

**MITIGATING RISK AND AMBIGUITY IN
SERVICE SYSTEMS**

QI, JIN

NATIONAL UNIVERSITY OF SINGAPORE

2014

**MITIGATING RISK AND AMBIGUITY IN
SERVICE SYSTEMS**

QI, JIN

(B.Eng, Tsinghua University (2006))

(M.Sc, Tsinghua University (2010))

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
DEPARTMENT OF DECISION SCIENCES
NATIONAL UNIVERSITY OF SINGAPORE

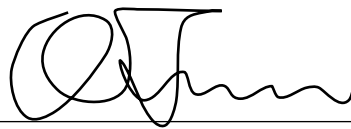
2014

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, appearing to read 'Qi, Jin', written over a horizontal line.

Qi, Jin
28 July 2014

ACKNOWLEDGEMENT

First and foremost, no words could express my heartfelt gratitude to my advisor and also a great friend, Melvyn Sim. His burning passion for the research, creative ideas, endless support and encouragement led me through this long, arduous but exciting PhD journey. Whenever I need help, he could always provide the best advice immediately, no matter whether it is at 1am or at the weekend. I am extremely lucky to have had this opportunity to work with him.

Great thanks are also due to the members of my committee: Jie Sun and Qiang Meng. Their tireless supports and insightful suggestions on my research are immensely valuable to me. I am quite privileged to work with my coauthors: Nicholas G. Hall, Patrick Jaillet, Defeng Sun, Xiaoming Yuan, Xin Chen. Without their outstanding contributions, I could not complete this thesis. I would like to call particular attention to Nicholas G. Hall, who has always been willing to share his experience in research and teaching with me, and Patrick Jaillet, who has kindly invited me to exchange at MIT for half a year. The experience there was wonderful and inspiring.

Department of Decision Sciences is a great home to me. Besides my advisor, I have also benefitted greatly from other remarkable faculty members: Chung-Piaw Teo, Jie Sun, Hanqin Zhang, Andrew Lim, Jussi Keppo, Mabel

Chou, Yaozhong Wu, Lucy Chen and Tong Wang. Thank you for creating such a nice environment for us to study. Special gratitude also goes to my friends in the department: Qingxia Kong, Vinit Kumar Mishra, Yuchuan Yuan, Zhichao Zheng, Junfei Huang, Meilin Zhang, Rohit Nishant, Li Xiao, Jeremy Chen, Zhi Chen, Sheng Zhao, Weijia Gu, Baiyu Li, Yini Gao, Shasha Han and Zhenzhen Yan. I will never forget the joyful moments that we had together.

I am deeply indebted to my parents Mingliang Qi, Jihong Guo and my brother Guanqun Qi. Although they are physically far away, this thesis would not have been possible without their unconditional love and fully supports. I also thank my late grandmother, Delan Zhang, to whom this thesis is dedicated.

Last but not least, I owe a great deal of gratitude to my husband Daniel Zhuoyu Long. We have been together for ten years. He is my best friend, soul mate and the greatest coauthor. I am so thankful for having him always be by my side and for giving purpose to my days.

CONTENTS

1. <i>Introduction</i>	1
1.1 Motivation and Literature Review	2
1.2 Structure of the Thesis	4
1.3 Notation	7
2. <i>Preferences for Travel Time under Risk and Ambiguity</i>	9
2.1 Introduction	9
2.2 Preferences for Travel Time	13
2.2.1 Ambiguity-aware CARA travel time (ACT)	16
2.2.2 Two uncertainty models for travel time	23
2.3 Path Selection under the ACT Criterion	30
2.4 Analysis of Network Equilibrium with Risk and Ambiguity Aware Travelers	33
2.4.1 Network equilibrium formulation	34
2.4.2 Inefficiency of network equilibrium	41
2.4.3 A network equilibrium example	47
2.5 Conclusion	54
3. <i>Routing Optimization with Deadlines under Uncertainty</i>	57

3.1	Introduction	57
3.2	Lateness Index	64
3.3	General Routing Optimization Problem with Deadlines	74
3.3.1	Model definition	75
3.3.2	Model reformulation	76
3.3.3	Solution procedure	86
3.4	Computational Study	98
3.4.1	Stochastic shortest path problem with deadline	99
3.4.2	Solution procedure illustration	105
3.4.3	General routing optimization problem	108
3.5	Extension: correlations between uncertain travel times	110
3.6	Conclusion	111
4.	<i>Mitigating Delays and Unfairness in Appointment Systems</i>	113
4.1	Introduction	113
4.2	Delay Unpleasantness Measure	120
4.3	Lexicographic Min-Max Fairness	124
4.4	Appointment Schedule Design	130
4.4.1	Stochastic optimization approach	133
4.4.2	Distributionally robust optimization approach	134
4.5	Appointment Sequence and Schedule Design	145
4.6	Computational Study	152
4.6.1	Comparison of quality measures	152
4.6.2	Distributional ambiguity	156
4.6.3	A sequencing and scheduling example	158

4.7 Conclusion	160
5. <i>Conclusions and Future Research</i>	162

ABSTRACT

This dissertation explicitly distinguishes between risk, where the frequency of outcomes is exactly known, and ambiguity, where it is not, and studies problems in two service systems: transportation system and healthcare system. At its core, we collectively address three issues: 1) how to properly model uncertainties to incorporate empirical data and reflect real-world concerns, 2) how to describe and prescribe individual preferences when facing uncertainties and account for behavior issues such as fairness, and 3) how to incorporate the two aspects in optimization or equilibrium models so that meaningful decisions can be obtained with modest computational effort.

In the transportation system, we first study the preferences for uncertain travel times in which probability distributions may not be fully characterized. In particular, we propose a new criterion named *ambiguity-aware CARA travel time* for evaluating uncertain travel times under various attitudes of risk and ambiguity, which is a preference based on blending the Hurwicz criterion and Constant Absolute Risk Aversion. More importantly, we show that when the uncertain link travel times are independently distributed, finding the path that minimizes travel time under the new criterion is essentially a shortest path problem. We also study the implications on Network Equilibrium model where travelers on the traffic network are characterized by their

knowledge of the network uncertainty as well as their risk and ambiguity attitudes. The results suggest that as uncertainty increases, the influence of selfishness on the inefficiency diminishes.

Based on the new criterion, we then consider a class of routing optimization problems on networks with deadlines imposed at a subset of nodes, and with uncertain arc travel times. We introduce the *lateness index* to evaluate the deadline violation level of a given policy for the network with multiple deadlines. We provide two mathematical programming formulations: a linear decision rule formulation, and a multi-commodity flow formulation and develop practically “efficient” algorithms involving Benders decomposition to find the exact optimal routing policy. The numerical results clearly demonstrate the benefit of the lateness index policies, and the practicality associated with the computation time of the solution methodology.

In the healthcare system, we study an appointment system design problem in which heterogeneous participants are sequenced and scheduled for service. As service times are uncertain, the aim is to mitigate the unpleasantness experienced by the participants in the system when their waiting times or delays exceed acceptable thresholds, and address fairness concerning the balancing of service levels among participants. In evaluating uncertain delays, we propose the *Delay Unpleasantness Measure* which accounts for the frequency and intensity of delays above a threshold, and introduce the concept of lexicographic min-max fairness to design appointment systems from the perspective of the worst-off participants. The optimal sequencing and scheduling decisions can be derived by solving a sequence of mixed-integer programming problems.

Thesis supervisor: Melvyn Sim

Title: Professor, Department of Decision Sciences

LIST OF FIGURES

2.1	A simple network with uncertain travel time.	22
2.2	Path preferences under different attitudes towards risk and ambiguity.	29
2.3	Two paths network with uncertain travel times.	47
2.4	Inefficiency of NE and DSO under the ACT criterion in Case 2 and 3 in two-nodes network.	52
2.5	Inefficiency of NE and DSO under the ACT criterion in Case 3 in five-nodes network.	55
3.1	An illustrative example explaining the difference between LDR and MCF formulations.	85
3.2	Performance comparison for stochastic shortest path problem when deadline varies.	104
3.3	An illustrative example on a five-nodes network.	106
4.1	Sequencing and scheduling decisions with various tolerances. .	159

LIST OF TABLES

2.1	Preferences for travel times under the ACT criterion.	23
2.2	Path preferences under the ACT criterion.	27
2.3	Travelers' profile in Case 3.	48
2.4	Flow patterns of NE and SO under the ACT criterion for three cases.	49
3.1	Performances of various selection criteria for stochastic shortest path problem with deadline.	103
3.2	Statistics of CPU time of two algorithms for stochastic shortest path problem with deadline.	105
3.3	Travel time information corresponding to Figure 3.3.	106
3.4	All feasible paths for the illustrative example without the deadline requirements.	106
3.5	Calculation procedure of lateness index model with different β	107
3.6	Arrival time comparison between paths 5 and 6.	108
3.7	CPU time (sec) on routing optimization problem with different settings.	109
3.8	Number of iterations on routing optimization problem with different settings.	109

4.1	Patients' optimal appointment time under two scheduling methods.	153
4.2	Delay performance under two scheduling methods (two-point).	154
4.3	Average performance analysis of two scheduling methods among 100 instances.	155
4.4	Statistics of consultation time from empirical data.	155
4.5	Delay performance under two scheduling decisions (empirical data).	156
4.6	Delay performance under uniform distribution.	157
4.7	Delay performance under beta distribution.	157
4.8	Characterization of heterogeneous patients.	158

1. INTRODUCTION

This dissertation focuses on the analytics of service systems, with the goals of eliciting operational insights and providing solutions for supporting decision-making in practice. At its core, it seeks to address three issues in service systems: 1) how to properly model uncertainties to incorporate empirical data and reflect real-world concerns, 2) how to describe and prescribe individual preferences when facing uncertainties and account for behavior issues such as fairness, and 3) how to incorporate these two aspects in optimization or equilibrium models so that meaningful decisions or insights can be obtained with modest computational effort. This dissertation clearly distinguishes the risk, in which the frequency of outcomes is exactly known, and ambiguity, in which it is not, and studies decision makers' preferences on the risk and ambiguity in three operational problems. It is a collection of interrelated essays, including the traffic equilibrium problem and vehicle routing problem in the transportation system, and the appointment scheduling problem in the healthcare system.

1.1 Motivation and Literature Review

Uncertainty is ubiquitous. In healthcare operations, the consultation time, patients' arrival rate and length of stay are uncertain. In the transportation area, the travel time is uncertain. To describe and analyze uncertainties, a popular and classic approach is using probability theory, which assumes that each uncertainty follows a known probability distribution. Based on that, researchers tend to use expected utility theory to capture decision makers' attitudes towards risk. However, in many cases, complete probability distribution of a random variable is seldom known exactly, and even the estimated one could be considerably affected by the sampling procedure. Moreover, if the probability distribution of a random variable is not fully known, then it would be impossible to establish the preferences based on the expected utility criterion. In fact, the distinction between risk, where the frequency of outcomes is known, and ambiguity, where it is not, can be retrospectively traced to Knight (1921): *But uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated. . . . It will appear that a measurable uncertainty, or "risk" proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all. We shall accordingly restrict the term "uncertainty" to cases of the non-quantitative type.*

Since then, risk and ambiguity have been extensively studied in economics (see for instance, Camerer and Weber 1992; Mukerji and Tallon 2003; Maccheroni et al. 2006; Gilboa et al. 2008; Wakker 2008), finance (see for instance, Dow and da Costa Werlang 1992; Chen and Epstein 2002; Epstein

and Schneider 2008; Bossaerts et al. 2010; Guidolin and Rinaldi 2013), and marketing (see for instance, Swait and Erdem 2007; Muthukrishnan et al. 2009). Ellsberg (1961) shows convincingly by means of paradoxes that ambiguity preference cannot be reconciled by classical expected utility theory. He argues that the ambiguity of information brings a degree of “confidence” in the estimation of the likelihood. Inspired by this seminal work, numerous experimental and theoretical studies spring up to verify and accommodate this behavior issue. Notably, in Hsu et al. (2005) groundbreaking experiments, economists and neuroscientists collaborate to establish significant physiological evidence via functional brain imaging that humans have varying and distinct attitudes towards risk and ambiguity. The results also indicate that people’s attitudes towards risk and ambiguity are not fully correlated, i.e., there exists a population of people that are ambiguity averse and risk-seeking, or ambiguity seeking and risk-averse.

From the normative perspective, ambiguity is also an active area of research within the domains of decision theory and operations research. Gilboa and Schmeidler (1989) consider ambiguity as a set of possible probability distributions, and present the Max-Min Expected Utility (MEU) model, which appeals to ambiguity averse decision makers. To accommodate the heterogeneity of ambiguity and risk attitudes found in the experiments, Ghirardato et al. (2004), based on Hurwicz criterion (Hurwicz 1951), axiomatize the α -MEU model, which represents a compromise via a convex combination of the worst and best case expected utility. The parameter α is an index of pessimism or optimism.

However, in the service industries, for example, transportation and health-

care, the majority of studies still assumes that the full knowledge of the uncertainties is known to every one. These assumptions on the uniformity of the agents and the known distribution are unrealistic in many operational problems and may also complicate the solution procedure. For example, in the traffic equilibrium problem, various travelers may have distinct information on the uncertain travel time and the attitudes towards it. A local resident, who is very familiar with the area, would be less ambiguous, compared to a tourist, in characterizing the uncertain travel times. Even different residents may have different information. In the appointment system design problem, it is generally hard to construct a probability distribution of the consultation time, that could be verified by the empirical data but also help us develop a tractable model.

Motivated by the evidence above, we aim to investigate the decision making in the service systems under both risk and ambiguity. Specifically, by clearly distinguishing between risk and ambiguity, we first study people's preferences and attitudes towards them. Then, we provide guidance for managers or central planners to make decisions based on these preferences. In this thesis, we focus on the transportation system and the healthcare system. The ideas and formulations can be generalized to other service systems.

1.2 Structure of the Thesis

The rest of the thesis is organized as follows.

- **Chapter 2: Preferences for Travel Time under Risk and Ambiguity.**

In this chapter, we study the preferences for uncertain travel times in which probability distributions may not be fully characterized. In evaluating an uncertain travel time, we explicitly distinguish between *risk* and *ambiguity*. In particular, we propose a new criterion called *ambiguity-aware CARA travel time* (ACT) for evaluating uncertain travel times under various attitudes of risk and ambiguity, which is a preference based on blending the Hurwicz criterion and Constant Absolute Risk Aversion (CARA). More importantly, we show that when the uncertain link travel times are independently distributed, finding the path that minimizes travel time under the ACT criterion is essentially a shortest path problem. We also study the implications on Network Equilibrium (NE) model where travelers on the traffic network are characterized by their knowledge of the network uncertainty as well as their risk and ambiguity attitudes under the ACT. We derive and analyze the existence and uniqueness of solutions under NE. Finally, we obtain the Price of Anarchy that characterizes the inefficiency of this new equilibrium. The computational study suggests that as uncertainty increases, the influence of selfishness on the inefficiency diminishes.

- **Chapter 3: Routing Optimization with Deadlines under Uncertainty.**

In this chapter, inspired by the ACT defined in Chapter 2, we consider a class of routing optimization problems on networks with deadlines imposed at a subset of nodes, and with uncertain arc travel times. The

problems are static in the sense that routing decisions are made prior to the realization of uncertain travel times. The goal is to find optimal routing policies such that arrival times at nodes respect deadlines “as much as possible”. We propose a precise mathematical framework for defining and solving such routing problems. We first introduce a performance measure, called *lateness index*, to evaluate the deadline violation level of a given policy for the network with multiple deadlines. The criterion can handle the risk, when probability distributions of the travel times are considered known, and ambiguity, when these distributions are partially characterized through descriptive statistics, such as means and bounded supports. We show that for the special case in which there is only one node with a deadline requirement, the corresponding shortest path problem with deadline can be solved in polynomial time under the assumption of stochastic independence between arc travel times. For the general case, we provide two mathematical programming formulations: a linear decision rule formulation, and a multi-commodity flow formulation. We develop practically “efficient” algorithms involving Lagrangian relaxation and Benders decomposition to find the exact optimal routing policy, and give numerical results from several computational studies, showing the attractive performance of lateness index policies, and the practicality associated with the computation time of the solution methodology.

- **Chapter 4: Mitigating Delays and Unfairness in Appointment Systems.**

In this chapter, we consider an appointment system design in the healthcare system, where heterogeneous participants are sequenced and scheduled for service. As service times are uncertain, the aim is to mitigate the unpleasantness experienced by the participants in the system when their waiting times or delays exceed acceptable thresholds, and address fairness concerning the balancing of service levels among participants. In evaluating uncertain delays, we propose the Delay Unpleasantness Measure (DUM) which takes into account the frequency and intensity of delays above a threshold, and introduce the concept of lexicographic min-max fairness to design appointment systems from the perspective of the worst-off participants. The model can be adapted in the robust setting when the underlying probability distribution is not fully available. To capture the correlation between uncertain service times, we suggest using mean absolute deviation as descriptive statistics in the distributional uncertainty set to preserve linearity of the model. The optimal sequencing and scheduling decisions could be derived by solving a sequence of mixed-integer programming problems and we report the insights from our computational studies.

- **Chapter 5: Conclusions and Future Research.** This chapter concludes the thesis and highlights future research.

1.3 Notation

We adopt the following notations throughout the thesis. We use boldface lowercase characters to represent vectors, for example, $\mathbf{x} = (x_1, x_2, \dots, x_n)$,

and \mathbf{x}' represents the transpose of a vector \mathbf{x} . Given a vector \mathbf{x} , we define (y_i, \mathbf{x}_{-i}) to be the vector with only the i th component being changed, i.e., the vector $(y_i, \mathbf{x}_{-i}) = (x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n)$. $\mathbf{x} \geq \mathbf{y}$ represents the element-wise comparison. We use tilde (\sim) to denote uncertain quantities, for example, \tilde{t} represents a random variable, and $\tilde{\mathbf{c}}$ represents a random vector. We model uncertainty \tilde{t} by a state-space Ω and a σ -algebra of events in Ω . We use \mathcal{V} to represent the set of all real-valued random variables. The inequality between two random variables $\tilde{x} \geq \tilde{y}$ denotes state-wise dominance, i.e., $x(\omega) \geq y(\omega)$ for all $\omega \in \Omega$. To model distributional ambiguity, instead of specifying the true distribution \mathbb{P} on (Ω, \mathcal{F}) , we assume that it belongs to a certain distributional uncertainty set \mathbb{F} , as $\mathbb{P} \in \mathbb{F}$. Accordingly, the case of knowing the exact probability distribution is incorporated in the assumption as well, where $\mathbb{F} = \{\mathbb{P}\}$. We denote by $\mathbb{E}_{\mathbb{P}}(\tilde{t})$ the expectation of \tilde{t} under the probability distribution \mathbb{P} . The cardinality of a set \mathcal{N} is denoted by $|\mathcal{N}|$. For notational simplicity, we use $k \in [1; N]$ and $k \in \{1, \dots, N\}$ interchangeably.

2. PREFERENCES FOR TRAVEL TIME UNDER RISK AND AMBIGUITY

2.1 Introduction

The travel time from an origin to a destination in an urban transportation network is almost always uncertain because of the traffic congestion, which is found to be one the most important factors in the path selection decisions (Abdel-Aty et al. 1995). Individuals' preferences greatly depend on their knowledge about the uncertain travel time as well as their attitudes towards uncertainty. In transportation literatures, an uncertain travel time is often associated with a random variable with the known probability distribution. In other words, the traveler knows the exact frequency of travel time outcomes, and his/her preference relies on his/her risk attitude, that is usually characterized by taking an expectation over a disutility function (an increase in the travel time amounts to a loss). Deliberating on reliability, Mirchandani (1976), Fan et al. (2005) and Nie and Wu (2009) consider the probability of punctuality as a preference criterion, which could be treated as a step disutility function. Unfortunately, since in general, computing the probability of a sum of random variables is NP-hard (Khachiyan 1989), it is a compu-

tationally intractable problem to find the path with the minimum expected disutility over a transportation network, which severely limits our analysis and implementation. Murthy and Sarkar (1998) consider a piece-wise linear concave disutility function, and solve the problem with certain enumeration algorithms. Loui (1983) and Eiger et al. (1985) consider disutility functions in the form of linear, quadratic or exponential, in which the resultant static path selection problems are computationally tractable. In particular, de Palma and Picard (2005) justify empirically the relevance of the exponential disutility function, which appeals to travelers with Constant Absolute Risk Aversion (CARA) and has the best fit on path selection behavior amongst common disutility functions.

Implications of risk in Network Equilibrium (NE) problems, which model a collective behavior of a large population of travelers, have also been studied. One stream suggests using disutility function to capture travel time uncertainty, and travelers' attitudes towards risk (see Mirchandani and Soroush 1987; Yin and Ieda 2001; Chen et al. 2002; Nagurney and Dong 2002; and Yin et al. 2004). The second stream discusses the travel time variability by adding the mean travel time with a safety margin, which can be described by a penalty function (see Noland and Polak 2002; Watling 2006), or the standard deviation (see Uchida and Iida 1993; Lo et al. 2006; Siu and Lo 2008; Connors et al. 2007). However, adding the safety margin in these ways may violate first-order stochastic dominance, and it generally cannot be separated by links, which makes the model hard to solve. We refer interested readers to the review papers of Noland and Polak (2002) and Connors and Sumalee (2009).

Nevertheless, the assumption that travelers know the exact frequency of travel time outcomes is unrealistic. In a real world, it is conceivable that a traveler is incapable of knowing the entire probability distributions of the transportation network. Major exceptional events (e.g., natural disasters) and minor regular events (e.g., minor accident, traffic signal) will incur uncertainty to travel time. Hence, complete distribution of travel time is seldom known exactly, and even the estimated one could be considerably affected by the sampling procedure. If the actual travel time probability distribution is not fully known, then it would be impossible to establish the preferences for travel times based on the expected disutility criterion. However, the discussion on travel time ambiguity is relatively new. Yu and Yang (1998) propose a worst-case shortest path problem over a set of discrete scenarios, which results in an NP -hard problem. Bertsimas and Sim (2003) introduce the “budget of uncertainty” in characterizing uncertain travel time and show that the worst-case shortest path problem is a tractable optimization problem. Ordóñez and Stier-Moses (2010) extend the work to address an NE problem. They generally consider three cases of equilibrium with uncertain travel times: α -percentile equilibrium, added-variability equilibrium, and robust Wardrop equilibrium. The α -percentile equilibrium assumes travelers minimize the α quantile (or Value-at-Risk) of their experienced travel times, which are generally computationally intractable optimization problems. Added-variability equilibrium provides a safety margin to the expected travel time as a proxy to account for risk-averse behavior, an approach that may not be coherent with decision analysis such as violating first order stochastic dominance. Robust Wardrop equilibrium borrows the idea

of Bertsimas and Sim (2003), and assumes that ambiguity averse travelers minimize the worst-case travel time given that the total variation is bounded by a certain parameter. However, the assumptions that the entire population of travelers are only ambiguity averse and not risk sensitive limit the application of this model.

In contrast to the aforementioned works that consider risk and ambiguity separately, our main contribution is to explicitly distinguish between risk and ambiguity in a unified framework in articulating travelers' preferences for travel times. We present a new criterion named *ambiguity-aware CARA travel time* (ACT) for evaluating uncertain travel times for travelers with various attitudes of risk and ambiguity. Apart from the behavioral relevance of the ACT, we also present a computational justification by showing that when the uncertain link travel times are independently distributed, finding the path that minimizes travel time under the ACT criterion is essentially a shortest path problem. We also study the implications on NE problem, in which travelers minimize their own travel times under the ACT criterion, and no traveler can improve his/her travel time under the ACT by unilaterally changing routes. Our new NE model under the ACT criterion shares similar properties with deterministic multi-class NE model, and can be solved by the traditional Frank-Wolfe algorithm. We also examine the inefficiency of this NE model compared with System Optimum (SO), which minimizes the aggregate travel time under the ACT criterion of all travelers, by deriving its Price of Anarchy. The computational study suggests that as uncertainty increases, the influence of selfishness on inefficiency diminishes. Moreover, when uncertainty is neglected in traffic equilibrium analysis, the social op-

timum solution may become more inefficient than the solution under selfish routing.

The remainder of this chapter is organized as follows. In Section 2.2, we formally define the ACT criterion and its properties. In Section 2.3, we investigate a path selection problem under the ACT criterion. In Section 2.4, we extend to the study of the NE problem under the ACT criterion and discuss its computational solvability when the uncertain link travel time is independent with each other. We also analyze the corresponding NE inefficiency by calculating its Price of Anarchy. Finally, in Section 2.5, we make our conclusions and some suggestions for future research.

2.2 Preferences for Travel Time

In the empirical study of de Palma and Picard (2005), they conclude that exponential disutility function, which is the unique disutility function that appeals to travelers with Constant Absolute Risk Aversion (CARA), aptly characterizes travelers' preferences for travel times under risk. Besides, Cheu and Kreinovich (2007) also verify that exponential disutility function is the only function that is consistent with common sense and could simplify the model. Hence, we first introduce the exponential disutility function in the following form,

$$u(x) = \begin{cases} \frac{1}{\lambda} \exp(\lambda x), & \text{when } \lambda \neq 0, \\ ax + b, & \text{when } \lambda = 0, \end{cases}$$

in which $a \in \mathfrak{R}_+$ and the parameter $\lambda \in \mathfrak{R}$ is known as the coefficient of absolute risk aversion. The corresponding certainty equivalent of \tilde{t} , $CE_\lambda(\tilde{t})$:

$\mathcal{V} \rightarrow \mathfrak{R}$ is defined as

$$u(CE_\lambda(\tilde{t})) = \mathbb{E}_\mathbb{P}(u(\tilde{t})).$$

The concept of certainty equivalent $CE_\lambda(\tilde{t})$ is popularized in economic literature, and represents a fixed interval of travel time that the traveler with risk tolerance parameter λ will view equally acceptable as the uncertain travel time \tilde{t} under disutility function $u(\cdot)$. When $u(\cdot)$ is exponential disutility function, we have

$$CE_\lambda(\tilde{t}) = \begin{cases} \frac{1}{\lambda} \ln \mathbb{E}_\mathbb{P}(\exp(\lambda\tilde{t})), & \text{when } \lambda \neq 0, \\ \mathbb{E}_\mathbb{P}(\tilde{t}), & \text{when } \lambda = 0. \end{cases}$$

Parameter λ specifies the traveler's risk attitude. If $\lambda > 0$, he/she is risk-averse and evaluates an uncertain travel time longer than its average. In contrast, a traveler with risk-seeking attitude has $\lambda < 0$ and perceives the uncertain travel time shorter than its average. At neutrality ($\lambda = 0$), the traveler is indifferent between the uncertain travel time and its mean. When travel time is deterministic, we have $CE_\lambda(\text{constant}) = \text{constant}$ for all $\lambda \in \mathfrak{R}$. When travel time follows certain probability distribution, function $CE_\lambda(\tilde{t})$ can be derived through calculating the moment generating function of random variable \tilde{t} . For example, if \tilde{t} is normally distributed $N(\mu, \sigma^2)$, we have $\mathbb{E}_\mathbb{P}(\exp(\lambda\tilde{t})) = \exp(\lambda\mu + \frac{1}{2}\sigma^2\lambda^2)$, and certainty equivalent $CE_\lambda(\tilde{t})$ is

$$CE_\lambda(\tilde{t}) = \mu + \frac{1}{2}\lambda\sigma^2,$$

which is consistent with mean-variance measure (Markowitz 1959) of uncertain travel time \tilde{t} . Note that $CE_\lambda(\tilde{t})$ is different from the mean-variance measure when \tilde{t} follows other kinds of distributions. Moreover, the nice thing about $CE_\lambda(\tilde{t})$ is it preserves first-order stochastic dominance (see for instance Föllmer and Schied 2011), which is violated by the mean-variance measure. Take two paths as an example, one with travel time equal to 1 or 2 with 0.5 probabilities and the other with travel time equal to 3 (with certainty). Though the first path stochastically dominates the second, mean-variance measure would favor the second path for an extremely risk-averse traveler, while the CARA model always supports the first path, as the certainty equivalent of the first is always less than that of the second.

If the actual travel time probability distribution is not fully known, then it would be impossible to establish preferences for travel times based on the expected disutility criterion. The CARA model could not reveal travelers' preferences when facing ambiguity. We study the preference for uncertain travel times in which the traveler is oblivious to the true probability distribution \mathbb{P} but knows the distributional uncertainty set \mathbb{F} , which can be characterized by certain descriptive statistics. The "size" of the set \mathbb{F} indicates the level of ambiguity perceived by the traveler. For instance, the distributional uncertainty set perceived by an informed traveler may be a subset of that perceived by a clueless traveler. To evaluate an ambiguity preference, the Hurwicz criterion (Hurwicz 1951) represents a compromise between the worst-case and the best-case evaluation of travel time under

distributional ambiguity as follows:

$$H_\alpha(\tilde{t}) = \alpha \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}),$$

where the parameter $\alpha \in [0, 1]$ indicates the level of optimism, with $\alpha = 0$ being the most optimistic and $\alpha = 1$ being the most pessimistic.

2.2.1 Ambiguity-aware CARA travel time (ACT)

Instead of considering risk and ambiguity separately, we explicitly distinguish between them in a unified framework for articulating travelers' preferences for travel times. We propose the ambiguity-aware CARA travel time (ACT) criterion for evaluating an uncertain travel time under various attitudes of risk and ambiguity, which is based on blending Hurwicz and Constant Absolute Risk Aversion (CARA) criteria.

The traveler has a distributional uncertainty set \mathbb{F} to characterize the uncertain travel time. Similar to the Hurwicz criterion, his/her attitude towards ambiguity is described by parameter $\alpha \in [0, 1]$ and risk attitude under CARA is given by parameter $\lambda \in \mathfrak{R}$. Accordingly, we identify the traveler under the ACT by $V = (\alpha, \lambda, \mathbb{F})$.

Definition 2.1. The ambiguity-aware CARA travel time $\text{ACT}_V(\tilde{t}) : \mathcal{V} \rightarrow \mathfrak{R}$

specified by the traveler with parameter $V = (\alpha, \lambda, \mathbb{F})$ is

$$\begin{aligned} & \text{ACT}_V(\tilde{t}) \\ = & \begin{cases} \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})), & \text{when } \lambda \neq 0, \\ \alpha \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}), & \text{when } \lambda = 0. \end{cases} \end{aligned}$$

Observing that if the probability distribution is known, i.e., $\mathbb{F} = \{\mathbb{P}\}$, we have

$$\begin{aligned} & \text{ACT}_V(\tilde{t}) \\ = & \text{ACT}_{(\alpha, \lambda, \mathbb{P})}(\tilde{t}) \\ = & \begin{cases} \alpha \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) + (1 - \alpha) \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})), & \text{when } \lambda \neq 0, \\ \alpha \mathbb{E}_{\mathbb{P}}(\tilde{t}) + (1 - \alpha) \mathbb{E}_{\mathbb{P}}(\tilde{t}), & \text{when } \lambda = 0. \end{cases} \\ = & \text{CE}_{\lambda}(\tilde{t}). \end{aligned}$$

Hence, the ACT criterion is a generalization of certainty equivalent function under CARA. It is a weighted sum of the best-case certainty equivalent and the worst-case certainty equivalent when the true probability distribution belongs to a distributional uncertainty set. $\alpha = 0$ represents an extremely ambiguity seeking traveler, while $\alpha = 1$ representing an extremely ambiguity averse traveler. To quantitatively characterize travelers' attitudes towards risk and ambiguity, economists have summarized the procedure to sought these two parameters α, λ in experimental studies (see for instance Wakker 2010; Abdellaoui et al. 2011). We believe this could shed some light on the future empirical studies on travelers' preferences. Next, we

provide some useful properties of the ACT criterion. For any given distributional uncertainty set \mathbb{F} , we first define the corresponding bound as $\bar{t}_{\mathbb{F}} = \inf \{t \in \mathfrak{R} | \mathbb{P}(\tilde{t} \leq t) = 1, \forall \mathbb{P} \in \mathbb{F}\}$ and $\underline{t}_{\mathbb{F}} = \sup \{t \in \mathfrak{R} | \mathbb{P}(\tilde{t} \geq t) = 1, \forall \mathbb{P} \in \mathbb{F}\}$.

Proposition 2.1.

(a) $\text{ACT}_V(\tilde{t})$ is nondecreasing in $\lambda \in \mathfrak{R}$ and $\alpha \in [0, 1]$, and

$$\lim_{\lambda \rightarrow +\infty} \text{ACT}_{(1, \lambda, \mathbb{F})}(\tilde{t}) = \bar{t}_{\mathbb{F}}, \quad \lim_{\lambda \rightarrow -\infty} \text{ACT}_{(0, \lambda, \mathbb{F})}(\tilde{t}) = \underline{t}_{\mathbb{F}}.$$

(b) For any $\tilde{x}, \tilde{y} \in \mathcal{V}$, if $\tilde{x} \geq \tilde{y}$, we have $\text{ACT}_V(\tilde{x}) \geq \text{ACT}_V(\tilde{y})$;

(c) Suppose $\tilde{t}_1, \dots, \tilde{t}_J$ are independent random variables, and $t_0 \in \mathfrak{R}$. Then

$$\text{ACT}_V\left(t_0 + \sum_{j=1}^J \tilde{t}_j\right) = t_0 + \sum_{j=1}^J \text{ACT}_V(\tilde{t}_j).$$

Proof. (a) Note that $\text{ACT}_V(\tilde{t})$ being nondecreasing in α follows directly from $\sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) \geq \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}))$ and $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}) \geq \inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t})$. Based on Jensen's inequality, for any $\lambda_1 \leq \lambda_2 < 0$ or $0 < \lambda_1 \leq \lambda_2$, we can get

$$\begin{aligned} & \frac{1}{\lambda_2} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda_2 \tilde{t})) \\ &= \frac{1}{\lambda_2} \ln \mathbb{E}_{\mathbb{P}}\left(\left(\exp(\lambda_1 \tilde{t})\right)^{\lambda_2/\lambda_1}\right) \\ &\geq \frac{1}{\lambda_2} \frac{\lambda_2}{\lambda_1} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda_1 \tilde{t})) = \frac{1}{\lambda_1} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda_1 \tilde{t})). \end{aligned}$$

When $\lambda_1 < 0 < \lambda_2$, we have

$$\frac{1}{\lambda_2} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_2 \tilde{t})) \geq \frac{1}{\lambda_2} \ln \exp (\mathbb{E}_{\mathbb{P}} (\lambda_2 \tilde{t})) = \mathbb{E}_{\mathbb{P}} (\tilde{t}) \geq \frac{1}{\lambda_1} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_1 \tilde{t})).$$

Therefore, for any $\lambda_1 \leq \lambda_2$,

$$\begin{aligned} \text{ACT}_{(\alpha, \lambda_2, \mathbb{F})} (\tilde{t}) &= \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda_2} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_2 \tilde{t})) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda_2} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_2 \tilde{t})) \\ &\geq \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda_1} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_1 \tilde{t})) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda_1} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda_1 \tilde{t})) \\ &= \text{ACT}_{(\alpha, \lambda_1, \mathbb{F})} (\tilde{t}). \end{aligned}$$

Equivalently, $\text{ACT}_V (\tilde{t})$ is nondecreasing in λ .

When $\alpha = 1$, the traveler is most pessimistic towards ambiguity, then

$$\text{ACT}_{(1, \lambda, \mathbb{F})} (\tilde{t}) = \begin{cases} \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda \tilde{t})), & \text{when } \lambda \neq 0, \\ \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\tilde{t}), & \text{when } \lambda = 0. \end{cases}$$

We have for any $\mathbb{P} \in \mathbb{F}$ and $\lambda \in \mathfrak{R} \setminus \{0\}$,

$$\frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}} (\exp (\lambda \tilde{t})) \leq \frac{1}{\lambda} \ln (\exp (\lambda \bar{t}_{\mathbb{F}})) = \bar{t}_{\mathbb{F}}.$$

Therefore,

$$\lim_{\lambda \rightarrow +\infty} \text{ACT}_{(1, \lambda, \mathbb{F})} (\tilde{t}) \leq \bar{t}_{\mathbb{F}}.$$

Moreover, according to the definition of $\bar{t}_{\mathbb{F}}$, for any $\epsilon > 0$, $\exists \mathbb{P} \in \mathbb{F}$ such that

$\mathbb{P}(\tilde{t} \in [\bar{t}_{\mathbb{F}} - \epsilon, \bar{t}_{\mathbb{F}}]) > 0$, we have

$$\begin{aligned}
& \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) \\
&= \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \bar{t}_{\mathbb{F}}) \exp(\lambda(\tilde{t} - \bar{t}_{\mathbb{F}}))) \\
&= \bar{t}_{\mathbb{F}} + \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda(\tilde{t} - \bar{t}_{\mathbb{F}}))) \\
&\geq \bar{t}_{\mathbb{F}} + \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln(\exp(\lambda(\bar{t}_{\mathbb{F}} - \epsilon - \bar{t}_{\mathbb{F}})) \mathbb{P}(\tilde{t} \in [\bar{t}_{\mathbb{F}} - \epsilon, \bar{t}_{\mathbb{F}}])) \\
&= \bar{t}_{\mathbb{F}} + \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln(\exp(-\lambda\epsilon)) + \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln(\mathbb{P}(\tilde{t} \in [\bar{t}_{\mathbb{F}} - \epsilon, \bar{t}_{\mathbb{F}}])) \\
&= \bar{t}_{\mathbb{F}} - \epsilon,
\end{aligned}$$

which means,

$$\lim_{\lambda \rightarrow +\infty} \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) \geq \lim_{\lambda \rightarrow +\infty} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) \geq \bar{t}_{\mathbb{F}} - \epsilon \quad \forall \epsilon > 0.$$

Combining these two inequalities together, we have

$$\lim_{\lambda \rightarrow +\infty} \text{ACT}_{(1, \lambda, \mathbb{F})}(\tilde{t}) = \bar{t}_{\mathbb{F}}.$$

Similarly, we can modify the above proof to show that

$$\lim_{\lambda \rightarrow -\infty} \text{ACT}_{(0, \lambda, \mathbb{F})}(\tilde{t}) = \underline{t}_{\mathbb{F}}.$$

(b) If $\tilde{x} \geq \tilde{y}$ i.e., $x(\omega) \geq y(\omega)$ for all $\omega \in \Omega$, we have when $\lambda = 0$,

$$\text{ACT}_V(\tilde{x}) = \alpha \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{x}) + (1-\alpha) \inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{x}) \geq \alpha \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{y}) + (1-\alpha) \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{y}) = \text{ACT}_V(\tilde{y}).$$

When $\lambda \neq 0$, noting that $\frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{x})) \geq \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{y}))$ for all $\mathbb{P} \in \mathbb{F}$, we have

$$\begin{aligned} \text{ACT}_V(\tilde{x}) &= \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{x})) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{x})) \\ &\geq \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{y})) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{y})) \\ &= \text{ACT}_V(\tilde{y}). \end{aligned}$$

(c) Since $\tilde{t}_1, \dots, \tilde{t}_J$ are independently distributed, we have

$$\begin{aligned} &\text{ACT}_V\left(t_0 + \sum_{j=1}^J \tilde{t}_j\right) \\ &= \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}\left(\exp\left(\lambda\left(t_0 + \sum_{j=1}^J \tilde{t}_j\right)\right)\right) \\ &\quad + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}\left(\exp\left(\lambda\left(t_0 + \sum_{j=1}^J \tilde{t}_j\right)\right)\right) \\ &= \alpha t_0 + \alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \left(\prod_{j=1}^J \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}_j))\right) \\ &\quad + (1 - \alpha)t_0 + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \left(\prod_{j=1}^J \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}_j))\right) \\ &= t_0 + \sum_{j=1}^J \left(\alpha \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}_j)) + (1 - \alpha) \inf_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}_j))\right) \\ &= t_0 + \sum_{j=1}^J \text{ACT}_V(\tilde{t}_j). \quad \square \end{aligned}$$

Remark 2.1. Property (a) is a trivial statement, it indicates that when a traveler is more risk-averse or ambiguity averse than the others, he/she perceives the uncertain travel time longer than the others' perception. The extreme

cases occur when $\lambda = \infty, \alpha = 1$ and $\lambda = -\infty, \alpha = 0$, respectively. When a traveler is extremely risk-averse and ambiguity averse, he/she pessimistically regards the uncertain travel time from the worst-case perspective, and the corresponding $\text{ACT}_V(\tilde{t})$ takes the largest possible value. Property (b) captures traveler's essential preference for a shorter travel time. His/her perceived travel time becomes longer when the travel time increases. Property (c) suggests that $\text{ACT}_V(\cdot)$ is additive for independent random variables. This property is quite helpful for modeling, since $\text{ACT}_V(\cdot)$ along a path could be easily separated by links.

