PATTERN SEEKING AND SERVICE SYSTEM DESIGN DRIVEN BY FACT-BASED DECISION-MAKING

KONG, QINGXIA

NATIONAL UNIVERSITY OF SINGAPORE

2012

PATTERN SEEKING AND SERVICE SYSTEM DESIGN DRIVEN BY FACT-BASED DECISION-MAKING

KONG, QINGXIA

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF DECISION SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENT

As I am writing the acknowledgement section of my PhD dissertation, I reflect upon the arduous journey of the past 5 years. It is drawing to an end (finally), but only to mark the beginning of a new journey. I am fortunate to have received exceptional support from many outstanding people, who made my journey a joyful and rewarding one.

First and foremost, I thank my supervisor, Prof. Chung-Piaw Teo, for his constant support, motivation, guidance and training. He guided me through difficult times in research and also in life, and has been a wonderful friend and mentor who has always amazed me with his integrity, wisdom, accessability, and above all his compassion and patience. Without his support, I would not have had this chance to write an acknowledgement for my PhD dissertation. I hope to be a supervisor just like him, to groom intellectual children of my own, and pass the baton onto them.

I would like to especially thank my coauthors who have contributed in significant ways to this work. Prof. Mabel Chou and Prof. Huan Zheng guided me on the first project on small number phenomenon. Mr. Zhichao Zheng worked closely with me on several projects and taught me the mathematics of co-positive cones. Prof. Nicolas Lambert guided me through the third project on Gambler's Fallacy. I learned so much about academic writ-

ing skills from Prof. Qiang Fu when I worked with him on a research on Joint Venture. Without their contributions and guidance, this dissertation would not be possible. I would also like to thank my committee member, Prof. Songfa Zhong, for his guidance and feedbacks on my research. I would particularly like to thank Prof. Yaozhong Wu, for the intriguing lessons in behavioral economics. Besides, I would like to Prof. Melvyn Sim, Prof. Jie Sun, Prof. Hanqin Zhang, Prof. John Buzacott, and Prof. Paul Zipkin, for being a source of inspiration and motivation through this long journey.

I also gratefully acknowledge the support that I have received from the staff in the PhD office and Decision Sciences department: Cheow Loo, Hamidah, Siew Geok, Dorothy, Chwee Ming and (another) Siew Geok.

My PhD journey would not have been wonderful without the amazing set of friends I have. I would like to thank especially Jin and Zhuoyu for always being there for me. They have been a constant source of support and motivation for me. I thank my friend Masia, Maggie, Cissy, Yu-chen, Shangtai for being my sounding boards, and for listening to my frustrations. I thank my land-lady Fiona who took care of me for the past 5 years. I thank my friend Robert, Jane, David who helped me to navigate the job market. Special mention must go to my life teacher and guide Pujyashri P. Rajagopalachari, and friends in the meditation group: Akash, Neelu, Juli, Nitin, Rajesh, Rakesh Pratima etc., for keeping me going through the PhD with a peaceful and loving mind. I particularly want to thank Shirish and Anu for being my most faithful friends.

I am fortunate to have steadfast support from my family during the course of my PhD studies. I would like to thank my loving and lovely father

Mr. Xianlai Kong, who plays basketball with me when I visit, who sings songs with me over the phone and who is always eager to learn the latest progress in my research, for being a constant source of strength. He helped me to understand my research work even better. I thank my sister Qinghui and brother-in-law Peng for hosting and providing a cozy hideout whenever I visit China. I thank my niece little Shuyan for being the sunshine in my life. I thank my younger brother Qingkang for growing up to be a responsible and reliable young man. I thank my 95-year old grandma who is a source of security and strength for me. My appreciation particularly goes to my uncle Dr. Xiantao Kong for encouraging me to pursue a PhD when I was unsure and lost.

Finally, no words can express my gratitude to my Mom for being the best mom ever. She never let me believe that there is anything I could not do. She encouraged me to follow my dreams. She taught me to be brave, to be persistent, to never give up. The values she passed onto me benefit me all through my life. No matter where she is, she is always a source of inspiration and motivation. To her this thesis belongs to.

25 March 2012 Qingxia Kong

Santiago, Chile

CONTENTS

ABSTRACT

The rapid development of Internet Technology has enabled convenient access to large amount of business transactions data and has advocated the era of "fact-based" decision-making. An integrated "fact-based" decision-making process consists of three phases : data collecting, pattern-seeking, and performance. Motivated by the importance of understanding the integrated "fact-based" decision process, I investigate three fundamental questions with focuses on different phases of the process.

- Are there simple and universal patterns when people make choices?
- How to utilize limited (insufficient) data to design a robust service system to ensure good performance under all possible situations?
- How system design affects behavioral patterns?

This dissertation adopts a multi-theoretic and multi-disciplinary approach to offer fresh insights on "fact-based" decision-making. I investigate three topics to tackle these questions. The first topic models a universal choice rule: the small number phenomenon. The second topic solves a robust appointment scheduling problem using a parsimonious set of information on the consultation durations. The third topic explores how information complexity and pattern recognition affect people's perception of randomness.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Nowadays, companies are confronting unprecedented challenges amid macroeconomics downturn, financial crisis, forces of globalization, fiercely competitive market, ever-demanding investors and customers. The basis for competition lies in the capability to make smarter decisions and design/execute high-performance business processes. Smarter decisions create value through higher efficiency or greater effectiveness, like reducing cost, increasing sales, lowering the risk, better allocation of resources and enhancing customer satisfaction. (cf.Davenport and Harris (2006), Laursen and Thorlund (2010) etc.) A thorough study on the performance of a range of companies in various industries from 1965 to 1995 concluded that a series of good decisions laid solid foundations for many companies in the study (Collins (2001)).

The rapid development of IT system has enabled convenient access to large amount of business transactions data and has advocated the era of "fact-based" decision-making. The financial crisis revealed the price of ungrounded assumptions. Good decisions flow from a consistent and thorough effort to confront "the brutal facts" (Collins (2001)). Extensive explorations of the data helps a company to identify the trends and patterns of its customers' behaviors, make predictive models and take prompt actions (e.g., re-designing its service system) to increase its profitability. The famous "beer and diapers"

phenomenon provides an excellent example on the benefits of data utilization. A US supermarket chain found a strong correlation between sales of beer and diapers through an investigation of their check-out receipts. This pattern was then readily exploited by displaying the two products side by side on the shelves. It would have been difficult to hypothesize through common sense the correlations of sales between two seemingly unrelated products without checking the sales data. Yet most companies sit on a priceless mountain of data, but fail to utilize it in any meaningful way (Davenport et al. (2010)). Therefore, the capacity to exploit the data and make fact-based, real-time decisions is crucial for a company to gain a competitive edge.

As a decision support system (DSS), Business Analytics (BA) is defined as "*the continuous and iterative exploration and investigation of past business performance to gain insights and drive business planning*". BA enables companies to aggressively leverage huge, noisy and messy data in key business decisions and promptly react to changing conditions around them. By delivering insights gleaned from data about customers, suppliers, operations, performance, and more, BA gives companies the tools to tackle complex business problem (SAS white paper). A survey conducted by Davenport and Harris (2006) confirmed a significant correlation between the adoption of BA system and high performance of companies.

Adoptions of BA systems can now be seen in many business domains. In health-care, for example, computer-based program is used to keep track of patients' electronic medical records (EMRs). EMR holds great promise to enhance efficiency, reduce costs and errors, and make records available anytime, anywhere. Analytical tools can be used to determine the optimal

operating room usage by specialty, scheduling appointment, explore clinical outcomes and risk tolerance to improve overall quality of patient care. Likewise, a BA system can create value for hospitality and gaming industry. From predicting demand to improving inventory and staffing level, to determining the optimal mix of tables at a restaurant, to simulating front desk before a redesign, BA can help with better understand operations to make decisions that increase revenues and profits.

An integrated fact-based decision-making process consists of three phases (see Figure 1.1): data, patterns, and performance.

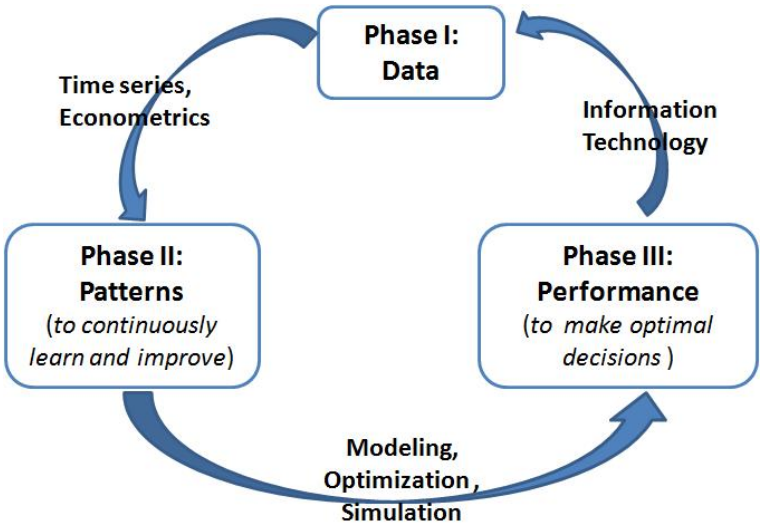

Fig. 1.1: The close loop of the integrated process in fact-based decision-making

Collecting relevant data to analyze a system and achieve certain target is the first and foremost phase of fact-based decision-making. Due to the fact most data are obnoxiously dirty and messy, one challenge of this phase is to acquire the *right* data, and if not, how to obtain cleaner data.

When (re-)designing a service system, prediction on customers' behaviors is crucial. Identifying patterns of a system is the starting point for improving decision-making and enhancing customer experience. In the second phase, patterns are to be sought from data, using various statistical tools (e.g. time series analysis, econometrics), to tell what is more likely to happen in the future. Once the future becomes more predictable, we can make effective plans accordingly or conduct more thorough analysis to optimize the outcomes. For example, Progressive Insure found that financial irresponsibility predicts reckless driving. They thus use customers' credit reports data as input to optimally provide offers and price different customers.

Phase III lies in effective performances–making best decisions and taking best actions. In this phrase, predictive model are to be built by incorporating outcomes in the second phrase, and approaches like optimization or simulation are to be used to find the best of all reasonable solutions to achieve certain target, like minimizing cost, maximizing revenue or best allocating limited resources etc.

An integrated "fact-based" decision-making is an iterative, and recursive process. So, it is important to return to phase I to close the loop. Once the system is optimally re-designed, the performance of the new system should be constantly monitored and new data should be collected and analyzed to identify new patterns. This iterative process not only keep track of performance of the new system but also enable agile responses to changing conditions.

Driven by the wave of the "fact-based" decision-making, in this dissertation, I study three fundamental questions with different focuses on the

integrated process:

- Whether there exists a simple and universal choice rule that depicts people's choice pattern? (with focus on Phase II)?
- How could we leverage limited (insufficient) data to glean patterns and design robust service system ensuring good performance under all possible situations? (with focus on Phase III)?
- How would different system designs affect people's behavioral patterns, especially, perception of randomness when uncertainty is involved in the system (with focus on the close loop)?

These questions are tackled by adopting a mix of research methodologies like conic programming, lab experiments, econometric model, simulation etc., and explore profound implications of the results in different business domains. Adopting multi-theoretic and multi-disciplinary approach, the contribution of this dissertation lies in bringing out fresh insights and opening new avenue for future research in the integrated fact-based decision-making.

1.1 Structure of the Dissertation

This dissertation is structured as three separate topics associated with different focuses on the integrated fact-based decision-making process. The three topics have separate theoretical underpinnings and implications in practice, and have contributed to the three fundamental question raised.

The first topic investigates an universal pattern revealed when people make choices: The small number phenomenon. I quantify this phenomenon

and examine its relation to the classical Benford's law. I use this connection to develop a choice model, and explore its implications in setting the appropriate sales limit in fixed-odd lottery games. The second topic starts with historical data on patients' consultations and develops a convex programming approach to solve a general class of robust appointment scheduling problem in a single server facility, using a "representative" worst case distribution matching the prescribed mean and covariance estimates of the service durations of the patients, to determine the optimal schedule and sequence. The third topic is to explore in depth how the deeply embedded human nature– pattern seeking– shapes the perceptions of randomness. I create a simple economic setting in which a sequence of random outcomes are generated and build a Bayesian Updating model to explore conditions under which the Hot-hand Fallacy appears. I collect two sets of field data from gaming industry to provide a solid foundation and verification of the insights gained from the model. These results have important implications in problem gambling, risk management, and lab experiments where random outcomes are involved.

2. TOPIC 1: SMALL NUMBER PHENOMENON

2.1 Abstract

In fixed-odds numbers games, the prizes and the odds of winning are known at the time of placement of the wager. Both players and operators are subject to the vagaries of luck in such games. Most game operators limit their liability exposure by imposing a sales limit on the bets received for each bet type, at the risk of losing the rejected bets to the underground operators. This raises a question - how should the game operator set the appropriate sales limit?

We argue that the choice of the sales limit is intimately related to the ways players select numbers to bet on in the games. There are ample empirical evidence suggesting that players do not choose all numbers with equal probability, but have a tendency to bet on (small) numbers that are closely related to events around them (e.g., birth dates, addresses, etc.). To the best of our knowledge, this is the first paper to quantify this phenomenon and examine its relation to the classical Benford's law. We use this connection to develop a choice model, and propose a method to set the appropriate sales limit in these games.

2.2 Introduction

Gambling is probably one of the oldest inventions in human history. In the ancient past, it was often organized around a fight between tribesmen. This ancient game of skill has proliferated into the different sports betting games that are now commonly played in many countries. Gambling can also take the form of a game of chance where the winners are determined via an external event - a toss of bones, whoever draws the short straw, and so on. In fact, such games are now routinely played at a national or state level, where players bet on which prize-winning numbers will be drawn using mechanical devices (cf. Lafaille and Simonis (2005)).

In many countries, number lotteries have become a popular source of revenue for governments. In 2005, the Hong Kong Jockey Club paid close to HK$12.4 billion to the SAR government in betting duties and profits tax. This is close to 8.6% of the total tax collected by the Inland Revenue Department of Hong Kong that year. In the same year, the Singapore government took in S$1.05 billion from the gaming operators in betting duties, against a total tax revenue of close to S$17 billion. These games are also popular in the West. A recent survey by the licensed operator of the UK National Lottery, Camelot, found that as many as 69% of the adult population in Britain played the lottery in 2005-06.[1] On the other hand, while there is no national lottery in the US, similar games are now played in more than 30 states in the country.

There are many ways in which number lottery games can be organized.

[1] The BBC news article on this can be accessed at http://news.bbc.co.uk/1/hi/uk/6174648.stm

In a parimutuel game, the players bet on the outcome of the draw of (random) prize-winning numbers, with the winner drawing a fixed portion of the total amount of bets received. The payout for the winners in such games depends on the total amount of bets received and the total number of winners. On the other hand, in a fixed-odds game, the winner receives a fixed payout for each winning wager, and the total payout for the winner is proportional to the amount of the wagers he makes in the game. For a fixed-odds game, the prize is fixed for each ticket, and hence the return for each player does not depend on how other players bet. However, the game operator now bears the risk of paying out a large sum in prizes if a very popular number is chosen as the winning number. Most game operators handle the risk exposure issue in their fixed-odds numbers games by imposing a liability limit on the sales of each number - all future bets on those numbers with accumulated sales hitting the limit will be rejected. This raises an associated question - how should a game operator set the liability limit? Note that this issue is particularly important to legalized game operator as a large chunk of their sales will have to be returned to the government as tax revenues at the end of each year. This prevents the operator from building up a large reserve to absorb the exposure risk.

Teo and Leong (2002) used the Markowitz model to argue that it is reasonable to use a common sales limit for all numbers/bet-types in the game. They exploited the design of a popular four-digit numbers game played in Singapore to demonstrate the benefits of risk pooling in the liability limits management system. However, they focused mainly on internal risk control mechanism and did not study the impact of external demand distribution

(i.e., how players select numbers) on the game. Interestingly, this turns out to have a huge impact on the effectiveness of the risk control mechanism.

2.2.1 Benford's Law and Number Selection in Fixed-Odds Numbers Game

There are numerous studies in the gaming literature on lottery numbers selection among the players. One group of studies (e.g., Simon (1999), Henze (1997), Haigh (1997), and Ziemba et al. (1986)) focuses on the lotto games (where players compete to pick, for instance, 6 winning numbers out of 45), and has revealed many interesting behavioral patterns showing how the players select their numbers. The most striking conclusion from these studies is that the players do not select their numbers randomly; that is, not all numbers are chosen with equal likelihood, and there is a tendency to select "auspicious" numbers (for instance, the number 7 is routinely chosen by players in the game in the UK; numbers below 31 are more popular than numbers above 31, etc.). Table 2.1 shows the proportion of bets received on each number from 1-45 (ranked from highest to lowest proportions), in a 1996 powerball game played in the UK (Tijms (2007)).

Another group of studies (Chernoff (1999), Halpern and Devereaux (1989)) focuses on the numbers game (where the players compete to pick the winning 3-digit or 4-digit number), which is also known as Pick-3 or Pick-4 in many states in the US. Halpern and Devereaux (1989) also observed that players in Pennsylvania favor small numbers in the 3-digit numbers game, where the winning number is drawn randomly from among the numbers 000 to 999. They observed that the bet volumes decrease rapidly from numbers in the

rank	number	proportion	rank	number	proportion	rank	number	proportion
1	7	0.036	16	25	0.026	31	20	0.019
2	9	0.033	17	15	0.025	32	33	0.018
3	5	0.033	18	21	0.025	33	35	0.016
4	3	0.033	19	17	0.024	34	32	0.015
5	11	0.031	20	16	0.024	35	40	0.015
6	12	0.030	21	26	0.024	36	34	0.014
7	8	0.030	22	14	0.024	37	42	0.014
8	4	0.029	23	24	0.024	38	36	0.013
9	10	0.029	24	27	0.023	39	41	0.013
10	2	0.029	25	19	0.023	40	44	0.012
11	6	0.028	26	30	0.023	41	39	0.012
12	23	0.027	27	18	0.022	42	45	0.012
13	13	0.026	28	31	0.02	43	43	0.012
14	22	0.026	29	28	0.02	44	38	0.011
15	1	0.026	30	29	0.02	45	37	0.01

Tab. 2.1: Popularity of the 45 numbers in a 6/45 powerball game.

100s to 400s, then slowly to the 900s. A similar phenomenon was also observed by Chernoff (1999) in his study of the 4-digit game in Massachusetts.

The sales data received on a particular draw in Pennsylvania was clearly presented in Halpern and Devereaux (1989), which allows us to quantify this phenomenon in the numbers games. Figure 2.1 shows the empirical distribution of the sum-of-three-digits statistics of the numbers chosen by the players in the Pennsylvania game. We compare the empirical distribution against the base case where all the 3-digit numbers are selected with equal probability (i.e., the uniform-choice model). Interestingly, the empirical distribution indicates a leftward shift from the base-case distribution, indicating a general preference for smaller digits in the number selections.

This empirical evidence indeed suggests that players favor small numbers. We call this the **small-number** phenomenon in the numbers game.

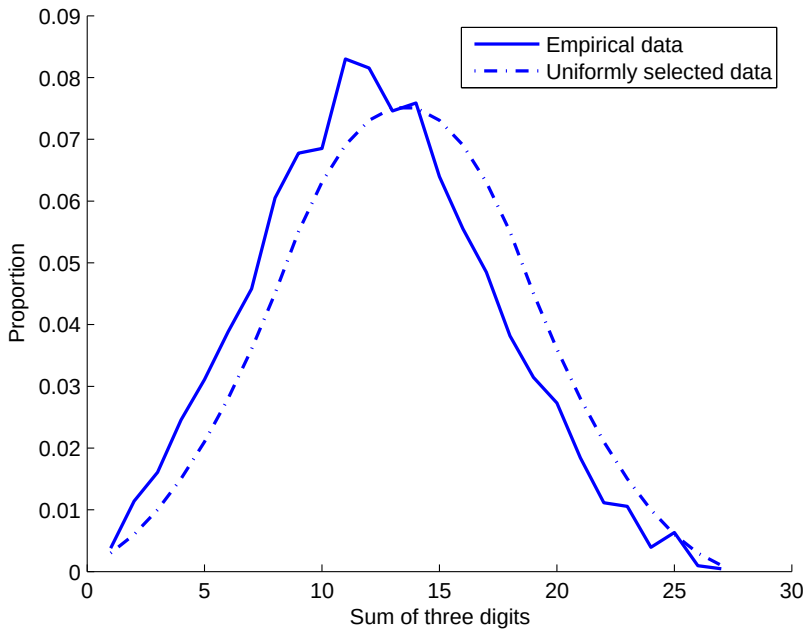

Fig. 2.1: Distribution of the sum-of-three-digits statistics in the 3-digit numbers game.

2.2.2 Explanations

Studies investigating cognitions of lottery ticket purchasers showed that people failed to recognize that each number on a ticket is independent of the others. For example, Ladouceur et al. (1996) showed that adults were more likely to select the "most random" perceived combinations, although in reality each ticket was as likely to win as the others. In addition, Langer (1975) asserted that factors in a chance situation which are typically associated with skill situations (such as choice, competition, and passive or active involvement) cause an individual to believe they have control over a situation that is completely governed by chance. Ladouceur et al. (1996) found that individuals who selected their own lottery ticket requested a larger sum of money in order to relinquish or sell back their ticket than those individuals who were randomly given a ticket (machine generated numbers). They concluded that participants who were able to select their own lottery ticket perceived their ticket as having a greater chance of winning and, as a result, assigned a higher monetary value to the ticket than individuals in the no-choice condition. Erroneous beliefs commonly held by adult gamblers were also identified in Hardoon et al. (1997) and Ladouceur and Walker (1996). Herman et al. (1998) studied the question as to when children's gambling behavior resembles that of adults. They showed that as children get older they are more specific in their beliefs that certain types of tickets are more likely to win than others.

There are a few explanations for the small-number phenomenon in lottery games. As stated in many studies (e.g., Halpern and Devereaux (1989),

Simon (1999), etc.), a large proportion of players tend to select numbers associated with special dates (e.g., birthdays, anniversaries, etc.), meaningful numbers (e.g., phone numbers, car numbers, address numbers, etc.), and special events (e.g., accidents and murders), and these numbers tend to start with smaller digits (e.g., there are only 12 months in a year, so that the numbers 1-12 should be more popular than the numbers 13-45 in many 6/45 lotto games). Another explanation put forth by researchers is the observation that human beings simply can not choose numbers in a uniform manner. Loetscher and Brugger (2007) demonstrated using experimental methods that there is indeed a cognitive bias towards the selection of small numbers by human beings, even when they are told to select numbers "randomly." In one of their studies, a total of 488 subjects were told to "name a sequence of digits with each digit chosen from 1 to 6 as randomly as possible," and they found a surplus of small digits (1, 2, or 3) in all their experiments.

These studies, unfortunately, offer only anecdotal evidence (through surveys and interviews) and rudimentary explanations for the existence of the small-number phenomenon, and do not provide an analytical framework to quantify and model this phenomenon.

Another factor influencing the choice of numbers is superstitious beliefs, widely held by players of lottery games. In Chinese culture, certain numbers are believed by some to be lucky (or unlucky) based on the similarity of their pronunciation to that of certain Chinese words. For instance, Chinese people usually associate the digit 8 with prosperity, and thus numbers containing the digit 8 are normally more popular[2]. On the other hand, the number 4

[2] In fact, China Mobile's Jiangxi branch held an auction to sell a "lucky" phone number

is considered unlucky in many cultures in Asia, since it sounds like the word for "death" in spoken Chinese. Such beliefs concerning lucky and unlucky numbers tend to affect the popularity of certain numbers in the lottery game, leading to uneven distribution of the wagers on the different numbers.

2.2.3 Modeling Empirical Data

There have been several attempts to model the biases in the choice model of players. Simon (1999) considered the impact of the "lucky-number" biases and developed a model to approximate the distribution of the number of times that a combination would be chosen in the UK national lottery game. This provided a more accurate choice model for the UK lottery game, and fitted the data better than the uniform-choice model. Stern and Cover (1989) obtained the choice probability for each number in a lotto game, from the empirical marginal frequencies, by solving a related entropy-minimization problem. Ziemba et al. (1986) used regression methods and empirical data to estimate the popularity of each number combination in the lotto game. Haigh (1997) used choice probabilities directly on a set of numbers to estimate the popularity of the number combinations. Unfortunately, all these methods took the empirical data as given and focused merely on finding a better choice model to fit the empirical data. Thus, these methods did not exploit the existence of the small-number phenomenon in their modeling approaches, nor did they try to quantify this phenomenon.

recently, and one such number - with six consecutive eights - was sold for RMB 44,000!

2.2.4 Contributions

In this topic, we investigate the small-number phenomenon in the numbers games (rather than the lotto games) and use it to address the liability limits management problem. Our contributions in this topic are as follows:

- We quantify the small-number phenomenon through a curious fact observed by Newcomb (1881) and independently by Benford (1938). Interestingly, while the classical Benford's law captures the proportion of bets on the first significant digit reasonably well, it fails to account for the self-replicating nature of the empirical data beyond the first significant digit. By carefully modeling the ways players compose the digits in the numbers game, we refine Benford's law to develop an alternate consumer-choice model for different bet types using a handful of parameters only. Surprisingly, this parsimonious choice model is already able to capture some of the most important characteristics of the data in the numbers game.

- We examine the consequences of the small-number phenomenon on the prize liability performance of the game operator. In particular, our analysis suggests that it will be fruitful for the operators to pursue strategies to reduce the effect of the small-number phenomenon; that is, to promote or encourage players to choose numbers randomly. On the other hand, we show that the debate on whether a sales limit should be imposed on the game can be examined from the demand side - if numbers are selected in a uniform manner, then it may be futile to impose any sales limit since the performance is very sensitive to the

total sales revenues of the game; that is, with a slight change in total sales revenues, the operator may swing from a situation with all numbers hitting the sales limit to a situation where all bets are accepted. Unless the total sales revenue can be accurately forecasted, it will be difficult to set the right sales limit in this environment. The small-number phenomenon in the choice process actually helps to stabilize this relationship between the total sales revenues and the proportion of numbers sold out. The imposition of a sales limit is thus more effective in such environment.

2.3 Modeling the Small-Number Phenomenon

Classical economic theory assumes that players behave like rational agents, and make decisions based on utility-maximization reasoning. As the returns from each number combination are identical, these players have no specific preference for any particular number and thus all numbers are selected with equal probability. We call these "Type 1" players.

However, recent empirical studies show that agents are not always seeking utility maximization in their decision making since framing, loss aversion, decision biases etc. can have major effects on players' decisions. To understand the small-number phenomenon, we need to augment the classical approach by incorporating the behavior of agents who pick their "lucky" numbers (arising from events in their daily life, or through superstitious beliefs) using reasoning which cannot be captured by any economic model. These players are superstitious, and have a general tendency to avoid betting

on certain digits.[3]. We call these the "Type 2" players.

We also assume that each player bets \$1 on each number chosen.

Definition 1. Let β_B and β_N denote the proportions of type 2 and type 1 players respectively, with

$$\beta_B + \beta_N = 1. \tag{2.1}$$

The challenge in our problem is to estimate the proportions of type 1 and type 2 agents in the population of players based on the aggregate sales data. To this end, we need to have a better understanding on how type 2 agents choose their numbers. As these "lucky" numbers are normally selected from data series arising in the daily life of these type 2 agents, we will exploit a curious property associated with these natural numbers.

2.3.1 Benford's law

Newcomb (1881) observed that the first few pages of books of logarithms were more worn than the last few and inferred that there might be more numbers starting with 1 or 2 than starting with larger numbers. Newcomb then drew a counter-intuitive conclusion that the first significant digits (i.e., first non-zero digits) of many data series in nature are not evenly distributed as expected, but follow a logarithmic law. Almost 50 years later, independently of Newcomb, Benford (1938) noticed the same phenomenon for categories of naturally occurring numerical data; for example, areas of rivers, atomic weights, numbers from Reader's Digest, and so on. He then came to the same

[3] Interestingly, our data suggests that players in Pennsylvania have an aversion to the digit 2, but favor digits 7 and 8.

conclusion, now known as Benford's law, which Newcomb had arrived at so many years previously. Both Newcomb (1881) and Benford (1938) proposed that the probability that a number has the first significant digit D_1 in a set [1..9], is given by

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right), \text{ for all } \ d_1 \in [1..9]. \tag{2.2}$$

Let D_i denote the i^{th} significant digit of a number. Hill (1995a) extended the above observation to a general version of Benford's law: for all $d_1 \in [1..9]$, and $d_k \in [0..9]$ for $k > 1$,

$$P(D_1...D_i = d_1...d_i) = \log_{10}\left(1 + \frac{1}{\sum_{j=1}^{i} d_j \times 10^{i-j}}\right). \tag{2.3}$$

Empirical evidence of Benford's law has appeared in a wide range of data; for example, stock index (Ley (1996)), income tax (Nigrini (1996)), mathematical series (Whitney (1972)), and so on. Benford (1938) analyzed the underlying causes of this logarithmic phenomenon using a heuristic argument. Other mathematicians and statisticians have offered various explanations for this phenomenon. Raimi (1976) gave a review of some of the more intuitive explanations. It wasn't until 1995 that Hill (1995a) provided a formal rigorous proof that Benford's law is the only probability distribution which is scale-invariant and base-invariant. Using modern mathematical probability theory, and the scale- and base-invariant proofs, Hill rigorously demonstrated that the "distribution of distributions" given by random samples taken from a wide variety of different distributions in fact satisfies Benford's law (cf. Hill

(1998)).

