_i = n \mid Z_i > n_0\} \right] P\{Z_i > n_0\} \Lambda.$$

A lower bound for the average distance traveled per customer served is

$$\sum_i E\left[\frac{L_{TSP}(Z_i + X_i)}{(Z_i + X_i)}\right] \Lambda = \sum_i E\left[E\left[\frac{L_{TSP}(Z_i + X_i)}{(Z_i + X_i)} \mid Z_i + X_i\right]\right] \Lambda \geq \varpi.$$

From observation 5.4, we know $E[X_i + Z_i] \leq \frac{2g(i)N\Lambda}{p_i}$ and $E[Z_i] \geq \frac{4}{3}(2\underline{w}N\Lambda - 1)$.

From the definition of $g(i)$, we know $\sum_i g(i)\Lambda = 1$.

Minimizing \underline{w} under above constraints leads to $\lim_{\rho \rightarrow 1} \sqrt{2Nd} \geq \sqrt{p(1-\rho_1)}(\beta - \varepsilon)$.

Let $\varepsilon \rightarrow 0$, we know $\lim_{\rho \rightarrow 1} \sqrt{2Nd} \geq \sqrt{p(1-\rho_1)}\beta$. (5.13)

Observation 5.6: (the cost of remaining idle when the system is not empty)

Now we make the following observation: the average idle time per demand served during each subsequent busy period(s) is bounded from below by $\frac{1}{\lambda\Lambda}(1-\rho_1)$. This comes from dividing the average interarrival time by the average number of customers served during a single busy period in an M/G/1 queue. So the average extra-cost due to idle periods per customer served for customers region R_i is equal to $\frac{1}{\lambda\Lambda}(1-\rho_1)(1-p_i)$ and the average extra-cost due to idle periods per overall demand served is bounded by

$$\sum_i (1-p_i) \frac{1}{\lambda\Lambda} (1-\rho_1)\Lambda = \frac{1-p}{\lambda\Lambda} (1-\rho_1). \quad (5.14)$$

From (5.13) and (5.14), we know that the average extra-cost due to switching and idling

is at least $\frac{\beta\sqrt{p(1-\rho_1)}}{\sqrt{2N}} + o\left(\frac{1}{\sqrt{2N}}\right) + \frac{1-p}{\lambda\Lambda}(1-\rho_1)$.

Because our system is in steady-state, we have the following inequality,

$$\frac{\lambda}{m} \left[\frac{1}{s} + \frac{\beta \sqrt{p(1-\rho_1)}}{v\sqrt{2N}} + o\left(\frac{1}{\sqrt{2N}}\right) + \frac{1-p}{\lambda\Lambda}(1-\rho_1) \right] < 1.$$

Algebraic manipulation leads to $\lim_{\rho \rightarrow 1} W \left(1 - \rho - m \frac{(1-p)(1-\rho_1)}{\Lambda} \right)^2 > \frac{\lambda\beta^2 p(1-\rho_1)}{2v^2 m^2}$.

We can show that when $p=1$, $w(p) = \frac{\lambda\beta^2 p(1-\rho_1)}{2v^2 m^2 \left(1 - \rho - m \frac{(1-p)(1-\rho_1)}{\Lambda} \right)^2}$ is minimized.

Finally, we know $\lim_{\rho \rightarrow 1} W(1-\rho)^2 > \frac{\lambda\beta^2(1-\rho_1)}{2v^2 m^2}$.

Because $\rho_1 \leq \Lambda$, when Λ is small enough, we have $\lim_{\rho \rightarrow 1} W(1-\rho)^2 \geq \frac{\lambda\beta^2}{2v^2 m^2}$.

5.5 Conclusion

We construct a class of algorithms and demonstrate that $BvR(n, k)^*$ is in this class and that it is optimal among algorithms in this class. Then we show that an algorithm in this class is asymptotically optimal. Therefore $BvR(n, k)^*$ is asymptotically optimal. If $BvR(n, k)^*$ is asymptotically optimal for the single vehicle case, it is also asymptotically optimal for the multiple vehicle case. Our results demonstrate the robustness of partition

algorithms for routing and scheduling problems. These results mirror those developed earlier for the traveling salesman problem (Karp, 1985).

