To my parents
ACKNOWLEDGMENTS

First of all, I would like to thank my wonderful parents for always being so supportive through each and every stage of my life, no matter how far away they are.

I would also like to thank the faculty members at the University of Florida (UF), who have provided me not only huge amount of knowledge and technical skills, but also good work ethics and the spirit of teamwork during my Ph.D. study. I would like to thank my advisor Dr. Yafeng Yin for constantly being a source of inspiration, an outstanding example as professional researcher and a good friend. I am grateful to the advices he has given me, both academic and otherwise. I would like to thank Dr. Siriphong Lawphongpanich for his valuable suggestions to this dissertation, and the high standard of professionalism he has set up for me by his own earnest practice. I would like to thank Dr. Lily Elefteriadou for being one of my role models as women professors, Dr. Scott Washburn for his influence on a light and positive attitude towards life and work, and Dr. Sivaramakrishnan Srinivasan for being an example of diligent researchers. I would also like to thank them for their helpful comments from various perspectives, which helped me to place this research in a broader context. I am also grateful to many other faculty members, such as Mr. William Sampson, Dr. Cole Smith, Dr. Ravindra Ahuja, and Dr. Panos Pardalos, for their valuable guidance and help during my study.

I would also like to thank all my friends and fellow students for making my UF experience pleasant and enjoyable, especially Lihui Zhang, Ziqi Song, Alexandra Kondyli, Dimitra Michalaka, Chao Cao and Jun Liu for their friendship and support. Special thanks also go to Ines Aviles-Spadoni, coordinator of the Center for Multimodal Solutions for Congestion Mitigation at UF, for her help.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ...............................................................................................................4
LIST OF TABLES ...........................................................................................................................8
LIST OF FIGURES .........................................................................................................................9
ABSTRACT ...................................................................................................................................11

CHAPTER

1 INTRODUCTION ..................................................................................................................13
  1.1 Background ...................................................................................................................13
  1.2 Problem Statement and Dissertation Objective .............................................................16
    1.2.1 Network Level Problem ....................................................................................16
    1.2.2 Facility Level Problem ....................................................................................17
  1.3 Dissertation Outline ......................................................................................................17

2 LITERATURE REVIEW .......................................................................................................18
  2.1 Economics Principle of Congestion Pricing .................................................................18
  2.2 Canonical Equilibrium-Based Pricing Models ..............................................................19
    2.2.1 Static Network Modeling ..................................................................................20
    2.2.1.1 First-best pricing with fixed demand ..................................................20
    2.2.1.2 First-best pricing with elastic demand ................................................23
    2.2.1.3 Extensions to first-best pricing ...........................................................24
    2.2.1.4 Second-best pricing ............................................................................25
    2.2.1.5 Robust pricing.....................................................................................26
    2.2.2 Time-Dependent Pricing Models ......................................................................27
  2.3 Traffic-Responsive Pricing Models ..............................................................................29
    2.3.1 Trial-and-Error Approach .................................................................................29
    2.3.2 Day-to-Day Dynamics ......................................................................................30
    2.3.3 Adaptive Pricing Scheme for Managed Lanes ..................................................30
      2.3.3.1 Managed lanes ....................................................................................31
      2.3.3.2 Current operation of HOT lanes .........................................................32
  2.4 Summary .......................................................................................................................34

3 HIERARCHICAL PRICING FRAMEWORK .......................................................................40
  3.1 Robust Pricing at Network Level ..................................................................................41
  3.2 Adaptive Pricing at Facility Level ................................................................................42
  3.3 Summary .......................................................................................................................43
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>ROBUST NETWORK PRICING WITH BOUNDEDLY RATIONAL USER EQUILIBRIUM</td>
<td>44</td>
</tr>
<tr>
<td>4.1</td>
<td>Bounded Rationality</td>
<td>45</td>
</tr>
<tr>
<td>4.2</td>
<td>Boundedly Rational User Equilibrium in a Static Network</td>
<td>47</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Definition of BRUE</td>
<td>47</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Properties of BRUE Flow Distributions</td>
<td>49</td>
</tr>
<tr>
<td>4.2.2.1</td>
<td>Link-based representation of BRUE</td>
<td>49</td>
</tr>
<tr>
<td>4.2.2.2</td>
<td>Existence, uniqueness and nonconvexity</td>
<td>53</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Finding Best- and Worst-Case BRUE Flow Distributions</td>
<td>54</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Numerical Examples</td>
<td>55</td>
</tr>
<tr>
<td>4.3</td>
<td>Robust Pricing with Boundedly Rational User Equilibrium</td>
<td>57</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Model Formulations</td>
<td>57</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Solution Algorithm</td>
<td>62</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Numerical Examples</td>
<td>64</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>OPTIMAL ADAPTIVE TOLLING STRATEGIES FOR HIGH OCCUPANCY/TOLL LANES: BASIC MODELS</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>Feedback Approach for HOT Pricing</td>
<td>83</td>
</tr>
<tr>
<td>5.2</td>
<td>Self-learning Approach for HOT Pricing</td>
<td>84</td>
</tr>
<tr>
<td>5.2.1</td>
<td>General Framework</td>
<td>84</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Calibration of Willingness-to-Pay</td>
<td>85</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Optimal Tolling Strategy with Point-Queue Model</td>
<td>86</td>
</tr>
<tr>
<td>5.3</td>
<td>Simulation Study</td>
<td>89</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Design of Simulation Experiments</td>
<td>89</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Comparison of Feedback and Self-learning Controllers</td>
<td>90</td>
</tr>
<tr>
<td>5.3.2.1</td>
<td>Performance of feedback controller</td>
<td>91</td>
</tr>
<tr>
<td>5.3.2.2</td>
<td>Performance of self-learning controller with point-queue model</td>
<td>92</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>OPTIMAL ADAPTIVE TOLLING STRATEGIES FOR HIGH OCCUPANCY/TOLL LANES: A UNIFIED FRAMEWORK OF EXTENDED SELF-LEARNING APPROACH</td>
<td>104</td>
</tr>
<tr>
<td>6.1</td>
<td>Incorporating More Realistic Traffic Dynamics</td>
<td>105</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Impacts of Lane-Changing Behaviors</td>
<td>105</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Impacts of HOT Slip Ramp Design</td>
<td>106</td>
</tr>
<tr>
<td>6.2</td>
<td>System Inference</td>
<td>107</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Traffic State Estimation</td>
<td>107</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Demand Learning</td>
<td>110</td>
</tr>
<tr>
<td>6.3</td>
<td>Toll Determination</td>
<td>112</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Deterministic Proactive Pricing Approach</td>
<td>112</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Robust Controller with Heterogeneous Motorists</td>
<td>114</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Robust Controller with Demand Uncertainties</td>
<td>116</td>
</tr>
<tr>
<td>6.4</td>
<td>Simulation Study</td>
<td>117</td>
</tr>
</tbody>
</table>
6.5 Summary .................................................................................................................................120

7 CONCLUSION ............................................................................................................................127

LIST OF REFERENCES ..................................................................................................................130

BIOGRAPHICAL SKETCH .............................................................................................................141
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Current onsite-accessible HOT facilities in the U.S.</td>
<td>39</td>
</tr>
<tr>
<td>4-1</td>
<td>Illustrative example of the restrictiveness of the link-based BRUE representation</td>
<td>78</td>
</tr>
<tr>
<td>4-2</td>
<td>Nonconvexity of the BRUE flow distribution set for the bridge network</td>
<td>79</td>
</tr>
<tr>
<td>4-3</td>
<td>Flow distributions for the nine-node network ($\alpha = 0.10$)</td>
<td>80</td>
</tr>
<tr>
<td>4-4</td>
<td>Pricing schemes and their performances for the nine-node network</td>
<td>81</td>
</tr>
<tr>
<td>5-1</td>
<td>Performance of feedback controllers with different $K$ values</td>
<td>103</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Economics Foundation of Marginal-Cost Pricing</td>
<td>36</td>
</tr>
<tr>
<td>2-2</td>
<td>Alternative HOT lane slip ramp configurations</td>
<td>37</td>
</tr>
<tr>
<td>2-3</td>
<td>I-394 toll rates on May 18, 2005 (data source: <a href="http://www.MnPASS.net">www.MnPASS.net</a>)</td>
<td>38</td>
</tr>
<tr>
<td>4-1</td>
<td>A bridge network</td>
<td>68</td>
</tr>
<tr>
<td>4-2</td>
<td>Relationship between three different flow sets</td>
<td>69</td>
</tr>
<tr>
<td>4-3</td>
<td>The nine-node network</td>
<td>70</td>
</tr>
<tr>
<td>4-4</td>
<td>Comparison of system performances of BRUE at the nine-node network</td>
<td>71</td>
</tr>
<tr>
<td>4-5</td>
<td>Comparison of system performances of BRUE at the Sioux Falls network</td>
<td>72</td>
</tr>
<tr>
<td>4-6</td>
<td>System performances with MC pricing scheme at the nine-node network</td>
<td>73</td>
</tr>
<tr>
<td>4-7</td>
<td>Three-parallel-link network</td>
<td>74</td>
</tr>
<tr>
<td>4-8</td>
<td>Comparison of worst-case total travel times with no toll, robust and MC tolls at the nine-node network</td>
<td>75</td>
</tr>
<tr>
<td>4-9</td>
<td>System performances with MC and RT1 at the nine-node network (1)</td>
<td>76</td>
</tr>
<tr>
<td>4-10</td>
<td>System performances with MC and RT1 at the nine-node network (2)</td>
<td>77</td>
</tr>
<tr>
<td>5-1</td>
<td>System configuration for the feedback-control approach</td>
<td>95</td>
</tr>
<tr>
<td>5-2</td>
<td>System sketch of the self-learning approach</td>
<td>96</td>
</tr>
<tr>
<td>5-3</td>
<td>Throughputs, queues and toll rates resulted by the feedback controller (Scenario 1, (K=0.1))</td>
<td>97</td>
</tr>
<tr>
<td>5-4</td>
<td>Throughputs, queues and toll rates resulted by the feedback controller (Scenario 2, (K=0.1))</td>
<td>98</td>
</tr>
<tr>
<td>5-5</td>
<td>Calibrated ( \alpha_1 ) &amp; ( \gamma ) by the self-learning controller (Scenario 1)</td>
<td>99</td>
</tr>
<tr>
<td>5-6</td>
<td>Throughputs, queues and toll rates resulted by the self-learning controller (Scenario 1)</td>
<td>100</td>
</tr>
<tr>
<td>5-7</td>
<td>Performance comparison of two controllers under scenario 1</td>
<td>101</td>
</tr>
<tr>
<td>Page</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>5-8</td>
<td>Throughputs, queues and toll rates resulted by the self-learning controller (Scenario 2)</td>
<td></td>
</tr>
<tr>
<td>6-1</td>
<td>Discretized freeway representation</td>
<td></td>
</tr>
<tr>
<td>6-2</td>
<td>Discretized freeway representation (barrier-separated)</td>
<td></td>
</tr>
<tr>
<td>6-3</td>
<td>Simulation settings</td>
<td></td>
</tr>
<tr>
<td>6-4</td>
<td>Calibration of Willingness to Pay</td>
<td></td>
</tr>
<tr>
<td>6-5</td>
<td>Optimal toll rate and its performance</td>
<td></td>
</tr>
</tbody>
</table>
To further advance road pricing to be a more efficient and pragmatic tool for congestion mitigation, this dissertation proposes a hierarchical congestion pricing framework for urban transportation networks. Within the framework, toll determination is decomposed into two levels: network and facilities. Empirical studies have discovered that travelers have a strong preference for simple system-wide pricing structures. For example, dynamic network pricing models are not only difficult to implement, more importantly, their pricing signals are also too complicated for travelers to understand and consequently change their travel behaviors. On the other hand, time-varying tolls at particular facilities, such as managed lanes, are acceptable and effective. Therefore, tolls at these two levels should follow different strategies due to their distinctive purposes and travelers’ different response abilities.

At the network level, we propose a robust static or time-of-day pricing policy to avoid complex toll structures while ensuring the network to perform reasonably well against a variety of uncertainties. Sources of uncertainty in transportation networks consist of not only randomness in demand and supply, but also travelers’ stochastic and irrational behaviors. This dissertation investigates one of the uncertainties resulting from boundedly rational route-choice
behaviors. Users with bounded rationality seek for acceptable paths rather than a necessarily minimum one. Boundedly rational user equilibrium (BRUE) flow distribution is generally non-unique and can be characterized as a non-convex and non-empty path flow set. A more restrictive link-based representation is also presented. A robust pricing scheme is determined by solving a nonlinear mathematical program with complementarity constraints to minimize the system travel time of the worst-case tolled BRUE flow distribution.

At some critical facilities, the toll scheme determined at the network level may be further adjusted in response to real-time traffic conditions. This dissertation focuses on developing pricing strategies for managed toll lanes. Adaptive tolls may be adopted in order to provide a superior free-flow travel service to the users of the toll lanes while maximizing the freeway’s throughput. Two sensible and practically implementable approaches, one feedback and one self-learning, are proposed. The self-learning approach monitors conditions of the facility through both direct observation and real-time estimation, and learns recursively motorists’ willingness to pay and short-term future demand by mining the traffic data from sensors. In determination of the tolls, a detailed modeling of drivers’ lane-choice behavior and traffic dynamics is adopted to explicitly consider their impacts on the performance of the facility.

In summary, based on practical considerations of pricing, robust time-of-day tolls are proposed for the entire network while adaptive tolls are advocated for special facilities. This composes a hierarchical congestion pricing framework for a transportation system. Feasibility of this framework is shown through discussions on its general inputs and concerns. Our investigation on the network-level robust pricing with boundedly rational user behavior and the facility-level traffic-responsive pricing of managed lanes demonstrates that the proposed hierarchical framework may be practical and promising for congestion mitigation.
CHAPTER 1
INTRODUCTION

1.1 Background

The continuously-growing traffic congestion remains one of the most severe problems many societies are facing today. In the United States, congestion has grown in cities of every size in the past ten years in extent, duration and intensity. The peak-period trips are on average seven percent longer; delay per traveler per year has increased from 40 hours to 47 hours; and the percent of congested freeway mileage has risen from 51 to 60 (FHWA, 2005). The need for congestion mitigation strategies has never been greater.

Three categories of strategies may be applied as congestion reliefs. While the traditional response of adding capacity is restricted by various constraints, the combination of advanced traffic operations (ATO) and travel demand management (TDM) has gained an increasing attention as a more effective and cost-efficient solution to congestion (FHWA, 2005). One of the key components of ATO and TDM is congestion pricing.

Pricing of road networks were common in Britain and United States during most of the nineteenth century (Lindsey, 2006). But the idea of employing pricing as an approach to improve efficiency did not initiate until the seminal work by Pigou (1920) and Knight (1924). This school of thoughts originated from the public goods theory and the key idea is to internalize the externality of a trip by charging up to the full price that the trip imposes to the society. Though it has long been advocated, congestion pricing was not adopted until recently perhaps due to technical and institutional barriers.

As an approach of TDM, congestion pricing has been proved to be efficient in reshaping travel patterns. For example, Singapore implemented its Area Licensing Scheme to restrict vehicular traffic into the central area of the city in 1975. Later in 1988 an electronic toll
collecting system, namely the Electronic Road Pricing (ERP), was installed to replace the original manual pricing scheme. During its first phase, the pricing scheme reduced the traffic entering the central area by 31% despite the growth of employment and driver population in the same period. Employment of ERP further reduced the congestion by 10-15% (Chin, 2002). More recently, the city of London introduced in February 2003 a five-pound (later increased to eight) daily fee on cars entering its city center. Compared to the traffic condition in 2002, traffic entering the charged zones in 2006 was 21% lower (Transport for London, 2007).

Congestion pricing has gained more relevance in the United States as an approach of advanced highway traffic operations. A prevalent form is managed lanes, more specifically, high occupancy/toll (HOT) lanes. By allowing lower occupancy vehicles to pay to gain access to the facility, the aim of HOT lanes is to make better use of the existing capacity of the original high occupancy vehicle (HOV) lanes and reduce the total delay of the highway. Since the first managed lane was implemented in 1995 on State Route 91 in Orange County, California, HOT lanes have become popular among governors and transportation officials, in state legislatures and the media (Orski, 2006). Evaluation of the existing HOT lanes indicates that they significantly reduced traffic delay, improved travel time reliability and the overall cost-benefit ratio is greater than one (e.g., Burris and Sullivan, 2006, Sullivan and Burris, 2006, Supernak et al., 2003a). In May 2006, the U.S. Department of Transportation (USDOT) further launched a new national congestion relief initiative promoting congestion pricing and variable tolling (USDOT, 2006).

Despite of its favorable developments and the potential as a promising solution to congestion problem, road pricing still faces many challenges. Though pricing strategies in practice are often effective, the magnitude of benefits depends largely on the details of the tolling structure (Santos and Fraser, 2006). However, tolls are mostly determined in an ad-hoc manner
in practice. For instance, the Land Transport Authority (LTA) of Singapore simply doubled the preset initial ERP charge and rate increment recently in response to drivers’ insensitiveness to the original pricing scheme (LTA, 2008); the Transport of London set the toll rate without estimating the marginal congestion cost (Santos and Fraser, 2006); and the tolls on Minnesota I-394 HOT lane are adjusted according to a “look-up” table created based on static traffic assignment with assumptions regarding travel demand and value of travel time (Halvorson et al., 2006). This suggests that the current practice of road pricing can be further improved.

Yet, most pricing models in the literature do not meet the practical needs well. On one hand, network-level models are either too simple to represent the real-world situation or too complex to implement. Canonical static models (e.g., Beckmann et al., 1956, Walters, 1961, and works in the same school) rely on assumptions of users’ perfect rationality and macroscopic network performance functions. These models are ideal as they simplify many human and traffic factors. Some dynamic models employ dynamic traffic assignment with assumptions on the time-dependent origin-destination (OD) demand (e.g., Wie and Tobin, 1998). But the resulting pricing schemes are often so complex and dramatically fluctuating that they are intractable for real operations in general networks. Moreover, Bonsall et al. (2007) discovered that people have a strong preference for simple network-level pricing structures. It is not reasonable to expect users to predict and respond to the prices if the tolling signals across the entire network are very complicated and change quite often. On the other hand, it is believed that time-varying tolls at facility-level are acceptable to the users. One good example is HOT lanes (Bonsall et al., 2007), whose performance can often be improved through dynamic pricing (Supernak et al., 2003b). However, studies on HOT pricing are limited. Related single-bottleneck models (e.g. Vickery, 1963; Arnott et al., 1998, etc.) are mostly idealized and cannot be applied to HOT lanes directly.
These challenges provide many opportunities for researchers to develop sensible and practical pricing schemes. In this dissertation, we propose a hierarchical pricing framework for congestion mitigation. Within the framework, toll determination is decomposed into two levels: network and facility. At the network level, robust static or time-of-day pricing policy will be developed to avoid complex toll structures while ensuring the network will perform reasonably well under various circumstances. A detailed description of traffic dynamics is avoided at the network level. At the facility level, tolls can be further adjusted to meet local objectives with a detailed modeling of drivers’ behavior and traffic dynamics. We now state the problem more specifically.

1.2 Problem Statement and Dissertation Objective

1.2.1 Network Level Problem

Being strategic, toll rates at the network level will not be revisited frequently and often extend over a certain period of time. A static time-of-day pricing structure will be easy for the general public to understand and for transportation authorities to implement. Nevertheless, we still expect this simple pricing scheme to function well in the presence of demand and supply uncertainties. Therefore, robustness becomes an important merit and consideration in the network-level toll optimization. We seek for robust pricing strategies against uncertain equilibrium flow distributions due to volatilities in both demand and supply side.

There are many sources of uncertainty in transportation networks. This dissertation primarily focuses on examining one of the uncertainties associated with users’ route choice behavior, namely the bounded rationality. This problem contains three tasks: 1) Propose rigorous definition of BRUE; 2) Develop mathematical model to characterize BRUE flow patterns; 3) Determine robust tolls to accommodate the resulting uncertain equilibrium flow distribution.
1.2.2 Facility Level Problem

We mainly focus on HOT lanes at the facility level. Consistent with the current HOT lane operating policies (FHWA, 2003) and the recent trend of variable pricing (USDOT, 2006), this dissertation aims to build adaptive tolling strategies that can provide a superior free-flow traffic service on the HOT lanes while maximizing the throughput rate of the entire freeway segment. Note that between these two objectives, the operators often give higher priority to the former, because HOV lanes are designed “first and foremost to provide less congested conditions for carpoolers and transit users” (Munnich, 2006). The pricing model should take into account drivers’ reaction to the toll rates, i.e. their willingness-to-pay (WTP) to use the HOT lane, the impact of lane changing at HOT lane entrance on traffic dynamics, and be capable of handling other practical issues.

1.3 Dissertation Outline

This dissertation is organized as follows. Chapter 2 reviews the existing pricing models in the literature, including canonical static and dynamic equilibrium-based models and traffic responsive models. The proposed general hierarchical pricing structure is presented in Chapter 3. Chapter 4 is devoted to robust network pricing models against users’ boundedly rational route choice behaviors. Basic models for HOT lane operation are discussed in Chapter 5, followed by a unified framework of extensions in Chapter 6 that addresses several critical practical issues. Conclusion with guidelines of future research is provided in Chapter 7.
CHAPTER 2
LITERATURE REVIEW

There is an extensive literature on congestion pricing in both fields of economics and transportation engineering. Early studies in economics have built up the theoretical foundation of marginal-cost (MC) pricing while more practical settings are taken into account by researchers in transportation engineering. Based on different concepts and assumptions adopted, congestion pricing models can be coarsely categorized into two types: canonical equilibrium-based and traffic-responsive.

This chapter mainly reviews various modeling approaches in the field of transportation engineering. We will briefly summarize the economics principle of congestion pricing and examine more carefully the canonical equilibrium-based and the traffic-responsive models respectively.