One of the main applications of Benford's law is in fraud detection, under the hypothesis that fabricating data which conform to Benford's law is difficult. Recent empirical evidence shows that true accounting data sets conform very closely to Benford's law (Thomas (1989), Nigrini (1996)). On the other hand, fabricated data rarely conform to Benford's law. Therefore, digital analysis based on Benford's law has been proposed as a new tool for fraud detection in recent years. Another application of Benford's law has been in the design of computers. Peter Schatte (1988) devised rules that optimize computer data storage, by allocating disk space according to the proportions dictated by Benford's law, based on the assumption that input request satisfy Benford's law. Furthermore, both Varian (1972) and Hill (1995b) suggested using Benford's law as a test of the reasonableness of forecasts of a proposed model. If real life data follows Benford's law, it seems reasonable to assume that a good mathematical model should also do so.

In this topic, we add to this growing list of applications by showing that Benford's law can be used to capture the number selection behavior of type 2 agents in our model.

2.3.2 Choice Model for Type 2 Agents

WLOG, we will develop the choice model based on a 3D game, using the sales data published earlier in Halpern and Devereaux (1989). We have cross validated this model on other empirical data in several other number games, but unfortunately, due to the sensitivity of the data, we could not report

the results here. For ease of exposition, we ignore the bets received for the number 000 from subsequent analysis; that is, we assume that none of the players will place a bet on the number 000. Using this assumption, the betting profiles of the type 1 players are drawn from a uniform distribution where all the numbers from 001 to 999 will have an equal chance of being selected. We focus next on the betting behavior of the type 2 players.

To ensure that the number selected has exactly 3 digits, we assume that the type 2 player may choose to compose a 3-digit number by padding the number he or she has chosen with leading zeros.[4]

Definition 2. Let γ_i denote the proportion of type 2 players who are betting on numbers with i significant digits.

By definition,

$$\sum_{i=1}^{3} \gamma_i = 1. \tag{2.4}$$

We first state a very simple consumer-choice model, where the classical Benford's law holds directly for the 3-digit numbers played.

Assumption 1. We assume that the type 2 player will choose to play the 3-digit number $d_1 \dots d_i$ ($d_1 > 0$), with $3 - i$ leading zeros, with probability

$$\gamma_i \log_{10}(1 + \frac{1}{d_1 \times 10^{i-1} + ... + d_i}). \tag{2.5}$$

[4] Note that this simplifying assumption may not hold in general, as some players may pad the numbers with trailing zeros, and some may simply duplicate the numbers to reach a 3-digit number. Halpern and Devereaux (1989) mentioned that triplets like 111 or 888 are very popular in the Pick-3 game in Pennsylvania. Unfortunately, it does not appear possible to incorporate such features into the model, without sacrificing the simplicity and tractability of the calibration model.

Note that this is none other than the classical Benford's law, except that we weigh it with a factor γ_i to account for the proportion of players who bet with i significant digits.

It is now easy to prove the following proposition.

Proposition 1. Under Assumption 1, the expected proportion of the betting volume on a 3-digit number with first significant digit i, denoted by $E[S(i)]$, is

$$E[S(i)] = \beta_B \times \log_{10}\left(1 + \frac{1}{i}\right) + \beta_N \times \frac{1}{9}, \text{ for all } i = 1, 2, \ldots 9. \tag{2.6}$$

Note that $E[S(i)]$ does not depend on γ_j. We can thus use this property to calibrate the value of β_B and β_N, by looking at the proportion of bets received for each significant digit. In the 3D data from Pennsylvania, the proportion of the type 2 and type 1 players are estimated to be 39.58% ($\beta_B = 0.3958$) and 60.42% ($\beta_N = 0.6042$), according to the least square model. We plot next the expected proportion of the first significant digit, given by the optimal parameter values, as shown in Figure 2.2, along with the empirical proportion. The prediction from Benford's law captures the general trend in the empirical data, although we observe a general preference for first significant digit 3, 7 and 8 among the players, whereas the digit 2 has lower frequency than expected. Although we can refine our model to build in these biases into the model, we opted not to do so because such preferences

do not appear to be universal across cultures.

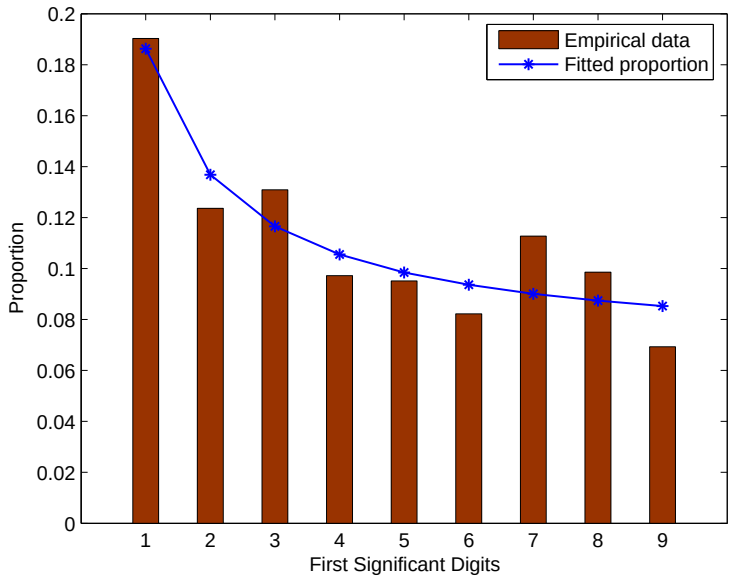

Fig. 2.2: Fitted proportion by the model.

While the simple model in Assumption 1 captures the behavior concerning the first significant digit rather accurately, we examine its ability to track the proportion of betting volume for the first two significant digits. We plot next the expected proportion of bets received and the empirical averages in Figure 2.3. Interestingly, our model is able to capture the declining popularity in the 3-digit numbers, as the first two significant digits grow from 10 to 99. This provides a partial explanation for the small-number phenomenon often observed in the games. Unfortunately, it could not explain the fact that the small-number phenomenon exists even in each decile (sub-block), as shown in Figure 2.3.

To understand the choice preferences beyond the first significant digit,

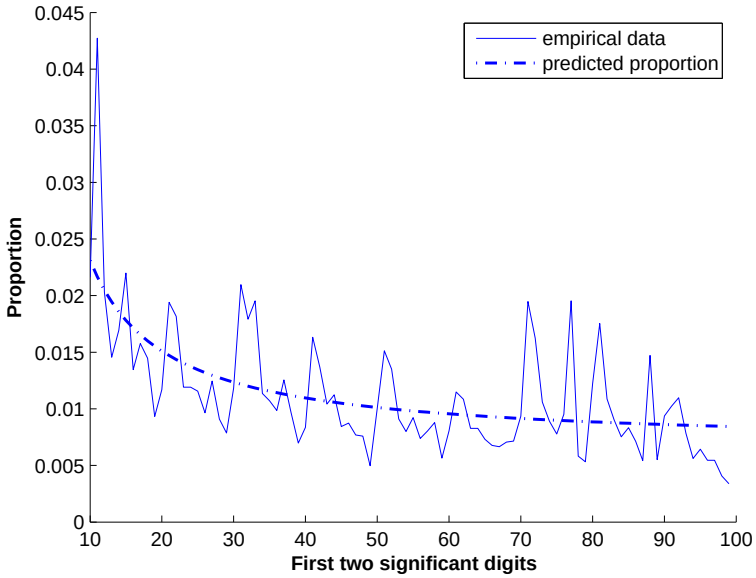

Fig. 2.3: Fitted proportion by the model for the first two significant digits.

we need to model an important characteristic in the way players compose the 3-digit numbers in the game. One such common strategy is to combine data from two different series to form a 3-digit number. For example, the number 246 could come from the 24th day of the month of June, or it could come from the address being level 6 of block unit number 24. The previous model assumes that the 3-digit numbers come from a single data series and hence fails to capture this switching behavior.

We notice that the probability distribution in our first assumption can be written in a different form:

$$\begin{aligned}
&\gamma_i \log_{10}(1+\frac{1}{d_1\times 10^{i-1}+\ldots+d_i}) \\
&= \gamma_i \log_{10}(1+\frac{1}{d_1})\frac{\log_{10}(1+\frac{1}{d_1\times 10+d_2})}{\log_{10}(1+\frac{1}{d_1})}\cdots\frac{\log_{10}(1+\frac{1}{d_1\times 10^{i-1}+\ldots+d_i})}{\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\ldots+d_{i-1}})}.
\end{aligned}$$

Here, γ_i represents the probability that the type 2 player will pick a number with i significant digits. $\log_{10}(1+\frac{1}{d_1\times 10^{i-1}+\ldots+d_i})/\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\ldots+d_{i-1}})$ represents the probability that the ith digit is d_i, given that the first $i-1$ digits are $d_1\ldots d_{i-1}$. To model the switching behavior, we refine the recursive approach in the following way:

- As before, $\log_{10}(1+\frac{1}{d_1})$ represents the probability that the first digit is d_1.
- Let
$$\frac{\log_{10}(1+\frac{1}{d_1\times 10^{i-1}+\ldots+d_i})}{\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\ldots+d_{i-1}})+\lambda}$$
denote the probability that the player will continue to generate the i^{th} digit d_i as if it comes from the same data series as the first $i-1$ digits, with parameter $\lambda>0$. Note that in this way, the players will switch to a different data series with a non-negative probability
$$\frac{\lambda}{\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\ldots+d_{i-1}})+\lambda}.$$
- If the players switch to a different data series, let p_0 denote the probability that they will switch to the digit "0." Otherwise, they will switch to digit i, with $i\in\{1,\ldots,9\}$, with probability $(1-p_0)\log_{10}(1+\frac{1}{i})$.

With a slight abuse of notation, we can write

$$\log_{10}(1+\frac{1}{0}):=\frac{p_0}{1-p_0},\ \text{and}\quad \lambda:=\frac{q}{1-q}.$$

We can now model the switching behavior in the 3-digit game in the following way:

Assumption 2. We assume that the type 2 player will choose to play the 3-digit number $d_1 \dots d_i$ ($d_1 > 0$), with $3 - i$ leading zeros, with probability

$$\gamma_i \log_{10}(1 + \frac{1}{d_1}) \frac{(1-q)\log_{10}(1 + \frac{1}{d_1 \times 10 + d_2}) + q(1-p_0)\log_{10}(1+\frac{1}{d_2})}{(1-q)\log_{10}(1+\frac{1}{d_1}) + q} \times \dots$$
$$\times \frac{(1-q)\log_{10}(1 + \frac{1}{d_1 \times 10^{i-1} + \dots + d_i}) + q(1-p_0)\log_{10}(1+\frac{1}{d_i})}{(1-q)\log_{10}(1 + \frac{1}{d_1 \times 10^{i-2} + \dots + d_{i-1}}) + q}.$$

In this way, we can interpret the parameters as follows.

Definition 3. Let q denote the switching probability. Let p_0 denote the probability that the digit will be switched to 0.

Let $E[S(i,j)]$ denote the expected proportion of bets with first two significant digits i and j respectively.

Proposition 2. Under Assumption (2),

$$E[S(i)] = \beta_B \times \log_{10}\left(1 + \frac{1}{i}\right) + \beta_N \times \frac{1}{9}, \text{ for all } i = 1, 2, \dots 9;$$

$$E[(S(i,j)] = \beta_B \log_{10}(1+\frac{1}{i})\left(\frac{(1-q)\log_{10}(1 + \frac{1}{i \times 10 + j}) + q(1-p_0)\log_{10}(1+\frac{1}{j})}{(1-q)\log_{10}(1+\frac{1}{i}) + q}\right) + \beta_N\left(\frac{1}{90}\right).$$

Note that the expected proportion of first significant digits remains unchanged under both assumptions. The parameters under Assumption 2 are

calibrated to be $q = 0.9105, p_0 = 0.1054$, to best fit the empirical data under the least square model.

We compare the expected frequencies of first two significant digits with those in empirical data respectively in Figure 2.4. The frequencies generated

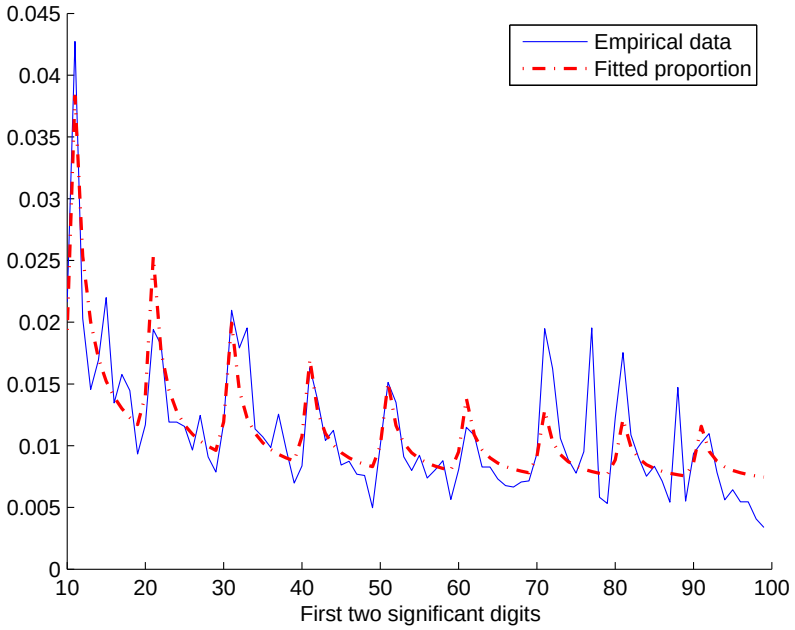

Fig. 2.4: Fitted proportion for the first two significant digits.

from this model closely fit the frequencies of the empirical data. More interestingly, this model is able to capture the small-number phenomenon in the second significant digit of the data series.

Note that so far the parameters γ_j did not feature in the analysis. This arises because we have fixed the number of significant digits. To complete our specification of the choice model, we need to estimate the values of these parameters. Let $\hat{\gamma}_j$ denote the sample average of the proportion of bets with

exactly $3-j$ leading zeros. We use $\hat{\gamma}_j$ to obtain an unbiased estimator of γ_j, using the following relationship:

$$\hat{\beta}_B \gamma_j + \hat{\beta}_N \frac{9 \times 10^{j-1}}{999} = \hat{\gamma}_j, \quad \text{for } j \geq 1. \tag{2.7}$$

2.3.3 Model Validation

We show next that the choice model under Assumption 2 proposed in the earlier section has the ability to track some of the most important characteristics of the betting data in the 3D game.

We first estimate the behavior for the sum-of-digits statistic, using data simulated according to Assumption 1 (with the calibrated parameters). We compare it with the empirical data (after removing the betting volumes on the 3-digit number 000). Figure 2.5 depicts the distributions of the sum-of-digits in three data series: the actual data, simulated data from our choice model, and the uniform-choice model.

The estimation of 39.58% type 2 and 60.42% type 1 players in the population seems right, as it captures the magnitude of the leftward shift in the empirical data reasonably well. Also, note that our choice model does not account for the superstitious beliefs observed in the empirical data (players generally avoid 2 and prefer 7 and 8). This partially explains why the proportions from our model are higher for smaller sum-of-digits (from 3 to 7) and lower for sum-of-digits around 15.

We track the performance of another statistic - the numbers of 3-digit

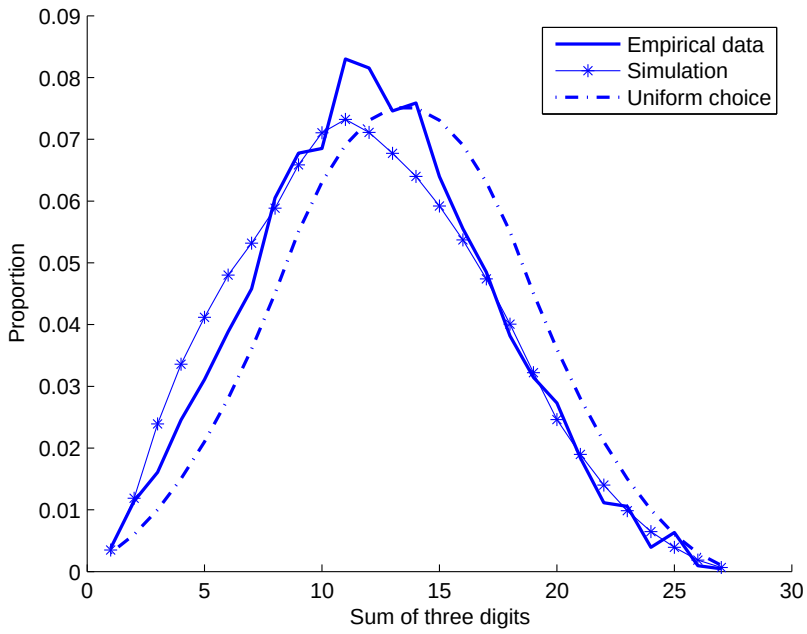

Fig. 2.5: Distributions of sum-of-digits in empirical data, simulated data with Assumption 2, and uniform choice

bets attaining a certain betting volume. Figure 2.6 shows the number of 3-digit numbers in the game with a given betting volume (specified in the horizontal axis). The distribution obtained from the uniform-choice model follows a binomial distribution, and centers mainly around the mean. This yields a poor fit for the empirical data. The distribution obtained from our choice model clearly has a better fit.

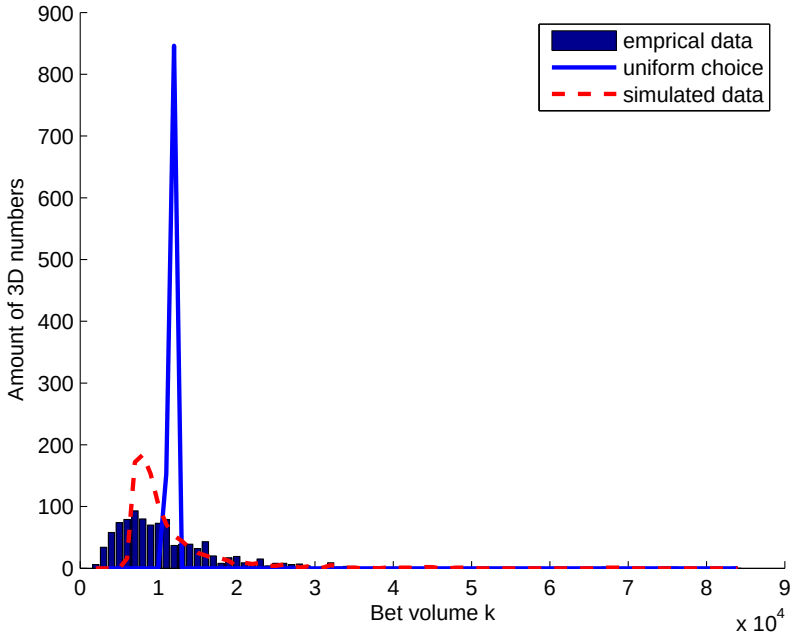

Fig. 2.6: Betting volume of actual data, simulated data with Assumption 2, and uniform choice.

2.4 Applications

The small-number phenomenon clearly has important implications for the operational risk management of the game. The numbers picked by the type 2 players introduce variability and skewness to the distribution of bets on the 3-digit numbers. The winning numbers, on the other hand, are randomly (i.e., uniformly) rolled out by a mechanical device, which implies that the hot numbers are chosen with the same probability as other numbers. The mismatch between the winning number distribution and the betting volume distribution leads to a significant operational risk: the operators may face a substantial payout if a popular number happens to be picked as the winning number! This is a phenomenon which often worried the game operators. In Quebec, according to Lafaille and Simonis (2005), "the first drawing caused a prize liability well in excess of the amount received in sales." Fortunately, "over the long run it all evened out and the projected prize percentage was achieved."

We show in this section that the small-number phenomenon plays a significant role in the large volatility of prize liability experienced by many operators in the game. We further exploit this observation to propose a method to help determine the sales limit in these games.

2.4.1 Volatility of prize liability

We examine the impact on the prize payout volatility by the proportion of type 2 players in the population of players. We compare the variability of payout in the 3D game, as we increase the proportion of type 2 players from

0% to 39.58% in the choice model (both with Assumption 1 and Assumption 2). For this study, we assume that the sales limit is higher than demands, so that all bets are accepted.

Consider a game with a prize P and N players, each betting \$1 on a number drawn from a respective distribution. Let $X_{\beta_B}(n)$ denote the amount of bets received on the number n when the proportion of type 2 players is equal to β_B. When the winning number for that prize is drawn uniformly among the 999 numbers (from 001 to 999, as we have ruled out the bets on the number 000), the expected payout in our choice model is simply

$$\frac{P}{999}\sum_{n=1}^{999} E(X_{\beta_B}(n)) = \frac{P}{999} \times N.$$

The second moment of the payout is

$$P^2\left(\frac{\sum_{n=1}^{999} E(X_{\beta_B}^2(n))}{999}\right).$$

Hence, the variance of payout is

$$P^2\left(\frac{\sum_{n=1}^{999} E(X_{\beta_B}^2(n))}{999} - \frac{N^2}{999^2}\right).$$

If all the N players choose their numbers independently, $X_{\beta_B}(n) \sim Bi(N, p_{\beta_B}(n))$, where $p_{\beta_B}(n)$ denote the probability that number n is picked in our choice model, given that the proportion of type 2 players is β_B. Hence,

$$E(X_{\beta_B}(n)^2) = N^2 p_{\beta_B}^2(n) + N p_{\beta_B}(n)(1 - p_{\beta_B}(n)).$$

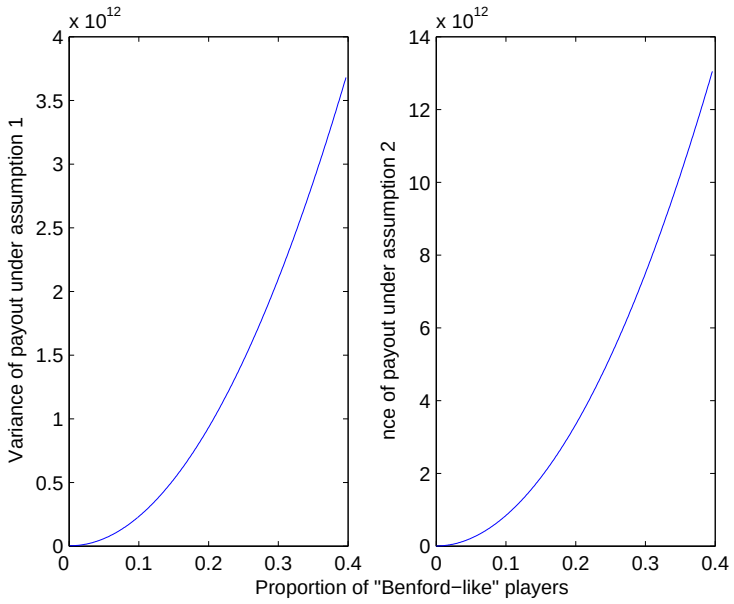

Fig. 2.7: Variance of payout as the proportion of type 2 players increases.

We can thus analytically compare the variance of the payout, under different values of β_B. As shown in Figure 2.7, under both assumptions, the variability of payout is increasing as the proportion of type 2 players increases. When β_B is equal to 0, that is, the demand is evenly distributed, the variance of payout is only 0.003×10^{12}. When β_B increases to 39.58%, the variance of payout under Assumption 1 is 3.6794×10^{12}, about 1216 times higher than that of the uniform-choice model. Since Assumption 2 captures more of the volatility of the data, the variance of payout is 13.049×10^{12} in this model, 4313 times bigger than that of the uniform-choice model.

Note that our choice model only includes 39.58% type 2 players and does not account for other random effects such as date or month effect. So, the

volatility of actual prize payout should be even worse. For the 3D game in Pennsylvania, we conclude that the standard deviation of the prize payout can be reduced by 65 times if the proportion of the type 2 agents (β_B) reduces to 0.

2.4.2 Liability Limit

In the rest of this section, we use the small-number phenomenon to set the appropriate liability limit for the 3D game. Let D_n denote the (random) demand of a 3-digit number n. The distribution of D_n depends on the proportion of type 1 and type 2 players in the game. Let C denote the corresponding sales limit. Let S_n denote the accepted sales for number n; i.e.,

$$S_n = \min(D_n, C).$$

Note that

$$E[S_n] = C \cdot P(D_n > C) + E(D_n | D_n \leq C) \cdot P(D_n \leq C).$$

Let $R(S_1, \ldots, S_N)$ denote the "risk exposure" when sales for the N numbers are given by $(S_1, \ldots, S_N)$. There are several ways to model the risk measure $R(\cdot)$, and it generally depends on the distribution of the winning numbers drawn.

Suppose the expected return given \$1 bet is r. We use the mean-risk trade-off to model the utility function of the game operator. The expected

utility function of the game operator is thus given by

$$r\sum_{n=1}^{N} E[S_n] - \lambda E\big\{R(S_1,\ldots,S_N)\big\},$$

where λ is an exogenous penalty term for risk exposure.

We can find C by solving the following maximizing problem:

$$\max_{C>0}\ r\sum_{n=1}^{N}[C\cdot P(D_n > C)+E(D_n|D_n \le C)\cdot P(D_n \le C)]-\lambda E\big\{R(\min(D_1,C),\ldots,\min(D_N,C))\big\}.$$

It can be easily shown that the objective function is convex. Thus, according to the first order condition, the optimal liability limit C satisfies:

$$\sum_{n=1}^{N} P(D_n > C) = \frac{\lambda}{r}E\left[\frac{\partial R(\min(D_1,C),\ldots,\min(D_N,C))}{\partial C}\right]. \tag{2.8}$$

Note that the left hand side corresponds to the expected number of hot numbers, i.e., the expected number of bet types reaching the sales limit in the draw. The sales limit can be set by merely choosing a sales limit C to control the number of hot numbers.

Suppose the total bets collected are to the value of $\$N$, and the cut-off limit is $\$C$ for each number. We next estimate the expected number of hot numbers (i.e., the numbers with betting volumes hitting the liability limit).

We define an indicator function $Y_{\beta_B}(n)$ as follows:

$$Y_{\beta_B}(n) = \begin{cases} 1 & \text{if } X_{\beta_B}(n) \ge C; \\ 0 & \text{otherwise.} \end{cases}$$

The expected number of hot numbers with liability limit $\$C$ is

$$\begin{aligned}\boldsymbol{E}\left(\sum_{n=1}^{999} Y_{\beta_B}(n)\right) &= \sum_{n=1}^{999} \boldsymbol{P}\left(X_{\beta_B}(n) \geq C\right) \\ &= \sum_{n=1}^{999}\left(1-\sum_{i=0}^{C-1}\binom{N}{i}(p_{\beta_B}(n))^i(1-p_{\beta_B}(n))^{N-i}\right)\end{aligned}$$

Note we can use a normal distribution $N(Np_{\beta_B}(n), \sqrt{Np_{\beta_B}(n)(1-p_{\beta_B}(n))})$ to approximate the binomial distribution $Bi(N, p_{\beta_B}(n))$, if N is large enough. Hence, we have

$$\boldsymbol{E}\left(\sum_{n=1}^{999} Y_{\beta_B}(n)\right) = \sum_{n=1}^{999}\left(1-\Phi\left(\frac{C-Np_{\beta_B}(n)}{\sqrt{Np_{\beta_B}(n)(1-p_{\beta_B}(n))}}\right)\right).$$

We can thus analytically compute the expected number of hot numbers given a liability limit $\$C$, and compare the results using different liability limits. Figure 2.8 shows the expected number of hot numbers under different liability limits in the case that 39.58% players are type 2 and 60.42% players are type 1, and in the ideal case that all players are type 1 agents.