Next, we will provide an example to illustrate travelers' preferences for travel times under the ACT criterion. Figure 2.1 shows three paths from the origin O to the destination D. Travel time on path A is deterministic, 1.5hrs; travel time on path B is stochastic and the duration is 1hr or 2hrs with equal probability; travel time on path C is uncertain, and bounded by 1hr and 2hrs. We present in Table 2.1 the path preferences induced by the ACT criterion under various attitudes and degrees of risk and ambiguity.

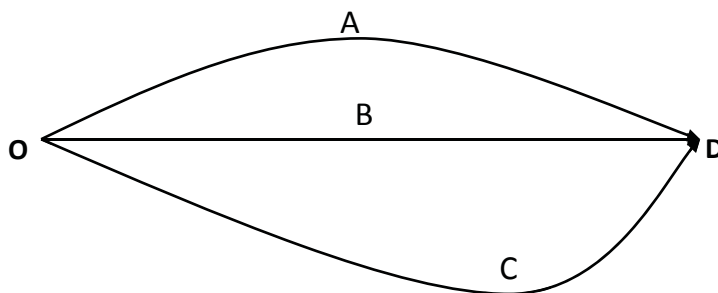


Fig. 2.1: A simple network with uncertain travel time.

When a traveler is extremely risk-averse and pessimistic towards ambiguity ($\lambda \rightarrow +\infty, \alpha = 1$) as property (a) described, he/she will perceive the uncertain travel time as taking the longest duration. Hence, path A is preferred as it has the smallest ACT. On the other hand, when the traveler is radically risk-seeking and optimistic towards ambiguity ($\lambda \rightarrow -\infty, \alpha = 0$), then path A would be least preferred. At risk neutrality, both paths A and B are equally preferred and the preference for path C depends on the traveler's attitude towards ambiguity. For instance, if he/she is optimistic towards ambiguity, then path C will be preferred over paths A and B.

Risk attitude λ	Ambiguity attitude α	$ACT_V(\tilde{t}_A)$	$ACT_V(\tilde{t}_B)$	$ACT_V(\tilde{t}_C)$	Preferences
$+\infty$	1	1.5	2	2	$A \succ B \sim C$
0	1	1.5	1.5	2	$A \sim B \succ C$
0	0	1.5	1.5	1	$C \succ A \sim B$
$-\infty$	0	1.5	1	1	$B \sim C \succ A$

Tab. 2.1: Preferences for travel times under the ACT criterion.

2.2.2 Two uncertainty models for travel time

If the probability distribution of an uncertain travel time \tilde{t} is completely known, there exists no ambiguity, and $ACT_V(\tilde{t})$ reduces to $CE_\lambda(\tilde{t})$, which can be calculated directly. When the probability distribution is not fully available, the characterization of uncertain travel time can be in various ways depending on the available information. We then propose two simple models for characterizing the uncertain travel time and provide analytical forms of the ACT criterion.

Uncertainty model I

Driven by pragmatism, the traveler may have a simple description of the uncertain travel time by providing the ranges in which travel time and average travel time would fall within. Specifically, the travel time takes values in $[\underline{t}, \bar{t}]$, $0 < \underline{t} \leq \bar{t}$ and the average travel time falls within the range $[\underline{\mu}, \bar{\mu}] \subseteq [\underline{t}, \bar{t}]$. Hence, the distributional uncertainty set \mathbb{F} of the uncertain travel time \tilde{t} is given by

$$\mathbb{F} = \{ \mathbb{P} \mid \mathbb{E}_{\mathbb{P}}(\tilde{t}) \in [\underline{\mu}, \bar{\mu}], \mathbb{P}(\tilde{t} \in [\underline{t}, \bar{t}]) = 1 \}. \quad (2.1)$$

Proposition 2.2. Given a distributional uncertainty set \mathbb{F} described by (2.1), the uncertain travel time under the ACT criterion is

$$\text{ACT}_V(\tilde{t}) = \begin{cases} \frac{\alpha}{\lambda} \ln \left(\frac{(\bar{t}-\bar{\mu}) \exp(\lambda \underline{t}) + (\bar{\mu}-\underline{t}) \exp(\lambda \bar{t})}{\bar{t}-\underline{t}} \right) + (1-\alpha)\underline{\mu}, & \text{when } \lambda > 0, \\ \alpha\bar{\mu} + \frac{1-\alpha}{\lambda} \ln \left(\frac{(\bar{t}-\underline{\mu}) \exp(\lambda \underline{t}) + (\underline{\mu}-\underline{t}) \exp(\lambda \bar{t})}{\bar{t}-\underline{t}} \right), & \text{when } \lambda < 0, \\ \alpha\bar{\mu} + (1-\alpha)\underline{\mu}, & \text{when } \lambda = 0. \end{cases}$$

Moreover,

$$\begin{aligned} \lim_{\lambda \rightarrow +\infty} \text{ACT}_V(\tilde{t}) &= \alpha\bar{t} + (1-\alpha)\underline{\mu}, \\ \lim_{\lambda \rightarrow -\infty} \text{ACT}_V(\tilde{t}) &= (1-\alpha)\underline{t} + \alpha\bar{\mu}. \end{aligned}$$

Proof. We first provide the analytical expressions for $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}))$ and $\inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t}))$. According to Proposition 3 in Brown et al. (2012),

$$\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\exp(\lambda \tilde{t})) = \begin{cases} \frac{(\bar{t}-\bar{\mu}) \exp(\lambda \underline{t}) + (\bar{\mu}-\underline{t}) \exp(\lambda \bar{t})}{\bar{t}-\underline{t}}, & \text{when } \lambda > 0, \\ \frac{(\bar{t}-\underline{\mu}) \exp(\lambda \underline{t}) + (\underline{\mu}-\underline{t}) \exp(\lambda \bar{t})}{\bar{t}-\underline{t}}, & \text{when } \lambda < 0. \end{cases}$$

To determine $\inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp (\lambda \tilde{t}))$, we note that by Jensen's inequality,

$$\mathbb{E}_{\mathbb{P}} (\exp (\lambda \tilde{t})) \geq \exp (\mathbb{E}_{\mathbb{P}} (\lambda \tilde{t})) = \exp (\lambda \mathbb{E}_{\mathbb{P}} (\tilde{t})),$$

consequently,

$$\inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp (\lambda \tilde{t})) \geq \begin{cases} \exp (\lambda \underline{\mu}), & \text{when } \lambda > 0, \\ \exp (\lambda \bar{\mu}), & \text{when } \lambda < 0. \end{cases}$$

Equality holds when \tilde{t} is deterministic,

$$\begin{cases} \mathbb{P} (\tilde{t} = \underline{\mu}) = 1, & \text{when } \lambda > 0; \\ \mathbb{P} (\tilde{t} = \bar{\mu}) = 1, & \text{when } \lambda < 0. \end{cases}$$

Note that this distribution also belongs to the distributional uncertainty set \mathbb{F} , and $\text{ACT}_V (\tilde{t})$ can be accordingly calculated. Based on L'Hôpital's rule,

when $\lambda > 0$,

$$\begin{aligned}
& \lim_{\lambda \rightarrow +\infty} \text{ACT}_V(\tilde{t}) \\
&= \lim_{\lambda \rightarrow +\infty} \left(\frac{\alpha}{\lambda} \ln \left(\frac{(\bar{t} - \bar{\mu}) \exp(\lambda \underline{t}) + (\bar{\mu} - \underline{t}) \exp(\lambda \bar{t})}{\bar{t} - \underline{t}} \right) + (1 - \alpha) \underline{\mu} \right) \\
&= \lim_{\lambda \rightarrow +\infty} \left(\frac{\alpha}{\lambda} \ln \left(\frac{(\bar{t} - \bar{\mu}) \exp(\lambda \underline{t}) + (\bar{\mu} - \underline{t}) \exp(\lambda \bar{t})}{\bar{t} - \underline{t}} \right) \right) + (1 - \alpha) \underline{\mu} \\
&= \alpha \lim_{\lambda \rightarrow +\infty} \left(\frac{(\bar{t} - \bar{\mu}) \exp(\lambda \underline{t}) \underline{t} + (\bar{\mu} - \underline{t}) \exp(\lambda \bar{t}) \bar{t}}{(\bar{t} - \bar{\mu}) \exp(\lambda \underline{t}) + (\bar{\mu} - \underline{t}) \exp(\lambda \bar{t})} \right) + (1 - \alpha) \underline{\mu} \\
&= \alpha \lim_{\lambda \rightarrow +\infty} \left(\frac{(\bar{t} - \bar{\mu}) \exp(\lambda(\underline{t} - \bar{t})) \underline{t} + (\bar{\mu} - \underline{t}) \bar{t}}{(\bar{t} - \bar{\mu}) \exp(\lambda(\underline{t} - \bar{t})) + (\bar{\mu} - \underline{t})} \right) + (1 - \alpha) \underline{\mu} \\
&= \alpha \bar{t} + (1 - \alpha) \underline{\mu}.
\end{aligned}$$

Likewise, the result could extend to

$$\lim_{\lambda \rightarrow -\infty} \text{ACT}_V(\tilde{t}) = (1 - \alpha) \underline{t} + \alpha \bar{\mu}. \quad \square$$

We further analyze paths preferences on the simple network depicted in Figure 2.1 as an example.

Example: In Figure 2.1, travel times on path A and C remain unchanged. As for path B, we now assume that the travel time is within 1hr to 2hrs, and the mean travel time is exactly 1.5hrs. Given the above information of three paths, travelers' preferences ranked by the ACT criterion are summarized in Table 2.2.

To show the results in Table 2.2, from Proposition 2.2, we calculate the travel time under the ACT criterion for each of the three paths. The

Risk attitude λ	Ambiguity attitude α	Preferences
$[0, +\infty)$	$[f(\lambda), 1]^1$	$A \succeq B \succeq C$
$[0, +\infty)$	$[\frac{1}{2}, f(\lambda)]$	$A \succeq C \succeq B$
$[0, +\infty)$	$[0, \frac{1}{2}]$	$C \succeq A \succeq B$
$(-\infty, 0]$	$[\frac{1}{2}, 1]$	$B \succeq A \succeq C$
$(-\infty, 0]$	$[g(\lambda), \frac{1}{2}]^2$	$B \succeq C \succeq A$
$(-\infty, 0]$	$[0, g(\lambda)]$	$C \succeq B \succeq A$

$$^1 f(\lambda) = \frac{\lambda}{3\lambda + 2\ln 2 - 2\ln(1 + \exp(\lambda))};$$

$$^2 g(\lambda) = \frac{2\ln(1 + \exp(\lambda)) - 2\ln 2}{\lambda + 2\ln(1 + \exp(\lambda)) - 2\ln 2}.$$

Tab. 2.2: Path preferences under the ACT criterion.

information is specified as follows:

$$\begin{aligned} \underline{t}_A &= 1.5, & \bar{t}_A &= 1.5, & \underline{\mu}_A &= 1.5, & \bar{\mu}_A &= 1.5, \\ \underline{t}_B &= 1, & \bar{t}_B &= 2, & \underline{\mu}_B &= 1.5, & \bar{\mu}_B &= 1.5, \\ \underline{t}_C &= 1, & \bar{t}_C &= 2, & \underline{\mu}_C &= 1, & \bar{\mu}_C &= 2. \end{aligned}$$

Therefore, the ACT can be calculated correspondingly,

$$\begin{aligned} \text{ACT}_V(t_A) &= \frac{3}{2}; \\ \text{ACT}_V(\tilde{t}_B) &= \begin{cases} \frac{\alpha}{\lambda} \ln \left(\frac{1}{2} \exp(\lambda) + \frac{1}{2} \exp(2\lambda) \right) + \frac{3}{2}(1 - \alpha), & \text{when } \lambda > 0, \\ \frac{3}{2}\alpha + \frac{1 - \alpha}{\lambda} \ln \left(\frac{1}{2} \exp(\lambda) + \frac{1}{2} \exp(2\lambda) \right), & \text{when } \lambda < 0, \\ \frac{3}{2}\alpha + \frac{3}{2}(1 - \alpha), & \text{when } \lambda = 0; \end{cases} \\ \text{ACT}_V(\tilde{t}_C) &= \begin{cases} \frac{\alpha}{\lambda} \ln(\exp(2\lambda)) + (1 - \alpha), & \text{when } \lambda > 0, \\ 2\alpha + \frac{1 - \alpha}{\lambda} \ln(\exp(\lambda)), & \text{when } \lambda < 0, \\ 2\alpha + (1 - \alpha), & \text{when } \lambda = 0 \end{cases} \\ &= 1 + \alpha. \end{aligned}$$

Since the travel time under the ACT criterion is nondecreasing in both λ and

α , the preference relationships between paths A and B, and between paths A and C can be readily established. When $\lambda \geq 0$, we have $A \succeq B$. Likewise, when $1 \geq \alpha \geq \frac{1}{2}$, then $A \succeq C$. Hence, we focus on the preferences between paths B and C.

$\text{ACT}_V(\tilde{t}_B) \geq \text{ACT}_V(\tilde{t}_C)$ implies

$$\begin{cases} \frac{\alpha}{\lambda} \ln \left(\frac{1}{2} \exp(\lambda) + \frac{1}{2} \exp(2\lambda) \right) + \frac{3}{2}(1 - \alpha) \geq 1 + \alpha, & \text{when } \lambda > 0, \\ \frac{3}{2}\alpha + \frac{1 - \alpha}{\lambda} \ln \left(\frac{1}{2} \exp(\lambda) + \frac{1}{2} \exp(2\lambda) \right) \geq 1 + \alpha, & \text{when } \lambda < 0, \\ \frac{3}{2} \geq 1 + \alpha, & \text{when } \lambda = 0. \end{cases}$$

Equivalently, path C is preferred to path B when

$$\begin{cases} \alpha \leq f(\lambda) = \frac{\lambda}{3\lambda + 2 \ln 2 - 2 \ln(1 + \exp(\lambda))}, & \text{when } \lambda > 0, \\ \alpha \leq g(\lambda) = \frac{2 \ln(1 + \exp(\lambda)) - 2 \ln 2}{\lambda + 2 \ln(1 + \exp(\lambda)) - 2 \ln 2}, & \text{when } \lambda < 0. \end{cases}$$

The preferences expressed by travelers with varied λ and α are depicted in Figure 2.2. When the traveler is risk-averse ($\lambda > 0$), he/she prefers path A over path B, and the converse is true when the traveler is risk-seeking. With α decreases from 1 to 0, the traveler's attitude towards ambiguity shifts from being pessimistic to optimistic, in which case, path C, which has complete ambiguity, will become more favorable. This example may suggest a way to empirically identify travelers' attitudes towards risk and ambiguity by providing travelers with different choice scenarios.

Uncertainty model II

In practice, the uncertain travel time only takes a set of discrete values. Next,

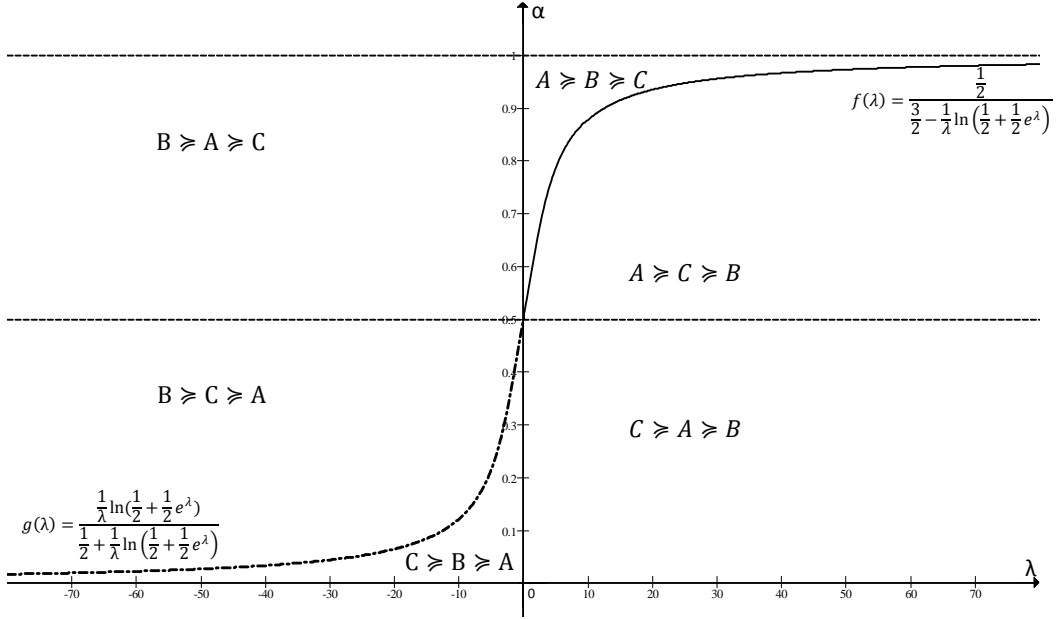


Fig. 2.2: Path preferences under different attitudes towards risk and ambiguity.

we present a general model for this discrete case where the uncertain travel time has finite realizations, for example, t_1, \dots, t_M , and more statistics on the moment are available, i.e., the distributional uncertainty set \mathbb{F} is given by

$$\mathbb{F} = \left\{ \mathbb{P} \left| \mathbb{E}_{\mathbb{P}}(\tilde{t}^{l_k}) \in [\underline{\mu}_k, \bar{\mu}_k], \quad k = 1, \dots, K, \quad \mathbb{P}(\tilde{t} \in \{t_1, \dots, t_M\}) = 1 \right. \right\}, \quad (2.2)$$

where $l_k \in \mathcal{Z}_+, k = 1, \dots, K$.

Proposition 2.3. If the distributional uncertainty set \mathbb{F} is described by (2.2), the uncertain travel time under the ACT criterion can be derived by solving

two linear optimization problems,

$$\begin{aligned} & \text{ACT}_V(\tilde{t}) \\ = & \begin{cases} \alpha \frac{1}{\lambda} \ln \left(\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) \right) + (1 - \alpha) \frac{1}{\lambda} \ln \left(\inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) \right), & \text{when } \lambda > 0, \\ (1 - \alpha) \frac{1}{\lambda} \ln \left(\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) \right) + \alpha \frac{1}{\lambda} \ln \left(\inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) \right), & \text{when } \lambda < 0, \\ \alpha \bar{\mu}_1 + (1 - \alpha) \underline{\mu}_1, & \text{when } \lambda = 0. \end{cases} \end{aligned}$$

where

$$\begin{aligned} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) &= \max_{(p_1, \dots, p_M) \in \mathcal{P}} \sum_{m=1}^M p_m \exp(\lambda t_m) \\ \inf_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\exp(\lambda \tilde{t})) &= \min_{(p_1, \dots, p_M) \in \mathcal{P}} \sum_{m=1}^M p_m \exp(\lambda t_m) \end{aligned}$$

and

$$\mathcal{P} = \left\{ (p_1, \dots, p_M) \left| \begin{array}{l} \sum_{m=1}^M p_m t_m^{l_k} \leq \bar{\mu}_k, \quad k = 1, \dots, K, \\ \sum_{m=1}^M p_m t_m^{l_k} \geq \underline{\mu}_k, \quad k = 1, \dots, K, \\ \sum_{m=1}^M p_m = 1, \\ p_m \geq 0, \quad m = 1, \dots, M. \end{array} \right. \right\}.$$

Proof. The proof for this proposition is rather straight forward. \square

2.3 Path Selection under the ACT Criterion

In this section, we study the problem of selecting the path that minimizes the ACT criterion when the link travel times on the network are uncertain. We consider a directed network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ and let \mathcal{R} be the set of all admissible

paths, which are sets of links connecting the origin node to the destination node. The uncertain travel time along the link $a \in \mathcal{A}$ is denoted by \tilde{t}_a .

The deterministic version of this path selection problem or shortest path problem is well known to be polynomial time solvable. When the travel times are uncertain, the path selection problem that minimizes the travel time under the ACT criterion is given by

$$\min_{r \in \mathcal{R}} \text{ACT}_V \left(\sum_{a \in r} \tilde{t}_a \right). \quad (2.3)$$

In Proposition 2.4 below, we show that the solvability of Problem (2.3) depends on whether the uncertain link travel times are correlated.

Proposition 2.4. (a) If the uncertain link travel times are independently distributed, then Problem (2.3) is a shortest path problem on the same network in which the link travel time on $a \in \mathcal{A}$ is given by $\text{ACT}_V(\tilde{t}_a)$.

(b) If the uncertain link travel times are correlated, then the recognition version of Problem (2.3) is NP-complete.

Proof. (a) According to Proposition 2.1, if the link travel times are independently distributed, the objective function in Problem (2.3) can be written additively as

$$\text{ACT}_V \left(\sum_{a \in r} \tilde{t}_a \right) = \sum_{a \in r} \text{ACT}_V(\tilde{t}_a).$$

In this case, we can regard the travel time under the ACT criterion along each link as the deterministic link travel time, and polynomially solve it by

the shortest path algorithm.

(b) We will prove its NP-complete by reduction from the following problem, which is proved to be NP-complete by Yu and Yang (1998):

$$\min_{r \in \mathcal{R}} \max \left\{ \sum_{a \in r} t_a^1, \sum_{a \in r} t_a^2 \right\}, \quad (2.4)$$

where t_a^1 and t_a^2 are two travel time scenarios on link $a \in \mathcal{A}$.

We construct an instance of Problem (2.3), in which the uncertain travel time on link a is

$$\tilde{t}_a = \frac{1}{2} (t_a^1 + t_a^2) + \frac{1}{2} (t_a^1 - t_a^2) \tilde{z}, \quad \forall a \in \mathcal{A},$$

that is, the travel times of all the links are influenced by a common random variable \tilde{z} , which we assume is $+1$ or -1 with equal probability. Hence, for an extremely risk-averse and pessimistic towards ambiguity traveler ($\lambda \rightarrow +\infty$, $\alpha = 1$), finding a path with minimum travel time under the ACT criterion from the origin node to the destination node can be written as

$$\min_{r \in \mathcal{R}} \limsup_{\lambda \rightarrow +\infty} \sup_{\mathbb{P} \in \mathbb{F}} \frac{1}{\lambda} \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\lambda \sum_{a \in r} \left(\frac{1}{2} (t_a^1 + t_a^2) + \frac{1}{2} (t_a^1 - t_a^2) \tilde{z} \right) \right) \right).$$

According to Proposition 2.1, it can be simplified further as

$$\min_{r \in \mathcal{R}} \sum_{a \in r} \frac{1}{2} (t_a^1 + t_a^2) + \max \left\{ \sum_{a \in r} \frac{1}{2} (t_a^1 - t_a^2), \sum_{a \in r} \frac{1}{2} (t_a^2 - t_a^1) \right\},$$

which could be equivalently written as Problem (2.4). Thus, Problem (2.3) is NP-complete. \square

Proposition 2.4 shows that when the link travel times are independently distributed, we can easily find the optimal path under the ACT criterion, which accounts for both risk and ambiguity. The result, though simple, shows that the ACT criterion not only is descriptive relevant by being able to account for a traveler's different attitudes of risk and ambiguity over uncertain travel times, but also can be used normatively to find the most preferred path using modest computational effort.

2.4 Analysis of Network Equilibrium with Risk and

Ambiguity Aware Travelers

We study the network equilibrium problem when travelers are sensitive to risk and ambiguity and evaluate the travel times along paths using the ACT criterion. In section 2.4.1, we characterize the network equilibrium such that no traveler could improve his/her travel time under the ACT criterion by unilaterally changing routes. In section 2.4.2, we investigate the inefficiency of the NE by comparing with the System Optimal solution that minimizes the total travel time under the ACT criterion of all travelers. We also provide a simple network equilibrium study in section 2.4.3.

2.4.1 Network equilibrium formulation

Given a network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, we let $\mathcal{W} \subseteq \mathcal{N} \times \mathcal{N}$ be a set of Origin-Destination (OD) pairs, and \mathcal{R}_w be a set of all simple paths connecting a given OD pair $w \in \mathcal{W}$. To derive a tractable model, we assume that the uncertain link travel times are independently distributed. We define the uncertain travel time along link $a \in \mathcal{A}$ as

$$\tilde{t}_a(v_a) = s_a(v_a)\tilde{z}_a + \tilde{\tau}_a,$$

where $s_a(v_a)$ is a differentiable, monotonically increasing function in its own link traffic flow v_a , and $\tilde{z}_a, \tilde{\tau}_a, a \in \mathcal{A}$ are independently distributed nonnegative random variables. The multiplicative uncertainty \tilde{z}_a can be interpreted as the flow dependent disturbance, while $\tilde{\tau}_a$, the additive uncertainty, is the flow independent disturbance.

For generality, we allow travelers to have different perceptions on uncertainty in link travel times. For example, a local resident, who is very familiar with the area, would be less ambiguous, compared to a tourist, in characterizing the uncertain travel times along the network links. To characterize the heterogeneity, we classify all travelers on the network into n types. The i th type of travelers, $i \in \mathcal{I} = \{1, \dots, n\}$ are characterized by their risk parameter λ_i , ambiguity parameter α_i , and their distributional uncertainty set \mathbb{F}_i of the travel times on the network. For notational convenience, we denote $V_i = (\lambda_i, \alpha_i, \mathbb{F}_i)$. Under the ACT criterion, the uncertain travel time $\tilde{t}_a(v_a)$

perceived by the i th type of travelers is given by

$$\begin{aligned} t_{ai}(v_a) &= \text{ACT}_{V_i}(\tilde{t}_a(v_a)) \\ &= \text{ACT}_{V_i}(s_a(v_a)\tilde{z}_a + \tilde{\tau}_a) \\ &= \text{ACT}_{V_i}(s_a(v_a)\tilde{z}_a) + \text{ACT}_{V_i}(\tilde{\tau}_a). \end{aligned}$$

For a given OD pair $w \in \mathcal{W}$, let d_{wi} be the number of trips made by the i th type of travelers and f_{ri} be the flow on path $r \in \mathcal{R}_w$ contributed by the i th type of travelers, and $\mathbf{f} = (f_{ri})_{r \in \mathcal{R}_w, w \in \mathcal{W}, i \in \mathcal{I}}$ is the vector of flows of all travelers along all paths. The aggregate flow on link $a \in \mathcal{A}$ is

$$v_a(\mathbf{f}) = \sum_{i \in \mathcal{I}} \sum_{w \in \mathcal{W}} \sum_{r \in \mathcal{R}_w} f_{ri} \delta_{ar},$$

where δ_{ar} equals 1 if the link a is along the path r and 0 otherwise. Moreover, since the travel time along any path r is given by

$$\tilde{c}_r(\mathbf{f}) = \sum_{a \in \mathcal{A}} \tilde{t}_a(v_a(\mathbf{f})) \delta_{ar},$$

the travel time along path $r \in \mathcal{R}_w$ under the ACT criterion perceived by the

i th type of travelers is given by

$$\begin{aligned}
c_{ri}(\mathbf{f}) &= \text{ACT}_{V_i}(\tilde{c}_r(\mathbf{f})) \\
&= \text{ACT}_{V_i}\left(\sum_{a \in \mathcal{A}} (s_a(v_a(\mathbf{f}))\tilde{z}_a + \tilde{\tau}_a)\delta_{ar}\right) \\
&= \sum_{a \in \mathcal{A}} (\text{ACT}_{V_i}(s_a(v_a(\mathbf{f}))\tilde{z}_a\delta_{ar}) + \text{ACT}_{V_i}(\tilde{\tau}_a\delta_{ar})) \\
&= \sum_{a \in \mathcal{A}} t_{ai}(v_a(\mathbf{f}))\delta_{ar}.
\end{aligned}$$

Let $\mathbf{c}(\mathbf{f}) = (c_{ri}(\mathbf{f}))_{r \in \mathcal{R}_w, w \in \mathcal{W}, i \in \mathcal{I}}$ be the vector of the travel time under the ACT criterion of all types of travelers over all paths, and \mathcal{F} be the feasible set of possible flows on all paths denoted by

$$\mathcal{F} = \left\{ \mathbf{f} \geq \mathbf{0} \mid \sum_{r \in \mathcal{R}_w} f_{ri} = d_{wi}, \quad w \in \mathcal{W}, i \in \mathcal{I} \right\},$$

in which the constraints are OD demand conservation conditions for all classes of travelers among all OD pairs. We then characterize the NE as follows.

Definition 2.2. A path flow $\mathbf{f}^* \in \mathcal{F}$ is a NE if and only if

$$\begin{aligned}
c_{ri}(\mathbf{f}^*) &\geq \mu_{wi}, & \forall r \in \mathcal{R}_w, w \in \mathcal{W}, i \in \mathcal{I}, \\
f_{ri}^*(c_{ri}(\mathbf{f}^*) - \mu_{wi}) &= 0, & \forall r \in \mathcal{R}_w, w \in \mathcal{W}, i \in \mathcal{I},
\end{aligned}$$

where $\mu_{wi} \geq 0$.

At NE, the travel time along any path connecting the OD pair w perceived by the i th type of travelers under the ACT criterion is at least μ_{wi} .

Moreover, on the paths that have been actually traveled ($f_{ri}^* > 0$), the perceived travel times are exactly at the minimum $c_{ri}(\mathbf{f}^*) = \mu_{wi}$. In other words, no traveler could improve his/her travel time under the ACT criterion by unilaterally changing routes.

Clearly, we can also formulate the NE by means of Variational Inequalities (VI). We let $\mathbf{v} = (v_{ai})_{a \in \mathcal{A}, i \in \mathcal{I}}$ be the vector of flows of all travelers along all links, and we have $v_a = \sum_{i \in \mathcal{I}} v_{ai}, a \in \mathcal{A}$. Let $\mathbf{t}(\mathbf{v}) = (t_{ai}(v_a))_{a \in \mathcal{A}, i \in \mathcal{I}}$ be the vector of travel time under the ACT criterion of all traveler types and along all links. The set of feasible link flows is represented by

$$\mathcal{V} = \left\{ \mathbf{v} \left| v_{ai} = \sum_{w \in \mathcal{W}} \sum_{r \in \mathcal{R}_w} f_{ri} \delta_{ar}, \quad a \in \mathcal{A}, i \in \mathcal{I}, \mathbf{f} \in \mathcal{F} \right. \right\}.$$

Proposition 2.5. The path flow of the NE can be equivalently characterized by the following VI problem:

Find $\mathbf{f}^* \in \mathcal{F}$, such that

$$\langle \mathbf{f} - \mathbf{f}^*, \mathbf{c}(\mathbf{f}^*) \rangle \geq 0, \quad \forall \mathbf{f} \in \mathcal{F},$$

where $\langle \cdot \rangle$ denotes the Euclidean inner product. Likewise, the link flow of NE is characterized by finding $\mathbf{v}^* \in \mathcal{V}$, such that

$$\langle \mathbf{v} - \mathbf{v}^*, \mathbf{t}(\mathbf{v}^*) \rangle \geq 0, \quad \forall \mathbf{v} \in \mathcal{V}. \quad (2.5)$$

Proof. This is an extension of the single class deterministic NE problem and we refer interested readers to Smith (1979) and Dafermos (1980). \square

If travelers are homogeneous, i.e., $n = 1$, the NE defined under the ACT criterion reduces to a single class deterministic NE model. For the general case, $n > 1$, we could adopt algorithms for solving the generic VI (see Nagurney 1998; Facchinei and Pang 2003).

Corollary 2.1. The link flow of NE exists, but may not be unique.

Proof. Since set \mathcal{V} is a compact set, and function $\mathbf{t}(\mathbf{v})$ is continuous, Problem (2.5) admits at least one solution \mathbf{v}^* . Furthermore, this link flow of NE may not be unique, as $\mathbf{t}(\mathbf{v})$ is not strictly monotone in \mathcal{V} . \square

For the special case in which uncertainty along links is flow independent, we show that the corresponding NE problem can be solved via a convex optimization problem. Under this case, the uncertain travel time on link $a \in \mathcal{A}$ can be simplified as

$$\tilde{t}_a(v_a) = s_a(v_a) + \tilde{\tau}_a,$$

and travel time perceived by the i th type of travelers under the ACT criterion is

$$t_{ai}(v_a) = s_a(v_a) + \text{ACT}_{V_i}(\tilde{\tau}_a).$$

Proposition 2.6. When the uncertainty is flow independent, we can compute

the NE traffic flow by solving the following convex optimization problem:

$$\min_{\mathbf{v} \in \mathcal{V}} \sum_{a \in \mathcal{A}} \int_0^{v_a} s_a(x) dx + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) v_{ai}. \quad (2.6)$$

Proof. First, set \mathcal{V} is convex and compact. Let $Z(\mathbf{v}) = \sum_{a \in \mathcal{A}} \int_0^{v_a} s_a(x) dx + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) v_{ai}$, we can easily verify that

$$\frac{\partial Z(\mathbf{v})}{\partial v_{ai}} = s_a(v_a) + \text{ACT}_{V_i}(\tilde{\tau}_a) = t_{ai}(v_a), \quad \forall a \in \mathcal{A}, i \in \mathcal{I},$$

and $Z(\mathbf{v})$ is convex in \mathbf{v} . Therefore, from the necessary optimality condition, we know \mathbf{v}^* is an optimal solution to the convex optimization problem

$$\min_{\mathbf{v} \in \mathcal{V}} Z(\mathbf{v}),$$

if and only if it solves VI Problem (2.5) when the uncertainty is flow independent. \square

We next derive the uniqueness of the NE traffic flow under the assumption that uncertainty along links is flow independent.

Corollary 2.2. If the travel time function is a strictly monotonically increasing function of its own link flow, then the optimal solution of aggregate flow on each link is unique.

Proof. Suppose two distinct link flow solutions \mathbf{v}^1 and \mathbf{v}^2 are both optimal solutions to Problem (2.6). That is, $\exists a \in \mathcal{A}, v_a^1 \neq v_a^2$, and $Z(\mathbf{v}^1) = Z(\mathbf{v}^2)$.

Then we will show the contradiction.

Since $s_a(x)$ is a strictly monotonic increasing function, $\int_0^{v_a} s_a(x)dx$ is a strictly convex function in v_a . For any $\eta \in (0, 1)$,

$$\begin{aligned}
& Z(\eta \mathbf{v}^1 + (1 - \eta) \mathbf{v}^2) - (\eta Z(\mathbf{v}^1) + (1 - \eta) Z(\mathbf{v}^2)) \\
&= \sum_{a \in \mathcal{A}} \int_0^{\eta v_a^1 + (1 - \eta) v_a^2} s_a(x) dx + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) (\eta v_{ai}^1 + (1 - \eta) v_{ai}^2) \\
&\quad - \eta \left(\sum_{a \in \mathcal{A}} \int_0^{v_a^1} s_a(x) dx + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) v_{ai}^1 \right) \\
&\quad - (1 - \eta) \left(\sum_{a \in \mathcal{A}} \int_0^{v_a^2} s_a(x) dx + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) v_{ai}^2 \right) \\
&= \sum_{a \in \mathcal{A}} \int_0^{\eta v_a^1 + (1 - \eta) v_a^2} s_a(x) dx - \left(\eta \sum_{a \in \mathcal{A}} \int_0^{v_a^1} s_a(x) dx + (1 - \eta) \sum_{a \in \mathcal{A}} \int_0^{v_a^2} s_a(x) dx \right) \\
&< 0,
\end{aligned}$$

it follows that

$$Z(\eta \mathbf{v}^1 + (1 - \eta) \mathbf{v}^2) < \eta Z(\mathbf{v}^1) + (1 - \eta) Z(\mathbf{v}^2) = Z(\mathbf{v}^1) = Z(\mathbf{v}^2).$$

Now we have a contradiction to the assumption that \mathbf{v}^1 and \mathbf{v}^2 are both optimal. Therefore, the optimal solution of aggregate flow on each link is unique. \square

We can interpret Problem (2.6) as a deterministic multi-class NE problem, which is easily solved by the traditional Frank-Wolfe algorithm (see for instance Frank and Wolfe 1956; Yang and Huang 2004).

2.4.2 Inefficiency of network equilibrium

Another concept accompanied with NE is to compare with the so-called System Optimum (SO) in which the aggregate travel time of all travelers is minimized (Nash 1951; Wardrop 1952). As travelers choose routes without considering about possible negative impacts on the system performance, it is obvious that the NE solution usually deviates from SO and is less efficient in attaining the minimum aggregate travel time. Led by the seminal work of Koutsoupias and Papadimitriou (2009), the loss of efficiency in NE is an active area of research. The authors propose the concept of Price of Anarchy, which is formally defined as the worst-case inefficiency or the ratio between the aggregate cost of NE and that of SO. In particular, Roughgarden and Tardos (2002) and Correa et al. (2004) present a surprising, but welcome result that NE is near optimal in the sense that the aggregate travel time of all travelers under NE is at most that under SO with double traffic in the same network. In addition, when the travel time function depends linearly on traffic flow, the aggregate travel time of all travelers under NE is at most $4/3$ times that under SO. A sequence of results with respect to a more general link travel time function are further developed by Roughgarden (2003), Chau and Sim (2003), Perakis (2007), Correa et al. (2008), Han et al. (2008) and Han et al. (2014). In this section, we derive similar results in the NE problem for the case when travelers are sensitive to risk and ambiguity. To obtain analytical results, we again assume that the uncertainty along links is flow independent. Since in the network, each traveler may not have complete information about uncertain travel times, his/her perceived travel time (the

travel time under the ACT criterion) is the only foundation to make route choice decisions. Therefore, to be consistent with travelers' choices, we define SO as the minimum aggregate perceived travel time, i.e., minimum aggregate travel time under the ACT criterion.

For a given traffic flow, $\mathbf{v} \in \mathcal{V}$, we represent the aggregate travel time under the ACT criterion on the entire network by

$$\mathcal{C}_{\mathbf{v}}(\mathbf{v}) = \langle \mathbf{t}(\mathbf{v}), \mathbf{v} \rangle = \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} t_{ai}(v_a) v_{ai} = \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (s_a(v_a) + \text{ACT}_{V_i}(\tilde{\tau}_a)) v_{ai}.$$

By defining $\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}) = \langle \mathbf{t}(\mathbf{v}^*), \mathbf{v} \rangle$, variational inequalities (2.5) can be replaced as $\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) \leq \mathcal{C}_{\mathbf{v}^*}(\mathbf{v})$, where $\mathbf{v}^* = (v_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ is traffic flow vector at NE for types of travelers along all links, and $\mathbf{v} \in \mathcal{V}$ is the vector of any feasible flows. Let $\mathbf{x}^* = (x_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ denote the traffic flow vector at SO, which minimizes aggregate travel time under the ACT criterion. We can analyze the inefficiency of NE by comparing $\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*)$ and $\mathcal{C}_{\mathbf{x}^*}(\mathbf{x}^*)$. In particular, we are interested in the Price of Anarchy, which is the worst-case ratio between the aggregate travel time of NE and that of SO under the ACT criterion.

Proposition 2.7. Consider an instance of Problem (2.6). The vectors $\mathbf{v}^* = (v_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ and $\mathbf{x}^* = (x_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ represent link flows at NE and SO, respectively.

- (a) Let vector $\mathbf{u} = (u_{ai})_{a \in \mathcal{A}, i \in \mathcal{I}}$ be a feasible flow for the same network but

with twice as many travelers of the same type. Then

$$\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) \leq \mathcal{C}_{\mathbf{u}}(\mathbf{u}).$$

- (b) If travel time function is a monomial function $s_a(v_a) = b_a(v_a)^m$ ($m \geq 0$), then

$$\frac{\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*)}{\mathcal{C}_{\mathbf{x}^*}(\mathbf{x}^*)} \leq (1 - m(m+1)^{-(m+1)/m})^{-1}.$$

- (c) If travel time function is a general continuous and nondecreasing function, we have

$$\frac{\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*)}{\mathcal{C}_{\mathbf{x}^*}(\mathbf{x}^*)} \leq \frac{1}{1 - \beta(\mathcal{A})},$$

where

$$\beta(\mathcal{A}) = \sup_{a \in \mathcal{A}} \sup_{v \geq 0} \frac{\max_{x \geq 0} x (s_a(v) - s_a(x))}{s_a(v)v}, \quad \text{and} \quad 0 \leq \beta(\mathcal{A}) \leq 1.$$

Proof. The proof of this result follows from Correa et al. (2004).