CHAPTER 6 CONCLUSION AND FUTURE RESERACH

6.1 Conclusion

Stochastic network optimization plays an important role in theoretical research and application. Our research examines several key problems: the probabilistic traveling salesman problem, the dynamic traveling salesman problem and the related M/G/1 queueing system with switchover costs and, the dynamic traveling repairman problem.

For the PTSP, we developed a quasi-polynomial c -approximation algorithm. Computational results suggest that solutions developed under our algorithm have the structure of the optimal or near-optimal solutions. Further, we show that if there exists a c_1 -approximation algorithm for the k -capacitated median problem, we can find a c_2 -approximation algorithm for PTSP. By reducing the PTSP to a well known optimization problem, we know that any advancement in the development of solution methods to that problem translate translated directly to the PTSP.

For the M/G/1 queueing model with switchover costs, we characterize the optimal algorithm. By doing this, we identify a near optimal algorithm for problem instances with certain characteristics. First, we develop a lower bound for the waiting time in these systems under any arbitrary algorithm, including those that are optimal. Next we examine systems in which service is provided according to a cyclic polling algorithm. We show that for the special case where the switchover costs are identical and traffic intensity is high, the average waiting time

of the cyclic polling algorithm is bounded by approximately 2 times the average waiting time of optimal algorithm. We also show that for this special case that when $\rho \rightarrow 1$ and x (the ratio of $\overline{\lambda s^2}$ to the switching cost for a single switchover) is very small, then cyclic polling is close to optimal. When $\rho \rightarrow 1$ and x and n , the number of individual queues in our system, are both very large, cyclic polling is also close to optimal. Under the special case of very low demand intensity, cyclic polling performs poorly. For that case we provide an alternative heuristic and a lower bound for the average waiting time of optimal algorithm. When $\lambda \rightarrow 0$, our heuristic is approximately optimal.

As noted by earlier researchers (for example, Bertsimas and van Ryzin, 1991), dynamic problems on a network are much more challenging than those on a metric space. Our research provides analytical results for the DTSP on a network. First, we examine a special case of networks in which in which the optimal TSP tour and the minimum spanning tree across customer locations involve only links of equal length. For this special case, we show that the average waiting time under the a priori cyclic polling algorithm is approximately bounded by $\frac{2-\rho_1}{1-\rho_1}$ times the average waiting time of the optimal algorithm. We also identify circumstances under which our bound is very tight. This implies that under certain conditions, cyclic polling is close to optimal.

Next, we introduce a heuristic algorithm for the DTSP on a general graph. We provide a lower bound on the waiting time for the optimal algorithm. We also identify an upper bound for the average waiting time under the optimal algorithm. Finally, when the arrival rate is

very low, we provide an alternative heuristic and show that it is approximately optimal as the arrival rate approaches zero. We also present some simulation results for randomly generated networks and demonstrate the robustness of the cyclic polling algorithm. By examining the average performance over randomly generated networks, we find that when ρ is sufficiently large, that cyclic polling algorithms perform better than the longest queue first algorithm which is known to be optimal for networks in which the switching costs are constant across nodes.

For the DTRP, we construct a class of algorithms and demonstrate that a partitioning based algorithm is asymptotically optimal when the measure of traffic intensity, ρ , approaches one. We also demonstrate that the asymptotically optimal algorithm for the single server case can be easily extended to an asymptotically optimal algorithm for the m server case. This is done by partitioning the service area into m sub-regions of equal size and assigning one server to work independently in a single sub-region.

6.2 Future Research

For the PTSP, we need explore the performance of the proposed class of heuristic algorithms using simulation based analysis. In addition, we need further investigate the related k – capacitated median problem. Any advancement on the k – capacitated median problem can be translated directly to the PTSP.

For the M/G/1 with switchover costs, we would like to generalize our result to include more general cases involving different arrival processes at nodes and node dependent switchover costs.