In the ideal case, because all numbers are selected with equal probability, the concentration of measure phenomenon kicks in and the expected number of hot numbers goes through a phase transition - dropping sharply from 999 (all sold out) to 0 (none sold out) for a narrow range of sales limit. This is most evident from Figure 2.8: when the total sales is \$9M, \$10M, \$11M, and \$12M respectively, the expected number of hot numbers drops sharply to zero when the liability limit is around \$10000, \$11000, \$12000, and \$13000

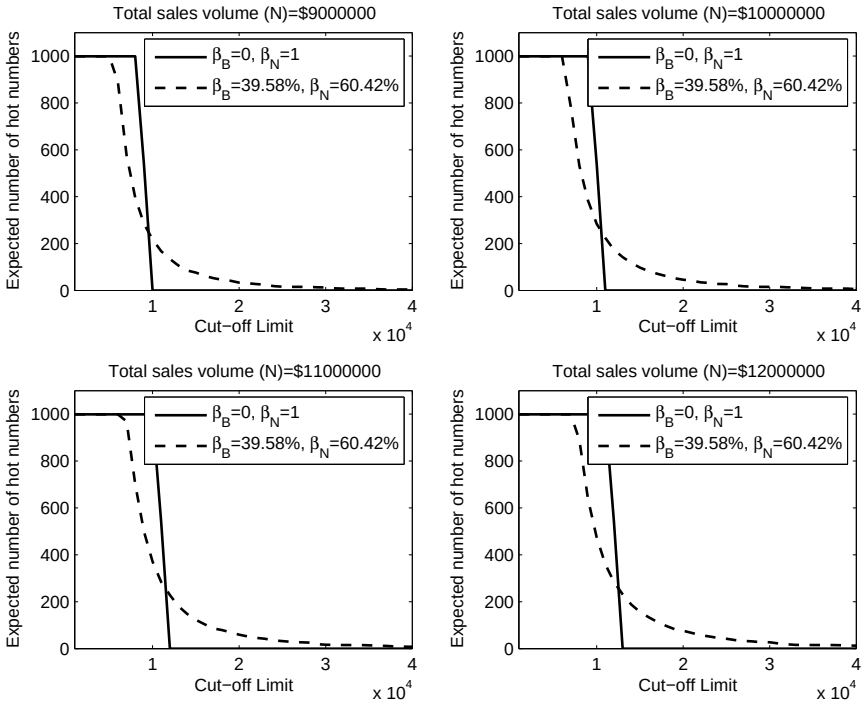

Fig. 2.8: Expected number of hot numbers using different liabilities.

respectively. In this environment, trying to find the appropriate sales limit to control the right level of hot numbers is almost impossible because this number depends critically on the total sales level, a number which normally fluctuates from draw to draw.

In the empirical sales data, we have $\beta_B \approx 0.3958$. In this environment, interestingly, the phase transition phenomenon disappears, and the relationship between the sales limit and the expected number of hot numbers is more stable. For a sales limit of \$1000, the hot numbers fluctuate from 200 to 400 when the total sales level changes from \$9M to \$12M. The relationship between the total sales and proportion of hot numbers are thus more stable.

2.5 Conclusion

In this topic, we have analyzed an interesting phenomenon in a popular numbers game. While it is by now folklore that players in these games prefer small numbers, this paper is arguably the first to quantify this behavior using Benford's law. The connection is forged by virtue of the argument that many natural data series satisfy Benford's law. We also take into account the choice behavior of the players, in particular the way the players compose the 3-digit number to obtain a refined choice model. Although we do not model additional phenomenon such as superstitious beliefs and date-month effect on the choice behavior, the simple model we built, using only a few parameters (i.e., the proportion of type 2 agents β_B, the probability of switching q, and the probability of padding the number with digit zero p_0), is already able to capture some of the most important characteristics of the empirical data.

While we have presented only the analysis using a set of publicly available data from the US, we have tested the model on an extensive series of data provided by a game operator in another region. Despite the differences in culture and beliefs, we found that the same underlying model can be used to describe the behavior of the aggregate data, with the main difference coming from the proportion of Benford-like players. We believe that the small-number phenomenon is a generic behavior inherent in many numbers games.

The proportion of type 2 players (β_N) has a tremendous impact on the variability of the prize liability, and to a certain extent affects the appropriate choice of sales limit in the numbers game. There are many ways to mitigate the small-number effect through demand shaping. One approach, already in use, is to use on-site computer terminals to help players to pick the numbers randomly. However, people often do not like random picks because they like to assume some control over the outcomes (cf. Langer (1975)). Therefore, there may be limits on the extent to which the industry can encourage random picks. Another approach is to re-design the game to encourage the players to bet on as many different permutations as possible. In Singapore, the introduction of a new iBet system (cf. http://www.singaporepools.com.sg) proves to be popular with the players. The new system allows the players to spread a dollar bet on as many permutations of the number combination as possible, with a corresponding reduction in the prize monies. It also helps to mitigate the effect of the small number phenomenon. Other possible approaches include posting the results of past winning numbers in the retail outlets to influence the selection of numbers by the players. The past win-

ning numbers are drawn in a random manner and thus will not exhibit the same feature as numbers picked by Benford's type player. Of course, recency bias may actually deters players from betting on recent winning numbers, and hence this approach may not be as effective in persuading players from moving away from their preferred numbers.

Interestingly, if the game operators are able to reduce the proportion of benford players, then the above analysis shows that the imposition of the sales limit may no longer be needed, since it will be difficult and futile to implement such a mechanism anyway.

3. TOPIC 2: APPOINTMENT SYSTEM DESIGN USING COPOSITIVE CONES

3.1 Abstract

In this topic, we investigate a stochastic appointment scheduling problem in an outpatient clinic with a single doctor. The number of patients and their sequence of arrivals are fixed, and the scheduling problem is to determine an appointment time for each customer. The service durations of the patients are stochastic, and only the mean and covariance estimates are known. We do not assume any exact distributional form of the service durations, and solve for distributionally robust schedules that minimize the expectation of the weighted sum of patients' waiting time and doctor's overtime. We formulate this scheduling problem as a convex conic optimization problem with a tractable semidefinite relaxation. Our model can be extended to handle additional support constraints of the service durations. Using the primal-dual optimality conditions, we prove several interesting structural properties of optimal schedules. Despite the required relaxation in computation, we can still obtain near optimal solutions compared to the existing literature. We apply our method in a realistic setting at an eye clinic and suggest new

schedules that can significantly improve the efficiency of the clinic.

3.2 Introduction

In many service delivery systems, the core operational activities are largely planned around the arrival times of the customers. The ability to regulate the arrival of customers through a suitable appointment system, is thus central to the performance of these systems. The FastPass service of Disney is a well known example. Customers in the park can obtain a pass to ensure fast service at certain rides if they return at the stipulated time. The temple of Tirumala in India has also used an online appointment system to convert its long waiting line into a virtual queue. This has helped improve service delivery and generated spillover economic benefits to businesses in the vicinity of the temple[1].

The appointment design problem is also a core problem for healthcare facilities such as outpatient clinics and operating rooms. The appointment system is used to regulate the usage of the costly equipment and precious resources in the system. In an eye-care facility that we have visited, there are two consultation sessions per day, each lasting four hours, and the number of doctors available per session is around two to seven. Each doctor has to handle 20 to 30 patients per session. The patients can be classified into "New" (20%) and "Repeat" (80%) patient types. The mean and variance of the consultation times of the new patients are noticeably higher than those of repeat patients, as the conditions of the new patients are hitherto unknown

[1] See http://www.iimahd.ernet.in/publications/data/2005-08-02nravi.pdf for a thorough discussion.

prior to the visit. There are also various operational details that complicate the situation. For instance, patients often have to go for a dilation test prior to seeing the doctor. This process adds to the complexity of finding an optimal appointment strategy for the system.

One key performance indicator in this system is the "Turnaround Time" (TAT), defined to be *the time from the moment the patient walks into the clinic, to the moment the patient leaves the clinic.* Figure 3.1 shows the overall median TAT, service time and waiting time (WT) of patients arriving in different time slots for two different sessions in the clinic, where TAT is the sum of service and waiting times. Clearly the patients are experiencing long turn around time, with waiting time far exceeding the actual service time.

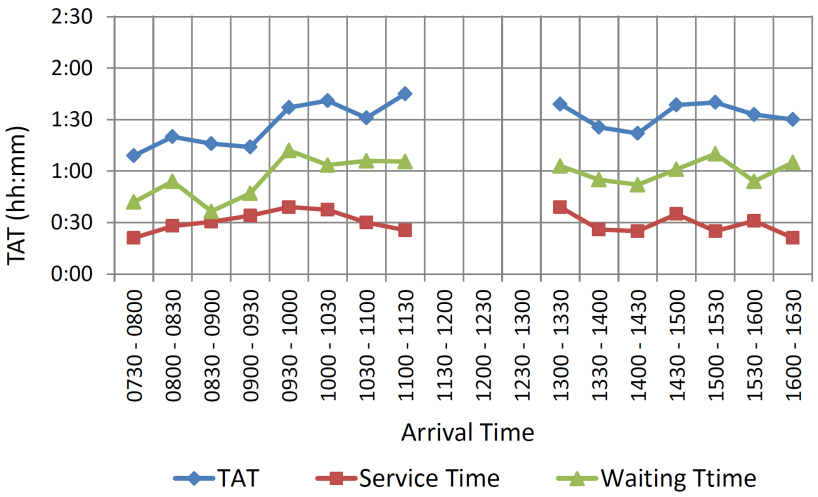

Fig. 3.1: Median time from registration to payment

We note that there are several pertinent features in this system: (i) New patients often have to undergo a series of checks (such as visual acuity,

and/or other advanced tests) after the consultation, some of which can take as much as 2.5 hours. To make sure that all the tests and consultations can be performed within the same day, the doctors prefer to see the new patients in the early portion of the morning session. Consequently, early morning slots are reserved primarily for new patients. (ii) The current appointment strategy is to allocate 5 minutes per patient slot for one hour, followed by a half hour break. This allows each doctor to see around 36 patients in each 4-hour session.

This leads to the central questions for this topic: is there any (near) optimal strategy to schedule and sequence the arrival of patients such that the waiting time of the patients and overtime work of the doctor are minimized? Furthermore, are there any "distributionally robust" solutions that perform well for a wide range of service time distributions?

The research on appointment system design over the past few decades has been driven largely by these issues. However, these problems are notoriously difficult. Standard queueing theory does not apply as we are interested in the transient performance measures of the system. It is technically challenging to calculate the expected waiting time of the n^{th} patient in the sequence, due to the difficulty of propagating the impact of earlier events on this patient. Recently, Begen and Queyranne (2009) show that the scheduling problem is solvable in polynomial time (in the size of the representation of the discrete distributions). However, this method works well only for discrete distributions with a small number of distinct values. To the best of our knowledge, simulation and stochastic programming methods are still the preferred approaches for the appointment design problem. Unfortunately,

the solutions obtained are often sensitive to the samples used to develop the schedules, and hence very little is known about the structure of the optimal policies, even in the simplest environment with one doctor and when patients arrive punctually according to the appointment time.

3.2.1 Contributions

In this topic, we develop a convex conic programming approach to solve the appointment scheduling problem. We show that this problem can be suitably reformulated as a two-stage stochastic optimization problem. In the second stage, we construct a network flow model to capture the waiting time of each patient, under a given scheduling policy (from the first stage problem). Our novelty comes in the solution to the first stage problem, which is a technically challenging problem. Instead of using a specific service time distribution to design the schedule, we employ a minimax approach so that the schedule is designed to minimize the maximum expected cost achieved by some distribution from a family of distributions. Next, we develop a conic optimization framework to transform the stochastic appointment scheduling problem into a single deterministic copositive programming problem (COP)[2].

Using the primal-dual optimality conditions, we prove several interesting structural properties of the optimal schedule. For instance, our analysis shows that when the appointment system is operating under the optimal schedule, other than the first slot and the last few ones (where the consultation intervals allocated are zero, i.e., patients are bunched together), the

[2] A copositive programming problem is a linear programming problem over the convex cone of the copositive matrices. Details of this optimization problem are discussed later in this topic.

chances of waiting for service in the clinic is identical for patients assigned to all other slots. Furthermore, our model can also handle the correlations between patients' service durations, which has been largely overlooked in literature.

Computationally, we solve a tractable semidefinite approximation to the COP. Although the schedule obtained using our model is optimal for a set of canonical service time distributions (called *worst case distributions*), our numerical results show that this schedule also works reasonably well for several other service duration distributions with the same moment conditions. We also find that the schedule obtained from solving the SDP approximation often satisfies the structural properties obtained from model analysis. Furthermore, with the help of existing semidefinite programming (SDP) packages, we can work out practical size appointment scheduling problems.

In a congested system with two types of patients, as in our eye clinic case, the optimal schedule often exhibits the pattern: "Bailey's Rule + Break"[3] - the optimal schedule allocates near zero time slot to the first few patients, which resembles the well known "Bailey's Rule", and a break is often inserted before switching from a class of patients with higher variability to another class of patients with lower variability. We use this observation and the solution from the SDP model to develop a simple and practical schedule for the eye clinic. Compared to the naive approach of allocating equal interval to

[3] "Bailey's Rule" refers to the scheduling strategy proposed in the seminal paper by Bailey (1952). It states that in a highly congested system, "*an optimum system seems to be as follows: the patients are given appointments at regular intervals equal to the average consultation time, and the consultant arrives at the same time as the second patient*". That means the first two patients are scheduled to come at the beginning of the consultation session at the same time.

each patient with a break in between (which is current practice in the clinic), our schedule can reduce the total system waiting cost by more than 35%. This approach has thus the potential of producing near optimal appointment schedules that can be deployed in practice.

Finally, we extend this approach to incorporate sequencing decision into the appointment design problem. We show that the problem can be solved approximately as a 0-1 SDP problem. Our numerical results give several insights into the structure of the optimal sequencing decisions. In particular, we observe that the optimal sequence may not follow the smallest variance first rule. In fact, in some instances, a U-shape rule is more efficient. This is surprising as it is counter-intuitive to put a high variability patient in front of the queue to minimize the total expected waiting time.

3.2.2 Structure of the Topic

In the next section, we briefly review the relevant literature for our problem. In Section 3.4, we describe the development of our conic optimization model in two steps, followed by several important extensions in Section 3.5 to address more practical issues. In Section 3.6, we analyze the structure and properties of the optimal scheduling policy, while in Section 3.7, numerical studies are presented to evaluate our approach under various circumstances as well as a case study of the eye clinic. We conclude in Section 3.9.

3.3 Literature Review

Since the pioneering work of Bailey (1952) and Welch & Bailey (1952), there have been extensive studies on the appointment design problem in the past six decades. In this section, we only briefly review some key results from this line of research that is most relevant to our topic, but refer the readers to Cayirli and Veral (2003), Gupta (2007), Gupta and Denton (2008) and Erdogan & Denton (2010) for more thorough reviews.

Denton & Gupta (2003) formulate the appointment scheduling problem as a two-stage stochastic linear program and used a sequential bounding approach to determine upper bounds of the problem. Kaandorp & Koole (2007) assume that the service durations follow an exponential distribution and that the patient arrivals can only be scheduled at discrete intervals. They used results in queueing theory to calculate the objective function for a given schedule of starting times and used a local search algorithm to find the optimal solution. Begen and Queyranne (2009) go a step further and argue that under mild assumptions, the discrete time version of the appointment scheduling problem could be solved in polynomial time, by showing that the objective function is an L-convex function. A recent paper by Begen et al. (2010) is based on the methodology developed in Begen and Queyranne (2009) but assume no prior knowledge of probability distributions on job durations. They re-constructe an empirical distribution of the consultation durations from a set of historical data and then developed a sampling-based approach and established the cost (numbers of samples needed) to obtain a near-optimal solution with high probability.

When patients are homogenous, the issues are simpler since scheduling rules are now the only concern. In practice, however, patients are distinct due to patient's classification (e.g., new/repeat, ages, types of procedures performed etc.). Patients in different classifications tend to give rise to different means and variability in consultation/service time durations. Higher percentage of more complicated cases (e.g., new patients) normally translates into higher variability in the system performance, and thus proper sequencing of patients become more valuable (cf. Vanden and Dietz (2000) and Cayirli et al. (2008)). Weiss (1990) is arguably the first to study the optimal sequencing problem analytically. He explored the optimal starting time and sequencing of surgical procedures, to best utilize medical resources like surgeons and operating rooms. He showed that sequencing lower-variance procedure first is optimal in the case of 2 procedures under exponential/uniform service time. Weiss also conjectured that the smaller-variance-first rule might be optimal in more complicated systems. Similar results were later reported for local-scale distributions like normal and uniform distributions (cf. Gupta (2007)).

In view of the analytical and computational difficulties of the appointment scheduling problem, we address the issue from a different angle, utilizing the concept of robust optimization. Evolving from the *minimax theorem* established by John Von Neuman in 1928, the concept was first brought into operations research area by Scarf (1958). Scarf solves an inventory problem with random demand by assuming only the mean and variance of the demand instead of a specific form of distribution. Noting that there could be multiple distributions that satisfy a given mean and variance, Scarf identifies a worst case distribution that would result in the highest expected total system

cost, and finds an inventory strategy to minimize this maximal cost. That is why another popular term describing this concept is called *distributionally robust*. Such a concept has recently been extensively studied and extended, and one stream of research is to exploit the connection between the theory of moments and semidefinite programming (SDP) (cf. Bertsimas et al. (2004), Bertsimas et al. (2006), Bertsimas et al. (2008), Vandenberge et al. (2007), etc.). Most recently, Natarajan et al. (2009) show that a robust mixed 0-1 linear program under objective uncertainty is equivalent to a convex conic program, which may help to deal with a second stage recourse function in a two stage stochastic programming framework.

3.4 A Two Stage Model with the Copositive Cone

3.4.1 Assumptions, Notations and Problem Formulation

To isolate the impact of scheduling on the system performance, we rule out the presence of other disruptions in the system. The basic assumptions are listed as follows:

1. The sequence of patient arrivals is fixed. Service occurs in the same sequence[4].

2. Patients arrive punctually at the scheduled appointment times[5].

3. There is a single doctor in the facility. The doctor arrives punctually

[4] This assumption is relaxed when we study the sequencing problem in section 3.8

[5] This assumption can be relaxed. In Section 3.5.2, we demonstrate how to extend our model to incorporate late arrivals.

and only serves the scheduled patients during the session. No break is taken during the time serving one patient.

4. Patients in the same class are homogenous in the distribution of consultation durations.

5. Walk-in and emergency patients are not considered.

Note that in a typical appointment scheduling problem, it is common for the patients to choose the appointment slots in a dynamic fashion, and their characteristics, such as mean and standard deviation of service time, are known only at the time of booking. The problem described above matches more the surgery scheduling environment. However, in certain appointment scheduling environments, patients are classified into distinct classes and each appointment slot in a single clinical session is pre-assigned to a dedicated class of patients. The slots are filled up when patients call in for appointments and their classifications are revealed. We assume that the clinic has enough volume to fill up the slots available in each day. In this way, the scheduling problem described here essentially addresses the design of the appointment system based on the patient classifications, not on the characteristics of individual patients.

Let $N = \{1, 2, \ldots, n\}$ be the index set for all patients, and the sequence of arrivals is $1, 2, \ldots, n$. Let $\tilde{u}_i$ be the random service time of patient i, $i = 1, 2, \ldots, n$. We define $\boldsymbol{s} = \{s_1, s_2, \cdots, s_n\}^T$, where s_i represents the length of time slot scheduled for i^{th} patient in the sequence. Therefore, the appointment time of the patients in the sequence is given by $\left\{0, s_1, s_1 + s_2, \ldots, \sum_{i=1}^{n-1} s_i\right\}$.

We assume that $\tilde{u}_i$ follows a distribution with mean μ_i and standard deviation σ_i, and $\mathbf{P}(\tilde{u}_i \geq 0) = 1$, i.e., $\tilde{u}_i$ has nonnegative support. Let w_i denote the waiting time of the i^{th} patient in the sequence. It is reasonable to assume that the first session starts at time zero, i.e., $w_1 = 0$. Define $\tilde{c}_i$ to be the difference between the actual consultation time and the allocated consultation interval of the i^{th} patient in the sequence, i.e., $\tilde{c}_i = \tilde{u}_i - s_i,\ i = 1, \ldots, n$. Then the waiting time of subsequent patients are given by the following recursions:

$$w_i = \max\left\{0, w_{i-1} + \tilde{c}_{i-1}\right\},\ i = 2, 3, \ldots, n.$$

More precisely,

$$w_i = \max\left\{0, \tilde{c}_{i-1}, \tilde{c}_{i-1} + \tilde{c}_{i-2}, \cdots, \sum_{k=1}^{i-1} \tilde{c}_k\right\},\ i = 2, 3, \ldots, n. \tag{3.1}$$

If there were an additional "auxiliary" patient (i.e., the $(n+1)^{st}$ patient) arriving at the end of the consultation session, then the doctor's overtime would be exactly the waiting time of this patient, i.e., $w_{n+1} = \max\left\{0, w_n + \tilde{c}_n\right\}$. In this topic, we will use the total patients' waiting time and doctor's overtime (i.e., $\sum_{k=1}^{n} w_k$ and w_{n+1}) as the key performance indicators of the appointment system. The objective of the appointment scheduling problem is to minimize the expectation of the weighted sum of the patients' waiting times and the doctor's overtime, i.e.,

$$\mathbf{E}\left[\sum_{i=1}^{n} \rho_i w_i + \rho_{n+1} w_{n+1}\right], \tag{3.2}$$

where ρ_i, $i = 1, 2, \ldots, n+1$ are the corresponding weights (or the unit waiting time/overtime cost). We first assume that $\rho_i = 1$ for all $i = 1, \ldots, n+1$, and then relax this assumption in Section 3.5.

Note that the doctor's total idle time during the session is also a crucial performance indicator of the appointment system. When the consultation interval (i.e., the session length, denoted as T) is pre-determined, the total idle time is $T + w_{n+1} - \sum_{i=1}^{n} \tilde{u}_i$. Hence, we do not include the doctor's idle time in the objective since adding the expected total idle time can only cause the objective function to differ by a constant and the weight of w_{n+1} to increase by 1.

The technical difficulty associated with the scheduling problem is partially due to the computation of

$$\mathbf{E}\left[w_i\right] = \mathbf{E}\left[max\left\{0, \tilde{c}_{i-1}, \tilde{c}_{i-1} + \tilde{c}_{i-2}, \cdots, \sum_{k=1}^{i-1} \tilde{c}_k\right\}\right], \ i = 2, 3, \ldots, n.$$

We introduce a two stage stochastic optimization framework to tackle this problem. In the first stage, the appointment scheduling decisions are made under the objective to minimize the expected total waiting time cost[6] defined in equation (3.2). In the second stage, the patients' service durations are realized and the system performance is determined. Let us consider the second stage problem first.

[6] In the rest of this topic, we use the phrase "the total waiting time (cost)" to include both of the waiting time (costs) of all the patients and the overtime (cost) of the doctor.

3.4.2 The Second Stage Problem

Given the schedule of the patients (i.e., $\boldsymbol{s}$ is known), the total waiting time cost in equation (3.2) can be computed by solving a network flow problem on a directed acyclic graph shown in Figure 3.2, with $n+1$ supply nodes and a sink node s. The cost on arc (i, s) is 0, and the cost on arc $(i+1, i)$ is $\tilde{c}_i(\boldsymbol{s}) = \tilde{u}_i - s_i$, where the notation $\tilde{c}_i(\boldsymbol{s})$ is used here to emphasize the dependencies of $\tilde{c}_i$ on the given schedule $\boldsymbol{s}$ (not in the figure). The capacities for all the arcs are infinite. Let y_i, $i = 1, 2, \ldots, n$, be the flows on arc $(i+1, i)$, and z_i, $i = 1, 2, \ldots, n+1$ be the flows on arc (i, s).

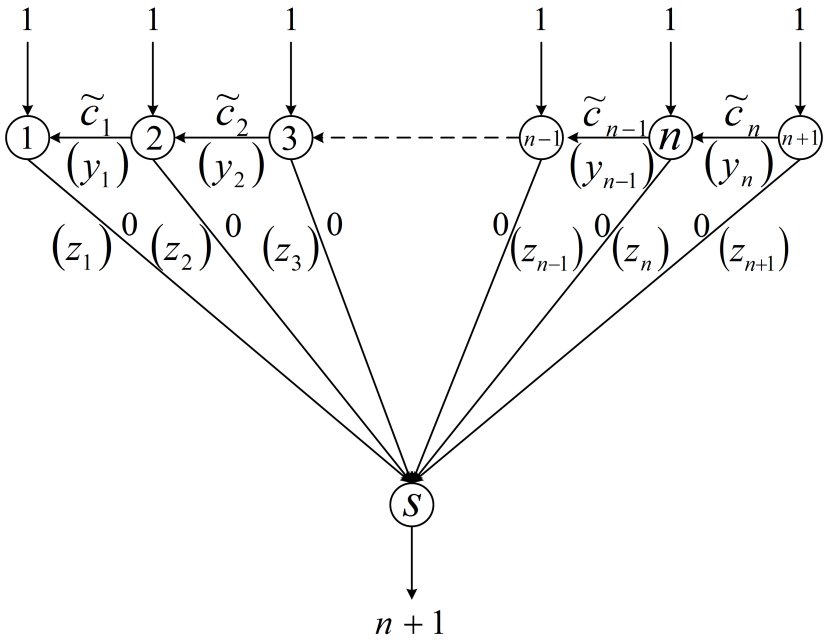

Fig. 3.2: Network flow representation of the appointment scheduling problem

Proposition 3. Given the schedule $\boldsymbol{s}$, the optimal cost of the following maximum cost flow problem equals the total waiting time cost of the system under

any realization of $\tilde{\boldsymbol{u}}$:

$$
\begin{aligned}
f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right) := \quad \max \quad & \sum_{i=1}^{n} \tilde{c}_i(\boldsymbol{s}) \cdot y_i \\
s.t. \quad & y_1 - z_1 = -1 \\
& y_i - y_{i-1} - z_i = -1, \, \forall i = 2, 3, \ldots, n \\
& -y_n - z_{n+1} = -1 \\
& y_i \geq 0, \, \forall i = 1, 2, \ldots, n \\
& z_i \geq 0, \, \forall i = 1, 2, \ldots, n+1
\end{aligned}
$$

Proof. The proposition can be easily verified through tracking the flow of each unit of supply at node $1, 2, \ldots n+1$. A detailed argument can be found in Appendix A. ■

Remark 1. Note that Proposition 3 is developed in the deterministic situation. In the second stage, the patients' service durations are realized, i.e., they can be considered as deterministic. Then the network optimization problem in Proposition 1 is proposed to find out the total waiting time cost under this realization. When the patients' service durations ($\tilde{\boldsymbol{c}}(\boldsymbol{s})$) become stochastic, the optimal value of the network flow problem ($f(\boldsymbol{s}, \tilde{\boldsymbol{u}})$) also becomes stochastic and depends on $\tilde{\boldsymbol{c}}(\boldsymbol{s})$.

Removing one redundant network flow conservation constraint and using the matrix notation, we rewrite $f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right)$ as follows for the ease of exposition:

$$
\begin{aligned}
f(\boldsymbol{s}, \tilde{\boldsymbol{u}}) = \quad & \max \quad \tilde{\boldsymbol{c}}^T(\boldsymbol{s})\,\boldsymbol{y} \\
& s.t. \quad \boldsymbol{a}(j)^T\boldsymbol{y} - \boldsymbol{e}(j)^T\boldsymbol{z} = -1,\ \forall j = 1, 2, \ldots n \\
& \qquad\ \ \boldsymbol{y}, \boldsymbol{z} \geq 0
\end{aligned}
$$

where $\tilde{\boldsymbol{c}}(\boldsymbol{s}) = (\tilde{c}_1(\boldsymbol{s}), \tilde{c}_2(\boldsymbol{s}), \ldots, \tilde{c}_n(\boldsymbol{s}))^T$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, and $\boldsymbol{z} = (z_2, z_3, \ldots, z_{n+1})^T$; and $\boldsymbol{e}(j) \in \mathbb{R}^n$ is the unit vector with its j^{th} entry being one; and $\boldsymbol{a}(j)_j = -1$ for $j = 1, \ldots n$, $\boldsymbol{a}(j)_{j+1} = 1$ for $j = 1, \ldots, n-1$, and $\boldsymbol{a}(j)_k = 0$ otherwise.

3.4.3 The First Stage Problem

As mentioned before, we will deploy the minimax approach in our modeling framework, which we need to address before solving the scheduling problem. Under a fixed schedule $\boldsymbol{s}$, when the service durations become stochastic, but with given moment conditions, the maximal expected total waiting time cost can be written as:

$$
\text{(P)} \qquad Z_P(\boldsymbol{s}) := \sup_{\tilde{\boldsymbol{u}} \sim (\boldsymbol{\mu}, \Sigma)^+} \{\mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}})]\}
$$

where $\tilde{\boldsymbol{u}} \sim (\boldsymbol{\mu}, \Sigma)^+$ denotes that the distribution of $\tilde{\boldsymbol{u}}$ lies in the set of feasible multivariate distributions supported on $\mathbb{R}^n_+$ with finite first moment $\boldsymbol{\mu}$ and finite second moment Σ. We assume this set to be nonempty. The challenge to solve (P) reduces to the following: *can one find a distribution for the random variable $\tilde{\boldsymbol{u}}$ in such a way that*

$$
\boldsymbol{P}(\tilde{\boldsymbol{u}} \geq 0) = 1,\ \boldsymbol{E}[\tilde{\boldsymbol{u}}] = \boldsymbol{\mu},\ \boldsymbol{E}\left[\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^T\right] = \Sigma,
$$

and a corresponding optimal solution $(\boldsymbol{y}(\boldsymbol{s},\tilde{\boldsymbol{u}}),\boldsymbol{z}(\boldsymbol{s},\tilde{\boldsymbol{u}}))$ *to* $f(\boldsymbol{s},\tilde{\boldsymbol{u}})$ *in* (P)*, so that* $\boldsymbol{E}\left[\tilde{\boldsymbol{c}}(\boldsymbol{s})^T\boldsymbol{y}(\boldsymbol{s},\tilde{\boldsymbol{u}})\right]$ *attains the maximum* $Z_P(\boldsymbol{s})$*? In general, if the maximum cannot be attained, can one find a sequence of random variables so that* Z_P *can be attained asymptotically?*

It turns out that this problem can be reformulated into a conic programming problem through a moment decomposition approach. Before showing the main result, we introduce some necessary notations, and briefly review related subjects on the conic optimization problem.