(a) Note that $s_a(v_a)$ is a differentiable, monotonically increasing function in v_a , and $\mathbf{u} = (u_{ai})_{a \in \mathcal{A}, i \in \mathcal{I}}$ is a feasible flow for the same network but with double demands. We have

$$\begin{aligned} s_a(u_a)u_a + s_a(v_a^*)v_a^* - s_a(v_a^*)u_a &\geq s_a(u_a)u_a \geq 0, & \text{if } u_a \leq v_a^*; \\ s_a(u_a)u_a + s_a(v_a^*)v_a^* - s_a(v_a^*)u_a &\geq s_a(v_a^*)v_a^* \geq 0, & \text{if } u_a \geq v_a^*. \end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathcal{C}_{\mathbf{u}}(\mathbf{u}) + \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) - \mathcal{C}_{\mathbf{v}^*}(\mathbf{u}) \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (t_{ai}(u_a)u_{ai} + t_{ai}(v_a^*)v_{ai}^* - t_{ai}(v_a^*)u_{ai}) \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (s_a(u_a)u_{ai} + s_a(v_a^*)v_{ai}^* + \text{ACT}_{V_i}(\tilde{\tau}_a)v_{ai}^* - s_a(v_a^*)u_{ai}) \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a)v_{ai}^* + \sum_{a \in \mathcal{A}} (s_a(u_a)u_a + s_a(v_a^*)v_a^* - s_a(v_a^*)u_a) \\
&\geq \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a)v_{ai}^* \\
&\geq 0.
\end{aligned}$$

Besides, we note that $\mathbf{u}/2 = (\frac{u_{ai}}{2})_{a \in \mathcal{A}, i \in \mathcal{I}}$ is a feasible flow for the original instance. From the NE property,

$$\mathcal{C}_{\mathbf{u}}(\mathbf{u}) \geq \mathcal{C}_{\mathbf{v}^*}(\mathbf{u}) - \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) = 2\mathcal{C}_{\mathbf{v}^*}\left(\frac{\mathbf{u}}{2}\right) - \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) \geq 2\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) - \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) = \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*).$$

(b) If travel time is a monomial function, defined as $s_a(v_a) = b_a(v_a)^m$ such that

$$t_{ai}(v_a) = b_a(v_a)^m + \text{ACT}_{V_i}(\tilde{\tau}_a).$$

Then, we have

$$\begin{aligned}
& \mathcal{C}_{\mathbf{v}^*}(\mathbf{x}) \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (t_a(v_a^*) + \text{ACT}_{V_i}(\tilde{\tau}_a)) x_{ai} \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) x_{ai} + \sum_{a \in \mathcal{A}} b_a(v_a^*)^m x_a \\
&\leq \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) x_{ai} + \sum_{a \in \mathcal{A}} b_a(x_a^{m+1} + m(m+1)^{-(m+1)/m}(v_a^*)^{m+1}) \\
&= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (b_a(x_a)^m + \text{ACT}_{V_i}(\tilde{\tau}_a)) x_{ai} + m(m+1)^{-(m+1)/m} \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} b_a(v_a^*)^m v_{ai}^* \\
&\leq \mathcal{C}_{\mathbf{x}}(\mathbf{x}) + m(m+1)^{-(m+1)/m} \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (b_a(v_a^*)^m + \text{ACT}_{V_i}(\tilde{\tau}_a)) v_{ai}^* \\
&= \mathcal{C}_{\mathbf{x}}(\mathbf{x}) + m(m+1)^{-(m+1)/m} \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*),
\end{aligned}$$

where the first inequality is tenable because the function $f(x) = v^m x - x^{m+1}$ ($x \geq 0$) will get its maximum $m(m+1)^{-(m+1)/m} v^{m+1}$ at $x = v(m+1)^{-1/m}$; and the second inequality holds because $\sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) v_{ai}^* \geq 0$. Then, since $\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) \leq \mathcal{C}_{\mathbf{v}^*}(\mathbf{x})$, we get

$$(1 - m(m+1)^{-(m+1)/m}) \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) \leq \mathcal{C}_{\mathbf{x}^*}(\mathbf{x}^*).$$

When $\mathbf{x}^* = (x_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ is the system optimum, we can find the Price of Anarchy bounded at $(1 - m(m+1)^{-(m+1)/m})^{-1}$, which is the same as that in deterministic cases.

(c) We could generalize the travel time function to continuous, nondecreasing

case.

$$\begin{aligned}
\mathcal{C}_{\mathbf{v}^*}(\mathbf{x}) &= \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} t_{ai}(v_a^*) x_{ai} \\
&= \sum_{a \in \mathcal{A}} s_a(v_a^*) x_a + \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} \text{ACT}_{V_i}(\tilde{\tau}_a) x_{ai} \\
&= \sum_{a \in \mathcal{A}} x_a (s_a(v_a^*) - s_a(x_a)) + \mathcal{C}_{\mathbf{x}}(\mathbf{x}) \\
&= \sum_{a \in \mathcal{A}} \frac{x_a (s_a(v_a^*) - s_a(x_a))}{s_a(v_a^*) v_a^*} s_a(v_a^*) v_a^* + \mathcal{C}_{\mathbf{x}}(\mathbf{x}) \\
&\leq \sum_{a \in \mathcal{A}} \beta(v_a^*, s_a(v_a^*)) s_a(v_a^*) v_a^* + \mathcal{C}_{\mathbf{x}}(\mathbf{x}) \\
&\leq \sup_{a \in \mathcal{A}} \beta(v_a^*, s_a(v_a^*)) \sum_{a \in \mathcal{A}} s_a(v_a^*) v_a^* + \mathcal{C}_{\mathbf{x}}(\mathbf{x}) \\
&\leq \beta(\mathcal{A}) \mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*) + \mathcal{C}_{\mathbf{x}}(\mathbf{x}),
\end{aligned}$$

where $\beta(v, s(v)) = \frac{1}{s(v)v} \max_{x \geq 0} \{x(s(v) - s(x))\}$, and $\beta(\mathcal{A}) = \sup_{a \in \mathcal{A}} \sup_{v \geq 0} \beta(v, s_a(v))$.

Since the travel time function $s(v)$ is a continuous nondecreasing function, the following relationship holds:

$$\begin{aligned}
0 &= \frac{v(s(v) - s(v))}{s(v)v} \leq \beta(v, s(v)) \\
&\leq \frac{\max_{0 \leq x \leq v} x(s(v) - s(x))}{s(v)v} \leq \frac{\max_{0 \leq x \leq v} xs(v)}{s(v)v} \leq \frac{vs(v)}{s(v)v} = 1.
\end{aligned}$$

Assuming $\mathbf{x}^* = (x_{ai}^*)_{a \in \mathcal{A}, i \in \mathcal{I}}$ is SO solution, we have the Price of Anarchy as

$$\frac{\mathcal{C}_{\mathbf{v}^*}(\mathbf{v}^*)}{\mathcal{C}_{\mathbf{x}^*}(\mathbf{x}^*)} \leq \frac{1}{1 - \beta(\mathcal{A})}. \quad \square$$

As far as we know, classical Price of Anarchy results on traffic equilibriums do not consider the influence of uncertainty on travelers' choice and our result is possibly the first attempt in this direction. It is interesting to observe

that after accounting for travelers preferences for risk and ambiguity in the traffic equilibrium problem, the Price of Anarchy results remain similar to the classical ones where travel times are deterministic.

2.4.3 A network equilibrium example

The following example explicitly illustrates the calculation of NE and SO under the ACT criterion, and demonstrates the inefficiency issues under various mixtures of travelers' profiles. It elucidates the importance of taking travelers' risk and ambiguity attitudes into account in analyzing traffic networks.

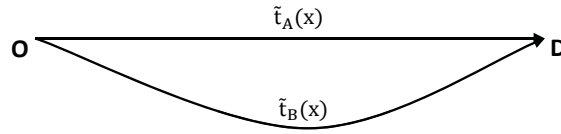


Fig. 2.3: Two paths network with uncertain travel times.

We consider a two paths network from origin O to destination D depicted in Figure 2.3. The traffic rate is assumed to be 1. The paths have travel times as follows:

$$\tilde{t}_A(v_A) = (v_A)^4 + \tilde{\tau}_A, \quad \tilde{t}_B(v_B) = \frac{6}{5},$$

where $\tilde{\tau}_A$ is uncertain. We assume that all travelers have the same information on the uncertain parameter $\tilde{\tau}_A$. Specifically, $\tilde{\tau}_A$ has a mean value of $\frac{1}{5}$ and support in $[0, \Delta]$, $\Delta > \frac{1}{5}$. Hence, the corresponding distributional

Type i	Demand	α_i	λ_i	$ACT_{V_i}(\tilde{\tau}_A)$
1	$\frac{2}{3}$	$\frac{4}{5}$	5	$\frac{1}{25} + \frac{4}{25} \ln\left(1 + \frac{1}{5\Delta} (\exp(5\Delta) - 1)\right)$
2	$\frac{1}{3}$	$\frac{1}{5}$	-5	$\frac{1}{25} - \frac{4}{25} \ln\left(1 + \frac{1}{5\Delta} (\exp(-5\Delta) - 1)\right)$

Tab. 2.3: Travelers' profile in Case 3.

uncertainty set is given by

$$\mathbb{F}(\Delta) = \left\{ \mathbb{P} \left| \mathbb{E}_{\mathbb{P}}(\tilde{\tau}_A) = \frac{1}{5}, \mathbb{P}(\tilde{\tau}_A \in [0, \Delta]) = 1 \right. \right\}.$$

Note that the parameter Δ represents the worst-case delay of $\tilde{\tau}_A$ and implies the level of uncertainty along Path A . On the other hand, Path B has deterministic travel time and is unaffected by Δ . With various compositions of travelers in terms of risk and ambiguity attitudes, the NE and SO under the ACT criterion will yield different flow patterns. To explore the impact of Δ on these flow patterns, we consider the following three cases:

Case 1: All travelers are risk-neutral and ambiguity neutral ($\lambda = 0, \alpha = \frac{1}{2}$);

Case 2: All travelers are extremely risk-averse and pessimistic towards ambiguity ($\lambda \rightarrow +\infty, \alpha = 1$).

Case 3: Travelers composition with profiles is shown in Table 2.3.

In Case 1, all travelers are risk and ambiguity neutral and they intuitively perceive the uncertain term as its mean value. Hence, the solutions are consistent with traditional deterministic NE and SO models. In Case 2, travelers who are radically risk-averse and pessimistic towards ambiguity consider the worst-case travel time in deciding between paths. In Case 3, type 1 travelers are risk-averse and pessimistic towards ambiguity, while type 2 travelers are risk-seeking and optimistic towards ambiguity. We derive flow

solutions of NE and SO under the ACT criterion in Table 2.4. For notational simplicity, in this example, we let $t_{A1} = \text{ACT}_{V_1}(\tilde{\tau}_A)$ and $t_{A2} = \text{ACT}_{V_2}(\tilde{\tau}_A)$. Note that Δ_1 and Δ_2 are the unique solutions satisfying $(\frac{6}{25} - \frac{1}{5}t_{A1})^{1/4} - \frac{1}{3} = 0$ and $(\frac{6}{5} - t_{A1})^{1/4} - \frac{1}{3} = 0$, and $\Delta_1 \approx 1.8136$, and $\Delta_2 \approx 1.8829$, respectively.

Case	Condition	Criterion	Type	Traffic flow		Path ACT ¹	
				A	B	A	B
1		NE		1	0	$\frac{6}{5}$	$\frac{6}{5}$
		SO		$5^{-1/4}$	$1 - 5^{-1/4}$	$\frac{2}{5}$	$\frac{6}{5}$
2	$\frac{1}{5} < \Delta \leq \frac{6}{5}$	NE		$(\frac{6}{5} - \Delta)^{1/4}$	$1 - (\frac{6}{5} - \Delta)^{1/4}$	$\frac{6}{5}$	$\frac{6}{5}$
		SO		$(\frac{6}{25} - \frac{1}{5}\Delta)^{1/4}$	$1 - (\frac{6}{25} - \frac{1}{5}\Delta)^{1/4}$	$\frac{6}{25} + \frac{4}{5}\Delta$	$\frac{6}{5}$
	$\frac{6}{5} \leq \Delta$	NE		0	1	Δ	$\frac{6}{5}$
		SO		0	1	Δ	$\frac{6}{5}$
3	$\frac{1}{5} < \Delta \leq \Delta_1$	NE	1	$(\frac{6}{5} - t_{A1})^{1/4} - \frac{1}{3}$	$1 - (\frac{6}{5} - t_{A1})^{1/4}$	$\frac{6}{5}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{6}{5} - t_{A1} + t_{A2}$	$\frac{6}{5}$
		SO	1	$(\frac{6}{25} - \frac{1}{5}t_{A1})^{1/4} - \frac{1}{3}$	$1 - (\frac{6}{25} - \frac{1}{5}t_{A1})^{1/4}$	$\frac{6}{25} + \frac{4}{5}t_{A1}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{6}{25} - \frac{1}{5}t_{A1} + t_{A2}$	$\frac{6}{5}$
	$\Delta_1 \leq \Delta \leq \Delta_2$	NE	1	$(\frac{6}{5} - t_{A1})^{1/4} - \frac{1}{3}$	$1 - (\frac{6}{5} - t_{A1})^{1/4}$	$\frac{6}{5}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{6}{5} - t_{A1} + t_{A2}$	$\frac{6}{5}$
		SO	1	0	$\frac{2}{3}$	$\frac{1}{81} + t_{A1}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{1}{81} + t_{A2}$	$\frac{6}{5}$
	$\Delta_2 \leq \Delta$	NE	1	0	$\frac{2}{3}$	$\frac{1}{81} + t_{A1}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{1}{81} + t_{A2}$	$\frac{6}{5}$
		SO	1	0	$\frac{2}{3}$	$\frac{1}{81} + t_{A1}$	$\frac{6}{5}$
			2	$\frac{1}{3}$	0	$\frac{1}{81} + t_{A2}$	$\frac{6}{5}$

¹ Path ACT refers to the travel time along the path under the ACT criterion;

Tab. 2.4: Flow patterns of NE and SO under the ACT criterion for three cases.

We now study the inefficiency of NE under the ACT criterion with respect to the parameter Δ . We represent the aggregate travel times under the ACT criterion in Case i under the NE and SO model by ACT_i^{NE} and ACT_i^{SO} respectively, and quantify the inefficiency of NE via the ratio $\frac{\text{ACT}_i^{NE}}{\text{ACT}_i^{SO}}$.

For these three cases, the ratios are calculated as:

$$\begin{aligned} \frac{\text{ACT}_1^{NE}}{\text{ACT}_1^{SO}} &= \frac{6}{6 - 4 \times 5^{-1/4}}; \\ \frac{\text{ACT}_2^{NE}}{\text{ACT}_2^{SO}} &= \begin{cases} \frac{6}{6 - 20 \left(\frac{6}{25} - \frac{1}{5}\Delta\right)^{5/4}}, & \text{when } \frac{1}{5} < \Delta \leq \frac{6}{5}, \\ 1, & \text{when } \frac{6}{5} \leq \Delta; \end{cases} \\ \frac{\text{ACT}_3^{NE}}{\text{ACT}_3^{SO}} &= \begin{cases} \frac{\frac{6}{5} - \frac{1}{3}t_{A1} + \frac{1}{3}t_{A2}}{\frac{6}{5} - \frac{1}{3}t_{A1} + \frac{1}{3}t_{A2} - 4 \left(\frac{6}{25} - \frac{1}{5}t_{A1}\right)^{5/4}}, & \text{when } 0.2 < \Delta \leq \Delta_1, \\ \frac{\frac{6}{5} - \frac{1}{3}t_{A1} + \frac{1}{3}t_{A2}}{\frac{4}{5} + \frac{1}{243} + \frac{1}{3}t_{A2}}, & \text{when } \Delta_1 \leq \Delta \leq \Delta_2, \\ 1, & \text{when } \Delta_2 \leq \Delta. \end{cases} \end{aligned}$$

Figure 2.4 depicts the ratios of Case 2 and 3. We observe that the ratios decrease with the increase of upper bound Δ . For this specific example, when the travel time becomes more uncertain, the change of traffic flow has less impact on the traveler's path choice decisions, correspondingly, the flow pattern at NE will approach to that at SO. In other words, it suggests that if the travel time along a traffic network is highly uncertain, then there is little benefit from having the system optimal solution in which the aggregate travel time under the ACT criterion is minimized.

Next, we highlight that it is essential for traffic managers to consider travelers' risk and ambiguity attitudes when determining the system optimal flow pattern. Specifically, if we ignore uncertainty and calculate the deterministic system optimal (DSO) flow pattern, the system performance may be worse than that of NE in terms of the aggregate travel time under the ACT criterion. We represent the DSO flow pattern by $\mathbf{u}^* = (u_a^*)_{a \in \mathcal{A}}$, which

is the unique optimal solution of

$$\begin{aligned}
\min \quad & \sum_{a \in \mathcal{A}} (s_a(u_a) + \mathbb{E}(\tilde{\tau}_a)) u_a \\
\text{s.t.} \quad & u_a = \sum_{w \in \mathcal{W}} \sum_{r \in \mathcal{R}_w} f_r \delta_{ar}, \quad \forall a \in \mathcal{A}, \\
& \sum_{r \in \mathcal{R}_w} f_r = \sum_{i \in \mathcal{I}} d_{wi}, \quad \forall w \in \mathcal{W}, \\
& f_r \geq 0, \quad \forall r \in \mathcal{R}_w, w \in \mathcal{W}.
\end{aligned}$$

Note that the flow pattern $\mathbf{u}^* = (u_a^*)_{a \in \mathcal{A}}$ only identifies the aggregate traffic flow on each link. Therefore, with the mixture of travelers, the traffic flow for each type of travelers on each link may not be unique. We represent its feasible set as

$$\mathcal{U} = \left\{ \mathbf{v} \in \mathcal{V} \mid \sum_{i \in \mathcal{I}} v_{ai} = u_a^*, \forall a \in \mathcal{A} \right\}.$$

Then, for any $\mathbf{v} \in \mathcal{U}$, we define $\text{ACT}^{DSO}(\mathbf{v})$ as the total travel time under the ACT criterion of all travelers when the traffic flow is \mathbf{v} , that is,

$$\text{ACT}^{DSO}(\mathbf{v}) = \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (s_a(u_a^*) + \text{ACT}_{V_i}(\tilde{\tau}_a)) v_{ai}.$$

Since $\text{ACT}^{DSO}(\mathbf{v})$ is a function of $\mathbf{v} \in \mathcal{U}$, we define its lower and upper bound by $\underline{\text{ACT}}^{DSO}$ and $\overline{\text{ACT}}^{DSO}$ respectively, where

$$\begin{aligned}
\underline{\text{ACT}}^{DSO} &= \min_{\mathbf{v} \in \mathcal{U}} \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (s_a(u_a^*) + \text{ACT}_{V_i}(\tilde{\tau}_a)) v_{ai}; \\
\overline{\text{ACT}}^{DSO} &= \max_{\mathbf{v} \in \mathcal{U}} \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}} (s_a(u_a^*) + \text{ACT}_{V_i}(\tilde{\tau}_a)) v_{ai}.
\end{aligned}$$

Hence, for any $\mathbf{v} \in \mathcal{U}$, $\text{ACT}^{DSO}(\mathbf{v}) \in [\underline{\text{ACT}}^{DSO}, \overline{\text{ACT}}^{DSO}]$. Similarly, we quantify the inefficiency of DSO under the ACT criterion via the ratios $\frac{\text{ACT}_i^{DSO}}{\text{ACT}_i^{SO}}$ and $\frac{\overline{\text{ACT}}_i^{DSO}}{\overline{\text{ACT}}_i^{SO}}$ as follows:

$$\frac{\text{ACT}_2^{DSO}}{\text{ACT}_2^{SO}} = \frac{\underline{\text{ACT}}_2^{DSO}}{\underline{\text{ACT}}_2^{SO}} = \frac{\overline{\text{ACT}}_2^{DSO}}{\overline{\text{ACT}}_2^{SO}} = \begin{cases} \frac{6 + 5^{3/4}(\Delta - 1)}{6 - 20\left(\frac{6}{25} - \frac{1}{5}\Delta\right)^{5/4}}, & \frac{1}{5} < \Delta \leq \frac{6}{5}, \\ \frac{6 + 5^{3/4}(\Delta - 1)}{6}, & \frac{6}{5} \leq \Delta; \end{cases}$$

$$\frac{\text{ACT}_3^{DSO}}{\text{ACT}_3^{SO}} = \begin{cases} \frac{\frac{6}{5} - 5^{-1/4} + \left(5^{-1/4} - \frac{1}{3}\right)t_{A1} + \frac{1}{3}t_{A2}}{\frac{6}{5} - \frac{1}{3}t_{A1} + \frac{1}{3}t_{A2} - 4\left(\frac{6}{25} - \frac{1}{5}t_{A1}\right)^{5/4}}, & \frac{1}{5} < \Delta \leq \Delta_1, \\ \frac{\frac{6}{5} - 5^{-1/4} + \left(5^{-1/4} - \frac{1}{3}\right)t_{A1} + \frac{1}{3}t_{A2}}{\frac{4}{5} + \frac{1}{243} + \frac{1}{3}t_{A2}}, & \Delta_1 \leq \Delta; \end{cases}$$

$$\frac{\overline{\text{ACT}}_3^{DSO}}{\overline{\text{ACT}}_3^{SO}} = \begin{cases} \frac{\frac{6}{5} - 5^{-1/4} + \frac{2}{3}t_{A1} + \left(5^{-1/4} - \frac{2}{3}\right)t_{A2}}{\frac{6}{5} - \frac{1}{3}t_{A1} + \frac{1}{3}t_{A2} - 4\left(\frac{6}{25} - \frac{1}{5}t_{A1}\right)^{5/4}}, & \frac{1}{5} < \Delta \leq \Delta_1, \\ \frac{\frac{6}{5} - 5^{-1/4} + \frac{2}{3}t_{A1} + \left(5^{-1/4} - \frac{2}{3}\right)t_{A2}}{\frac{4}{5} + \frac{1}{243} + \frac{1}{3}t_{A2}}, & \Delta_1 \leq \Delta. \end{cases}$$

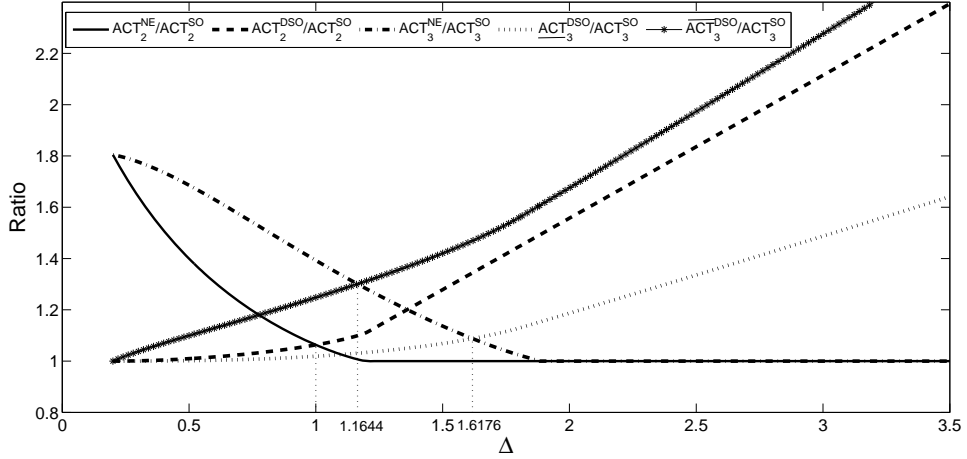


Fig. 2.4: Inefficiency of NE and DSO under the ACT criterion in Case 2 and 3 in two-nodes network.

Figure 2.4 demonstrates the inefficiency of NE and DSO under the ACT

criterion of Cases 2 and 3. In Case 2, for the network where travelers are extremely risk-averse and pessimistic towards ambiguity, with the increase of Δ , the NE flow pattern under the ACT criterion becomes less inefficient, while the inefficiency of DSO grows increasingly severe. When $\Delta > 1$, $\text{ACT}_2^{DSO} > \text{ACT}_2^{NE}$ suggests if we instruct the traffic flow following DSO criterion, which does not account for travelers' attitudes towards risk and ambiguity, the performance will turn worse than its original anarchy state. Similarly, in Case 3, with two types of travelers, the ratio $\frac{\text{ACT}_3^{DSO}(\mathbf{v})}{\text{ACT}_3^{SO}}$ lies between the two curves $\frac{\text{ACT}_3^{DSO}}{\text{ACT}_3^{SO}}$ and $\frac{\overline{\text{ACT}}_3^{DSO}}{\text{ACT}_3^{SO}}$. The increase of upper bound Δ will cut down the inefficiency of NE, but result in the deterioration of DSO in terms of system performance. Moreover, when the level of travel time uncertainty increases to some specific value, the DSO performance will be no better than the NE performance, which suggests this guidance effort would be in vain.

Following the same strategy, we extend our computational study from this two links small network to a five-nodes complete network, which includes 5 nodes, and 20 links. Since calculating $\underline{\text{ACT}}^{DSO}$ and $\overline{\text{ACT}}^{DSO}$ is generally a hard problem, we only use this simple network for illustrative purpose. The demand on each OD pair for each type of travelers is uniformly generated from the set $\{101, 102, \dots, 800\}$. Uncertain travel time on each link is written as

$$\tilde{t}_a(v_a) = s_a(0) + 0.15 \left(\frac{v_a}{c_a} \right)^4 + \tilde{\tau}_a, \quad \forall a \in \mathcal{A}.$$

Free flow travel time $s_a(0)$ follows uniform distribution $U(2, 6)$, and capacity c_a is generated from uniform distribution $U(200, 1000)$. Instead of determin-

istic travel time, we assume that uncertainties occur on each link independently. Moreover, the disturbance is flow independent, with the mean equal to 20% of free flow travel time on that link, and lower bound equal to zero. We vary the upper bound of uncertainties by Δ . The uncertainty $\tilde{\tau}_a$ is characterized as

$$\mathbb{F} = \{\mathbb{P}[\mathbb{E}_{\mathbb{P}}(\tilde{\tau}_a) = 0.2s_a(0), \mathbb{P}(\tilde{\tau}_a \in [0, \Delta \times \mathbb{E}_{\mathbb{P}}(\tilde{\tau}_a)]) = 1, \forall a \in \mathcal{A}]\}.$$

Travelers' characteristics are consistent with Case 3. We randomly generate 50 instances, and summarize the average performance. The inefficiency results of NE and DSO under the ACT criterion of five-nodes network are listed in Figure 2.5. Similar conclusions could be derived here. When the flow independent disturbance on travel time becomes highly uncertain, the influence of selfishness on inefficiency diminishes.

2.5 Conclusion

This chapter studies the preferences for uncertain travel times in which the probability distributions may not be fully characterized. By explicitly distinguishing risk and ambiguity concepts, we propose a new criterion called ambiguity-aware CARA travel time for ranking the uncertain travel time, which systematically integrates the travelers' inability to capture the exact information of uncertain travel times, and their attitudes towards risk and ambiguity. This setting is based on the Hurwicz criterion and constant absolute risk aversion, which is empirically supported and provides computational

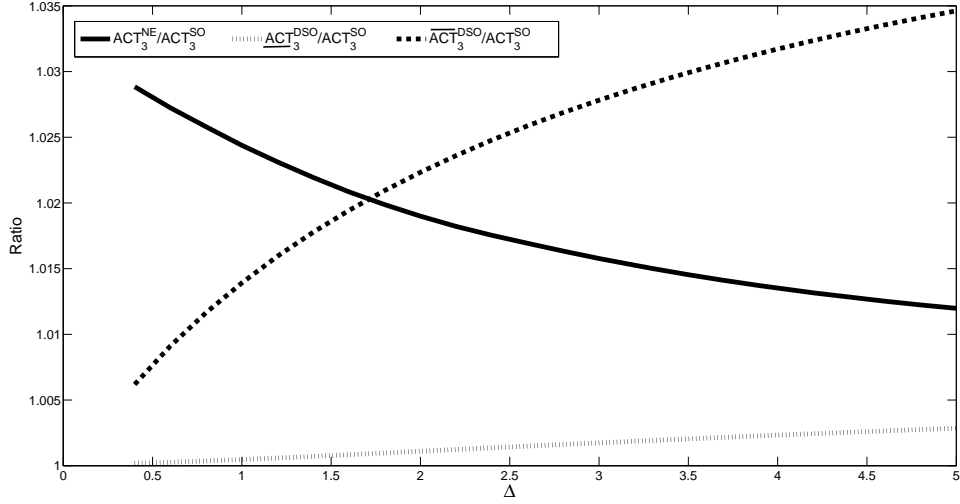


Fig. 2.5: Inefficiency of NE and DSO under the ACT criterion in Case 3 in five-nodes network.

benefits.

With this criterion, we explore computational solvability of the path choice problem on a network where travel times are uncertain. We show that finding a path with the minimum travel time under the ACT criterion is polynomially solvable when link travel times are independently distributed. We also prove that the problem becomes intractable when link travel times are correlated. Focusing on independently distributed link travel times, we present the general VI formulation of NE under the ACT criterion. We analyze the case when the uncertainty along links is flow independent and show that it can be addressed as a convex optimization problem. We also determine the inefficiency of NE by deriving the Price of Anarchy, which is similar to the deterministic NE case.

The ACT criterion could potentially enhance the predictive capability

of path choice and traffic equilibrium. First, it does not require travelers to know the probability distributions of the network. Second, it has the potential to incorporate risk and ambiguity in travelers' decision making. Third, the path choice problem and network equilibrium established retain the computational tractability of their deterministic counterparts. It will be valuable to establish empirically the risk and ambiguity profiles of a population of travelers residing in different cities and possibly having different cultures. We hope that our work could encourage future research in this direction. This is joint work with Melvyn Sim, Defeng Sun and Xiaoming Yuan.

3. ROUTING OPTIMIZATION WITH DEADLINES UNDER UNCERTAINTY

3.1 Introduction

Routing optimization problems on networks consist of finding paths (either simple paths, closed paths, tours, or walks) between nodes of the networks in an efficient way. These problems and their solutions have proved to be essential ingredients for addressing many real-world decisions in applications as diverse as logistics, transportation, computer networking, internet routing, to name a few.

In many of these routing applications, specially those imposing deadlines on when to visit nodes, the presence of uncertainty in the networks (e.g., presence or not of some of the nodes, arc travel times) is a critical issue to consider explicitly if one hopes to provide solutions of practical values to the end users. There are two related issues: (i) how to properly model uncertainty in order to reflect real-world concerns, and (ii) how to do so in models which will be computationally tractable? In this chapter, we provide novel ways to address such issues for a subclass of these routing problems under uncertainty.

More specifically, we study routing problems on networks with deadlines imposed at a subset of nodes, and with uncertain arc travel times that can be characterized by exact distributions, or by a distributional uncertainty set incorporating ambiguity. Our model is static in the sense that routing decisions are made prior to the realization of uncertain travel times. To incorporate ambiguity, instead of defining an exact probability distribution \mathbb{P} for an arc travel time, we assume its true distribution lies in a distributional uncertainty set denoted by \mathbb{F} , which is characterized by some descriptive statistics, e.g., means and bounded supports. The goal is to find optimal routing policies such that arrival times at nodes respect deadlines “as much as possible”, in a mathematically precise way under an appropriately defined performance measure which takes into account such distributional uncertainty assumptions.

This framework can be applied to transportation networks, for example, for delivery service providers to route their vehicles, where multiple vehicles and uncertain service time could be incorporated, or for individuals to make their travel plans. It can also be employed to solve problems arising from telephone networks or electronic data networks.

The deterministic version of many routing optimization problems (e.g., shortest path problems, traveling salesman problems, vehicle routing problems) have been studied extensively over many decades (see the literature reviews of Toth and Vigo 2001; Öncan et al. 2009; and Laporte 2010, to name a few). Due to the recognized practical importance of incorporating uncertainty, the uncertain versions of routing problems have also attracted increasing attention. Researchers have formulated various problems depending

on the uncertainty under consideration; for example, uncertainty in customer presence (Jaillet 1988; Jaillet and Odoni 1988; Campbell and Thomas 2008), uncertainty in demand (Bertsimas 1992; Bertsimas and Simchi-Levi 1996), and uncertainty in travel time (see below). A comprehensive overview can be found in Cordeau et al. (2006) and Häme and Hakula (2013).

In this chapter, our particular attention is on the uncertainty of travel times, and we now review the literature specific to this area, first concentrating on the shortest path problems with deadline. Under uncertainty about arc travel times, and given a deadline at the destination node, these problems consist of finding paths from the origin to the destination in such a way that the deadline is “effectively” met. At the heart of this problem, one has to (i) model the uncertainty, and (ii) explicitly and quantitatively define the word “effectively”. Researchers have established distinct selection criteria.

One intuitive and well-discussed way is to select a route with the largest probability of arriving on time (see Frank 1969; Mirchandani 1976; Nie and Wu 2009). However, maximizing the arrival probability fails to take the delay level into account. Everything else being equal, a path with a probability of 0.01 of incurring a delay may not be better than another one with a probability 0.011 if the delays are 10 hours and 10 minutes, respectively. Furthermore, evaluating the probability of a weighted sum of random variables is generally difficult, as we have discussed in Chapter 2. By assuming that travel times follow normal distributions, researchers have reformulated the problem of maximizing the arrival probability using different techniques. Chen et al. (2012) formulate the problem as a multi-criteria shortest path finding problem. Nikolova (2009) and Xiao et al. (2012) equivalently formu-

late the problem as maximizing the earliness per unit of standard deviation (i.e., a “punctuality ratio” criterion). Nevertheless, the problem is still a computationally hard problem, and the chosen criterion does not respect the first order stochastic dominance property, which essentially makes it not so appealing. Kosuch and Lissner (2010) minimize the delay excess penalty for the stochastic shortest path problem under normally distributed arc travel times. They embed a stochastic projected gradient method within a branch-and-bound framework to solve it. Nie et al. (2012) approach the problem using a second order stochastic method, and suggest sampling and dynamic programming techniques.

The above approaches necessitate a complete distributional knowledge about uncertain travel times. However, in reality, it is hard to figure out the exact frequency associated with an uncertain event. Additionally, because of complex phenomena due to traffic congestion, weather conditions, and drivers’ behaviors, travel times cannot easily be modeled using simple probability distributions, and cannot even be estimated accurately without the “bias” of a chosen sampling procedure. Henceforth, with only limited information about travel time distributions, robust approaches have been proposed to model the uncertainty.

Researchers have either considered that each arc travel time is associated with an interval or with a discrete set of scenarios, and have suggested different optimization criteria and methodologies. Kouvelis and Yu (1997) use the min-max approach to find an optimal path such that its worst-case performance across all possible realizations is superior to that of any other path. The problem is proved to be NP-hard. Karaşan et al. (2001), Monte-

manni et al. (2004), Averbakh and Lebedev (2004), Catanzaro et al. (2011) study a relative robustness criterion, which finds a path minimizing the maximal robust deviation. Gabrel et al. (2013) propose to use bw-robustness, in which robust solutions should provide good performances in most scenarios without being too bad for the rest, to formulate the shortest path problem into a large scale integer linear programming problem. However, all the robust formulations introduced above do not inherit one key property of its deterministic version, i.e., polynomial solvability. The only exception comes from Bertsimas and Sim (2003, 2004). To adjust the conservatism level, they introduce a parameter Γ , named the budget of uncertainty, to represent the maximum number of coefficients that could deviate from the nominal value. With their formulation, the optimal solution can be obtained by solving only a small number of deterministic shortest path problems.

The modeling issues are further complicated when we have a subset of nodes with deadline requests. Campbell and Thomas (2008) show that the problem of incorporating deadlines is much more computationally complex than the version without deadlines. To the best of our knowledge, only few studies consider such general routing problems with deadlines in the presence of uncertain travel times.

Laporte et al. (1992) consider a multiple vehicle routing problem with stochastic travel times and service times. Each vehicle has a targeted time to complete the route. They propose a chance constrained model and a stochastic programming model, and suggest branch and cut algorithms to solve moderate-size problems. Kenyon and Morton (2003) mainly focus on the length of the longest route traveled by multiple vehicles and develop

two versions of the model by minimizing the expected completion time or maximizing the probability of completion within a given deadline, and finally solve the model by branch-and-cut scheme.

Jula et al. (2006), Chang et al. (2009), and Mazmnyan et al. (2009) consider a stochastic routing problem with time windows in which they seek a solution guaranteeing that the probability of violating the latest time is no larger than a threshold. To estimate the arrival time at each node, Jula et al. (2006) approximate its first two moments based on dynamic programming, while Chang et al. (2009) and Mazmnyan et al. (2009) impose a normal distribution assumption on arrival times. Russell and Urban (2008) study the problem with time windows, assuming the travel time follows a shifted gamma distribution. After investigating several different functions of penalty incurred from the deviation of the time window, they develop a tabu-search meta-heuristic. Li et al. (2010), Taş et al. (2013) solve the stochastic vehicle routing problem with time windows based on certain known probability distributions.

To achieve robust performances, Montemanni et al. (2007) assume that the travel times take a range of possible values, and propose several exact algorithms and heuristics to find a route minimizing the robust deviation. Cho et al. (2010) consider the uncertain travel time as an interval, and propose a modified Soyster's model. To adjust conservatism level, they introduce a common parameter to interpolate along the range of data. Following the robust formulation suggested by Bertsimas and Sim (2003), Sungur (2007), Souyris et al. (2013), Agra et al. (2013), and Lee et al. (2012) formulate the vehicle routing problem with time windows, and propose different ap-

proaches, for example, Dantzig-Wolfe decomposition approach and dynamic programming, to solve the problem.

Our main contribution comes from two aspects.

- **Introduction of lateness index:** We propose a new criterion, which we call the *lateness index*, to evaluate how well the arrival times at the nodes, which are uncertain, could meet the deadlines. The criterion can handle risk, where probability distributions of the travel times are known, and ambiguity, where these distributions are partially characterized through descriptive statistics such as means and supports. The criterion possesses important properties including monotonicity, punctuality satisficing, non-abandonment, and convexity. Moreover, it can easily be computed and incorporated in the general routing optimization problem under uncertainty and with deadlines.
- **Optimization of routing problems with lateness index:** To solve the general routing optimization problem under uncertainty and with deadlines, we provide two mathematical programming formulations (a linear decision rule formulation, and a multi-commodity flow formulation) to improve upon a big-M formulation. We show that with the lateness index, we can develop practically “efficient” algorithms to find the exact optimal routing policy through decomposition techniques. We also show the “effectiveness” of our approach through computational studies where we benchmark against other methods.

The chapter is structured as follows. In Section 3.2, we introduce the lateness index as a performance measure for evaluating how arrival times

at nodes meet the deadlines, and present some of its important characteristics. We also explain a special case, in which only one node has a deadline requirement. The associated shortest path problem with deadline is polynomial-time solvable when travel times are independent of each other. In Section 3.3, we formulate the general routing optimization problem with deadlines, and provide a solution methodology. In Section 3.4, we perform several computational studies with encouraging results on the performance of lateness index policies. In Section 3.5, we extend the model to account for correlation between uncertain travel times.