2.1 Economics Principle of Congestion Pricing

The standard approach in Walters (1961) may serve as a landmark of MC road pricing. With a number of assumptions, the model suggests that in order to maximize the social benefit, road users should pay the difference between the marginal social cost and private cost. Figure 2-1 illustrates this model. Let $Q$ denote the flow along a given uniform stretch of road. Without pricing, equilibrium occurs at the intersection of the inverse demand curve $ID$ and the average cost curve $AC$, at which user’s WTP is equal to the average cost. However, social optimum is achieved at the point $B$ where $ID$ and marginal social cost curve $MSC$ intersect. This point can be achieved with a toll whose magnitude is represented by $BC$ and can be mathematically calculated as:

$$MSC(Q_0) - AC(Q_0) = \left. \frac{d[Q \cdot AC(Q)]}{dQ} \right|_{Q=Q_0} - AC(Q_0) = Q_0 \cdot \left. \frac{dAC(Q)}{dQ} \right|_{Q=Q_0}$$
With this MC pricing, total social surplus is increased by area $ABD$. The travelers lose their surplus by area $ABB'E'$, while the government collects a toll revenue of $BCC'B'$. Total net social benefit is thus $CEE'C' - ABE$, which is equal to area $ABD$. Such a pricing scheme is often referred to as first-best pricing because there are no constraints on the pricing instrument and it is assumed that the government and other markets are perfectly efficient so that the collected toll revenue can be optimally transferred (Small and Verhoef, 2007).

Though the standard approach received many critiques, mainly on its over-simplification, it is nonetheless a good starting point of congestion pricing. Debates on how to cover common costs, how to allocate excess revenues, whether and how to compensate tolled-off users, and whether to privatize highways have never stopped among economists (Lindsey, 2006). These issues are beyond the scope of this dissertation. We will assume an efficient society and focus on how to set up practically implementable and effective tolls within this context.

2.2 Canonical Equilibrium-Based Pricing Models

Canonical equilibrium-based models usually assume that travel demand (functions) and network performance functions are known. The pricing problems were first studied in a relatively ideal situation where demand for each OD pair is fixed and travel time function of each link is separable (travel time of a link is only affected by the flow on that particular link). Users are assumed to be homogeneous, perfectly informed and rational in making their travel choices (route, departure time and mode, etc.), i.e. each user attempts to minimize his own general travel cost. Many relaxations to these assumptions, including elastic demand, non-separable link travel time functions, imperfect information and heterogeneity among users can be found in the literature. Canonical models often try to determine congestion tolls that maximize the total social benefit in the short-run equilibrium (transportation and activity systems do not change),
although the measure of social benefit may differ in different contents, particularly when other externalities such as environmental, financial and maintenance cost are considered.

Canonical models can be further classified into static and time-dependent ones. Static models only consider user’s route choice, relying on the standard traffic assignment to describe the tolled equilibrium flow distribution. Another dimension of choice, namely departure time, is included in the time-dependent models. Dynamic traffic assignment models are employed to predict tolled dynamic equilibrium flows.

2.2.1 Static Network Modeling

To facilitate the presentation, let \( N \) and \( L \) denote the sets of nodes and links in a road network, and \( W \) the set of OD pairs. A link is represented by its head and tail nodes, i.e., an element of \( L \) is denoted as \((i, j)\). For OD pair \( w \), \( x_{ij}^w \) represents the flow on link \((i, j)\). The vector \( x^w \) denotes the flow vector for OD pair \( w \) with \( x_{ij}^w \) as its element, and \( v = \sum_w x^w \) is the associated aggregate link flow vector. For each link \((i, j) \in L\), \( t_{ij}(v_{ij}) \) denotes the separable link travel time function that is continuous and monotonically increasing with the aggregate link flow, \( v_{ij} \). Let \( o(w) \) and \( d(w) \) represent the origin and destination of OD pair \( w \), \( \Delta \) the node-link incidence matrix and \( D^w \) the demand vector satisfying that \( D_{o(w)}^w = q^w \), \( D_{d(w)}^w = -q^w \) and \( D_i^w = 0 \) for all other node \( i \), where \( q^w \) is the given demand of OD pair \( w \).

2.2.1.1 First-best pricing with fixed demand

Wardrop (1952) introduced two behavior principles, which was later termed user equilibrium (UE) and system optimum (SO) respectively in the work of Dafermos (Patriksson, 1994). Wardrop’s first principle states that travel time on all used routes are equal and less than those on any unused routes. Note that this principle is equivalent to the assumption that users are
perfectly rational. Under this condition, no user can further minimize his travel cost, thus users will have no incentive to change their route choices. Wardrop’s second principle is limited to the fixed demand case and suggests the total system travel time is a minimum. SO principle can be generalized to total social benefit maximization when elastic demand is considered. Static traffic assignment models are formulated to find link flow distributions such that UE or SO principle is satisfied. Accordingly, the first-best pricing problem is to determine toll schemes such that the tolled UE is SO.

Beckmann et al. (1956) formulated the static traffic assignment problems as mathematical programs, and proved the equivalence between optimality conditions of the proposed UE model and Wardrop’s first principle. The UE flow pattern can be obtained by solving the following mathematical program:

\[
\min_{v,x} \sum_{ij \in L} \int_0^{t_{ij}(v)} L_{ij} dv
\]

s.t. \[\Delta x^w = D^w \quad \forall w \in W\] \[v = \sum_w x^w\] \[x \geq 0\] (2.1) (2.2) (2.3) (2.4)

Similarly, SO traffic assignment can be formulated as:

\[
\min_{v,x} \sum_{ij \in L} t_{ij}(v_j) v_{ij}
\]

s.t. Conditions (2.2)-(2.4)

Note that \[\sum_{ij \in L} t_{ij}(v_j) \cdot v_{ij} = \sum_{ij \in L} \int_0^{t_{ij}(v)} \left( t_{ij}(v) + v \frac{dt_{ij}(v)}{dv} \right) dv\], suggesting the SO problem is equivalent to a UE problem with modified travel cost \[\tilde{t}_{ij}(v_j) = t_{ij}(v_j) + v \frac{dt_{ij}(v_j)}{dv_{ij}}\], which is exactly the marginal social cost of having one more traveler on link \((i, j)\). Further more,
following the Karush-Kuhn-Tucker (KKT) first-order optimality conditions of the UE and SO formulations, it can be shown that link-based MC toll is able to drive the system from UE to SO. Mathematically, the MC link toll $\tau_{ij}$ can be expressed as $v_{ij}^{SO} \cdot \frac{dt_{ij}(v_{ij})}{dv_{ij}|_{v_{ij}=v_{ij}^{SO}}}$, where $v_{ij}^{SO}$ is the SO link flow for $(i, j)$. This can be viewed as a network generalization of the marginal cost pricing model developed in Walters (1961).

For decades, the only first-best pricing scheme discussed in the literature is the above MC toll. However, first-best tolls that can drive UE to SO are not necessarily unique (Hearn and Ramana, 1998). Consider the tolled UE condition:

\[
\begin{align*}
\chi_{ij} \left( t_{ij} + \tau_{ij} + \pi_{i}^{w} - \pi_{j}^{w} \right) &= 0 \quad \forall ij \in L, \forall w \in W \\
t_{ij} (v_{ij}) + \tau_{ij} + \pi_{i}^{w} - \pi_{j}^{w} &\geq 0 \quad \forall ij \in L, \forall w \in W
\end{align*}
\]  

Conditions (2.2)-(2.4)

where $\pi_{i}^{w}$ represents the node potential of node $i$ for OD pair $w$. This set of conditions is exactly the KKT condition of tolled UE problem and is equivalent to Wardrop’s first principle with the presence of link tolls (see e.g., Patriksson, 1994). Since the tolled UE with first-best pricing is SO, a set of link tolls is first best if and only if there exists a set of node potentials such that conditions (2.2)-(2.4) and (2.6)-(2.7) are satisfied with $v_{ij} = v_{ij}^{SO}$ and $\chi_{ij}^{w}$ equal to some corresponding commodity link flow. More concisely, summing equation (2.6) over all links and OD pairs yields (2.8), where $v^{SO}$, $t(v)$, $\tau$ and $\pi^{w}$ are vectors with $v_{ij}^{SO}$, $t_{ij}(v_{ij})$, $\tau_{ij}$ and $\pi_{i}^{w}$ as their respective components. When the SO flow pattern is unique, the first-best toll set can be characterized as a polyhedron described in conditions (2.8)-(2.9).
\[
(t(v^{SO}) + \tau)^T v^{SO} + \sum_w (D^w)^T \pi^w = 0
\]  
(2.8)

\[
t(v^{SO}) + \tau + \Delta^T \pi^w \geq 0 \quad \forall w \in W
\]  
(2.9)

Since there exist multiple first-best toll patterns that can force the system from UE to SO, it becomes possible to consider a secondary objective within the first-best toll set. Secondary objectives may be minimizing, e.g., total toll revenue, the largest link toll, or the number of tolled links (Hearn and Ramana, 1998). Specific algorithms have been developed for finding the first-best toll that also minimizes toll revenue (Dial, 1999a, 2002 and Bai et al., 2004).

2.2.1.2 First-best pricing with elastic demand

The fixed-demand models can be easily extended to the case of elastic demand. The equilibrium OD demand is now a function of the equilibrium general OD travel cost, which includes both the travel time and the toll. Let \( d^w(q^w) \) denote the inverse demand function for OD pair \( w \). The UE problem with elastic demand is formulated as follows:

\[
\begin{align*}
\max_{q,v,x} & \quad \sum_w \int_0^{q^w} d^w(q) dq - \sum_{ij \in L} \int_0^{v^w} t_{ij}(v) dv \\
\text{s.t.} & \quad \text{Conditions (2.2)-(2.4)}
\end{align*}
\]  
(2.10)

while the SO problem is in the following form (Beckmann et al., 1956):

\[
\begin{align*}
\max_{q,v,x} & \quad \sum_w \int_0^{q^w} d^w(q) dq - \sum_{ij \in L} t_{ij}(v) v_{ij} \\
\text{s.t.} & \quad \text{Conditions (2.2)-(2.4)}
\end{align*}
\]  
(2.11)

The first-best toll set with elastic demand (Hearn and Yildirim, 2002) can then be described as

\[
\begin{align*}
(t(v^{SO}) + \tau)^T v^{SO} = \sum_w d^w(q^{wSO}) q^{wSO} \\
t(v^{SO}) + \tau + \Delta^T \pi^w \geq 0 \quad \forall w \in W \\
d^w(q^{wSO}) \leq \pi^w_{\alpha(w)} - \pi^w_{\beta(w)} \quad \forall w \in W
\end{align*}
\]  
(2.12-2.14)
where \( q_{w}^{SO} \) is the SO demand for OD pair \( w \). Similar to the case of fixed demand, the MC toll belongs to the first-best toll set as well. Furthermore, any toll vector in the first-best toll set can be interpreted as a MC pricing scheme, as the total price charged along any utilized path for any toll pattern are equal, and matches the difference between the marginal social cost and the average cost along this path (Yin and Lawphonpanich, 2009).

### 2.2.1.3 Extensions to first-best pricing

The literature offers several extensions to the basic first-best pricing models from various perspectives, such as incorporating users’ perception errors of travel time, travelers’ heterogeneity, and different performance measures other than system surplus, etc.

The assumption that travelers are perfectly informed are relaxed in the stochastic user equilibrium (SUE) models, where each traveler is assumed to have a random perception error of link travel times (Sheffi, 1985). Smith et al. (1994) provided the conditions of the existence of link tolls that can support SO as tolled SUE, where SO is considered as minimizing total system travel time. Yang (1999) explored the feasibility of MC pricing when social benefit is considered as the SO objective instead. First-best pricing with secondary objectives under SUE is further studied in Stewart and Maher (2006).

User heterogeneity has been taken into account in forms of multiple classes of travelers. These classes may differ in travel cost functions, values of travel times, vehicle types, and travel modes (Dafermos, 1973; Dial, 1999b; Yin and Yang, 2004; Chu and Tsai, 2004; and Hamdouch et al., 2007).

Some models try to determine tolls for objectives different from the simple measure of system surplus. These problems can be formulated as bi-level programming models or mathematical programs with equilibrium constraints (MPEC) with the UE or SUE models as the
lower-level problem or embedded equilibrium constraints (see, e.g., Ferrari, 2002 and Chiou, 2004). When the alternative objective reduces to a performance requirement depending only on the link flows, a simple UE traffic assignment model with additional constraint is able to provide information on the link tolls (Larsson and Patriksson, 1999).

Examples of other extensions are non-separable link travel cost (Yang and Huang, 2005) and capacity constraints on link flows (see, e.g., Ferrari, 1995; Yang and Bell, 1997).

2.2.1.4 Second-best pricing

So far we have discussed models for first-best pricing problems where there is no constraint on the toll. However, a first-best toll might not be practical in real-world implementation. For instance, every link has to be charged under the MC pricing scheme, which may not be practical. In the case of heterogeneous users, discriminatory tolls may be the only solution to first-best pricing problem (Chu and Tsai, 2004). But it might be difficult to distinguish different groups of travelers to apply discriminatory tolls. Therefore, pricing problems with constraints on the toll itself are of practical importance. These problems are known as second-best pricing problems.

Second-best pricing models are often formulated as bi-level programming (Yang and Lam, 1996; Brotcorne et al., 2001; Chen et al., 2004) or MPEC (Lawphongpanich and Hearn, 2004) where explicit constraints on the pricing instrument are employed. The constraints on the tolls can originate from either the network spatial or the user heterogeneity aspect.

One problem from the network spatial aspect is to select a limited set of toll points and determine the optimal toll rate. Verhoef (2002a) derived a general analytical solution for second-best problem where only a subset of the links can be tolled. Heuristics for integrated toll location and rate problem were developed based on the optimal solution for a given tollable sub-network (Verhoef, 2002b; May et al., 2002; Shepherd et al., 2004). More generally, Han and
Yang (2008) provided a coarse efficiency evaluation of any given second-best pricing scheme. Another problem important in practice is the cordon design. Since the cordon design is more restrictive in that a cordon has to be a closed toll ring where users are charged each time they cross the ring, many adopted meta-heuristic algorithms (Yang and Zhang, 2002; Zhang and Yang, 2004; Sumalee, 2004) and others proposed simplified or approximation models (Mum et al., 2003 and Ho et al., 2005). A third problem is for area-based pricing, where trip-chain might be involved. Area-based pricing is similar to cordon pricing but a traveler may have unlimited access to the controlled zone after paying the toll. Lawphongpanich and Yin (2008a) developed the link-based equilibrium condition under the area-based pricing and formulate the pricing problem as a MPEC. Comparison between cordon and area-based pricing can be found in Murayama and Sumalee (2007).

From a different perspective, economists demonstrated that accounting for user heterogeneity is important in evaluating constrained pricing policies (Small and Yan, 2001; Verhoef andSAMLL, 2004). However, the literature on constructing second-best pricing scheme where user groups cannot be differentiated is very limited. Verhoef et al. (1995) derived for a simple case the second-best toll as a weighted average of the first-best tolls for each group.

**2.2.1.5 Robust pricing**

A recent trend of static network-level pricing models is to address uncertainties of the transportation system. This is meaningful as the system may change within the planning horizon. For example, Li et al. (2007) developed a bi-level program model to improve travel time reliability with the presence of both random demand and capacity. Considering only demand uncertainty, Gardner et al. (2008) proposed three methods to minimize the weighted sum of the mean and the variance of total system travel time.
It should be noted that although robust pricing models are not extensively studied, numerous studies have been done in network design problems under demand uncertainties. Some studies apply the notion of reliability-based design (e.g., Yin and Ieda, 2002; Lo and Tung, 2003; Chootinan et al., 2005 and Sumalee et al., 2006) while others seek a robust design that functions reasonably well in extreme cases (Chen et al., 2003; Ukkusuri et al., 2007; Chen et al., 2007; Yin and Lawphongpanich, 2007; and Yin et al., 2008). The models and solution methods in these robust network design problems can be readily applied to robust pricing problems.

2.2.2 Time-Dependent Pricing Models

In this sub-section, we will review time-dependent equilibrium models in which another dimension of travel decision, namely departure time choice, is incorporated. Under the time-dependent setting, the UE condition can be interpreted as that no user has incentive to change his combined departure time and route choice.

To demonstrate the dispersion in departure time choice, the seminal work by Vickrey (1969) assumed that there is only a single road connecting the origin and destination with a bottleneck of fixed capacity. Traffic congestion was assumed to take the form of a point queue. An analytical solution was developed for the single bottleneck model with the travel cost as a linear function of travel time, time early and time late disutility. Accordingly, the first-best time-dependent toll takes a triangular form rising linearly with time up to some threshold and declining linearly back to zero. Under this pricing scheme, travelers will not experience any queuing delay but a toll equal to the queuing delay under UE is charged in order to guarantee the tolled equilibrium condition.

Newell (1987) provided a general analysis of single bottle neck model considering heterogeneous users. By limiting the heterogeneity only in the values of time, Arnott et al. (1988, 1992 and 1994) further developed a close-form first-best pricing scheme for the case with
two groups of users. Their analysis implies that under UE, users order themselves according to their relative values of travel time to schedule delay, while under SO the self-assorting only depends on the value of the schedule delay because the queuing delay is eliminated by tolling.

In addition to the spatial and user-class restrictions, there is one more consideration for the second-best pricing in time-dependent models: the toll may be restricted not to vary continuously with time. Both uniform toll and step tolls (Braid, 1989; Laih, 1994; Chu, 1999 and De Palma et al., 2005) are examined in the literature. Though not as effective as first-best tolls in eliminating queues, second-best tolls can reduce congestion to a certain level.

Extending single-bottleneck model to networks of bottlenecks is not very straightforward. Similar analytical solutions can only be derived for limited simple cases (Arnott et al., 1998). Some studies focused on modeling a more realistic interaction between bottlenecks by considering physical queues (Mun, 1999; Lo and Szeto, 2005). For general networks, dynamic traffic assignment models are incorporated to obtain optimal pricing schemes (Carey and Srinivasan, 1993; Huang and Yang, 1996; Wie and Tobin, 1998; Joksimovic et al., 2005; Wie 2007). However, these models are extremely difficult to solve due to the ill-behaved properties of dynamic traffic assignment, such as non-convexity and non-differentiability, arising from the need to realistically model flow dynamics (Peeta and Yang, 2003). Moreover, instead of solving for departure time jointly, some of the models require a full knowledge of dynamic OD demand, a type of information difficult to obtain.

Robust pricing models in time-dependent setting are limited. Nagae and Akamatsu (2006) developed a stochastic singular control model for dynamic revenue management under demand uncertainty for a private road. The resulting model is complicated and no efficient solution approach is provided for networks of a realistic size.
2.3 Traffic-Responsive Pricing Models

We have discussed canonical equilibrium-based models (static and time-dependent) in the previous section. These models usually assume known demand and link performance functions, and ignore the evolution of the system to the stationary equilibrium state. In contrast, in traffic-responsive models, travel demand is usually unknown. Moreover, this approach features an observation-adaption feedback loop which takes into account travelers’ learning process in response to the toll charges. The objective in traffic-responsive models is not necessarily system-wide social benefit but some local target.

2.3.1 Trial-and-Error Approach

All the equilibrium-based models require the knowledge of travel demand functions, which is usually unknown and difficult to estimate in practice. Trial-and-error approach is proposed particularly to handle the situation where travel demand is unknown.

Vickery (1993) and Downs (1993) first argued that MC congestion pricing can be applied with the absence of demand function on a trial-and-error basis. Li (1999, 2002) materialized this concept and developed a bisection implementation. The MC toll is calculated in the same way as in the static model with an initial target flow level. After applying this toll and the realized flow is observed, the new target flow is set as the average of the originally intended and the realized flows. The procedure then proceeds back and forth until the realized flow coincides with the target flow. Yang et al. (2004) further enhanced the procedure by modifying the updating rule of the target flow and provided a theoretical foundation of the feasibility of the trial-and-error approach. However, both models assumed the observed flow is the stationary state under the current toll. This assumption again ignores the evolution of system state to the equilibrium and consequently the loss of system benefit within this process.
2.3.2 Day-to-Day Dynamics

Day-to-day dynamics of travelers’ response to the change of travel cost and the planners’ follow-on adjustments to the tolls are considered important in the implementation of congestion pricing. Whether and how the system will achieve SO is not a concern in the equilibrium-based models but is the main subject of several studies on modeling pricing strategies under day-to-day dynamics.

Sandholm (2002) was among the first to study evolutionary implementation of congestion pricing. A model based on potential game theory (Sandholm, 2001) was developed. It is assumed that the planners only have limited knowledge of travelers’ preference and behavior, and update the pricing scheme with new observations of users’ response. Friesz et al. (2004) developed a disequilibrium pricing scheme to maximize the net social benefit. Travelers’ adjustment to experienced travel time is assumed proportional to excess travel cost, known as the proportional-switch adjustment process originally proposed by Smith (1984). The resulting day-to-day dynamic tolls do not necessarily settle the system at SO but some other tolled UE flow pattern. On the other hand, Yang and Szeto (2006) proposed a model that guarantees the final stationary condition to be SO, regardless of users’ underlying adjustment process. More recently, Yang et al. (2007) suggested a new approach of maximizing the decreasing rate of total system travel time at each particular day. This approach not only ensures the convergence to SO but also expedites the evolution process substantially.

2.3.3 Adaptive Pricing Scheme for Managed Lanes

Traffic-responsive pricing strategies are more widely applied to facility-level problems than network-level ones. One plausible reason is that implementation of traffic-responsive tolls at facility-level is much easier. Another reason is that through traffic-responsive tolls, local objectives can be better achieved (Supernak et al., 2003b).
tolls may confuse the travelers is barely an issue at the facility level (Bonsall et al., 2007). The most common traffic-responsively-priced facility in the U.S. is managed lanes.