Notations and A Brief Review of Conic Optimization

The trace of a matrix A, denoted by $tr(A)$, is the sum of the diagonal entries of A. The inner product between matrices A and B of the same dimensions is denoted as $A \bullet B = tr(A^T B)$. I_n represents the identity matrix of dimension $n \times n$, while $\mathbf{0}_{m\times n}$ is used to denote the zero matrix of dimension $m \times n$. We may drop the subscript when it represents a zero vector of an appropriate dimension that is obvious.

For any cone $\mathcal{K}$, its dual cone is denoted as $\mathcal{K}^*$. Let S_n denote the cone of $n \times n$ *symmetric matrices*, and $\mathcal{S}_n^+$ denote the cone of $n \times n$ *positive semidefinite matrices*. $A \succeq 0$ indicates that the matrix A is positive semidefinite, and $B \succeq A$ is equivalent to $B - A \succeq 0$. Similarly, $A \geq 0$ indicates that the matrix A has nonnegative entries, and $B \geq A$ is equivalent to $B - A \geq 0$.

Two cones of special interest are the cone of *completely positive matrices* and the cone of *copositive matrices*. The cone of $n \times n$ completely positive

matrices is defined as

$$\mathcal{CP}_n := \left\{A \in \mathcal{S}_n : \exists V \in \mathbb{R}_+^{n\times k}, \text{ such that } A = VV^T\right\} = \text{conv}\left\{\boldsymbol{v}\boldsymbol{v}^T : \boldsymbol{v} \in \mathbb{R}_+^n\right\},$$

where "conv" means the convex hull. The cone of $n \times n$ *copositive matrices* is defined as

$$\mathcal{CO}_n := \left\{A \in \mathcal{S}_n : \forall \boldsymbol{v} \in \mathbb{R}_+^n,\ \boldsymbol{v}^T A \boldsymbol{v} \geq 0\right\}.$$

$A \succeq_{cp} (\succeq_{co}) 0$ indicates that matrix A is completely positive (copositive). These two cones are both closed, convex, pointed and of course, and have nonempty interior. Moreover, they are duals of each other (cf. Berman and Shaked-Monderer (2003)). A linear program over the cone of copositive matrices is called a *copositive program (COP)*, whose dual problem is a linear program over the cone of completely positive matrices known as a *completely positive program (CPP)*.

Despite the nice properties of these two cones, it is widely believed that their membership status is $\mathcal{NP}$-hard to check. For instance, the problem of testing if a given matrix is copositive is known to be co-$\mathcal{NP}$-complete (cf. Murty et al. (1987)). In a recent paper, Dickinson & Gijben (2011) showed that the membership problems for both copositive and completely positive cones are $\mathcal{NP}$-hard. Fortunately, there are well-known hierarchies of linear and semidefinite representable cones that approximate the copositive and completely positive cones (cf. Bomze et al. (2000), Klerk et al. (2002), Parrilo (2000)). In this topic, we restrict our attention to the simplest relaxations

of CPP and COP for the numerical experiments, i.e.,

$$\begin{cases} A \succeq_{cp} 0 \approx A \succeq 0, \text{ and } A \geq 0 \\ A \succeq_{co} 0 \approx \exists\, A_1 \succeq 0, \text{ and } A_2 \geq 0, \text{ such that } A = A_1 + A_2. \end{cases} \tag{3.3}$$

More information on CPP and COP can be found in Berman and Shaked-Monderer (2003).

Moment Decomposition and Conic Representation

For ease of exposition, we define $\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T$, and rewrite the network flow constraints as $A\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq 0$, where

$$A = \begin{pmatrix} \boldsymbol{a}(1)^T & -\boldsymbol{e}(1)^T \\ \boldsymbol{a}(2)^T & -\boldsymbol{e}(2)^T \\ \vdots & \vdots \\ \boldsymbol{a}(n)^T & -\boldsymbol{e}(n)^T \end{pmatrix}, \text{ and } \boldsymbol{b} = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}.$$

Since A has full rank, the only feasible solution to $A\boldsymbol{x} = \boldsymbol{0}$ and $\boldsymbol{x} \geq 0$ is $\boldsymbol{x} = \boldsymbol{0}$.

Let

$$\mathcal{D} := \text{conv}\left\{ \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix} \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix}^T : \pi \geq 0,\, \boldsymbol{t} \in \mathbb{R}^n_+,\, \boldsymbol{v} \in \mathbb{R}^{2n}_+,\, A\boldsymbol{v} = \boldsymbol{b}\pi \right\}. \tag{3.4}$$

From the definition of $\mathcal{CP}_n$, we know that $\mathcal{D}$ is indeed the intersection of the

completely positive cone, $\mathcal{CP}_{3n+1}$ with a hyperplane in $\mathbb{R}^{2n+1}$ projected onto $\mathbb{R}^{3n+1}$ (i.e., a polyhedral cone in $\mathbb{R}^{3n+1}$). Furthermore, if $\pi = 0$, then $A\boldsymbol{v} = \mathbf{0}$ and consequently $\boldsymbol{v} = \mathbf{0}$. Therefore, every $Z \in \mathcal{D}$ can be expressed as

$$Z = \sum_{k \in K_+} \pi(k)^2 \begin{pmatrix} 1 \\ \frac{\boldsymbol{t}(k)}{\pi(k)} \\ \frac{\boldsymbol{v}(k)}{\pi(k)} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\boldsymbol{t}(k)}{\pi(k)} \\ \frac{\boldsymbol{v}(k)}{\pi(k)} \end{pmatrix}^T + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \boldsymbol{t}(k) \\ \mathbf{0}_{2n\times 1} \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{t}(k) \\ \mathbf{0}_{2n\times 1} \end{pmatrix}^T, \quad (3.5)$$

where K_+ and K_0 are the corresponding indicator sets, and they can be chosen to be finite[7] (c.f. Berman and Shaked-Monderer (2003)).

If $Z_{1,1} = 1$, then $\pi(k)^2$ can be interpreted as the probability of the k^{th} scenario with service duration $\tilde{\boldsymbol{u}} = \boldsymbol{t}(k)/\pi(k)$, and solution $\boldsymbol{x}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) = (\boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}), \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}})) = \boldsymbol{v}(k)/\pi(k)$. The corresponding objective function in the k^{th} scenario is given by $\sum_{i=1}^{n}(\tilde{u}_i - s_i)y(\boldsymbol{s}, \tilde{\boldsymbol{u}})_i$. Averaging over all the scenarios each with probability $\pi(k)^2$, we get the objective function given by $Y(\boldsymbol{s}) \bullet Z$, where $Y(\boldsymbol{s})$ is a $(3n+1) \times (3n+1)$ symmetric matrix defined as

$$Y(\boldsymbol{s}) = \begin{pmatrix} 0 & \mathbf{0}_{1\times n} & -\frac{\boldsymbol{s}^T}{2} & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \frac{I_n}{2} & \mathbf{0}_{n\times n} \\ -\frac{\boldsymbol{s}}{2} & \frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix}.$$

The second term in the expression for Z in (3.5) can be viewed as a characterization of the null set for the corresponding probability space. With

[7] Indeed, not only they could be finite, but also bounded. This is related to the concept of *cp-rank*, details of which can be found in Berman and Shaked-Monderer (2003).

such moment decomposition interpretation, we get the following optimization problem by incorporating other moment conditions:

$$\begin{aligned} \text{(C)} \qquad Z'_P(\boldsymbol{s}) := \quad & \max \quad Y(\boldsymbol{s}) \bullet Z \\ & s.t. \quad Z_{1,1} = 1,\ Z_{1,i+1} = \mu_i,\ Z_{i+1,j+1} = \Sigma_{i,j},\ \forall i,j = 1,\ldots,n \\ & \qquad Z \in \mathcal{D} \end{aligned}$$

Furthermore, we can prove that the above conic optimization problem is indeed equivalent to problem (P).

Proposition 4. For any given schedule $\boldsymbol{s}$, $Z'_P(\boldsymbol{s}) = Z_P(\boldsymbol{s})$.

Proof. There are two steps involved in the proof. Firstly, we show that problem (C) provides an upper bound for (P), i.e. $Z'_P(\boldsymbol{s}) \geq Z_P(\boldsymbol{s}), \forall \boldsymbol{s}$. Next, through a constructive approach, we find a sequence of random vectors, $\tilde{\boldsymbol{u}}^*_\epsilon$ that satisfies the moment conditions in the limiting sense and $\mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*_\epsilon)]$ converges to $Z'_P(\boldsymbol{s})$ when ϵ converges to zero, i.e., the bound provided by (C) is tight. The technical details are omitted here but available in Appendix B.■

Now we have a conic maximization problem that solves problem (P) exactly. To incorporate the scheduling decision $\boldsymbol{s}$, we still need one more step, which is taking the dual of problem (P).

Remark 2. Note that our conic optimization model resembles the results of Natarajan et al. (2009) from a different perspective. Instead of separating the moment requirement on $\tilde{\boldsymbol{u}}$ and the feasibility conditions on $\boldsymbol{x}$ and then

enforcing their relationship through a lifting constraint, we directly characterize the cone $\mathcal{D}$ from the moment decomposition angle. One of the advantages of the new perspective is that it makes the regularity condition for the strong conic duality self-evident, in particular, Slater's constraint qualification. We will elaborate more on the dual problem in the following analysis. Another important advantage is that it can be extended into a much more general conic framework for a stochastic optimization problem, which will be discussed in Section 3.5.4.

Conic Duality and Copositive Program

Let $\mathcal{D}^*$ denote the dual cone of $\mathcal{D}$, i.e., $\mathcal{D}^* = \{W : Z \bullet W \geq 0, \forall Z \in \mathcal{D}\}$. Then the dual of problem (C), denoted by $Z_D(\boldsymbol{s})$, can be written as follows:

$$
\begin{aligned}
Z_D(\boldsymbol{s}) := \quad & \min \quad \Sigma \bullet \Gamma + \boldsymbol{\mu}^T \boldsymbol{\beta} + \alpha \\
& s.t. \quad W = \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \mathbf{0}_{1\times 2n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} - Y(\boldsymbol{s}) = \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\boldsymbol{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\boldsymbol{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} \\
& \qquad W \in \mathcal{D}^*
\end{aligned}
$$

where $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ are the corresponding dual variables of the moment constraints.

By the definition of $\mathcal{D}^*$, for all $(1, \tilde{\boldsymbol{u}}, \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}), \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}))^T$ satisfying

$$A \begin{pmatrix} \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \end{pmatrix} = \boldsymbol{b},\ \tilde{\boldsymbol{u}} \geq 0,\ \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \geq 0,\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \geq 0,$$

we have

$$\begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \\ \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\boldsymbol{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\boldsymbol{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \\ \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \end{pmatrix} \geq 0,$$

i.e.,

$$\begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \end{pmatrix} \geq (\tilde{\boldsymbol{u}} - \boldsymbol{s})^T \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}).$$

Hence, for any distribution of the service durations, with probability 1,

$$\begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\boldsymbol{u}} \end{pmatrix} \geq \max \left\{ (\tilde{\boldsymbol{u}} - \boldsymbol{s})^T \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) : A \begin{pmatrix} \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \end{pmatrix} = \boldsymbol{b},\ \tilde{\boldsymbol{u}} \geq 0,\ \boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \geq 0,\ \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}) \geq 0 \right\}.$$

Then the weak duality $Z_D(\boldsymbol{s}) \geq Z_P(\boldsymbol{s})$ follows immediately. Furthermore, since problem (P) is obviously bounded, so is (C). Then as long as $\mathcal{D}$ has a nonempty relative interior, by the Slater's constraint qualification, there is

no duality gap between the primal $Z_P(\boldsymbol{s})$ and its dual $Z_D(\boldsymbol{s})$. Note that $\mathcal{D}$ needs not be full dimensional for the strong duality to hold. We use a simple example in Appendix C to illustrate this.

To convert $Z_D(\boldsymbol{s})$ into a copositive programming problem, we need to analyze the structure of the cone $\mathcal{D}$ and $\mathcal{D}^*$. Let $Z \in \mathcal{D}$, and

$$M_i = \begin{pmatrix} b_i^2 & \mathbf{0}_{1\times n} & -b_i \boldsymbol{A}_i^T \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times 2n} \\ -b_i \boldsymbol{A}_i & \mathbf{0}_{2n\times n} & \boldsymbol{A}_i \boldsymbol{A}_i^T \end{pmatrix} = \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \boldsymbol{A}_i \end{pmatrix} \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \boldsymbol{A}_i \end{pmatrix}^T, \; i = 1, 2, \ldots, n,$$

where $\boldsymbol{A}_i^T$ is the i^{th} row vector of A, i.e., $\boldsymbol{A}_i^T = \begin{pmatrix} \boldsymbol{a}(i)^T & -\boldsymbol{e}(i)^T \end{pmatrix}$. Note that

$$\begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix}^T M_i \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix} = (\boldsymbol{A}_i^T \boldsymbol{v} - b_i \pi)^2 = 0 \text{ if and only if } \boldsymbol{A}_i^T \boldsymbol{v} = b_i \pi.$$

Hence, with a simple justification, we get

$$\mathcal{D} = \left\{ Z : Z \bullet M_i = 0, \, \forall i = 1, 2, \ldots, n, \; Z \in \mathcal{CP}_{3n+1} \right\}, \tag{3.6}$$

and it can be easily verified that

$$\mathcal{D}^* = \left\{ W : W = V + \sum_{i=1}^{n} \gamma_i M_i, \; V \in \mathcal{CO}_{3n+1}, \; \gamma_i \in \mathbb{R}, \; i = 1, 2, \ldots, n \right\}. \tag{3.7}$$

Therefore, we obtain the follow formulation for the appointment scheduling problem:

$$
\begin{aligned}
\text{(S)}\quad Z_S := \quad \min \quad & \Sigma \bullet \Gamma + \boldsymbol{\mu}^T \boldsymbol{\beta} + \alpha \\
s.t. \quad & \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\boldsymbol{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\boldsymbol{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} + \sum_{i=1}^{n} \gamma_i \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \boldsymbol{a}(i) \\ -\boldsymbol{e}(i) \end{pmatrix} \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \boldsymbol{a}(i) \\ -\boldsymbol{e}(i) \end{pmatrix}^T \succeq_{co} 0 \\
& \boldsymbol{s} \in \Omega_{\boldsymbol{s}}
\end{aligned}
$$

where the decision variables are $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^n$, $\Gamma \in \mathbb{R}^{n\times n}$, $\boldsymbol{\gamma} \in \mathbb{R}^n$ and $\boldsymbol{s} \in \mathbb{R}^n$. The last constraint confines the choice of $\boldsymbol{s}$ to a feasible set $\Omega_{\boldsymbol{s}}$. For example, $\boldsymbol{s} \in \Omega_{\boldsymbol{s}}$ in our case is

$$
\sum_{i=1}^{n} s_i \leq T, \text{ and } s_i \geq 0, \forall i = 1, 2, \ldots, n, \tag{3.8}
$$

which means the time slots must be nonnegative and the total scheduled time cannot exceed the session time T. We assume $T > 0$.

We have thus obtained the central result in this topic:

Theorem 1.

$$
\min_{\boldsymbol{s}\in\Omega_{\boldsymbol{s}}} \left\{ \sup_{\tilde{\boldsymbol{u}}\sim(\boldsymbol{\mu},\Sigma)^+} \left\{ \mathbf{E}\left[f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right) \right] \right\} \right\} = Z_S
$$

Remark 3. With the compact formulation of Z_S, we are then able to analytically investigate the structure of the optimal scheduling policy. Furthermore,

computationally we can solve the relaxation of Z_S as a semidefinite programming problem as mentioned before.

3.5 Extensions

In this section, we show that our model can be extended to capture more features of the practical appointment scheduling problem, while still maintaining a formulation that is a compact convex conic optimization problem.

3.5.1 General Waiting Time Costs

In the earlier discussion, we have assumed $\rho_i = 1$ for all patients. The network flow model used in the second stage problem can be extended to cope with general waiting time costs ρ_i. This can be achieved by simply changing the in-flow at each node i from 1 to ρ_i, and the out-flow at node s from $n+1$ to $\sum_{i=1}^{n+1} \rho_i$. The reader can easily verify that the total waiting time cost is now mapped to the maximum cost flow problem in the network with the new supply and demand parameters.

3.5.2 Eye test before consultation (Late arrivals)

Suppose that the i^{th} patient in the sequence has to undertake a test prior to the consultation. The test is often handled by a nurse and can be administered immediately upon arrival. The duration of the test is random and denoted by the random variable $\tilde{l}_i$. We define the waiting time of the patients to be *the waiting time needed to consult the doctor* **after** *the test is*

administered. We also assume that the patients are seen by the doctor in the same sequence based on the appointment time, i.e., the sequence of the patients seen by the doctors is the same as the sequence of arrival. In this case, we can also use the network flow model to capture the impact of the test on the performance of the system. This is achieved by changing the cost on arcs (i, s), $i = 1, 2, \ldots, n$, from 0 to the random variables, $\tilde{l}_i$. Then the network flow solution in our model corresponds to the total waiting time cost in the system, offset by $\sum_{i=1}^{n} \tilde{l}_i$, i.e.,

$$
\begin{aligned}
f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{l}}\right) = \quad \max \quad & \sum_{i=1}^{n} \tilde{c}_i(\boldsymbol{s}) \cdot y_i + \sum_{i=1}^{n} \tilde{l}_i z_i - \sum_{i=1}^{n} \tilde{l}_i \\
s.t. \quad & y_1 - z_1 = -1 \\
& y_i - y_{i-1} - z_i = -1, \, \forall i = 2, 3, \ldots, n \\
& -y_n - z_{n+1} = -1 \\
& y_i \geq 0, \, \forall i = 1, 2, \ldots, n \\
& z_i \geq 0, \, \forall i = 1, 2, \ldots, n+1
\end{aligned}
$$

To see this, note that when $z_i = 1$, the i^{th} patient finishes the eye test and finds the doctor to be idling. This patient gets to consult the doctor at time $\tilde{l}_i$ after arrival. The waiting time is thus zero. This starts a new busy period, with the initial consultation duration given by $\tilde{l}_i + \tilde{c}_i(\boldsymbol{s})$, so we need to offset the objective by $\tilde{l}_i$. On the other hand, if after the test, the patient finds the doctor to be busy, then $z_i = 0$ in the network flow solution, and hence the waiting time is simply the length of the longest path originating from node i deducted by $\tilde{l}_i$.

Then it is clear that we can extend the definition of the cone $\mathcal{D}$[8] to capture the impact of $\tilde{\boldsymbol{l}}$ just as $\tilde{\boldsymbol{u}}$, and finally we can still arrive at a convex

[8] More precisely, the new dimension of $\mathcal{D}$ is $(4n + 1) \times (4n + 1)$.

conic optimization formulation for the appointment scheduling problem with random prior tests. Note that the effect of such tests is exactly the same as late arrivals, i.e. patients arriving at a random time after the scheduled appointment. Thus, we can also address the issue of late arrivals with the same approach described above.

3.5.3 Relationship to Scenario Planning

In our model, we assume that only the moments and covariance parameters of the service durations are known. Then our model constructs a set of scenarios, the associated probability functions, and a solution which attains the (worst case) performance objective under this set of scenarios. Our approach can be easily augmented to include specific scenarios when describing the uncertainty set for the service durations. More specifically, suppose that the system planner would like to construct the optimal schedule under the additional restrictions to include N scenarios $\boldsymbol{u}^L$ with probability p_L, such that $\sum_{L=1}^{N} p_L = p \leq 1$. Furthermore, the conditional first and second moments for the remaining scenarios are denoted by $(\boldsymbol{\mu}, \Sigma)^+$. Then our model reduces to

$$Z_P(\boldsymbol{s}) = (1-p) \sup_{\tilde{\boldsymbol{u}} \sim (\boldsymbol{\mu}, \Sigma)^+} \{\mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}})]\} + \sum_{L=1}^{N} p_L f_L(\boldsymbol{s}, \tilde{\boldsymbol{u}})$$

where $f(\boldsymbol{s}, \tilde{\boldsymbol{u}})$ is defined as before and

$$
\begin{aligned}
f_L(\boldsymbol{s}, \tilde{\boldsymbol{u}}) = \quad \max \quad & \sum_{i=1}^{n} (u_i^L - s_i) \cdot y_i^L \\
s.t. \quad & y_1^L - z_1^L = -1 \\
& y_i^L - y_{i-1}^L - z_i^L = -1, \, \forall i = 2, 3, \ldots, n \\
& -y_n^L - z_{n+1}^L = -1 \\
& y_i^L \geq 0, \, \forall i = 1, 2, \ldots, n \\
& z_i^L \geq 0, \, \forall i = 1, 2, \ldots, n+1
\end{aligned}
$$

In this way, we use a small set of scenarios to ensure that the optimal solution constructed will not perform too badly for these typical scenarios, and hence will not be overly conservative. Note that the dual to the above second stage problem can be written using the approach described earlier, together with standard linear programming duality.

When $p = 1$, Z_P reduces to the conventional stochastic optimization problem solved via the sampling method. Hence, this framework can be viewed as a bridge between the traditional stochastic optimization and modern robust optimization.

3.5.4 Generalized Conic Framework for More Support Information

For the random service time, except the moment conditions, we only require that they must be nonnegative. In general, there may be other conditions that the system planner would like to impose on the random service time, like a boundedness condition, etc. Our model provides a natural way to incorporate more support information through the construction of the cone

$\mathcal{D}$. Recall in equation (3.6), we express $\mathcal{D}$ as

$$\mathcal{D} = \{Z : Z \bullet M_i = 0,\, \forall i = 1, 2, \ldots, n,\, Z \in \mathcal{CP}_{3n+1}\}\,.$$

We can view $\mathcal{D}$ as the intersection of the completely positive cone $\mathcal{CP}_{3n+1}$ with

$$\mathcal{M}^i := \{Z : Z \bullet M_i = 0\}\,,\; i = 1, 2, \ldots, n.$$

While the network conservation constraints are embedded within $\mathcal{M}^i$, $\mathcal{CP}_{3n+1}$ captures both the non-negativity constraints for the network flow variables and nonnegative support requirement of the random service time. Thus, it appears intuitive for us to augment $\mathcal{CP}_{3n+1}$ if we want to incorporate more support conditions. In order to develop a more general framework, we need the following lemma, which can be easily verified by the definition of a dual cone.

Lemma 1. Suppose $\mathcal{K}^k \subseteq \mathbb{R}^{n\times n}$, $k = 1, 2, \ldots, m$, are closed convex cones. Let the dual cone of $\mathcal{K}^k$ be $\mathcal{K}^{k*}$. Then the dual cone of the following cone

$$\mathcal{K}_n := \bigcap_{k=1}^{m} \mathcal{K}^k = \left\{A \in \mathbb{R}^{n\times n} : A \in \mathcal{K}^k,\, k = 1, 2, \ldots, m\right\}$$

is

$$\mathcal{K}_n^* := \sum_{k=1}^{m} \mathcal{K}^{k*} = \left\{ A \in \mathbb{R}^{n\times n} : \exists A_k \in \mathcal{K}^{k*},\, k = 1, 2, \ldots, m,\, \text{such that } A = \sum_{k=1}^{m} A_k \right\}.$$

With Lemma 1, one can easily derive the expression of the dual cone of $\mathcal{D}$ as shown in equation (3.7) by recognizing that the dual cone of $\mathcal{M}^i$ is $\mathcal{M}^{i^*} := \{\gamma_i M_i : \gamma_i \in \mathbb{R}\}$. Thus, as long as the extra support conditions can be characterized with some conic constraints and their dual cones are compactly representable, we could still obtain a single conic optimization formulation for the appointment scheduling problem.

For example, if the system planner would like to add some boundedness conditions for the the random service time, which is characterized by the following ellipsoid constraint, i.e.,

$(\tilde{\boldsymbol{u}} - \overline{\boldsymbol{u}})^T \overline{Q} (\tilde{\boldsymbol{u}} - \overline{\boldsymbol{u}}) \leq r$ with probability 1, for some $\overline{Q} \in S_n \subseteq \mathbb{R}^{n\times n}, \overline{\boldsymbol{u}} \in \mathbb{R}^n$ and $r \in \mathbb{R}$.

This constraint restricts the random service time to lie in an ellipsoid of size r centered at $\overline{\boldsymbol{u}}$. Using the probabilistic interpretation of $Z \in \mathcal{D}$, we can transform this condition into the following conic constraint on Z,

$$Z \in \Theta :=$$
$$\operatorname{conv}\left\{ \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix} \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix}^T : \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix}^T \begin{pmatrix} r - \overline{\boldsymbol{u}}^T\overline{\boldsymbol{u}} & \overline{\boldsymbol{u}}^T\overline{Q} & \mathbf{0}_{1\times 2n} \\ \overline{Q}\overline{\boldsymbol{u}} & -\overline{Q} & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \begin{pmatrix} \pi \\ \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix} \geq 0, \begin{array}{r} \pi \in \mathbb{R} \\ \boldsymbol{t} \in \mathbb{R}^n \\ \boldsymbol{v} \in \mathbb{R}^{2n} \end{array} \right\}.$$

Then the dual cone of Θ can be easily obtained using S-Lemma, i.e.,

$$\Theta^* := \left\{ V \in \mathbb{R}^{(3n+1)\times(3n+1)} : \exists \tau \geq 0, \text{ such that } V - \tau \begin{pmatrix} r - \overline{\boldsymbol{u}}^T \overline{\boldsymbol{u}} & \overline{\boldsymbol{u}}^T \overline{Q} & \mathbf{0}_{1\times 2n} \\ \overline{Q}\overline{\boldsymbol{u}} & -\overline{Q} & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \succeq 0 \right\},$$

which will translate into an extra semidefinite constraint in the final formulation of the appointment scheduling problem, since the resulted dual cone of $\mathcal{D}$ becomes

$$\mathcal{D}^* = \left\{ W : W = V_1 + V_2 + \sum_{i=1}^{n} \gamma_i M_i,\ V_1 \in \mathcal{CO}_{3n+1},\ V_2 \in \Theta^*,\ \gamma_i \in \mathbb{R},\ i = 1, 2, \ldots, n \right\}.$$

Following the similar argument in the proof of Proposition 4, one can easily verify that the main result of our model (i.e., Theorem 1) still holds with the modified $\mathcal{D}^*$ as shown above.

3.6 Model Analysis

Our model provides a single deterministic convex formulation to solve a two stage stochastic optimization problem. To the best of our knowledge, this model is the first of its kind. Furthermore, as shown in the development of the conic optimization model, the optimal solution to problem (C) has a natural probabilistic interpretation under the worst case distribution. Note that we can obtain the values of those (primal) variables in (C) by taking the dual of (S). Together with the network flow formulation of the waiting

time experienced, they provide a new way to obtain some insights into the structure of the optimal appointment schedule. In the rest of this section, we show that the solution obtained from this deterministic model retains many of the intuitive properties of the optimal schedule under more realistic probabilistic consultation service distributions. To maintain the flow of this topic as well as to keep it succinct, we relegate all of the proofs in this section to Appendix D. In terms of notation, we use the asterisk sign $(*)$ to indicate the respective optimal solution. For example, s_i^* denotes the optimal solution of s_i in problem (S).

We show first that if there is a need to bunch the arrival of patients together, then it is optimal to bunch the arrivals at the end of the session. This is intuitive because whenever the consultation time is modeled by a non-negative distribution, if bunching occurs for the $(i-1)^{st}$ and i^{th} patient, but not the $(i+1)^{st}$ patient, then it is optimal to schedule the arrival of the i^{th} patient slightly later and keep the schedule of the $(i+1)^{st}$ patient unchanged. The reason is obvious since the i^{th} patient has to wait almost surely if she comes at the same time as the $(i-1)^{st}$ patient. The optimal schedule in our model retains this feature.

Proposition 5. Let the waiting time costs and overtime cost be strictly positive. In any optimal solution $\boldsymbol{s}^*$ to problem (S), let I be the set of allocated service times, which are zero, i.e., $I := \{i : s_i^* = 0\}$. Then $I = \{n - |I| + 1, \ldots, n-1, n\}$, i.e., I is the last $|I|$ members of $\{1, 2, \ldots, n\}$.