3.2 Lateness Index

We consider a directed network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N} = \{1, \dots, n\}$ represents the set of nodes and \mathcal{A} denotes the set of arcs in the network. We will use (i, j) and a interchangeably to represent an arc in \mathcal{A} . We define $\mathcal{N}_R \subseteq \mathcal{N}$ as the set of nodes that we need to visit. In addition, among these nodes to be visited, we define the subset $\mathcal{N}_D \subseteq \mathcal{N}_R$ as the set of nodes with deadline impositions. Without loss of generality, node $1 \in \mathcal{N}_R \setminus \mathcal{N}_D$ and node $n \in \mathcal{N}_D$ represent the origin and destination nodes respectively. Two common special cases for the set \mathcal{N}_R are $\mathcal{N}_R = \mathcal{N}$, which requires all the nodes in the network to be visited, and $\mathcal{N}_R = \mathcal{N}_D \cup \{1\}$, which corresponds to the situation where only the deadline nodes are required to be visited. For any node set $\mathcal{N}^0 \subset \mathcal{N}$, we define the following arc sets

$$\delta^+(\mathcal{N}^0) \triangleq \{(i, j) \in \mathcal{A} : i \in \mathcal{N}^0, j \in \mathcal{N} \setminus \mathcal{N}^0\},$$

and

$$\delta^-(\mathcal{N}^0) \triangleq \{(i, j) \in \mathcal{A} : i \in \mathcal{N} \setminus \mathcal{N}^0, j \in \mathcal{N}^0\}.$$

Hence, we have $\delta^+(\{n\}) = \emptyset$ and $\delta^-(\{1\}) = \emptyset$.

We consider an off-line routing problem where the routing decisions are made at the beginning before the realization of uncertainty, and they will not change dynamically in response to information updates along the network. Since the travel times along the arcs are uncertain, the actual arrival time, say at each node $i \in \mathcal{N}$, denoted by \tilde{t}_i , is also uncertain. If $i \in \mathcal{N}_D$, then it would be ideal for the uncertain travel time, \tilde{t}_i to always fall below the pre-specified deadline, τ_i . However, as such idealistic solution may not always be feasible, our goal is to find an optimal routing solution such that arrival times at nodes respect deadlines “as much as possible”, while keeping the optimization problem tractable from a practical point of view. In order to do so, we introduce a new performance measure, named *lateness index*, to evaluate how the uncertain arrival times respect the corresponding deadlines from a systematic point of view. Let function $\varphi(\boldsymbol{\alpha})$ be a sub-differentiable mapping $[0, +\infty]^{|\mathcal{N}_D|} \rightarrow [0, +\infty]$ that is convex in $\boldsymbol{\alpha} \geq \mathbf{0}$. Besides, function $\varphi(\boldsymbol{\alpha})$ is non-decreasing in α_i for all $i \in \mathcal{N}_D$, with boundary conditions $\varphi(\mathbf{0}) = 0$ and for all $j \in \mathcal{N}_D$, $\varphi((+\infty, \boldsymbol{\alpha}_{-j})) = \lim_{\alpha_j \rightarrow +\infty} \varphi((\alpha_j, \boldsymbol{\alpha}_{-j})) = +\infty$. The lateness index is formally defined as follows.

Definition 3.1. (Lateness Index) Let $\boldsymbol{\tau} = (\tau_i)_{i \in \mathcal{N}_D}$ represent the deadlines pre-specified on the network, let $\tilde{\boldsymbol{t}} = (\tilde{t}_i)_{i \in \mathcal{N}_D}$ represent the arrival times at corresponding nodes associated with a given routing policy, and let $\boldsymbol{\alpha} =$

$(\alpha_i)_{i \in \mathcal{N}_D}$ be a vector of nonnegative real-valued parameters. The lateness index $\rho_\tau(\tilde{\mathbf{t}}) : \mathcal{V} \rightarrow [0, +\infty]$ is defined as

$$\rho_\tau(\tilde{\mathbf{t}}) = \inf \{ \varphi(\boldsymbol{\alpha}) \mid C_{\alpha_i}(\tilde{t}_i) \leq \tau_i, \alpha_i \geq 0, i \in \mathcal{N}_D \},$$

or $+\infty$ if no such $\boldsymbol{\alpha}$ exists, where $C_\alpha(\tilde{t})$ is the worst-case certainty equivalent under exponential disutility defined as

$$C_\alpha(\tilde{t}) = \begin{cases} \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}}{\alpha} \right) \right), & \text{if } \alpha > 0, \\ \lim_{\gamma \downarrow 0} C_\gamma(\tilde{t}), & \text{if } \alpha = 0. \end{cases}$$

Note that the lateness index involves minimization of a nondecreasing, convex function of the risk tolerance parameters, $\boldsymbol{\alpha}$, while constraining the worst-case certainty equivalent of arrival times within the corresponding deadlines.

Lemma 3.1. The worst-case certainty equivalent has some rather well known and useful properties that we list here:

- (a) $C_\alpha(\tilde{t})$ is decreasing in $\alpha \geq 0$ and strictly decreasing when \tilde{t} is not constant. Moreover,

$$\lim_{\alpha \downarrow 0} C_\alpha(\tilde{t}) = \bar{t}_{\mathbb{F}}, \quad \lim_{\alpha \rightarrow +\infty} C_\alpha(\tilde{t}) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}),$$

where $\bar{t}_{\mathbb{F}} = \inf\{t \in \mathfrak{R} \mid \mathbb{P}(\tilde{t} \leq t) = 1, \forall \mathbb{P} \in \mathbb{F}\}$;

(b) For any $\lambda \in [0, 1]$, $\tilde{t}_1, \tilde{t}_2 \in \mathcal{V}$, and $\alpha_1, \alpha_2 \geq 0$,

$$C_{\lambda\alpha_1 + (1-\lambda)\alpha_2}(\lambda\tilde{t}_1 + (1-\lambda)\tilde{t}_2) \leq \lambda C_{\alpha_1}(\tilde{t}_1) + (1-\lambda)C_{\alpha_2}(\tilde{t}_2);$$

(c) If the random variables $\tilde{t}_1, \tilde{t}_2 \in \mathcal{V}$ are independent from each other,

then for any $\alpha \geq 0$,

$$C_{\alpha}(\tilde{t}_1 + \tilde{t}_2) = C_{\alpha}(\tilde{t}_1) + C_{\alpha}(\tilde{t}_2).$$

Proof. (a) Kaas et al. (2001) has shown that function $\alpha \ln \mathbb{E}_{\mathbb{P}}\left(\exp\left(\frac{\tilde{t}}{\alpha}\right)\right)$ is decreasing in α and strictly decreasing when \tilde{t} is constant. Besides,

$$\lim_{\alpha \downarrow 0} \alpha \ln \mathbb{E}_{\mathbb{P}}\left(\exp\left(\frac{\tilde{t}}{\alpha}\right)\right) = \text{ess sup}(\tilde{t}), \quad \lim_{\alpha \rightarrow \infty} \alpha \ln \mathbb{E}_{\mathbb{P}}\left(\exp\left(\frac{\tilde{t}}{\alpha}\right)\right) = \mathbb{E}_{\mathbb{P}}(\tilde{t}).$$

Hence, taken the supremum over the distributional uncertainty set \mathbb{F} preserves the monotonicity, besides,

$$\lim_{\alpha \downarrow 0} C_{\alpha}(\tilde{t}) = \sup_{\mathbb{P} \in \mathbb{F}} \text{ess sup}(\tilde{t}) = \bar{t}_{\mathbb{F}}, \quad \lim_{\alpha \rightarrow \infty} C_{\alpha}(\tilde{t}) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}).$$

(b) For any $\lambda \in [0, 1]$, we define $\alpha_\lambda = \lambda\alpha_1 + (1 - \lambda)\alpha_2$, we have

$$\begin{aligned}
& C_{\alpha_\lambda}(\lambda\tilde{t}_1 + (1 - \lambda)\tilde{t}_2) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \alpha_\lambda \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\lambda\tilde{t}_1 + (1 - \lambda)\tilde{t}_2}{\alpha_\lambda} \right) \right) \\
&= \alpha_\lambda \sup_{\mathbb{P} \in \mathbb{F}} \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\lambda\alpha_1}{\alpha_\lambda} \frac{\lambda\tilde{t}_1}{\lambda\alpha_1} + \frac{(1 - \lambda)\alpha_2}{\alpha_\lambda} \frac{(1 - \lambda)\tilde{t}_2}{(1 - \lambda)\alpha_2} \right) \right) \\
&\leq \alpha_\lambda \sup_{\mathbb{P} \in \mathbb{F}} \left(\frac{\lambda\alpha_1}{\alpha_\lambda} \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\lambda\tilde{t}_1}{\lambda\alpha_1} \right) \right) + \frac{(1 - \lambda)\alpha_2}{\alpha_\lambda} \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{(1 - \lambda)\tilde{t}_2}{(1 - \lambda)\alpha_2} \right) \right) \right) \\
&\leq \lambda \sup_{\mathbb{P} \in \mathbb{F}} \alpha_1 \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_1}{\alpha_1} \right) \right) + (1 - \lambda) \sup_{\mathbb{P} \in \mathbb{F}} \alpha_2 \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_2}{\alpha_2} \right) \right) \\
&= \lambda C_{\alpha_1}(\tilde{t}_1) + (1 - \lambda) C_{\alpha_2}(\tilde{t}_2),
\end{aligned}$$

where the first inequality holds based on the Holder's inequality.

(c) Since \tilde{t}_1 and \tilde{t}_2 are independently distributed, we have

$$\begin{aligned}
& C_\alpha(\tilde{t}_1 + \tilde{t}_2) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_1 + \tilde{t}_2}{\alpha} \right) \right) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \left(E_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_1}{\alpha} \right) \right) \times E_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_2}{\alpha} \right) \right) \right) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \left(E_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_1}{\alpha} \right) \right) \right) + \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \left(E_{\mathbb{P}} \left(\exp \left(\frac{\tilde{t}_2}{\alpha} \right) \right) \right) \\
&= C_\alpha(\tilde{t}_1) + C_\alpha(\tilde{t}_2). \quad \square
\end{aligned}$$

Remark 3.1. Property (a) shows that function $C_\alpha(\cdot)$ is monotonic in α , the smaller risk tolerance parameter α , the larger certainty equivalent will be.

Property (b) indicates that function $C_\alpha(\tilde{t})$ is jointly convex in (α, \tilde{t}) . Property

(c) explains a very attractive property for optimization, $C_\alpha(\tilde{t})$ is additive for independent random variables.

Function $\varphi(\boldsymbol{\alpha})$ is defined in a general sense, and the travelers could designate their own suitable functions to evaluate the performance based on its properties. Two special cases are $\varphi(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{N}_D} \alpha_i$, and $\varphi(\boldsymbol{\alpha}) = \max_{i \in \mathcal{N}_D} \alpha_i$.

To motivate the lateness index as a reasonable criterion for evaluating how well uncertain arrival times meet deadlines, we next present important properties of this criterion.

Proposition 3.1. Lateness index $\rho_\tau(\tilde{\mathbf{t}})$ satisfies the following properties:

- (a) **Monotonicity:** if $\tilde{\mathbf{t}}, \tilde{\mathbf{v}} \in \mathcal{V}$ and $\tilde{\mathbf{t}} \geq \tilde{\mathbf{v}}$, then $\rho_\tau(\tilde{\mathbf{t}}) \geq \rho_\tau(\tilde{\mathbf{v}})$;
- (b) **Punctuality satisficing:** $\rho_\tau(\boldsymbol{\tau}) = 0$. For any $\tilde{\mathbf{t}}, (\tilde{v}_j, \tilde{\mathbf{t}}_{-j}) \in \mathcal{V}$, if $\tilde{t}_j \leq \tilde{v}_j \leq \tau_j$, then $\rho_\tau((\tilde{t}_j, \tilde{\mathbf{t}}_{-j})) = \rho_\tau((\tilde{v}_j, \tilde{\mathbf{t}}_{-j})) = \rho_\tau((\tau_j, \tilde{\mathbf{t}}_{-j}))$;
- (c) **Non-abandonment:** if there exists $j \in \mathcal{N}_D$, such that $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}_j) > \tau_j$, then $\rho_\tau(\tilde{\mathbf{t}}) = +\infty$;
- (d) **Convexity:** for any $\tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_2 \in \mathcal{V}$ and $\beta \in [0, 1]$, $\rho_\tau(\beta\tilde{\mathbf{t}}_1 + (1 - \beta)\tilde{\mathbf{t}}_2) \leq \beta\rho_\tau(\tilde{\mathbf{t}}_1) + (1 - \beta)\rho_\tau(\tilde{\mathbf{t}}_2)$;
- (e) **Probabilistic envelope:** For any $i \in \mathcal{I}$, let $\rho_i^* = \rho_{\tau_i}(\tilde{t}_i)$, then the probabilistic envelope for the deadline violation is

$$\mathbb{P}(\tilde{t}_i \geq \tau_i + \theta\rho_i^*) \leq \exp(-\theta), \quad \forall \theta \geq 0.$$

Proof. (a). *Monotonicity:* if $\tilde{\mathbf{t}} \geq \tilde{\mathbf{v}}$, for any $\alpha_i \geq 0$ satisfying $C_{\alpha_i}(\tilde{t}_i) \leq \tau_i$, we have $C_{\alpha_i}(\tilde{v}_i) \leq \tau_i$. Therefore, the monotonicity of function $\varphi(\boldsymbol{\alpha})$ indicates $\rho_{\boldsymbol{\tau}}(\tilde{\mathbf{t}}) \geq \rho_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$.

(b). *Punctuality satisfying:* since $C_{\alpha_i}(\tau) = \tau$ for any $\alpha_i \geq 0$, we have $\rho_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \varphi(\mathbf{0}) = 0$. Therefore, if $\tilde{t}_j \leq \tilde{v}_j \leq \tau_j$, we could observe that $C_{\alpha_j}(\tilde{t}_j) \leq C_{\alpha_j}(\tilde{v}_j) \leq \tau_j$ for any $\alpha_j \geq 0$. It follows that

$$\begin{aligned} \rho_{\boldsymbol{\tau}}(\tilde{\mathbf{t}}) &= \inf \{ \varphi(\boldsymbol{\alpha}) \mid C_{\alpha_i}(\tilde{t}_i) \leq \tau_i, \alpha_i \geq 0, i \in \mathcal{N}_D \} \\ &= \inf \{ \varphi(0, \boldsymbol{\alpha}_{-j}) \mid C_{\alpha_i}(\tilde{t}_i) \leq \tau_i, \alpha_i \geq 0, i \in \mathcal{N}_D, i \neq j \} \\ &= \rho_{\boldsymbol{\tau}}((\tilde{v}_j, \tilde{\mathbf{t}}_{-j})) \\ &= \rho_{\boldsymbol{\tau}}((\tau_j, \tilde{\mathbf{t}}_{-j})). \end{aligned}$$

(c). *Non-abandonment:* if there exists $j \in \mathcal{N}_D$, such that $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{t}_j) > \tau_j$, then $C_{\alpha_j}(\tilde{t}_j) > \tau_j$ for any $\alpha_j \geq 0$. It follows that

$$\rho_{\boldsymbol{\tau}}(\tilde{\mathbf{t}}) = \inf \{ \varphi(+\infty, \boldsymbol{\alpha}_{-j}) \mid C_{\alpha_i}(\tilde{t}_i) \leq \tau_i, \alpha_i \geq 0, i \in \mathcal{N}_D, i \neq j \} = +\infty.$$

(d). *Convexity:* let $\rho_{\boldsymbol{\tau}}(\tilde{\mathbf{t}}_1) = \varphi(\boldsymbol{\alpha}_1)$ and $\rho_{\boldsymbol{\tau}}(\tilde{\mathbf{t}}_2) = \varphi(\boldsymbol{\alpha}_2)$, we have

$$C_{\alpha_{1i}}(\tilde{t}_{1i}) \leq \tau_i, \quad C_{\alpha_{2i}}(\tilde{t}_{2i}) \leq \tau_i, \quad \forall i \in \mathcal{N}_D.$$

Since the worst-case certainty equivalent satisfies for any $\beta \in [0, 1]$ and $i \in$

\mathcal{N}_D ,

$$\begin{aligned} C_{\beta\alpha_{1i}+(1-\beta)\alpha_{2i}}(\beta\tilde{t}_{1i}+(1-\beta)\tilde{t}_{2i}) &\leq \beta C_{\alpha_{1i}}(\tilde{t}_{1i})+(1-\beta)C_{\alpha_{2i}}(\tilde{t}_{2i}) \\ &\leq \beta\tau_i+(1-\beta)\tau_i \\ &= \tau_i, \end{aligned}$$

we have

$$\begin{aligned} \rho_\tau(\beta\tilde{\mathbf{t}}_1+(1-\beta)\tilde{\mathbf{t}}_2) &\leq \varphi(\beta\boldsymbol{\alpha}_1+(1-\beta)\boldsymbol{\alpha}_2) \\ &\leq \beta\varphi(\boldsymbol{\alpha}_1)+(1-\beta)\varphi(\boldsymbol{\alpha}_2) \\ &= \beta\rho_\tau(\tilde{\mathbf{t}}_1)+(1-\beta)\rho_\tau(\tilde{\mathbf{t}}_2). \end{aligned}$$

(e). *Probabilistic envelope*: For any $i \in \mathcal{I}$, since $\rho_i^* = \rho_{\tau_i}(\tilde{t}_i)$, we have for any $\theta \geq 0$,

$$\begin{aligned} \mathbb{P}(\tilde{t}_i \geq \tau_i + \theta\rho_i^*) &= \mathbb{P}\left(\frac{\tilde{t}_i - \tau_i}{\rho_i^*} \geq \theta\right) \\ &= \mathbb{P}\left(\exp\left(\frac{\tilde{t}_i - \tau_i}{\rho_i^*}\right) \geq \exp(\theta)\right) \\ &\leq \frac{\mathbb{E}_{\mathbb{P}}(\exp((\tilde{t}_i - \tau_i)/\rho_i^*))}{\exp(\theta)} \\ &\leq \exp(-\theta). \end{aligned}$$

The first inequality holds because of the Markov inequality, while the second inequality holds since $C_{\rho_i^*}(\tilde{t}_i) \leq \tau_i$. \square

Remark 3.2. Monotonicity captures travelers' intrinsic preferences for a shorter travel time, that is, if for any $i \in \mathcal{N}_D$, the travel time \tilde{t}_i is state-wise greater than its counterpart \tilde{v}_i , the lateness index returns a larger value for

\tilde{t} . Punctuality satisficing indicates that an arrival time that is guaranteed to meet the deadline is most preferred. However, any improvement on that arrival time will not affect the lateness index. For the lateness index to be finite, the non-abandonment property requires all the arrival times to meet the corresponding deadlines in expectation. The convexity property serves two purposes. First, it is synonymous with risk pooling and diversification preference in the context of risk management. If two arrival profiles, \tilde{t}_1 and \tilde{t}_2 are preferred over the profile \tilde{t}_3 , then any convex combination of these two profiles will be preferred over \tilde{t}_3 . Moreover, as we will later illustrate, it has important ramifications in the context of formulating a computationally attractive problem which we can use to find optimal solutions via standard solvers.

When only one node has a deadline requirement, i.e., $\mathcal{N}_D = \{n\}$, the lateness index reduces to

$$\rho_\tau(\tilde{t}) = \inf \{ \varphi(\alpha) \mid C_\alpha(\tilde{t}) \leq \tau, \alpha \geq 0 \},$$

or $+\infty$ if no such α exists. This criterion is similar to the riskiness index of Aumann and Serrano (2008). It is a particular case of the satisficing measure proposed by Brown and Sim (2009) and Brown et al. (2012) for evaluating uncertain monetary outcomes and has been applied in project selection by Hall et al. (2014). We use this lateness index as an optimization criterion to

formulate a shortest path problem under uncertainty with deadline prescribed at the destination node n only.

Assuming the deadline at node n is given by τ , we are seeking a path from node 1 to node n that minimizes the lateness index. Given the definition of the lateness index, we formulate the problem as follows:

$$\begin{aligned} \inf \quad & \varphi(\alpha) \\ \text{s.t.} \quad & C_\alpha(\tilde{\mathbf{c}}\mathbf{x}) \leq \tau, \\ & \alpha \geq 0, \\ & \mathbf{x} \in \mathcal{X}_{SP}, \end{aligned}$$

where,

$$\mathcal{X}_{SP} = \left\{ \mathbf{x} \in \{0, 1\}^{|\mathcal{A}|} \left| \sum_{a \in \delta^+(i)} x_a - \sum_{a \in \delta^-(i)} x_a = \begin{cases} 1, & \text{when } i = 1, \\ -1, & \text{when } i = n, \\ 0, & \text{otherwise} \end{cases} \right. \right\},$$

with the standard convention that a sum of an empty set of indices is 0. Since functions $\varphi(\alpha)$ and $C_\alpha(\cdot)$ are monotone in α , bisection can be used to calculate the optimal α if, for any given $\alpha \geq 0$, we can solve the following sub-problem:

$$\min_{\mathbf{x} \in \mathcal{X}_{SP}} C_\alpha(\tilde{\mathbf{c}}\mathbf{x}) \quad (3.1)$$

When the travel time on each arc is independent of each other, similar to Chapter 2, the problem is a classical shortest path problem. In the next section, we show that the worst-case certainty equivalent can be calculated

under both a given probabilistic distribution and distributional uncertainty set of travel time. Therefore, as far as we know, the shortest path problem based on minimizing the lateness index is possibly the only formulation that incorporates a deadline, accounts for both probabilistic and ambiguous distributions of travel times and retains a polynomial time complexity. Nevertheless, when the travel times are correlated, Chapter 2 has shown that the recognition version of Problem (3.1) is NP-complete. In Section 3.5, we provide one formulation to address the correlation issue. Observe that in the presence of multiple deadlines, the bisection process would not be generalizable. Hence, in the following section, we explore a different solution approach to address the general routing problem.

3.3 General Routing Optimization Problem with Deadlines

We propose here a general routing optimization model when there is a subset of nodes with deadline requirements. Our objective is to determine a routing policy such that the route (a) starts at the origin node 1, ends at the destination node n , (b) visits each node in set \mathcal{N}_R exactly once, and the rest of nodes at most once, and (c) effectively respects the deadlines specified at nodes in set \mathcal{N}_D . We first assume the travel time on each arc is independent of each other.

3.3.1 Model definition

We formulate a general routing optimization problem as follows.

$$\begin{aligned}
& \inf \quad \rho_{\tau}(\tilde{\mathbf{t}}) \\
& \text{s.t.} \quad \tilde{t}_j \geq \tilde{t}_i + \tilde{c}_{ij}x_{ij} - (1 - x_{ij})M, \quad (i, j) \in \mathcal{A}, & \text{(a)} \\
& \quad \tilde{t}_1 = 0, & \text{(b)} \\
& \quad \mathbf{x} \in \mathcal{X}_{RO},
\end{aligned} \tag{3.2}$$

where

$$\mathcal{X}_{RO} = \left\{ \mathbf{x} \in \{0, 1\}^{|\mathcal{A}|} \left| \begin{array}{ll} \sum_{a \in \delta^+(i)} x_a = 1, & i \in \mathcal{N}_R \setminus \{n\}, \\ \sum_{a \in \delta^-(i)} x_a = 1, & i \in \mathcal{N}_R \setminus \{1\}, \\ \sum_{a \in \delta^+(i)} x_a \leq 1, & i \in \mathcal{N} \setminus \mathcal{N}_R, \\ \sum_{a \in \delta^-(i)} x_a - \sum_{a \in \delta^+(i)} x_a = 0, & i \in \mathcal{N} \setminus \mathcal{N}_R \end{array} \right. \right\}.$$

The objective is to minimize the lateness index for all the nodes with deadline requirements. Constraint (3.2a) uses a big-M method to calculate the arrival time at each node by linking it to its successive node's arrival time, eventually eliminating subtours. Constraint (3.2b) specifies that the starting time at node 1 is zero. Set \mathcal{X}_{RO} represents flow conservation constraints, which enforces that each node in set \mathcal{N}_R should be visited exactly once, while the other nodes can be visited at most once.

When there is a subset of nodes required to be visited, i.e., $\mathcal{N}_R \subseteq \mathcal{N}$, one intuitive way to formulate this problem is to convert the current network into

a standard network, in which all the nodes belong to \mathcal{N}_R , and the arc travel time is represented by the shortest paths between each pair of nodes. However, it is worth pointing out that even if the original network is sparse, this transformation will lead to a complete graph with $|\mathcal{N}_R|(|\mathcal{N}_R| - 1)/2$ arcs, which may increase the number of decision variables substantially. Interested readers could refer to Cornuéjols et al. (1985) for more details. Besides, the new arc travel times in the transformed network may not necessarily be independent, even though they were independent in the original one, since the shortest paths between different pairs of nodes may share common arcs.

According to the definition of the lateness index, Problem (3.2) is equivalent to

$$\begin{aligned}
& \inf \quad \varphi(\boldsymbol{\alpha}) \\
& \text{s.t.} \quad C_{\alpha_i}(\tilde{t}_i) \leq \tau_i, & i \in \mathcal{N}_D, \\
& \quad \alpha_i \geq 0, & i \in \mathcal{N}_D, \\
& \quad \tilde{t}_j \geq \tilde{t}_i + \tilde{c}_{ij}x_{ij} - (1 - x_{ij})M, \quad (i, j) \in \mathcal{A}, \\
& \quad \tilde{t}_1 = 0, \\
& \quad \boldsymbol{x} \in \mathcal{X}_{RO}.
\end{aligned} \tag{3.3}$$

3.3.2 Model reformulation

In Problem (3.3), the choice of M could pose serious computational issues. Smaller M may rule out the actual optimal solution from the feasible set, while larger M may lead to longer computation time. Moreover, when the arc travel time \tilde{c}_{ij} follows a continuous probability distribution, the uncertainty of travel time \tilde{t}_i and \tilde{t}_j may yield an infinite number of constraints. In this section, we propose efficient ways to address these issues. Two formulation

techniques are introduced.

Linear decision rule formulation

The first formulation is inspired by the linear decision rule, a common approach in robust optimization to address problems with recourse. We introduce auxiliary variables $\mathbf{s}_{LDR}^i \in \mathfrak{R}_+^{|\mathcal{A}|}$ for all $i \in \mathcal{N}$, and define a $|\mathcal{A}| \times |\mathcal{N}|$ matrix $\mathbf{s}_{LDR} = (\mathbf{s}_{LDR}^i)_{i \in \mathcal{N}}$. The linear decision rule formulation is provided as follows.

Proposition 3.2. Problem (3.3) can be equivalently formulated as

$$\begin{aligned}
 & \inf \quad \varphi(\boldsymbol{\alpha}) \\
 & \text{s.t.} \quad C_{\alpha_i} (\tilde{\mathbf{c}}' \mathbf{s}_{LDR}^i) \leq \tau_i, \quad i \in \mathcal{N}_D, \\
 & \quad \quad \alpha_i \geq 0, \quad i \in \mathcal{N}_D, \\
 & \quad \quad (\mathbf{x}, \mathbf{s}_{LDR}) \in \mathcal{S}_{LDR},
 \end{aligned} \tag{3.4}$$

where

$$\mathcal{S}_{LDR} = \left\{ \begin{array}{l} \mathbf{x} \in \{0, 1\}^{|\mathcal{A}|} \\ \mathbf{s} \in \mathbb{R}_+^{|\mathcal{A}| \times |\mathcal{N}|} \end{array} \left| \begin{array}{ll} \mathbf{x} \in \mathcal{X}_{RO}, & \text{(a)} \\ s_a^j - s_a^i \leq 1 - x_{ij}, & a, (i, j) \in \mathcal{A}, a \neq (i, j), \text{ (b)} \\ s_a^j - s_a^i \geq x_{ij} - 1, & a, (i, j) \in \mathcal{A}, a \neq (i, j), \text{ (c)} \\ s_a^i = x_a, & a \in \delta^-(i), i \in \mathcal{N} \setminus \{1\}, \text{ (d)} \\ s_a^i = 0, & a \in \delta^+(i), i \in \mathcal{N} \setminus \{1, n\}, \text{ (e)} \\ s_a^1 = 0, & a \in \mathcal{A}, \text{ (f)} \\ \sum_{a \in \mathcal{A}} s_a^i \leq |\mathcal{A}| \sum_{a \in \delta^-(i)} x_a, & i \in \mathcal{N} \setminus \mathcal{N}_R, \text{ (g)} \end{array} \right. \right\}. \quad (3.5)$$

Proof. For notational simplicity, let us omit the subscript “LDR”. First, we prove by contradiction that, with this linear decision rule formulation, there exists no subtour in a feasible solution. Suppose there exists a subtour going through $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_{k-1} \rightarrow i_k \rightarrow i_1$, that is, $x_{i_1 i_2} = \dots = x_{i_{k-1} i_k} = x_{i_k i_1} = 1$. Constraints (3.5d) and (3.5e) indicate $s_{i_1 i_2}^{i_2} = 1, s_{i_1 i_2}^{i_1} = 0$, while constraints (3.5b) and (3.5c) suggest $s_{i_1 i_2}^{i_k} = s_{i_1 i_2}^{i_1} = 0$ and $s_{i_1 i_2}^{i_k} = s_{i_1 i_2}^{i_{k-1}} = \dots = s_{i_1 i_2}^{i_2} = 1$, respectively, which generates the contradiction.

Now, we prove that the arrival time at each deadline node can be written as $\tilde{t}_i = \tilde{\mathbf{c}}' \mathbf{s}^i$. After the decision variable \mathbf{x} is selected, we observe that the arrival time \tilde{t}_i representing the path travel time between the origin node 1 and node i is only a recourse variable. We prove it by induction. Suppose \mathbf{x} is given, representing a path $i_0 \rightarrow i_1 \rightarrow \dots \rightarrow i_{k-1} \rightarrow i_k$, in which $i_0 = 1, i_k = n$, and $\mathcal{N}_D \subset \{i_0, i_1, \dots, i_k\}$. Correspondingly, $x_{i_0 i_1} = \dots = x_{i_{k-1} i_k} = 1$, and all

others are 0. To start, we have $\tilde{t}_{i_0} = \tilde{t}_1 = 0 = \tilde{\mathbf{c}}' \mathbf{s}^1$. Since $x_{i_l i_{l+1}} = 1$, for all $l = 0, \dots, k-1$, constraints (3.5b)~(3.5e) indicate that

$$\begin{aligned} s_a^{i_{l+1}} &= s_a^{i_l}, \quad a \neq (i_l, i_{l+1}), a \in \mathcal{A}, \\ s_{i_l i_{l+1}}^{i_{l+1}} &= 1, \\ s_{i_l i_{l+1}}^{i_l} &= 0. \end{aligned}$$

If $\tilde{t}_{i_l} = \tilde{\mathbf{c}}' \mathbf{s}^{i_l}$, we could get the formulation of $\tilde{t}_{i_{l+1}}$ from the above as

$$\begin{aligned} \tilde{t}_{i_{l+1}} &= \tilde{t}_{i_l} + \tilde{c}_{i_l i_{l+1}} = \tilde{\mathbf{c}}' \mathbf{s}^{i_l} + \tilde{c}_{i_l i_{l+1}} = \sum_{a \neq (i_l, i_{l+1}), a \in \mathcal{A}} \tilde{c}_a s_a^{i_l} + \tilde{c}_{i_l i_{l+1}} s_{i_l i_{l+1}}^{i_l} + \tilde{c}_{i_l i_{l+1}} \\ &= \sum_{a \neq (i_l, i_{l+1}), a \in \mathcal{A}} \tilde{c}_a s_a^{i_{l+1}} + 0 + \tilde{c}_{i_l i_{l+1}} s_{i_l i_{l+1}}^{i_{l+1}} = \tilde{\mathbf{c}}' \mathbf{s}^{i_{l+1}}. \end{aligned}$$

Finally, we observe that when the feasible solution \mathbf{x} is given, the solution \mathbf{s} is uniquely determined, and $\mathbf{s} \in \{0, 1\}^{|\mathcal{A}| \times |\mathcal{M}|}$. If $i \notin \{i_0, i_1, \dots, i_k\}$, based on constraint (3.5g), $\sum_{a \in \mathcal{A}} s_a^i \leq |\mathcal{A}| \sum_{a \in \delta^-(i)} x_a = 0$, which leads to $\mathbf{s}^i = \mathbf{0}$. While $i = i_{l+1}, l = 0, \dots, k-1$, if $a \neq (i_0, i_1), \dots, (i_l, i_{l+1}), a \in \mathcal{A}$, constraints (3.5b), (3.5c) and (3.5f) suggest $s_a^{i_{l+1}} = s_a^{i_l} = \dots = s_a^{i_0} = 0$. Otherwise, from constraints (3.5b)~(3.5e), we have

$$\begin{aligned} s_{i_0 i_1}^{i_{l+1}} &= s_{i_0 i_1}^{i_l} = \dots = s_{i_0 i_1}^{i_1} = 1, \quad s_{i_0 i_1}^{i_0} = 0, \\ s_{i_1 i_2}^{i_{l+1}} &= s_{i_1 i_2}^{i_l} = \dots = s_{i_1 i_2}^{i_2} = 1, \quad s_{i_1 i_2}^{i_1} = s_{i_1 i_2}^{i_0} = 0, \\ &\vdots \\ s_{i_l i_{l+1}}^{i_{l+1}} &= 1, \quad s_{i_l i_{l+1}}^{i_l} = s_{i_l i_{l+1}}^{i_{l-1}} \dots = s_{i_l i_{l+1}}^{i_0} = 0. \end{aligned}$$

The solution satisfies all constraints (3.5b)~(3.5g). \square

In this formulation, we have a total of $|\mathcal{N}||\mathcal{A}|+|\mathcal{N}_D|$ continuous variables, $|\mathcal{A}|$ binary variables, and $2|\mathcal{A}|^2 + |\mathcal{A}| + 3|\mathcal{N}| - |\mathcal{N}_R| + |\mathcal{N}_D| - 2$ ($\approx O(|\mathcal{A}|^2)$) constraints.

Multi-commodity flow formulation

Apart from the linear decision rule formulation, we can also adapt the multi-commodity flow (MCF) formulation of the traveling salesman problem to reformulate Problem (3.3). We add auxiliary variables $\mathbf{s}_{MCF}^i \in \mathfrak{R}_+^{|\mathcal{A}|}$ for all $i \in \mathcal{N}$, and define a $|\mathcal{A}| \times |\mathcal{N}|$ matrix $\mathbf{s}_{MCF} = (\mathbf{s}_{MCF}^i)_{i \in \mathcal{N}}$. The formulation is as follows.

Proposition 3.3. Problem (3.3) can be equivalently written as

$$\begin{aligned}
 & \inf \quad \varphi(\boldsymbol{\alpha}) \\
 & \text{s.t.} \quad C_{\alpha_i} (\tilde{\mathbf{c}}' \mathbf{s}_{MCF}^i) \leq \tau_i, \quad i \in \mathcal{N}_D, \\
 & \quad \alpha_i \geq 0, \quad i \in \mathcal{N}_D, \\
 & \quad (\mathbf{x}, \mathbf{s}_{MCF}) \in \mathcal{S}_{MCF},
 \end{aligned} \tag{3.6}$$

where

$$\mathcal{S}_{MCF} = \left\{ \begin{array}{l} \mathbf{x} \in \{0, 1\}^{|\mathcal{A}|} \\ \mathbf{s} \in \mathfrak{R}_+^{|\mathcal{A}| \times |\mathcal{N}|} \end{array} \middle| \begin{array}{l} \mathbf{x} \in \mathcal{X}_{RO}, \quad (\text{a}) \\ \sum_{a \in \delta^-(u)} s_a^i - \sum_{a \in \delta^+(u)} s_a^i = 0, \quad i \in \mathcal{N} \setminus \{1\}, \\ u \in \mathcal{N} \setminus \{1, n, i\}, \quad (\text{b}) \\ \sum_{a \in \delta^+(1)} s_a^i = \sum_{a \in \delta^-(i)} x_a, \quad i \in \mathcal{N} \setminus \{1\}, \quad (\text{c}) \\ \sum_{a \in \delta^-(i)} s_a^i - \sum_{a \in \delta^+(i)} s_a^i = \sum_{a \in \delta^-(i)} x_a, \quad i \in \mathcal{N} \setminus \{1\}, \quad (\text{d}) \\ s_a^i \leq x_a, \quad i \in \mathcal{N} \setminus \{1\}, \\ a \in \mathcal{A}, \quad (\text{e}) \\ s_a^1 = 0, \quad a \in \mathcal{A}, \quad (\text{f}) \end{array} \right\}. \quad (3.7)$$

Proof. As in the proof of the validity of the LDR formulation, we first prove by contradiction that the feasible solution does not contain a subtour. We observe that constraint (3.7a) coupled with (3.7e) indicate that constraint (3.7d) could be equivalently written as

$$s_a^i = 0, \quad a \in \delta^+(i), i \in \mathcal{N} \setminus \{1\}, \quad (3.7g)$$

$$s_a^i = x_a, \quad a \in \delta^-(i), i \in \mathcal{N} \setminus \{1\}. \quad (3.7h)$$

If there exists a subtour $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k \rightarrow i_1$ in the feasible solution, we could infer that nodes $1, n \notin \{i_1, i_2, \dots, i_k\}$ since $\delta^+(n) = \emptyset$ and $\delta^-(1) = \emptyset$.

Therefore, we have

$$\begin{aligned}
\text{Constraint (3.7a)} &\implies 1 = x_{i_k i_1} = \sum_{a \in \delta^-(i_1)} x_a, \\
\text{Constraint (3.7h)} &\implies x_{i_k i_1} = s_{i_k i_1}^{i_1}, \quad \sum_{a \in \delta^-(i_1)} x_a = \sum_{a \in \delta^-(i_1)} s_a^{i_1}, \\
\text{Constraints (3.7a), (3.7e)} &\implies s_{i_k i_1}^{i_1} \leq \sum_{a \in \delta^+(i_k)} s_a^{i_1} \leq \sum_{a \in \delta^+(i_k)} x_a \leq 1, \\
\text{Constraint (3.7b)} &\implies \sum_{a \in \delta^+(i_k)} s_a^{i_1} = \sum_{a \in \delta^-(i_k)} s_a^{i_1}
\end{aligned}$$

With the above constraints, we could show subsequently,

$$\begin{aligned}
1 &= \sum_{a \in \delta^-(i_1)} x_a = \sum_{a \in \delta^-(i_1)} s_a^{i_1} = s_{i_k i_1}^{i_1} = \sum_{a \in \delta^+(i_k)} s_a^{i_1} = \sum_{a \in \delta^-(i_k)} s_a^{i_1} = \dots \\
&= \sum_{a \in \delta^-(i_2)} s_a^{i_1} = \sum_{a \in \delta^+(i_1)} s_a^{i_1},
\end{aligned}$$

whereas constraint (3.7g) shows $\sum_{a \in \delta^+(i_1)} s_a^{i_1} = 0$.

Next, we explain that when the feasible solution is given, the artificial decision variables \mathbf{s} are uniquely determined. Suppose there exists a feasible path $i_0 \rightarrow i_1 \rightarrow \dots \rightarrow i_{k-1} \rightarrow i_k$, in which $i_0 = 1, i_k = n$, and we define arc set $\mathcal{A}_x = \{(i_0, i_1), \dots, (i_{k-1}, i_k)\}$, correspondingly,

$$x_a = \begin{cases} 1, & \text{when } a \in \mathcal{A}_x, \\ 0, & \text{otherwise.} \end{cases}$$

From constraints (3.7e) and (3.7f), we observe $s_a^1 = 0, a \in \mathcal{A}$, and $s_a^i = 0, i \in \mathcal{N} \setminus \{1\}, a \in \mathcal{A} \setminus \mathcal{A}_x$, which has $|\mathcal{A}| + (|\mathcal{A}| - k)(|\mathcal{N}| - 1)$ zero variables. For the

rest of decision variables $s_a^i, i \in \mathcal{N} \setminus \{1\}, a \in \mathcal{A}_x$, we could derive the solution from constraints (3.7b)~(3.7e) as:

For any $i_t, t \in \{1, \dots, k\}$,

$$\text{constraint (3.7c)} \implies s_{i_0 i_1}^{i_t} = \sum_{a \in \delta^-(i_t)} x_a = 1,$$

$$\text{constraint (3.7b)} \implies s_{i_{t-1} i_t}^{i_t} = \dots = s_{i_0 i_1}^{i_t} = 1,$$

$$\text{constraint (3.7d)} \implies s_{i_t i_{t+1}}^{i_t} = 0,$$

$$\text{constraint (3.7b)} \implies s_{i_t i_{t+1}}^{i_t} = \dots = s_{i_{k-1} i_k}^{i_t} = 0.$$

For any $i \notin \{i_0, i_1, \dots, i_k\}$,

$$\text{constraint (3.7c)} \implies s_{i_0 i_1}^i = \sum_{a \in \delta^-(i)} x_a = 0,$$

$$\text{constraint (3.7b)} \implies s_{i_{k-1} i_k}^i = s_{i_{k-2} i_{k-1}}^i = \dots = s_{i_0 i_1}^i = 0.$$

Clearly, this solution satisfies all constraints (3.7b)~(3.7f). Hence, \mathbf{s} is a uniquely determined integer solution when \mathbf{x} is a feasible solution.