2.3.3.1 Managed lanes

The FHWA defines managed lanes as “a lane or lanes designed and operated to achieve stated goals by managing access via user group, pricing, or other criteria. A managed lane facility typically provides improved travel conditions to eligible users.” (FHWA, 2003) Although the concept of managed lanes has not gained significant interest until recently, the practice started decades ago. Since the first bus exclusive lane was opened in Washington, D.C. in 1969 (FHWA, 1990), HOV lanes became widely accepted and implemented in the U.S as the most prevalent form of managed lanes. The early projects have demonstrated the potential of managed lanes in improving the efficiency of the transportation system. Nowadays, managed lanes are considered as a cost-efficient operational strategy to effectively address the mobility needs under limited ability of capacity expansion.

Besides various exclusive/bypass/reversible facilities, the majority of existing managed lanes are still HOV/HOT lanes. There are over 2600 miles of HOV/HOT facilities in U.S. and Canada by the end of 2005 (see Fuhs, 2005 for a more detailed inventory), and more HOV/HOT projects are upcoming these years (see, e.g. FHWA, 2008 for a summary of new HOT projects). Based on their physical design, managed lanes can be classified into three general categories: barrier-separated, wide-buffer-separated and non-separated. The ingress and egress designs of managed lanes depend largely on both its physical separation, location and management strategies. A variety of different weaving and buffering configurations of the ingress/egress are possible. Figure 2-2 illustrates the three most typical weaving designs for HOT facilities (FHWA, 2003) and can be generalized to other managed lanes as well.
2.3.3.2 Current operation of HOT lanes

While in a broad sense the concept of managed lane covers various management strategies as discussed above, it refers only to HOT facilities in a narrower sense (Obenberger, 2004). These facilities have the ability to adapt their management strategies (in a real-time manner) to the changing traffic conditions. This dynamic characteristic is what differentiates HOT lanes with other facilities as a recent trend of the latest managed-lane technology (Obenberger, 2004).

Currently there are seven HOT facilities in operation in the U.S. All the facilities allow eligible vehicles to pay onsite to gain access except the I-15 express lane in Utah, which provides a limited number of permits on a monthly basis (Utah DOT, 2009). Table 2-1 summarizes other six HOT facilities. These managed lanes have different time periods of operation and the pricing schemes vary from facility to facility. Note that the pricing schemes are categorized into three classes: fixed tolling scheme charges a constant usage fee throughout the entire operation period; time-dependent tolls vary with time (usually change every 30 minutes to one hour) but the rates are pre-determined; and in adaptive pricing, tolls are adjusted real time in response to the current traffic condition and may vary every several minutes. For example, tolls on I-394 in Minnesota vary from $0.25 to $4.00. When changes in density are detected, tolls can be adjusted as often as every three minutes based on a “look-up” table (Halvorson et al., 2006). Figure 2-3 illustrates the toll rates for one hour of a particular day on I-394. Similarly, tolls on I-15 vary from $0.50 to $4.00 with the highest value at $8.00, and can be adjusted up to every six minutes based on traffic counts. Recently, a new distance-based pricing system was launched on I-15. The total toll rate depends on the distance the user travels, while the per-mile toll rate is calculated responsively to the current traffic condition (San Diego Association of Governments, 2002). Another newly opened HOT facility on I-95, Florida also has the capability of adaptive tolling.
The proposed adaptive scheme responds to volume changes and adjusts toll rates each time interval by a preset margin (Wilbur Smith, 2007).

Although HOT has been launched for more than ten years and achieved visible benefits, there is still a heated debate on when and where to implement HOT, HOV and general purpose lanes in the transportation community. While Dahlgren (1998) concluded that adding a HOV lane is not always more effective than general purpose lanes, Dahlgren (2002) suggested that an added ideal (in the sense that toll can be correctly set such that the HOT lane can be fully utilized but not congested) HOT lane performs (in terms of reducing delay) as well as or better than an HOV lane in all circumstances. Practical tools are developed to assess the conversion of HOV to HOT as well (e.g., Eisele et al., 2006). Besides original conditions on HOV and general-purpose lanes, tolling structure, driver’s behaviours and performance of HOT, factors considered in implementation also include facility cross section and access design, enforcement and institutional considerations (Eisele et al., 2006 and Ungemah and Swisher, 2006). Same as in general road pricing, equity issue has always been a concern for HOT projects. Equity among different socio-demographic groups and users of different modes, and other equity issues caused by location and access of the HOT have attracted a lot of attention among all equity concerns (Weinstein and Sciara, 2006).

Regardless of all the debates about HOT projects, there is no doubt the tolling scheme plays a key role in the success of HOT lane operations. Often time operation policies of HOT lanes are to provide a superior free-flow traffic service on the toll lanes while maximizing the throughput rate of the freeway, i.e., the combined throughput of both regular and toll lanes (FHWA, 2003). To achieve the objectives efficiently, tolls should be adjusted real time in response to changes in traffic conditions. Empirical studies also show that adaptive tolling
structure outperforms fixed ones (Supernak et al., 2003b and Minnesota Department of Transportation, 2006). Supernak et al. (2003b) found the dynamic traffic-responsive tolls on I-15 significantly increase the utilization of the facility, create the desirable redistribution of peak-hour traffic and maintain a level of service C virtually at all times, while early phase of fixed tolls were not able to achieve the same performance. However, the literature on adaptive tolling strategies is very limited. Previous studies on time-varying tolls for bottlenecks (e.g., Arnott et al., 1998; Chu, 1995; Liu and McDonald, 1999; Yang and Huang, 1997) only considered hypothetical and idealized situations where system performance under equilibrium condition is the major concern, thus do not fit in the adaptive control category. It is worth mentioning that although the idea of adaptive control is rarely used in the content of pricing, it has been successfully applied to another freeway management strategy, ramp metering (see Papageorgiou and Kotsialos, 2002 for a review).

2.4 Summary

This chapter has briefly reviewed the advantages, limitations and applicability of the pricing models in the literature. The models are coarsely classified into three categories.

Canonical equilibrium-based static models provide a general guidance for planning level pricing applications. Most of the static models are developed based on the first- and second-best pricing concept in economics. Although some models attempted to incorporate more sophisticated human and network factors such as heterogeneity and uncertainty, most of them have oversimplified the real-world situation. For instance, ideal assumption of users’ perfect rationality and employment of simple macroscopic network performance functions are common. Also, sources of network uncertainties are limited only to variable demand and capacity.

Time-dependent equilibrium-based models enhance the description of traffic dynamics by applying dynamic traffic assignment models with known time-dependent OD demands. More
comprehensively, some models combine the departure time and route choice together as the underlying behavioral representation. One major drawback of these models is that the resulting pricing schemes are usually so complicated that they are not suitable for real operations in general networks.

Traffic-responsive pricing models fill the gap from the implementation of tolls to reaching equilibrium by taking into account the evolutionary process in between. More importantly, the observation-adaption approach is applicable to the cases where travel demand function is not known. Traffic-responsive models are particularly useful in managed-lane operations. The feedback feature also makes it possible to achieve other local objectives more effectively. However, studies on developing practical adaptive pricing models for managed lanes are limited.

We conclude that in order to be applicable to a general urban network, system-level pricing strategies should be static or at least static during each time period (an hour or a few hours) of a day. The current static equilibrium-based pricing models can be further enhanced by improving the representation of travelers’ behaviors and network performances to provide more efficient and robust pricing schemes for system-wide implementation. At certain critical facilities, the pricing can be traffic-responsive for better local performance. The proliferation of managed lanes has imposed a pressing need for more efficient traffic-responsive pricing models.
Figure 2-1. Economics Foundation of Marginal-Cost Pricing
Figure 2-2. Alternative HOT lane slip ramp configurations
Figure 2-3. I-394 toll rates on May 18, 2005 (data source: www.MnPASS.net)
Table 2-1. Current onsite-accessible HOT facilities in the U.S.

<table>
<thead>
<tr>
<th>Facility</th>
<th>SR 91</th>
<th>I-15</th>
<th>I-10</th>
<th>I-394</th>
<th>I-25</th>
<th>I-85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Orange County, CA</td>
<td>San Diego, CA</td>
<td>Houston, TX</td>
<td>Minneapolis, MN</td>
<td>Denver, CO</td>
<td>Miami, FL</td>
</tr>
<tr>
<td>Number of Lanes</td>
<td>2 each direction</td>
<td>2 reversible</td>
<td>2 each direction</td>
<td>2 reversible</td>
<td>1 each direction</td>
<td>2 reversible</td>
</tr>
<tr>
<td>Route Length (mi)</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>3</td>
<td>EB 8 / WB 6</td>
</tr>
<tr>
<td>Operation Period</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon-Thu</td>
<td>24 hr</td>
<td>5:30am-7:00am</td>
<td>1:00pm</td>
<td>24 hr</td>
<td>6:00am-7:00am</td>
<td>2:00pm-5:00pm</td>
</tr>
<tr>
<td>Fri</td>
<td>24 hr</td>
<td>5:00am-11:00am</td>
<td>2:00pm-8:00pm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weekends</td>
<td>NA</td>
<td>24 hr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicles Eligibility</td>
<td>Free</td>
<td>HOV3+, zero emission vehicles, motorcycles, disabled plates / veterans</td>
<td>HOV2+, motorcycles, electric hybrid vehicles</td>
<td>Transit; HOV2+, motorcycles (Only in HOV hours)</td>
<td>HOV2+, motorcycles</td>
<td>HOV2+, motorcycles</td>
</tr>
<tr>
<td></td>
<td>Buy-in</td>
<td>Others except large trucks and vehicles with trailer / boat</td>
<td>Others except trucks with 2+ axles and vehicles with trailer / boat</td>
<td>Others</td>
<td>SOV, small truck</td>
<td>Others</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Others except trucks with 3+ axles</td>
</tr>
</tbody>
</table>
CHAPTER 3
HIERARCHICAL PRICING FRAMEWORK

This chapter elaborates the hierarchical pricing framework we proposed to advance road pricing to be a more efficient and pragmatic tool for congestion mitigation. The essential idea is to decompose the decision of toll rates into two levels: network and facility. Tolls at these two levels serve distinctive purposes and address different objectives.

Tolls at the network level aim to manage the travel demand of the system. These tolls should be simple and logical so that they can be well understood by the travelers, which is a necessary condition for them to respond to the charges and shift their travel pattern towards the direction desired by the planners. Complicated network-level time-dependent tolls are not only difficult to calculate and implement but also hard for the travelers to follow, thus less likely to affect users’ travel pattern (Bonsall et al, 2007). In addition, network-level tolls are expected to accommodate to a certain extent the uncertain nature of the system. A system-level toll strategy should be robust enough against demand surge, random network failures (due to incidents or adverse weather) and uncertainties arising from travelers’ stochastic and irrational travel behaviors. These issues have not been adequately addressed in the literature.

Facility-level tolls, especially when applied to managed lanes, serve as the tickets of an extra travel option. Travelers are able to buy-in based on their overall evaluation of the managed facilities. Tolls for a certain facility can be more adaptive because travelers are usually able to adjust their lane choice locally to respond to the varying toll rates. The tolling objective at the facility level are often operational, and thus a dynamic pricing scheme may serve the purpose better through iteratively learning information from both the demand and the supply side, including the demand function of the managed lanes and the traffic condition along the facility.
The literature does not offer sensible and practical pricing models for managed lanes, leaving a gap to fill.

In the following two sections, we briefly identify the key issues that may need to be addressed to materialize the proposed hierarchical pricing concept. General data requirements and implementation concerns of both the network- and the facility-level pricing schemes are described.

### 3.1 Robust Pricing at Network Level

The following technical issues are to be solved in order to develop a simple but well-functioning network-level pricing scheme.

- **Determine times-of-day for pricing.** The dispersion of congestion during the day needs to be analyzed for a complete understanding of the current traffic pattern and the congestion problem. Since many problems of transportation systems occur during the peak period, differentiation of peak and off-peak periods is necessary in developing appropriate pricing strategies. In many situations, the congestion externality in the off-peak periods may not be severe and thus the pricing may not be necessary.

- **Estimate OD travel demand for each period.** OD travel demand is one critical data input to network-level pricing models. Time-of-day OD information can either be predicted from the planning models of metropolitan planning organizations (e.g., FHWA, 1997) or estimated using prior OD data combined with observed link flows (e.g., Chen et al., 2004). To obtain a robust pricing scheme, demand uncertainty should be proactively considered in the pricing models. The uncertainty set or likelihood region of OD demand may need to be specified, as discussed in Yin et al. (2008).

- **Select a primary system performance measure.** The performance measure of major concern may vary among different agencies, which will shape the toll scheme they may implement. Total system travel time, total social benefit, reliability or some other environmental measures may all serve as valid performance measures.

- **Identify and model uncertainties in the system.** The sources of uncertainty for a transportation system may reside in various aspects. Common sources are travel demand, incidents, infrastructure conditions and travelers’ behaviors etc. It is important to investigate which one(s) has major effects on the performance of the network so that the modeling effort can be correctly put into those that need the most.

- **Toll determination.** The pricing scheme should be robust in the presence of uncertainties in order to perform well in most circumstances. Practical and physical constraints on tolls should be considered when determining where and how much to charge.
• Communicate with general public and implement the pricing plan. The importance of public acceptance of a pricing scheme cannot be overly emphasized. It is the key to the success of the pricing strategy. Before actual launch of the pricing program, public outreach in many ways and a trial period will be necessary.

The above are among the major issues of designing a network-level toll. This dissertation will focus on the most technically-challenging issue, i.e., modeling the uncertainties and determining robust tolls.

### 3.2 Adaptive Pricing at Facility Level

In order to better achieve the local objectives at the facility level, tolls obtained from the network level can be further adjusted. More microscopic details of traffic dynamics and travelers’ choice behaviors should be considered in order to adjust the tolls in a real-time manner to effectively respond to the changes in traffic condition and demand. Some of the major issues are as follows.

• Define local objectives. Local objectives of a particular facility often depend on its design functionality. As for HOT lanes, the objective is usually to provide a superior free-flow traffic service on the toll lanes while maximizing the throughput rate of the freeway, i.e., the combined throughput of both regular and toll lanes (FHWA, 2003).

• Estimate traffic state. Difference between the current and the desired traffic conditions of the facility governs what toll should be charged for managed lanes. It is crucial to have a complete image of the freeway’s traffic state at the current time interval based on a limited amount of measurement data at sensor locations (e.g., Wang and Papageorgiou, 2005).

• Learn users’ WTP. Travelers’ choice of whether to use the optional tolled facility depends largely on their WTP. This information is usually unknown in advance because it varies among different locations, time of day and driver populations. During the operation, this information can be gradually learned by mining the loop detector data, and then be applied to determine optimal tolls for achieving control objectives.

• Predict short-term future demand. The demand of the tolled facility also depends on the total traffic demand, which often fluctuates and is one of the uncertainties in facility level pricing. It is important to predict traffic demand such that toll determination can be more proactive. The forecast is instrumented by Bayesian learning based on archived and real-time loop data (e.g., Lin, 2006).

• Toll determination. Based on the estimated current traffic condition, calibrated users’ WTP, and the predicted demand, optimal adaptive tolls can be determined to achieve local
objectives. The tolls determined from the network level should be the basis for deciding the feasible ranges of facility tolls. When optimizing the tolls, detailed traffic dynamics can be modeled to predict the effect of pricing.

- Coordination of tolls in time and space. In order to avoid charges that fluctuate dramatically in time, tolls can be determined proactively considering the future traffic demand. A rolling horizon framework may be adopted to coordinate the tolls in time. Coordination of several segments of the managed facility may help reduce the inequity that may arise among users who access the facility from different entrances.

- Enforcement. Managed lanes are easier for the general public to accept, because instead of “tolling off” travelers from original travel routes, they provide an additional option. Compared with network-level tolls, institutional difficulties in implementation may be reduced. But violations of managed lane policies may occur due to ineffective design of the facility or insufficient enforcement. This issue is among the most important practical concerns (Ungemah and Swisher, 2006).

For facility-level adaptive pricing, this dissertation will focus on developing a practical and sensible pricing scheme for HOT lanes.

3.3 Summary

This chapter has presented a hierarchical pricing framework for system-wide congestion pricing. Based on practical needs at the network and facility levels respectively, robust time-of-day tolls are proposed for the network-wide implementation while adaptive tolls are advocated for critical facilities. General inputs and technical issues of this hierarchical framework are discussed. We believe the proposed framework is feasible and has great potential for actual implementation. In the following chapters of this dissertation, we will address main technical issues at both levels to further materialize the framework.
CHAPTER 4
ROBUST NETWORK PRICING WITH BOUNDEDLY RATIONAL USER EQUILIBRIUM

This chapter presents a static model for determining optimal pricing scheme at the network level. Since network-level tolls need to be simple and logical for users to follow, it may be less frequently revisited and often extends over a certain period of time (e.g., one year). Therefore, robustness becomes a more important merit and concern in the toll optimization because we expect the pricing scheme to function reasonably well under various circumstances. Sources of uncertainty in transportation networks consist of not only randomness in demand and supply, but also travelers’ stochastic and irrational behaviors. Note that uncertainties in demand and supply can be easily incorporated using similar methods that have been applied to robust network design problems (see Section 2.2.1.5). This chapter will instead focus on one particular uncertainty arising from traveler’s bounded rational route choice behavior. Under such behavior, users do not necessarily choose a shortest or cheapest route when doing so does not reduce their travel time by a significant amount. The objective is to study the implication of BRUE in transportation systems planning and develop an analytical framework to proactively account for boundedly rational travel behaviors in the context of congestion pricing.

In Section 4.1, we will first review the literature on bounded rationality. Section 4.2 then mathematically defines BRUE and discusses the property of the set of all possible BRUE flow patterns. The problems of finding the best- and worst-case system travel times among the possible BRUE flow patterns are formulated as mathematical programs with complementarity constraints (MPCC) and then illustrated with examples. Section 4.3 formulates congestion pricing models that minimizes the worst-case system travel time and develops a heuristic solution algorithm to the models. Numerical examples are then presented to demonstrate and validate the models and solution algorithm. Summary of this chapter is provided in Section 4.4.
4.1 Bounded Rationality

Congestion pricing models in the literature (see Chapter 2) typically assume that users are perfectly or unboundedly rational, i.e., they always choose shortest (or cheapest) routes possible. However, when doing so only lead to a small or negligible improvement in their travel times, some users may not be sufficiently encouraged to change routes in practice. Behaving in this manner, users are said to have a “bounded rationale.” The literature in psychology and economics has provided a wide range of evidence that bounded rationality is important in many contexts, particularly in the context of day-to-day choices (see Conlisk, 1996 for a recent review on this topic).

In transportation, the number of models that allow for bounded rationality is minuscule. Mahmassani and Chang (1987) study the travel and trip timing decisions by boundedly rational travelers in an idealized setting that consists of one OD pair and a single route. Later, Jayakrishnan et al. (1994), Hu and Mahmassani (1997), Mahmassani and Liu (1997), and Mahmassani (2000) apply the concept and similar models to study, e.g., the effects of advanced traffic information and management systems on road systems. Similar to Conlisk (1996), Nakayama et al. (2001) concludes that their experimental study indicates a need to evaluate the validity of the perfectly rational assumption in the traffic equilibrium analysis. Although referred to as “tolerance-based,” Szeto and Lo (2006) are “forced” to consider bounded rationality in their dynamic traffic assignment problem because the problem may be infeasible when travelers are perfectly rational. In their paper, Szeto and Lo compare and contrast queuing paradigms in a dynamic setting, a setting under which many problems in transportation can be massively large, cannot be solved analytically, and often have to rely on computer simulation for solutions.
In contrast to previous researches on BRUE in dynamic setting, this chapter addresses boundedly rational route choice behaviors in static networks where BRUE arises whenever all users’ travel costs are not sufficiently larger than the best available ones and no user has an incentive to switch his or her route. The focus is to study the implication of BRUE in transportation systems planning and develop an analytical framework to proactively account for boundedly rational travel behaviors in the context of congestion pricing. We also note that there are other network equilibrium models in the literature that attempt to describe more realistically users’ route choice behaviors, such as stochastic user equilibrium models (e.g., Sheffi, 1985), risk-taking models (e.g., Mirchandani and Sourosh, 1987; Yin and Ieda, 2001), travel time budget models (e.g., Lo et al. 2006; Shao et al., 2006). The purpose of studying BRUE is to offer another alternative model that enriches the literature and to fully explore the properties and consequences of bounded rationality in route choices.

As Mahmassani and Chang (1987) pointed out, BRUE flow distributions in a static network may not be unique. In one interpretation, the flow distribution we observe at one particular day is just a particular realization from the set of multiple feasible BRUE flow patterns. Such an uncertainty incurred by boundedly rational travel behaviors may deteriorate the effectiveness of traditional congestion pricing strategies. For example, the link-based MC pricing scheme commonly advocated in the congestion pricing literature may not necessarily reduce congestion to its minimum level because users may not switch to routes with the least generalized cost. In this chapter, we seek a tolling pattern that minimizes the worse-case system travel delay among all the possible BRUE flow patterns based on generalized costs (time plus tolls).
4.2 Boundedly Rational User Equilibrium in a Static Network

4.2.1 Definition of BRUE

As explained in Mahmassani and Chang (1987) and Chen et al. (1997), travelers with bounded rationale still follow the behavior that exhibits a tendency toward utility maximization, but not necessarily to the absolute maximum level. We thus define travelers with bounded rationality as those who (a) always choose routes with no cycle and (b) do not necessarily switch to the shortest (cheapest) routes when the difference between the travel times (or costs) on their current route and the shortest one is no larger than a threshold value. Using the terminology in Ahuja et al. (1992), all utilized routes under bounded rationality must correspond to paths. (In the literature, others refer to paths with no cycle as “simple paths.” Ahuja et al., 1992, refer to routes with cycles as “walks.”) The above consideration motivates to the following definitions:

**Definition 4.1**: A path is considered “acceptable” if the difference between its travel time or cost and that of the shortest or least-cost path is no larger than a pre-specified threshold value.