Remark 4. Note that in the numerical studies that we will present in Section

3.7, there is also a bunching effect appearing at the beginning of the session. However, those slots at the beginning of the session are not exactly zero, but very small positive values. Such phenomenon would rise in the optimal policy for a heavily congested system, where every patient has to wait almost surely (with probability close to 1), details of which will be analyzed in Section 3.7.

In a practical settings, the nonnegativity constraints on the consultation slots (i.e., $s_i \geq 0, \forall i = 1, 2, \ldots, n$) enforce that all the appointment times are within the consultation session (T). Intuitively, if the system is heavily congested, it may be optimal to schedule some patients to arrive after time T. To incorporate this into our model, we may remove the nonnegativity constraints on the consultation slots. The next proposition shows that if these nonnegativity constraints are removed, only the last slot (s_n) can be negative in the optimal schedule as long as the costs of waiting time and overtime are strictly positive. Note that the scheduled arrival time of the n^{th} patient is $\sum_{i=1}^{n-1} s_i$ and is therefore larger than T if $s_n < 0$, because $\sum_{i=1}^{n} s_i = T$. Furthermore, the constraint $\sum_{i=1}^{n} s_i = T$ ensures that the counting of the doctor's overtime starts from time T, and $s_n < 0$ in the network flow structure indicates that the doctor's overtime (i.e., the $(n+1)^{th}$ patient's waiting time) is at least $-s_n > 0$.

Proposition 6. Suppose the nonnegativity constraints on consultation slots (i.e., the second set of constraints in equation (3.8)) are removed. When the waiting time costs and overtime cost are strictly positive, in the optimal solution to problem (S), there is at most one negative slot. Furthermore, if

this negative slot exists, it must be the last one, i.e., $s_i^* > 0$, $\forall i = 1, 2, \ldots, n-1$, and $s_n^* < 0$.

We investigate next the probability of a patient arriving at the scheduled time to find the system busy. From Figure 3.2, the flow y_i merges with ρ_i at node i. The probability that this combined flow goes through arc $(i, i-1)$ is exactly the probability that the i^{th} patient has to wait. Otherwise, the flow on arc $(i, i-1)$ would be zero, which indicates that the waiting time cost is zero for the i^{th} patient since arc (i, s) has zero flow cost. More precisely,

$$\begin{aligned}
\mathbf{E}[y_{i-1}(\boldsymbol{s}, \tilde{\boldsymbol{u}})] &= \mathbf{E}[\mathbf{E}[y_{i-1}(\boldsymbol{s}, \tilde{\boldsymbol{u}})|y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}})]] \\
&= \mathbf{E}[(y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}}) + \rho_i) \cdot Pr\{i^{th} \text{ patient has to wait}\}] \\
&= (\mathbf{E}[y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}}] + \rho_i) \cdot Pr\{i^{th} \text{ patient has to wait}\}
\end{aligned}$$

$$\implies Pr\{i^{th} \text{ patient has to wait}\} = \frac{\mathbf{E}[y_{i-1}(\boldsymbol{s}, \tilde{\boldsymbol{u}})]}{\mathbf{E}[y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}})] + \rho_i}.$$

Since the optimal $\boldsymbol{s}^*$ is selected to minimize

$$\mathbf{E}\left[f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right)\right] = \mathbf{E}\left[\sum_{i=1}^{n} \tilde{c}_i(\boldsymbol{s}) y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}})\right] = \mathbf{E}\left[\sum_{i=1}^{n} (\tilde{u}_i - s_i) y_i(\boldsymbol{s}, \tilde{\boldsymbol{u}})\right],$$

From the first order optimality conditions, we expect that at the optimal $\boldsymbol{s}^*$, if $s_{i-1}^* > 0$ and $s_i^* > 0$, then $\mathbf{E}[y_{i-1}(\boldsymbol{s}^*, \tilde{\boldsymbol{u}})] = \mathbf{E}[y_i(\boldsymbol{s}^*, \tilde{\boldsymbol{u}})]$. This holds indeed for the optimal schedule obtained using our model.

Proposition 7. If in the optimal solution to problem (S), the allocated service time slots are strictly positive, (i.e., $s_i^* > 0$, $\forall i \in I \subseteq 1, 2, \ldots, n$), then the

network flow solution must satisfy $\mathbf{E}[y_i(\boldsymbol{s}^*, \tilde{\boldsymbol{u}})] \equiv K$, $\forall i \in I$, where K is some nonnegative constant.

Combining the propositions established thus far, we can derive an important optimality condition for an appointment system:

Theorem 2. Suppose in the optimal solution to problem (S), the allocated consultation slots are strictly positive for the first k patients, (i.e., $s_i^* > 0$, $\forall i = 1, 2, \ldots, k$, where $0 < k \leq n$). Furthermore, if $\rho_i \equiv \rho$, for some constant $\rho > 0$, for all $i = 1, 2, \ldots, k$, then the probabilities of waiting for the service are the same for all the patients from $i = 2, \ldots, k$, under the optimal worst case distribution.

Remark 5. Note that the optimality condition stated in the above theorem is independent of the sequence of the patients. This property of the optimal schedule is particularly useful for the patients: there is little incentive to choose between the slots in the clinical session if the objective is to minimize the chances of waiting for the service.

3.7 Computational Results

All the computational studies are carried out in MATLAB on a Dell desktop (Core 1.86 GHz and 3GB of RAM). We solve the simplest form of SDP relaxation of the COP and CPP as shown in equation (3.3). In MATLAB,

we use YALMIP as the programming interface with SDPT3 as the underlying SDP solver (cf. Löfberg (2004), Toh et al. (1999), Tutuncu et al. (2003)).

Note that expressing a problem as a COP or CPP and relaxing it only partially resolve the difficulty of the problem, because even solving a large-scale SDP can be computationally inhibitive. Since our model lifts the original problem into a cone with higher dimensions, the current computational power limits the size of the problem instance we can solve to around 36 patients. While it is an interesting challenge to push the computational limit of this approach further, we leave this to future research. By "large-scale problem", we mean problems that involve hundreds or even thousands of variables. Fortunately, in practice we usually will not encounter such large sized problems. In the eye clinic case, we only need to schedule 36 patients for the whole morning session.

In what follows, we use extensive numerical experiments to provide a glimpse into the performance of the optimal scheduling solutions obtained using our model.

3.7.1 Comparison with near-optimal solutions

In this section, we test the performance of our model against a set of near optimal solutions given in Denton & Gupta (2003). Table 3.1 lists the near optimal schedules given in that paper, for 7 jobs with identically independent distributed service time (Uniform$(0, 2)$) under different cost structures and fixed session length $T = 7$. The waiting time costs are identical among all the patients. In their numerical results, the optimality gap is less than 1%.

We compute problem (S) to obtain the optimal schedule that minimizes the worst-case cost under all distributions with mean 1 and standard deviation $1/\sqrt{3}$. The results of our model are presented in Table 3.2. Note that in Denton & Gupta (2003), the objective function is the weighted sum of total waiting time, idle time and overtime of the doctor, while in our model the objective function does not include the cost of idle time. According to Proposition 1 in Denton & Gupta (2003) (similar to our argument in Section 3.4.1), we can transform the optimal scheduling problem in Denton & Gupta (2003) equivalently into our problem by combining the cost of idle time and overtime. Since Denton & Gupta (2003) allows negative schedules, we remove the non-negativity constraints in equation (3.8) when solving problem (S) for a fair comparison.

(ρ_1, ρ_{n+1})	(3,14)	(5,12)	(7,10)	(3,12)	(5,10)	(7,8)	(3,10)	(5,8)	(7,6)
s_1	0.61	0.83	1.06	0.65	0.88	1.14	0.72	1.00	1.25
s_2	1.09	1.18	1.27	1.11	1.22	1.34	1.13	1.25	1.38
s_3	1.08	1.20	1.26	1.11	1.24	1.31	1.12	1.25	1.38
s_4	1.09	1.20	1.27	1.13	1.22	1.32	1.13	1.25	1.38
s_5	1.07	1.10	1.21	1.05	1.14	1.25	1.08	1.19	1.35
s_6	0.94	1.00	1.16	0.96	1.01	1.20	0.94	1.07	1.24
s_7	1.14	0.50	-0.23	1.01	0.31	-0.56	0.89	-0.01	-0.98

Tab. 3.1: Optimal schedules from Denton & Gupta (2003) under different cost structures

Next, we compare the total waiting time costs under the schedules given in Tables 3.1 and 3.2 through Monte Carlo simulation. In evaluating our model, the service duration of each patient is generated under four common distributions used in practice: uniform, normal, two-point and Gamma distribution, with mean 1 and standard deviation $1/\sqrt{3}$. All 9 different cost

(ρ_1, ρ_{n+1})	(3,14)	(5,12)	(7,10)	(3,12)	(5,10)	(7,8)	(3,10)	(5,8)	(7,6)
s_1	0.35	0.87	0.94	0.52	0.89	0.99	0.76	0.92	1.05
s_2	1.32	1.09	1.16	1.22	1.10	1.20	1.08	1.13	1.26
s_3	1.05	1.17	1.25	1.08	1.19	1.30	1.11	1.22	1.38
s_4	1.12	1.29	1.38	1.16	1.31	1.44	1.21	1.35	1.53
s_5	1.20	1.31	1.36	1.23	1.31	1.42	1.26	1.33	1.50
s_6	1.17	1.27	1.20	1.20	1.20	1.25	1.24	1.18	1.33
s_7	0.79	0.00	-0.29	0.58	0.00	-0.61	0.33	-0.14	-1.04

Tab. 3.2: Optimal schedules from our model under different cost structures

structures are tested. 50,000 rounds of simulation are executed for each of the 36 scenarios (4 distributions $\times$ 9 costs structures)[9]. The average total costs under different scenarios are then compared with the corresponding benchmark schedules given by Denton & Gupta (2003) under the uniform distribution. As shown in Table 3.3 the schedules obtained from our model work phenomenally well when evaluated against the benchmarks. The average total costs under our model is close to that of Denton & Gupta (2003) even under different distributions. The gaps are within 2% and most of them are less than 1%. Moreover, it is worthwhile to point out that the average total costs of our schedules do not vary much under different distributions.

(ρ_1, ρ_{n+1})	(3,14)	(5,12)	(7,10)	(3,12)	(5,10)	(7,8)	(3,10)	(5,8)	(7,6)
Benchmark	23.32	27.03	28.50	21.42	24.51	25.02	19.43	21.69	20.94
Uniform	23.55	27.62	28.89	21.55	24.79	25.48	19.60	21.94	21.48
Normal	23.57	27.77	28.98	21.63	24.92	25.55	19.72	22.03	21.53
Two point	24.00	28.64	30.20	21.95	25.81	26.56	20.23	22.91	21.89
Gamma	22.73	27.53	28.87	20.93	25.08	25.84	19.48	22.10	22.21

Tab. 3.3: Comparison of the average total costs between the schedules obtained by our model and Denton & Gupta (2003) under different distributions

[9] We obtain similar results through the test under a larger set of distributions as well, but only the four most commonly used distribution are reported in this paper.

3.7.2 Empirical Study in an Eye Clinic

In this subsection, we examine the performance of the appointment system in the eye clinic and apply the methodology we develop in section 3.5.4 to improving the performance of the system. We present numerical results based on data collected from the eye clinic and discuss pertinent managerial insights from our model.

Tan Tock Sin hospital, built in 1844 by the entrepreneur Tan Tock Sin, is the second largest hospital in Singapore. The hospital's specialist clinics serve around 1,500 patients daily and its Emergency Department attends around 400 patients daily, making it one of the busiest clinics and emergency departments. Hence, efficient appointment systems are crucial for the hospital operation, to to regulate the usage of the costly equipment and precious resources in the system, and to enhance its service level and increase customer satisfaction.

Our research is motivated by a visit to an eye clinic in Tan Tock Sin hospital. In the clinic, there are two consultation sessions per day, each lasting four hours from 8am to 12pm and from 1pm to 5pm. The number of doctors available per session is around two to seven. Each doctor has to handle 20 to 30 patients per session. Patients are classified into "new patients" and "repeat patients". New patients refer to those who visit the doctor for the treatment of a new problem and whose eye conditions are unknown to the doctors they see; while repeat patients are those who have visited the clinic before and who go there for follow-up checks. Their appointment strategy is to 1) assign new patients to come at the beginning of the session; 2) allocate

5 minutes per patient slot for one hour, followed by a half hour break.

We collect a dataset on the visits of 1021 patients during 7 working days from 22 May, 2006 to 30 May, 2006. Among them, there are 201 new patients (around 20%) and 820 repeat patients (around 80%). The data set consists of the date, patient type, patient's NRIC, type of services, starting and ending time of each service, service durations etc. The clinic uses the "Turnaround Time" (TAT) as one key performance indicator. The number of doctors varies from session to session. Sometimes there are only 2-3 doctors on duty while some day there are 6-7 doctors. The number of patients varies accordingly. Table 3.4 summarizes patients classifications, workload, and median TAT in each of the 7 days[10]. From the table we can see that the ratio of new patients to repeat is roughly around 1:4 and the workload affects the TAT significantly.

Date	22-05	23-05	24-05	25-05	26-05	29-05	30-05
No. of Doctors	5	4.5	5	5	3	4	5.5
No. of Patients	136	162	155	126	106	146	190
New Patients	25	36	25	28	18	19	50
Repeat Patients	111	126	130	98	88	127	140
Median TAT	1:27	1:39	1:33	1:10	1:06	1:32	1:45

Tab. 3.4: Patients Classifications and Median TAT

A typical visit to the eye clinic usually includes the following several steps[11]:

- Registration
- Vision Test (Visual Acuity, Eye Pressure etc.)

[10] Note that the number of doctors per day takes average on the number of patients in the morning session and afternoon session.

[11] Please refer to Liang (2006)for more detailed descriptions.

- Consultation
- Specialized Eye Tests (Refraction, Dilation, Visual Field Test etc.)
- Payment and scheduling of follow-up appointment

Besides, a patient (especially a new one) may go through several consultation during one visit. There is usually one major consultations and normally takes the longest time. We record the duration of the longest consultation of each patient and assume it is the length of a major consultation. Figure 3.3 depicts the variations of consultation durations of new and repeat patients. It can be seen that the mean and variance of the consultation times of the new patients are noticeably higher than those of repeat patients. Specifically, the mean and standard deviation of the consultation time of the repeat patients are 6.24 minutes and 6.0 minutes respectively, while those for the new patients are noticeably higher, with a mean of 9.97 minutes and a standard deviation of 7.6 minutes.

In the following, I apply the methodology we develop in section 3.5.4 to obtaining optimal schedules for the eye clinic. I discuss the structural properties of the optimal schedules under different overtime cost and propose implementable heuristics. Then I run computational studies to compare the performances of the optimal schedules, heuristic solutions and the current appointment strategy.

In this experiment, we assume that one session lasts for 150 minutes. This mimics the current practice with one hour block, followed by a half hour break and then another one hour block. During one session, 24 patients are scheduled to arrive in the clinic, with 5 new patients arriving before 19

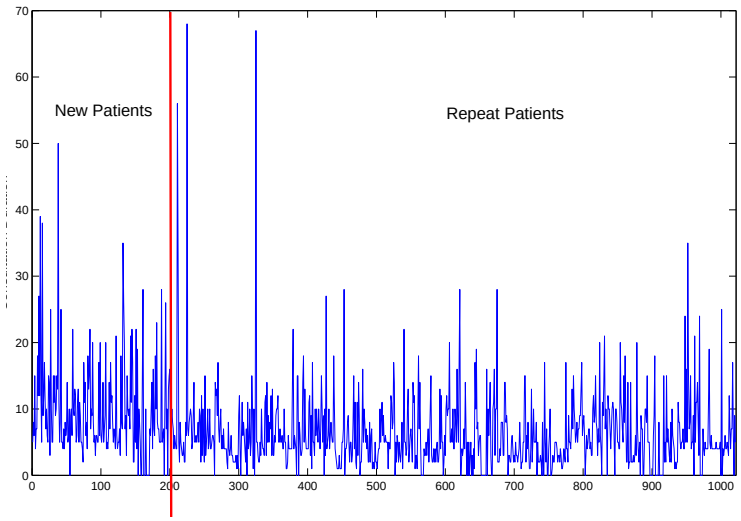

Fig. 3.3: Consultation durations of new and repeat patients

repeat patients. The consultation durations follow the distributions with the mean and standard deviation as estimated by the empirical data. Note that the sum of mean service durations of all patients is 168.41 minutes, which is larger than the session length. This indicates that the system may be heavily congested.

The patient's waiting time cost (ρ_i) is assumed to be identical among all the patients and normalized to 1. We test various overtime costs, i.e., $\rho_{n+1} = 1$, 20 or 40. Figure 3.4 plots the optimal schedules obtained by our model under different ρ_{n+1}.

It is interesting to note that the optimal schedules exhibit the pattern of "Bailey's Rule + Break". First, the optimal schedule allocates near zero time slot to the first few patients. Although Proposition 5 indicates that all the zero time slots should be placed at the end of the session, the time slots

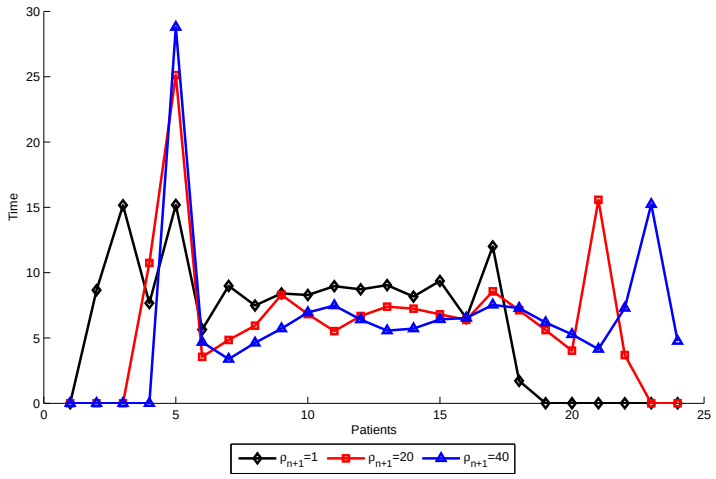

Fig. 3.4: Optimal schedule when ρ_{n+1} is equal to 1, 20 and 40, given ρ_1=1

for the first few patients are indeed strictly positive but extremely small, as the system is heavily congested and the overtime costs are large enough to induce such scheduling rules. The second outstanding feature is that, after serving the group of new patients, a break is inserted before switching to the group of repeat patients with lower variability. To confirm this feature, we run another group of experiments with 3 classes of patients. Similar patterns are observed - breaks are inserted after serving the first and the second class of patients.

One drawback of the optimal schedule is that it is generally not practical and is non-intuitive. To fix this problem, we try to use the above insights to develop a simple but effective appointment schedule. In the current practice, each patient is assigned with an equal interval of 5 minutes and a 30 minutes

break is inserted after seeing 12 patients. We simply modify the "*Current Practice*" by replacing the 30 minutes break after serving all the new patients, i.e. after the 5^{th} patient. We call this schedule "*Modified Practice*".

In a more advanced system design, we allow the allocated service intervals to vary according to the mean service duration of each patient, denoted as the "*Varying Interval*" schedule. To resemble the optimal schedule (under $\rho_{n+1} = 1$), we assign zero time slots to the first patient and the last six patients. Other patients are assigned with time slots by rounding up their mean service durations, i.e., 10 minutes for a new patient and 7 minutes for a repeat patient. The remaining time is combined and inserted after the 5^{th} patient as a break.

	Uniform	Normal	Two-points	Gamma
Optimal Schedule	352.58	349.80	355.37	352.78
Current Practice	564.13	560.18	570.31	535.37
Modified Practice	485.36	479.95	491.95	462.44
Varying Interval	358.24	353.61	363.60	354.83

Tab. 3.5: Average total waiting time cost under different scheduling policies when $\rho_1 = 1$ and $\rho_{n+1} = 1$

The simulated performance of various policies under different service time distributions are shown in Table 4. Implementing a schedule resembling the optimal solution dramatically decreases the total waiting time cost by about 35% as compared to the current practice. Interestingly, it seems that one can significantly improve the performance of the system by simply inserting a break after serving one class of patients in the optimal scheduling. The easily implemented "Varying Interval" strategy makes it quite attractive for the practical considerations.

Note that the above simulation results are obtained under $\rho_{n+1} = 1$. In most environment, the overtime cost ρ_{n+1} is likely to be large and should be proportional to the number of patients seen in the clinic. The choice of $\rho_{n+1} = 1$ is thus a conservative estimate and assumes the doctor places small penalty on the overtime work. In what follows, we summarize the features of optimal schedules when ρ_{n+1} increases. The pattern of "Bailey's Rule + Break" seems to be quite robust no matter how the the overtime cost changes. Besides this, Figure 3.4 also illustrates several interesting features: As ρ_{n+1} increases,

- more patients are assigned with near zero consultation time slots at the beginning of the session;
- fewer patients are assigned with zero time slots at the end of the session;
- a longer time slot is assigned to the last patient.

Intuitively, all these features benefit a clinic that prefers a shorter overtime. Consequently, patients may suffer from longer waiting times as a result.

As we can see, the optimal properties persist as the overtime cost ρ_{n+1} increases. One question is that whether we can still design efficient appointment systems with the help of the optimal properties under different overtime costs. To answer this question, we first solve the optimal schedules when ρ_{n+1} is 2, 5, 10, 20, 50, and 100, and then create the "Varying Interval" schedules using the following heuristics: allocate zero time slots to those clustering patients (with zero or close to zero time slots) at the beginning and the end of the session, assign the rest of the patients their mean consultation durations,

and insert the remaining time as a break after the 5th patient.We simulate the total costs of the "Varying Interval" schedules and compare with current practice. Table 3.7.2 records the efficiency gains under different overtime costs ρ_{n+1}. The percentage savings decrease as rho_{n+1} is increased. The efficiency gain drops to around 30% when $\rho_{n+1} = 2$, to around 13% when $\rho_{n+1} = 10$, and to around 10% when $\rho_{n+1} = 100$. Since a higher overtime cost indicates larger total cost, a 10% efficiency gain when $\rho_{n+1} = 100$ can save around 360 minutes in total waiting time. Hence, although efficiency gain drops as ρ_{n+1} increases, employing the "Varying Interval" schedule can still ensure significant efficiency improvements in the clinic. We simulate the performances of the "Varying Interval" schedules and compare them against those under the current strategy.

	Percentage Increase			
Overtime	Uniform	Normal	Two-points	Gamma
1	36.5%	37.1%	36.3%	33.7%
2	31.7%	32%	31.4%	28.5%
5	24%	24.3%	23.9%	21%
10	13.3%	13.4%	13.3%	10.1%
20	7.1%	7.3%	7.3%	4.2%
50	7.3%	6.8%	7.9%	5.5%
100	11.4%	11.5%	12.2%	9.7%

Tab. 3.6: Efficiency gains under different overtime costs

3.8 Sequencing Problem

We have shown that the scheduling problem can be effectively solved using a simple convex program. We discuss in the rest of this section some insights on the optimal sequence of arrival of the patients. We assume $s_i = \mu_i$ to

remove the needs to address the scheduling decision, and focus solely on the sequencing problem. We want to determine the sequence to minimize the total waiting time, with $\rho_i = 1$ for $i = 1, \ldots, n+1$, in our appointment problem. In particular, we address the question: Is it optimal to sequence the patients with smaller variance to arrive earlier in the session?

Although many current research conjecture that the smaller-variance-first rule might be optimal, the following example, unfortunately, shows that this is not true in general.

Assume the service durations $\{\tilde{\mu}_i\}$ are independent. Let $\tilde{\mu}_1 = 0$ or 2 with equal probability, $\tilde{\mu}_2, \tilde{\mu}_3 = 0$ or 4 with equal probability, and $\tilde{\mu}_k = 0$ or 6 with equal probability for $k > 3$. In this case, $P(\tilde{c}_1 = \pm 1) = \frac{1}{2}$, $P(\tilde{c}_j = \pm 2) = \frac{1}{2}$, for $j = 2, 3$, and $P(\tilde{c}_k = \pm 3) = \frac{1}{2}$, for $k = 4, \ldots, n$. We compare the performance of the sequence $\{1, 2, \cdots, n\}$ and another obtained by switching 1 and 2 in the sequence. Note that patients in the first sequence are ordered in non-decreasing order of the variances. We ran simulation and plot the difference in the performance, (i.e., the difference in total waiting time), as a function of the number of patients n, in Figure 3.5.

As the number of patients is small, say n below 20, scheduling patients with smaller variance first is generally better in this example. Surprisingly, this behavior changes as n increases, and for a large enough n, putting patient 2 in front of patient 1 is now beneficial in reducing total waiting time! Consequently, sequencing patients in increasing variance is no longer optimal. The simulation result also suggests that the optimal

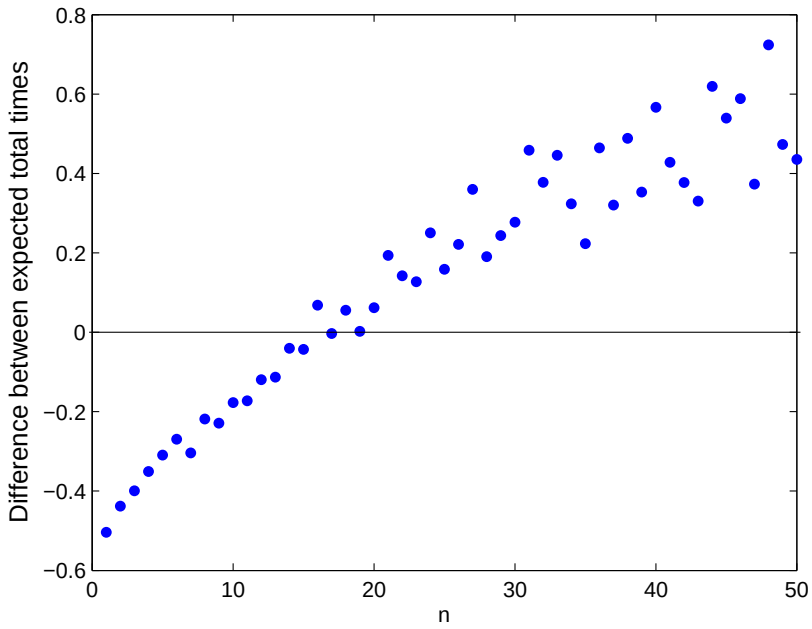

Fig. 3.5: Difference in Performance w.r.t n

sequence is affected by the number of patients in each class, and hence sequencing patients by looking at pair of patients in isolation through stochastic ordering is probably a futile attempt.

To add to the perplexity of the results, we show next that under the deterministic model where s_i is set to a constant (e.g. patients scheduled to arrive in constant interval), then knowing the service duration $\tilde{\mu}_i$ *in advance* does not make the sequencing problem any easier! In fact, under this deterministic model, Vanden (1997) have shown earlier that when the objective coefficients ρ_i are allowed to take arbitrary values, to determine the optimal sequence is equivalent to solve a nonlinear knapsack problem and is thus NP-hard. Surprisingly, we show next that the problem in fact remains NP-hard even when ρ_i's are identical.

Theorem 3. The appointment scheduling problem is NP-hard in the strong sense, even if the allocated appointment interval S_j is constant for all j, and $\rho_i = 1$ for all i.

We refer the readers to Appendix I for a formal proof of this result.

In the rest of this section, we describe how the proposed approach can be used to address this class of sequencing problem, even when scheduling decision has to be made in conjunction with the sequencing decision. Let

(P)' $$Z'_P = \sup_{\boldsymbol{U}\sim(\boldsymbol{\mu},\Sigma)^+} \{\mathbf{E}[f(\boldsymbol{s},\sigma)]\}$$

where $f(\boldsymbol{s},\sigma)$ is the cost of the second stage network flow model after fixing schedule $\boldsymbol{s}$ and sequence σ.