In addition, due to the integer property of \mathbf{s} , $\tilde{\mathbf{c}}' \mathbf{s}^i$ actually represents the cost of sending this unit flow, which is equivalently interpreted as the travel time from node 1 to node i , i.e., $\tilde{t}_i = \tilde{\mathbf{c}}' \mathbf{s}^i$. Therefore,

$$C_{\alpha_i}(\tilde{t}_i) = C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}_{MCF}^i). \quad \square$$

In this MCF formulation, the additional non-negative variable s_a^i is defined as the amount of commodity i passing through arc a , and constraints (3.7b)~(3.7e) ensure that $\sum_{a \in \delta^-(i)} x_a$ unit of commodity i travels from source

node 1 to sink node i , with capacity bound x_a on arc a . Constraint (3.7b) enforces the requirement that for commodity i , the incoming flow to node l , which is different from node i , should be equal to the outgoing flow. Constraint (3.7c) ensures that for all these $|\mathcal{N}| - 1$ commodities, node 1 is their common source node. Constraint (3.7d) represents that node i in set $\mathcal{N} \setminus \{1\}$ is the sink node for commodity i . Constraint (3.7e) describes that commodity flow can only go through the selected arcs with maximal amount 1. This MCF formulation was first proposed by Claus (1984), and has been verified as a relative strong formulation for the traveling salesman problem in terms of LP relaxation (Öncan et al. 2009). Letchford et al. (2013) also extend this formulation to the Steiner traveling salesman problem. In total, the MCF formulation has $|\mathcal{N}||\mathcal{A}| + |\mathcal{N}_D|$ continuous variables, $|\mathcal{A}|$ binary variables, and $|\mathcal{N}||\mathcal{A}| + |\mathcal{N}|^2 + |\mathcal{N}_D| - 1$ ($\approx O(|\mathcal{N}||\mathcal{A}|)$) constraints.

For the feasible sets \mathcal{S}_{LDR} and \mathcal{S}_{MCF} , we relax the binary constraints for \mathbf{x} , such that $\mathbf{x} \in [0, 1]^{|\mathcal{A}|}$, and define the corresponding feasible sets after linear relaxations as P_{LDR} and P_{MCF} , respectively. We provide counter examples to show that

$$P_{LDR} \subsetneq P_{MCF} \quad \text{and} \quad P_{MCF} \subsetneq P_{LDR}.$$

We only consider a five nodes network shown in Figure 3.1, where $\mathcal{N} = \{1, 2, 3, 4, 5\}$, $\mathcal{N}_R = \{1, 2, 5\}$, $\mathcal{N}_D = \{2, 5\}$. The solution $(\mathbf{x}, \mathbf{s}_{LDR})$ given as follows satisfies $(\mathbf{x}, \mathbf{s}_{LDR}) \in P_{LDR}$ and $(\mathbf{x}, \mathbf{s}_{LDR}) \notin P_{MCF}$, since $s_{14}^2 = \frac{1}{2}$ violates

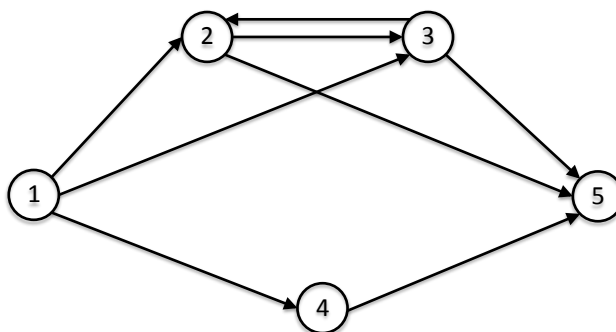


Fig. 3.1: An illustrative example explaining the difference between LDR and MCF formulations.

constraint (3.7e).

(i, j)	(1, 2)	(1, 3)	(2, 3)	(2, 5)	(3, 2)	(3, 5)	(1, 4)	(4, 5)
\mathbf{x}	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
\mathbf{s}_{LDR}^1	0	0	0	0	0	0	0	0
\mathbf{s}_{LDR}^2	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0
\mathbf{s}_{LDR}^3	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0
\mathbf{s}_{LDR}^4	0	0	0	0	0	0	0	0
\mathbf{s}_{LDR}^5	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0

Besides, we could also construct a solution $(\mathbf{x}, \mathbf{s}_{MCF})$ as

(i, j)	(1, 2)	(1, 3)	(2, 3)	(2, 5)	(3, 2)	(3, 5)	(1, 4)	(4, 5)
\mathbf{x}	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
\mathbf{s}_{MCF}^1	0	0	0	0	0	0	0	0
\mathbf{s}_{MCF}^2	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	0	0
\mathbf{s}_{MCF}^3	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0
\mathbf{s}_{MCF}^4	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0
\mathbf{s}_{MCF}^5	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0

and $(\mathbf{x}, \mathbf{s}_{MCF}) \in P_{MCF}$, $(\mathbf{x}, \mathbf{s}_{MCF}) \notin P_{LDR}$, since $s_{23}^4 = \frac{1}{2}$ violates constraint (3.5g).

3.3.3 Solution procedure

With the above formulations, the general routing optimization problem is still complicated since the function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ is non-linear in α_i , and involves the uncertain travel time $\tilde{\mathbf{c}}$. In this section, we further study the function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ in Problems (3.4) and (3.6), and develop algorithms to solve them. As the approach is applicable to both the LDR and MCF formulations, we will drop the subscript for notational simplicity. To guarantee the feasibility of the problem, we impose the requirement for the deadline $\boldsymbol{\tau}$, such that there exists a feasible solution $\mathbf{s} \in \mathcal{S}$ satisfying

$$\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{\mathbf{c}}' \mathbf{s}^i) \leq \tau_i, \quad i \in \mathcal{I}.$$

This implies that the deadline must be set to guarantee that there exists a feasible solution in which we can stay within the deadline in expectation. This assumption is reasonable since violating it indicates the optimal value is infinite, and hence, the deadline is irrational. Now, the constraint set is updated as

$$\mathcal{X} = \mathcal{S} \cap \left\{ (\mathbf{x}, \mathbf{s}) \left| \sum_{a \in \mathcal{A}} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} (\tilde{c}_a) s_a^i \leq \tau_i, \quad i \in \mathcal{N}_D \right. \right\}. \quad (3.8)$$

Given $(\mathbf{x}, \mathbf{s}) \in \mathcal{S}$, we define function $f(\mathbf{s})$ as

$$\begin{aligned} f(\mathbf{s}) = \inf \quad & \varphi(\boldsymbol{\alpha}) \\ \text{s.t.} \quad & C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) \leq \tau_i, \quad i \in \mathcal{N}_D, \\ & \alpha_i \geq 0, \quad i \in \mathcal{N}_D. \end{aligned} \quad (3.9)$$

Observing that functions $\varphi(\boldsymbol{\alpha})$ and $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ are both convex in α_i , Problem (3.9) is a classical convex problem, which could be solved efficiently. We next show the convexity of function $f(\mathbf{s})$ and concentrate on the calculation of its subgradient.

Calculation of the subgradient of $f(\mathbf{s})$

The Lagrange function $L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$ of Problem (3.9) is given by

$$L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \varphi(\boldsymbol{\alpha}) + \sum_{i \in \mathcal{N}_D} \lambda_i (C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) - \tau_i),$$

where λ_i is the Lagrange multiplier associated with the inequality constraint $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) \leq \tau_i$. We next show that the subgradient of $f(\mathbf{s})$ can be calculated through its Lagrange function.

Proposition 3.4. Function $f(\mathbf{s})$ is convex in \mathbf{s} , and if the vector $\begin{pmatrix} d_{\mathbf{s}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \\ d_{\boldsymbol{\alpha}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \end{pmatrix}$ is the subgradient of function $L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*)$ at $(\mathbf{s}, \boldsymbol{\alpha}^*)$, and $d_{\boldsymbol{\alpha}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) = 0$, then $d_{\mathbf{s}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$ is the subgradient of $f(\mathbf{s})$, where

$$(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \in Z(\mathbf{s}) = \left\{ (\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\lambda}}) \mid L(\mathbf{s}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\lambda}}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \right\}.$$

Proof. Let $\boldsymbol{\alpha}^s, \boldsymbol{\alpha}^y$ be the optimal solution of $f(\mathbf{s})$ and $f(\mathbf{y})$ respectively, such that $f(\mathbf{s}) = \varphi(\boldsymbol{\alpha}^s)$ and $f(\mathbf{y}) = \varphi(\boldsymbol{\alpha}^y)$. With the convexity of function $C_{\alpha_i}(\tilde{t}_i)$ described in Lemma 3.1, function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ is jointly convex in (α_i, \mathbf{s}^i) . It implies that for any $i \in \mathcal{I}$,

$$\begin{aligned} c_{\beta\alpha_i^s + (1-\beta)\alpha_i^y}(\tilde{\mathbf{c}}'(\beta\mathbf{s}^i + (1-\beta)\mathbf{y}^i)) &\leq \beta C_{\alpha_i^s}(\tilde{\mathbf{c}}' \mathbf{s}^i) + (1-\beta) C_{\alpha_i^y}(\tilde{\mathbf{c}}' \mathbf{y}^i) \\ &\leq \beta\tau_i + (1-\beta)\tau_i \\ &= \tau_i. \end{aligned}$$

In other words, solution $\beta\boldsymbol{\alpha}^s + (1-\beta)\boldsymbol{\alpha}^y$ satisfies the constraints

$$C_{\beta\alpha_i^s + (1-\beta)\alpha_i^y}(\tilde{\mathbf{c}}'(\beta\mathbf{s}^i + (1-\beta)\mathbf{y}^i)) \leq \tau_i, \quad i \in \mathcal{N}_D.$$

Hence, with the convexity of $\varphi(\boldsymbol{\alpha})$,

$$f(\beta \mathbf{s} + (1-\beta) \mathbf{y}) \leq \varphi(\beta \boldsymbol{\alpha}^s + (1-\beta) \boldsymbol{\alpha}^y) \leq \beta \varphi(\boldsymbol{\alpha}^s) + (1-\beta) \varphi(\boldsymbol{\alpha}^y) = \beta f(\mathbf{s}) + (1-\beta) f(\mathbf{y}),$$

which indicates $f(\mathbf{s})$ is convex in \mathbf{s} .

Since

$$\lim_{\alpha_i \rightarrow +\infty} C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{\mathbf{c}}' \mathbf{s}^i) = \sum_{a \in \mathcal{A}} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{c}_a) s_a^i \leq \tau_i, \quad i \in \mathcal{N}_D,$$

with the monotonicity of function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ in α , there exists an $\boldsymbol{\alpha} > \mathbf{0}$ such that

$$C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) < \tau_i, \quad i \in \mathcal{N}_D.$$

Observing that Problem (3.9) is a convex problem, we have strong duality

$$f(\mathbf{s}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} g(\mathbf{s}, \boldsymbol{\lambda})$$

holds because the constraint qualification (in particular, Slater's condition) can be satisfied.

Since for all $i \in \mathcal{N}_D$, function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ is jointly convex in (α_i, \mathbf{s}^i) , as an immediate conclusion, function $L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$ is also jointly convex in $(\mathbf{s}, \boldsymbol{\alpha})$

given $\boldsymbol{\lambda} \geq \mathbf{0}$. Therefore, based on strong duality,

$$\begin{aligned}
f(\mathbf{y}) - f(\mathbf{s}) &= \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) - \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \\
&\geq \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*) - \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*) \\
&= L(\mathbf{y}, \boldsymbol{\alpha}^y, \boldsymbol{\lambda}^*) - L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \\
&\geq d_{\mathbf{s}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)'(\mathbf{y} - \mathbf{s}) + d_{\boldsymbol{\alpha}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)'(\boldsymbol{\alpha}^y - \boldsymbol{\alpha}^*) \\
&\geq d_{\mathbf{s}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)'(\mathbf{y} - \mathbf{s}),
\end{aligned}$$

where $\boldsymbol{\alpha}^y \in \arg \inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*)$, $(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \in Z(\mathbf{s})$ and vector $\begin{pmatrix} d_{\mathbf{s}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \\ d_{\boldsymbol{\alpha}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) \end{pmatrix}$ is the subgradient of function $L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*)$ at $(\mathbf{s}, \boldsymbol{\alpha}^*)$, and $d_{\boldsymbol{\alpha}}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$.

The second inequality holds as $L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*)$ is jointly convex in $(\mathbf{s}, \boldsymbol{\alpha})$. The last inequality holds since $\boldsymbol{\alpha}^*$ is the optimal solution of problem $\inf_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^*)$.

□

To calculate the subgradient of $f(\mathbf{s})$, Proposition 3.4 suggests we could equivalently calculate the subgradient of $L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$. Given $(\mathbf{x}, \mathbf{s}) \in \mathcal{X}$, after solving Problem (3.9), we separate the set \mathcal{N}_D into two sets \mathcal{N}_{D1} and \mathcal{N}_{D2} , such that $\mathcal{N}_{D1} = \{i \in \mathcal{N}_D \mid \alpha_i^* > 0\}$ and $\mathcal{N}_{D2} = \{i \in \mathcal{N}_D \mid \alpha_i^* = 0\}$.

Proposition 3.5. The subgradient of $f(\mathbf{s})$ with respect to s_a^i for all $i \in \mathcal{N}_D$, $a \in \mathcal{A}$ can be calculated as

$$d_{s_a^i}^f(\mathbf{s}) = \begin{cases} -\frac{d_{s_a^i}^c(\alpha_i^*, \mathbf{s}^i)}{d_{\alpha_i}^c(\alpha_i^*, \mathbf{s}^i)} d_{\alpha_i}^{\varphi}(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}}), & \text{when } i \in \mathcal{N}_{D1}, a \in \mathcal{A}, \\ 0, & \text{when } i \in \mathcal{N}_{D2}, a \in \mathcal{A}, \end{cases} \quad (3.10)$$

where $d_{s_a^i}^c(\alpha_i^*, \mathbf{s}^i)$ and $d_{\alpha_i}^c(\alpha_i^*, \mathbf{s}^i)$ is the subgradient of $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ with respect to s_a^i and α_i at point $(\alpha_i^*, \mathbf{s}^i)$, and $d_{\alpha_i}^\varphi(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}})$ is the subgradient of $\varphi(\mathbf{0}, (\alpha_i)_{i \in \mathcal{N}_{D1}})$ at point $(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}})$.

Proof.

We first study set \mathcal{N}_{D2} . Since $\varphi(\boldsymbol{\alpha})$ is non-decreasing in $\boldsymbol{\alpha} \geq \mathbf{0}$, for any $i \in \mathcal{N}_{D2}$, i.e., $\alpha_i^* = 0$, we have

$$f((y^i, \mathbf{s}^{-i})) - f((s^i, \mathbf{s}^{-i})) = f((y^i, \mathbf{s}^{-i})) - \varphi(\mathbf{0}, (\alpha_i)_{i \in \mathcal{N}_{D1}}) \geq 0,$$

consequently, for any $i \in \mathcal{N}_{D2}$, $a \in \mathcal{A}$,

$$d_{s_a^i}^f(\mathbf{s}) = 0.$$

We next study the set \mathcal{N}_{D1} . Since when $i \in \mathcal{N}_{D2}$, $\alpha_i^* = 0$, with the monotonicity property of function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$, we have $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) \leq \tau_i$ for any $\alpha_i \geq 0$. Hence, for given $\mathbf{s} \in \mathcal{S}$, Problem (3.9) can be equivalently formulated as

$$\begin{aligned} f(\mathbf{s}) &= \inf \quad \varphi(\mathbf{0}, (\alpha_i)_{i \in \mathcal{N}_{D1}}) \\ \text{s.t.} \quad & C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) \leq \tau_i, \quad i \in \mathcal{N}_{D1}, \\ & \alpha_i \geq 0, \quad i \in \mathcal{N}_{D1}. \end{aligned}$$

We then calculate the subgradient by the KKT condition. Note that for $i \in \mathcal{N}_{D1}$, $a \in \mathcal{A}$,

$$\frac{\partial}{\partial s_a^i} L(\mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \lambda_i d_{s_a^i}^c(\alpha_i^*, \mathbf{s}^i),$$

so we focus on the calculation of $\boldsymbol{\lambda}^*$. Since the strong duality holds, the

KKT conditions are both necessary and sufficient to characterize the optimal solutions. Therefore, according to the generalized KKT Theorem, $\boldsymbol{\alpha}^*$ is primal optimal if and only if

$$\begin{cases} d_{\alpha_i}^{\varphi}(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}}) + \lambda_i^* d_{\alpha_i}^c(\alpha_i^*, \mathbf{s}^i) = 0, & i \in \mathcal{N}_{D1}, \\ \lambda_i^* (C_{\alpha_i^*}(\tilde{\mathbf{c}}' \mathbf{s}^i) - \tau_i) = 0, & i \in \mathcal{N}_{D1}, \\ C_{\alpha_i^*}(\tilde{\mathbf{c}}' \mathbf{s}^i) - \tau_i \leq 0, & i \in \mathcal{N}_{D1}. \end{cases}$$

For any $i \in \mathcal{N}_{D1}$, i.e., $\alpha_i^* > 0$, then

$$C_{\alpha_i^*}(\tilde{\mathbf{c}}' \mathbf{s}^i) = \tau_i, \quad \text{and} \quad \lambda_i^* = -\frac{d_{\alpha_i}^{\varphi}(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}})}{d_{\alpha_i}^c(\alpha_i^*, \mathbf{s}^i)}.$$

After obtaining $\boldsymbol{\alpha}^*$ and $\boldsymbol{\lambda}^*$, we can calculate the subgradient

$$d_{s_a^i}^f(\mathbf{s}) = d_{s_a^i}^L(\mathbf{s}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) = \lambda_i^* d_{s_a^i}^c(\alpha_i^*, \mathbf{s}^i) = -\frac{d_{s_a^i}^c(\alpha_i^*, \mathbf{s}^i)}{d_{\alpha_i}^c(\alpha_i^*, \mathbf{s}^i)} d_{\alpha_i}^{\varphi}(\mathbf{0}, (\alpha_i^*)_{i \in \mathcal{N}_{D1}}). \quad \square$$

We have shown how to calculate $f(\mathbf{s})$ and its subgradient. Since $f(\mathbf{s})$ is a convex function, we next approximate it with a piece-wise linear function, and use Benders decomposition algorithm to solve problem $\inf_{(\mathbf{x}, \mathbf{s}) \in \mathcal{X}} f(\mathbf{s})$.

Proposition 3.6. For any $(\mathbf{v}, \mathbf{y}) \in \mathcal{S}$, we have

$$f(\mathbf{y}) = \sup_{(\mathbf{x}, \mathbf{s}) \in \mathcal{X}} \{f(\mathbf{s}) + d_{\mathbf{s}}^f(\mathbf{s})'(\mathbf{y} - \mathbf{s})\}, \quad (3.11)$$

where $d_{\mathbf{s}}^f(\mathbf{s})$ is the vector of subgradient of $f(\mathbf{s})$ with respect to \mathbf{s} .

Proof.

According to Proposition 3.4, we have

$$f(\mathbf{y}) \geq f(\mathbf{s}) + d_{\mathbf{s}}^f(\mathbf{s})'(\mathbf{y} - \mathbf{s}), \quad \forall (\mathbf{x}, \mathbf{s}) \in \mathcal{X},$$

therefore,

$$f(\mathbf{y}) \geq \sup_{(\mathbf{x}, \mathbf{s}) \in \mathcal{X}} \{f(\mathbf{s}) + d_{\mathbf{s}}^f(\mathbf{s})'(\mathbf{y} - \mathbf{s})\}.$$

Since

$$f(\mathbf{y}) = f(\mathbf{y}) + d_{\mathbf{y}}^f(\mathbf{y})'(\mathbf{y} - \mathbf{y}),$$

proposition is proved. \square

As the size of the set \mathcal{X} defined in Equation (3.8) is generally too large for us to directly tackle the problem. We use Benders decomposition method and summarize the entire algorithm as follows.

Algorithm RO

1. Select any $(\mathbf{x}, \mathbf{s}) \in \mathcal{X}$, and define the set $\mathcal{U} = \{(\mathbf{x}, \mathbf{s})\}$.
2. Given current solution (\mathbf{x}, \mathbf{s}) , solve the convex problem (3.9) and find the optimal $\boldsymbol{\alpha}$. Calculate the subgradient function $d_{\mathbf{s}}^f(\mathbf{s})$ according to Equation (3.10).
3. Solve the following subproblem

$$\begin{aligned} \inf \quad & w \\ \text{s.t.} \quad & w \geq f(\mathbf{s}) + d_{\mathbf{s}}^f(\mathbf{s})'(\mathbf{y} - \mathbf{s}), \quad \forall (\mathbf{x}, \mathbf{s}) \in \mathcal{U}, \\ & (\mathbf{v}, \mathbf{y}) \in \mathcal{X}, \end{aligned} \tag{3.12}$$

and denote the solution by $(\mathbf{v}^*, \mathbf{y}^*)$ and the optimal value w^* .

4. If $w^* = f(\mathbf{y}^*)$, then output the optimal value and optimal solution $(\mathbf{v}^*, \mathbf{y}^*)$, and stop.
5. If $w^* < f(\mathbf{y}^*)$, update, $\mathcal{U} = \mathcal{U} \cup \{(\mathbf{v}^*, \mathbf{y}^*)\}$, and go to step 2.

Proposition 3.7. Algorithm RO finds an optimal solution to Problem (3.11) in a finite number of steps.

Proof.

When the algorithm terminates, we have

$$w^* = f(\mathbf{y}^*) \geq f(\mathbf{s}) + d_{\mathbf{s}}^f(\mathbf{s})'(\mathbf{y}^* - \mathbf{s}), \quad \forall (\mathbf{x}, \mathbf{s}) \in \mathcal{X}.$$

Hence, $(\mathbf{v}^*, \mathbf{y}^*, w^*)$ is feasible for Problem (3.11). Since Problem (3.12) is a relaxation of Problem (3.11), $(\mathbf{v}^*, \mathbf{y}^*)$ is also optimal for Problem (3.11). Moreover, since \mathcal{U} at most includes all feasible solution, it is finite, and for each iteration, it increases by one element, the algorithm will terminate in a finite number of steps. \square

Now the only difficulty left is to calculate the subgradient of $C_{\alpha_i^*}(\tilde{\mathbf{c}}' \mathbf{s}^i)$, which undoubtedly depends on the information set of uncertain travel time. For notational simplicity, we drop the script i .

Calculation of $C_\alpha(\tilde{\mathbf{c}}' \mathbf{s})$ with different distributional uncertainty sets

Since $\tilde{\mathbf{c}} = (\tilde{c}_a)_{a \in \mathcal{A}}$ is a vector of independently distributed random variables, we have

$$C_\alpha(\tilde{\mathbf{c}}' \mathbf{s}) = \sum_{a \in \mathcal{A}} C_\alpha(\tilde{c}_a s_a) = \sum_{a \in \mathcal{A}} C_\alpha(\tilde{c}_a) s_a,$$

where the first equality holds since $C_\alpha(\cdot)$ is additive for independent random variables, while the second equality holds because s_a is a binary decision variable.

Known distribution

When the probability distribution of the random variable \tilde{c}_a is completely known, the function $C_\alpha(\tilde{c}_a)$ can be calculated through the moment generating functions. For example, if \tilde{c}_a follows a normal distribution $N(\mu_a, \sigma_a)$, its certainty equivalent is

$$C_\alpha(\tilde{c}_a s_a) = \alpha \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{c}_a s_a}{\alpha} \right) \right) = \alpha \ln \left(\exp \left(\frac{\mu_a s_a}{\alpha} + \frac{\sigma_a^2 s_a^2}{2\alpha^2} \right) \right) = \mu_a s_a + \frac{\sigma_a^2 s_a^2}{2\alpha},$$

and the subgradient can be calculated sequentially as

$$\begin{aligned} d_{s_a}^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial s_a} C_\alpha(\tilde{\mathbf{c}}' \mathbf{s}) = \frac{\partial}{\partial s_a} C_\alpha(\tilde{c}_a s_a) = \mu_a + \frac{\sigma_a^2}{\alpha} s_a, \\ d_\alpha^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial \alpha} C_\alpha(\tilde{\mathbf{c}}' \mathbf{s}) = \sum_{a \in \mathcal{A}} \frac{\partial}{\partial \alpha} C_\alpha(\tilde{c}_a s_a) = - \sum_{a \in \mathcal{A}} \frac{\sigma_a^2}{2\alpha^2} s_a^2. \end{aligned}$$

Discrete distribution with known samples

Suppose that we know the random variable \tilde{c}_a can only take the discrete values $\tilde{c}_a \in \{c_{a1}, \dots, c_{aK_a}\}$ and we may have the moment information on \tilde{c}_a

as follows.

$$\mathbb{F}_a = \left\{ \mathbb{P} \left| \mathbb{E}_{\mathbb{P}}(\mathbf{g}(\tilde{c}_a)) \in [\underline{\boldsymbol{\eta}}_a, \bar{\boldsymbol{\eta}}_a], \mathbb{P}(\tilde{c}_a \in \{c_{a1}, \dots, c_{aK_a}\}) = 1 \right. \right\},$$

where function $\mathbf{g}(\tilde{c}_a) = (g_l(\tilde{c}_a))_{l \in \mathcal{L}}$, and $g_l(\tilde{c}_a)$ can be any power of the random variable \tilde{c}_a , i.e., $g_l(\tilde{c}_a) = \tilde{c}_a^m, m \in \mathcal{Z}$. The certainty equivalent $C_\alpha(\tilde{c}_a) = \alpha \ln \sup_{\mathbb{P} \in \mathbb{F}_a} \mathbb{E}_{\mathbb{P}}(\exp(\tilde{c}_a/\alpha)) = \alpha \ln \mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a/\alpha))$, where the probability distribution \mathbb{Q}_a is the optimal solution of the following linear optimization problem, i.e., $\mathbb{Q}_a \in \arg \sup_{\mathbb{P} \in \mathbb{F}_a} \mathbb{E}_{\mathbb{P}}(\exp(\tilde{c}_a/\alpha))$.

$$\begin{aligned} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{c}_a}{\alpha} \right) \right) &= \sup \sum_{k=1}^{K_a} p_{ak} \exp \left(\frac{z_{ak}}{\alpha} \right) \\ \text{s.t.} \quad &\sum_{k=1}^{K_a} p_{ak} \mathbf{g}(z_{ak}) \leq \bar{\boldsymbol{\eta}}_a, \\ &\sum_{k=1}^{K_a} p_{ak} \mathbf{g}(z_{ak}) \geq \underline{\boldsymbol{\eta}}_a, \\ &\sum_{k=1}^{K_a} p_{ak} = 1, \\ &p_{ak} \geq 0, \quad k = 1, \dots, K_a. \end{aligned}$$

Hence, we could calculate the subgradient as

$$\begin{aligned} d_{s_a}^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial s_a} C_\alpha(\tilde{\mathbf{z}}' \mathbf{s}) = \frac{\partial}{\partial s_a} C_\alpha(\tilde{c}_a s_a) = \frac{\partial}{\partial s_a} \{ \alpha \ln \mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a s_a/\alpha)) \} \\ &= \frac{\mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a s_a/\alpha) \tilde{c}_a)}{\mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a s_a/\alpha))}, \\ d_\alpha^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial \alpha} C_\alpha(\tilde{\mathbf{z}}' \mathbf{s}) = \sum_{a \in \mathcal{A}} \frac{\partial}{\partial \alpha} C_\alpha(\tilde{c}_a s_a) = \sum_{a \in \mathcal{A}} \frac{\partial}{\partial \alpha} \{ \alpha \ln \mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a/\alpha)) \} s_a \\ &= \sum_{a \in \mathcal{A}} \left(\ln \mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a/\alpha)) - \frac{\mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a/\alpha) \tilde{c}_a)}{\alpha \mathbb{E}_{\mathbb{Q}_a}(\exp(\tilde{c}_a/\alpha))} \right) s_a \end{aligned}$$

Continuous distribution with certain descriptive statistics

When the random variable \tilde{c}_a is a continuous random variable, and the uncertainty set

$$\mathbb{F}_a = \left\{ \mathbb{P} \mid \mathbb{E}_{\mathbb{P}}(\tilde{c}_a) \in [\underline{\mu}_a, \bar{\mu}_a], \mathbb{P}(\tilde{c}_a \in [\underline{c}_a, \bar{c}_a]) = 1 \right\}, \quad (3.13)$$

where $[\underline{c}_a, \bar{c}_a]$ is bounded support.

Lemma 3.2. If the distributional uncertainty set of random variable \tilde{c}_a is given as Equation (3.13), then

$$\begin{aligned} C_\alpha(\tilde{c}_a) &= \sup_{\mathbb{P} \in \mathbb{F}} \alpha \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{c}_a}{\alpha} \right) \right) \\ &= \begin{cases} \alpha \ln \left(g(\tilde{c}_a) \exp \left(\frac{\underline{c}_a}{\alpha} \right) + h(\tilde{c}_a) \exp \left(\frac{\bar{c}_a}{\alpha} \right) \right), & \text{when } \alpha > 0, \\ \bar{c}_a, & \text{when } \alpha = 0. \end{cases}, \end{aligned}$$

where $g(\tilde{c}_a) = \frac{\bar{c}_a - \bar{\mu}_a}{\bar{c}_a - \underline{c}_a}$ and $h(\tilde{c}_a) = \frac{\bar{\mu}_a - \underline{c}_a}{\bar{c}_a - \underline{c}_a}$.

Proof. Please refer to Proposition 2.2 in Chapter 2.

Immediately, as the function $C_\alpha(\tilde{\mathbf{c}}' \mathbf{s})$ is differentiable, we calculate its

gradient with respect to s_a as

$$\begin{aligned}
d_{s_a}^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial s_a} C_\alpha(\tilde{\mathbf{z}}' \mathbf{s}) \\
&= \frac{\partial}{\partial s_a} C_\alpha(\tilde{c}_a s_a) \\
&= \frac{\partial}{\partial s_a} \{ \alpha \ln (g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha)) \} \\
&= \frac{g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) \underline{c}_a + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha) \bar{c}_a}{g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha)}.
\end{aligned}$$

When $s_a = 0$, we have $\left. \frac{\partial C_\alpha(\tilde{\mathbf{z}}' \mathbf{s})}{\partial s_a} \right|_{s_a=0} = \bar{\mu}_a$. Meanwhile, the gradient of $C_\alpha(\tilde{\mathbf{z}}' \mathbf{s})$ with respect to α is

$$\begin{aligned}
d_\alpha^c(\alpha, \mathbf{s}) &= \frac{\partial}{\partial \alpha} C_\alpha(\tilde{\mathbf{z}}' \mathbf{s}) \\
&= \sum_{a \in \mathcal{A}} \frac{\partial}{\partial \alpha} C_\alpha(\tilde{c}_a s_a) \\
&= \sum_{a \in \mathcal{A}} \left(\frac{\ln (g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha))}{\frac{g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) \underline{c}_a + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha) \bar{c}_a}{g(\tilde{c}_a) \exp(\underline{c}_a s_a / \alpha) + h(\tilde{c}_a) \exp(\bar{c}_a s_a / \alpha)}} \frac{s_a}{\alpha} \right).
\end{aligned}$$

3.4 Computational Study

In this section, we conduct computational studies intending to address two concerns. First, whether this newly proposed lateness index model could provide us with a reasonable policy under uncertainty. Second, as the deterministic version of the general routing optimization problems is already hard to solve, whether this lateness index model is practically solvable. The program is coded in python and run on a Intel Core i7 PC with a 3.40 GHz CPU by calling CPLEX 12 as ILP solver.

3.4.1 Stochastic shortest path problem with deadline

We carry out the first experiment to make a comparative study on the validity of the lateness index as a performance measure. For a randomly generated network, we solve a shortest path problem with deadline under uncertainty, in which $\mathcal{N}_D = \{n\}$ and $\mathcal{N}_R = \{1, n\}$. We investigate several classical selection criteria to find optimal paths, and then use out-of-sample simulation to compare the performances of these paths. We summarize four selection criteria which appeared in the literature.

Minimize average travel time

For a network with uncertain travel time, the simplest way to find a path is by minimizing the average travel time, which can be formulated as a deterministic shortest path problem.

$$\min_{\mathbf{x} \in \mathcal{X}_{SP}} \boldsymbol{\mu}'\mathbf{x},$$

where \mathcal{X}_{SP} is the feasible set for the shortest path problem defined in Problem (3.2). This problem is polynomially solvable, but the optimal path does not depend on the deadline.

Maximize arrival probability

The second selection criterion is to find a path that gives the largest probability to arrive on time, which is formulated as follows:

$$\max_{\mathbf{x} \in \mathcal{X}_{SP}} \mathbb{P}(\tilde{\mathbf{c}}'\mathbf{x} \leq \tau).$$

Since the problem is generally intractable (Khachiyan 1989), we adopt a sampling average approximation method to solve it. Assuming the sample size is K , then we solve

$$\begin{aligned} \max \quad & \frac{1}{K} \sum_{k=1}^K I_k \\ \text{s.t.} \quad & \mathbf{x}' \mathbf{c}^k \leq M(1 - I_k) + \tau, \quad k = 1, \dots, K, \\ & I_k \in \{0, 1\}, \quad k = 1, \dots, K, \\ & \mathbf{x} \in \mathcal{X}_{SP}, \end{aligned}$$

where M is a big number.

Maximize punctuality ratio

The third selection criterion is to maximize the punctuality ratio, which is defined as

$$\max_{\mathbf{x} \in \mathcal{X}_{SP}} \frac{\tau - \boldsymbol{\mu}' \mathbf{x}}{\sigma(\tilde{\mathbf{c}}' \mathbf{x})}$$

where $\sigma(\cdot)$ represents the standard deviation. The idea is to find a path that can give a shorter and less uncertain travel time. When the travel time on each arc is independently normally distributed, maximizing the arrival probability is in fact equivalent to maximizing the punctuality ratio, since

$$\mathbb{P}(\tilde{\mathbf{c}}' \mathbf{x} \leq \tau) = \mathbb{P}\left(\frac{\tilde{\mathbf{c}}' \mathbf{x} - \boldsymbol{\mu}' \mathbf{x}}{\sigma(\tilde{\mathbf{c}}' \mathbf{x})} \leq \frac{\tau - \boldsymbol{\mu}' \mathbf{x}}{\sigma(\tilde{\mathbf{c}}' \mathbf{x})}\right) = \Phi\left(\frac{\tau - \boldsymbol{\mu}' \mathbf{x}}{\sigma(\tilde{\mathbf{c}}' \mathbf{x})}\right),$$

in which, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable $N(0, 1)$. As this problem is not a convex problem, we use the algorithm proposed by Nikolova et al. (2006) to solve it.

Maximize budget of uncertainty

By introducing a parameter Γ , named budget of uncertainty, Bertsimas and Sim (2004) successfully provide a new robust formulation to flexibly adjust the level of conservatism while withstanding the parameter uncertainty. This formulation can also be applied readily to discrete optimization problem (Bertsimas and Sim 2003). Hence, the robust shortest path problem is formulated as

$$\min_{\mathbf{x} \in \mathcal{X}_{SP}} \max_{\tilde{\mathbf{c}} \in \mathcal{W}_\Gamma} \tilde{\mathbf{c}}' \mathbf{x}$$

in which, $\mathcal{W}_\Gamma = \left\{ \boldsymbol{\mu} + \mathbf{s} \mid \mathbf{0} \leq \mathbf{s} \leq \bar{\mathbf{c}} - \boldsymbol{\mu}, \sum_{a \in \mathcal{A}} \frac{s_a}{\bar{c}_a - \mu_a} \leq \Gamma \right\}$, for all $\Gamma \geq 0$. $\Gamma = 0$ represents the nominal case. Given the deadline τ , we could transform the problem to find a path that could return the maximal Γ while respecting the deadline. The formulation is given as

$$\begin{aligned} \Gamma^* = \max \quad & \Gamma \\ \text{s.t.} \quad & \max_{\tilde{\mathbf{c}} \in \mathcal{W}_\Gamma} \tilde{\mathbf{c}}' \mathbf{x} \leq \tau, \\ & \mathbf{x} \in \mathcal{X}_{SP}. \end{aligned}$$

Following the calculation procedure suggested by Bertsimas and Sim (2003), we first define $0 = \bar{c}_{|\mathcal{A}|+1} - \mu_{|\mathcal{A}|+1} \leq \bar{c}_{|\mathcal{A}|} - \mu_{|\mathcal{A}|} \leq \dots \leq \bar{c}_1 - \mu_1 \leq \infty$, and the above problem is equivalent to

$$\begin{aligned} \Gamma^* = \max \quad & \Gamma \\ \text{s.t.} \quad & \min_{l=1, \dots, |\mathcal{A}|+1} \{\Gamma(\bar{c}_l - \mu_l) + Z_l\} \leq \tau, \end{aligned}$$

where $Z_l = \min_{\mathbf{x} \in \mathcal{X}_{SP}} \left(\boldsymbol{\mu}' \mathbf{x} + \sum_{j=1}^l ((\bar{c}_j - \mu_j) - (\bar{c}_l - \mu_l)) x_j \right)$, for all $l = 1, \dots, |\mathcal{A}| + 1$. Calculating Z_l is a classical shortest path problem, and

$$\Gamma^* = \max_{l=1, \dots, |\mathcal{A}|+1} \frac{\tau - Z_l}{\bar{c}_l - \mu_l}.$$

Since some selection criteria introduced above could not handle distributional ambiguity, to make a fair comparison, we assume that the probability distribution of the uncertain travel time is perfectly known, and each follows a two-point distribution. For each instance, we randomly generate a directed network with 300 nodes, and with a number of arcs around 1,500 on a 1×1 square, where node $(0, 0)$ is the origin node, and node $(1, 1)$ is the destination node. Using some screening procedure, we guarantee that there exists at least one path going from the origin to the destination. The mean travel time on each arc is given by the Euclidean distance between the two nodes, and the corresponding upper and lower bounds are randomly generated. In order to ensure the problem feasibility, we artificially set the deadline for the destination node as $\tau = \eta \min_{\mathbf{x} \in \mathcal{X}_{SP}} \boldsymbol{\mu}' \mathbf{x} + (1 - \eta) \min_{\mathbf{x} \in \mathcal{X}_{SP}} \bar{\mathbf{c}}' \mathbf{x}$. In this example, $\eta = 0.8$. Of course, if the deadline is exogenous, we could check the feasibility for this deadline by computing the shortest average travel time. We calculate the optimal paths under the five selection criteria, and use out-of-sample simulation to analyze the performances. Table 3.1 summarizes the average performances among 50 instances. For notational clarity, we only show the performance ratio, which is the original performance divided by the performance of minimizing the lateness index. Therefore, all the performance ratios for the lateness index model are one, and a ratio greater than

one indicates a better performance for the lateness index model.

Selection criteria	Performance measures							
	Mean	LP ¹	STD ²	EL ³	CEL ⁴	VaR @95% ⁵	VaR @99%	CPU time
Minimize average travel time	0.985	1.124	1.206	1.397	1.228	1.006	1.014	0.027
Maximize arrival probability	1.006	1.549	1.116	1.873	1.202	1.017	1.021	926.14
Maximize punctuality ratio	0.986	1.033	1.150	1.201	1.157	1.002	1.008	1.255
Maximize budget of uncertainty	0.990	1.125	1.155	1.325	1.160	1.005	1.010	44.872
Minimize lateness index	1	1	1	1	1	1	1	1

¹ **LP** refers to lateness probability;

² **STD** refers to standard deviation;

³ **EL** refers to expected lateness, $EL = \mathbb{E}_{\mathbb{P}} \left((\tilde{c}'\mathbf{x}^* - \tau)^+ \right)$;

⁴ **CEL** refers to conditional expected lateness, $CEL = \mathbb{E}_{\mathbb{P}} \left((\tilde{c}'\mathbf{x}^* - \tau)^+ | \tilde{c}'\mathbf{x}^* > \tau \right)$;

⁵ **VaR@ γ** refers to value-at-risk, $VaR@_{\gamma} = \inf\{\nu \in \mathcal{R} | \mathbb{P}(\tilde{c}'\mathbf{x}^* - \nu) \leq 1 - \gamma\}$.