**Definition 4.2**: A path flow distribution is at BRUE if it is feasible or compatible with the travel demands and every user uses an acceptable path.

Unlike the conventional or perfectly rational user equilibrium (PRUE) the flow distribution under BRUE as defined above may not utilize any shortest or least-cost path.

To mathematically define BRUE, we follow the notation in Chapter 2. We further let $P^w$ be the set of paths for OD pair $w$. For each path $r \in P^w$, $f_r^w$ and $c_r^w$ denote the corresponding path flow and path travel time. We use $f^w$ to represent a vector of path flows for OD pair $w$, with $f_r^w$ as its elements, and $f = (f^1, \ldots, f^w, \ldots, f^{|P|})'$, where $|\bullet|$ denotes the cardinality of a set. We may similarly define $c^w$ and $c$ for the vectors of path travel times. Assuming that users of
the same OD pair have the same threshold value, the result below describes a BRUE flow
distribution algebraically.

**Theorem 4.1**: Assume that the threshold value for OD pair \( w \) is \( \bar{c}^w \). Then, a feasible path flow distribution \( f \) is at BRUE if and only if there exist \( \lambda^w \in R \) for each \( w \) such that the following conditions hold:

\[
\begin{align*}
  c_r^w & \geq \lambda^w & \forall r \in P^w, w \in W \\
  c_r^w & \leq \lambda^w + \bar{c}^w & \forall r \in P^w = \{ r : f_r^w > 0, r \in P^w \} \\
  \sum_{r \in P^w} f_r^w & = q^w & \forall w \in W \\
  f_r^w & \geq 0 & \forall r \in P^w, w \in W
\end{align*}
\]

**Proof**: For a given feasible path flow distribution \( f^w \), denote the corresponding minimum
OD travel time as \( \bar{c}^w \).

**Necessity**: If \( f \) is a BRUE flow distribution, it is feasible by definition, thus conditions (4.3) and (4.4) are satisfied. Conditions (4.1) and (4.2) can be satisfied by setting \( \lambda^w = \bar{c}^w \).

**Sufficiency**: If \( f \) is a flow distribution with \( \lambda^w \) satisfying the above conditions, then conditions (4.3) and (4.4) guarantee that the path flow distribution is feasible. Condition (4.2) implies \( \lambda^w \) is the lower bound of \( c_r^w \). Together with condition (4.2), we have

\[
c_r^w \leq \lambda^w + \bar{c}^w \leq \min_{r \in P^w} \{ c_r^w \} + \bar{c}^w = \bar{c}^w + \bar{c}^w
\]

for every utilized path \( r \), suggesting that the travel time of any utilized path may differ at most the threshold value from the minimum OD travel time. \( \square \)

The above conditions requiring knowing utilized paths. As an alternative, the ones below use slack variables \( \varepsilon_r^w \) and do not require the utilized-path information.
\begin{align*}
&c^w_r - \lambda^w - \varepsilon^w_r = 0 \quad \forall r \in P^w, w \in W \quad (4.5) \\
&f^w_r (\varepsilon^w - \varepsilon^w_r) \geq 0 \quad \forall r \in P^w, w \in W \quad (4.6) \\
&\sum_{r \in P^w} f^w_r = q^w \quad \forall w \in W \quad (4.7) \\
&f^w_r \geq 0 \quad \forall r \in P^w, w \in W \quad (4.8) \\
&\varepsilon^w_r \geq 0 \quad \forall r \in P^w, w \in W \quad (4.9)
\end{align*}

Conditions (4.5)-(4.9) are equivalent to conditions (4.1)-(4.4), thus equivalent to the BRUE definition. We therefore refer equations (4.5)-(4.9) as the BRUE conditions.

Note that when \( \varepsilon^w = 0 \) for all \( w \), that the BRUE conditions reduce to the PRUE conditions. To illustrate, combining conditions (4.6), (4.8) and (4.9) yields \( f^w_r \varepsilon^w_r = 0 \), which together with condition (4.5) leads to \( f^w_r (c^w_r - \lambda^w) = 0 \), \( \forall r \in P^w, w \in W \). On the other hand, a PRUE flow distribution always satisfies conditions (4.5)-(4.9) with \( \varepsilon^w_r = 0 \) and \( \lambda^w \) as the equilibrium O-D travel time. Thus, a PRUE distribution is always a BRUE flow distribution.

4.2.2 Properties of BRUE Flow Distributions

When users are perfectly rational, it is well-known (see, e.g., Patriksson, 1994) that there exist equivalent link-based equilibrium conditions and the set of all PRUE flow patterns is convex. Below, we show that these properties may not hold when users are boundedly rational. More specifically, the link-based representation of BRUE proposed in this dissertation is not equivalent to the path-based definition and the set of BRUE flows may not be unique.

4.2.2.1 Link-based representation of BRUE

Let the notations follow those defined in Chapter 2. Moreover, we assume \( t_\phi(v_\phi) > 0 \) when \( v_\phi > 0 \). The following conditions (4.10)-(4.15) provide a link-based representation of BRUE, analogous to those under the perfectly rational assumption:
\[ t_{ij}(v_{ij}) + \pi_i^w - \pi_j^w - \epsilon_{ij}^w = 0 \quad \forall w \in W, \ (i, j) \in L \tag{4.10} \]
\[ x_{ij}^w(-\epsilon_{ij}^w + \mu_i^w - \mu_j^w) \geq 0 \quad \forall w \in W, \ (i, j) \in L \tag{4.11} \]
\[ -\mu_{d(w)}^w \leq \bar{\epsilon}^w - \mu_{o(w)}^w \quad \forall w \in W \tag{4.12} \]
\[ \Delta x^w = D^w \quad \forall w \in W \tag{4.13} \]
\[ \nu = \sum_w x^w \tag{4.14} \]
\[ x, \nu \geq 0 \tag{4.15} \]

where \( \pi^w \in R^{|\mathcal{V}|}, \epsilon^w \in R^{|\mathcal{L}|} \) and \( \mu^w \in R^{|\mathcal{V}|} \) are vectors of node potentials associated with travel time, travel times in excess of the minimum, and node potential associated with the excess travel times respectively.

Using a similar argument previously for the path-based BRUE definition, it can be verified that when \( \bar{\epsilon}^w = 0 \), conditions (4.10)-(4.15) reduce to PRUE conditions in the literature when users are perfectly rational. When \( \bar{\epsilon}^w > 0 \), we prove in the following that a flow distribution satisfying conditions (4.10)-(4.15) is a BRUE flow (see Definition 4.2). However, as a counterexample below illustrates, the converse is not true.

**Lemma 4.1**: Define a sub-network \( G^w \) with respect to each OD pair \( w \), which consists of all the links with strictly positive OD flows \( x_{ij}^w \). If there exist a link flow distribution \( \nu \), vectors \( \pi^w \in R^{|\mathcal{V}|}, \epsilon^w \in R^{|\mathcal{L}|} \) and \( \mu^w \in R^{|\mathcal{V}|} \) satisfying conditions (4.10)-(4.15), then the sub-network \( G^w \) is acyclic.

**Proof**: Suppose that there is a cycle in \( G^w \). By definition, \( x_{ij}^w > 0 \) for each link \((i, j) \in G^w \). Condition (4.11) then requires \(-\epsilon_{ij}^w + \mu_i^w - \mu_j^w \geq 0 \). Thus, summing conditions (4.10) and (4.11) along the cycle leads to

\[ \sum_y t_{ij}(v_{ij}) = \sum_y \epsilon_{ij} \leq 0 \tag{4.16} \]
However, since $v_{ij} \geq x_{ij}^w > 0$, we have $t_{ij}(v_{ij}) > 0$, which contradicts condition (4.16). Therefore $G^w$ is acyclic. □

Note that Lemma 4.1 applies to PRUE as well, under which the flow-carrying sub-network is acyclic if the assumption of link travel time functions holds.

**Theorem 4.2**: For a link flow distribution $v$, if there exist $\pi^w \in R^{[\pi]}$, $\varepsilon^w \in R^{[\varepsilon]}$ and $\mu^w \in R^{[\mu]}$ such that conditions (4.10)-(4.15) hold, then any of its underlying path flow patterns satisfies Definition 4.2.

**Proof**: Assume that a link flow distribution $v$ satisfies conditions (4.10)-(4.15). As $v$ satisfies conditions (4.13)-(4.15), it is a feasible flow distribution. According to Lemma 4.1, the flow-carrying sub-networks are acyclic, implying the underlying utilized routes are all paths. Let $\delta_{ij}^r$ (equals 0 or 1) indicates whether link $(i, j)$ is on path $r$. It follows from conditions (4.10) and (4.15) that for any path between O-D pair $w$,

$$
\pi_{d(w)}^w - \pi_{o(w)}^w = \sum_{ij} \delta_{ij}^r t_{ij}(v_{ij}) - \sum_{ij} \delta_{ij}^r \varepsilon_{ij}^w \leq \sum_{ij} \delta_{ij}^r t_{ij}(v_{ij})
$$

Therefore, $\pi_{d(w)}^w - \pi_{o(w)}^w$ is the lower bound of $\bar{c}^w$, the minimum travel time between a given OD pair $w$. For any utilized path $r \in P^w_+$ with respect to the given link flow distribution, the flow on each link along the path must be positive, i.e., $x_{ij}^w > 0$ if $\delta_{ij}^r = 1$. Thus summing together condition (4.11) for links on a utilized path $r \in P^w_+$ yields:

$$
\sum_{ij} \delta_{ij}^r \varepsilon_{ij}^w \leq \mu_{o(w)}^w - \mu_{d(w)}^w
$$

Essentially $\mu_{o(w)}^w - \mu_{d(w)}^w$ is the upper bound on the amount in excess of the shortest travel time.

Combining conditions (4.12), (4.17) and (4.18) for a utilized path $r$ leads to:

$$
\sum_{ij} \delta_{ij}^r t_{ij}(v_{ij}) - \bar{c}^w \leq \sum_{ij} \delta_{ij}^r t_{ij}(v_{ij}) - (\pi_{d(w)}^w - \pi_{o(w)}^w) = \sum_{ij} \delta_{ij}^r \varepsilon_{ij}^w \leq \mu_{o(w)}^w - \mu_{d(w)}^w \leq \bar{c}^w
$$
which implies that for any utilized path, its travel time may differ at most the threshold value from the minimum OD travel time. Therefore, any underlying path pattern of \( v \) is at BRUE. □

On the other hand, it is possible that a BRUE flow distribution that satisfies conditions (4.5)-(4.9) does not satisfy conditions (4.10)-(4.15). This implies that the link-based representation of BRUE is more restrictive than the path-based definition. Consider the bridge network in Figure 4-1 in which there is only one OD pair (1, 4) with a demand of 3 and a threshold value of 4. The link travel time functions are reported in Table 4-1 together with a BRUE path flow pattern.

The minimum OD travel time is 9 and all the utilized paths have travel times of no more than \( 9 + 4 = 13 \). Therefore, the path flow pattern is a valid BRUE flow. However, the corresponding link flows do not satisfy conditions (4.10)-(4.15) as the following system of equations does not admit a solution.

\[
\begin{align*}
5 - \varepsilon_{13} + \pi_1 - \pi_3 &= 0 \\
8 - \varepsilon_{12} + \pi_1 - \pi_2 &= 0 \\
1 - \varepsilon_{23} + \pi_2 - \pi_3 &= 0 \\
6 - \varepsilon_{24} + \pi_2 - \pi_4 &= 0 \\
1 - \varepsilon_{32} + \pi_3 - \pi_2 &= 0 \\
4 - \varepsilon_{34} + \pi_3 - \pi_4 &= 0 \\
\varepsilon_{12} + \varepsilon_{24} &\leq 4 \\
\varepsilon_{12} + \varepsilon_{23} + \varepsilon_{34} &\leq 4 \\
\varepsilon_{13} + \varepsilon_{32} + \varepsilon_{24} &\leq 4 \\
\varepsilon_{13} + \varepsilon_{34} &\leq 4 \\
\varepsilon_{ij} &\geq 0
\end{align*}
\]

The nonequivalence between the two sets of conditions originates from the relationship between path and link flows. Under BRUE, it is possible that a path flow pattern results in a cyclic flow-carrying sub-network (the original network in the above example) while with the link-based BRUE definition, all the flow-carrying sub-networks must be acyclic (Lemma 4.1).
Moreover, for a non-utilized path that only consists of links with positive flows (path 1-2-4 in the above example), the link-based BRUE definition still requires its travel time to be within the indifference band even that the path is not utilized (conditions (4.10) and (4.11)). Note that both issues do not arise in the PRUE because (a) the flow-carrying sub-networks are acyclic under UE; and (b) all links with positive flows must be on shortest paths. For a non-utilized path that consists of only positive-flow links, its travel time will be minimum OD travel time and conditions (4.10) and (4.11) alike will be satisfied automatically.

In this sub-section, we provided a more restrictive link-based representation of BRUE flow. Figure 4-2 illustrates the relationship between the feasible flow set, the path-based and the link-based BRUE flow set. Note that all three sets contain the PRUE flow distribution.

4.2.2.2 Existence, uniqueness and nonconvexity

In a static network with continuous travel time functions, a BRUE flow distribution always exists since a PRUE distribution is a BRUE flow distribution with $\varepsilon^w = 0$ and the existence of the former is ensured by the continuity of travel time functions (e.g., Patriksson, 1994). If the threshold value, $\bar{\varepsilon}^w$, equals zero for each OD pair, BRUE reduces to PRUE. In this case, BRUE link flow distribution is unique as we have assumed that travel time functions are monotonically increasing (e.g., Patriksson, 1994). In general, however, BRUE link flow distributions are not unique. For example, if the threshold value is sufficiently large, every feasible path flow distribution is a BRUE distribution and the set of BRUE flows is a polyhedron. We further use the following example to illustrate that in general the BRUE set is not convex in path flows.

Consider the same network in Figure 4-1 with different travel time functions reported in Table 4-2 with one OD pair (1, 4) as before. However, other problem parameters are new. The
travel time functions are as shown in Table 4-2. The demand and a threshold value are 6 and 15, respectively.

Table 4-2 displays two BRUE flow distributions, BRUE-1 and BRUE-2. The maximum differences between the lengths of utilized paths and the shortest paths are 13 for BRUE-1 and 14 for the other, i.e., both differences are within the threshold value. However, the convex combination of these two flow distributions $0.5 \times (\text{BRUE-1} + \text{BRUE-2})$ is not a BRUE flow distribution because the maximum difference now is 24. This implies that the BRUE flow set is not convex. Indeed, the non-convexity of the BRUE flow set is largely due to the nonlinear condition (4.2) in the BRUE definition. Similarly, the restrictive link-based BRUE set is not convex either due to the complementarity condition (4.7).

### 4.2.3 Finding Best- and Worst-Case BRUE Flow Distributions

In practice, traffic assignment models in the four-step process are based on PRUE and typically provide a unique equilibrium solution. In other words, PRUE models typically provide a point estimate of traffic flow distribution. On the other hand, there are generally many BRUE flow distributions and models based on BRUE provide an interval estimate instead. The problem of finding best-case BRUE flow distribution can be formulated as follows:

$$\text{BC/WC-BRUE: } \min_{\{f, \lambda, \xi, \zeta\}} \max \left\{ c(f)^T f \right\}$$

s.t. Conditions (4.5)-(4.9)

Clearly, the above two problems identify BRUE flow distributions that minimize and maximize the total travel time in the system.

To avoid generating paths, we henceforth work with the more restrictive link-based BRUE representation and define the analogous best- and worse-case problems as follows. The formulation contains auxiliary variables, $z_{ij}^w$, and additional constraints to highlight the complementarity structure in the problems.
BC/WC-ABRUE:

\[
\begin{align*}
\min / \max_{(v,x,e,\pi,\mu)} & \quad t(v)^T v \\
\text{s.t.} & \quad t_i(v_j) + \pi_i^w - \pi_j^w - e_{ij}^w = 0 \quad \forall w \in W, (i, j) \in L \\
& \quad x_{ij}^w z_{ij} = 0 \quad \forall w \in W, (i, j) \in L \\
& \quad z_{ij}^w \geq e_{ij}^w - (\mu_i^w - \mu_j^w) \quad \forall w \in W, (i, j) \in L \\
& \quad z_{ij}^w \geq 0 \quad \forall w \in W, (i, j) \in L
\end{align*}
\]

Conditions (4.12)-(4.15)

As formulated, both BC- and WC-ABRUE are MPCC, a class of optimization problems difficult to solve because the problems are generally non-convex and violate the Mangasarian-Fromovitz constraint qualification (MFCQ) at any feasible point (see, e.g., Scheel and Scholtes, 2000). In the literature, many (see, e.g., Luo et al. 1996 and references cited therein) have proposed special algorithms to solve MPCC. Lawphongpanich and Yin (2008b) applied concepts from manifold sub-optimization and proposed a new algorithm that converges to a strongly stationary solution in a finite number of iterations. The algorithm can be applied with multiple initial solutions to solve both formulations for good strongly stationary solutions.

4.2.4 Numerical Examples

This sub-section presents two numerical examples to demonstrate the relevance of BRUE. BC-ABRUE and WC-ABRUE are solved on two networks: nine-node (see, e.g., Hearn and Ramana, 1998) and Sioux Falls (see, e.g., LeBlanc et al., 1975) using GAMS (Brook et al., 2003) and CONOPT (Drud, 1995).

**Nine-node problem:** Figure 4-3 displays the underlying network. The network data can be found in Hearn and Ramana (1998). There are four OD pairs and the demands are \( q^{1-3} = 10 \), \( q^{1-4} = 20 \), \( q^{2-3} = 30 \) and \( q^{2-4} = 40 \). The threshold value is assumed to be \( \bar{e}^w = \alpha \cdot t_{UE}^w \) where \( \alpha \) is a parameter and \( t_{UE}^w \) is the PRUE travel time between a given OD pair \( w \).
Results for the nine-node problem are presented below. Table 4-3 compares the PRUE, SO, best- and worst-case BRUE flow distributions with $\alpha = 0.10$. Note that the SO flow distribution in a network with boundedly rational behaviors is the same as standard SO that minimizes total system travel time. Figure 4-4 displays the system travel times of SO, best- and worst-case BRUE flow distributions with $\alpha$ varying from 0 to 0.20. It can be observed that when $\alpha = 0$, the BRUE flow distribution reduces to the unique PRUE flow and there is no difference between the best- and worst-case performances. As $\alpha$ increases, the difference becomes more substantial. For example, the difference rises from 9.2% to 21.0% when $\alpha$ changes from 0.10 to 0.20. As the SO flow distribution is still the one that minimizes total system travel time, it does not change with the threshold value.

**Sioux Falls problem:** The Sioux Falls network (see, LeBlanc et al., 1975) contains 76 links, 24 nodes, and 528 OD pairs. The original network parameters in LeBlanc et al (1975) are used while original demands are divided by 11. We assume the same threshold value for users departing from the same origin, i.e., $\bar{\epsilon}^o = \alpha \cdot \tilde{t}^o_{UE}$ where $\tilde{t}^o_{UE}$ is the maximum PRUE travel times among the OD pairs that share the same origin $o$.

Figure 4-5 presents the discrepancy between the best and worst system performances of BRUE flow distributions for Sioux Falls with $\alpha$ varying from 0 to 0.20. Although the Sioux Falls network is not as congested as the nine-node network, a similar pattern can be observed and the difference increases up to 6.7%.

The above two examples reveal a difficulty of evaluating or designing improving strategies to highway networks as the resulting performances are likely to be uncertain (Mahmassani and Chang, 1987). Such uncertainty should be accommodated proactively in the decision making process. Note that the true range of possible total system travel times may be larger than those
shown in Figure 4-4 and Figure 4-5, for two reasons. First, the link-based conditions (4.10)-(4.15) only characterize a subset of BRUE flows. Second, the algorithms we implemented, like most algorithms for MPCC, do not guarantee local or global optimality, regardless of the fact that we initiated or “warm-started” the algorithms with many initial solutions, randomly generated and otherwise.

4.3 Robust Pricing with Boundedly Rational User Equilibrium

4.3.1 Model Formulations

To motivate the models, we implement a traditional MC pricing scheme on the nine-node network. Under MC pricing, tolls are in the form of \( t_y (v_y^{SO}) \cdot v_y^{SO} \) where \( v_y^{SO} \) is the SO flow. Imposing such a pricing scheme may not necessarily evolve the system to SO, because the tolled BRUE flow distributions may not be unique. Figure 4-6 presents the best and worst performances that MC pricing scheme may achieve as compared with those without pricing. As expected, MC pricing may result in a range of system performance and the difference can be as high as 21.3% when \( \alpha = 0.20 \). Although MC pricing achieves SO at its best performance (i.e., under the scenario with perfect rationality), it may potentially make the traffic condition worse because users are not perfectly rational. More specifically, it can be observed from Figure 4-6 that when \( \alpha \) is greater than 0.08, the worst-case system travel time with MC pricing will be greater than the best-case travel time without tolling.

Although Figure 4-6 does not show a case that MC pricing leads to an inferior worse-case performance, we do observe it when \( \alpha \) is large. This suggests that MC pricing does not necessarily ensure an improvement in the worst-case performance. To further illustrate, consider a network with three parallel links connecting one O-D pair in Figure 4-7. Assume that the O-D demand is 2 and the threshold value is 3. It is easy to verify that the worst-case BRUE flow
distribution is \( v = (2,0,0) \), with a total system travel time of 8. The SO flow for the problem is \( v^{SO} = (1,0,1) \). This yields the MC toll \( \tau^{MC} = (2,0,1) \). The worst-case BRUE flow distribution with MC tolls is \( (0,2,0) \) and has a system travel time of 10, an amount larger than the one without any toll.