For the DTSP, we make the following observation for the more general case in which there exist i and j for which λ_i is not equal to λ_j . One possibility is the following algorithm: Use a clustering algorithm to identify a partition such that each piece of the partition holds approximately the same expected demand; second, select the center of each partition as a representative; third, obtain an a priori tour over the representatives; finally, travel according to the a priori tour, providing service to each group according to a TSP tour over that group. We conjecture that if implemented correctly, this algorithm should have very good performance.

For the DTRP, further research should consider the situation in which customer locations are generated according to a general probability distribution function instead of uniform distribution.

REFERENCES

Arora, S., Nearly Linear Approximation Schemes for Euclidean TSP and Other Geometric Problems, Proceedings of 38th IEEE Symposium on Foundations of Computer Science, pp. 554-563, 1997.

Arora, S., P. Raghavan and S. Rao, Approximation Schemes for Euclidean k-medians and Related Problems, Proceedings of 30th Annual ACM Symposium on theory of Computing, pp. 106-113, 1998.

Ausiello, G., E. Feuerstein, S. Leonardi, L. Stougie, M. Talamo, Algorithms for the On-line Travelling Salesman, Algorithmica 29, pp. 560-81, 2001.

Beardwood, J., J. Halton and J. M. Hammersley, The Shortest Path Through Many Points, Proceedings of the Cambridge Philosophical Society, 55, pp.229-327, 1959.

Berman, O., R.C. Larson and S. Chiu, Optimal Server Location on a Network Operating as an M/G/1 queue, Operations Research 33, pp.746-771, 1985.

Bertsekas, D. and R. Gallager, **Data Networks**, Prentice Hall, NJ, 1992.

Bertsimas, D., Traveling Salesman Facility Location Problems, Transportation Science 23, pp. 184-191, 1989.

Bertsimas, D., A Vehicle Routing Problem with Stochastic Demand, Operations Research, 40, pp.574-585, 1992.

Bertsimas, D., P. Chervi and M. Peterson, Computational Approaches to Stochastic Vehicle Routing Problems, Transportation Science 29, pp. 342-352, 1995.

Bertsimas, D., and M. Grigni, Worst-case Examples for the Spacefilling Curve Heuristic for the Euclidean Traveling Salesman Problem, Operations Research Letters, 8, pp.241-244, 1989.

Bertsimas, D. and L.H. Howell, Further Results On the Probabilistic Traveling Salesman Problem, European Journal of Operational Research, 65, pp.68-95, 1993.

Bertsimas, D., P. Jaillet and A.R. Odoni, A Priori Optimization, Operations Research, 38, pp. 1019-1033, 1990.

- Bertsimas, D., and D. Simchi-Levi, A New Generation of Vehicle Routing Research: Robust Algorithms, Addressing Uncertainty, *Operations Research*, 44, pp. 286-304, 1996.
- Bertsimas, D., and G. van Ryzin, A Stochastic and Dynamic Vehicle Routing Problem in the Euclidean Plane, *Operations Research*, 39, pp. 601-615, 1991.
- Bertsimas, D., and G. van Ryzin, Stochastic and Dynamic Vehicle Routing Problem with General Demand and Interarrival Time Distributions, *Advances in Applied Probability*, 25, pp. 947-978, 1993a.
- Bertsimas, D., and G. van Ryzin, Stochastic and Dynamic Vehicle Routing Problem in the Euclidean Plane with Multiple Capacitated Vehicles, *Operations Research*, 41, pp. 60-76, 1993b.
- Borst, S.C. and O. J. Boxma, Polling Models with and without Switchover Times, *Operations Research* 45, pp. 536-543, 1997.
- Bramel, J. and D. Simchi-Levi, **The Logic of Logistics**, Springer-Verlag, New York, 1997.
- Charikar, M. and S. Guha, Improved Combinatorial Algorithms for the Facility Location and k -median Problems, 40th Annual Symposium on Foundations of Computer Science, pp.378-88, 1999.
- Clarke, G. and J. Wright, Scheduling of Vehicles from a Central Depot to a Number of Delivery Points, *Operations Research*, 12, pp. 568-581, 1964.
- Cooke, K. and E. Halsey, The Shortest Route Through a network with Time-dependent Internodal Transit Times, *Journal of mathematical analysis and applications*, 14, pp. 493-498, 1966.
- Cooper, B.R., S.C. Niu and M.M. Srinivasan, A Decomposition Theorem for Polling Models: The Switchover Times are Effectively Additive, *Operations Research*, 44, pp. 629-633, 1996.
- Desrosiers J., Y. Dumas, M. Solomon and F. Soumis, Time Constrained Routing and Scheduling, in **Network Routing**, M.O.Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser (eds), pp. 35-139, 1995.
- Dror, M. G. Laporte and P. Trudeau, Vehicle Routing with Stochastic Demands: Properties and Solution Frameworks, *Transportation Science*, 23, pp.166-176, 1989.
- Duenyas, I. And M. P. van Oyen, Heuristic Scheduling of Parallel Heterogeneous Queues with Set-ups, *Management Science*, pp.814-829, 1996.