Tab. 3.1: Performances of various selection criteria for stochastic shortest path problem with deadline.

In terms of the mean arrival time measure, we observe that the lateness index model gives a larger mean than the other selection criteria, but it provides a path with significantly lower standard deviation, expected lateness and conditional expected lateness. Hence, by slightly increasing the expected travel time, the lateness index model can better mitigate the risk of tardiness. In addition, since solving stochastic shortest path problem under the lateness index only requires solving a small sequence of deterministic shortest path problems, the CPU time is relatively short compared to the other methods, except for the selection criterion of minimizing the average travel time. For maximizing the arrival probability, since we use a sampling average approximation, the calculation takes quite a long time even with a small sample size ($K = 80$), and the performance is worse even in terms of the lateness

probability.

By varying the coefficient η , we also alter the deadline at the destination node, and summarize the performance ratio of each selection criterion in Figure 3.2. We exclude the selection criterion of maximizing the arrival

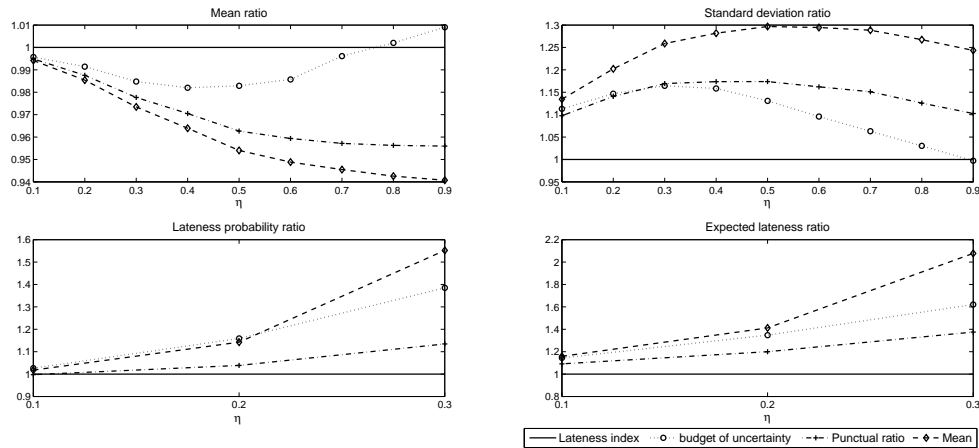


Fig. 3.2: Performance comparison for stochastic shortest path problem when deadline varies.

probability, as a small sample size resulted in inconsistent solutions for comparison. Among the remaining four selection criteria, the lateness index model outperforms the others, especially in terms of standard deviation. It is worthwhile to point out that in terms of the lateness probability ratio and expected lateness ratio, η is only used with values 0.1, 0.2, 0.3, since when η is greater than 0.3, the lateness probability and expected lateness under lateness index solution are very close to 0. Similar conclusion could be derived when the travel times are uniformly distributed.

Since the shortest path problem with deadline is a special case of our more general routing problem, we could also test the algorithm RO of Section 3.3 on it, though it is not necessarily polynomial time. We randomly

generate 50 instances, and compare the statistics on CPU time of these two algorithms for a network with 300 nodes and 1,500 arcs. Table 3.2 suggests the calculation time of RO algorithm is longer than the bisection method, but is still attractive. It provides an encouraging result for the employment of RO algorithm in the general routing optimization problem.

Statistics	Bisection	RO algorithm	
	CPU time (sec)	CPU time (sec)	Number of iterations
Average	0.396	1.211	3.32
Maximum	0.512	4.951	14
Minimum	0.165	0.356	1
Standard deviation	0.059	1.093	3.01

Tab. 3.2: Statistics of CPU time of two algorithms for stochastic shortest path problem with deadline.

3.4.2 Solution procedure illustration

We next consider an example on a simple network with 5 nodes and 12 arcs shown in Figure 3.3, and provide a detailed description of the results obtained using this new performance measure, as well as the computational characteristics of our proposed solution methodologies. For simplicity, we use the function $\varphi(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{N}_D} \alpha_i$. The travel time information is specified in Table 3.3. The travel time uncertainties along the arcs vary according to the parameter β . Note that arc 6 is distinct from the rest. Our aim is to find a path from node 1 to node 5, that visits each node exactly once, and meets the specific deadline requirements $\tau_3 = \tau_5 = 14.5$. Correspondingly, $\mathcal{N}_D = \{3, 5\}$ and $\mathcal{N}_R = \mathcal{N}$. In this simple network, if we ignore the deadline constraints, all the feasible paths can be easily enumerated as in Table 3.4.

By substituting the uncertain travel times with their mean values, paths

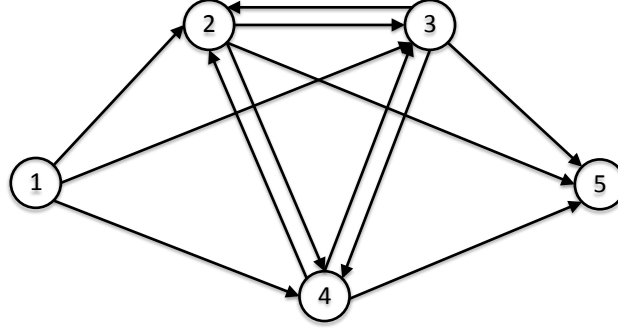


Fig. 3.3: An illustrative example on a five-nodes network.

Index	Arc	Lower bound	Mean	Upper bound
1	(1, 2)	$2(1 - \beta)$	2	$2(1 + \beta)$
2	(1, 3)	$2(1 - \beta)$	2	$2(1 + \beta)$
3	(1, 4)	$2(1 - \beta)$	2	$2(1 + \beta)$
4	(2, 3)	$3(1 - \beta)$	3	$3(1 + \beta)$
5	(2, 4)	$7(1 - \beta)$	7	$7(1 + \beta)$
6	(2, 5)	$4(1 - 1.5\beta)$	4	$4(1 + 1.5\beta)$
7	(3, 2)	$2(1 - \beta)$	2	$2(1 + \beta)$
8	(3, 4)	$2(1 - \beta)$	2	$2(1 + \beta)$
9	(3, 5)	$1 - \beta$	1	$1 + \beta$
10	(4, 2)	$6(1 - \beta)$	6	$6(1 + \beta)$
11	(4, 3)	$4(1 - \beta)$	4	$4(1 + \beta)$
12	(4, 5)	$7(1 - \beta)$	7	$7(1 + \beta)$

Tab. 3.3: Travel time information corresponding to Figure 3.3.

Index	Path
1	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$
2	$1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 5$
3	$1 \rightarrow 3 \rightarrow 2 \rightarrow 4 \rightarrow 5$
4	$1 \rightarrow 3 \rightarrow 4 \rightarrow 2 \rightarrow 5$
5	$1 \rightarrow 4 \rightarrow 2 \rightarrow 3 \rightarrow 5$
6	$1 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 5$

Tab. 3.4: All feasible paths for the illustrative example without the deadline requirements.

1, 2, 4, 5, and 6 are all feasible paths that could meet the deadline requirements. Instead, when the travel times take their worst values, we could see that, if $\beta = 0.1$, both paths 5 and 6 would satisfy the deadline requirements. If $\beta = 0.2$, only path 5 is feasible, and no path is feasible when $\beta = 0.3, 0.4$.

The result indeed illustrates that the worst case approach may be overly conservative. With the lateness index, when $\beta = 0.1, 0.2$, the selection decisions are the same as the worst-case method, and the associated objective value is 0. When $\beta = 0.3, 0.4$, the calculation procedure is listed in Table 3.5.

β	Iteration	optimal solution (path number)	objective value w^*	optimal alpha (α_2^*, α_5^*)	summation of optimal alpha $f(\mathbf{y}^*)$
0.3	0	5		(0, 0.448)	0.448
	1	6	-1.024	(0, 0.710)	0.710
	2	1	0.191	(0, 5.844)	5.844
	3	2	0.360	(1.785, 6.209)	7.994
	4	5	0.448	(0, 0.448)	0.448
0.4	0	5		(0.439, 1.137)	1.576
	1	6	-5.650	(0, 1.551)	1.551
	2	1	-0.459	(0, 10.464)	10.464
	3	2	0.678	(3.397, 11.109)	14.506
	4	6	1.551	(0, 1.551)	1.551

Tab. 3.5: Calculation procedure of lateness index model with different β .

Several interesting results can be observed from this computational study. With the increase of β , travel time becomes more uncertain, and the optimal path changes from path 5 to path 6. Observing that node 3 has the same deadline as the destination node 5, intuitively, travelers may expect that as long as node 3 could be reached before the destination node, the actual time of arrival would be inconsequential. However, the obtained result is not so trivial. When $\beta = 0.3$, as shown in Table 3.6, even the worst-case arrival time at node 3 through both path 5 and path 6 can meet the presumed deadline. Therefore, with the punctuality satisficing property of the lateness index measure, the selection decision only depends on whether the arrival time meets the deadline at node 5, and path 5 is calculated as optimal. Similarly, when $\beta = 0.4$, the value of lateness index of path 6 only depends

on the performance at node 5. Nonetheless, when travelling through path 5, the lateness index should account for both node 3 and node 5. Accordingly, path 6 becomes the optimal path.

β	Node	Path 5			Path 6		
		Lower bound	Mean	Upper bound	Lower bound	Mean	Upper bound
0.3	3	7.7	11	14.3	4.2	6	7.8
	5	8.4	12	15.6	7.8	12	16.2
0.4	3	6.6	11	15.4	3.6	6	8.4
	5	7.2	12	16.8	6.4	12	17.6

Tab. 3.6: Arrival time comparison between paths 5 and 6.

3.4.3 General routing optimization problem

The formulation of the routing optimization problem implies that the computation time greatly depends on the network structure, $|\mathcal{N}|$, $|\mathcal{A}|$, and the properties of sets \mathcal{N}_R and \mathcal{N}_D . Additionally, the deadline setting will also tremendously affect the size of the feasible set, and so, the number of iterations. In this part, we mainly focus on the influence of the number of nodes and arcs on the computation time and the number of iterations, and show the results in Table 3.7 and Table 3.8 respectively. We randomly generate the arcs for a network while ensuring the existence of a Hamiltonian path, and the information of uncertain travel times includes means and supports. To set reasonable deadlines, we first derive a feasible path that minimizes the total average travel time. With this path, we calculate the corresponding mean arrival time and worst-case arrival time for each node with a deadline requirement, and set the deadline in between. For each case, we randomly generate 20 instances, and present the average values.

$(\mathcal{N} , \mathcal{A})$		$\mathcal{N}_R = \mathcal{N}$ $\mathcal{N}_D = \mathcal{N} \setminus \{1\}$				$\mathcal{N}_R = \mathcal{N}$ $\mathcal{N}_D = \{\lfloor n/2 \rfloor, n\}$				$\mathcal{N}_R = \mathcal{N}_D \cup \{1\}$ $\mathcal{N}_D = \{\lfloor n/2 \rfloor, n\}$			
		Avg	Max	Min	STD	Avg	Max	Min	STD	Avg	Max	Min	STD
(10, 30)	LDR	1.1	2.7	0.2	0.7	0.7	1.2	0.2	0.4	0.3	0.7	0.2	0.2
	MCF	0.8	2.0	0.2	0.5	0.5	1.0	0.2	0.3	0.2	0.4	0.1	0.1
(10, 50)	LDR	36.4	123	1.5	35.3	13.3	44.4	1.0	13.1	1.6	7.6	0.3	1.7
	MCF	6.06	17.7	0.7	5.6	1.9	5.5	0.3	1.4	0.4	1.2	0.2	0.3
(10, 70)	LDR	526	3477	9.20	797	214	1316	0.84	314	10.2	64.3	0.6	15.9
	MCF	21.7	135	0.7	31.8	5.8	26.0	0.4	6.5	0.6	1.2	0.3	0.3
(20, 60)	LDR	13.5	43.8	1.2	12.2	4.2	11.1	1.1	3.1	10.4	59.3	0.8	17.9
	MCF	8.6	28.6	1.2	7.6	3.0	8.1	1.2	1.9	1.6	6.8	0.4	1.7
(30, 90)	LDR	112	663	5.3	186	49.0	259	4.6	78.3	208	939	2.4	265
	MCF	55.8	310	5.5	69.3	24.0	96.6	3.4	28.2	6.2	23.7	1.8	6.2
(40, 120)	LDR	1645	7405	31.9	2572	346	1694	18.1	500	4241	13712	8.7	4750
	MCF	854	5002	21.3	1436	134	718	11.8	202	13.3	36.1	4.3	10.1

Tab. 3.7: CPU time (sec) on routing optimization problem with different settings.

$(\mathcal{N} , \mathcal{A})$	$\mathcal{N}_R = \mathcal{N}$ $\mathcal{N}_D = \mathcal{N} \setminus \{1\}$				$\mathcal{N}_R = \mathcal{N}$ $\mathcal{N}_D = \{\lfloor n/2 \rfloor, n\}$				$\mathcal{N}_R = \mathcal{N}_D \cup \{1\}$ $\mathcal{N}_D = \{\lfloor n/2 \rfloor, n\}$			
	Avg	Max	Min	STD	Avg	Max	Min	STD	Avg	Max	Min	STD
(10, 30)	4.4	10	1	2.5	3	6	1	1.7	1.5	5	1	1.0
(10, 50)	11.2	30	1	8.8	6.5	12	1	3.6	2.8	8	1	2.1
(10, 70)	10.9	47	1	12.1	6.1	18	1	5.1	2.8	8	1	1.9
(20, 60)	11.9	31	1	9.5	4.1	11	1	3.1	3.0	12	1	2.9
(30, 90)	17.1	43	2	11.5	7.9	27	1	7.4	2.6	9	1	2.0
(40, 120)	36.8	133	3	36.8	9.6	27	1	7.5	2	5	1	1.3

Tab. 3.8: Number of iterations on routing optimization problem with different settings.

Table 3.7 demonstrates that the RO algorithm could solve moderate-size problems within a reasonable time range, and the MCF formulation is more appealing computationally. While setting the time limit as 7200 seconds, with the MCF formulation, the RO algorithm can solve a network with 100 nodes, and 450 arcs for the case where $\mathcal{N}_R = \mathcal{N}_D \cup \{1\}, \mathcal{N}_D = \{\lfloor n/2 \rfloor, n\}$. Table 3.8 shows that on average, we only need a relatively small number of iterations. If more efficient algorithms could be implemented for solving the subproblem, the computation time could be remarkably improved.

3.5 Extension: correlations between uncertain travel times

All the models introduced above are based on a stochastic independence assumption between travel times on arcs. We now extend the model to the case in which travel times are correlated. To model the correlation relationships, instead of specifying the commonly used covariance matrix, we assume that the travel time on each arc is an affine function of independently distributed factors $\tilde{z}_1, \dots, \tilde{z}_K$, i.e.,

$$\tilde{c}_a = c_a^0 + \sum_{k=1}^K c_a^k \tilde{z}_k, \quad \forall a \in \mathcal{A},$$

in which the factor coefficients $c_a^0, c_a^1, \dots, c_a^K$ are known. These parameters can be estimated from a linear regression technique. Correspondingly,

$$\begin{aligned} C_\alpha(\tilde{\mathbf{c}}' \mathbf{x}) &= C_\alpha \left(\sum_{a \in \mathcal{A}} \left(c_a^0 + \sum_{k=1}^K c_a^k \tilde{z}_k \right) x_a \right) \\ &= C_\alpha \left(\mathbf{x}' \mathbf{c}^0 + \sum_{k=1}^K \mathbf{x}' \mathbf{c}^k \tilde{z}_k \right) \\ &= \mathbf{x}' \mathbf{c}^0 + \sum_{k=1}^K C_\alpha(\mathbf{x}' \mathbf{c}^k \tilde{z}_k). \end{aligned}$$

To solve the shortest path problem with deadline under such uncertainty, Problem (3.1) can be equivalently written as

$$\min_{\mathbf{x} \in \mathcal{X}_{SP}} \mathbf{x}' \mathbf{c}^0 + \sum_{k=1}^K C_\alpha(\mathbf{x}' \mathbf{c}^k \tilde{z}_k). \quad (3.14)$$

Different from our previous discussion on the stochastic shortest path problem when travel times are independently distributed, this case cannot be polynomially solvable. For any fixed $\alpha \geq 0$, the problem reduces to a convex integer optimization problem, in which Benders decomposition can be adopted to solve the problem.

For the general routing problem, the only difference from the model with stochastic independence assumption lies in the calculation of the function $C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i)$ and its subgradient. With the linear factor-based model and distributional uncertainty set \mathbb{F} , we have

$$C_{\alpha_i}(\tilde{\mathbf{c}}' \mathbf{s}^i) = \sup_{\mathbb{P} \in \mathbb{F}} \alpha_i \ln \mathbb{E}_{\mathbb{P}} \left(\exp \left(\frac{\tilde{\mathbf{c}}' \mathbf{s}^i}{\alpha_i} \right) \right) = (\mathbf{s}^i)' \mathbf{c}^0 + \sum_{k=1}^K C_{\alpha} \left((\mathbf{s}^i)' \mathbf{c}^k \tilde{z}_k \right).$$

Accordingly, we could calculate the subgradient function $d_{\mathbf{s}}^f(\mathbf{s})$, and then use Algorithm RO to solve the general routing problem when the travel times are correlated.

3.6 Conclusion

We study a vehicle routing problem with uncertain travel times. The aim is to find an optimal routing policy to meet the deadline requirements imposed on a subset of nodes in the network. We introduce a new performance measure called lateness index to evaluate how the uncertain arrival times meet the deadlines and propose an algorithm using Benders decomposition to solve the general problem.

In this chapter, we only consider a special case where only one vehicle is

available. The framework and performance measure could easily be extended to multiple capacitated vehicles, while also incorporating the uncertain service time. Besides, the framework could also be applied in the uncertain demand case. Since this chapter mainly focuses on the mathematics framework, we do not go to detail to discuss them. Interested readers could refer to Adulyasak and Jaillet (2014) for certain extensions. This is joint work with Patrick Jaillet and Melvyn Sim.

4. MITIGATING DELAYS AND UNFAIRNESS IN APPOINTMENT SYSTEMS

4.1 Introduction

In any service system, due to the uncertainty in service times, waiting times or delays experienced by the participants are inevitable. However, long waiting time that occurs in a scheduled appointment is an annoyance and leads to poor quality of service. We focus our study in the healthcare industry where the participants are patients and the physician. Decisions associated with the appointment systems include the sequencing of patients and the scheduling of their appointment times, where these patients are distinguished by their service time characteristics. The goal of this chapter is to design an appointment system that mitigates the unpleasantness experienced by the patients while waiting for the treatment and by the physician in having to work overtime. The model is applicable in outpatient clinics to design consultation slots and operating theatres to deliver an efficient and smooth schedule.

The study of appointment systems stems from the pioneering work of Bailey (1952). Before that, service providers typically allocate each patient

a slot with the same fixed time length. Bailey (1952) designs an appointment scheduling rule which assigns two patients at the first slot, followed by other patients' arrivals evenly spaced. This minor change effectively reduces physician's idle time by overcoming the problem of patients no-show or lateness without compromising on the patients' waiting time. Since then, many researchers have started to explore the optimal appointment system settings under various conditions. For comprehensive literature reviews, we refer readers to Cayirli and Veral (2003) and Gupta and Denton (2008), which highlight the current status and challenges in resolving appointment problems.

Patrick and Aubin (2013) mention that patient access decisions generally involve two-stage planning. The first stage is advance scheduling, which decides how many patients to assign within a fixed session, while the second stage named appointment scheduling allocates time slot for each patient. In this chapter, our appointment scheduling refers to the second stage, where the information about patients who need appointment is known, and all the decisions must be made prior to the commencement of a clinic session. Though the appointment for outpatient services is generally made in a dynamic fashion, this model serves as a reference table with designed time slots for different types of patients. When patients call in, service providers could pull up patients' archived information and schedule them into suitable slots. Hence, all of the following analysis concentrates on the static case only. Now, we begin with discussing several concerns related to appointment system design problems.

The first concern regards to characterizing patients' experience of wait-

ing, which is an integral aspect of service quality in a hospital environment. One commonly used service quality measure for describing this preference on uncertain waiting time is the expectation, which corresponds to the average delay experienced by the patient over potentially infinite number of visits under the same identical conditions. However, the expected waiting time criterion may not adequately distinguish patients' attitudes towards uncertain delays. From patients' perspectives, the unpleasantness on waiting process may not proportionally accord to the length of waiting time (see Camacho et al. 2006), and certain waiting time is considered acceptable among patients (see Cartwright and Windsor 1992; McCarthy et al. 2000). In the survey conducted by Hill and Joonas (2006), 86% respondents consider 30 minutes or less as an acceptable threshold. Huang (1994) empirically shows that, for patients arriving on the appointment time, they appear reasonably satisfied if they wait no more than an average of 37 minutes, and their patience may steeply decline when the service delay exceeds this threshold. From service providers' perspectives, their key performance indicator lies on the percentage of patients seen within certain time threshold, instead of total expected waiting time. For example, patients in UK can expect to be seen within 30 minutes of their given appointment time (National Health Service, UK). The Ministry of Health Malaysia has proposed one of the key performance indicators as "percentage of patients seen within 30 minutes of appointment time by the dental specialist in specialist clinics should not be less than 50%, provided the patient was not late" (Toh and Sern 2011). Following these empirical results, we could use a reasonable unpleasantness tolerance threshold to describe the patient's satisfaction on waiting processes, and take

the frequency of delays above this threshold as an alternative service quality measure. Nonetheless, several non-negligible drawbacks have hampered the wide application of this measure. One disadvantage lies in the intensity of delay, for its inability to distinguish waiting processes with the same frequency of surpassing the patient's tolerance threshold but with different length of delay. Moreover, the computational intractability of this probability measure also arises due to lack of convexity. Thus, we need to establish a new service quality measure which could in some extent reflect people's real attitudes towards delay process, in particular, could account for both the frequency and intensity of the delay over the threshold.

The optimization criterion for an appointment system involves multiple participants including patients and physicians. Currently, majority of studies take a weighted average of the combinations among patients' waiting time, the physician's idle time and overtime as an optimization criterion, and exploit different methods to solve. Three main streams are based on queueing theory (see Wang 1993; Wang 1999; Green and Savin 2008; Hassin and Mendel 2008), stochastic programming (see Robinson and Chen 2003; Denton and Gupta 2003), and robust optimization (see Mittal and Stiller 2011; Kong et al. 2013; Mak et al. 2013) frameworks. However, as the decisions are very sensitive to the prescribed weight for each participant, how to provide an accurate interpretation and estimation of these weights is a crucial issue (Mondschein and Weintraub 2003). Additionally, minimizing a weighted combination of expectations of patients' waiting time, physician's idle time and overtime fails to accommodate the fairness issue highlighted by Cayirli and Veral (2003). In layman terms, fairness regards to distinguish-

ing a strategy of keeping say 20 patients each waiting for 2 minutes and its counterpart of keeping only one of them waiting for 40 minutes (Klassen and Rohleder 1996). Cayirli and Veral (2003) have highlighted the phenomenon that current appointment system is unfair to the patient at the last position, as waiting time tends to progressively build up. The notion of “fairness” has been widely studied in economics literatures (see Young 1995; Sen and Foster 1997) and industrial applications, especially resource allocation problems (see Bertsimas et al. 2011 and references therein), but few papers focus on the appointment scheduling problems except Cox et al. (1985), Yang et al. (1998). For this reason, an effective appointment system should be able to guarantee the uniformity of qualities across multiple participants.

To cope with the difficulties of eliciting the exact probability distributions for patients’ consultation times, robust optimization techniques have also been applied in appointment problems (see Mittal and Stiller 2011; Kong et al. 2013; Mak et al. 2013). In these papers, the optimization criteria are based on a weighted sum of patients’ expected waiting times, physician’s idle time and overtime. Mittal and Stiller (2011) consider the scheduling problem where only the bound support of service time is provided. To minimize the sum of waiting time cost and idle time cost, they present a global balancing heuristic, and prove that it will deliver an optimal schedule under certain mild condition. Kong et al. (2013) assume lower bound, mean, and covariance of the service time are known, and formulate a robust min-max problem, which could be solved by a semidefinite programming relaxation. Mak et al. (2013) investigate the scheduling problem by assuming the knowledge of marginal moments of uncertain service time, and derive a computationally tractable

conic programming formulation.

In general, consultation times among different types of patients such as new and repeated one are not necessarily homogenous. Since the physician would be familiar with the medical history of repeated patients, their consultation times tend to be shorter than new ones. By exploiting the information of patients' classification, appointment systems would inevitably rely on the sequencing decisions on these various types of patients. Due to the difficulty of the problems, few papers investigate the sequencing and scheduling decisions simultaneously. Weiss (1990) is the first to examine this problem and provides analytical results for a two patients case with general service time distribution, however, the conclusions could not be simply extended to multiple patients case. Wang (1999) addresses the problem with a specific assumption that patients' service time follows exponential distribution with different rates, and infers that the optimal service sequence is in the descending order of service rates. Bosch and Dietz (2000, 2001) classify the patients into different categories according to their service times that follow different phase-type distributions. They approximately solve the scheduling problem by shifting the appointment time to incrementally improve the objective value for a given sequence, and then swap the sequence pairwise until it terminates. Denton et al. (2007) jointly formulate the sequencing and scheduling problem into a two-stage stochastic programming model, and suggest an interchange heuristic with the sampling average approximation technique. Gupta (2007) uses stochastic programming to model this problem and mainly highlights the complication of problem by investigating the case with two patients only.

To fully characterize all the above perspectives in appointment system design, especially, to mitigate the delay and unfairness in the appointment system, this chapter first proposes a new service quality measure named Delay Unpleasantness Measure (DUM) to demonstrate the dependency of individual participant's attitude towards his/her delay process based on their corresponding acceptable levels. The acceptable level is an exogenous factor, and varies according to patients' demographic profiles. For example, the tolerable threshold of elderly patients is much longer (Moschis and Bellinger 2003). Besides, as the consultation time for repeated patients is relatively short, in certain cases, they may deserve a shortened waiting process, which corresponds to a small threshold. We could use survey or interview methods to study patients' thresholds based on different medical departments, ages, frequency of visit etc. (see for instance McCarthy et al. 2000; Hill and Joonas 2006). Unlike the probability measure, DUM collectively accounts for the frequency and intensity of delay over a threshold. Secondly, we present the concept of lexicographic min-max fairness to tackle the fairness concern arising in appointment system design. We lexicographically minimize the worst DUM, the second worst DUM, and so on. Thirdly, by assuming patients' sequence is predetermined, we develop a scheduling model that can be adapted in the robust setting. Different from the conventional distributional uncertainty set, in which covariance matrix is used to capture the correlation among uncertain service times, we propose mean absolute deviation of summation over service times as the information that could help retain linearity of the model. Therefore, the optimal decisions are derived by solving a small sequence of linear optimization problems. Fourthly, this model could be ex-

tended to incorporate sequencing decisions when patients are heterogeneous.

The rest of the chapter is organized as follows. In Section 4.2, we show how a participant's behavior in delay process can be characterized by the DUM. In Section 4.3, we introduce the concept of lexicographic min-max fairness and propose the solution procedure under the DUM. In Section 4.4, we propose a scheduling model for appointment systems by assuming patients' sequence is fixed, and demonstrate how the resulting model can be solved. In Section 4.5, we extend our model to solve both sequencing and scheduling problems. In Section 4.6, we perform several computational studies with encouraging results on the DUM regarding the fairness concern. Finally, in Section 4.7, we provide conclusions and managerial insights.

4.2 Delay Unpleasantness Measure

In this section, we will motivate and introduce a new service quality measure to evaluate uncertain waiting time (service delay) of patients and overtime (off-work delay) of physicians. We start with defining Delay Unpleasantness Measure (DUM) for individual participant (patient or physician) in the appointment system. We assume that each participant has his/her own tolerance threshold τ on waiting time, and the real uncertain delay is represented by \tilde{w} . DUM takes into account of both the frequency and intensity of delay over the threshold and is defined as follows.

Definition 4.1. Given an uncertain delay $\tilde{w} \in \mathcal{L}$ and tolerance threshold $\tau \in \mathfrak{R}_+$, the Delay Unpleasantness Measure is a function $\rho_\tau : \mathcal{L} \rightarrow [0, 1]$

defined as

$$\rho_\tau(\tilde{w}) = \inf\{\alpha \geq 0 \mid \varphi_\alpha(\tilde{w}) \leq \tau\},$$

(or 1 if no such α exists), where

$$\varphi_\alpha(\tilde{w}) = \min_{\nu \in \mathfrak{R}} \left(\nu + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w} - \nu)^+) \right), \quad \alpha \in (0, 1].$$

This definition is similar to Shortfall aspiration level criterion in Chen and Sim (2009) and Definition 5 in Brown and Sim (2009) in the monetary context. Function $\varphi_\alpha(\tilde{w})$ is the worst-case Conditional Value-at-Risk (CVaR) (see Zhu and Fukushima 2009 and Natarajan et al. 2010) when we only have information that the true distribution \mathbb{P} lies in a distributional uncertainty set \mathbb{F} . CVaR (Rockafellar and Uryasev 2000) is a measure with specific focus on the tail distribution, and has become a major reference in the area of financial mathematics with its endearing properties. It is also shown to be the best convex conservative approximation of frequency of delay over the threshold (Nemirovski and Shapiro 2006). In hospital settings, Dehlendorff et al. (2010) use simulation models and suggest that CVaR is a reliable measure for the waiting time. In definition 4.1, $\varphi_\alpha(\tilde{w})$ denotes the worst-case expected waiting time in the conditional distribution of its upper α tail (Rockafellar 2007). Therefore, roughly speaking, DUM represents the smallest upper 100α percentile, such that the worst-case average of α longest delay is no more than patient's tolerable threshold. Several properties of DUM are listed in Proposition 4.1.

Proposition 4.1. The DUM, ρ_τ has the following properties:

- (a) Monotonicity: if $\tilde{w}_1 \leq \tilde{w}_2$, then $\rho_\tau(\tilde{w}_1) \leq \rho_\tau(\tilde{w}_2)$;
- (b) Threshold Satisficing: if $\tilde{w} \leq \tau$, then $\rho_\tau(\tilde{w}) = 0$;
- (c) Tardiness Intolerance: if $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w}) > \tau$, then $\rho_\tau(\tilde{w}) = 1$;
- (d) Upper bound of tardiness probability: $\rho_\tau(\tilde{w}) \geq \mathbb{P}(\tilde{w} > \tau)$ for all $\mathbb{P} \in \mathbb{F}$;
- (e) If $\mathbb{P}(\tilde{w} < \tau) > 0$ for all $\mathbb{P} \in \mathbb{F}$, then

$$\rho_\tau(\tilde{w}) = \inf_{a > 0} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((a(\tilde{w} - \tau) + 1)^+).$$

Proof. (a) Monotonicity: if $\tilde{w}_1 \leq \tilde{w}_2$, we have for any $\alpha \in (0, 1]$, $\varphi_\alpha(\tilde{w}_1) \leq \varphi_\alpha(\tilde{w}_2)$ because of monotonicity property of $\varphi_\alpha(\tilde{w})$ function. Therefore, $\rho_\tau(\tilde{w}_1) \leq \rho_\tau(\tilde{w}_2)$.

(b) Threshold Satisficing: if $\tilde{w} \leq \tau$, $\rho_\tau(\tilde{w}) \leq \rho_\tau(\tau) = \inf\{\alpha \geq 0 \mid \varphi_\alpha(\tau) \leq \tau\} = 0$. With the bound that $\rho_\tau(\tilde{w}) \in [0, 1]$, we could immediately conclude $\rho_\tau(\tilde{w}) = 0$.

(c) Tardiness Intolerance: we first prove that $\varphi_1(\tilde{w}) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w})$. According to the definition of $\varphi_\alpha(\tilde{w})$, $\varphi_1(\tilde{w}) \leq 0 + \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w} - 0)^+) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w})$. Moreover, since

$$\varphi_1(\tilde{w}) = \min_{\nu \in \mathfrak{R}} \left\{ \sup_{\mathbb{P} \in \mathbb{F}} (\nu + (\tilde{w} - \nu)^+) \right\} \geq \min_{\nu \in \mathfrak{R}} \left\{ \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\nu + \tilde{w} - \nu) \right\} = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w}),$$

we have $\varphi_1(\tilde{w}) = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w})$. Therefore, $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}(\tilde{w}) > \tau$ is equivalent

to $\varphi_1(\tilde{w}) > \tau$. According to monotonicity property of function $\varphi_\alpha(\tilde{w})$, there exists no $\alpha \geq 0$ satisfying $\varphi_\alpha(\tilde{w}) \leq \tau$, which leads to $\rho_\tau(\tilde{w}) = 1$.

(d) The proof can be referred to Theorem 3 in Brown and Sim (2009).

(e) Given $\mathbb{P}(\tilde{w} > \tau) > 0$ for all $\mathbb{P} \in \mathbb{F}$, we could obtain for any $\nu \geq 0$, $\nu + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w} - \tau - \nu)^+) > 0$. Hence,

$$\begin{aligned}
\rho_\tau(\tilde{w}) &= \inf \{ \alpha \geq 0 \mid \varphi_\alpha(\tilde{w}) \leq \tau \} \\
&= \inf \left\{ \alpha \geq 0 \mid \exists \nu \in \mathfrak{R}, \nu + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w} - \tau - \nu)^+) \leq 0 \right\} \\
&= \inf \left\{ \alpha \geq 0 \mid \exists \nu < 0, -\nu \geq \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w} - \tau - \nu)^+) \right\} \\
&= \inf \left\{ \alpha \geq 0 \mid \exists a > 0, \frac{1}{a} \geq \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\left(\tilde{w} - \tau + \frac{1}{a} \right)^+ \right) \right\} \\
&= \inf \left\{ \alpha \geq 0 \mid \alpha \geq \inf_{a > 0} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((a(\tilde{w} - \tau) + 1)^+) \right\} \\
&= \inf_{a > 0} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((a(\tilde{w} - \tau) + 1)^+). \quad \square
\end{aligned}$$

Remark 4.1. Property (a) captures participant's essential preference to a shorter delay, i.e., if the waiting time \tilde{w}_1 is state-wise greater than its counterpart \tilde{w}_2 , then the former is not more preferred under the DUM. Property (b) indicates participant's desire to be served within the threshold and any uncertain delay that always meets the deadline will be most preferred. In contrast, Property (c) indicates the intolerance to any delay always exceeds the threshold in expectation. Property (d) suggests a close relationship between the DUM and frequency of delay over a threshold. We could guarantee that the frequency of delay over the threshold is less than the corresponding

DUM. Property (e) demonstrates that the DUM can be written as a form of an optimized expected utility, where the utility function is convex.

Next, we provide a simple illustration of the DUM. Given two options A and B on delay, where

$$\tilde{w}_A = \begin{cases} 10 \text{ minutes,} & \text{with probability } 0.89; \\ 30 \text{ minutes,} & \text{with probability } 0.11. \end{cases}$$

$$\tilde{w}_B = \begin{cases} 10 \text{ minutes,} & \text{with probability } 0.9; \\ 60 \text{ minutes,} & \text{with probability } 0.1. \end{cases}$$

When the tolerance threshold $\tau = 29$ minutes, the outcome of minimizing frequency of delay over a threshold suggests option B is better than A with $\mathbb{P}(\tilde{w}_B > 29) = 0.1 < \mathbb{P}(\tilde{w}_A > 29) = 0.11$, which indicates that this quality measure only focuses on the violation probability without taking the delay level into consideration. Instead, the use of the DUM can avoid these disadvantages with its outcome suggests that option A is more preferable than B as $\rho_{29}(\tilde{w}_A) = \frac{11}{95} \leq \frac{5}{19} = \rho_{29}(\tilde{w}_B)$.

4.3 Lexicographic Min-Max Fairness

The service quality of an appointment system depends on the participants' experiences on delays and we can formulate this as a multiple criteria optimization problem in which participants' DUMs are minimized, i.e.,

$$\min_{\tilde{w} \in \mathcal{W}} \{\rho_{\tau}(\tilde{w})\},$$

where $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) = (\rho_{\tau_1}(\tilde{w}_1), \dots, \rho_{\tau_N}(\tilde{w}_N))$ and \mathcal{W} represents the space of feasible waiting times experienced by the participants. Among the Pareto optimal solutions, we would like to mitigate unfairness and avoid discriminating a subset of participants in terms of their service experiences in the appointment system. We adopt the lexicographic min-max fairness solution approach (see Young 1995).

Definition 4.2. Let $\rho_i(\tilde{\boldsymbol{w}})$ and $\rho_i(\tilde{\boldsymbol{v}})$, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}} \in \mathcal{W}$ be the i th largest elements of $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}})$ and $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$ respectively. We say $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}})$ is lexicographically equivalent to $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$, denoted by

$$\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$$

if and only if $\rho_h(\tilde{\boldsymbol{w}}) = \rho_h(\tilde{\boldsymbol{v}})$ for all $h \in [1; N]$. Moreover, $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}})$ is lexicographically less than $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$, denoted by

$$\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$$

if and only if there exists $i^* \in [1; N]$ such that $\rho_h(\tilde{\boldsymbol{w}}) = \rho_h(\tilde{\boldsymbol{v}})$ for $h \in [1; i^* - 1]$ and $\rho_{i^*}(\tilde{\boldsymbol{w}}) < \rho_{i^*}(\tilde{\boldsymbol{v}})$. Similarly, we denote by

$$\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$$

if either $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$ or $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\boldsymbol{v}})$.

The lexicographic ordering shows that the participant with the worst

value of DUM has the highest priority in preference ranking among solutions in \mathcal{W} . Subsequently, if these values among different solutions are the same, then the next worst value will be used in deciding preference. We explore some characteristics of lexicographic ordering of participants' DUMs and link them to issues of fairness in an appointment system.

Proposition 4.2. The following properties hold for $\tilde{\mathbf{w}}, \tilde{\mathbf{v}} \in \mathcal{W}$:

- (a) Monotonicity: if $\tilde{\mathbf{w}} \leq \tilde{\mathbf{v}}$, then

$$\rho_{\mathcal{T}}(\tilde{\mathbf{w}}) \preceq_{\text{lex}} \rho_{\mathcal{T}}(\tilde{\mathbf{v}}).$$

- (b) Threshold Satisficing: let $\mathcal{S} \subset [1; N]$ and $\bar{\mathcal{S}}$ be the complement set.

Suppose $\tilde{v}_j = \tilde{w}_j$ for all $j \in \mathcal{S}$ and $\tilde{v}_j \leq \tilde{w}_j \leq \tau_j$ for all $j \in \bar{\mathcal{S}}$, then

$$\rho_{\mathcal{T}}(\tilde{\mathbf{w}}) =_{\text{lex}} \rho_{\mathcal{T}}(\tilde{\mathbf{v}}).$$

- (c) Discrimination Resistance: let

$$\mathcal{S}_1 = \{i \in [1; N] \mid \rho_{\tau_i}(\tilde{w}_i) = 1\} \text{ and } \mathcal{S}_2 = \{i \in [1; N] \mid \rho_{\tau_i}(\tilde{v}_i) = 1\}.$$

Suppose $|\mathcal{S}_1| < |\mathcal{S}_2|$ then

$$\rho_{\mathcal{T}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \rho_{\mathcal{T}}(\tilde{\mathbf{v}}).$$

Proof. (a) Monotonicity: if $\tilde{\mathbf{w}} \leq \tilde{\mathbf{v}}$, i.e., $\tilde{w}_n \leq \tilde{v}_n$ for all $n \in [1; N]$, with the monotonicity property of $\rho_{\tau_n}(\tilde{w}_n)$, we have for all $n \in [1; N]$, $\rho_{\tau_n}(\tilde{w}_n) \leq \rho_{\tau_n}(\tilde{v}_n)$. Therefore, $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{w}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{v}})$.