In real-world applications, decision makers and planners tend to be risk averse and may be more concerned with the worst-case performance. Following the notion of robust optimization (e.g., Ben-Tal and Nemirovski, 2002), we attempt to determine a pricing scheme to minimize the worst-case system travel time among all the tolled BRUE distributions. Then, the problem of finding an optimal pricing scheme can be formulated as the following min-max program:

RTP-1: \[
\min_{\tau} \max_{(v,x,\varepsilon,\pi,\mu)} \quad t(v)^T v \\
\text{s.t.} \quad t_{ij}(v_{ij}) + \tau_{ij} - \varepsilon_w^v + \pi^w_i - \pi^w_j = 0 \quad \forall w \in W, \ (i,j) \in L \quad (4.19) \\
0 \leq \tau \leq \tau_{\max} \quad (4.20)
\]

Conditions (4.11)-(4.15)

where, \( \tau \) denotes a toll vector with a link toll \( \tau_{ij} \) as its element and \( \tau_{\max} \) is the maximum toll vector that can be possibly charged to each link.

For convenience, we represent the feasible region of the inner problem or the maximization part of RTP-1 as \( \Xi(\tau) \), the decision variable \( (v,x,\varepsilon,\pi,\mu) \) as \( \xi \) and thus the objective function as \( \varphi(\xi) \). The inner problem can then be written as:

RTP-1-IN: \[
\psi(\tau) = \max_{\xi} \{ \varphi(\xi) : \xi \in \Xi(\tau) \} \quad (4.21)
\]

It then follows that RTP-1 is essentially in a form of \( \min_{0 \leq \tau \leq \tau_{\max}} \psi(\tau) \), which is a generalized semi-infinite min-max problem (see, e.g., Polak and Royset, 2005), because the feasible set of the inner problem depends on the decision variables of the outer problem. A generalized semi-infinite optimization problem is more difficult to solve than its ordinary counterpart (e.g., Lopez
and Still, 2007). There are only a few studies dealing with numerical algorithms for such a
generalized problem (see, e.g., Still, 1999), and the methods presented may not be applicable to
RTP-1 as its inner problem is a MPCC and the MFCQ is not satisfied at any feasible point of
\( \Xi(\tau) \).

The optimum solution to RTP-1 is expected to reduce the worst-case system travel time
but without any guarantee for its best-case performance. Alternatively, it may be meaningful to
find a robust toll whose best-case performance is guaranteed. To this aim, recall that Hearn and
Ramana (1998) derived a “valid” toll set for PRUE, which consists of an infinite number of valid
tolls and each valid toll can evolve the system from PRUE to SO when users are perfectly
rational. Clearly, any toll vector in Hearn and Ramana’s valid toll set will achieve SO as its best-
case performance under the BRUE setting. However, other toll vectors not in the valid toll set
may also admit SO as their tolled best-case performance. We denote all these tolls as
“optimistic” tolls.

We now derive the optimistic toll set. Let \( S := \arg \min \{ (\nu)^T \nu \mid \nu \in V \} \) where

\[
V := \left\{ \nu \mid \Delta \nu = \sum_w D^w \quad \text{and} \quad \nu \geq 0 \right\}
\]
denotes the feasible aggregate link flow set, and

\[
X(\nu) := \left\{ x \sum_w x^w = \nu \quad \text{and} \quad \Delta x^w = D^w, x^w \geq 0, \forall w \in W \right\}.
\]

**Lemma 4.2:** Let \( \tau \) be a toll vector, \( \hat{\nu} \) be a feasible aggregate flow and \( \hat{x} \in X(\hat{\nu}) \). Then \( \hat{x} \)
is a tolled (link-based) BRUE flow distribution with \( \tau \) being the toll vector if and only if there
exist vectors \( \rho^w \in R^{|N|}, \eta^w \in R^{|L|} \) and \( \nu^w \in R^{|N|} \) such that the following holds:
In the above, the triplet \((\rho, \eta, \nu)\) plays the role of \((\epsilon, \pi, \mu)\) in the link-based BRUE conditions (4.10)-(4.15). More specifically, \(\rho, \eta, \nu\) are vectors of node potentials associated with travel time, travel times in excess of the minimum, and node potentials associated with the excess travel times respectively in the tolled BRUE.

**Theorem 4.3:** Let \(T(\hat{x})\) denote the \(\tau\) part of the polyhedron (4.22)-(4.25), then the optimistic toll set is equivalent to \(\tau = T(\hat{x})\).

**Proof:**

*Necessity:* If \(\tau\) is an optimistic toll, then there must exist some feasible by-commodity (OD pair) link flow \(\hat{x}\) in the tolled BRUE flow set whose aggregate link flow \(\hat{v}\) is SO. By Lemma 4.2, \(\tau \in T(\hat{x})\). Together with that \(\hat{v} \in S\) and \(\hat{x} \in X(\hat{v})\), we have \(\tau \in \tau\).

* Sufficiency: *If \(\tau \in \tau\), \(\tau\) must belong to some set \(T(\hat{x})\) where the aggregate link flow corresponding to \(\hat{x}\) is the SO flow. By Lemma 4.2, \(\hat{x}\) is a tolled link-based BRUE flow with \(\tau\) being the toll vector. Therefore, \(\tau\) is an optimistic toll since the tolled best-case BRUE can achieve SO.

The optimistic toll set \(\tau\) is not necessarily convex although all its component sets \(T\) are polyhedrons. Even when the set \(S\) is singleton under the assumption that the link travel time function is strict monotone, set \(\tau\) is still a union of an infinite number of subsets \(T\) because the by-commodity link flow will not be unique. Only under the extreme case where there is only
one commodity, the optimistic toll set \( \mathcal{T} \) reduces to a polyhedron. As previously indicated, Hearn and Ramana’s valid toll set is a subset of the optimistic toll set \( \mathcal{T} \).

All optimistic tolls achieve SO at their best-case performance, but their worse-case performances are not necessarily the same. We then formulate an alternative model to find the most robust toll vector from the optimistic toll set. The solution will achieve SO at its best-case performance but probably lead to a higher worst-case travel time compared with the solution to RTP-1. Since \( \mathcal{T} \) is a complicated set, we instead use its subset, the valid toll set in Hearn and Ramana (1998), in formulating the robust toll problem for a proof of concept. With the assumption of strictly monotonic link travel time functions, \( S \) is singleton. The alternative robust pricing problem can be formulated as follows:

\[
\min_{(\tau, \rho)} \psi(\tau)
\]

s.t.
\[
\nu_{SO}^T \left( t(\nu_{SO}) + \tau \right) + \sum_w D_w^T \rho_w = 0
\]
\[
t_{ij}(\nu_{ij}^{SO}) + \tau_{ij} + \rho_i^w - \rho_j^w \geq 0 \quad \forall w \in W, (i, j) \in L
\]

As formulated, RTP-2 is another generalized semi-infinite min-max problem, which is of the same complexity as RTP-1. The inner problems are the same, defining the worst-case tolled BRUE flow distribution, as shown in equation (4.21). RTP-2 incorporates some additional linear constraints (4.26)-(4.27) in its outer problem, which are separable from its inner problem. The linear constraints define the valid toll set (Hearn and Ramana, 1998), and ensure that SO flow satisfies the tolled PRUE conditions with \( \tau \) being the toll vector and \( \rho \) the corresponding node potentials. Consequently, SO flow will be a tolled BRUE flow distribution induced by the toll vector \( \tau \), i.e., the toll achieves SO at its best-case performance.
4.3.2 Solution Algorithm

We propose a heuristic procedure to solve RTP-1 and RTP-2. The key idea is to use a differentiable penalty function to remove the constraints of the inner problem involving the decision variables of the outer problem, thereby transforming a generalized semi-infinite min-max problem into an ordinary semi-infinite optimization problem. The procedure then applies a cutting-plane scheme (e.g., Lawphongpanich and Hearn, 2004) to solve a sequence of ordinary finite optimization problems, each one better approximating the original problem than its predecessors. In the following, we use solving RTP-1 as an example to describe the procedure.

We formulate a penalized inner problem as follows:

\[
P_{WS}(\tau) : \max_{(v,x,d,\pi,\mu)} t(v)^T v - M \sum_{w} \sum_{j} (t_{ij} + \tau_{ij} - \tau_{ij}^w + \pi_i^w - \pi_j^w)^2
\]

s.t. Conditions (4.11)-(4.15)

where \(M\) is a sufficiently-large penalty parameter. For convenience, we further denote the feasible region of the above problem as \(\Omega\) and the objective function as \(\hat{\phi}(\xi,\tau)\). With a finite penalty, \(P_{WS}(\tau)\) provides an upper bound to the original inner problem. Given a toll vector \(\tau\), assume that \(\xi\) is the optimal solution to the original problem RTP-1-IN, and let \(\hat{\xi}\) denote the optimal solution to the above penalized problem for some \(M > 0\). It follows that

\[
\phi(\xi) = \hat{\phi}(\xi,\tau) \leq \hat{\phi}(\hat{\xi},\tau) .
\]

The first equality is due to that no penalty is associated with \(\xi\) and the second inequality is due to the fact that \(\hat{\xi}\) is feasible to the penalized problem. As the penalty coefficient \(M\) goes to infinity, \(P_{WS}(\tau)\) reduces to the original inner problem.

We then formulate a penalized version of robust toll problem RTP-1 as

\[
\min_{0 \leq r \leq r_{max}} \max_{\xi \in \Omega} \hat{\phi}(\xi,\tau).
\]

Because the objective is to minimize the upper bound of total system travel time, the solution may reduce the worst-case system travel time as the upper bound is pretty tight when the penalty
Note that the penalized robust optimization problem is an ordinary semi-infinite min-max problem because the feasible region of its inner problem doesn’t depend on the decision variables of the outer problem. Moreover, it is equivalent to the following ordinary semi-infinite optimization problem:

\[
\begin{align*}
\text{P-RTP-1} & \quad \min_{(\tau, \sigma)} \quad \sigma \\
\text{s.t.} & \quad \sigma \geq \phi(\xi, \tau) \\
& \quad 0 \leq \tau \leq \tau_{\max} \\
& \quad \forall \xi \in \Omega
\end{align*}
\]

We now discuss a cutting-plane scheme to solve P-RTP-1. Assume that \( \xi^1, \xi^2, \ldots, \xi^n \) are elements of \( \Omega \). Then, a relaxed penalized robust toll problem (RP-RTP-1) can be written as:

\[
\begin{align*}
\text{RP-RTP-1} & \quad \min_{(\tau, \sigma)} \quad \sigma \\
\text{s.t.} & \quad \sigma \geq \phi(\xi^i, \tau) \\
& \quad 0 \leq \tau \leq \tau_{\max} \\
& \quad \forall i = 1, \ldots, n
\end{align*}
\]

The RP-RTP-1 problem stated above is simply the original P-RTP-1 problem with \( \Omega \) approximated by the discrete set \( \Omega = \{ \xi^1, \ldots, \xi^n \} \). Let \( (\tilde{\tau}, \tilde{\sigma}) \) be a global optimal solution to the above relaxed toll problem. If \( (\tilde{\tau}, \tilde{\sigma}) \) is feasible to the original P-RTP-1 problem, it is an optimal solution. To check its feasibility, one may solve P-WS(\( \tilde{\tau} \)), i.e., \( \max_{\xi \in \Omega} \phi(\xi, \tilde{\tau}) \). If \( \tilde{\xi} \), a global optimal solution to P-WS(\( \tilde{\tau} \)) is such that \( \phi(\tilde{\xi}, \tilde{\tau}) \leq \tilde{\sigma} \), then \( (\tilde{\tau}, \tilde{\sigma}) \) is a feasible and optimal solution to P-RTP-1. On the other hand, if \( \phi(\tilde{\xi}, \tilde{\tau}) > \tilde{\sigma} \), then an improved solution may be obtained by solving the RP-RTP-1 problem with an expanded discrete set \( \Omega \cup \{ \tilde{\xi} \} \).

Below is a detailed description of the procedure we outlined above:

- **Step 0**: Solve the WC-BRUE problem without tolls to obtain an initial solution \( \xi^1 \). Set \( n = 1 \) and \( \Omega^1 = \{ \xi^1 \} \).
• **Step 1:** Solve the RP-RTP-1 problem with the discrete set $\Omega^n$ and let $(\tau^n, \sigma^n)$ denote the resulting optimal solution.

• **Step 2:** Solve P-WS($\tau^n$) and let $\zeta^{(n+1)}$ denote the resulting optimal solution.

• **Step 3:** If $\hat{\phi}(\zeta^{(n+1)}, \tau^n) \leq \sigma^n$, stop and $\tau^n$ is an optimal robust pricing scheme. Otherwise, set $\Omega^{(n+1)} = \Omega^n \cup \{\zeta^{(n+1)}\}$ and $n = n + 1$. Go to Step 1.

Assume that both $(\tau^n, \sigma^n)$ and $\zeta^{(n+1)}$ globally solve the RP-RTP and P-WS($\tau^n$) problems in Steps 1 and 2, respectively. If the above algorithm stops at some finite iteration $n$, it is easy to see that $(\tau^n, \sigma^n)$ is an optimal solution to the original P-RTP problem. When the algorithm generates an infinite sequence, any of its subsequential limits is optimal to the P-RTP-1 problem under a set of strong conditions discussed in Yin and Lawphongpanich (2007). Because it is difficult in practice to verify those conditions and ensure global optimality for solving RP-RTP-1 and P-WS($\tau^n$) problems, the above algorithm should be regarded as heuristic in general.

### 4.3.3 Numerical Examples

The above algorithm was implemented using GAMS to solve both RTP-1 and RTP-2 for the nine-node network. We set the value of $M$ to be 10000 based on various preliminary tests and the upper bound for the toll on each link is set as the minimum integer number that is greater than the MC toll. Other settings are the same as in Section 4.2.4. RTP-1 was solved for four scenarios with $\alpha = 0.05, 0.10, 0.15$ and 0.20 while RTP-2 was solved with $\alpha = 0.20$ for comparison purpose.

Since the outer problem is a nonconvex quadratic program and the inner problem is a MPCC, multiple (10 and 4 respectively) initial solutions were used in Step 2 and 3 in order to obtain better local optimal solutions. The computation time thus increased significantly. Although not particularly efficient, the cutting-plane algorithm generated some reasonable solutions. When $\alpha = 0.05$, the robust toll from solving RTP-1 was the same as the MC toll.
(reported in Table 4-4) while different tolls were obtained for the other three scenarios. Table 4-4 reports the robust toll vector for $\alpha = 0.20$, denoted as RT1.

Figure 4-8 compares the worst-case total travel time with robust tolls (RT in the figure), the MC toll and no toll (NT in the figure). When compared to the situation with no toll, the figure shows that robust tolls reduce the worst-case system travel time significantly, e.g., the reduction is approximately 10.2% when $\alpha = 0.20$. When compared to MC tolls, robust tolls have superior worst-case performances when $\alpha$ is large, in particular when $\alpha \geq 0.05$. For smaller $\alpha$, BRUE does not deviate much from PRUE, thereby implying that the worst-case performances of robust and MC tolls are similar.

The solution to RTP-2 with $\alpha = 0.20$ is also reported in Table 4-4, denoted as RT2. Table 4-4 also presents the best- and worst-case system travel times of MC, RT1 and RT2 with $\alpha = 0.20$. From the table, RT2 achieves SO at its best-case performance while reducing the worst-case system travel time from 2793 to 2662, a 4.7% reduction. If the full set of optimistic tolls is considered, a higher reduction rate can be expected. RT2 also dominates MC in the sense that both tolls have the same best-case performances while the former performs better against the worst-case scenarios.

The parameter $\alpha$ reflects the level of rationality in travelers’ route choice decision making, which may be estimated by conducting a stated preference survey. However, given the fact that the estimate of $\alpha$ is likely to be biased or inaccurate, it makes sense to examine how a robust toll designed for one particular level of rationality performs at other levels of rationality. In Figure 4-9, tolls from RT1 with $\alpha = 0.20$ reduces both the best- and worst-case system travel times significantly for all $\alpha$ values compared with the no-toll case. Compared with the MC toll, the same toll leads to a superior worst-case performance with $\alpha \geq 0.08$ but inferior otherwise.
Intuitively, the smaller $\alpha$ is, the more rational travelers’ route choices are. Thus the first-best tolls are expected to perform well with small $\alpha$ values.

We further note that although robust tolls are designed to guard against the worst-case scenarios, implementation of robust tolls actually leads to a more stable system performance in the sense that the span between the best and worst performances is much narrower. Figure 4-10 displays the difference between the worse- and best-case travel time as a percentage (or fraction) of the latter. In all three cases (no toll, MC and RT1 tolls), the difference increases as $\alpha$ increases. However, the difference for the case with the robust toll (see the bottom most graph) is significantly smaller than the other two for large $\alpha$. This suggests that there is less variability in travel time under robust tolls, i.e., the tolls are more effective and reliable at inducing travelers to use the road network more efficiently, especially when the threshold value is relatively large.

4.4 Summary

This chapter investigates static or time-of-day robust network-level pricing problem with bounded rationality in travelers’ route choices. As an alternative source of network uncertainties, BRUE is not fully explored in previous literatures. This chapter enriches the literature by providing a rigorous model of BRUE, a discussion on the implication of BRUE and a robust pricing scheme with BRUE. More specifically, a general path-based definition of BRUE is presented, followed by a more restrictive link-based definition. The BRUE flow distributions are characterized as non-empty and usually non-convex sets. Robust pricing models have then been formulated to minimize the maximum system travel time realized from the tolled BRUE set. Numerical experiments demonstrate that system performance may vary substantially within the BRUE set and the proposed robust pricing models are able to protect against the worst-case scenario effectively.
Our future research will develop more efficient solution approaches, e.g., based on meta-heuristic algorithms. Moreover, path-based algorithms may be developed to solve the path-based counterparts of the proposed formulations. We will further investigate how to price under other more realistic network settings with boundedly rational behaviors, such as elastic demand or combined route and departure time choices.
Figure 4-1. A bridge network
Figure 4-2. Relationship between three different flow sets
Figure 4-3. The nine-node network
Figure 4-4. Comparison of system performances of BRUE at the nine-node network
Figure 4-5. Comparison of system performances of BRUE at the Sioux Falls network
Figure 4-6. System performances with MC pricing scheme at the nine-node network
Figure 4-7. Three-parallel-link network

\[ t_1 = 2v_1 \]
\[ t_2 = 5 \]
\[ t_3 = v_3 + 2 \]
Figure 4-8. Comparison of worst-case total travel times with no toll, robust and MC tolls at the nine-node network
Figure 4-9. System performances with MC and RT1 at the nine-node network (1)
Figure 4-10. System performances with MC and RT1 at the nine-node network (2)
Table 4-1. Illustrative example of the restrictiveness of the link-based BRUE representation

<table>
<thead>
<tr>
<th>Link</th>
<th>Travel time function</th>
<th>Flow</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 3)</td>
<td>( t_{13} = 3 + v_{13} )</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>( t_{12} = 7 + v_{12} )</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>(2, 3)</td>
<td>( t_{23} = v_{23} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(2, 4)</td>
<td>( t_{24} = 5 + v_{24} )</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>(3, 2)</td>
<td>( t_{32} = v_{32} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(3, 4)</td>
<td>( t_{34} = 2 + v_{34} )</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Path

<table>
<thead>
<tr>
<th>Path</th>
<th>Flow</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3-4</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>1-3-2-4</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1-2-3-4</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>1-2-4</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 4-2. Nonconvexity of the BRUE flow distribution set for the bridge network

<table>
<thead>
<tr>
<th>Link</th>
<th>Travel time function</th>
<th>BRUE-1</th>
<th>BRUE-2</th>
<th>0.5×(BRUE-1+BRUE-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Flow</td>
<td>Time</td>
<td>Flow</td>
</tr>
<tr>
<td>(1, 3)</td>
<td>( t_{13} = 10v_{13} )</td>
<td>3</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>( t_{12} = 50 + v_{12} )</td>
<td>3</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>(2, 3)</td>
<td>( t_{23} = v_{23} )</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>(2, 4)</td>
<td>( t_{24} = 10v_{24} )</td>
<td>3</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>(3, 2)</td>
<td>( t_{32} = 10 + v_{32} )</td>
<td>0</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>(3, 4)</td>
<td>( t_{34} = 50 + v_{34} )</td>
<td>3</td>
<td>53</td>
<td>2</td>
</tr>
</tbody>
</table>

Path

<table>
<thead>
<tr>
<th>Path</th>
<th>Flow</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3-4</td>
<td>3</td>
<td>83</td>
</tr>
<tr>
<td>1-3-2-4</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>1-2-3-4</td>
<td>0</td>
<td>106</td>
</tr>
<tr>
<td>1-2-4</td>
<td>3</td>
<td>83</td>
</tr>
<tr>
<td>1-2-4</td>
<td>3</td>
<td>83</td>
</tr>
<tr>
<td>1-3-2-4</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>1-2-3-4</td>
<td>0</td>
<td>106</td>
</tr>
<tr>
<td>1-2-4</td>
<td>3</td>
<td>83</td>
</tr>
</tbody>
</table>
Table 4-3. Flow distributions for the nine-node network ($\alpha = 0.10$)