- Eisenberg, M., The Polling System with a Stopping server, *Queueing Systems* 18, pp. 387-431, 1994.
- Federgruen, A. and Z. Katalan, The Impact of Setup Times on the Performance of Multiclass Service and Production Systems, *Operations Research* 44, pp. 989-1001, 1996.
- Few, L. The Shortest Path and the Shortest Road through n Points, *Mathematika*, 2, pp. 141-144, 1955.
- Fisher, M., Vehicle Routing, in **Network Routing**, M.O.Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser (eds), pp. 1-33, 1995.
- Friesz, T.L., J. Luque, R. Tobin and B. Wie, Dynamic Network Traffic Assignment Considered as a Continuous Time Optimal Control Problem, *Operations Research* 37, pp. 893-901, 1989.
- Garey, M.R. and D. S. Johnson, **Computers and Intractability: A Guide to the Theory of NP-Completeness**, Freeman and Company, New York, 1985.
- Golden, B. L. and W.R. Steward, Empirical Analysis of Heuristics, in **The Traveling Salesman Problem, A Guided Tour of Combinatorial Optimization**, Lawler E.L., Lenstra J.K., Rinnooy Kan A.H.G. and Shmoy(eds.), pp. 207-249, John Wiley & Sons Ltd., 1985.
- Hofri, M. and K.W. Ross, On the Optimal Control of two Queues with Server Setup Times and its Analysis, *SIAM Journal on Computing* 6, pp. 399-420, 1987.
- Irani. S. and Anna R. Karlin, Online Computation, in **Approximation Algorithms for NP-hard problems**, D.S. Hochbaum (eds), PWS Publishing Company, Boston, 1997.
- Irani, S. X. Lu and A.C. Regan, The Online Algorithms for the Dynamic Traveling Repairman Problem, *Proceedings of the 2002 Symposium on Discrete Algorithms SODA*, under review, 2001.
- Jaillet P. Probabilistic Traveling Salesman Problems, Ph.D. Dissertation, Department of Civil Engineering, Massachusetts Institute of Technology, 1985.
- Jaillet, P. A priori Solution of a Traveling Salesman Problem in which a Random Subset of the Customers Are Visited. *Operations Research*, 36, pp. 929-936, 1988.
- Jaillet, P. Analysis of Probabilistic Combinatorial Optimization Problems in Euclidean Spaces, *Mathematics of Operations Research*, 18, pp.51-70, 1993.
- Jaillet, P., and A. Odoni, The Probabilistic Vehicle Routing Problem, In **Vehicle Routing: Methods and Studies**, B.L. Golden and A.A. Assad (eds.), North-Holland, Amsterdam, 1988.