To apply our approach on the appointment design problem, we introduce scheduling and sequencing decision variables $\boldsymbol{s}$ and $P = (p_{ij})$ into the model in the following way:

Theorem 4. (P)' can be solved as the following completely positive program, i.e., $Z'_P = Z'_C$:

$$
\begin{aligned}
\text{(C)'} \qquad Z'_C = \max \quad & X \bullet P - \boldsymbol{s}^T \boldsymbol{y} \\
s.t. \quad & \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix}^T \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = -1, \, \forall j = 1, 2, \ldots n \\
& \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix}^T \begin{pmatrix} Y & W^T \\ W & Z \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix} = 1, \, \forall j = 1, 2, \ldots n \\
& \begin{pmatrix} 1 & \boldsymbol{\mu}^T & \boldsymbol{y}^T & \boldsymbol{z}^T \\ \boldsymbol{\mu} & \Sigma & X^T & V^T \\ \boldsymbol{y} & X & Y & W^T \\ \boldsymbol{z} & V & W & Z \end{pmatrix} \succeq_{cp} 0
\end{aligned}
$$

where the decision variables are $\boldsymbol{y}$, $\boldsymbol{z} \in \mathbb{R}^n$, Y, W, Z, X, $V \in \mathbb{R}^{n \times n}$, and $P = (p_{i,j}) \in \{0, 1\}^{n \times n}$ is a known permutation matrix given by

$$
p_{i,j} = \begin{cases} 1 & , \text{ if } \sigma(j) = i \\ 0 & , \text{ otherwise} \end{cases}, \; i, j = 1, 2, \ldots, n.
$$

This formulation models the expected waiting cost of the appointment system, when the sequencing decision $P = (p_{i,j})$ and scheduling decision $\boldsymbol{s}$ are fixed. It replaces the objective function $tr(X)$ by $X \bullet P$, due to the sequencing consideration. The corresponding copositive cone program is now:

$$
\begin{aligned}
(S^2)' \quad Z'_{S^2} := \quad & \min \quad \Sigma \bullet \Gamma + \boldsymbol{\mu}^T \boldsymbol{\beta} + \alpha \\
s.t. \quad & \begin{pmatrix}
\sum_{j=1}^{n} (-u_j + v_j) + \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{0} \end{pmatrix}^T - \sum_{j=1}^{n} u_j \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix}^T}{2} \\
\frac{\boldsymbol{\beta}}{2} & \Gamma & \begin{pmatrix} -P/2 \\ O_n \end{pmatrix}^T \\
\frac{\begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{0} \end{pmatrix} - \sum_{j=1}^{n} u_j \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix}}{2} & \begin{pmatrix} -P/2 \\ O_n \end{pmatrix} & -\sum_{j=1}^{n} v_j \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_j \\ -\boldsymbol{e}_j \end{pmatrix}^T
\end{pmatrix} \succeq_{co} 0 \\
& \sum_{j=1}^{n} p_{i,j} = \sum_{j=1}^{n} p_{j,i} = 1, \, \forall i = 1, 2, \ldots, n \\
& p_{i,j} \in \{0, 1\}, \, \forall i, j = 1, 2, \ldots, n \\
& \boldsymbol{s} \in \Omega_{\boldsymbol{s}}
\end{aligned}
$$

When the sequencing becomes parts of the decision variables, due to its discrete nature (n^2 binary variables), the time consumed in searching for optimal sequence (e.g. using a Branch and Bound (B&B) type method) increases exponentially in the size of the instance. We developed a simple B&B code to take advantage of the special structure of the problem by adding some symmetry breaking constraints, and can solve the sequencing problem for up to 8 customers efficiently.

3.8.1 Numerical Results

Our earlier numerical examples have debunked the conjecture on the optimality of the smaller-variance-first rule. However, what is the structure of the optimal sequencing policy? We use a set of numerical experiments to provide a glimpse to the answer of this question.

The numerical example assumes 6 patients in the system, with identical mean consultation duration of 5 time units, but with different standard deviations ([1 1 2 2 3 3]). By fixing the allocated service time to be the mean ($\mu = 5$), we obtain the optimal sequences by solving the CPCMM model. In Figure 3.6, we observe that when the ratio between waiting time cost (ρ_i) and overtime cost (ρ_{n+1}) is 1 : 1, smaller-variance-first rule is indeed optimal. However, as the overtime cost is sufficiently high as compared to the waiting time cost (e.g., 1:100), the optimal sequence appears to be "U-shaped" in terms of the variability of the service durations. Namely, the patients with larger variances are either assigned to the beginning or the end of the session.

Monte Carlo simulation results indeed show that under the "U-shape" sequencing rule, the expected total cost is smaller than that under the smaller-variance-first rule. The above structure continues to hold for many different sets of parameters in the experiments. We conjecture that it holds in general:

> [Conjecture]: When the allocated service time for each patient is set to the mean service time and the overtime cost is sufficiently high, the optimal sequence exhibits a U-shape pattern.

We leave the resolution to the above conjecture to future research.

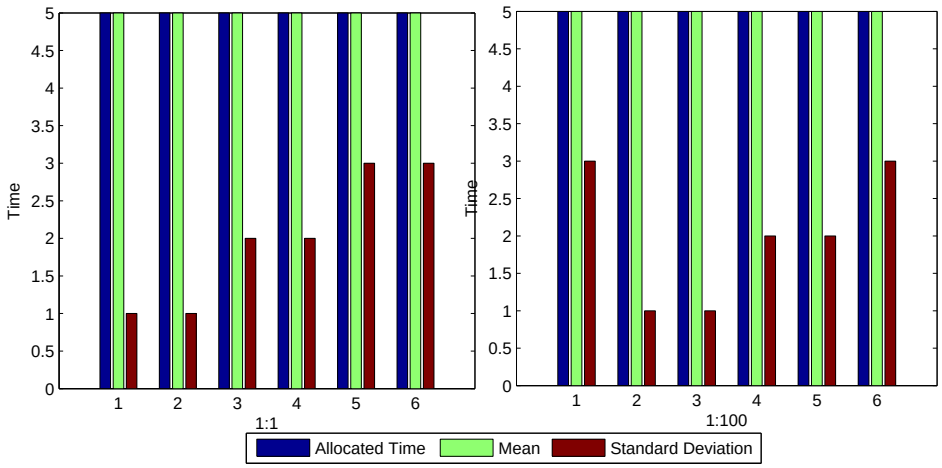

Fig. 3.6: Optimal sequencing under different cost structure and fixed schedule

3.9 Conclusion

We propose a novel approach to deal with the difficult patient scheduling and sequencing problem. Instead of planning against a fixed service distribution, we plan against a canonical set of service distributions with the same mean and covariance parameters. The canonical distribution is "constructed" via a copositive cone program. In this way, we reduce a difficult two stage stochastic programming problem into a single stage convex programming problem. Through extensive simulations we show that the optimal schedules solved under the "worst case" give near-optimal solutions when the objective is to minimize expected total cost. This approach allows us to shed some light on the structure of the optimal schedule and sequence, which we can readily modify to obtain practical and efficient schedule and sequence.

The approach can be generalized to deal with the situation when the patients need to undergo a test (random duration) prior to the consultation,

a pertinent feature in many eye clinic. The network flow approach can also be conceivably extended to deal with other practical considerations in a clinical environment. There are however several limitations with this approach - the computational difficulty associated with solving large scale SDP limits our ability to solve large scale scheduling problem. Furthermore, we need to devise a specialized Branch&Bound algorithm to deal with the case when sequencing decision is involved. However, we hope that larger instances of the sequencing problem can be solved if we move to a commercial platform to deal with the 0-1 problem. We leave this for future research.

4. TOPIC 3: PATTERN RECOGNITION AND BIASED PERCEPTION OF RANDOMNESS

4.1 Abstract

Loss of control generates the drive for human beings to seek patterns in their daily lives. In this paper, we investigate how our perceptions of randomness are shaped by this innate desire to find patterns out of chaos. I use both theoretical and empirical methods to show that lottery players can be influenced to believe erroneously that the winning probabilities of past winning numbers are higher in the current draw, even though the events are independent (i.e. Hot-Hand Fallacy prevails). This result is surprising as works by Clotfelter and Cook (1993) and Terrell (1994) have documented the presence of Gambler's fallacy in the US lottery market instead - the amount of money bet on a particular number in a pick-3 or pick-4 game falls sharply after the number is drawn. I use two sets of lottery game data in Asia to show that Hot-Hand Fallacy can prevail in a pick-3 or pick-4 game, and conclude that the design of lottery games (e.g. prize structures) can influence the perception of randomness, and hence the two fallacies may dominate under different gaming environments. These results have important implications as it indicates that

people's perception of randomness, and thus behavior, can be manipulated through appropriate design in the lottery game. Our results also provide an explanation to a question raised in the "lucky store" effect paper by Guryan and Kearney (2008) - why lottery players believe that lightning will strike twice in the case of lottery vendors, but not in the case of numbers.

4.2 Introduction

Ever since human ancestors looked at the starry sky and wondered how different combinations of stars foretold the future, pattern seeking has been entrenched in our daily lives. Our brains are well wired to constantly interpret meanings out of (random) events happening in our lives and we search for patterns to reassure us that life is "under control". Both anecdotal evidence and research findings show that when facing complex/uncertain situations, our sense of control is threatened and we are more likely to impose relationships among unrelated events and perceive (illusory) patterns. Pattern seeking undoubtedly has its survival value in the evolutional history of human beings (cf. Whiteson and Galinsky (2008), Proulx and Heine (2009), and Beitman (2010) etc.) However, pattern seeking inevitably affects human perceptions of randomness since it imposes connections between otherwise possibly uncorrelated events. Important decisions under uncertainty heavily rely on predictions of (random) future events. Therefore, careful investigation should be made to understand how the perception of randomness is shaped by this deeply imbedded human nature.

People are known to rely on heuristics or simple mental models, rather than theoretical models, to interpret random events (cf. Kahneman and Tversky (1971), Hastie et al (2009) etc.). Among many others, Gambler's Fallacy and Hot-Hand Fallacy are the most common biases in the perception of randomness. Gambler's Fallacy is an erroneous belief in the negative correlation of independent outcomes generated by a random process; while Hot-Hand Fallacy refers to the belief in positive correlations in the inde-

pendent outcomes of a random process. Although the two fallacies seem to contradict each other, current literature posits that the two fallacies are related. They may arise from the same mechanism: belief in "the law of small numbers" or "local representativeness" where small samples are used to represent the characteristics of the total population. (cf. Rabin (2002), Rabin and Vayanos (2010) etc.)

The earlier works focus on understanding the phenomenon that the Gambler's and Hot-Hand Fallacy co-exist in the same individual. They identify the importance of the length of streaks (sequence of repeated events) in influencing the perception of randomness, i.e., Gambler's Fallacy prevails in short streaks, but as the streaks lengthen, beliefs in Hot-Hand dominate (cf. Asparouhva et al. (2009), Rabin and Vayanos (2010), Jorgensen et al (2011)). On the other hand, observations that Gambler's Fallacy is more prevalent in lottery games, and Hot-hand Fallacy more prevalent in games involving skills, are often attributed to the "intentions" of the subjects involved. In the recent work of Caruso et al. (2010), through a set of lab experiments, they predict continuation of a streak when subjects involved are considered to be intentional. Hence, when the streak is caused by some kind of mechanical device (as in lottery games), people believe the streak will end, i.e., Gambler's Fallacy dominates.

Recent filed observation, however, suggests that Hot-Hand Fallacy can prevail in a lottery game under appropriate conditions. In a pick-4 numbers games, where there are multiple winning numbers in each draw, I find that punters bet more on previous winning numbers instead of avoiding them. Figure 4.1 shows the 25th percentile, the median, and the 75th percentile

of betting proportions of 2300 winning numbers on the day they are drawn and the following 56 draws. Punters participating this game seem to believe that the winning probabilities of past winning numbers are higher in the current draw, even though the events are independent (i.e. Hot-Hand Fallacy prevails). This result is surprising, as works by Clotfelter and Cook (1993) and Terrell (1994) have documented the presence of Gambler's Fallacy in the US lottery market instead - the amount of money bet on a particular number in a pick-3 or pick-4 game falls sharply after the number is drawn.

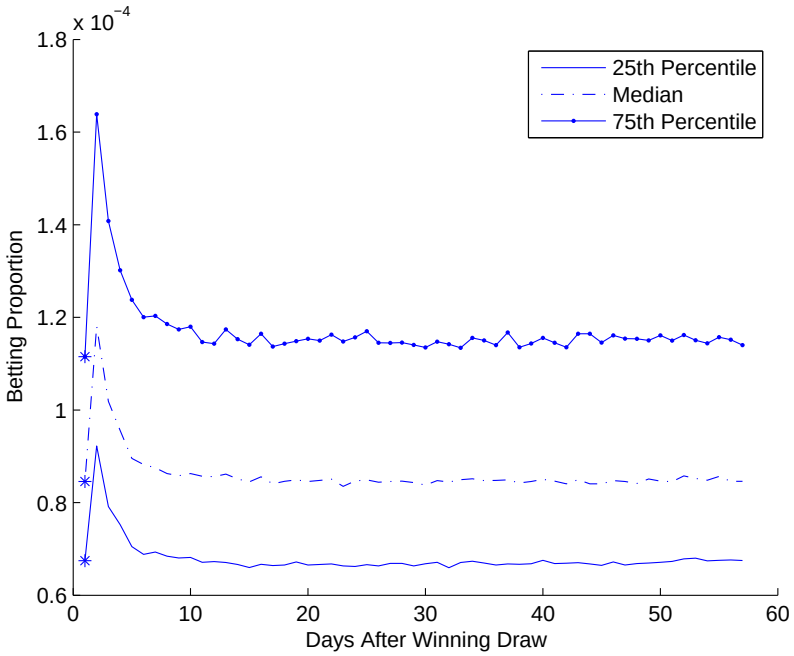

Fig. 4.1: Betting on Previous winning numbers

To reinforce the field observations, I collect data from two different lottery games played in the same region. The two games are of similar designs, under similar operations and played by two populations that are considered

to be similar in both race and culture. The only difference between the two games comes from the size of the prize-winning numbers in each draw. In one game, single winning number is drawn whereas multiple (23) winning numbers are drawn in the other. Through careful statistical analysis, the empirical data validate that Gambler's Fallacy is observed in the lottery game with a single winner; while Hot-Hand Fallacy arises in the one with more complex, multiple outcomes.

Motivated by these observations, I propose in this paper a formal theory to investigate the role of different game designs in "shaping" our perceptions of randomness in the long run, by incorporating our innate desire to seek patterns out of chaos.

I develop a Bayesian updating model to investigate conditions under which pattern seeking leads to Hot-Hand Fallacy. In our setting, I suppose a *theory*, such as a (possibly randomized) computer program, produces a sequence of independent and identical outcomes. The agent has a prior belief over a finite set of possible data-generating theories. The set in consideration may or may not contain the true data-generating theory. The agent observes realizations of random events in each period and relies on limited history to update his/her beliefs.

One question to ask is whether the observed data provides enough information to infer a "relatively small" set of plausible data-generating theories. Unless one imposes some particular restrictions on conceivable theories, the answer to this question is *no*. Recent research in statistics and economic theory shows that there exists a set of theories that can be accepted as the true

theory with arbitrarily high probability (e.g., Fudenberg and Levine (1999), Lehrer (2001), Olszewski and Sandroni (2008), Sandroni et al. (2003), Sandroni (2003), Shmaya (2008)). This gives credence to the fact that people might hold different beliefs to explain a sequence of random outcomes, which may possibly conflict with each other.

I show that as long as the agent considers the true theory as a possible explanation, he/she will eventually make correct inferences as evidence accumulates. What of more interests is whether there exist consistent predictions about the agent's belief if his/her initial belief is biased, namely, the true theory is not considered plausible[1]. In the model, I assume that there are two distinct sets of theories. One set of theories predict higher chance for the pattern outcomes (identified from past data) to occur; while the other predict less. I show that the former theory prevails when the environment is complex - when the ratio of pattern outcomes is relatively high compared to the total number of possible outcomes (as in the multiple-number lottery game).

Our results have important implications in problem gambling, risk management, and lab experiments where random outcomes are involved. They indicate that the perception of randomness can be manipulated, and hence behavior can be nudged with the appropriate design. For instance, several countries have attempted to tamper with commuter's behavior by offering incentive schemes for commuters to earn credit for each journey taken (triple credit for off-peak journeys) to earn a chance of cash prizes in weekly lot-

[1] Some punters in the lottery games subscribe to the conspiracy theory and believe that winning numbers drawn are tweaked to favor the operators.

teries. The success of these schemes hinges on the insights of behavioral economics that the average person is risk seeking when the stake is small. So a 1 in 1000 chance to win $100 is more attractive than a cash award of $0.10. Our results say that we can do better, through designing the lottery games to induce Hot-Hand Fallacy in the commuters, and offering them the chance to choose their own "lucky" numbers to bet on. The belief (distortion) that the winning probabilities of certain numbers are higher than theoretical average, and the ability for the players to bet on these numbers, is enough to make what is actually a 1 in 1000 chance of winning appears to be a much safer bet, and thus more attractive to the players.

4.3 Literature Review

4.3.1 Uncertainty and pattern seeking

It has long been observed that in situations lacking control, people may tend to believe that some mysterious, unseen mechanisms are secretly at work. Numerous anecdotal evidence and research findings circumstantially support the assumption that more complex and uncertain situations lead to stronger pattern seeking.

Fishermen who fish in the deep sea and whose lives are usually threatened by the unpredictable weather and water conditions have much more complicated superstitious beliefs than those who fish in the shallow water (Malinowski [1948]). In March 2010, when the red shirt protests propagated and economy receded, people in Thailand were reported to seek help from

fortune tellers more frequently, probably to feel assured about the future.

Facing complex situation, it is a human instinct to look for patterns, as illustrated by Proulx and Heine (2009). They made a group of subjects read an absurd story by Kafka. To make it more complicated to comprehend, they added to the story more inconsistent and bizarre illustrations, which was meant to throw the subject in the world of complexity and chaos. While another set of subjects read a more consistent version of the story. In the end, when students played the game to find patterns in strings of letters, students who read the absurdist story were reported to find much more patterns in strings that those who read the consistent version. Those students tried to combat the chaos presented by the absurd story, to find the patterns in it, and to make sense of it. This primed their brains to look harder for patterns elsewhere, such as in strings of letters.

Whiteson and Galinsky (2008) study the relationship between lack of control and illusory pattern seeking, rituals, and superstitions. They create situations under which a group of subjects experience lack of control. They find that compared to the rest, those subjects see more false patterns in all type of data, like imagining trends in stock markets, seeing faces in static and develop superstitions etc.

In this paper, I make a natural assumption that people identify more patterns from more complex historical results.

4.3.2 Gambler's Fallacy and Hot-Hand Fallacy

Both Gambler's Fallacy and Hot-Hand Fallacy have long been observed in the field. Among many others, Clotfelter and Cook (1993) observe that lottery players in 3D game in United States are subject to Gambler's Fallacy. Figure 4.2 shows the percentiles of betting ratios (betting volume index on particular day over average index) on different days after the numbers are selected as winners. Once a 3D number is drawn as the winner, the subsequent betting volumes drop immediately and then gradually pick up. The immediate drop after the winning number is drawn provides a strong evidence of Gambler's Fallacy in the game. Meanwhile, Camerer (1989) show that betting markets for basketball games exhibit a small Hot-Hand bias. Guryan and Kearney (2008) examine the sales of lottery outlets selling winning jackpot tickets and show that those stores experience significant increase in game-specific ticket sales, exhibiting the "lucky store" effect. Jorgensen et al (2011) examine a set of panel data on lotto games and found that people usually avoid numbers that have recently been drawn (exhibiting Gambler's Fallacy"); while they tend to bet more on those winning numbers in streaks (that have been drawn several time in a row), suggesting the existence of Hot-Hand Fallacy.

Gambler's Fallacy and Hot-Hand Fallacy may co-exist in the same agent and appear within the same setting. One line of research posits that they both arise from representativeness bias or belief in "the law of small numbers". Tversky and Kahneman (1971) coine this term and examine its connection with Gambler's Fallacy. Rabin (2002) and Rabin and Vayanos (2010) build theoretical models to model "the law of small numbers" which directly

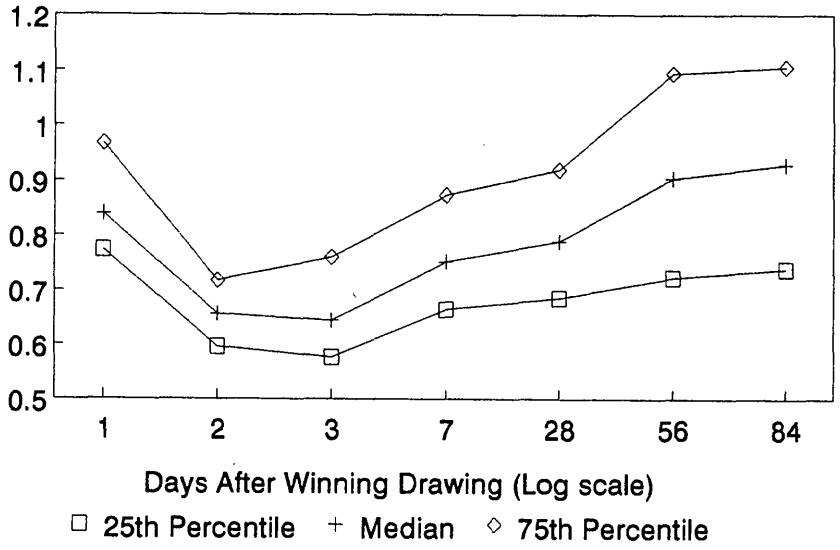

Fig. 4.2: Betting Ratios on Previous Winning Numbers in 3-Digit Numbers Game (Clotfelter and Cook (1993)

leads to Gambler's Fallacy. These works identify the importance of the length of streaks (sequence of repeated events) in influencing the perception of randomness, i.e., Gambler's Fallacy prevails in short streaks, but as the streaks lengthen, beliefs in Hot-Hand dominate. Asparouhva et al. (2009) run a set of lab experiments on binary choice game and the results support the work of Rabin (2002). In recent work of Kendall (2010), the author uses a discount factor to refine the model presented in Rabin (2002) and propose a simple but effective model to explain the two fallacies.

Different from this line of research, our paper examine the role of game designs in "shaping" the perception of randomness in a long run. I show that due to pattern seeking, either Gambler's Fallacy or Hot-Hand Fallacy can arise under appropriate conditions. I can actually manipulate the behaviors through careful system design. Besides, our paper generalizes the cur-

rent conclusions on the streaks of binary outcomes. In our paper, "streaks" becomes one particular pattern agents might recognize from previous outcomes. When the streak short (i.e., the "streak" pattern has not been recognized yet), people tend to believe in Gambler's Fallacy; while as the streak lengthens and become an obvious pattern, Hot-Hand fallacy occurs. In the single-winner lottery games, however, streaks are extremely unlikely, thus Gambler's Fallacy is observed.

In spite of the excellent work on the "the law of small numbers" to examine the two fallacies, another stream of research provides different explanations. They argue that the nature of a random event generator, i.e., whether it is an inanimate device or a human being, determines the occurrence of Gambler's Fallacy or Hot-Hand Fallacy. (c.f. Ayton and Fisher (2004), Burns and Corpus (2004)). If the random process is believed to be generated by a machine, people will expect negative recency; while a positive recency is expected if human beings generate the sequences. These results are reinforced by field observations made in casinos and lottery store(cf. Sundali and Croson (2006) and Guryan and Kearney (2008)). In the recent work of Caruso et al. (2010), through a set of lab experiments, they predict continuation of a streak when subjects involved are considered to be intentional. This work shares similar flavor to our work in the way that an intentional mind shows some mechanism is at work and patterns related to the intension can be perceived.

According to their theory, Gambler's Fallacy would occur in lottery games as the outcomes are generally believed to be generated without human involvements. It, unfortunately, contradicts the field observation I made. Our

model explore this issue from the innate human nature of pattern seeking in complex settings and establish conditions under which Gambler's Fallacy and Hot-Hand Fallacy occur.

4.4 Field Evidence

In this section, I first exhibit the empirical evidence to show that different fallacy and decision biases may prevail under different gaming environments, even if the games may appear to be similar in design. I collect first-hand data from two fixed-odds lottery games to demonstrate this phenomenon. The first is a 3-Digit (3D) lottery game played in China, the scheme of which resembles "pick 3" lottery game reported in Clotfelter and Cook (1993). It draws a single prize-winning number each day, 7 times per week. The other game, a 4-Digit (4D) lottery game played in South East Asia, draws 23 winning numbers instead, with 3 draws organized per week. The two number games are of similar design, and played by two populations that are similar in both race and culture. They differ mainly in the number of winning numbers drawn. This provides a perfect test bed for the theory and predictions that I will develop in section 4.5.

4.4.1 Gambler's Fallacy in 3D numbers game

In the 3D numbers game, a single winning number is randomly drawn from 000 to 999 with equal probability at 8:30pm each day. Lottery tickets on the current draw are sold until 8:00pm. Every wager costs 2 local currency. The payout structure and winning odds are shown in Table 4.1.

Tab. 4.1: Prize Structure and Winning Odds in 3D Numbers Game

Bet Type	Match to Win	Payout per Wager	Odd
Straight	Match the exact order	1000	1 in 1000
Box 3	Match any order (2 identical digits)	320	1 in 333
Box 6	Match any order (3 unique digits)	160	1 in 167

A direct test of Gambler's Fallacy requires a full data set consisting of the sales of each number in each draw. These data, however, are rarely made available to the public. I collect regularly released data on the official web site of the game operator[2]. The lottery company posts regularly on its web site (i) the winning number drawn, (ii) number of winning wagers (of three play types), and (iii) total sales in each draw. I collect the online data of 2272 draws over a 76-month period, from May 8, 2005 to September 15, 2011. From the data set, I construct a sub-sample of 95 draws that contains winning numbers repeating within 7 weeks. In analyzing the data, I adopt a similar methodology used in Terrell (1994).

With the information of winning wagers and payout structure, I can easily calculate the payout of each winning numbers. To adjust for the day of the week effect, I define the payout rate R as the ratio of the payout to total sales volume in a draw. R indicates the popularity of a winning number. I want to find out the impact on sales in subsequent draws after a 3D number has been drawn as a winning number. Since I only have information on winning numbers, I check the payout rates of those winning numbers repeating within 7 weeks.

Table 4.2 shows the payout statistics (mean, standard deviation and median) of winning numbers repeating within certain period. The overall

[2] See *http://www.zhcw.com/3d/kaijiangshuju/index.shtml?type=0.*

Tab. 4.2: Payout Statistic of Repeating Winning Numbers

	Number	Mean	STD	Median
winning numbers repeating within 1 week	12	0.436	0.52	0.259
winning numbers repeating between 1 and 2 weeks	12	0.358	0.15	0.346
winning numbers repeating between 2 and 4 weeks	25	0.384	0.22	0.303
winning numbers repeating between 4 and 7 weeks	46	0.475	0.23	0.418
Winner not repeating within 7 weeks	2177	0.503	0.31	0.430
Total	2272	0.500	0.31	0.427

mean payout rate is 0.5 and the median is 0.427. Table 4.2 indicates that both the mean and the median of repeating winning numbers within 7 weeks are lower than those statistics of all winning numbers. The median payout of winning numbers that repeat within one week drops to around 40% and then gradually picks up[3]. Two sample t-test shows that the mean of those numbers repeating within 1 to 5 weeks is significantly smaller than 0.5.

Let $LogR_i$ denote the logarithmic value of the payout rate of winning number i. To analyze the effects of the winning number on sales, I build a linear regression model using $LogR_i$ as the dependent variable. Let D_i denote the *inverse* of the number of draws it takes for a winning number i to show up again as a winning number. A positive coefficient of D indicates that a winning draw promotes the popularity of a winning number (Hot-Hand Fallacy) while a negative one undermines the popularity of a number (Gambler's Fallacy). Besides, I define the other independent variable $LogLastR_i$. It is the logarithmic value of the payout rate when the number i last won.

[3] Note that such a tendency does not stand out in the mean statistic since the mean of numbers repeating within one week is really high due to an outlier. 3D number 149 was drawn as the winner on Feb 13, 2008. One week later, when the same number 149 came out again as the winner, the payout ratio shoot 200.25%, the highest level recorded in history. This noise explains the high mean and standard deviations among winner repeating within one week

Tab. 4.3: Results of Linear Regression on 3D Numbers Game

		3D
Number of Observations		95
Constant	α	-0.296^{***}
Delay	β	$\mathbf{-.762}^{***}$
Previous Payout	γ	$.228^{**}$
Adjusted R^2		0.156

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

A positive coefficient of $LogLastR_i$ shows that the popularity of a number persists. I specify the following estimation for a repeating winner i:

$$LogR_i = \alpha + \beta D_i + \gamma LogLastR_i + \epsilon_i. \tag{4.1}$$

Among the parameters, α is a constant, β captures the effect of a winning draw, γ indicates the popularity of a number. Consistent with Table 4.2, I only select winning numbers that repeat within 7 weeks. The regression results are shown in Table 4.3. The coefficient β is estimated to be $-.762$ and is significantly negative at the 1% level. It indicates that the expected payout value of a number drops by $17.3\%(= 10^{-.762})$ one draw after it wins the prize, keeping other variables constant. In all, if the payout rate of a number is 0.5, the regression model predicts a mean payout of the same number is 0.07 one draw after it wins, 0.362 ten draws after it wins and 0.395 twenty draws after it wins. Both the statistical data and regression results suggest that Gambler's Fallacy exists among the players in the 3D lottery game.