<table>
<thead>
<tr>
<th>Link</th>
<th>UE</th>
<th>WC-BRUE</th>
<th>BC-BRUE</th>
<th>SO</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 5)</td>
<td>8.2</td>
<td>4.0</td>
<td>11.5</td>
<td>9.4</td>
</tr>
<tr>
<td>(1, 6)</td>
<td>21.8</td>
<td>26.0</td>
<td>18.5</td>
<td>20.6</td>
</tr>
<tr>
<td>(2, 5)</td>
<td>47.4</td>
<td>53.2</td>
<td>43.0</td>
<td>38.3</td>
</tr>
<tr>
<td>(2, 6)</td>
<td>22.6</td>
<td>16.8</td>
<td>27.0</td>
<td>31.7</td>
</tr>
<tr>
<td>(5, 6)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(5, 7)</td>
<td>27.8</td>
<td>29.2</td>
<td>26.4</td>
<td>21.3</td>
</tr>
<tr>
<td>(5, 9)</td>
<td>27.7</td>
<td>28.0</td>
<td>28.0</td>
<td>26.4</td>
</tr>
<tr>
<td>(6, 5)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(6, 8)</td>
<td>44.5</td>
<td>42.8</td>
<td>45.6</td>
<td>39.5</td>
</tr>
<tr>
<td>(6, 9)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>12.8</td>
</tr>
<tr>
<td>(7, 3)</td>
<td>38.2</td>
<td>34.0</td>
<td>31.7</td>
<td>29.6</td>
</tr>
<tr>
<td>(7, 4)</td>
<td>17.4</td>
<td>23.2</td>
<td>22.8</td>
<td>20.8</td>
</tr>
<tr>
<td>(7, 8)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(8, 3)</td>
<td>1.8</td>
<td>6.0</td>
<td>8.3</td>
<td>10.4</td>
</tr>
<tr>
<td>(8, 4)</td>
<td>42.6</td>
<td>36.8</td>
<td>37.2</td>
<td>39.2</td>
</tr>
<tr>
<td>(8, 7)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(9, 7)</td>
<td>27.7</td>
<td>28.0</td>
<td>28.0</td>
<td>29.1</td>
</tr>
<tr>
<td>(9, 8)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>10.2</td>
</tr>
<tr>
<td>Link</td>
<td>No Toll</td>
<td>MC Toll</td>
<td>RT1</td>
<td>RT2</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>(1, 5)</td>
<td>-</td>
<td>1.13</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td>(1, 6)</td>
<td>-</td>
<td>6.16</td>
<td>4.43</td>
<td>6.00</td>
</tr>
<tr>
<td>(2, 5)</td>
<td>-</td>
<td>2.59</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>(2, 6)</td>
<td>-</td>
<td>3.62</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>(5, 6)</td>
<td>-</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(5, 7)</td>
<td>-</td>
<td>16.88</td>
<td>17.00</td>
<td>15.81</td>
</tr>
<tr>
<td>(5, 9)</td>
<td>-</td>
<td>5.13</td>
<td>4.37</td>
<td>4.61</td>
</tr>
<tr>
<td>(6, 5)</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(6, 8)</td>
<td>-</td>
<td>7.37</td>
<td>2.42</td>
<td>7.81</td>
</tr>
<tr>
<td>(6, 9)</td>
<td>-</td>
<td>0.11</td>
<td>1.00</td>
<td>0.61</td>
</tr>
<tr>
<td>(7, 3)</td>
<td>-</td>
<td>3.54</td>
<td>0.00</td>
<td>4.00</td>
</tr>
<tr>
<td>(7, 4)</td>
<td>-</td>
<td>2.01</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>(7, 8)</td>
<td>-</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(8, 3)</td>
<td>-</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>(8, 4)</td>
<td>-</td>
<td>2.50</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>(8, 7)</td>
<td>-</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(9, 7)</td>
<td>-</td>
<td>3.75</td>
<td>4.00</td>
<td>3.20</td>
</tr>
<tr>
<td>(9, 8)</td>
<td>-</td>
<td>0.06</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Best-case system travel time</td>
<td>2309</td>
<td>2254</td>
<td>2306</td>
<td>2254</td>
</tr>
<tr>
<td>Worst-case system travel time</td>
<td>2793</td>
<td>2735</td>
<td>2509</td>
<td>2662</td>
</tr>
</tbody>
</table>
CHAPTER 5
OPTIMAL ADAPTIVE TOLLING STRATEGIES FOR HIGH OCCUPANCY/TOLL LANES: BASIC MODELS

This chapter is devoted to the facility-level pricing problem. Different from network-wide pricing, it is found that travelers are able to respond to dynamic tolls charged at some critical facilities, e.g., HOT facilities (Bonsall et al., 2007). Literature review in Chapter 2 also reveals that several agencies in the U.S. have applied traffic-responsive tolls on their HOT facilities. Although empirical studies shows that adaptive tolling outperforms fixed time-of-day tolling on HOT facilities (e.g., Burris and Sullivan, 2006, Sullivan and Burris, 2006, Supernak et al., 2003a, b), most of the adaptive tolling schemes implemented are simple and heuristic, a fact that provides many opportunities for improvements. This chapter aims to enhance adaptive pricing for HOT lanes such that they are more effective and more attractive to both transportation authorities and motorists.

Two sensible and pragmatic approaches are proposed for determining time-varying tolls in response to the detected traffic condition. The first approach adjusts the toll rate based on the concept of the feedback control, while the second approach would ‘learn’ in a sequential fashion both the demand and supply information of the HOT facility and then determine pricing strategies to explicitly achieve the operating objectives. This chapter presents the basic models of these two approaches in Section 5.1 and 5.2 respectively. A simulation experiment is provided in Section 5.3 to demonstrate and compare the proposed approaches. In the prototype model of the self-learning approach presented in this chapter, only one particular component of information learning, namely learning users’ WTP, and a simple point-queue model for traffic dynamics are employed for a proof of concept. A couple of extensions within a unified framework of the self-learning pricing approach will be discussed in the next chapter.
Without loss of generality and to facilitate the presentation of the essential idea, this dissertation considers a simple setting of HOT facilities shown in Figure 5-1 and Figure 5-2 where there are one HOV/HOT lane and one regular lane, and a bottleneck is activated downstream. Moreover, it is a single segment of HOT lane with one entry, and there is no on/off ramp in between. We assume that HOVs have shifted to the HOT lane before the toll gate and each motorist on regular lane must pay to access the toll lane. Also note that starting from this chapter, a set of notations different than that in Chapter 2 and 4 will be adopted.

5.1 Feedback Approach for HOT Pricing

The first approach is based on the concept of feedback control and requires one loop-detector station (or other types of sensors) located downstream of the toll tag reader, as illustrated in Figure 5-1 (in a real-world implementation, several loop-detector stations need be installed downstream in order to better detect the traffic condition along the segment.)

In the feedback control, the toll rate at time interval \( t+1 \) depends on the toll at interval \( t \) and the occupancy detected by the downstream loop detectors. Mathematically the control logic can be stated as follows:

\[
\beta(t+1) = \beta(t) + K \cdot (o(t) - o^*)
\]

where \( \beta(t) \) and \( \beta(t+1) \) are the toll rates at interval \( t \) and \( t+1 \) respectively; \( K \) is a regulator parameter for adjusting the disturbance of the feedback control; \( o^* \) is the desired occupancy of the toll lane, which is typically set equal to or slightly less than the critical occupancy (corresponding to the critical density), and \( o(t) \) is the measured occupancy.

Note that a similar concept has been successfully applied to ramp metering known as ALINEA or Asservissement Lineaire d'Entree Autoroutiere (Papageorgiou et al., 1997). Therefore, we anticipate that this approach will perform reasonably well and thus provide a
benchmark for evaluating the other approach. Moreover, this approach is cost efficient and easy to implement.

5.2 Self-learning Approach for HOT Pricing

5.2.1 General Framework

The objective of HOT lane operation is to maximize the total freeway throughput while maintaining a free-flow condition on the toll lanes. In other words, it is to attract the “just right” amount of paid users to make full use of the HOT lane without making it congested. To do so, information from both the demand and supply side is needed to determine the right amount of toll. The essential idea of the self-learning approach is that in the operation, the information needed can be gradually learned by mining the loop detector data, and the tolls can be determined based on the updated information to better serve the local operation objectives. More specifically, the proposed framework of the self-learning approach consists of two steps:

- **System Inference.** From mining real-time traffic data (such as speed, flow or occupancy) collected at a regular intervals from loop detectors (often at limited locations and inaccurate), the first step learns travelers’ WTP in order to predict motorists’ reaction to tolls, composes a full picture of current traffic condition for the entire freeway, and forecasts short-term traffic demand.

- **Toll Determination.** The attained knowledge will then be used in the second step to maximize the throughput rate for the entire freeway while maintaining a superior service on the toll lanes. Given that future demand is likely uncertain, a stochastic programming approach is adopted to determine robust tolls whose performance is not very sensitive to different realizations of uncertain traffic demand. Moreover, a robust optimization approach can be applied to accommodate users’ heterogeneity.

Figure 5-2 sketches the framework for proactive self-learning dynamic pricing.

In the following sub-sections, we demonstrate the framework of the self-learning approach by building a prototype model which employs the WTP learning component of system inference and a point-queue model in describing the traffic dynamics for toll determination.
5.2.2 Calibration of Willingness-to-Pay

Understanding motorists’ WTP is critical for toll determination. While stated-preference surveys often overestimate motorists’ WTP, the data collected during the tolling operation reflect actual lane choice behaviors when travelers are facing the tolls being charged. Such revealed-preference information can be used to estimate more accurately motorists’ WTP. Assuming homogeneous motorists with the same WTP whose decision on whether to pay to gain access to the HOT lanes follows a logit model, the relationship between the approaching flow rates and the flows on HOT and regular lanes can be stated as follows:

\[
\frac{\lambda_T(t) - \mu_T(t)}{\mu_R(t)} = \frac{1}{1 + \exp(\alpha_1(c_T(t) - c_R(t)) + \alpha_2\beta(t) + \gamma)}
\]

(5.1)

where \(\mu_T(t)\) and \(\mu_R(t)\) represent the approaching flow rates on HOV and regular lanes during time interval \(t\) respectively; \(\lambda_T(t)\) and \(\lambda_R(t)\) are the flow rates after the lane choice, and \(c_T(t)\) and \(c_R(t)\) are the (average) travel times on HOT and regular lanes at time interval \(t\). In equation (5.1), there are three parameters to be estimated: \(\alpha_1\), \(\alpha_2\) and \(\gamma\), where \(\alpha_1\) and \(\alpha_2\) indicate respectively the marginal effect of travel time and toll on motorists’ utility, and \(\gamma\) encapsulates other factors affecting motorists’ WTP. Note that \(\alpha_1/\alpha_2\) represents motorists’ trade-off between time savings and tolls, i.e., the value of travel time. Other variables in equation (5.1) can either be obtained directly from loop detectors or estimated using traffic flow models, and the toll rate \(\beta(t)\) is set by the operator. It should be pointed out that here we use volume splits \((\lambda_T(t) - \mu_T(t))/\mu_R(t)\) to approximate the probabilities of lane choices.

In real-time operation, a recursive least-squares technique or discrete Kalman filtering (KF) (Kalman, 1960) can be used to estimate the constant parameters, \(\alpha_1\), \(\alpha_2\) and \(\gamma\). To do so, equation (5.1) can be reformulated as follows:
\[
\ln \left( \frac{\mu_R(t)}{\lambda^*(t) - \mu_R(t)} - 1 \right) = \alpha_1 (c_T(t) - c_R(t)) + \alpha_2 \beta(t) + \gamma \quad (5.2)
\]

Applying discrete KF technique to estimate the parameters, we have:

\[
\begin{bmatrix}
\hat{\alpha}_1(t + 1) \\
\hat{\alpha}_2(t + 1) \\
\hat{\gamma}(t + 1)
\end{bmatrix} = \begin{bmatrix}
\hat{\alpha}_1(t) \\
\hat{\alpha}_2(t) \\
\hat{\gamma}(t)
\end{bmatrix} + G_{WTP}(t) \begin{bmatrix}
\ln \left( \frac{\mu_R(t)}{\lambda^*(t) - \mu_R(t)} - 1 \right) - (c_T(t) - c_R(t)) \\
\beta(t) \\
1
\end{bmatrix} \begin{bmatrix}
\hat{\alpha}_1(t) \\
\hat{\alpha}_2(t) \\
\hat{\gamma}(t)
\end{bmatrix}
\]

\[
G_{WTP}(t) = P_{WTP}(t-1) \begin{bmatrix}
c_T(t) - c_R(t) \\
\beta(t) \\
1
\end{bmatrix} \begin{bmatrix}
(c_T(t) - c_R(t)) \\
\beta(t) \\
1
\end{bmatrix} + \sigma_{WTP}^{-1}
\]

\[
P_{WTP}(t) = \left[ I - G_{WTP}(t)(c_T(t) - c_R(t)) \beta(t) \right] P_{WTP}(t-1)
\]

where the variables with "\(^\wedge\)" are estimates. \(\sigma_{WTP}\) is the variance of the left-hand side of equation (5.2) computed from the known variance of random detector measurement error (the value could vary with different types of sensors), \(P_{WTP}\) is the expected covariance matrix of the estimation errors and \(G_{WTP}\) is the Kalman gain. With an initialization of \(\alpha_1(0), \alpha_2(0), \gamma(0)\) and \(P_{WTP}(0)\), equation (5.3) can update estimates of \(\alpha_1, \alpha_2\) and \(\gamma\) real time with newly-obtained information. As time evolves, the impact of the initialization will be diminishing, and accurate estimates of \(\alpha_1, \alpha_2\) and \(\gamma\) are expected.

### 5.2.3 Optimal Tolling Strategy with Point-Queue Model

The calibrated motorists’ WTP at interval \(t\) provides a basis for determining a toll for interval \(t+1\). Consistent with the prevailing operation policies of HOT lanes, we now attempt to specify toll rates to maximize the throughput of the corridor while ensuring a superior traffic condition on the toll lane.

Assume in rush hours when the total demand is much higher than the capacity of the freeway segment, the queuing delay dominates other factors that may affect the travel time of the freeway segment, such as lane-changing maneuvers. Therefore, we adopt a simple point-queue
model to describe traffic dynamics in this prototype model of the self-learning approach. The objectives of maintaining free-flow condition on the HOT lanes and maximizing the throughput of the entire freeway are approximately equivalent to operating the HOT lane at a throughput close to its capacity while keeping it from being congested. This is because general purpose lanes are assumed to be congested and their discharge rates are equal to the capacity of the downstream bottleneck. Consequently, based on the measured condition on the managed lane at interval $t$, the desired inflow at interval $t+1$, denoted as $\tilde{\lambda}_r(t+1)$, can be determined such that the managed lane is uncongested but efficiently utilized.

With homogeneous motorists, it is straightforward to determine the toll rate for time interval $t+1$ by using the following relationship:

$$
\beta(t+1) = \frac{\ln \left( \frac{\mu_r(t+1) - \tilde{\lambda}_r(t+1)}{\tilde{\lambda}_r(t+1)} \right) - \alpha_t(t+1)(c_r(t+1) - c_R(t+1)) - \gamma(t+1)}{\alpha_s(t+1)}
$$

where $c_r(t+1)$ and $c_R(t+1)$ are estimated using traffic flow models, and $\mu_r(t+1)$ is obtained from loop detectors.

The above toll determination is flexible in incorporating various models of flow dynamics to estimate travel times as long as lane-changing behavior is neglectable. For the purpose of the proof of concept, we employ the point-queue concept (Kuwahara and Akamatsu, 1997) to describe traffic dynamics, determining $\tilde{\lambda}_r(t+1)$ and calculating $c_r(t+1)$ and $c_R(t+1)$ at each time interval.

It is assumed that the average travel time during time interval $t$ can be decomposed into a flow-independent cruise time and an average queuing delay at the end of the segment, written as:

$$
c(t) = c_o + \frac{v(t-1) + v(t)}{2s}
$$

(5.5)
where \( c_0 \) is the flow-independent cruise time, and \( v(t-1) \) and \( v(t) \) are the numbers of the queuing vehicles by the end of the time interval \( t-1 \) and \( t \) respectively and \( s \) is the capacity of the bottleneck.

The evolution of queue length can be captured by the following equation:

\[
v(t) = \max\{(v(t-1) + (\lambda(t) - s) \cdot \Delta t), 0\}
\]

(5.6)

where \( \Delta t \) is the duration of the time interval. Huang and Lam (2002) proved that equation (5.6) is consistent with the first-in-first-out discipline, which ensures that a vehicle must leave the segment in the same order as it arrived at that segment.

Equations (5.5) and (5.6) apply to both the managed and regular lanes. Again, at current interval \( t \), \( \lambda_R(t) \) and \( \lambda_T(t) \) are obtained directly from loop detectors, and then \( v(t) \) can be estimated from (5.6). Consequently, \( c_T(t) \) and \( c_R(t) \) are estimated from equation (5.5). Up to this point, we are ready to apply equation (5.3) to estimate the motorists’ willingness to pay.

To maintain a superior free-flow traffic condition on the managed lane, the desired inflow \( \tilde{\lambda}_T(t+1) \) should be regulated as follows:

\[
\tilde{\lambda}_T(t+1) = \begin{cases} 
  s_T & v_T(t) = 0 \\
  s_T - v_T(t)/\Delta t & v_T(t)/s_T \leq \Delta t \\
  0 & \text{Others}
\end{cases}
\]

(5.7)

Note that the case of \( \tilde{\lambda}_T(t+1) = 0 \) implies that the managed lane is already congested and the toll rate will be set to its predetermined maximum value. Given \( \tilde{\lambda}_T(t+1) \) and the estimated willingness to pay, the toll can be determined according to equation (5.4).
5.3 Simulation Study

5.3.1 Design of Simulation Experiments

Given a test environment on real facilities is not readily available, we conducted simulation experiments to validate and compare the proposed approaches. The developed simulation platform consists of three major components: simulator, monitor and controller. The simulator attempts to replicate the motorists’ lane choice behaviors and traffic dynamics. The monitor serves as a surveillance system, collecting information of traffic condition at each interval including total arrival, flow rates before and after lane choices, and travel times. The controller implements the operations of the proposed approaches, namely the feedback and the self-learning approaches.

- **Simulator.** A Poisson arrival is generated based on the given average arrival flow rate. At each interval, based on the toll rate specified by the controller and the instantaneous travel times from the monitor, the simulator applies the Logit model with the true values of $\alpha_1, \alpha_2$ and $\gamma$ to compute the percentages of the motorists for choosing the HOT lane. Moreover, the simulator applies traffic flow models to replicate traffic dynamics. With application of point-queue model, equations (5.5) and (5.6) are adopted to track the queue evolution and estimate the travel times on both lanes.

- **Monitor.** At each time interval, the monitor collects the total arrival, flows entering the managed and general purpose lanes, the travel conditions and travel times generated from the simulator. In a real-world implementation, this information could be directly measured, although not necessarily adequate or accurate. In this case, the monitor will generate perturbed measurements in the simulation to mimic the real-world situation, and it is then important to estimate the actual traffic conditions through the partial or inaccurate measurements. Traffic state estimation will be discussed in Chapter 6.

- **Controller.** The controller implements the operations of the proposed approaches. For the feedback controller, the occupancy is not available in the simulation when the point-queue model is used. Instead, the inflow rate of HOT lane will be used to adjust the toll rate. More specifically, $\beta(t+1) = \beta(t) + K(\lambda_r(t) - \bar{\lambda}_r(t))$. For the self-learning approach, the controller first uses information of flows and travel times from the monitor to calibrate the value of $\alpha_1, \alpha_2$ and $\gamma$ according to equation (5.3). Toll rate is then determined based on equations (5.4)-(5.7) with point-queue model.
5.3.2 Comparison of Feedback and Self-learning Controllers

This sub-section examines the performances of the feedback and the self-learning controllers, and compares them under the same settings. Note that as described above, the simulator mimics individual motorist’s lane choice behavior based on the Logit model. Since actual lane-choice behaviors are much more complicated, the simulator may not replicate the real-world situation. Moreover, since the controller and the simulator adopt the same behavior assumption, i.e. the simulator represent the actual traffic dynamics accurately, the performance of the self-learning approach may be overestimated. However, there seems no better way unless we do a real-world experiment or a “human-in-the-loop” simulation.

The downstream bottleneck has a capacity of 1800 vehicles per hour per lane. We simulate a peak 20-minutes period and the duration is equally divided into discrete time intervals with two minutes each. For the computation simplicity, we assume in the Logit model (5.1) that $\alpha_2$ is known as one. Hence, only two parameters are left to be estimated for motorists’ utility of lane choice. The true values of $\alpha_1$ and $\gamma$ used in the simulator are 1 and 0.2. We represent the travel time difference in the unit of one time interval (2 minutes), hence, $\alpha_1/\alpha_2 = 1$ means a value of travel time of one dollar per two minutes, i.e., 30 dollars per hour. $\alpha_1(0)$, $\gamma(0)$ and $P(0)$ are initialized as 0.8, 0.1 and an identity matrix respectively. The upper bound of the toll rate is set to eight dollars. A variety of demand scenarios are created to test the performance of the two controllers. In the following, we report the simulation results for two selected scenarios. Scenario 1 assumes an average arrival rate of 3600 vehicles per hour for the regular lane and 300 vehicles per hour for the HOT lane throughout the simulated ten intervals, and scenario 2 has arrival rates of 3600 and 3000 vehicles per hour at first and second five intervals respectively.
5.3.2.1 Performance of feedback controller

We first examine the performance of the feedback controller with homogenous motorists. Figure 5-3 displays the time-varying throughputs and queues on both the HOT and regular lanes, resulted by the feedback controller with $K=0.1$ under Scenario 1. Since the arrival flow is far beyond the capacity of the bottleneck, it is inevitable that queues constantly increase on the regular lane. At the same time, the feedback controller is able to adjust the toll to avoid severe queuing on the managed lane as well as maintain a fairly high and stable throughput. In fact, the average throughput for the corridor is 1642 vehicles per hour per lane, and the average queue length on the managed lane is 7.7 vehicles, a descent performance for operating the managed lane.

Figure 5-4 illustrates the performance of the same controller under demand scenario 2. Although the average total arrival (both HOV and lower-occupancy-vehicle flows) during the simulation period is equal to the capacity of the bottleneck, the queues can not be cleared due to the random arrival. Again, the controller is able to achieve the operating objectives fairly well, resulting in an average throughput of 1602 vehicles per hour per lane and an average queue length on the managed lane of 5.27 vehicles.