- Karp, R.M. and Steele J.M, Probabilistic Analysis of Heuristics, in **The Traveling Salesman Problem, A Guided Tour of Combinatorial Optimization**, Lawler E.L., Lenstra J.K., Rinnooy Kan A.H.G. and Shmoy (eds.), pp. 181-205, John Wiley & Sons Ltd., 1985.
- Laporte, G. The Traveling Salesman Problem: An Overview of Exact and Approximate Algorithms, *European Journal of Operational Research*, 59, 231-247, 1992.
- Larson, R. and A. Odoni, **Urban Operation Research**, Prentice Hall, 1981.
- Lawler E.L., Lenstra J.K., Rinnooy Kan A.H.G. and Shmoy (eds.), **The Traveling Salesman Problem, A Guided Tour of Combinatorial Optimization**, John Wiley & Sons Ltd., 1985.
- Lin, S. and B. Kernighan, An Effective Heuristic Algorithm for the Traveling Salesman Problem, *Operations Research*, 21, pp. 221-227, 1973.
- Lindley J. A., Urban Freeway Congestion: Quantification of the Problem and Effectiveness of Potential Solutions, *Institute of Transportation Engineers Journal*, 57, 27-32, 1987.
- Liu, Z., P. Nain and D. Towsley, On Optimal Polling Policies, *Queueing Systems*, 11, pp. 59-83, 1992.
- Lu, X., A.C. Regan, S. Irani, The M/G/1 Queue with Switchover Costs: An Examination of Alternative Heuristics, *Queueing Systems*, under review, 2001.
- Lu, X., S. Irani, A.C. Regan, The Dynamic Traveling Salesman Problem: An Examination of Alternative Heuristics, *Transportation Science*, under review, 2001.
- Lu, X., A.C. Regan, S. Irani, An Asymptotically Optimal Algorithm for the Dynamic Traveling Repair Problem, *Networks*, under review, 2001.
- Lu, X., S. Irani, A.C. Regan, A Heuristic Algorithm for the Probabilistic Traveling Salesman Problem, *Mathematics of Operations Research*, under review, 2001.
- Powell, W., A Stochastic Model for Dynamic Vehicle Allocation Problem, *Transportation Science*, 20, pp.119-129, 1986.
- Powell, W., P., Jaillet and A. Odoni, Stochastic and Dynamic Networks and Routing, in **Network Routing**, M.O.Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser (eds), pp. 141-294, 1995.
- Psaraftis, H. N., Dynamic Vehicle Routing Problems, in **Vehicle Routing: Methods and Studies**, B.L. Golden and A.A. Assad (eds), pp. 223-248, Elsevier Science Publishers, 1988.

Psaraftis, H. N., On the practical Importance of Asymptotic Optimality in Certain Heuristic Algorithms, *Networks* 14, pp. 587-596, 1984.

Psaraftis, H. N., and J. Tsitsiklis, Dynamic Shortest Paths in Acyclic networks with Markovian Arc Costs, *Operations Research*, 41, pp. 91-101, 1993.

Rao, B. L.S. P., **Asymptotic Theory of Statistical Inference**, John Wiley & Sons, 1987.

Regan, A., J. Herrmann and X. Lu, The Relative Performance of Heuristics for the Dynamic Traveling Salesman Problem, *Transportation Research Record*, under review, 2001.

Ross, S.M., **Stochastic Processes**, John Wiley & Sons, 1983.

Srinivasan, M.M. and D. Gupta, When Should a Roving Server be Patient?, *Management Science*, 42, pp. 437-451, 1996.

Spaccamela, A. M., A.H.G. Rinnooy Kan and L. Sougje, Hierarchical Vehicle Routing Problems, *Networks* 14, pp. 571-586, 1984.

Srinivasan, M.M., S.C. Niu and B.R. Cooper, Relating Polling Models with Zero and Nonzero Switchover Times, *Queueing Systems*, 19, pp. 149-168, 1995.

Stewart, W. and B. Golden, Stochastic Vehicle Routing: A Comprehensive Approach, *European Journal of Operational Research*, 14, pp. 371-385, 1983.

Tillman, F. The Multiple Terminal Delivery Problem with Probabilistic Demands, *Transportation Science*, 3, pp. 192-204, 1969.