(b) Threshold Satisficing: Since $\tilde{w}_n = \tilde{v}_n$ for all $n \in \mathcal{S}$, we have $\rho_{\tau_n}(\tilde{w}_n) = \rho_{\tau_n}(\tilde{v}_n)$. For any $j \in \bar{\mathcal{S}}$, $\tilde{w}_j, \tilde{v}_j \leq \tau_j$, then according to Threshold Satisficing of DUM, $\rho_{\tau_j}(\tilde{w}_j) = \rho_{\tau_j}(\tilde{v}_j) = 0$. Therefore, $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{v}})$.

(c) Discrimination Resistance: if $|\mathcal{S}_1| < |\mathcal{S}_2|$, we have $\rho_i(\tilde{\mathbf{w}}) = \rho_i(\tilde{\mathbf{v}})$ for all $i \in [1; |\mathcal{S}_1|]$. For $i = |\mathcal{S}_1| + 1 \leq |\mathcal{S}_2|$, we have $\rho_i(\tilde{\mathbf{w}}) < 1 = \rho_i(\tilde{\mathbf{v}})$. Therefore, $\boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\mathcal{T}}(\tilde{\mathbf{v}})$. \square

Remark 4.2. Monotonicity ensures consistency so that reduction in delays for all participants will be favorably valued. Threshold Satisficing property ensures that the participants whose delays are always within their thresholds, then any improvement of their delays do not contribute to the lexicographic ordering. A participant is discriminated if the appointment system cannot guarantee his/her average waiting time below the threshold, which corresponds to the DUM taking value of one. Hence, Discrimination Resistance induces preferences for solutions that have fewer participants being discriminated. This property is in accord with the hospital's key performance indicator, to keep the number of patients who experiences the worst waiting process as small as possible. ¹

¹ In the context of earlier example provided by Klassen and Rohleder (1996), if each patient's tolerable threshold is 3 minutes, the number of patients whose DUMs equal to 1 is 20 to the strategy that keeps only one patient waiting for 40 minutes, while that to the other strategy is 0.

Since lexicographic order is complete, we can rank solutions and replace the multiple criteria optimization by the following lexicographic minimization problem

$$\text{lex min}_{\tilde{w} \in \mathcal{W}} \{\rho_{\tau}(\tilde{w})\},$$

where the optimal solution $\tilde{w}^* \in \mathcal{W}$ satisfies

$$\rho_{\tau}(\tilde{w}^*) \preceq_{\text{lex}} \rho_{\tau}(\tilde{v}), \quad \forall \tilde{v} \in \mathcal{W}.$$

Though this may not be a standard mathematical programming problem, we can obtain the optimal solution by solving a sequence of optimization problems (see Isermann 1982 and Ogoryczak et al. 2005) as follows:

Algorithm: Lexicographic Minimization Procedure

1. Set $h := 1, \mathcal{G}_0 := [1; N]$,

$$\begin{aligned} \alpha_1 &:= \min_{\tilde{w} \in \mathcal{W}} \max_{n \in \mathcal{G}_0} \rho_{\tau_n}(\tilde{w}_n), \\ \mathcal{I}_1 &:= \left\{ j \in \mathcal{G}_0 : \min_{\tilde{w} \in \mathcal{W}} \left\{ \rho_{\tau_j}(\tilde{w}_j) \mid \max_{n \in \mathcal{G}_0} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_1 \right\} = \alpha_1 \right\}. \end{aligned}$$

2. Set $\mathcal{G}_h := \mathcal{G}_{h-1} \setminus \mathcal{I}_h$. If $\mathcal{G}_h = \emptyset$, algorithm terminates and outputs solution. Otherwise, set $h := h + 1$,

$$\begin{aligned} \alpha_h &:= \min_{\tilde{w} \in \mathcal{W}} \left\{ \max_{n \in \mathcal{G}_{h-1}} \rho_{\tau_n}(\tilde{w}_n) \mid \max_{n \in \mathcal{I}_i} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_i, i \in \{1, \dots, h-1\} \right\}, \\ \mathcal{I}_h &:= \left\{ j \in \mathcal{G}_{h-1} : \min_{\tilde{w} \in \mathcal{W}} \left\{ \rho_{\tau_j}(\tilde{w}_j) \mid \begin{array}{l} \max_{n \in \mathcal{G}_{h-1}} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_h, \\ \max_{n \in \mathcal{I}_i} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_i, i \in \{1, \dots, h-1\} \end{array} \right\} = \alpha_h \right\}. \end{aligned}$$

3. Go to Step 2.

In this algorithm, we minimize the maximum DUM among a set of participants and elicit the subset of participants that attain the worst value. Hence, the optimum solution, $\tilde{\mathbf{w}}^* \in \mathcal{W}$ satisfies

$$\rho_{\tau_n}(\tilde{w}_n^*) = \alpha_i, \quad n \in \mathcal{I}_i,$$

for all $i \in [1; h]$. Observe that the problem to derive α_h is the same as

$$\begin{aligned} \alpha_h = \min \quad & \alpha \\ \text{s.t.} \quad & \rho_{\tau_n}(\tilde{w}_n) \leq \alpha, \quad n \in \mathcal{G}_{h-1}, \\ & \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_i, \quad n \in \mathcal{I}_i, i \in [1; h-1], \\ & \tilde{\mathbf{w}} \in \mathcal{W}. \end{aligned}$$

According to the definition of $\rho_{\tau_n}(\tilde{w}_n)$, we could equivalently solve

$$\begin{aligned} \inf \quad & \alpha \\ \text{s.t.} \quad & \nu_n + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{G}_{h-1}, \\ & \nu_n + \frac{1}{\alpha_i} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{I}_i, i \in [1; h-1], \quad (4.1) \\ & \alpha \in (0, 1], \\ & \tilde{\mathbf{w}} \in \mathcal{W}. \end{aligned}$$

Though the problem is nonlinear in α , we observe that $\frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+)$ is monotonic in α and hence we could use binary search procedure to find the optimal solution in which α is minimized. Similarly, we can determine

\mathcal{I}_i by performing a sequence of binary search procedures.

4.4 Appointment Schedule Design

We first consider an appointment scheduling problem with one physician serving N patients under the following assumptions:

Assumptions

- Schedules have to be made before the commencement of the session.
- Patients may be heterogenous and are characterized by their service time distributions and tolerance thresholds.
- The consultation sequence of patients is pre-determined.
- Patients arrive on time.²
- Physician will start his/her session promptly. Hence, the first patient experiences no delay.

Model parameters and decision variables

- N : total number of patients to be scheduled;
- L : session length pre-determined for the consultation of N patients;
- τ_n : the tolerance threshold of delay for the patient at n th position, $n \in [1; N]$;

² According to data collection of Harper and Gamlin (2003) and Zhu et al. (2011), majority of patients arrive earlier than they are expected. This assumption avoids the complexity of modeling due to potential change in sequence.

- τ_{N+1} : physician's tolerance on his/her overtime;
- \tilde{s}_n : consultation time of the n th patient;
- \tilde{w}_n : waiting time of the n th patient, $n \in [1; N]$;
- \tilde{w}_{N+1} : physician's overtime;
- x_n : decision variable, appointment time for the n th patient. For notational simplicity, we let $x_1 = 0, x_{N+1} = L$, and its vector notation $\mathbf{x} = (x_1, \dots, x_N, x_{N+1})'$.

We first specify the feasible set of waiting times, \mathcal{W} as follows:

$$\mathcal{W} = \left\{ \tilde{\mathbf{w}} \left| \begin{array}{l} \tilde{w}_1 = 0, \\ \tilde{w}_n = \max \{x_{n-1} + \tilde{w}_{n-1} + \tilde{s}_{n-1} - x_n, 0\}, \quad n \in [2; N + 1], \\ \mathbf{x} \in \mathcal{X} \end{array} \right. \right\},$$

where set \mathcal{X} is defined as

$$\mathcal{X} = \left\{ \mathbf{x} \left| \begin{array}{l} x_1 = 0, \\ x_{n-1} \leq x_n, \quad n \in [2; N + 1], \\ x_{N+1} = L \end{array} \right. \right\}.$$

The first two constraints in the set \mathcal{W} recursively calculate the delays experienced by the patients and the physician, while the set \mathcal{X} ensures sequencing compliance. Accordingly as in Denton and Gupta (2003), we further simplify the formulation by defining the difference between the real service time and

scheduled interval as \tilde{t}_n for the n th patient

$$\tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in [1; N]. \quad (4.2)$$

It follows that the n th patient's waiting time and physician's overtime can be represented by

$$\tilde{w}_n = \max \left\{ 0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k \right\}, \quad n \in [2; N + 1]. \quad (4.3)$$

Since the lexicographic minimization procedure requires solving a sequence of similar problems, we will focus on solving Problem (4.1) as a representative instance. To derive the optimal scheduling decisions, we formulate Problem (4.1) as

$$\begin{aligned} & \inf \quad \alpha \\ & \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{G}_{h-1}, \\ & \quad \nu_n + \frac{1}{\alpha_i} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{I}_i, i \in [1; h - 1], \\ & \quad \tilde{w}_n = \max \left\{ 0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k \right\}, \quad n \in [2; N + 1], \\ & \quad \tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in [1; N], \\ & \quad \alpha \in (0, 1], \\ & \quad \mathbf{x} \in \mathcal{X}. \end{aligned} \quad (4.4)$$

Since the first patient's waiting time is zero, we have $\rho_{\tau_1}(\tilde{w}_1) = 0$ for any nonnegative threshold τ_1 . Therefore, we can define $\mathcal{G}_0 = [2; N + 1]$.

We first focus on the simplification of function $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+)$, which is complicated by the recursive property of uncertain waiting times. In conjunction with Equations (4.2) and (4.3), we observe that

$$\begin{aligned}
& \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\left(\max \left\{ 0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k \right\} - \nu_n \right)^+ \right) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{t}_{n-1} - \nu_n, \dots, \sum_{k=1}^{n-1} \tilde{t}_k - \nu_n \right\} \right) \\
&= \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n \right\} \right).
\end{aligned}$$

The calculation of this function inevitably depends on the information we possess about the uncertain service time \tilde{s}_n , $n \in [1; N]$. Next, we will classify the information set we could have on \tilde{s}_n and provide different reformulation and solution techniques.

4.4.1 Stochastic optimization approach

For the case of known discrete distribution (i.e. $\mathbb{F} = \{\mathbb{P}\}$) in which there are M sets of service times, $\{s_1^m, \dots, s_N^m\}$, each occurring with probability p_m , $m \in [1; M]$, we have

$$\begin{aligned}
& \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n \right\} \right) \\
&= \sum_{m=1}^M p_m \max \left\{ 0, -\nu_n, s_{n-1}^m - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (s_k^m - (x_{k+1} - x_k)) - \nu_n \right\}.
\end{aligned}$$

Therefore, by adding decision variables q_{mn} , $m \in [1; M], n \in [2; N + 1]$, Problem (4.4) is equivalent to

$$\begin{aligned}
& \inf \quad \alpha \\
& \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \sum_{m=1}^M p_m q_{mn} \leq \tau_n, & n \in \mathcal{G}_{h-1}, \\
& \quad \nu_n + \frac{1}{\alpha_i} \sum_{m=1}^M p_m q_{mn} \leq \tau_n, & n \in \mathcal{I}_i, i \in [1; h-1], \\
& \quad q_{mn} + \nu_n \geq 0, & n \in [2; N+1], m \in [1; M], \\
& \quad q_{mn} + \nu_n + x_n - x_l \geq \sum_{k=l}^{n-1} s_k^m, & l \in [1; n-1], n \in [2; N+1], m \in [1; M], \\
& \quad q_{mn} \geq 0, & n \in [2; N+1], m \in [1; M], \\
& \quad \alpha \in (0, 1], \\
& \quad \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

Whenever α is fixed, the feasible set is a polyhedron comprising $O(MN)$ decision variables and $O(MN^2)$ constraints. In practice, this approach is amiable to empirical distributions where M is relatively small.

4.4.2 Distributionally robust optimization approach

We also propose a distributional robust optimization approach with the goal of preserving linearity of the model. We assume the family of service times distributions are characterized based on their bounded supports $\mathbb{P}(\tilde{s}_k \in [\underline{s}_k, \bar{s}_k]) = 1$, means $\mathbb{E}_{\mathbb{P}}(\tilde{s}_k) = \mu_k$, $\mu_k \in (\underline{s}_k, \bar{s}_k)$ and bounds of mean absolute deviation $\mathbb{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \leq \sigma_k$, $\sigma_k > 0$ for all $k \in [1; N]$. Intuitively, the worst case probability distributions may result in highly correlated

service times, which may not be realistic and lead to conservative solutions. To impose correlation, the conventional approach is to specify covariance within the distributional uncertainty set, i.e. the descriptive statistics of $\mathbb{E}_{\mathbb{P}}((\tilde{s}_r - \mu_r)(\tilde{s}_k - \mu_k))$ for all $r, k \in [1; N]$, $r \leq k$. However, this will necessarily lead to nonlinear optimization models, which are harder to solve (Kong et al. 2013; Mak et al. 2013). To avoid nonlinearity, we propose a different approach of capturing correlation. We note that the waiting time of a participant may be influenced by the aggregation of uncertain service times of earlier participants. Hence, in our distributional uncertainty set, we use the descriptive statistics of $\mathbb{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^k \frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right)$ for all $r, k \in [1; N]$ and $r \leq k$. Observe that

$$\mathbb{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^k \frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \leq \sum_{m=r}^k \mathbb{E}_{\mathbb{P}}\left(\left|\frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \leq k - r + 1,$$

in which the first equality is achieved under perfect correlation. As a proxy for modeling correlation, we impose the constraints,

$$\mathbb{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^k \frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \leq \epsilon_{rk}, \quad r, k \in [1; N], r \leq k,$$

where $\epsilon_{rk} \in (0, k - r + 1]$. Without loss of generality, we define $\epsilon_{kk} = 1$ that is equivalent to the information $\mathbb{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \leq \sigma_k$. These constraints set the bound for the dispersion of the total uncertain service times for $k - r + 1$ consecutive patients, and enable us to specify less conservative uncertainty set while keeping the model linear. Now, the distributional uncertainty set

can be written as

$$\mathbb{F} = \left\{ \mathbb{P} \left| \begin{array}{l} \mathbb{E}_{\mathbb{P}}(\tilde{s}_k) = \mu_k, \mathbb{P}(\tilde{s}_k \in [\underline{s}_k, \bar{s}_k]) = 1, \mathbb{E}_{\mathbb{P}} \left(\left| \sum_{m=r}^k \frac{\tilde{s}_m - \mu_m}{\sigma_m} \right| \right) \leq \epsilon_{rk}, \\ r, k \in [1; N], r \leq k \end{array} \right. \right\}.$$

For convenience, we let $\tilde{z}_k = (\tilde{s}_k - \mu_k)/\sigma_k$, and define \mathbb{F}_z as

$$\mathbb{F}_z = \left\{ \mathbb{P} \left| \begin{array}{l} \mathbb{E}_{\mathbb{P}}(\tilde{z}_k) = 0, \mathbb{P}(\tilde{z}_k \in [\underline{z}_k, \bar{z}_k]) = 1, \mathbb{E}_{\mathbb{P}} \left(\left| \sum_{m=r}^k \tilde{z}_m \right| \right) \leq \epsilon_{rk}, \\ r, k \in [1; N], r \leq k \end{array} \right. \right\},$$

and we have

$$\begin{aligned} & \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n \right\} \right) \\ &= \sup_{\mathbb{P} \in \mathbb{F}_z} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \sigma_{n-1} \tilde{z}_{n-1} + \mu_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \right. \right. \\ & \quad \left. \left. \sum_{k=1}^{n-1} (\sigma_k \tilde{z}_k + \mu_k - (x_{k+1} - x_k)) - \nu_n \right\} \right) \end{aligned}$$

Proposition 4.3. For a given $\mathbf{x} \in \mathcal{X}$ and $n \in [2; N + 1]$, the problem

$$Z_P = \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n \right\} \right)$$

corresponds to the optimal value of the following linear optimization problem

$$\begin{aligned}
Z_D = \inf \quad & f_0 + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk} \\
\text{s.t.} \quad & f_0 + \sum_{k=1}^{n-1} (\underline{z}_k u_k^0 - \bar{z}_k v_k^0) \geq 0, \\
& f_0 + \nu_n + \sum_{k=1}^{n-1} (\underline{z}_k u_k^n - \bar{z}_k v_k^n) \geq 0, \\
& f_0 + \nu_n + x_n - x_l + \sum_{k=1}^{n-1} (\underline{z}_k u_k^l - \bar{z}_k v_k^l) \geq \sum_{k=l}^{n-1} \mu_k, \quad l \in [1; n-1], \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k = 0, \quad k \in [1; n-1], l = 0, n, \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k = 0, \quad k, l \in [1; n-1], k \leq l-1, \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k = -\sigma_k, \quad k, l \in [1; n-1], l \leq k, \\
& b_{rk}^l + c_{rk}^l - g_{rk} = 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& u_k^l, v_k^l, b_{rk}^l, c_{rk}^l, g_{rk} \geq 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n].
\end{aligned} \tag{4.5}$$

Proof. To justify our claim, we first notice that the calculation of function

$$\begin{aligned}
& Z_P \\
= \quad & \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n \right\} \right)
\end{aligned}$$

can be equivalently written as an optimization problem as follows

$$\begin{aligned}
Z_P = \sup \quad & \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \dots, \sum_{k=1}^{n-1} (\sigma_k \tilde{z}_k + \mu_k - (x_{k+1} - x_k)) - \nu_n \right\} \right) \\
\text{s.t.} \quad & \mathbb{E}_{\mathbb{P}} (\tilde{z}_k) = 0, \quad k \in [1; n-1], \\
& \mathbb{E}_{\mathbb{P}} \left(\left| \sum_{m=r}^k \tilde{z}_m \right| \right) \leq \epsilon_{rk}, \quad r, k \in [1; n-1], r \leq k, \\
& \mathbb{P} \{ \tilde{z}_k \in [\underline{z}_k, \bar{z}_k], k \in [1; n-1] \} = 1.
\end{aligned} \tag{4.6}$$

Its dual form can be written as

$$\begin{aligned}
Z_1 = \min \quad & f_0 + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk} \\
\text{s.t.} \quad & f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1} \sum_{r=1}^k g_{rk} \left| \sum_{m=r}^k \tilde{z}_m \right| \geq 0, \\
& \quad \quad \quad \forall z_k \in [\underline{z}_k, \bar{z}_k], k \in [1; n-1], \\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1} \sum_{r=1}^k g_{rk} \left| \sum_{m=r}^k z_m \right| \geq -\nu_n, \\
& \quad \quad \quad \forall z_k \in [\underline{z}_k, \bar{z}_k], k \in [1; n-1], \\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1} \sum_{r=1}^k g_{rk} \left| \sum_{m=r}^k z_m \right| \geq \sum_{k=l}^{n-1} (\sigma_k z_k + \mu_k - (x_{k+1} - x_k)) - \nu_n, \\
& \quad \quad \quad \forall z_k \in [\underline{z}_k, \bar{z}_k], k, l \in [1; n-1], \\
& g_{rk} \geq 0, \quad r, k \in [1; n-1], r \leq k,
\end{aligned} \tag{4.7}$$

in which weak duality holds (see Isii 1963), and hence, $Z_P \leq Z_1$. Observe that each constraint in Problem (4.7) is the robust counterpart of a linear optimization problem with bounded box uncertainty set. Hence, Problem (4.7) is feasible and objective is finite, i.e., $Z_1 < \infty$. Moreover, the dual form

of the linear optimization problem

$$\begin{aligned}
\min \quad & \sum_{k=1}^{l-1} f_k z_k + \sum_{k=l}^{n-1} (f_k - \sigma_k) z_k + \sum_{k=1}^{n-1} \sum_{r=1}^k g_{rk} \left| \sum_{m=r}^k z_m \right| \\
\text{s.t.} \quad & z_k \geq \underline{z}_k, & k \in [1; n-1], \\
& z_k \leq \bar{z}_k, & k \in [1; n-1],
\end{aligned}$$

is equivalently written as

$$\begin{aligned}
\max \quad & \sum_{k=1}^{n-1} (\underline{z}_k u_k - \bar{z}_k v_k) \\
\text{s.t.} \quad & u_k - v_k + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm} - c_{rm}) = f_k, & k \in [1; l-1], \\
& u_k - v_k + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm} - c_{rm}) = f_k - \sigma_k, & k \in [l; n-1], \\
& b_{rk} + c_{rk} = g_{rk}, & r, k \in [1; n-1], r \leq k, \\
& u_k, v_k, b_{rk}, c_{rk} \geq 0, & r, k \in [1; n-1], r \leq k.
\end{aligned}$$

Combining all these analysis parts together, we could derive the optimization problem (4.5) in the proposition, and $Z_P \leq Z_1 = Z_D$. To show that strong duality holds for the primal problem (4.6) and the dual problem (4.5), we cannot directly use the result of Isii (1963). To prove it, we derive the dual

of Problem (4.5) as

$$\begin{aligned}
Z_2 = \max \quad & -\lambda_n \nu_n + \sum_{l=1}^{n-1} \lambda_l \left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1} \mu_k \right) + \sum_{l=1}^{n-1} \sum_{k=l}^{n-1} \kappa_{lk} \sigma_k \\
\text{s.t.} \quad & \sum_{l=0}^n \lambda_l = 1, \\
& \sum_{l=0}^n \kappa_{lk} = 0, & k \in [1; n-1], \\
& -\kappa_{lk} + \lambda_l \underline{z}_k \leq 0, & k \in [1; n-1], l \in [0; n], \\
& \kappa_{lk} - \lambda_l \bar{z}_k \leq 0, & k \in [1; n-1], l \in [0; n], \\
& -\eta_{rk}^l + \sum_{m=r}^k \kappa_{lm} \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& -\eta_{rk}^l - \sum_{m=r}^k \kappa_{lm} \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& \sum_{l=0}^n \eta_{rk}^l \leq \epsilon_{rk}, & r, k \in [1; n-1], r \leq k, \\
& \lambda_l \geq 0, & l \in [0; n].
\end{aligned} \tag{4.8}$$

Since strong duality holds in this linear optimization problem, we have $Z_D = Z_2 \in \mathfrak{R}$. Since, $\mu_k \in (\underline{s}_k, \bar{s}_k)$, we have $0 \in (\underline{z}_k, \bar{z}_k)$ for all $k \in [1; n-1]$. Therefore, solution $\lambda_l = \frac{1}{n+1}$, $\kappa_{lk} = 0$, $\eta_{rk}^l = \frac{\epsilon_{rk}}{n+2}$, $r, k \in [1; n-1]$, $r \leq k$, $l \in [0; n]$ is strictly feasible. Since Problem (4.8) is a linear optimization problem with finite objective and non-empty relative interior, there exists a sequence of interior feasible solutions whose objectives asymptotically coverage to opti-

mum. Hence, we have

$$\begin{aligned}
Z_2 = \sup & \quad -\lambda_n \nu_n + \sum_{l=1}^{n-1} \lambda_l \left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1} \mu_k \right) + \sum_{l=1}^{n-1} \sum_{k=l}^{n-1} \kappa_{lk} \sigma_k \\
\text{s.t.} & \quad \sum_{l=0}^n \lambda_l = 1, \\
& \quad \sum_{l=0}^n \kappa_{lk} = 0, & k \in [1; n-1], \\
& \quad -\kappa_{lk} + \lambda_l \bar{z}_k \leq 0, & k \in [1; n-1], l \in [0; n], \\
& \quad \kappa_{lk} - \lambda_l \bar{z}_k \leq 0, & k \in [1; n-1], l \in [0; n], \\
& \quad -\eta_{rk}^l - \sum_{m=r}^k \kappa_{lm} \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& \quad -\eta_{rk}^l + \sum_{m=r}^k \kappa_{lm} \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& \quad \sum_{l=0}^n \eta_{rk}^l \leq \epsilon_{rk}, & r, k \in [1; n-1], r \leq k, \\
& \quad \lambda_l > 0, & l \in [0; n].
\end{aligned}$$

Since $\lambda_l > 0$, by defining $\zeta_{lk} = \kappa_{lk}/\lambda_l$, $l \in [0; n]$, $k \in [1; n-1]$, the above

problem is equivalent to

$$\begin{aligned}
Z_2 &= \sup -\lambda_n \nu_n + \sum_{l=1}^{n-1} \lambda_l \left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1} (\mu_k + \zeta_{lk} \sigma_k) \right) \\
\text{s.t. } &\sum_{l=0}^n \lambda_l = 1, \\
&\sum_{l=0}^n \lambda_l \zeta_{lk} = 0, & k \in [1; n-1], \\
&-\zeta_{lk} \leq -\underline{z}_k, & k \in [1; n-1], l \in [0; n], \\
&\zeta_{lk} \leq \bar{z}_k, & k \in [1; n-1], l \in [0; n], \\
&-\eta_{rk}^l - \sum_{m=r}^k \zeta_{lm} \lambda_l \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
&-\eta_{rk}^l + \sum_{m=r}^k \zeta_{lm} \lambda_l \leq 0, & r, k \in [1; n-1], r \leq k, l \in [0; n], \\
&\sum_{l=0}^n \eta_{rk}^l \leq \epsilon_{rk}, & r, k \in [1; n-1], r \leq k, \\
&\lambda_l > 0, & l \in [0; n], \\
&= \sup -\lambda_n \nu_n + \sum_{l=1}^{n-1} \lambda_l \left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1} (\mu_k + \zeta_{lk} \sigma_k) \right) \\
\text{s.t. } &\sum_{l=0}^n \lambda_l \zeta_{lk} = 0, & k \in [1; n-1], \\
&\sum_{l=0}^n \lambda_l \left| \sum_{m=r}^k \zeta_{lm} \right| \leq \epsilon_{rk}, & r, k \in [1; n-1], r \leq k, \\
&\sum_{l=0}^n \lambda_l = 1, \\
&\zeta_{lk} \in [\underline{z}_k, \bar{z}_k], & k \in [1; n-1], l \in [0; n], \\
&\lambda_l > 0, & l \in [0; n].
\end{aligned} \tag{4.9}$$

We observe that the feasible solution in Problem (4.9) can be translated to \tilde{z}_k being discrete distributed that takes values of ζ_{lk} with probability λ_l ,

$l \in [0; n]$ for all $k \in [1; n - 1]$. Moreover, the objective of Problem (4.9)

satisfies

$$\begin{aligned} & -\lambda_n \nu_n + \sum_{l=1}^{n-1} \lambda_l \left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1} (\mu_k + \zeta_{lk} \sigma_k) \right) \\ \leq & \sum_{l=0}^n \lambda_l \left(\max \left\{ 0, -\nu_n, \dots, \sum_{k=1}^{n-1} (\sigma_k \zeta_{lk} + \mu_k - (x_{k+1} - x_k)) - \nu_n \right\} \right). \end{aligned}$$

Therefore, $Z_P \leq Z_1 = Z_D = Z_2 \leq Z_P$ and strong duality follows. \square

Correspondingly, Problem (4.4) is equivalent to

$$\begin{aligned}
& \inf \quad \alpha \\
& \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \left(f_0^n + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk}^n \right) \leq \tau_n, & n \in \mathcal{G}_{h-1}, \\
& \quad \nu_n + \frac{1}{\alpha_i} \left(f_0^n + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk}^n \right) \leq \tau_n, & n \in \mathcal{I}_i, i \in [1; h-1], \\
& \quad f_0^n + \sum_{k=1}^{n-1} (\underline{z}_k u_k^{0n} - \bar{z}_k v_k^{0n}) \geq 0, & n \in [2; N+1], \\
& \quad f_0^n + \nu_n + \sum_{k=1}^{n-1} (\underline{z}_k u_k^{nn} - \bar{z}_k v_k^{nn}) \geq 0, & n \in [2; N+1], \\
& \quad f_0^n + \nu_n + x_n - x_l + \sum_{k=1}^{n-1} (\underline{z}_k u_k^{ln} - \bar{z}_k v_k^{ln}) \geq \sum_{k=l}^{n-1} \mu_k, & l \in [1; n-1], n \in [2; N+1], \\
& \quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, & k \in [1; n-1], l = 0, n, n \in [2; N+1], \\
& \quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, & k, l \in [1; n-1], k \leq l-1, n \in [2; N+1], \\
& \quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = -\sigma_k, & k, l \in [1; n-1], k \geq l, n \in [2; N+1], \\
& \quad b_{rk}^{ln} + c_{rk}^{ln} - g_{rk}^n = 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], n \in [2; N+1], \\
& \quad u_k^{ln}, v_k^{ln}, b_{rk}^{ln}, c_{rk}^{ln}, g_{rk}^n \geq 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], n \in [2; N+1], \\
& \quad \alpha \in (0, 1], \\
& \quad \mathbf{x} \in \mathcal{X}.
\end{aligned} \tag{4.10}$$

Problem (4.10) is quite complicated at a first glance, however, for any $\alpha \in$

$(0, 1]$, we observe that the problem reduces to a linear feasibility problem including $O(N^4)$ continuous decision variables and $O(N^4)$ constraints. When α decreases to zero, $\varphi_\alpha(\tilde{w})$ approaches the upper limit of \tilde{w} . We assume that it is onus of the decision maker to select the threshold values so that Problem (4.10) is feasible at $\alpha = 1$. Otherwise, the delay thresholds are not attainable in expectation and should be adjusted accordingly to reflect what is realistically achievable in practice.

It is worthy pointing out that the above scheduling formulation preserves linearity, and greatly reduces the computational complexity. Each approach only requires solving a sequence of linear optimization problems.

4.5 Appointment Sequence and Schedule Design

We now generalize the scheduling model to incorporate the realistic situation with sequencing decisions for heterogeneous patients. First, we clarify some extra parameters and decision variables.

- J : number of patient types. Patients with the same type have same mean μ_j , mean absolute deviation σ_j of the consultation time, and same tolerance threshold;
- N_j : number of j th type patients, where $\sum_{j=1}^J N_j = N$;
- β_j : the tolerance threshold of delay for j th type patients, $j \in [1; J]$;
- \tilde{s}_{nj} : uncertain service time associated with the n th patient if he/she belongs to j th type;

- y_{nj} : binary decision variable, if the j th type patient is scheduled in the n th position, then $y_{nj} = 1$, otherwise, $y_{nj} = 0$. Its matrix form is $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)' \in \{0, 1\}^{N \times J}$.

Correspondingly, with the sequencing decisions, the patient at position $n \in [1; N]$ has uncertain service time $\sum_{j=1}^J \tilde{s}_{nj} y_{nj}$ and tolerance threshold $\tau_n = \sum_{j=1}^J \beta_j y_{nj}$. We can formulate Problem (4.1) with both sequencing and scheduling decisions as follows:

$$\begin{aligned}
& \inf \quad \alpha \\
& \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{G}_{h-1}, \\
& \quad \nu_n + \frac{1}{\alpha_i} \sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_n - \nu_n)^+) \leq \tau_n, \quad n \in \mathcal{I}_i, i \in [1; h-1], \\
& \quad \tilde{w}_n = \max \left\{ 0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k \right\}, \quad n \in [2; N+1], \\
& \quad \tilde{t}_n = \sum_{j=1}^J \tilde{s}_{nj} y_{nj} - (x_{n+1} - x_n), \quad n \in [1; N], \\
& \quad \alpha \in (0, 1], \\
& \quad (\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathcal{Y},
\end{aligned} \tag{4.11}$$

in which

$$\mathcal{Y} = \left\{ (\boldsymbol{\tau}, \boldsymbol{x}, \mathbf{Y}) \left| \begin{array}{l} \sum_{j=1}^J \beta_j y_{nj} = \tau_n, \quad n \in [1; N], \\ \sum_{n=1}^N y_{nj} = N_j, \quad j \in [1; J], \\ \sum_{j=1}^J y_{nj} = 1, \quad n \in [1; N], \\ y_{nj} \in \{0, 1\}, \quad n \in [1; N], j \in [1; J], \\ \boldsymbol{x} \in \mathcal{X}. \end{array} \right. \right\}.$$

Set \mathcal{Y} guarantees that each patient is assigned to a position, and each position allotted to only one patient.

To solve this problem, we can implement similar procedures described in Section 4.4. The difference lies in the calculation of function $\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} ((\tilde{w}_i - \nu_i)^+)$, which is equivalent to

$$\sup_{\mathbb{P} \in \mathbb{F}} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \dots, \sum_{k=1}^{n-1} \left(\sum_{j=1}^J \tilde{s}_{kj} y_{kj} - (x_{k+1} - x_k) \right) - \nu_n \right\} \right) \quad (4.12)$$

For known discrete distribution case in which there are M sets of service time, $(s_{nj}^m)_{n \in [1; N], j \in [1; J]}$ with probability $p_m, m \in [1; M]$, Problem (4.12) can be formulated as

$$\sum_{m=1}^M p_m \max \left\{ 0, -\nu_n, \dots, \sum_{k=1}^{n-1} \left(\sum_{j=1}^J s_{kj}^m y_{kj} - (x_{k+1} - x_k) \right) - \nu_n \right\}.$$

By adding decision variables $q_{mn}, n \in [2; N + 1], m \in [1; M]$, Problem (4.11)

is equivalent to

$$\begin{aligned}
& \inf \quad \alpha \\
& \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \sum_{m=1}^M p_m q_{mn} \leq \tau_n, & n \in \mathcal{G}_{h-1}, \\
& \quad \nu_n + \frac{1}{\alpha_i} \sum_{m=1}^M p_m q_{mn} \leq \tau_n, & n \in \mathcal{I}_i, i \in [1; h-1], \\
& \quad q_{mn} + \nu_n \geq 0, & n \in [2; N+1], m \in [1; M], \\
& \quad q_{mn} + \nu_n + x_n - x_l - \sum_{k=l}^{n-1} \sum_{j=1}^J s_{kj}^m y_{kj} \geq 0, & l \in [1; n-1], n \in [2; N+1], \\
& & m \in [1; M], \\
& \quad q_{mn} \geq 0, & n \in [2; N+1], m \in [1; M], \\
& \quad \alpha \in (0, 1], \\
& \quad (\boldsymbol{\tau}, \boldsymbol{x}, \mathbf{Y}) \in \mathcal{Y}.
\end{aligned}$$

Similarly, binary search algorithm is used for finding optimal solution. For any fixed $\alpha \in (0, 1]$, the problem becomes a mixed-integer programming problem, including $N \times J$ binary decision variables, $O(MN)$ continuous decision variables, and $O(MN^2)$ constraints.

To obtain an amicably tractable robust optimization model, we assume that the uncertain service times $\tilde{s}_{1j}, \dots, \tilde{s}_{Nj}$ are respectively affinely dependent on a set of factors, $\tilde{z}_1, \dots, \tilde{z}_N$ for all patient types $j \in [1; J]$. Moreover, the centrality and dispersion of \tilde{s}_{nj} are characterized by the patient type, i.e.,

$$\tilde{s}_{nj} = \tilde{z}_n \sigma_j + \mu_j,$$

for all $n \in [1; N]$ and $j \in [1; J]$. Furthermore, the factors have the same support and its distributional uncertainty set is given as follows:

$$\mathbb{F}_z = \left\{ \mathbb{P} \left| \begin{array}{l} \mathbb{E}_{\mathbb{P}}(\tilde{z}_k) = 0, \mathbb{P}(\tilde{z}_k \in [\underline{z}, \bar{z}]) = 1, \mathbb{E}_{\mathbb{P}}\left(\left|\sum_{m=i}^k \tilde{z}_m\right|\right) \leq \epsilon_{rk}, \\ r, k \in [1; N], r \leq k, \end{array} \right. \right\}.$$

With this linear formulation, Problem (4.12) is written as

$$\sup_{\mathbb{P} \in \mathbb{F}_z} \mathbb{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \sum_{j=1}^J (\tilde{z}_{n-1} \sigma_j + \mu_j) y_{n-1,j} - (x_n - x_{n-1}) - \nu_n, \dots, \sum_{k=1}^{n-1} \left(\sum_{j=1}^J (\tilde{z}_k \sigma_j + \mu_j) y_{k,j} - (x_{k+1} - x_k) \right) - \nu_n \right\} \right), \quad (4.13)$$

Proposition 4.4. For any fixed decisions $(\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathcal{Y}$ and $n \in [2; N + 1]$, Problem (4.13) corresponds to the optimal value of the following linear

optimization problem

$$\begin{aligned}
\min \quad & f_0 + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk} \\
\text{s.t.} \quad & f_0 + \sum_{k=1}^{n-1} (\underline{z}u_k^0 - \bar{z}v_k^0) \geq 0, \\
& f_0 + \nu_n + \sum_{k=1}^{n-1} (\underline{z}u_k^n - \bar{z}v_k^n) \geq 0, \\
& f_0 + \nu_n + x_n - x_l - \sum_{k=l}^{n-1} \sum_{j=1}^J \mu_j y_{kj} + \sum_{k=1}^{n-1} (\underline{z}u_k^l - \bar{z}v_k^l) \geq 0, \quad l \in [1; n-1], \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k = 0, \quad k \in [1; n-1], l = 0, n, \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k = 0, \quad k, l \in [1; n-1], k \leq l-1, \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^l - c_{rm}^l) - f_k + \sum_{j=1}^J \sigma_j y_{kj} = 0, \quad k, l \in [1; n-1], k \geq l, \\
& b_{rk}^l + c_{rk}^l - g_{rk} = 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], \\
& u_k^l, v_k^l, b_{rk}^l, c_{rk}^l, g_{rk} \geq 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n].
\end{aligned}$$

Proof. The proof is similar to that of Proposition 4.3. □

Henceforth, Problem (4.11) is equivalent to

$$\begin{aligned}
& \inf \quad \alpha \\
& \text{s.t.} \quad \nu_n + \frac{1}{\alpha} \left(f_0^n + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk}^n \right) \leq \tau_n, & n \in \mathcal{G}_{h-1}, \\
& \nu_n + \frac{1}{\alpha_i} \left(f_0^n + \sum_{k=1}^{n-1} \sum_{r=1}^k \epsilon_{rk} g_{rk}^n \right) \leq \tau_n, & n \in \mathcal{I}_i, i \in [1; h-1], \\
& f_0^n + \sum_{k=1}^{n-1} (\underline{z}u_k^{0n} - \bar{z}v_k^{0n}) \geq 0, & n \in [2; N+1], \\
& f_0^n + \nu_n + \sum_{k=1}^{n-1} (\underline{z}u_k^{nn} - \bar{z}v_k^{nn}) \geq 0, & n \in [2; N+1], \\
& f_0^n + \nu_n + x_n - x_l - \sum_{k=l}^{n-1} \sum_{j=1}^J \mu_j y_{kj} + \sum_{k=1}^{n-1} (\underline{z}u_k^{ln} - \bar{z}v_k^{ln}) \geq 0, \\
& & l \in [1; n-1], n \in [2; N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, \\
& & k \in [1; n-1], l = 0, n, n \in [2; N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, \\
& & k, l \in [1; n-1], k \leq l-1, n \in [2; N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1} \sum_{r=1}^k (b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n + \sum_{j=1}^J \sigma_j y_{kj} = 0, \\
& & k, l \in [1; n-1], k \geq l, n \in [2; N+1], \\
& b_{rk}^{ln} + c_{rk}^{ln} - g_{rk}^n = 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], n \in [2; N+1], \\
& u_k^{ln}, v_k^{ln}, b_{rk}^{ln}, c_{rk}^{ln}, g_{rk}^n \geq 0, \quad r, k \in [1; n-1], r \leq k, l \in [0; n], n \in [2; N+1], \\
& \alpha \in (0, 1], \\
& (\boldsymbol{\tau}, \boldsymbol{x}, \mathbf{Y}) \in \mathcal{Y}.
\end{aligned}$$

Given $\alpha \in (0, 1]$, the sequencing and scheduling problem reduces to check the

feasibility of a mixed-integer optimization problem with $N \times J$ binary decision variables, $O(N^4)$ continuous decision variables, and $O(N^4)$ constraints.

4.6 Computational Study

In this section, we carry out three computational studies. In the first study, we investigate the problem of scheduling homogeneous patients, and compare performances under two strategies: (1) lexicographic minimization of DUM (L-DUM) and (2) minimization of total expected delays (TED). The second study explores the performance of appointment scheduling model under distributional ambiguity. In the third study, we solve a sequencing and scheduling problem for two patient types and provide some practical insights. The program is coded in python and run on a Intel Core i7 PC with a 3.40 GHz CPU by calling CPLEX 12 as ILP solver.