The sensitivity and performance of the feedback controller may depend on the setting of the regulator parameter $K$. By adjusting the value of $K$, we examine the dependency under demand scenario 1. For each $K$, 30 runs (a sufficient sample size determined from preliminary ten runs with an error tolerance of one) of simulation are conducted and Table 5-1 reports the average queue length and average throughput of the managed lane. Within our expectation, when $K$ is small, the toll adjustment may not be adequate enough to adapt to the traffic condition, thereby leading to severe queuing on the managed lane but a higher throughput (partially due to the use of the point-queue model). When $K$ gets larger, we see a reverse tendency with fewer
queues and less throughput. However, a large $K$ does not necessarily lead to fewer queues on the managed lane. In fact, in our simulation experiments with this particular demand scenario, the controller with $K=0.12$ outperforms the others with $K>0.12$ in both operating objectives.

5.3.2.2 Performance of self-learning controller with point-queue model

We then examine the performance of the self-learning controller under the same setting. Figure 5-5 is a proof of concept that motorists’ WTP can be gradually learned. The figure presents the estimates of parameters $\alpha_1$ and $\gamma$ with homogenous motorists. As time evolves, the estimates converge from the initial values of 0.8 for $\alpha_1$ and 0.1 for $\gamma$ to the true values of 1 and 0.2 respectively.

Figure 5-6 presents the performance of the self-learning controller under scenario 1, a counterpart to Figure 5-3. The self-learning controller results in an average throughput of 1764 vehicles per hour per lane and a queue length of 4.1. Compared with the feedback controller, it reduces the queue on the managed lane by 47% and increases the throughput by 7%. To further visualize the performance difference, Figure 5-7 plots the resulting average throughput and queue length over ten runs (sample size determined similarly as for the feedback controller) by the self-learning controller versus averages over 30 runs by the feedback controller with different values of $K$. The latter forms a frontier in terms of the two control objectives, and the self-learning controller actually outperforms the feedback controller with $K \geq 0.05$. When $K<0.05$, the feedback controller results in higher throughput but longer queues compared to the self-learning controller.

Figure 5-8 further reports the simulation results of the self-learning controller under scenario 2, a counterpart to Figure 5-4. The average throughput is 1773 vehicles per hour per
lane and the average queue length is 2.8 vehicles on the managed lane, superior to the performance of the feedback controller.

In summary, the self-learning controller is able to achieve a high throughput as well as prevent the managed lane from being congested. In view of the priority that the HOT lane operator put is to maintain a free-flow traffic condition, we conclude that the self-learning controller generally outperforms the feedback controller.

5.4 Summary

This chapter has investigated the facility-level pricing problem particularly for HOT lane operations. Two approaches, feedback and self-learning, are proposed to determine dynamic traffic-responsive pricing strategies for HOT lanes to provide superior free-flow travel services and efficiently utilize the capacity. The proposed approaches are feasible and easy to implement. Simulation experiments in this chapter have validated both models and verified their effectiveness. It is concluded that the prototype model of the self-learning controller outperforms the feedback controller. However, the latter could be more cost efficient, and easier to implement.

The basic model of the self-learning approach developed in this chapter has simplified some practical concerns in HOT lane operation. It only employs the WTP-learning component of system inference, while assuming the true traffic condition along the freeway and the future inflow are known. In the second step of toll determination, the underlying point-queue model does not capture the complicated characteristics of traffic flow caused by lane-changing behaviors. Also, the assumption that travelers are homogeneous is not very realistic. Another shortcoming from which the basic models of both the feedback and the self-learning approaches suffer is that the toll determination only focuses on the time interval $t+1$ and responds reactively to measured flows from interval $t$. Although the controllers are able to achieve high throughput
rates, they may lead to a wildly fluctuating toll pattern that frustrates motorists and cause unstable traffic condition on the regular lanes.

In Chapter 6, we will discuss a couple of possibilities for improving the self-learning approach. Other components in its general framework will be developed, including more realistic description of traffic dynamics, integrated traffic state estimation and WTP learning, and proactive robust toll optimization in coupled with demand prediction. These enhancements addressing several critical practical issues will advance the self-learning approach to be more realistic and efficient.
Figure 5-1. System configuration for the feedback-control approach
Figure 5-2. System sketch of the self-learning approach
Figure 5-3. Throughputs, queues and toll rates resulted by the feedback controller (Scenario 1, $K=0.1$)
Figure 5-4. Throughput, queues and toll rates resulted by the feedback controller (Scenario 2, \( K=0.1 \))
Figure 5-5. Calibrated $\alpha_1$ & $\gamma$ by the self-learning controller (Scenario 1)
Figure 5-6. Throughputs, queues and toll rates resulted by the self-learning controller (Scenario 1)
Figure 5-7. Performance comparison of two controllers under scenario 1
Figure 5-8. Throughputs, queues and toll rates resulted by the self-learning controller (Scenario 2)
Table 5-1. Performance of feedback controllers with different $K$ values

<table>
<thead>
<tr>
<th>$K$</th>
<th>Avg. Queue Length ( # vehicle)</th>
<th>Throughput (vehicle per hour per lane)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>29.47</td>
<td>1776.78</td>
</tr>
<tr>
<td>0.03</td>
<td>21.81</td>
<td>1773.27</td>
</tr>
<tr>
<td>0.05</td>
<td>14.64</td>
<td>1756.56</td>
</tr>
<tr>
<td>0.10</td>
<td>8.33</td>
<td>1659.99</td>
</tr>
<tr>
<td>0.11</td>
<td>7.96</td>
<td>1635.21</td>
</tr>
<tr>
<td>0.12</td>
<td>7.86</td>
<td>1632.81</td>
</tr>
<tr>
<td>0.13</td>
<td>8.07</td>
<td>1608.39</td>
</tr>
<tr>
<td>0.14</td>
<td>7.87</td>
<td>1608.12</td>
</tr>
<tr>
<td>0.15</td>
<td>8.12</td>
<td>1592.64</td>
</tr>
<tr>
<td>0.16</td>
<td>8.04</td>
<td>1592.37</td>
</tr>
<tr>
<td>0.17</td>
<td>8.43</td>
<td>1584.57</td>
</tr>
<tr>
<td>0.18</td>
<td>8.73</td>
<td>1569.63</td>
</tr>
<tr>
<td>0.19</td>
<td>9.39</td>
<td>1554.48</td>
</tr>
<tr>
<td>0.20</td>
<td>10.68</td>
<td>1536.45</td>
</tr>
</tbody>
</table>
CHAPTER 6
OPTIMAL ADAPTIVE TOLLING STRATEGIES FOR HIGH OCCUPANCY/TOLL LANES:
A UNIFIED FRAMEWORK OF EXTENDED SELF-LEARNING APPROACH

Two approaches, one feedback and one self-learning, have been presented in Chapter 5 for traffic-responsive pricing on HOT lanes. Between those two, the self-learning approach takes into consideration many human and traffic factors that are ignored by the simple feedback logic. Chapter 5 has demonstrated that the prototype reactive self-learning dynamic pricing approach is able to learn motorists’ WTP in a sequential fashion and explicitly optimizes the toll rate for the next rolling horizon. In this chapter, we will further advance this promising method to be a well-polished and readily-implementable approach.

The general framework of the self-learning approach consists of two steps, system inference and toll determination (see Section 5.2.1). The first step reveals the information from both demand and supply side by mining the real-time traffic data, and the second step adjusts the toll rate based on the updated information. Chapter 5 focuses on learning users’ WTP and reactive toll determination. Within the framework, we will further address several critical issues of the self-learning approach in this chapter, including more realistic description of the traffic dynamics, integrated traffic state estimation and WTP learning, and proactive robust toll optimization in coupled with demand learning.

Section 6.1 explores the possibility of incorporating more realistic traffic dynamics into the unified self-learning approach framework. Impacts of the lane-changing behaviors and the physical design of the HOT lane slip ramp are considered. Two other components of the system-inference step in the framework, the traffic state estimation and the demand prediction, are investigated in Section 6.2. Section 6.3 further presents several improvements of the toll-determination step in the framework, including a proactive pricing approach to provide smoothly changing toll rates, a robust approach to accommodate users’ heterogeneity, and a stochastic
method to incorporate demand uncertainty. A simulation study is conducted in Section 6.4 to demonstrate that the extended framework is feasible, effective and efficient. Finally, Section 6.5 summarizes this chapter.

6.1 Incorporating More Realistic Traffic Dynamics

In Chapter 5, the point-queue model is adopted to describe the traffic dynamics. However, the true characteristics of the traffic flow are more complicated due to the lane-changing behaviors of the lower-occupancy vehicles. These vehicles shifting to the HOT lane from regular lanes usually have a velocity less than that of the prevailing traffic on the HOT lane, and they are referred to as moving bottlenecks as they may cause queues to form behind them. Impact of moving bottlenecks and lane-changing behaviors on traffic conditions has attracted much attention in the literature (see, e.g., Newell, 1998 and Laval and Daganzo, 2006). In order to describe the traffic dynamics more realistically, a multi-lane hybrid traffic flow model will be incorporated in the framework to improve the toll determination. Moreover, the physical design of the HOT lane has a non-neglectable impact on the traffic characteristics as well. To take into account this factor, alternative HOT lane slip ramp design will be modeled. The enhanced traffic flow model serves as a basis for both the system inference and toll determination in the framework of the self-learning approach. The details will be discussed in the Sections 6.2 and 6.3.

6.1.1 Impacts of Lane-Changing Behaviors

Motivated by the postulate that the lane-changing behaviors before the entry points to the HOT lane may create voids in traffic streams and reduce the throughput of the lane, we adopt the multi-lane hybrid cell transmission model (CTM) proposed by Laval and Daganzo (2006) for a more realistic representation of traffic dynamics. The multi-lane hybrid CTM is able to capture both the mandatory and the discretionary lane-changing behaviors, and to treat lane changing
vehicles as moving bottlenecks. This way, the impacts of the lane-changing behaviors on the
freeway throughput and travel time can be explicitly considered.

In Laval and Daganzo’s multi-lane hybrid model, each lane is modeled as a separate
kinematic wave (KW) stream interrupted by lane-changing particles that completely block the
traffic. The flow transfers are predicted by using the incremental-transfer (IT) principle
(Daganzo et al., 1997) for multiple-lane KW problems, coupled with a one-parameter model for
discretionary lane-changing demand. In this dissertation, the lane changes made by the lower-
occupancy vehicles that want to pay to access the HOT lane are considered as mandatory lane-
changing. The lane-changing demand is given by equation (5.1). If the demand plus the arrival
flow on the HOV lane exceeds available capacity of the HOT lane, the IT principle is used to
prorate that available capacity\(^1\). To implement the hybrid model, all lanes are partitioned into
small cells of length \(\Delta x\), in addition to discretizing the time into time intervals \(\Delta t\). See Figure
6-1 for a discretized freeway representation.

### 6.1.2 Impacts of HOT Slip Ramp Design

It is known that the lane-changing behaviors before the entrance of the HOT lanes will be
largely affected by the physical configuration of access of the HOT lane. Figure 2-2 presents
three typical designs of HOT lane slip ramp (FHWA, 2003). The cell representation scheme
presented in Section 6.1.1 is relatively simple, and may reflect the reduced narrow buffer-
separated (no weave lane) design but with a strong assumption that all the lane changes will be
made at one location right before the end of the access segment. More realistic cell
representation may be used, and the modeling of lane changes will be updated accordingly.

\(^1\) In this case, the number of lane changes estimated from the loop detector data does not necessarily represent the
lane-changing demand. Therefore, when the HOT lane is congested, the detected flows will not be used to update the
estimates of motorists’ willingness to pay.
Figure 6-2 shows a cell representation scheme for the barrier-separated design option. The additional cells between the regular and the HOT lane represent the entrance (can be viewed as an acceleration lane). The weaving area may not need to be modeled since of more interest is the effect of the lane-changing behaviors on the HOT lane rather than the regular lane. The demand for the HOT lane is still calculated based on equation (5.1). However, the lane changes from the acceleration lane to the HOT lane will then be modeled as discretionary behaviors, using the approach proposed in Laval and Daganzo (2006).

Lou et al. (2007) conducted preliminary numerical experiments based on this more realistic HOT slip ramp cell representation and the corresponding multi-lane hybrid CTM. It is found that the total freeway throughput increases with the barrier-separated HOT slip ramp design. This is due to the fact that vehicles are able to accelerate before they merge into the HOT lane, and thus the voids created in the HOT lane flow are significantly reduced. These results suggest that HOT lane slip ramp configurations have substantial impacts on the performance of HOT lanes and the proposed modeling approach is able to capture those impacts.

6.2 System Inference

The first step in the framework of the self-learning approach, system inference, is to analyze traffic data to estimate the freeway system state and learn characteristics of travel demand and supply. The step includes three components, WTP learning, traffic state estimation and demand forecast. Discussion on WTP learning can be found in Section 5.2.2. With the enhanced traffic flow model presented in Section 6.1, this section elaborates the other two components in the system inference step.

6.2.1 Traffic State Estimation

Traffic state estimation is to deliver a complete image of freeway traffic conditions based on available traffic data at limited locations. It is the basis for facility performance monitoring
and providing initial conditions for the toll optimization in the unified framework. Data obtained from the detectors serves as direct measures with random measurement errors, while the predicted states from traffic flow models are viewed as an indirect measure which might not be very accurate either. The basic idea is to recover the full image by establishing a state that balances both the direct and indirect measurements. Various filtering techniques can be applied based on the different traffic flow models adopted (e.g., Wang and Papageorgiou 2005; Antoniou et al., 2007; Mihaylova et al. 2007).

To more realistically capture traffic dynamics along the tolling facility, Section 6.1 discusses the use of the multi-lane hybrid CTM proposed by Laval and Daganzo (2006) in the framework. However, the model is a non-linear non-differentiable transformation from one traffic state to another, particularly with the lane-choice probability involved. We thus apply unscented Kalman Filter (UKF) (Julier et al., 1995) to predict the new state with unscented sampling.

The state-space model of this filtering problem can be written as:

\[
\begin{align*}
    d(t+1) &= f_1(d(t)) + e_1 \\
    d_{\text{avg}}(t+1) &= f_2(d(t)) + e_2 \\
    z(t+1) &= h \cdot d_{\text{avg}}(t+1) + e_z
\end{align*}
\]

In model (6.1), \(d(t)\) is the vector of the densities along the freeway for both HOT and regular lanes at time point \(t\), \(f_1(\bullet)\) represents the transformation from \(d(t)\) to \(d(t+1)\), following the multi-lane hybrid CTM. Note that the measurements from the traffic detectors are usually aggregated, i.e., directly measured density is an average value across a certain time period. We then let \(d_{\text{avg}}(t)\) represents the average densities for both lanes between two data-pulling time points \(t-1\) and \(t\), computed from function \(f_2(\bullet)\) based on the same multi-lane hybrid CTM.
$z(t)$ is the vector of limited direct measurements, and $h$ denote the matrix indicating locations of the detectors. $e_1$ and $e_2$ are the corresponding process errors, and $e_z$ is the measurement error. While $e_z$ can be safely assumed to be Gaussian white noises with known variance matrix $\sigma_z$, further investigation (empirical study) of the traffic flow model is required to make reasonable assumptions on $e_1$ and $e_2$. In model (6.1), the parameters of the multi-lane hybrid CTM, such as the free flow speed, jam density and capacity, as well as the WTP are not treated as unknown states for better efficiency. Instead, the parameters for the traffic flow model can be calibrated offline, and the WTP can be estimated independently with the minimum requirements of detectors (see Figure 5-2). The first two state equations can be combined by augmenting the state variables to $x(t) = [d^T(t), d_{avg}^T(t)]^T$, grouping two CTM transformations and process errors to $f(\bullet) = [f_1^T(\bullet), f_2^T(\bullet)]^T$ and $e_x = [e_1^T, e_2^T]^T$, and expanding $h$ to $H$ accordingly. A more concise form of the state-space model (6.1) becomes:

$$
\begin{align*}
    x(t + 1) &= f(x(t)) + e_z \\
    z(t + 1) &= H \cdot x(t + 1) + e_z
\end{align*}
$$

Given the estimates of the mean and the covariance of state variable $x$ at time $t$ as $\hat{x}(t)$ and $P_x(t)$, the following $2n_x$ points with the same weight $1/2n_x$ can be generated, where $n_x$ is the dimension of the state variable $x$, and $(\sqrt{n_x P_x(t)})_i$ is the $i^{th}$ row or column of the matrix square root of $n_x P_x(t)$:

$$
\begin{align*}
    \bar{x}_{2i-1}(t) &= \hat{x}(t) - (\sqrt{n_x P_x(t)})_i \\
    \bar{x}_{2i}(t) &= \hat{x}(t) + (\sqrt{n_x P_x(t)})_i
\end{align*}
\quad \forall i = 1, 2, \ldots, n_x
$$

Equation (6.2) is called unscented sampling because these points have the same sample mean and covariance as $\hat{x}(t)$ and $P_x(t)$. Applying UKF, the new traffic state estimates are:
\[
\begin{align*}
\tilde{x}_i(t+1) &= f(\tilde{x}_i(t)) \quad \forall i = 1,2,...,2n_x \\
\tilde{x}_{\text{mean}}(t+1) &= \frac{1}{2n_x} \sum_{i=1}^{2n_x} \tilde{x}_i(t+1) \\
\hat{x}(t+1) &= \tilde{x}_{\text{mean}}(t+1) + G_x(t+1)[z(t+1) - H \cdot \tilde{x}_{\text{mean}}(t+1)] \\
G_x(t+1) &= P_{xx}(t+1)P_z^{-1}(t+1) \\
P_{xx}(t+1) &= \frac{1}{2n_x} \sum_{i=1}^{2n_x} (\tilde{x}_i(t+1) - \tilde{x}_{\text{mean}}(t+1))(H \cdot \tilde{x}_i(t+1) - H \cdot \tilde{x}_{\text{mean}}(t+1))^T \\
P_z(t+1) &= \frac{1}{2n_x} \sum_{i=1}^{2n_x} (H \cdot \tilde{x}_i(t+1) - H \cdot \tilde{x}_{\text{mean}}(t+1))(H \cdot \tilde{x}_i(t+1) - H \cdot \tilde{x}_{\text{mean}}(t+1))^T + \sigma_x \\
\tilde{P}_x(t+1) &= \frac{1}{2n_x} \sum_{i=1}^{2n_x} (\tilde{x}_i(t+1) - \tilde{x}_{\text{mean}}(t+1))(\tilde{x}_i(t+1) - \tilde{x}_{\text{mean}}(t+1))^T + \sigma_x \\
P_x(t+1) &= \tilde{P}_x(t+1) - G_x(t+1)P_z(t+1)G_x^T(t+1)
\end{align*}
\]

where \(\sigma_x\) is the covariance matrix of the process error \(e_x\).

Two most time-consuming steps in model (6.3) are applying multi-lane CTM to each sample point \(\tilde{x}_i(t)\) and the matrix inversion. Efficiency of the former depends on the dimension of the traffic state variable, and the latter the number of detectors installed. In order to improve the efficiency of the traffic state estimation, a coarser discretization of the site or simplification of traffic state estimation is necessary. Two possible simplified traffic state estimation procedures are: to generate less sample points in UKF, and use linear interpolation. Note that in the former simplified procedure, although the variance of the sample points after transformation may not be preserved, the sample mean remains unscented.

### 6.2.2 Demand Learning

Another component of system inference is to forecast short-term traffic demand such that toll determination can be more proactive. The forecast is instrumented by Bayesian learning. It is assumed that future traffic arrival follows a Poisson process, with unknown average arrival rate, denoted as \(\Lambda\). As time evolves, the detected arrival rates are used to adjust the estimate of the average arrival rate \(\Lambda\), and thus the distribution of the arrival rate. The operator can then
make use of the short-term demand forecast to determine tolls to better achieve the HOT operating objectives. Lin (2006) applies the above Bayesian learning concept to dynamically price a product to maximize the expected profit of a firm.

Based on the assumption, the number of vehicles arrived during time \((0,t]\), denoted as \(N\), will follow the Poisson distribution: 
\[
P(N = n|\Lambda = \lambda, t) = \frac{e^{-\lambda t} \cdot (\lambda t)^n}{n!}
\]
Since \(\Lambda\) is unknown, we may estimate its prior distribution from historical data as the following gamma distribution:

\[
pdf_{\Lambda} (\lambda) = \frac{ae^{-a \lambda} \cdot (a \lambda)^{k-1}}{\Gamma(k)}
\]

where \(\Gamma(k) = \int_{0}^{\infty} e^{-x} x^{k-1} dx\) is the gamma function, and parameters \(a\) and \(k\) are selected to fit the historical sample mean and variance of \(\Lambda\): \(a = E(\Lambda)/Var(\Lambda)\) and \(k = E^2(\Lambda)/Var(\Lambda)\). To update the prior distribution of \(\Lambda\) during operations, assume that during \((0,t], i\) number of vehicles have been observed from loop detectors. By Bayes’ theorem, posterior probability density function of \(\Lambda\) is:

\[
pdf_{\Lambda|N(t)=i} (\lambda) = \frac{pdf_{\Lambda} (\lambda) \cdot P(N(t) = i|\Lambda = \lambda)}{\int_{0}^{\infty} pdf_{\Lambda} (\lambda) \cdot P(N(t) = i|\Lambda = \lambda) d\lambda}
\]

\[
= \frac{ae^{-a \lambda} \cdot (a \lambda)^{k-1} \cdot e^{-\lambda t} \cdot (\lambda t)^i}{\Gamma(k)} \cdot \frac{i!}{\Gamma(k+i)}
\]

\[
= \frac{(a + t) e^{-(a+t) \lambda} ((a+t) \lambda)^{k+i-1}}{\Gamma(k+i)}
\]

which is another gamma distribution with parameters \((a+t, k+i)\). With the updated distribution of \(\Lambda\), the number of vehicles that may arrive during time interval \((t, t+\Delta t]\) can be estimated as below:
\[
P(N(t + \Delta t) - N(t) = n | N(t) = i) = \int_0^\infty e^{-\lambda \cdot \Delta t} \cdot (\lambda \cdot \Delta t)^n \cdot (a + t)e^{-(a + t)\lambda}(a + t)\lambda^{k+i-1} \cdot \frac{\Gamma(k+i)}{n!} \cdot \frac{\Delta t}{a + t + \Delta t}^n \cdot \left(\frac{a + t}{a + t + \Delta t}\right)^k \cdot \left(\frac{\Delta t}{a + t + \Delta t}\right)^n d\lambda
\]

Note that when \( k \) is integer, the above reduces to a negative binomial distribution:

\[
P(N(t + \Delta t) - N(t) = n | N(t) = i) = \binom{k + i + n - 1}{n} \cdot \left(\frac{a + t}{a + t + \Delta t}\right)^k \cdot \left(\frac{\Delta t}{a + t + \Delta t}\right)^n
\]

6.3 Toll Determination

The first step in the framework, system inference, has revealed the current system condition and predicted future traffic demand. With this information, the second step is to determine toll rates to maximize the total throughput of the entire freeway while ensuring the HOT lane operate in free flow condition. The multi-lane hybrid CTM can be incorporated in a rolling horizon scheme for toll determination.