Tab. 4.4: Prize Table and Winning Odds in 4D Numbers Game

Prize Category	Payout per Wager	Odds
1st Prize	2000	1 in 10000
2nd Prize	1000	1 in 10000
3rd Prize	500	1 in 10000
10 Starter	250	1 in 1000
10 Consolation	60	1 in 1000

4.4.2 Hot-Hand Fallacy in 4D numbers game

In the 4D lottery game, players bet on numbers selected from 0000 to 9999. Sales for each draw start a week before and close at 6pm on the draw day. Minimum cost of a bet is 1 local currency. 23 4D numbers are drawn as winning numbers 3 times a week. The 23 winning numbers are generated with replacement by rolling 4 boxes that contain 10 balls in each [4]. There are 5 prize categories. See in Table 4.4 the payout to each prize category and winning odds.

I obtained data from a game operator in the South East Asia region, with information on 156 draws that covers a one-year period. The dataset consists of 23 winning numbers and sales volumes of each 4D number in each draw. I want to investigate how a winning draw affects the sales of the winning numbers in subsequent draws. To account for daily effects, I use betting proportion as an indicator and it is defined as the ratio of the sales volume of a 4D number to the total sales volume. I examine all winning numbers from draw 1 to draw 100 and calculate their betting proportion respectively on the day they are drawn, and subsequent 56 draws. About 40% increase in sales immediately follows a winning draw, suggesting Hot-Hand Fallacy

[4] This means that with slight chance there might be repetitions among the 23 winning numbers.

Tab. 4.5: Results of Linear Regression on 4D Numbers Game

Number of Observations		157
Constant	α	$-.452^{***}$
Delay	β	$\mathbf{0.114}^{***}$
Previous Payout	γ	0.886^{***}
Adjusted R^2		0.738

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

prevail in the aggregate level.

Next, I run a regression model with the data in 4D numbers game. The model is similar to the one defined in equation (4.1). Due to the presence of different prize categories, I use the *betting proportion* instead of payout ratio to indicate the popularity of a number [5]. Again, let $LogR_i$ denote the log value of betting proportion and $LogLastR_i$ the log value of betting proportion when winner i last won. Similar to the analysis in the 3D game, I use data that repeats within 7 weeks (21 draws). The regression results are shown in Table 4.5 - contrary to the 3D game, the parameter in our regression model is significantly positive, which shows that winning in the current draw has a positive impact on subsequent sales. These compelling evidence suggests that players in 4D numbers game are subject to Hot-Hand Fallacy.

I use the 4D sales data to validate further the basic assumption of this paper: that the players recognize basic patterns from the winning numbers appearing in the past draws. I say that a 4D number i is a "near miss" in the current draw, if it is "similar" to a winning number, but not identical. I specify two patterns of "near miss": Permutation, and Digit Replacement. If

[5] Note that in the 3-D numbers game, the payout ratio is simply the betting proportion times the prize.

a player bets on a number that gets all the 4 digits right but in a wrong order, a Permutation near miss occurs. On the other hand, if a number matches 3 digits of a winning number in the right position correctly, the bet is considered a Digit-Replacement near miss. Both field and lab research evidence indicate that near miss can encourage more participation in lottery games (cf. Reid (1986), Kassinove and Schare (2001)). These results are supported by Clark et al. (2009) in their recent research in neuroscience. They also report that game operators manipulate the frequency of near misses to generate more sales. The near miss effect is a prime suspect to a fundamental psychological factor leading to problem gambling.

In the following, I test whether there exists "near miss" effect in the 4D sales data. I draw insights using data from the 8th draw to the 146th draw. I calculate the betting proportions of the near-miss numbers on the day the corresponding winning numbers are drawn, 7 draws before these numbers are drawn, and 1st, 2nd, 3rd, 7th and 14th draw after these number are drawn. Table 4.6 lists the median value of the betting proportions of the first-prize winning numbers, 23 prize winning numbers, and near miss numbers in specific draws. As illustrated in Table 4.6, both patterns of near miss gain popularity after it hits a draw. This effect is more evident for first prize winning numbers, with surge of relative sales to about 102%. Hence the hot hand fallacy on winning numbers spill over even to numbers that are a "near miss", although the effect on sales is less significant.

Tab. 4.6: Lagged Average Betting Proportion of All Winning Numbers

Lag	23 winning numbers	1st Prize Winner	Permutation	Digit Replacement
-7	0.0838	0.0842	0.0991	0.099
0	0.0846	0.0848	0.0992	0.0992
1	0.1185	0.1712	0.1175	0.1068
2	0.1025	0.1434	0.1116	0.1049
3	0.0959	0.1272	0.1093	0.1041
7	0.0884	0.1135	0.1052	0.1021
14	0.0868	0.1075	0.1032	0.101
20	0.0852	0.096	0.1021	0.1004

4.5 The Model

In this section, I create a simple model in which a sequence of random outcomes are generated and build a Bayesian updating model to investigate under what conditions different game designs might lead to different biases in the perception of randomness.

In each period t, suppose that a set of random outcomes W_t is produced according to a stochastic data-generating process ρ, under which the outcomes are independently and identically distributed (i.i.d.). Denote Ω as the set that consists of all possible outcomes. W_t is a subset of Ω. For example, Ω contains all the possible values of returns of a mutual fund and W_t is the realized return in period t. Alternatively, in a 4D lottery game, Ω consists of all integers ranging from 0 to 9999 and W_t is a set that contains 23 integers from 0 to 9999 - the 23 winning numbers drawn in period t.

Both current literature and anecdotal evidence demonstrate that agents exhibit different kind of beliefs even when facing purely random outcomes. To account for these differences, I assume that an agent possesses beliefs taking form of a finite collection of *theories* $\mathcal{T}$. Each theory offers an explanation

of the outcomes observed. Theories that attempt to explain the data can be interpreted as probability distributions (Olszewski and Sandroni (2008)).

Denote by ΔW the set of probability distributions over W and $\mathcal{H}_t = \{W_1, W_2, \cdots, W_t\}$ the set of possible histories. A *theory* is formally defined as a function $\tau : \cup_{t\geq 0}\mathcal{H}_t \mapsto \Delta W$. τ defines a probability distribution of possible outcomes in next period given the set of possible history $\mathcal{H}_t$. I assume that τ satisfies the following conditions:

1. τ is a stationary process,

2. τ is a Markov processes of order r,

3. τ gives positive probability to all finite-length histories of outcomes.

Condition 1 and 2 imply that theories are time-invariant and the agent has such a limited memory that he only recalls the outcomes in the previous r periods. Note that the third condition involves no loss of generality. Any theory that assigns zero probability to certain possible outcomes would be eventually discarded with probability one after sufficient many periods.

The agent has a prior belief q_0 over theories: $q_0 \in \Delta\mathcal{T}$. $q_0(\tau)$ gives a *priori* likelihood that τ explains the output sequence. For the ease of exposition, I assume that the agent starts to update his/her belief after observing the first r outcomes. For a given output sequence, let q_t be the agent's posterior belief after observing the first t $(t \geq r)$ outcomes. $q_t(\tau) = Pr(\tau \mid W_{t-r+1}, \ldots, W_t)$. I denote by $\tau(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)$ the likelihood of getting an outcome W_{t+1} given the previous sets of outcomes $W_{t-r+1}, \ldots, W_t$ under theory τ:

$$\tau(W_{t+1} \mid W_{t-r+1}, \ldots, W_t) = Pr(W_{t+1} \mid W_{t-r+1}, \ldots, W_t; \tau).$$

Assume the agent is perfectly Bayesian. Then the posteriors are computed according to Bayes' rule:

$$q_{t+1}(\tau) = q_t(\tau) \cdot \frac{\tau(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\sum_{\hat{\tau} \in \mathcal{T}} q_t(\hat{\tau}) \cdot \hat{\tau}(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)} \ . \tag{4.2}$$

$q_t(\tau)$ can be interpreted as the weight assigned to theory τ after observing the first t outcomes.

One question to ask is to which extent different theories explain a sequence of outcomes. By condition (3), I know that no theory refutes a finite sequence of outcomes. Two probability distributions P and Q are said to be *orthogonal* whenever there exists an event $\mathcal{E}$ such that $\mathcal{E}$ occurs almost surely under P and almost never under Q. The following lemma demonstrates that two distinct (orthogonal) theories satisfying conditions (1)–(3) provide explanations for disjoint collections of output sequences.

Lemma 2. Any two distinct processes τ_1 and τ_2 that satisfy conditions (1)–(3) are such that, for all periods t and finite histories $W_1, \ldots, W_t$, the conditional distributions $\tau_1(\cdot \mid W_{t-r+1}, \ldots, W_t)$ and $\tau_2(\cdot \mid W_{t-r+1}, \ldots, W_t)$ are orthogonal.

Proof.

See Appendix II. ■

Exploiting the result in Lemma 1, I show in the following proposition that the outcomes will eventually be interpreted by the truth data-generating process ρ , as long as the agent considers ρ as a possible theory, i.e., $\rho \in \mathcal{T}$.

Proposition 8. For every sequence of outcomes $W_1, W_2, \ldots$ generated by ρ, as long as $\rho \in \mathcal{T}$, the posterior weight that the agent assigns to the true data-generating process converges towards one, i.e., $\lim_{t\to+\infty} q_t(\rho) = 1$.

Proof. See Appendix II. ■

Proposition 1 states that when the agent is perfectly Bayesian and the true theory is considered possible, he/she will eventually learn the truth for almost all sequences. I stress that this fact is true regardless of the prior belief of the agent, as long as the agent assigns a positive weight, even light, to the true data generating process. Concretely, the result means that, even if an agent starts with a strong bias towards alternative theories, the bias eventually self-corrects through Bayesian revision as observations accumulate.

What will happen if the agent's prior belief does not assign any weight to the truth, i.e., $\rho \notin \mathcal{T}$? Will the agent's belief converge, and if so, which biased belief will it converge to? In what follows, I incorporate "pattern seeking" in our model and explore the conditions under which Gambler's Fallacy or Hot-Hand Fallacy prevails.

Human subjects tend to seek certain types of patterns among elements in and between the (possibly random) outcomes. These patterns capture the idea that one element is intuitively related to another. In basketball game, agents readily believe that a series of shots taken by the same player are related, even though complex analysis show they are statistically independent of each other. To capture this notion of relationship, I use a *relationship func-*

tion π to model such patterns. In particular, $\pi \equiv \pi(W_{t-r+1}, \ldots, W_t)$ records a set of possible outcomes that manifest the patterns an agent might recognize from the previous outcomes $W_{t-r+1}, \ldots, W_t$. These outcomes will be accorded special attention and treated differently from the outcomes outside of the set.

Let Π be the set of all relationship functions assumed plausible. I call $\Pi(W_{t-r+1}, \ldots, W_t)$ the "pattern set". It should be noted that the patterns recognized from a set of outcomes are generally subjective, and I do not attempt to explicitly describe the elements in Π. Intuitively, the larger the set of outcomes, the more likely a recognized pattern is reinforced by new "evidence". Due to limited attention of the agents, Π is kept reasonably small. As I show below, it is precisely the quantity of available patterns–or to be exact, their probability mass–that eventually determines whether an agent becomes subject to the Hot-Hand or Gambler's Fallacy in the long run.

Define an indicator

$$\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1}) \equiv \chi\{W_{t+1} \bigcap \Pi(W_{t-r+1}, \ldots, W_t) \neq \emptyset\}.$$

$\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1}) = 1$ if the outcome in period $t+1$ and the previous r outcomes are related. Since the outcomes W_i's in each period are independently and identically generated by ρ, the indicator $\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1})$ is independent of t.

To illustrate how the size of the outcomes affects the pattern set and $E(\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1}))$, suppose I randomly generate n 3-digit integers ranging from 0 and 999 in each period. In each round, the agent observes

n outcomes and forms a pattern set. For sake of simplicity, assume the agent has a memory of only one period, i.e., $r = 1$. Suppose the relationship function contains "identity", "permutation", and "one-digit replacement" functions. The "identity" function assigns the same n outcomes to the pattern set Π. The "one-digit replacement" function keeps two digits of each of the 3-Digit outcomes and replaces the last with a different digit chosen from 0 to 9. For example, this function assign an outcome 123 to $\{023, 223, 323, 423, 523, 623, 723, 823, 923, 103, 113, 133, 143, 153, 163, 173, 183, 193, 120, 121, 122, 124, 125, 126, 127, 128, 129\}$. The "permutation" function assigns the outcome 123 to $\{132, 213, 231, 312, 321\}$. As n increase, the size of the pattern set constructed using the above relationship functions will also increase, and hence the probability of observing an outcome intersecting the pattern set in the next period will increase. Figure 4.3 demonstrates the numerical results on the size of the pattern sets and the value $E(\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1}))$ under different n. Obviously, as the size of the outcomes n increase, the size of the pattern set increase, so does the probability that two sets are related.

In the following, I assume that the agents do not assign any weight to the true theory ρ. Without loss of generality, I divide the theories into two classes. The first class of theories $\mathcal{T}_1$ "promote" related outcomes. A theory τ_1 in the first class predicts a relatively high probability for outcomes in the next period that is deemed to be related to those in the previous periods. Theory τ_1 is specified by a parameter α_1, $\alpha_1 > 1$. The value α_1 indicates the probability ratio of the related outcome to the unrelated one. With a slight abuse of notation, I define $R(t+1, r) \equiv \{W_{t+1} : W_{t+1} \bigcap \Pi(W_{t-r+1}, \ldots, W_t) \neq \emptyset \mid W_{t-r+1}, \ldots, W_t\}$. $R(t+1, r)$ happens with $\rho-$probability $x(t+1, r)$. For

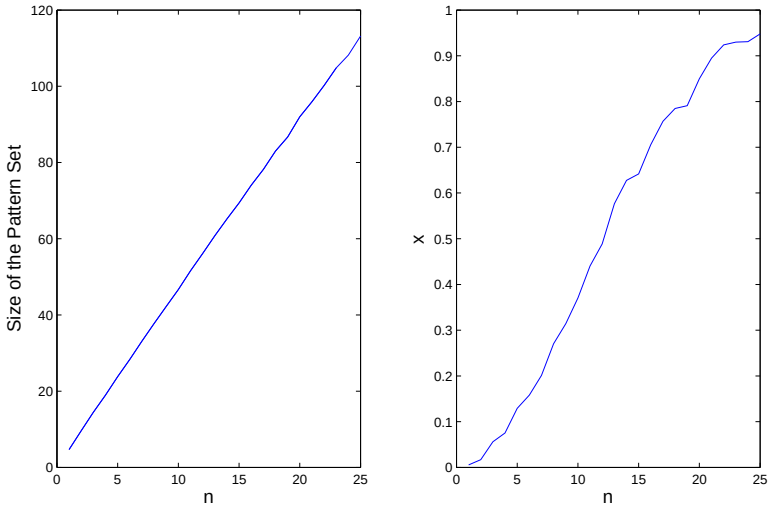

Fig. 4.3: Expected size of the pattern sets and estimated $E(\mathcal{R}(W_{t-r+1}, \ldots, W_{t+1}))$ (denoted by x) under different n

$t = r + 1, r + 2, \ldots,$

$$\tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t) = \begin{cases} \frac{\alpha_1 \rho(W_{t+1})}{1+(\alpha_1-1)x(t+1,r)} & \text{if } R(t+1,r)\text{occurs,} \\ \frac{\rho(W_{t+1})}{1+(\alpha_1-1)x(t+1,r)} & \text{otherwise.} \end{cases} \tag{4.3}$$

In contrast, the second class of theories $\mathcal{T}_2$ "undermine" related outcomes. A theory τ_2 in this class predicts that patterns are less likely than under the true data generating process. This class contains theories that lead to Gambler's Fallacy. Similarly, τ_2 is characterized by a parameter α_2, $0 < \alpha_2 < 1$. For $t = r, r + 1, \ldots,$

$$\tau_2(W_{t+1} \mid W_{t-r+1}, \ldots, W_t) = \begin{cases} \frac{\alpha_2 \rho(W_{t+1})}{1+(\alpha_2-1)x(t+1,r)} & \text{if } R(t+1,r)\text{occurs,} \\ \frac{\rho(W_{t+1})}{1+(\alpha_2-1)x(t+1,r)} & \text{otherwise.} \end{cases} \tag{4.4}$$

The set of plausible theories is then defined as $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$.

Further, two regularity assumptions on parameters $\alpha_1,\ \alpha_2$ are made.

Assumption 1. $0 < \alpha_1 + \alpha_2 - 2 < 2(\alpha_1 - 1)(1 - \alpha_2)$.

Assumption 2. $\log \frac{\alpha_1}{\alpha_2} < \alpha_1 - \alpha_2$ and $\log \frac{\alpha_1}{\alpha_2} < \frac{\alpha_1 - \alpha_2}{\alpha_1 \alpha_2}$.

The implications of Assumption 1 are two folds. The left-hand sided inequality guarantees that the sum of α_1 and α_2 is not too small. A smaller α_1 indicates that theory τ_1 is closer to the true theory; while a smaller α_2 implies that theory τ_2 is more distinct from the true theory. If this constraint is violated, theory τ_1 dominates in the long run. On the other hand, the second constraint regulate the two parameters to ensure τ_2 does not dominate. Assumption 2 requires that α_1 is big enough while α_2 is small enough. This assumption makes sure that both theories are sufficiently different from the true theory to make the results more meaningful.

Define a function $f(x) = x \log \frac{\alpha_1}{\alpha_2} + \log \frac{1+(\alpha_2-1)x}{1+(\alpha_1-1)x}$. Assumption 1 and 2 lead to the following lemma.

Lemma 3. When Assumption 1 and 2 hold, a unique solution $x^*(0 < x^* < 1)$ exists that solves $f(x^*) = 0$. Besides, $f(x) < 0$ when $x \in (0, x^*)$ and $f(x) > 0$ when $x \in (x^*, 1)$.

Proof. See Appendix II. ■

Figure 4.4 depicts function $f(x)$ given that $\alpha_1 = 5$, $\alpha_2 = 0.2$. The cut-off probability is around 0.5. As we can see, $f(x) > 0$ when $x > 0.5$ while $f(x)$ is below zero when $x < 0.5$.

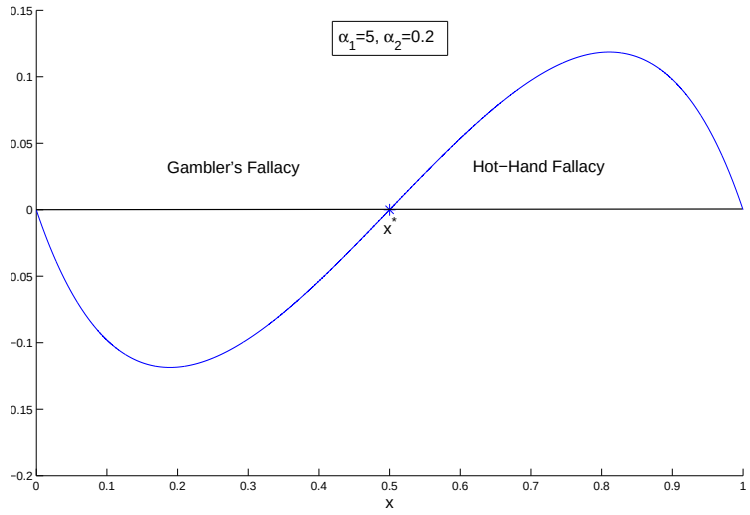

Fig. 4.4: One illustration of Proposition 2

For simplicity, I assume that the agent considers only τ_1 and τ_2 plausible. His prior belief assigns positive weights to both theories, i.e., $q_0(\tau_1) > 0,\ q_0(\tau_2) > 0,\ q_0(\tau_1) + q_0(\tau_2) = 1$. The agent starts updating his/her beliefs after observing r outcomes according to the Bayes' Rule and $q_r(\tau_1) = q_0(\tau_1),\ q_r(\tau_2) = q_0(\tau_2)$. The posterior belief period $t+1$ is:

$$q_{t+1}(\tau_1) = \frac{q_t(\tau_1) \cdot \tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\sum_{i=1,2} q_t(\tau_i) \cdot \tau_i(W_{t+1} \mid W_{t-r+1}, \ldots . W_t)} \tag{4.5}$$

Which biased belief would the Bayesian updating process converge to? The following proposition states sufficient conditions under which the Bayesian updating converges to the theory corresponding to Hot-Hand Fallacy or Gam-

bler's Fallacy. Recall that $x(t+1,r)$ is the probability that the outcomes in period $t+1$ and the previous outcomes are related. Define a critical value

$$\beta = \frac{-E\left[f(x(t+1,r)) \mid x(t+1,r) < x^*\right]}{E\left[f(x(t+1,r)) \mid x(t+1,r) \geq x^*\right] - E\left[f(x(t+1,r)) \mid x(t+1,r) < x^*\right]}$$

Proposition 9. Suppose that the theory set $\mathcal{T}$ contain τ_1 and τ_2. For every sequence of $\rho-$almost outcomes $W_1, W_2, \ldots$, the following convergence results hold:

1) The agent's belief converges to τ_1, i.e., $\lim_{t\to+\infty} q_t(\tau_1) = 1$ a.e. as long as $\rho(x(t+1,r) \geq x^*) > \beta,\ \forall\ t \geq r$;

2) The agent's belief converges to τ_2 when $\rho(x(t+1,r) \geq x^*) < \beta,\ \forall\, t \geq r$.

Proof. See Appendix II. ∎

Which belief will the updating process converge to if α_1 and α_2 change? Figure 4.5 plots the relationships between the cut-off probability x^* and $\alpha_1 \times \alpha_2$. I can see that as $\alpha_1 \times \alpha_2$ increases, the cut-off probability x^* becomes larger. Larger x^* suggests that the conditions in result 2) of Proposition 2 is easier to achieve, i.e., agent' belief will more likely to converge to Gambler's Fallacy. This is because a bigger $\alpha_1 \times \alpha_2$ indicates that τ_1 promotes related outcomes with a higher ratio than τ_2 promotes unrelated ones, which means τ_2 is closer to the true theory. Therefore, the agent's belief is more likely to converge to τ_2, which corresponds Gambler's Fallacy.

In conclusion, bigger $\alpha_1 \times \alpha_2$ results in larger x^*, and Gambler's Fallacy is more likely to occur in the long run. On the other hand, Hot-Hand Fallacy dominates when $\alpha_1 \times \alpha_2$ is small.

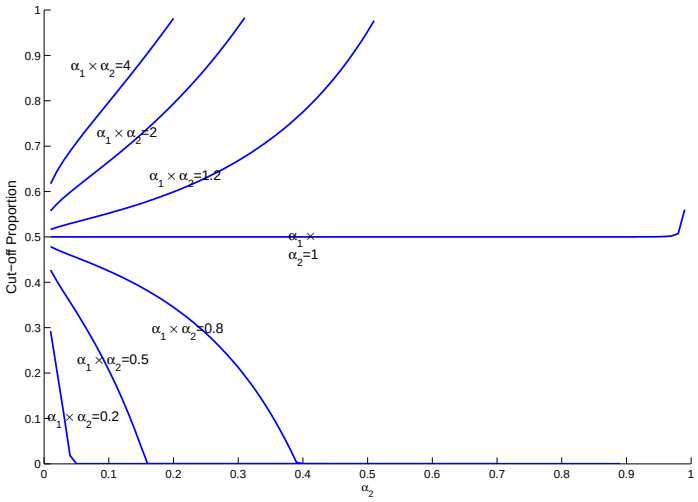

Fig. 4.5: The Cut-off Proportions as a function of $\alpha_1 \times \alpha_2$

I simulate the following scenarios and test the implications of the theory developed. In the simulation,

1. N balls are in a urn and are numbered from 0 to $N-1$;
2. In each period, n balls are drawn independently from the urn. The drawing process is with replacement, i.e., after a ball is drawn and its number taken down, it is put back to the urn and then another ball is drawn;
3. $n \ll N$;
4. Agents have a limited memory of only one period;

5. "Equality", "Permutation" and "One-digit Replacement" are the possible relationships an agent identifies among the numbers in two sets of outcomes.

Note that I assume that the winning numbers in the same period are drawn *with* replacement, which means the same number may be drawn as a winning number more than once. This case rarely happens though, under the assumption that $n << N$.

I define the set of outcomes in period t as $W_t = \{\omega_{1t}, \omega_{2t}, \ldots, \omega_{nt}\}$, where $\omega_{jt},\ j = 1, \ldots, n$ denotes a winning number drawn in period t. Since each number is randomly and independently drawn with replacement, $Pr(\omega_{j,t}) = 1/N$ for $j = 1, 2, \ldots, n$ and $P(W_t) = (\frac{1}{N})^n$.

Denote $\Pi(W_t)$ as the set of numbers that are related to W_t and $m_t = \mid \Pi(W_t) \mid$ the size of the set. An agent believes in two possible theories τ_1 and τ_2. In particular, τ_1 and τ_2 are defined by:

$$\tau_1(\omega_{j,t+1} \mid W_t) = \begin{cases} \frac{\alpha_1}{N+(\alpha_1-1)m_t} & \text{if } \omega_{j,t+1} \in \Pi(W_t) \\ \frac{1}{N+(\alpha_1-1)m_t} & \text{if } \omega_{j,t+1} \notin \Pi(W_t) \end{cases}. \tag{4.6}$$

$$\tau_2(\omega_{j,t+1} \mid W_t) = \begin{cases} \frac{\alpha_2}{N+(\alpha_2-1)m_t} & \text{if } \omega_{j,t+1} \in \Pi(W_t) \\ \frac{1}{N+(\alpha_2-1)m_t} & \text{if } \omega_{j,t+1} \notin \Pi(W_t) \end{cases}, \tag{4.7}$$

Assume that in the first period when no history is observed, $\tau_1(\omega_j, 1) = \tau_2(\omega_j, 1) = 1/N$.

In the following, I run a set of matlab experiments under a set of parameters $(N, n, \alpha_1, \alpha_2)$ to gain a glimpse of the relationships between the direction

of convergence and the value of the parameters. I first fix $N = 1000, \alpha_1 = 5$ and $\alpha_2 = 0.2$, and change n from 1 to 100. Consistent with the prediction of the model, the experimental results show that as n increases, the process is more likely to converge to the theory that corresponds to Hot-Hand Fallacy. Figure 4.6 demonstrates two typical updating processes for $q_t(\tau_1)$ under different size of winning numbers. In this case, note that $x^* = \beta = 0.5$. When $n = 1$, the number of outcomes in the pattern set, at time t, is at most 33[6]. Hence $x(t+1, r) < 33/1000$ irrespective of the outcome drawn in period t. Thus $\rho(x(t+1, r) > 0.5) = 0$ and hence our result indicates that Gambler's Fallacy will dominate. On the other hand, when n increases to 23, the number of outcomes in the pattern set increases drastically, and could be as high as $33 \times 23 = 759$. It is easy to check that $\rho(x(t+1, r) > 0.5) > 0.5$ in this case and hence Hot-Hand Fallacy dominates.

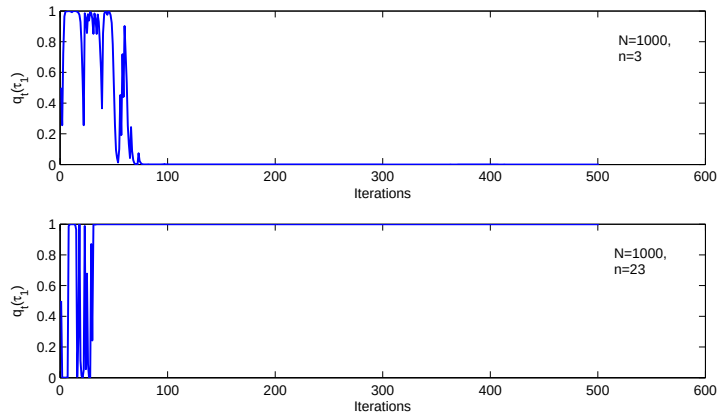

Fig. 4.6: The updating process of $q_1(\tau_1)$ Given $N = 1000$ and $n = 3, 23$.

[6] 1 from the identity relation, at most 5 more from the permutation relation, and at most 27 more from the single digit replacement function.

4.6 Conclusion and Discussions

This paper studies how the innate human nature to seek pattern out of chaos leads to biased perceptions of randomness. I develop a Bayesian updating model to examine how beliefs evolve through time. Our model predicts that when the set of outcomes are relatively large for agents to infer sufficient patterns, the Hot-Hand beliefs may dominate. The theoretical results are reinforced by two datasets collected from two different lottery games played in the region. One lottery game with single winning number per draw exhibits Gambler's Fallacy among players, while the other with multiple winning numbers shows Hot-Hand Fallacy. Field data show that once people recognize enough patterns among the outcomes, they tend to believe that the same outcome or related outcomes will repeat in the future. This confirms the prediction from the Bayesian updating model.