4.6.1 Comparison of quality measures

We compare the performance of two appointment system models: the L-DUM model and the TED model, which is formulated as follows:

$$\begin{aligned}
 \min \quad & \sum_{n=2}^{N+1} \mathbb{E}_{\mathbb{P}}(\tilde{w}_n) \\
 \text{s.t.} \quad & \tilde{w}_n = \max \left\{ 0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k \right\}, \quad n \in [2; N+1], \\
 & \tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in [1; N], \\
 & \mathbf{x} \in \mathcal{X}.
 \end{aligned}$$

We consider the case of scheduling seven homogeneous patients who have the same delay thresholds. We assume patients' consultation times are independent and identically distributed with two-point distributions. Hence, we have a number $2^8 = 256$ of scenarios, which could allow us to enumerate all possible realizations, and calculate the exact optimal scheduling decisions. Later on, we will extend to other distributions. We first study in detail an instance and analyze the performance by varying patients' and physician's delay thresholds. Afterwards, we randomly generate 100 instances and investigate their average performances. For each instance, we (a) generate the corresponding parameters for two-point distributions, (b) enumerate all the possible realizations of service time combination, (c) solve the scheduling problem by the L-DUM and the TED strategies, and (d) compute each participant's corresponding delay to summarize the performances.

In the first instance, two-point distribution is specified with realizations 1 and 4, and mean as 2. Total session length is 16. We obtain the scheduling decisions with different thresholds in Table 4.1. We consider four performance

	TED	L-DUM $(\tau_p, \tau_d)^1$					
		(1.5, 1.5)	(2, 2)	(2.5, 2.5)	(3, 3)	(3.5, 3.5)	(4, 4)
Patient 1	0	0	0	0	0	0	0
Patient 2	1	1	1	1	1	1	1
Patient 3	5	3.37	3.37	3.18	2.94	2.74	2.72
Patient 4	9	5.79	5.77	5.68	5.76	5.83	5.57
Patient 5	10	8.38	8.38	8.32	8.17	8.01	8.09
Patient 6	14	10.88	10.88	10.84	10.86	10.88	10.82
Patient 7	15	13.47	13.47	13.45	13.34	13.23	14.82

¹ τ_p : patients' delay threshold; τ_d : physician's delay threshold.

Tab. 4.1: Patients' optimal appointment time under two scheduling methods.

measures: expected delay, frequency of delay over the threshold, standard deviation of delay, and expected delay over the threshold.

	Delay performance of the worst-off participants				Total expected delays
	Expected delay	Frequency of delay over the threshold ¹	Standard deviation of delay	Expected delay over the threshold ²	
L-DUM(1.5,1.5) ³	1.24	56%	1.74	0.57	8.43
TED	2.40	61%	2.26	1.48	6.74
L-DUM(2,2)	1.25	33%	1.73	0.44	8.44
TED	2.40	61%	2.26	1.17	6.74
L-DUM(2.5,2.5)	1.34	33%	1.72	0.32	8.57
TED	2.40	61%	2.26	0.86	6.74
L-DUM(3,3)	1.48	26%	1.71	0.24	8.65
TED	2.40	17%	2.26	0.56	6.74
L-DUM(3.5,3.5)	1.59	11%	1.74	0.20	8.74
TED	2.40	17%	2.26	0.47	6.74
L-DUM(4,4)	1.60	11%	1.81	0.14	8.36
TED	2.40	17%	2.26	0.39	6.74

¹ Frequency of delay over the threshold: $\mathbb{P}(\tilde{w} > \tau)$;

² Expected delay over the threshold: $\mathbb{E}_{\mathbb{P}}((\tilde{w} - \tau)^+)$;

³ L-DUM(τ_p, τ_d).

Tab. 4.2: Delay performance under two scheduling methods (two-point).

Table 4.2 summarizes the delay performance of the worst-off participants (including all patients and the physician). Since the findings are similar, for convenience and clarity, we report the numerical performance for the case with patients' and physician's threshold taking the value of two. In terms of total expected delays, we observe that the TED method performs better than the L-DUM model. However, this performance comes at the price of sacrificing the service levels of some participants. From the fairness perspective, when we pay particular attention to the most discriminated participants, our model makes a significant improvement over the TED model. The maximal average delay reduces from 2.40 to 1.25, and the frequency of delay over the threshold improves from 61% to 33%.

Thenceforth, we study the average performance of 100 randomly generated instances. The parameters determining the two-point distribution

\tilde{s} are specified as $\underline{s} = 3\varphi_1$, $\bar{s} = 3 + 5\varphi_2$, and $\mathbb{P}(\tilde{s} = \bar{s}) = 0.5\varphi_3$, where $\varphi_1, \varphi_2, \varphi_3$ are independently uniformly distributed, $U(0, 1)$. The average service time, μ is therefore determined. Total session length is $L = 6\mu + \bar{s}$. The delay thresholds are set to three levels, namely, low, medium, and high, where $\tau_d(\text{low}) = \tau_p(\text{low}) = \underline{s}$, $\tau_d(\text{medium}) = \tau_p(\text{medium}) = \mu$, and $\tau_d(\text{high}) = \tau_p(\text{high}) = \bar{s}$. For each instance, we calculate the delay performance of the worst-off participants under the L-DUM model, and normalize it by the corresponding performance in the TED model. We summarize the average ratio in Table 4.3. The values less than one favor L-DUM model.

Threshold level	Delay performance of the worst-off participants				Total expected delays
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
Low	0.6813	0.8162	0.8494	0.4794	1.3134
Medium	0.6352	0.6185	0.8464	0.2892	1.31
High	0.7753	0.1886	0.8676	0.0867	1.2956

Tab. 4.3: Average performance analysis of two scheduling methods among 100 instances.

We also test our model using the empirical consultation data collected from the clinics in a local hospital in Singapore from March to May, 2012. The historical data during March and April (802 samples) is considered as the information to make scheduling decisions, while data in May (435 samples) is used for performance testing. The statistics of consultation time are summarized in Table 4.4.

Statistics	Average	Maximum	Minimum	Mean absolute deviation	Standard deviation
minutes	13.84	107	1	6.52	9.41

Tab. 4.4: Statistics of consultation time from empirical data.

Our appointment design problem is to schedule ten patients within 150 minutes session length. The performance derived with similar procedures is listed in Table 4.5, which also manifests our conclusions for two-point distributions.

	Delay performance of the worst-off participants				Total expected delays
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
L-DUM(15,15)	13.37	37%	17.33	4.61	94.88
TED	24.12	63%	18.57	11.21	66.65
L-DUM(25,25)	14.45	16%	17.29	2.95	98.60
TED	24.12	35%	18.57	6.51	66.65
L-DUM(35,35)	15.07	9%	17.25	1.81	107.09
TED	24.12	19%	18.57	3.60	66.65

Tab. 4.5: Delay performance under two scheduling decisions (empirical data).

In general, compared with the TED method, the L-DUM model provides a less discriminating solution that mitigates the unpleasantness of delays in the appointment system.

4.6.2 Distributional ambiguity

In this experiment, we study the performance of the L-DUM model under distributional ambiguity. We schedule seven homogeneous patients and compare the delay performance of the worst-off ones under three scheduling decisions. The first two are derived by both stochastic optimization approach and distributionally robust optimization approach in the L-DUM model. Sampling average approximation is employed for stochastic optimization approach, and the information of bound support, mean, and mean absolute deviation for robust optimization approach is calculated accordingly. The third scheduling

decision is derived from the TED method by using sampling average approximation scheme. Total session length is 7. We consider two types of distributions: uniform distribution $U(0, 2)$ and beta distribution $3 \times \text{Beta}(2, 4)$. Sample size for the L-DUM model and the TED model is 500 and 2000, respectively. The delay performance is listed in Table 4.6 and 4.7.

Approach	Delay performance of the worst-off participants				Total expected delays
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
L-DUMs(1.2,1.2) ¹	0.90	35%	0.86	0.21	5.62
L-DUMr(1.2,1.2)	1.00	40%	0.89	0.24	6.15
TED	1.54	64%	0.82	0.52	3.46
L-DUMs(1.4,1.4)	0.99	29%	0.87	0.16	5.84
L-DUMr(1.4,1.4)	1.02	31%	0.89	0.19	6.26
TED	1.55	55%	0.83	0.41	3.46
L-DUMs(1.6,1.6)	0.95	21%	0.86	0.12	5.75
L-DUMr(1.6,1.6)	1.12	28%	0.91	0.17	6.53
TED	1.54	46%	0.83	0.30	3.46

¹ L-DUMs represents stochastic optimization approach, and L-DUMr represents robust optimization approach.

Tab. 4.6: Delay performance under uniform distribution.

Approach	Delay performance of the worst-off participants				Total expected delays
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
L-DUMs(1.2,1.2)	0.89	28%	0.84	0.18	5.18
L-DUMr(1.2,1.2)	1.00	34%	0.86	0.21	5.79
TED	1.47	58%	0.80	0.45	3.18
L-DUMs(1.4,1.4)	0.93	20%	0.84	0.14	5.29
L-DUMr(1.4,1.4)	1.02	29%	0.86	0.16	5.89
TED	1.46	48%	0.79	0.34	3.18
L-DUMs(1.6,1.6)	0.83	16%	0.84	0.10	5.24
L-DUMr(1.6,1.6)	1.14	26%	0.88	0.14	6.20
TED	1.46	39%	0.79	0.26	3.18

Tab. 4.7: Delay performance under beta distribution.

We observe the performance between stochastic optimization approach

and robust optimization approach in the L-DUM model is very close, and much better than that of the TED method. With the distributional uncertainty set we proposed, the L-DUM model provides a comparatively good performance that is immunized against distributional ambiguity. It is particularly worth mentioning that the computation time for distributional robust optimization approach is relatively short. To solve each minimization problem, stochastic optimization approach requires 44 seconds, while distributional robust optimization approach only requires 8 seconds.

4.6.3 A sequencing and scheduling example

We also investigate the sequencing and scheduling problem with heterogeneous patients. By calculating the optimal solutions, we hope to deliver some useful insights for managers to make decisions in a unified manner. For simplicity, we only consider two patient types: new and repeated patients. Their demographics are collected from real data and shown in Table 4.8, and the information of mean absolute deviation is given as, for $i < k, i, k \in [1; N]$,

$$\epsilon_{ik} = \begin{cases} 1.71, & \forall i = k - 1, \\ 2.20, & \forall i = k - 2, \\ 2.52, & \forall i = k - 3. \end{cases}$$

Type	N_j	μ_j	σ_j	$[\underline{z}, \bar{z}]$
New patient ($j = 1$)	1	18	7	[-2,12]
Repeated patient ($j = 2$)	3	13	6	[-2,12]

Tab. 4.8: Characterization of heterogeneous patients.

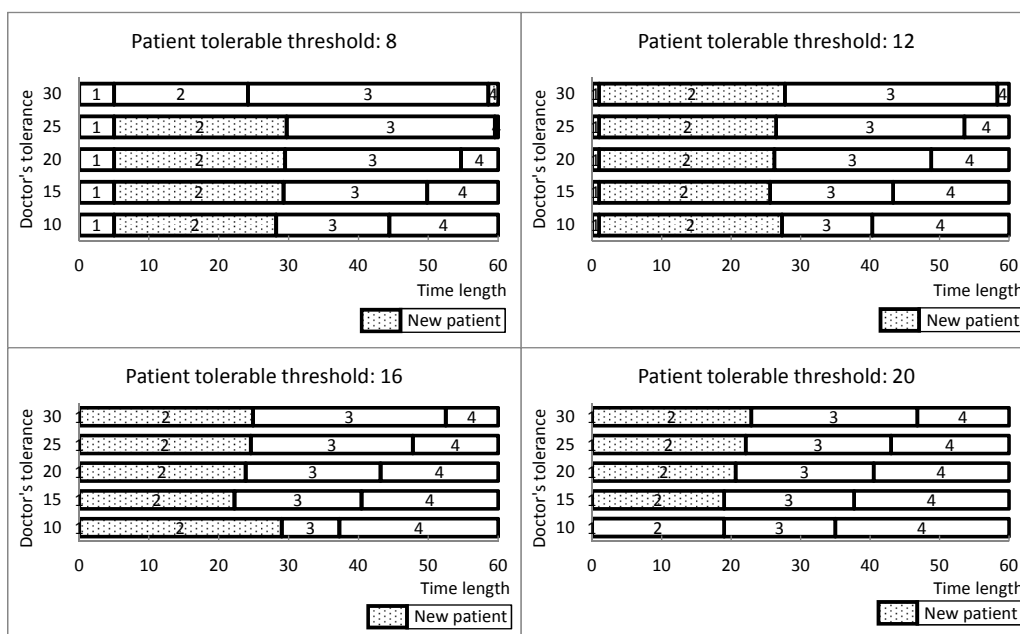


Fig. 4.1: Sequencing and scheduling decisions with various tolerances.

The sequencing and scheduling decisions are illustrated in Figure 4.6.3. For decades, researchers have debated whether to first schedule repeated patients (smallest variance), or new ones (largest variance). Our computational study actually suggests such universal rule may not be optimal, and the decisions may differ as participants' tolerable thresholds vary. For instance, as shown in the first graph of Figure 4.6.3, we generally observe that if the physician's tolerance threshold is low, his/her delay can better be mitigated under L-DUM model if new patient, who may have longer and more uncertain consultation times, is scheduled first. On the other hand, if patients' waiting tolerance is low, for example, in Pediatrics clinic, the L-DUM method will arrange the new patient to arrive at the last position, such that his/her uncertain consultation time will not influence other patients' waiting as they are scheduled to arrive earlier. Our program could easily solve a 10 patients'

sequencing and scheduling problem within seconds.

4.7 *Conclusion*

In this chapter, we study an appointment design problem in the healthcare system. We propose a new quality measure named Delay Unpleasantness Measure (DUM) to describe individual's dissatisfaction attitude towards a waiting process, and then lexicographically minimizes the worst DUM to mitigate the delay and unfairness in the appointment system. The contributions stem from three key aspects:

Firstly, we develop the quality measure DUM to describe individual participant's behavior towards delay process. By taking each participant's tolerance threshold as an exogenous factor, DUM could not only provide an upper bound for the frequency of delay over a threshold, but also account for its intensity.

Secondly, we introduce lexicographic min-max concept to address the issue of fairness in the appointment system. As far as we are aware, this is the first analytical paper taking the fairness subject as the principle aim. Our model allows the decision maker of the appointment system to adjust participants' thresholds based on their needs and in accordance to their service times.

Thirdly, we provide formulation and solution techniques to encompass different information of uncertain service times. When the distributional information is completely known or with historical data, stochastic optimization approach is suggested for solving the problem. In our distributional

uncertainty set, apart from support, and mean, we suggest using mean absolute deviation as descriptive statistics, which could capture the correlation and retain linearity of the nominal problem. The computational study suggests that even if distributions are known, the robust formulations, which are computationally more efficient, can be calibrated to provide competitive solutions to the stochastic programming problem.

5. CONCLUSIONS AND FUTURE RESEARCH

The concept of risk and ambiguity has been extensively studied, however, their applications in service systems are rather limited. Especially, how to develop a tractable model that could describe the distributional ambiguity while also capturing various people's preferences for it is still a thorny issue. In this thesis, we try to solve the above issue collectively, and study two problems in the transportation system and one problem in the healthcare system. Besides the directions of further research listed at the end of each chapter, we could also explore several directions peripheral to the general issue.

- **Description on distributional ambiguity.** As empirical data become increasingly important in assisting decision-making, how to harness these data into the model is an essential question. Probability theory is a popular and classic approach to analyze the uncertainty embedded in the data, but is not necessarily the only one. An alternative is the robust optimization theory, which offers certain advantages over probability theory. I believe that it can be valuable to future research that involves empirical data. Additionally, while various methods can be adopted to describe uncertainties within the robust optimization framework, and various statistics could be estimated or derived from

the empirical data, it is still unclear which one is better than the others. I believe the distributional information that we could use for the optimization model greatly depends on the problem structure. With studies using empirical data, the advantages and disadvantages of different methods can be analyzed.

- **Behavior issues in service systems.** The main difference between the service system and the manufacturing system is that service delivery is labor intensive and cannot be automated easily. Essentially, the main difficulty to study and improve the delivery process in service systems is human beings' behavior issues and concerns, which is interesting to observe but also challenging to analyze. The empirical data could allow us to explore these behavior issues and then develop more meaningful models. For example, in the Emergency Department (ED) in hospitals, doctor's service rate is not a constant, but is first decreasing, then increasing, and then decreasing with the increase of the number of patients in ED. We could analyze the reasons for this behavior, we could also take this behavior in the optimization model. Another example is the fairness issue. In the manufacturing system, machines cannot complain about the unfairness, but human beings can in the service system. Doctors' workload must be balanced in staff scheduling, while patients' waiting times should also be adjusted in scheduling appointments.

BIBLIOGRAPHY

- Abdel-Aty, Mohamed A, Ryuichi Kitamura, Paul P Jovanis. 1995. Investigating effect of travel time variability on route choice using repeated-measurement stated preference data. *Transportation Research Record* (1493) 39–45.
- Abdellaoui, Mohammed, Aurélien Baillon, Laetitia Placido, Peter P Wakker. 2011. The rich domain of uncertainty: source functions and their experimental implementation. *The American Economic Review* **101**(2) 695–723.
- Adulyasak, Yossiri, Patrick Jaillet. 2014. Models and algorithms for stochastic and robust vehicle routing with deadlines .
- Agra, Agostinho, Marielle Christiansen, Rosa Figueiredo, Lars Magnus Hvattum, Michael Poss, Cristina Requejo. 2013. The robust vehicle routing problem with time windows. *Computers & Operations Research* **40**(3) 856–866.
- Aumann, Robert J, Roberto Serrano. 2008. An economic index of riskiness. *Journal of Political Economy* **116**(5) 810–836.
- Averbakh, Igor, Vasilij Lebedev. 2004. Interval data minmax regret network optimization problems. *Discrete Applied Mathematics* **138**(3) 289–301.
- Bailey, Norman T J. 1952. A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society* **14** 185–199.
- Bertsimas, Dimitris, Vivek F Farias, Nikolaos Trichakis. 2011. The price of fairness. *Operations research* **59**(1) 17–31.
- Bertsimas, Dimitris, Melvyn Sim. 2003. Robust discrete optimization and network flows. *Mathematical programming* **98**(1-3) 49–71.
- Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.
- Bertsimas, Dimitris J. 1992. A vehicle routing problem with stochastic demand. *Operations Research* **40**(3) 574–585.
- Bertsimas, Dimitris J, David Simchi-Levi. 1996. A new generation of vehicle routing research: robust algorithms, addressing uncertainty. *Operations Research* **44**(2) 286–304.
- Bosch, Peter M Vanden, Dennis C Dietz. 2000. Minimizing expected waiting in a medical appointment system. *IIE Transactions* **32**(9) 841–848.
- Bosch, Peter M Vanden, Dennis C Dietz. 2001. Scheduling and sequencing arrivals to an appointment system. *Journal of Service Research* **4**(1) 15–25.

- Bossaerts, Peter, Paolo Ghirardato, Serena Guarnaschelli, William R Zame. 2010. Ambiguity in asset markets: theory and experiment. *Review of Financial Studies* **23**(4) 1325–1359.
- Brown, David B., Enrico De Giorgi, Melvyn Sim. 2012. Aspirational preferences and their representation by risk measures. *Management Science* **58**(11) 2095–2113.
- Brown, David B, Melvyn Sim. 2009. Satisficing measures for analysis of risky positions. *Management Science* **55**(1) 71–84.
- Camacho, F, R Anderson, A Safrit, AS Jones, P Hoffmann. 2006. The relationship between patient’s perceived waiting time and office-based practice satisfaction. *North Carolina Medical Journal* 409–413.
- Camerer, Colin, Martin Weber. 1992. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of risk and uncertainty* **5**(4) 325–370.
- Campbell, Ann M, Barrett W Thomas. 2008. Probabilistic traveling salesman problem with deadlines. *Transportation Science* **42**(1) 1–21.
- Cartwright, A, J Windsor. 1992. Outpatients and their doctors. *London: Department of Health Institute for Social Studies in Medical Care* .
- Catanzaro, Daniele, Martine Labbé, Martha Salazar-Neumann. 2011. Reduction approaches for robust shortest path problems. *Computers & operations research* **38**(11) 1610–1619.
- Cayirli, Tugba, Emre Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12**(4) 519–549.
- Chang, Tsung-Sheng, Yat-wah Wan, Wei Tsang Ooi. 2009. A stochastic dynamic traveling salesman problem with hard time windows. *European Journal of Operational Research* **198**(3) 748–759.
- Chau, Chi Kin, Kwang Mong Sim. 2003. The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. *Operations Research Letters* **31**(5) 327–334.
- Chen, Anthony, Zhaowang Ji, Will Recker. 2002. Travel time reliability with risk-sensitive travelers. *Transportation Research Record: Journal of the Transportation Research Board* **1783**(1) 27–33.
- Chen, Bi Yu, William HK Lam, Agachai Sumalee, Zhi-lin Li. 2012. Reliable shortest path finding in stochastic networks with spatial correlated link travel times. *International Journal of Geographical Information Science* **26**(2) 365–386.
- Chen, Wenqing, Melvyn Sim. 2009. Goal-driven optimization. *Operations Research* **57**(2) 342–357.
- Chen, Zengjing, Larry Epstein. 2002. Ambiguity, risk, and asset returns in continuous time. *Econometrica* **70**(4) 1403–1443.

- Cheu, Ruey L, Vladik Kreinovich. 2007. Exponential disutility functions in transportation problems: a new theoretical justification .
- Cho, Nayoung, Samuel Burer, Ann Melissa Campbell. 2010. Modifying soysters model for the symmetric traveling salesman problem with interval travel times .
- Claus, A. 1984. A new formulation for the travelling salesman problem. *SIAM Journal on Algebraic Discrete Methods* **5**(1) 21–25.
- Connors, Richard D, Agachai Sumalee. 2009. A network equilibrium model with travellers perception of stochastic travel times. *Transportation Research Part B: Methodological* **43**(6) 614–624.
- Connors, Richard D, Agachai Sumalee, David P Watling. 2007. Sensitivity analysis of the variable demand probit stochastic user equilibrium with multiple user-classes. *Transportation Research Part B: Methodological* **41**(6) 593–615.
- Cordeau, Jean-François, Gilbert Laporte, Martin WP Savelsbergh, Daniele Vigo. 2006. Vehicle routing. *Transportation, handbooks in operations research and management science* **14** 367–428.
- Cornuéjols, Gérard, Jean Fonlupt, Denis Naddef. 1985. The traveling salesman problem on a graph and some related integer polyhedra. *Mathematical programming* **33**(1) 1–27.
- Correa, José R, Andreas S Schulz, Nicolás E Stier-Moses. 2004. Selfish routing in capacitated networks. *Mathematics of Operations Research* **29**(4) 961–976.
- Correa, José R, Andreas S Schulz, Nicolás E Stier-Moses. 2008. A geometric approach to the price of anarchy in nonatomic congestion games. *Games and Economic Behavior* **64**(2) 457–469.
- Cox, Trevor F, John P Birchall, Henry Wong. 1985. Optimizing the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics* **12**(2) 113–126.
- Dafermos, Stella. 1980. Traffic equilibrium and variational inequalities. *Transportation science* **14**(1) 42–54.
- de Palma, Andre, Nathalie Picard. 2005. Route choice decision under travel time uncertainty. *Transportation Research Part A: Policy and Practice* **39**(4) 295–324.
- Dehlendorff, Christian, Murat Kulahci, Søren Merser, Klaus Kaae Andersen. 2010. Conditional value at risk as a measure for waiting time in simulations of hospital units. *Quality Technology and Quantitative Management* **7**(3) 321–336.
- Denton, Brian, Diwakar Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35**(11) 1003–1016.
- Denton, Brian, James Viapiano, Andrea Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science* **10**(1) 13–24.

- Dow, James, Sergio Ribeiro da Costa Werlang. 1992. Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica* 197–204.
- Eiger, Amir, Pitu B Mirchandani, Hossein Soroush. 1985. Path preferences and optimal paths in probabilistic networks. *Transportation Science* **19**(1) 75–84.
- Ellsberg, Daniel. 1961. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics* **75**(4) 643–669.
- Epstein, Larry G, Martin Schneider. 2008. Ambiguity, information quality, and asset pricing. *The Journal of Finance* **63**(1) 197–228.
- Facchinei, Francisco, Jong-Shi Pang. 2003. *Finite-dimensional variational inequalities and complementarity problems*, vol. 1. Springer.
- Fan, YY, RE Kalaba, JE Moore II. 2005. Arriving on time. *Journal of Optimization Theory and Applications* **127**(3) 497–513.
- Föllmer, Hans, Alexander Schied. 2011. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, Berlin, Germany.
- Frank, H. 1969. Shortest paths in probabilistic graphs. *Operations Research* **17**(4) 583–599.
- Frank, Marguerite, Philip Wolfe. 1956. An algorithm for quadratic programming. *Naval research logistics quarterly* **3**(1-2) 95–110.
- Gabrel, Virginie, Cécile Murat, Lei Wu. 2013. New models for the robust shortest path problem: complexity, resolution and generalization. *Annals of Operations Research* **207**(1) 97–120.
- Ghirardato, Paolo, Fabio Maccheroni, Massimo Marinacci. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* **118**(2) 133–173.
- Gilboa, Itzhak, Andrew W Postlewaite, David Schmeidler. 2008. Probability and uncertainty in economic modeling. *The Journal of Economic Perspectives* 173–188.
- Gilboa, Itzhak, David Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* **18**(2) 141–153.
- Green, Linda V, Sergei Savin. 2008. Reducing delays for medical appointments: a queueing approach. *Operations Research* **56**(6) 1526–1538.
- Guidolin, Massimo, Francesca Rinaldi. 2013. Ambiguity in asset pricing and portfolio choice: a review of the literature. *Theory and decision* **74**(2) 183–217.
- Gupta, Diwakar. 2007. Surgical suites' operations management. *Production and Operations Management* **16**(6) 689–700.
- Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: challenges and opportunities. *IIE transactions* **40**(9) 800–819.
- Hall, NG, Z Long, J Qi, M Sim. 2014. Managing underperformance risk in project portfolio selection. Tech. rep., Working paper.

- Häme, Lauri, Harri Hakula. 2013. Dynamic journeying under uncertainty. *European Journal of Operational Research* **225**(3) 455–471.
- Han, Deren, Hong K Lo, Jie Sun, Hai Yang. 2008. The toll effect on price of anarchy when costs are nonlinear and asymmetric. *European Journal of Operational Research* **186**(1) 300–316.
- Han, Deren, Jie Sun, Marcus Ang. 2014. New bounds for the price of anarchy under nonlinear and asymmetric costs. *Optimization* **63**(2) 271–284.
- Harper, PR, HM Gamlin. 2003. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum* **25**(2) 207–222.
- Hassin, Refael, Sharon Mendel. 2008. Scheduling arrivals to queues: a single-server model with no-shows. *Management Science* **54**(3) 565–572.
- Hill, C Jeanne, Kishwar Joonas. 2006. The impact of unacceptable wait time on health care patients' attitudes and actions. *Health marketing quarterly* **23**(2) 69–87.
- Hsu, Ming, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, Colin F Camerer. 2005. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**(5754) 1680–1683.
- Huang, Xiao-Ming. 1994. Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Research* **7**(1) 2–8.
- Hurwicz, Leonid. 1951. Some specification problems and applications to econometric models. *Econometrica* **19**(3) 343–44.
- Isermann, H. 1982. Linear lexicographic optimization. *Operations-Research-Spektrum* **4**(4) 223–228.
- Isii, K. 1963. On the sharpness of chebyshev-type inequalities. *Annals of the Institute of Statistical Mathematics* (12) 185–197.
- Jaillet, Patrick. 1988. A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research* **36**(6) 929–936.
- Jaillet, Patrick, A Odoni. 1988. The probabilistic vehicle routing problem. *Vehicle routing: methods and studies. North Holland, Amsterdam* .
- Jula, Hossein, Maged Dessouky, Petros A Ioannou. 2006. Truck route planning in nonstationary stochastic networks with time windows at customer locations. *Intelligent Transportation Systems, IEEE Transactions on* **7**(1) 51–62.
- Kaas, Rob, Marc Goovaerts, Jan Dhaene, Michel Denuit. 2001. *Modern actuarial risk theory*, vol. 328. Springer.
- Karaşan, OE, MC Pinar, H Yaman. 2001. The robust shortest path problem with interval data.
- Kenyon, Astrid S, David P Morton. 2003. Stochastic vehicle routing with random travel times. *Transportation Science* **37**(1) 69–82.

- Khachiyan, LG. 1989. The problem of calculating the volume of a polyhedron is enumerably hard. *Russian Mathematical Surveys* **44**(3) 199–200.
- Klassen, Kenneth J, Thomas R Rohleder. 1996. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management* **14**(2) 83–101.
- Knight, Frank H. 1921. *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston, MA.
- Kong, Qingxia, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* **61**(3) 711–726.
- Kosuch, Stefanie, Abdel Lissier. 2010. Stochastic shortest path problem with delay excess penalty. *Electronic Notes in Discrete Mathematics* **36** 511–518.
- Koutsoupias, Elias, Christos Papadimitriou. 2009. Worst-case equilibria. *Computer science review* **3**(2) 65–69.
- Kouvelis, Panos, Gang Yu. 1997. *Robust Discrete Optimization and Its Applications*, vol. 14. Springer.
- Laporte, Gilbert. 2010. A concise guide to the traveling salesman problem. *Journal of the Operational Research Society* **61**(1) 35–40.
- Laporte, Gilbert, François Louveaux, Hélène Mercure. 1992. The vehicle routing problem with stochastic travel times. *Transportation science* **26**(3) 161–170.
- Lee, Chungmok, Kyungsik Lee, Sungsoo Park. 2012. Robust vehicle routing problem with deadlines and travel time/demand uncertainty. *Journal of the Operational Research Society* **63**(9) 1294–1306.
- Letchford, Adam N, Saeideh D Nasiri, Dirk Oliver Theis. 2013. Compact formulations of the steiner traveling salesman problem and related problems. *European Journal of Operational Research* **228**(1) 83–92.
- Li, Xiangyong, Peng Tian, Stephen CH Leung. 2010. Vehicle routing problems with time windows and stochastic travel and service times: models and algorithm. *International Journal of Production Economics* **125**(1) 137–145.
- Lo, Hong K, XW Luo, Barbara WY Siu. 2006. Degradable transport network: travel time budget of travelers with heterogeneous risk aversion. *Transportation Research Part B: Methodological* **40**(9) 792–806.
- Loui, Ronald Prescott. 1983. Optimal paths in graphs with stochastic or multidimensional weights. *Communications of the ACM* **26**(9) 670–676.
- Maccheroni, Fabio, Massimo Marinacci, Aldo Rustichini. 2006. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* **74**(6) 1447–1498.
- Mak, Ho-Yin, Ying Rong, Jiawei Zhang. 2013. Appointment scheduling with limited distributional information. Available at SSRN 2317332 .
- Markowitz, Harry. 1959. *Portfolio Selection: Efficient Diversification of Investments*. New York: John Wiley & Sons, Inc.

- Mazmanyany, Lilit, Dan Trietsch, KR Baker. 2009. Stochastic traveling salesperson models with safety time .
- McCarthy, K, HM McGee, CA O’Boyle. 2000. Outpatient clinic waiting times and non-attendance as indicators of quality. *Psychology, health & medicine* **5**(3) 287–293.
- Mirchandani, Pitu, Hossein Soroush. 1987. Generalized traffic equilibrium with probabilistic travel times and perceptions. *Transportation Science* **21**(3) 133–152.
- Mirchandani, Pitu B. 1976. Shortest distance and reliability of probabilistic networks. *Computers & Operations Research* **3**(4) 347–355.
- Mittal, Shashi, Sebastian Stiller. 2011. Robust appointment scheduling. *Proceedings of the MSOM Annual Conference*.
- Mondschein, Susana V, Gabriel Y Weintraub. 2003. Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Management* **12**(2) 266–286.
- Montemanni, Roberto, János Barta, Monaldo Mastrolilli, Luca Maria Gambardella. 2007. The robust traveling salesman problem with interval data. *Transportation Science* **41**(3) 366–381.
- Montemanni, Roberto, Luca Maria Gambardella, Alberto V Donati. 2004. A branch and bound algorithm for the robust shortest path problem with interval data. *Operations Research Letters* **32**(3) 225–232.
- Moschis, G P, D N Bellinger. 2003. What influences the mature customer? *Marketing Health Care Servicest* **23** 16–21.
- Mukerji, Sujoy, Jean-Marc Tallon. 2003. An overview of economic applications of david schmeidler’s models of decision making under uncertainty. *Department of Economics, University of Oxford* .
- Murthy, Ishwar, Sumit Sarkar. 1998. Stochastic shortest path problems with piecewise-linear concave utility functions. *Management Science* **44** 125–136.
- Muthukrishnan, AV, Luc Wathieu, Alison Jing Xu. 2009. Ambiguity aversion and the preference for established brands. *Management Science* **55**(12) 1933–1941.
- Nagurney, Anna. 1998. *Network economics: A variational inequality approach*, vol. 10. Springer.
- Nagurney, Anna, June Dong. 2002. A multiclass, multicriteria traffic network equilibrium model with elastic demand. *Transportation Research Part B: Methodological* **36**(5) 445–469.
- Nash, John. 1951. Non-cooperative games. *Annals of mathematics* 286–295.
- Natarajan, Karthik, Melvyn Sim, Joline Uichanco. 2010. Tractable robust expected utility and risk models for portfolio optimization. *Mathematical Finance* **20**(4) 695–731.
- Nemirovski, Arkadi, Alexander Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization* **17**(4) 969–996.

- Nie, Yu Marco, Xing Wu. 2009. Shortest path problem considering on-time arrival probability. *Transportation Research Part B: Methodological* **43**(6) 597–613.
- Nie, Yu Marco, Xing Wu, Tito Homem-de Mello. 2012. Optimal path problems with second-order stochastic dominance constraints. *Networks and Spatial Economics* **12**(4) 561–587.
- Nikolova, Evdokia, Jonathan A Kelner, Matthew Brand, Michael Mitzenmacher. 2006. Stochastic shortest paths via quasi-convex maximization. *Algorithms-ESA 2006*. Springer, 552–563.
- Nikolova, Evdokia Velinova. 2009. Strategic algorithms. Ph.D. thesis, Massachusetts Institute of Technology.
- Noland, Robert B, John W Polak. 2002. Travel time variability: a review of theoretical and empirical issues. *Transport Reviews* **22**(1) 39–54.
- Ogoryczak, W, Michal Pióro, Artur Tomaszewski. 2005. Telecommunications network design and max-min optimization problem. *Journal of telecommunications and information technology* 43–56.
- Öncan, Temel, I Kuban Altinel, Gilbert Laporte. 2009. A comparative analysis of several asymmetric traveling salesman problem formulations. *Computers & Operations Research* **36**(3) 637–654.
- Ordóñez, Fernando, Nicolás E Stier-Moses. 2010. Wardrop equilibria with risk-averse users. *Transportation Science* **44**(1) 63–86.
- Patrick, Jonathan, Anisa Aubin. 2013. Models and methods for improving patient access. *Handbook of Healthcare Operations Management*. Springer, 403–420.
- Perakis, Georgia. 2007. The “price of anarchy” under nonlinear and asymmetric costs. *Mathematics of Operations Research* **32**(3) 614–628.
- Robinson, Lawrence W, Rachel R Chen. 2003. Scheduling doctors’ appointments: optimal and empirically-based heuristic policies. *IIE Transactions* **35**(3) 295–307.
- Rockafellar, R Tyrrell. 2007. Coherent approaches to risk in optimization under uncertainty. *Tutorials in operations research, INFORMS* .
- Rockafellar, R Tyrrell, Stanislav Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk* **2** 21–42.
- Roughgarden, Tim. 2003. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences* **67**(2) 341–364.
- Roughgarden, Tim, Éva Tardos. 2002. How bad is selfish routing? *Journal of the ACM* **49**(2) 236–259.
- Russell, RA, TL Urban. 2008. Vehicle routing with soft time windows and erlang travel times. *Journal of the Operational Research Society* **59**(9) 1220–1228.
- Sen, Amartya Kumar, James E Foster. 1997. *On economic inequality*. Oxford University Press.

- Siu, Barbara WY, Hong K Lo. 2008. Doubly uncertain transportation network: degradable capacity and stochastic demand. *European Journal of Operational Research* **191**(1) 166–181.
- Smith, MJ. 1979. The existence, uniqueness and stability of traffic equilibria. *Transportation Research Part B: Methodological* **13**(4) 295–304.
- Souyris, Sebastián, Cristián E Cortés, Fernando Ordóñez, Andres Weintraub. 2013. A robust optimization approach to dispatching technicians under stochastic service times. *Optimization Letters* **7**(7) 1549–1568.
- Sungur, Ilgaz. 2007. The robust vehicle routing problem. Ph.D. thesis, University of Southern California.
- Swait, Joffre, Tülin Erdem. 2007. Brand effects on choice and choice set formation under uncertainty. *Marketing Science* **26**(5) 679–697.
- Taş, Duygu, Nico Dellaert, Tom Van Woensel, Ton De Kok. 2013. Vehicle routing problem with stochastic travel times including soft time windows and service costs. *Computers & Operations Research* **40**(1) 214–224.
- Toh, Loke Shuet, Cheong Wai Sern. 2011. Patient waiting time as a key performance indicator at orthodontic specialist clinics in selangor. *Malaysian Journal of Public Health Medicine* **11** 60–69.
- Toth, Paolo, Daniele Vigo. 2001. *The vehicle routing problem*. SIAM.
- Uchida, Takashi, Yasunori Iida. 1993. Risk assignment: a new traffic assignment model considering the risk of travel time variation. *Transportation and Traffic Theory* 89–105.
- Wakker, Peter P. 2010. *Prospect theory: For risk and ambiguity*, vol. 44. Cambridge University Press Cambridge.
- Wakker, PP. 2008. Uncertainty. the new palgrave: A dictionary of economics.
- Wang, P Patrick. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics* **40**(3) 345–360.
- Wang, P Patrick. 1999. Sequencing and scheduling n customers for a stochastic server. *European journal of Operational Research* **119**(3) 729–738.
- Wardrop, John Glen. 1952. Road paper. some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions*, vol. 1. Thomas Telford, 325–362.
- Watling, David. 2006. User equilibrium traffic network assignment with stochastic travel times and late arrival penalty. *European Journal of Operational Research* **175**(3) 1539–1556.
- Weiss, Elliott N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE transactions* **22**(2) 143–150.
- Xiao, Ying, Krishnaiyan Thulasiraman, Xi Fang, Dejun Yang, Guoliang Xue. 2012. Computing a most probable delay constrained path: Np-hardness and approximation schemes. *IEEE Transactions on Computers* **61**(5) 738–744.

-
- Yang, Hai, Hai-Jun Huang. 2004. The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. *Transportation Research Part B: Methodological* **38**(1) 1–15.
- Yang, Kum Khiong, Mun Ling Lau, Ser Aik Quek. 1998. A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics* **45**(3) 313–326.
- Yin, Yafeng, Hitoshi Ieda. 2001. Assessing performance reliability of road networks under nonrecurrent congestion. *Transportation Research Record: Journal of the Transportation Research Board* **1771**(1) 148–155.
- Yin, Yafeng, William HK Lam, Hitoshi Ieda. 2004. New technology and the modeling of risk-taking behavior in congested road networks. *Transportation Research Part C: Emerging Technologies* **12**(3) 171–192.
- Young, H Peyton. 1995. *Equity: in theory and practice*. Princeton University Press.
- Yu, Gang, Jian Yang. 1998. On the robust shortest path problem. *Computers & Operations Research* **25**(6) 457–468.
- Zhu, Shushang, Masao Fukushima. 2009. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research* **57**(5) 1155–1168.
- Zhu, Zhecheng, Heng Bee Hoon, Teow Kiok Liang. 2011. Reducing consultation waiting time and overtime in outpatient clinic: challenges and solutions. *Management Engineering for Effective Healthcare Delivery: Principles and Applications* 229.