In this section, the toll determination is improved in several aspects. In order to avoid dramatically fluctuating toll rates resulting from the prototype reactive pricing model, Section 6.3.1 proposes a proactive approach. Since robustness in traffic-adaptive setting is as important as in the static setting, Section 6.3.2 and 6.3.3 investigate the two major sources of uncertainty, users’ heterogeneity and random demand, in the HOT lane operations respectively. Two approaches of robust modeling are employed.

6.3.1 Deterministic Proactive Pricing Approach

The prototype model of the self-learning approach presented in Section 5.2.3 determines a toll rate for each time interval in response to the inflows measured at that particular time. The tolls may thus fluctuate dramatically, which may cause safety issues in reality. If a driver who decides to access the toll lane observes a sudden price jump, he or she may become reluctant and make abrupt maneuvers that possibly lead to a rear-end accident. A proactive approach may be
adopted to address the issue. Assume for now that future traffic arrivals are known and deterministic. This knowledge can be incorporated in advance into the rolling horizon framework, which simulates the traffic for a period longer than the toll interval to allow appropriate flow propagation, obtain a more accurate estimate of the resulting control objectives, and provide some lead time for adjusting toll rates in a smoother manner to respond to the changes in traffic demand and maintain a high throughput.

Suppose the length of the rolling horizon is set to $V$ tolling intervals. Within the horizon, the decision is now a vector of toll rates, which has $V$ components. Let $\bar{q}_T$ and $\bar{q}_R$ represent the average throughput of HOT and regular lanes at the downstream end of the freeway segment respectively during the rolling horizon, and $\bar{v}_T$ the average speed of HOV/HOT lane. The control objectives can now be more specifically stated as maximizing the sum of the average throughputs at the downstream end while attempting to keep the average speed of HOV/HOT lane as free flow speed $v_f$. These measurements of performance can be computed from the hybrid CTM with mandatory lane-changing demand at the HOT entrance calculated by equation (5.1). With $\beta$ being the toll vector, $\mu_T$ and $\mu_R$ the singleton inflow value or the vector of future inflows, computing the objective measures is essentially a mapping $\psi$:

$$\psi(q_T, q_R, v_T) = \psi(\alpha_1, \alpha_2, \gamma, \beta, \mu_T, \mu_R)$$

Consequently, the toll optimization problem can be written as:

$$\max_{\beta} \left( q_T + q_R \right) + \theta \cdot \min\{|v_T - v_f, 0|\}$$

s.t. \hspace{1cm} 0 \leq \beta_i \leq \beta_{\text{max}} \hspace{1cm} \forall i = 1, 2, \ldots, V

$$-\Delta \beta \leq \beta_i - \beta_{i-1} \leq \Delta \beta \hspace{1cm} \forall i = 1, 2, \ldots, V$$

Condition (6.4)
where $\theta$ is a penalty parameter, $\beta_0$ is the toll rate of the time interval previous to the rolling horizon, $\Delta \beta$ and $\beta_{\text{max}}$ are the maximum margin of toll variation for two consecutive intervals (due to safety consideration) and the absolute maximum toll rate specified by the tolling authority. The second term in the objective function is a penalty for speed reduction. If $\theta$ is selected appropriately, this term should be equal to zero with the optimal solution, ensuring that the average speed of HOT lane is free flow speed. Note that the objective function is continuous and bounded above, and the feasibility set is compact, therefore optimal solution exists. A variety of numerical algorithms, such as the golden-section method, can be used to search for a local optimum.

To illustrate this proactive concept, Lou et al. (2007) conducts a numerical example upon the simplest cell representation. The numerical study compares the reactive model, which determines the toll rates interval by interval in response to the current detected inflow, with the proactive model that computes a toll vector each time based on the predicted future inflows. The study suggests that the proactive approach leads to a smoother toll pattern without sacrifice the control objectives.

### 6.3.2 Robust Controller with Heterogeneous Motorists

Previous studies have identified substantial heterogeneity in motorists’ values of travel time and its important implications for road pricing policy (e.g. see Small et al., 2005). In this section, heterogeneous motorists with different WTP are considered, i.e., there will be distributions associated with $\alpha_1, \alpha_2$ and $\gamma$ across the population. To accommodate the heterogeneity of motorists, we seek for a robust pricing strategy that makes the control performance insensitive to the variations of the three WTP parameters. Since the distributions of the WTP parameters are difficult to obtain, we apply the notion of robust optimization (e.g., Ben-
Tal and Nemirovski, 2002; El-Ghaoui, 2003) which does not require any distribution information on the uncertain parameters. This approach seeks a robust solution minimizing the worst-case performance against an ellipsoidal uncertainty set that reflects decision makers’ attitudes toward risk. It has been proved that the robust solution with ellipsoidal uncertainty set is less conservative than, e.g., the polyhedral uncertainty set, and has properties that are computationally advantageous.

With the presence of user heterogeneity, the learning procedure of users’ WTP discussed in Section 5.2.2 is still able to estimate approximately the means and variances of $\alpha$, $\beta$, and $\gamma$. Without requiring the knowledge about the underlying distributions of $\alpha$, $\beta$, and $\gamma$, we assume these three parameters vary within an ellipsoidal likelihood region. In other words, motorists’ WTP is confined by the following ellipsoid:

$\begin{align*}
Y_{WTP}(t+1) &= \{y \in R^3 \mid y = (\hat{\alpha}(t), \hat{\beta}(t), \hat{\gamma}(t))^T + \eta \cdot M(t) \cdot u, \|u\|_2 \leq 1\} \\
\end{align*}$

Here $M(t)$ is a diagonal matrix with standard deviations of $\alpha$, $\beta$, and $\gamma$ as its elements, and $\eta$ is the multiplier, adjusting the size of the ellipsoid to represent the operator’s trade-off between efficiency and robustness. We then determine a toll vector that maximizes the minimum control objective (improves the worst-case performance) incurred by the WTP parameters in the set.

Mathematically, the robust optimal toll rate solves the following max-min problem (a special case of the least-square problems with uncertain data discussed in El-Ghaoui and Lebret, 1997):

$\begin{align*}
\max_{\beta} \min_{\alpha, \gamma, y} \left( (\overline{y}_T + \overline{y}_R) + \theta \cdot \min \{v_T - v_f, 0\} \right) \\
\text{s.t.} \quad \text{Constraints (6.4), (6.6)-(6.7)}
\end{align*}$

Efficient iterative procedures can be developed to solve the problem. A cutting plane scheme similar to the one in Section 4.3.2 can be applied. The algorithm solves a sequence of relaxed max-min formulations that approximate the ellipsoidal uncertainty set with a set of
discrete parameter scenarios. The algorithm starts from an initial parameter scenario set that only contains the estimates at current time interval, and optimizes toll against the set. It then finds the worst-case parameters $\alpha_1, \alpha_2 \text{ and } \gamma$ under the toll rate, and then includes those parameters into the scenario set, against which the toll rate is then optimized again. This procedure will repeat until the optimal solution converges.

Yin and Lou (2009) conducts numerical examples using the point-queue model to illustrate the concept of robust pricing. It is demonstrated that the robust controller is able to produce WTP estimates that are satisfactorily close to the actual values, and deliver more stable performance.

6.3.3 Robust Controller with Demand Uncertainties

Given the future traffic demand is likely uncertain, this sub-section further addresses this issue in addition to the user heterogeneity. Similar robust optimization with ellipsoidal uncertainty set can be applied to model random future demand. However, while the distributions of WTP parameters are difficult to estimation, the distribution of the future arrival can be updated in real time using Bayesian learning discussed in Section 6.1.2. Therefore, another stream of optimization methods under uncertainty, namely stochastic programming (Birge and Louveaux, 1997), can be employed to seek for a toll rate that performs well under most of the possible scenarios.

With the demand learning, the probability density function of the arrival distribution can be approximated by a discrete set of possible scenarios, $s = 1, 2, \ldots, S$, where each scenario specifies the traffic arrival rates for each interval in the rolling horizon, denoted as $\mu_r^+(t)$ and $\mu_r^-(t)$, with a probability of occurrence as $p_s$. Since most decision makers are risk averse, a scenario-based stochastic program is formulated to minimize the expected loss (penalty minus
throughput) incurred by high-consequence scenarios, which is called conditional value-at-risk (CVaR) in financial engineering (Rockafellar and Uryasev, 2000). The decision makers’ attitude toward risk can be controlled by a parameter $\delta$. More specifically, only scenarios with loss greater than the $\delta$-percentile (say, 90%) value are considered, and the expected loss of those scenarios is called $\delta$-CVaR. The scenario-based stochastic toll optimization problem is written as follows:

$$\min_{(\beta, \xi, L_t)} \xi + \frac{1}{1-\delta} \sum_{s=1}^{S} p_s \cdot \max(L_s - \xi, 0)$$

s.t.

$$L_s = -\left(\bar{q}_T^s + \bar{q}_R^s\right) + \theta \cdot \max\left\{v_f - \bar{v}_T^s, 0\right\}$$

$$\left(\bar{q}_T^s, \bar{q}_R^s, \bar{v}_T^s\right) = \psi(\alpha_1, \alpha_2, \gamma, \beta, \mu_T^s, \mu_R^s)$$

Constraints (6.6)-(6.7)

It can be proved that the optimal value of the objective function is the minimum $\delta$-CVaR, and the solution $\xi$ is the $\delta$-percentile loss (Rockafellar and Uryasev, 2002).

Note that the above stochastic formulation is flexible and can be extended to incorporate other sources of uncertainty. For example, if user heterogeneity can be approximated by a number of discrete scenarios, it is possible to integrate it with the scenarios of traffic demand, and the above stochastic formulation will be able to consider both types of the uncertainty. On the other hand, the robust formulation can also be expanded to incorporate demand uncertainty by modifying the uncertainty set correspondingly.

### 6.4 Simulation Study

We conduct simulation experiments to demonstrate and validate the proposed approach. Multiple components of the system inference and toll determination are activated to illustrate that they can be integrated to a unified framework. Within the simulation platform, the simulator
attempts to replicate the motorists’ lane-choice behaviors and traffic dynamics; the controller implements the proposed framework of system inference and toll optimization; and the monitor mimics a real-world surveillance system, generate perturbed measurements of actual traffic conditions simulated with actual parameters. In the following, to facilitate the key concepts of the proposed framework, we implemented the system inference components of WTP learning and traffic state estimation, and the deterministic proactive version of toll optimization.

The simulation site is a freeway segment shown in Figure 6-3a. The site is 2 miles long, with a barrier-separated HOT slip ramp. It is assumed that each lane obeys the same triangular fundamental diagram. The relevant parameters are reported in Figure 6-3b, and the small triangle is for the downstream bottleneck. Additionally, all lane-changing vehicles are assumed to have an acceleration rate of 12.22 ft/s², i.e., it will take a vehicle 7.2 seconds to accelerate from zero speed to free-flow speed. The true values of $\alpha_1$, $\alpha_2$ and $\gamma$ used in the simulator are 0.5, 1 and 0.2. $\alpha_1/\alpha_2 = 0.5$ indicates a value of travel time of 15 dollars per hour since the travel time difference is represented in the unit of one tolling interval (2 minutes).

The entire simulation duration is 20 minutes. The discrete simulation interval in hybrid CTM model is 0.6 seconds, and all lanes are partitioned into small cells of 0.01 mile, leading to a vector with a total number of 420 elements in the traffic state estimation. To be consistent with the practice, the monitor reports aggregated traffic measurements every single minute, and the traffic state estimation is conducted once new measurements are received. Moreover, the toll rate varies every two minutes, and the rolling horizon for toll optimization is 4 minutes (2 time intervals). The weighting factor $\theta$ in model (6.5) is set as one and the toll optimization problem is solved using the golden-section method. In the simulation, random arrivals are generated from a virtual source with an average rate of 2400 vph for the regular lane, 600 vph and 1200 vph.
during the first and last 10 minutes for the HOV lane. $\alpha_1(0)$, $\alpha_2(0)$, $\gamma(0)$ and $P_{WTP}(0)$ are initialized as 1, 2, 0 and an identity matrix respectively. Initial values of traffic conditions are generated randomly, and $P_x(0)$ is set to identity matrix.

Figure 6-4 shows that travelers’ WTP can be gradually learned within the unified framework during the tolling operation. The calibrated values of parameters $\alpha_1$, $\alpha_2$ and $\gamma$ are able to converge from the initial values to the true values within 6 minutes. The calibration time for each interval is less than 0.003 seconds. Since the size of the traffic state vector is very large, we do not present the detailed comparison between the estimated and the actual values directly. Instead, we will compare the control results to the base case where actual traffic condition is known precisely.

The resulting toll rates and the performances of the freeway segment are presented in Figure 6-5. It is demonstrated that the controller is able to adjust the toll rate in response to the estimated traffic condition in real time. The proactive toll optimization ensures that the toll rate changes smoothly from interval to interval. Figure 6-5 also compares the resulting average speed and throughputs from two cases where HOT lane is operated respectively under optimal toll rate and without any control. The latter case means that LOV can use the HOT lane without paying any toll. It can be observed that the controller is able to maintain a high and stable throughput while preventing the HOV/HOT lane from being congested. The minimum average speed along the HOV/HOT lane is 55.41 mph under optimal toll rates; while in the no-control case, the average speed can be as low as 34 mph. On the other hand, the average throughput is 3291 vph under optimal toll rates, which is 91% of the downstream bottleneck capacity, and only slightly lower than the throughput of 3350 vph without any control, while 50% of the capacity would be wasted if LOV is not allowed to access the HOT lane. One of the reasons that the
HOT lane throughput does not reach the capacity of the downstream bottleneck is that lane-changing vehicles act as moving bottlenecks while accelerating to the speed prevailing on the HOT lane. Those vehicles create gaps in the flow in front of them that propagate forward, thereby reducing the throughput. In the simulation, additional loss of throughput is due to initial inaccurate estimates of motorists’ WTP.

In the above numerical example, the computation time of the toll rates for each time interval is less than 9 seconds. However, updating the estimates of traffic conditions once every minute takes more than 90 seconds due to the large number of unscented samples required in the UKF method. Therefore, the two simplified traffic state estimation procedures proposed in Section 6.2.1 are tested. The impacts of different traffic state estimators on the control objectives are investigated by comparing them to the base case where actual traffic condition is known precisely. We conducted multiple runs and our statistical tests indicate that different versions of traffic state estimation do not affect the resulting average speed of HOT lane and the freeway throughput significantly. One plausible reason is that the simplified traffic state estimators are still able to capture the prevailing traffic conditions along the freeway because the multilane hybrid CTM is mainly a first-order traffic flow model. On the other hand, the two simplified versions are far less demanding in computation with an average of less than 8 and 0.001 seconds respectively. Their computational efficiency is a desirable attribute in real-time operation.

6.5 Summary

In this chapter, the unified framework of proactive self-learning dynamic pricing for HOT lanes is further elaborated. This framework incorporates two major steps, system inference and toll optimization, and is able to generate robust and proactive dynamic pricing strategies. The system inference applies a variety of techniques such as regular KF, UKF and Bayesian learning to mine traffic data to gain better understanding of motorists’ WTP, traffic state and short-term
traffic demand. The attained knowledge contributes in the second step in determining optimal toll rates that lead to efficient utilization of freeway capacity and superior travel services for HOT lanes. Three toll optimization formulations have been proposed to enhance the proactiveness and the robustness of the self-learning controller. Simulation experiments confirm that the integrated framework is feasible, efficient and effective.

The proposed unified framework can be extended in multiple ways to take into account more considerations. For example, another more advanced discrete choice model, such as a mixed logit model, can be incorporated as an alternative to describe the travelers’ heterogeneous lane-choice behaviors. Our future research will also integrate ramp metering with dynamic pricing, and develop coordinated pricing strategies for freeways with multiple HOT segments. We also plan to enhance the capability of the microscopic traffic simulation software CORSIM in simulating HOT lane operations, and then conduct simulation experiments to evaluate the proposed framework in a microscopic simulation platform.
Figure 6-1. Discretized freeway representation
Figure 6-2. Discretized freeway representation (barrier-separated)
Figure 6-3. Simulation settings
Figure 6-4. Calibration of Willingness to Pay
Figure 6-5. Optimal toll rate and its performance
CHAPTER 7
CONCLUSION

This dissertation develops a hierarchical congestion pricing framework for transportation network in view of the fact that travelers have limited response capabilities to the tolling signals at different spatial levels. Toll determination is hence decomposed into two levels: network and facility. Robust static or time-of-day pricing scheme is proposed at the network level to avoid complex toll structures while ensuring the network will perform reasonably well against various uncertainties. At the facility level, tolls determined from the network level can be further adjusted based on real-time traffic conditions to meet local objectives with a detailed modeling of drivers’ behavior and traffic dynamics. Through systematically reviewing the general inputs and technical issues of this hierarchical framework, it is demonstrated that the proposed framework is feasible and has great potential for actual implementation. Moreover, the proposed pricing approaches for the two levels take into account the corresponding key features, and complete the missing pieces of robust pricing with users’ boundedly rational behaviors and adaptive pricing for managed lane operations in the literature. Simulation studies lead to promising results, suggesting that the proposed hierarchical framework may be efficient and effective for congestion mitigation in practice.

At the network level, this dissertation explores static or time-of-day robust network-level pricing problem with bounded rationality in travelers’ route choices. The static or time-of-day feature makes the tolling signals easy for the travelers to understand and follow. The investigation of bounded rationality, an important alternative source of network uncertainties, enriches the literature by providing a rigorous definition, models of traffic assignment with bounded rationality, implications of the resulting flow distribution, and a robust pricing scheme. It is found that system performance may vary substantially within the possible equilibrium flow
set under bounded rationality and the proposed robust pricing models are able to protect against the worst-case scenario effectively. More efficient solution approaches will be developed in the future. Moreover, further study will investigate how to determine a robust network-level tolling scheme under other more realistic network settings with boundedly rational behaviors, such as elastic demand or combined route and departure time choices.

At the facility level, this dissertation develops one feedback and one self-learning approach particularly for HOT lane operations. Both approaches are dynamic and traffic-adaptive, providing superior free-flow travel services for HOT lanes while efficiently utilizing the capacity of the freeway. The self-learning approach incorporates several human and traffic factors and has the potential to address several critical issues in practical HOT operations. Through mining the real-time data of traffic measurements, this approach is able to learn information from both the demand and supply side of the HOT facility and calculate proactive and robust toll rates based on the updated information. Components of system inference, such as travelers’ willingness-to-pay learning, traffic state estimation, and short-term demand prediction, are integrated in a unified framework to serve as a basis of the toll determination. These enhancements help the self-learning approach become more realistic, efficient and effective. Both of the proposed approaches are feasible and easy to implement. Simulation experiments demonstrate that the self-learning approach to be a more promising method for HOT operations. It can be further extended in multiple ways to take into account more considerations as well. For example, more advanced models can be adopted to describe the travelers’ lane-choice behaviors more realistically. Future research may also examine the possibility of integrating ramp metering with dynamic pricing, and develop coordinated pricing strategies for freeways with multiple HOT segments. Another direction is to enhance the capability of microscopic traffic
simulation software, such as CORSIM, in simulating HOT lane operations and conduct simulation experiments to evaluate the proposed framework in a microscopic simulation platform.

In summary, a hierarchical framework is developed in this dissertation based on practical considerations of pricing. Robust time-of-day tolls are proposed for the entire network while adaptive tolls are advocated for special facilities. The framework is feasible and promising for congestion mitigation, and can be further improved to be more realistic and practical.
LIST OF REFERENCES


131


Ungemah, D. and Swisher, M (2006) *So you want to make a high-occupancy toll lane?-Project manager’s guide for conversion from high-occupancy vehicle lane to high-occupancy toll lane.* Transportation Research Record, 1960, 94-98.
U.S. Department of Transportation (2006) *National strategy to reduce congestion on America's transportation network.*

Utah Department of Transportation (2009) *Using the express lanes.*


BIOGRAPHICAL SKETCH

Yingyan Lou received two bachelor’s degrees in both engineering science and economics from Peking (Beijing) University, China, in 2005; and earned her master’s degree in civil engineering from the University of Florida in 2007.

Yingyan Lou’s primary research interest is transportation systems modeling and optimization, with applications on system-wide congestion pricing, traffic-responsive tolling for managed-lane operations, dynamic origin-destination demand estimation, robust transportation network design, freeway incident response planning, and infrastructure asset management. She also has a keen interest in traffic flow theory and operations.

During her Ph.D. study at the Transportation Research Center in University of Florida, Yingyan Lou has co-authored seven papers and made eight presentations at various conferences. She won the second prize of student paper competition at the 2nd International Symposium of Freeway and Tollway Operations in 2009, the best poster award at the 2007 Institute of Transportation Engineers Florida District Annual Meeting, and was awarded the graduate scholarship of Women’s Transportation Seminar in 2008. She also won numerous travel grants from various sources and was awarded the title of “Outstanding International Students” by the University of Florida twice.