The notion that human can be manipulated to believe in the hot hand fallacy, through appropriate game design, has important ramifications in various fields. First, it provides a behavioral explanation to the "medium prizes puzzle" (cf. Haruvy et al (2001)) - why did lottery game operators typically offer prize distribution of a few large prizes and a large number of medium one? This is puzzling especially if we assume gamblers are typically risk seeking. There are two common explanations for this puzzling observation. One is through the lens of prospect theory - a large number of medium prizes reduces the probability of losing from near certainty to some smaller probability. Another explanation follows the line of adaptive learning - that human behavior is best captured by simple adaptive learning models, and actions

that did better in the past will tend to be adopted more frequently compared to actions that did worse. Thus the presence of medium prizes slows down the punter's inclination to gamble less. Both explanations, however, failed to account for the decision biases in the Gambler's and Hot-Hand fallacies. It could not explain, for instance, why players in these games normally prefer to choose their own numbers to bet on. Our theory of pattern seeking provides another explanation - that a large number of medium prizes can induce more players to believe in Hot-Hand Fallacy (that a selected set of numbers have a larger than the true probability of winning the lottery). This reduces the inclination of the players to quit the game and also to bet on those numbers they believe to have a larger probability of winning.

Second, it provides guidelines in designing lottery games to induce desirable behavior. There is a recent trend for governments to encourage good civic behavior through the use of lottery games. Richard Thaler expounds recently on the merits of this approach [7]. New Taipei City in Taiwan recently initiats a lottery as an inducement for dog owners (and other citizens) to clean up after their pets, in order to win gold ingots worth as much as $2000. The Singapore government is also experimenting with a new incentive scheme for commuters to earn credit for each journey taken (triple credit for off-peak journeys) for a chance to win cash prizes in weekly lotteries. Our study highlights several features for these lottery games to be more effective in influencing behavior - that there should be a sufficiently large number of medium prizes (to induce more hot hand believers), and also to find a mech-

[7] See the article "Making Good Citizenship Fun" on New York Times, February 13th, 2012

anism to allow the players to bet on numbers they believe to have a higher probability of winning (thus increasing their incentive to participate in the lottery games through good civic behavior). One way to do this is to use personalized numbers (instead of random numbers) that the players can easily relate to draw for the winners. The Dutch government uses this principle very effectively - one of its state lotteries is based on postal codes. The idea is to play on people's feelings of regret, and also to exploit the "lucky store" effect that has been shown to exist in various lottery games - your chances of winning will be higher if your postal code has been drawn before (or near miss) in the previous draws!

Our results also shed light on a question raised in Guryan and Kearney (2008): why lottery players believe that lightning will strike twice in the case of lottery vendors, but not in the case of numbers? Guryan and Kearney (2008) have documented a sharp increase in sales for stores that have sold a winning lottery ticket in the Lotto Texas game, whereas Clotfelter and Cook (1993) have documented that sales in a pick-3 game falls sharply after the number is drawn. One speculative explanation offered by Guryan and Kearney (2008) is along the line of animate/inanimate versus intended/non-intended distinction - the winning number is drawn from a mechanical device, but the location of the winning outlet is chosen deliberately by the person buying the winning ticket. Another possible explanation can be offered along the line of pattern set introduced in this paper - there were 669 retail outlets in Texas, and 68 winning jackpots over 2.5 year period of the study. Furthermore, the data indicates that lucky store effect can linger for as long as 40 over weeks after a store sold the winning ticket. This means

that the impression of a store selling a winning ticket has a long memory, and thus a richer pattern set. This is enough to induce some players in the game to believe in the hot-hand phenomenon, leading to an increase in sales for the lucky store. This theory also partially explains why the lucky store effect is less pronounced in other lottery games played in Texas, such as the Cash 5 and Texas Two-Step games. The jackpot sizes of these games are much smaller, and winning the jackpot is less sensational. The memory of the winners is shorter, and thus the pattern set is more sparse.

5. CONCLUSION AND DISCUSSIONS

The objective of this dissertation has been to investigate three phases in a typical fact-based decision making process. The three phases include data collecting, pattern seeking, and performance improvement. The dissertation is motivated by several industrial data sets. It explores (behavioral) patterns in the pertinent system, then incorporates these patterns into service system design. As a close loop, this dissertation have gone one step further to investigate how system design affects behavioral patterns, in particular, customers' biased perceptions of randomness. This dissertation focuses on two areas: health-care management and behavioral economics.

The dissertation consists of three topics, which examine the following three research questions: 1) Are there simple and universal patterns when people make choices? 2) How to utilize limited (insufficient) data to design a robust service system to ensure good performance under all possible situations? 3) How system design affects behavioral patterns?

The first topic titled "Benford's Law and Number Selection in Fixed-Odds Numbers Game" investigates a universal pattern revealed when people make random choices: the small number phenomenon. There are ample empirical evidence suggesting that players do not choose all numbers with equal probability, but have a tendency to bet on (small) numbers that are

closely related to events around them (e.g., birth dates, addresses, etc.). To the best of my knowledge, this topic is the first to quantify this phenomenon and examine its relation to the classical Benford's law. I use this connection to develop a choice model that incorporates this universal phenomenon. In fixed-odds numbers games, the prizes and the odds of winning are known at the time of placement of the wager. Both players and operators are subject to the vagaries of luck in such games. Most game operators limit their liability exposure by imposing a sales limit on the bets received for each bet type, at the risk of losing the rejected bets to the underground operators. I argue that the choice of the sales limit is intimately related to the ways players select numbers to bet on in the games. I exploit the choice model we built to optimally decide appropriate sales limit to control the risk for the game operator.

The second topic entitled "Appointment System Design using Copositive Cones" concerns the design of appointment systems to regulate the usage of precious resources in a service system. In particular, I investigate a stochastic appointment scheduling and sequencing problem in an outpatient clinic with a single doctor. The number of patients is fixed, and the problem is to determine the arrival time and order for each customer. I have collected data on the service durations of 1021 patients in an local eye clinic. In the model, I assume that the service durations of the patients are stochastic, and only the mean and covariance estimates are known. I do not assume any exact distributional form of the service durations, and solve for distributionally robust schedules that minimize the expectation of the weighted sum of patients' waiting time and doctor's overtime. The scheduling prob-

lem is formulated as a convex conic optimization problem with a tractable semi-definite (SDP) relaxation. Using the primal-dual optimality conditions, I prove several interesting structural properties of optimal schedules. Despite the required relaxation in computation, I can still obtain near optimal solutions compared to the existing literature. I apply this method in a realistic setting at an eye clinic and suggest new schedules that can improve the efficiency of the clinic by around 35%. Further, this approach can be extended to solve the appointment and sequencing problem can be simultaneously, which can be approximated by a 0-1 SDP problem.

In the third topic I focus on a close-loop of the fact-based decision making process and investigate how system design in turn affects people's behaviorial patterns. Facing a sequence of random outcomes, people may erroneously impose positive or negative correlations on independent outcomes. The Hot-Hand Fallacy (belief in positive recency) and the Gambler's Fallacy (belief in negative recency) are two most common biases in the perceptions of randomness. I shows that game designs might shape biased perception of randomness due to one embedded human nature, pattern seeking. I develop an economic setting in which a sequence of random outcomes are generated and build a Bayesian Updating model to explore conditions under which the Hot-hand Fallacy or Gambler's Fallacy appears. Our model implies that the more complex the information set (historical outcomes), where pattern-seeking has its stronger appeal, the more likely that an agent converges to beliefs giving rise to the Hot-Hand Fallacy. I collect two sets of field data from gaming industry to provide a solid foundation and verification of the insights gained from the model. These results have important implications in problem gambling, risk

management, and lab experiments where random outcomes are involved.

In summary, the three topics, viewing the issues in fact-based decision making through multiple methodologies, have contributed to the discipline of Business Analytics. In addition to the contributions that have already been highlighted in each topic, this dissertation brings out important perspectives for future BA research. First, the dissertation exhibits the effectiveness of combining multiple research methodologies in investigating an integrated fact-based decision making process. These methodologies include empirical study, experimental research, optimization, and simulation. A planned future research approached in different ways will certainly bring out fresh insights in both theory and practice. Second, this dissertation highlights an imperative role of incorporating human behaviors into system design. Continuous improvement in BA requires investigating the interface between behavioral economics and service science. It is crucial to examines human behavioral patterns before trying to improve the service system. Besides, while human behaviors may affect the system performance, the system may cultivate certain behavioral patterns in the system. Only taking this recursive effect into consideration could continuous improvement be made possible.

Future research can continue exploiting the framework built in this dissertation. For example, with individual-level data on price-plan choice and consumption, I can study patterns in customer choice when a bundle of products are offered using empirical and experimental approach. Next, performance improvement can be reached through exploring the corresponding pricing strategy of a company utilizing the optimization and simulation tools. This dissertation, although a preliminary exploration in BA research, is a sig-

nificant step in laying the foundations of the framework and providing a road map for future research in this area.

BIBLIOGRAPHY

Asparouhva, E., Hertzel, M., and Lemmon, M. 2009. Inference from streaks in random outcomes: experimental evidence on beliefs in regime shifting and the law of small numbers, *Management Science*, **55**, 1766-1782.

Ayton, P., and Fisher, I. 2004. The Gamblers Fallacy and the Hot-Handed Fallacy: Two faces of subjective randomness, *Memory & Cognition*, **32**, 1369-1378.

Bailey, N. 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.

Baron, G., and Leider, S. 2010. The role of experience in Gambler's Fallacy, *Journal of Behavioral Decision Making,* ,**23(1)**, 117-129.

Begen, M. A., Levi, R., Queyranne, M. (2010) *A Sampling-Based Approach to AppointmentScheduling*, Working Paper.

Begen, M. A., Queyranne, M. 2009 Appointment scheduling with discrete random durations, Proceedings of Symposium on Discrete Algorithms (SODA) 2009, 845-854.

Beitman, B.D. 2010. Brains seek patterns in coincidences/, *Psychiatric Annals*,

39(5), 255-264.

Benford, F. 1938. The law of anomalous numbers, *Proceedings of the American Philosophical Society*, **78**, 551-572.

Berman, A., Shaked-Monderer, N. 2003. Completely Positive Matrices, *World Scientific*.

Bertsimas, D., Doan, X. V., Natarajan, K., Teo, C. P. 2008. Models for minimax stochastic linear optimization problems with risk aversion, To appear in *Mathematics of Operations Research*.

Bertsimas, D., Natarajan, K., Teo, C. P. 2004. Probabilistic combinatorial optimization: moments, semidefinite programming and asymptotic bounds, *SIAM Journal of Optimization* , **15**, 185-209.

Bertsimas, D., Natarajan, K., Teo, C. P. 2006. Persistence in discrete optimization under data uncertainty, *Mathematical Programming*, **108**, 251-274.

Buhler, Joe. P., Elwyn Berlekamp. 2006. Puzzles Column. *Emissary, a newsletter published by the MSRI group in Berkeley* ,**Spring/Fall**, 13–13.

Burns, B.D., and Corpus, B. 2004 Randomness and inductions from steaks: "Gambler's fallacy" versus "hot hand", *Psychonomic Bulletin and Review* , **11**, 179-184.

Bomze, I. M., Dür, M., Klerk, E. D., Roos, C., Quist, A. J., Terlaky, T. (2000) On copositive programming and standard quadratic optimization problems, *Journal of Global Optimization* , **18**, 301-320.

Boyd, S., Vandenberghe, L. (2004) Convex Optimizatioin, *Cambridge University Press* .

Burer, S. 2009. On the copositive representation of binary and continuous nonconvex quadratic programs, *Mathematical Programming* , **120**, 479-495.

Camerer, C.F. 1989 Does the basketball market believe in the hot hand?/, *American Economic Review* , **79**, 1257C1261.

Caruso, E.M., Waytz, M., and Epley, N. 2010 The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue/, *Cognition* , **116**, 149C153.

Cayirli, T., Veral, E. (2003) *Outpatient-scheduling in health care: A review of literature*, Production and Operations Management, **12**, 519-549.

Cayirli, T., Veral, E., Rosen, H. (2008) Assessment of patient classification in appointment system design, *Production and Operations Management* , **17**, 47-58.

Chernoff, Herman. 1999. How to beat the Massachusetts number game: An application of some basic ideas in probability and statistics, *The Mathematical Intelligencer*, **3**, 166-175.

Clark, L., Lawrence, A.J., Astley-Jones F., and Gray, N. 2009 Gambling near-misses enhance motivation to gamble and recruit win-related brain circuitry/, *Neuron*, **61**, 481-490.

Clotfelter, Charles, Philip Cook. 1989. *Selling Hopes: State Lotteries in America*,

Cambridge, MA: Harvard University Press.

Clotfelter, C., and Cook, D. 1993. The "Gambler's Fallacy' in lottery play, *Management Science*, **39**, 1521-1525.

Clotfelter, Charles, Philip Cook. 1999. Lotteries in the real world, *Journal of Risk and Uncertainty*, **4(3)** 227–232.

Collins, J. 2001 Good to Great: Why some companies make the leap... and others don't. HaperCollines, New York.

Davenport, T.H., and Harris, J.G. 2006 Competing on analytics: The new science of winning. Harvard Business School Press, Boston, Massachusetts.

Davenport, T.H., Harris, J.G., and Morison, R. 2010. Analytics at work: Smarter decisions, better results. Harvard Business School Press, Boston, Massachusetts.

Dawid, A.P. 1982. The well-calibrated Bayesian, *Journal of the American Statistical Association*, **77**, 605-610.

Denton, B., Gupta, D. 2003. A sequential approach for appointment appointment scheduling, *IIE Transactions*, **35**, 1003-1016.

Denton, B.T., Miller, A., Balasubramanian, H., Huschka, T. (2010) *Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty*, *Operations Research*, **58(4)**,802-816.

Dickinson, P. J. C., Gijben, L. (2011) *On the Computational Complexity of Membership Problems for the Completely Positive Cone and its Dual*, Working

Paper.

Dür, M. (2009) Copositive programming: a survey, Available online at (http://www.optimization-online.org/DB_HTML/2009/11/2464.html).

Erdogan, S., Denton, B. (2010) *Surgery Planning and Scheduling: A Liaterature Review*, Wiley Encyclopedia of Operations Research and Management Science.

Fudenberg, D., and Levine, D.K. 1999. The well-calibrated Bayesian, *Journal of the American Statistical Association*, **29(1-2)**, 131-137.

Gary, M. R., Johnson, D.S. (1979) Computers and Intractability: A guide to the theory of NP-completeness , W. H. Freeman and Company, New York .

Giles, D.E. 2006. Benford's law and naturally occurring prices in certain ebay auctions, *Applied Economics Letters*, **14**, 157-161.

Grendar, M., Judge,G., and Schechter, L. 2007. An empirical non-parametric likelihood family of data-based benford-like distributions, *Physica A*, **380**, 429-438.

Gupta, D. 2007. Surgical suites' operations research, *Production and Operations Management* , **16**, 689-700.

Gupta, D., Denton, B. 2008. Appointment scheduling in health care: Challenges and opportunities, *IIE Transactions* , **40**, 800-819.

Guryan, J., and Kearney, M.S. 2008. Gambling at Lucky Stores: Empirical Evidence from State Lottery Sales, *American Economic Review*, **98(1)**, 458-473.

Haigh, J. 1997. The statistics of National Lottery *Journal of the Royal Statistical Society. Series A* , **160 (2)**, 187-206.

Halpern, A.R., Devereaux, S.D. 1989. Lucky numbers: Choice strategies in the Pennsylvania daily number game, *Bulletin of the Psychonomic Society*, **27(2)**, 167-170.

Hardoon, K., Baboushkin, H., Gupta, R., Powell, G. J., and Derevensky, J. L. 1997. Underlying cognitions in the selection of lottery tickets, Paper presented at the Annual Meeting of the Canadian Psychological Association, Toronto, Ontario, June.

Haruvy E., I. Erev, and D. Sonsino (2001). The Medium Prizes Paradox: Evidence from a Simulated Casino, *Journal of Risk and Uncertainty* **22**, 251-261.

Hastie, R., McClelland, G., Oskarsson, A., and Van Boven, L. 2009. What's next? Judging sequences of binary events, *Psychological Bulletin*, **135**, 262-285.

Henze, N. 1997. A statistical and probabilistic analysis of popular lottery tickets, *Statistica Neerlandica*, **51(2)**, 155-163.

Herman, J., Gupta, R., and Derevensky J. L. 1998. Children's cognitive perceptions of 6/ 49 lottery tickets, *Journal of Gambling Studies* ,**14**, 227-244.

Hill, T. 1995a. Base-invariance implies benford's law, *Proceedings of the American Mathematical Society*, **123**, 887-895.

Hill, T. P. 1995b. A statistical derivation of the significant-digit law, *Statistical Science*, **10**, 354-363.

Hill, T. P. 1998. The first digit phenomenon, *The American Scientist*, **10(4)**, 354-363.

Jorgensen, C.B., Suetens, C., and Tyran, J.R. 2011. Predicting Lotto Numbers, *Working Paper*.

Kaandorp, G. C., Koole, G. 2007 *Optimal outpatient appointment scheduling*, Health Care Management Science, **10**, 217-29.

Kahneman, D., and Tversky, A. 1971. Belief in the law of small numbers, *Psychological Bulletin*, **15**, 105-110.

Kassinove, J.I., and Schare, M.L. 2001. Effects of the near miss and the big win on persistence at slot machine gambling. *Psychology of Addictive Behaviors.* **15**, 155-158.

Kassinove, J.I., and Schare, M.L. 2010. Effects of the near miss and the big win on persistence at slot machine gambling. *Working Paper, University of British Columbia.* 2010.

Klerk, E. de, Pasechnik, D. V. (2002) Approximation of the stability number of a graph via copositive programming, *SIAM Journal on Optimization* , **12**, 875-892.

Kuiper, N. H. 1959. Alternative proof of a theorem of birnhaum and pyke, *Mathematical Statistics /* **30** , 251-252.

Ladouceur, R., Dube, D., Giroux, I., Legendre, N., and Gaudet, C. 1996. Cognitive biases and playing behavior on American roulette and the 6/49 lottery,

Unpublished manuscript, Universite Laval.

Ladouceur, R., and Walker, M. 1996. A cognitive perspective on gambling, In P.M. Salkovskis (Ed.), *Trends in cognitive and behavioral therapies*, New York: Wiley, 89-120.

Lafaille, J.M., G. Simonis. 2005. Dissected re-assembled: An analysis of gaming .

Langer, E. J. 1975. The illusion of control, *Journal of Personality and Social Psychology*, **32**, 311-328.

Laursen, G.H.N., and Thorlund, J. 2010. Businesss Analytics for managers:Taking Business Intelligence beyond reporting, John Wiley & Sons, Inc.

Lehrer, E. 2002. Any inspection rule is manipulable, *Econometrica*, **69(5)**, 1333-1347.

Ley, E. 1996. On the peculiar distribution of the u.s. stock indexes' digits, *The American Statistician* , **50**, 311-313.

Liang, J.J. 2006. Intelligent Appointment Scheduling to Reduce Turnaround Time (TAT), *Master Thesis, Singapore-MIT Alliance* .

Liptser, R.S., and Shiryaev, A.N. (2001) Statistics of random processes: General theory, Springer Verlag, 2001 .

Loetscher, T., Brugger, P. 2007. Exploring number space by random digit generation, *Experimental Brain Research* , **180**, 655-665.

Löfberg, J. (2004) *YALMIP : A Toolbox for Modeling and Optimization in*

MATLAB, In Proceedings of the CACSD Conference, Taipei, Taiwan, (http://control.ee.ethz.ch/ joloef/yalmip.php).

Murty, K. G., Kabadi, S. N. (1987) Some NP-complete problems in quadratic and nonlinear programming, *Mathematical Programming* , **39**, 117-129.

Natarajan, K., Teo, C. P., Zheng, Z. (2009) Mixed zero-one linear programs under objective uncertainty: a completely positive representation, Forthcoming in *Operations Resaerch*.

Newcomb, S. 1881. Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematician*, **4**, 39-40.

Nigrini, M.J. 1996. A taxpayer compliance application of benford's law, *Journal of the American Taxation Association*, **18**, 72-91.

Olszewski, W., and Sandroni, A. 2008. Manipulability of future-independent tests, *Econometrica*, **76(6)**, 1437-1466.

Parrilo, P. A. (2000) *Structured Semidefinite Programs and Semi-algebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Available online at: http://www.cds.caltech.edu/˜pablo/.

Proulx, T., and Heine, S.J. 2009. Connections from Kafka: Exposure to meaning threats improves implicit learning of an artificial grammar, *Psychological Science*, **20(9)**, 1125-1131.

Rabin, M. 2002. Inference by Believers in the Law of Small Numbers, *Quarterly Journal of Economics*, **117**, 775-816.

Rabin, M., and Vayanos, D. 2010. The Gamble's and Hot-Hand Fallacies: Theory and Applications, *Review of Economic Studies*, **77**, 730-778.

Raimi, R. 1976. The first digit problem , *The American Mathematical Monthly*, **83**, 887-895.

Rapoport, A., and Budescu,D. 1992. Generation of Random Binary Series in Strictly Competitive Games, *Journal of Experimental Psychology: General*, **121**, 352-364.

Rapoport, A., and Budescu, D. 1997. Randomization in Individual Choice Behavior, *Psychological Review*, **104**, 603-617.

Reid, R.L. 1986. The psychology of the near miss, *Journal of Gambling Studies*, **2**, 32-39.

Sandroni, A. 2003. The reproducible properties of correct forecasts, *International Journal of Game Theory*, **32(1)**, 151-159.

Sandroni, A., Smorodinsky, R., and Vohra, R.V. 2003. Calibration with many checking rules, *Mathematics of Operations Research*, **28(1)**, 141-153.

Scarf, H. (1958) A min-max solution of an inventory problem, *Studies in The Mathematical Theory of Inventory and Production (Arrow, K., Karlin, S. and Scarf, H., Eds) Stanford University Press, California* , 201-209.

Schatte, P. 1988. On mantissa distributions in computing and benford's law, *Journal of Information Processing and Cybernetics*, **24**, 443-445.

Shiryaev, A. 1995. *Probability (2nd edn)*, Springer: New York, NY.

Shmaya, E. 2008. Many inspections are manipulable, *Theoretical Economics*, **3(3)**, 151-159.

Simon, Jonathan. 1999. An analysis of the distribution of combinations chosen by UK national lottery players, *Journal of Risk and Uncertainty*, **17(3)**, 243-276.

Stephens, M.A. 1970. Use of the Kolmogorov-Smirnov, Cramer-von mises and related statistics without extensive tables, *Journal of the Royal Statistical Society Series B*, **32(1)**, 115-122.

Stern, H. and Cover, T. M. 1989. Maximum entropy and the lottery, *Journal of the American Statistical Association* , **84**, 980-985.

Sundali, J. and Croson, R. 2006. Biases in casino betting: The hot hand and the gambler's fallacy, *Judgement and Decision Making*, **1**, 1-12.

Teo, C.P. and S.M. Leong. 2002. Managing Risk in a Four Digit Number Game, *SIAM Review*, **44(4)**, 601-615.

Thomas, J.K. 1989. Unusual patterns in reported earnings, *The Accounting Review*, **64** 773–787.

Tijms, Henk 2007. Understanding Probability: Chance Rules in Everyday Lives, Cambridge University Press.

Toh, K. C., Todd, M. J., Tutuncu, R. H. (1999) *SDPT3 — a Matlab software package for semidefinite programming*, Optimization Methods and Software, **11**, 545-581.

Tversky, A. and Kahneman, D. 1971. *Belief in the Law of Small Numbers*, *Psy-*

chological Bulletin , **76-2**, 105-110.

Terrell, Dek. 1994. A Test of the Gambler's Fallacy-Evidence from Pari-Mutuel Games, *Journal of Risk and Uncertainty*, **8**, 309-317.

Tversky, A. and Kahneman, D. 1974. *Judgement under uncertainty: heuristics and biases*, *Science* , **185**, 1124C1131.

Tutuncu, R. H., Toh, K. C., Todd, M. J. (2003) Solving semidefinite-quadratic-linear programs using SDPT3, *Mathematical Programming*, **95**, 189-217.

Vanden, B. (1997) *Scheduling and sequencing arrivals to a stochastic service system*, Ph.D. Thesis, Air Force Institute of Technology.

Vanden, B., Dietz, C. D. (2000) *Minimizing expected waiting in a medical appointment system*, IIE Transactions, **32**, pp. 841–848.

Vandenberghe, L., Boyd, S., Comanor, K. (2007) *Generalized Chebyshev bounds via semidefinite programming*, SIAM Review, **49**, pp. 52–64.

Varian, H.R. 1972. Benford's law, *The American Statistician*, **26**, 65-66.

Weiss, N. E. (1990) *Models for determining estimated start times and case orderings in hospital operational rooms*, IIE Transactions, **22**, 143-150.

Welch, J. D., Bailey, N. (1952) *Appointment Systems in Hospital Outpatient Departments*, The Lancet, pp. 1105–1108.

Whiteson, J.A., and Galinsky, A.D. 2008 Lacking control increases illusory pattern perception, *Science*, **322**, 115C117.

Whitney, R.E. 1972. Initial digits for the sequence of primes, *American Mathematical Monthly*, **79**, 150-152.

Ziemba, William T., Shelby L. Brumelle, Antoine Gautier, and Sandra L. Schwartz. 1986. Dr. Z's 6/49 Lotto Guidebook, *Vancouvrr and Los Angeles: Dr. Z. Investments. Inc.*

6. APPENDICES

6.1 Appendix I: Proofs in Topic 2

Appendix A. Proof of Proposition 3

Proof. Recall equation (3), the waiting time of the i^{th} patient in the appointment system is given by

$$w_i = \max\left\{0, \tilde{c}_{i-1}, \tilde{c}_{i-1} + \tilde{c}_{i-2}, \cdots, \sum_{k=1}^{i-1} \tilde{c}_k\right\}.$$

In the optimal network flow solution, the unit supply from node i will find a path to destination s, by maximizing the flow cost among the paths:

$$(i \to s),\ (i \to i-1 \to s),\ \ldots,\ (i \to i-1 \to \ldots 1 \to s).$$

Hence the flow cost attained by the supply from node i is just w_i. ■

Appendix B. Proof of Proposition 4

Proof. The proof consists of two parts. In the first part, I show that problem (C) provides an upper bound to problem (P), i.e. $Z'_P(\boldsymbol{s}) \geq Z_P(\boldsymbol{s})$, $\forall \boldsymbol{s}$. Next, through a constructive approach, I find a sequence of random variables, $\tilde{\boldsymbol{u}}^*_\epsilon$ that satisfies the moment conditions in the limiting sense and $\mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*_\epsilon)]$ converges to $Z'_P(\boldsymbol{s})$ when ϵ converges to zero, i.e., the bound provided by (C) is tight.

Step 1. $Z'_P(\boldsymbol{s}) \leq Z_P(\boldsymbol{s})$, $\forall \boldsymbol{s}$.

For any random variable $\tilde{x}$ in $\mathbb{R}_+$, it is well-known that there exists a sequence of simple random variables that increasingly converges to $\tilde{x}$ in every sample point. Hence, for any valid distribution of the service durations, $\tilde{\boldsymbol{u}}$, I can find a sequence of discrete random vectors that increasingly converge to $\tilde{\boldsymbol{u}}$ in every sample point. Denoted the sequence of discrete random vectors as $\left\{\tilde{\boldsymbol{u}}^k\right\}_{k=1}^{\infty}$. Obviously, $f(\boldsymbol{s}, \tilde{\boldsymbol{u}})$ is continuous and increasing in $\tilde{\boldsymbol{u}}$, so for any schedule $\boldsymbol{s}$,

$$\tilde{\boldsymbol{u}}^k(\omega) \uparrow \tilde{\boldsymbol{u}}(\omega), \forall \omega \in \Omega \implies f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}^k\right)(\omega) \uparrow f(\boldsymbol{s}, \tilde{\boldsymbol{u}})(\omega), \forall \omega \in \Omega,$$

where $\boldsymbol{\Omega}$ denotes the sample space. Then from the Monotone Convergence Theorem,

$$\lim_{k\to\infty} \mathbf{E}\left[f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}^k\right)\right] = \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}})\right].$$

Since the feasible space of problem (C) is closed and $\mathbf{E}\left[f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}^k\right)\right]$ is attainable by (C) for every k, $\mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}})]$ is attainable by (C), too. Hence for any appointment schedule and any service time distribution, $Z'_P(\boldsymbol{s}) \geq$

$\mathbf{E}\left[f\left(\boldsymbol{s},\tilde{\boldsymbol{u}}\right)\right]$. Therefore, $Z'_P\left(\boldsymbol{s}\right) \geq Z_P\left(\boldsymbol{s}\right)$, $\forall \boldsymbol{s}$.

Step 2. $Z'_P\left(\boldsymbol{s}\right) \leq Z_P\left(\boldsymbol{s}\right)$, $\forall \boldsymbol{s}$.

Let Z^* be an optimal solution to problem (C). As shown before, I can decompose Z^* into

$$Z^* = \sum_{k\in K_+} \pi(k)^{*2} \begin{pmatrix} 1 \\ \frac{\boldsymbol{t}(k)^*}{\pi(k)^*} \\ \frac{\boldsymbol{v}(k)^*}{\pi(k)^*} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\boldsymbol{t}(k)^*}{\pi(k)^*} \\ \frac{\boldsymbol{v}(k)^*}{\pi(k)^*} \end{pmatrix}^T + \sum_{k\in K_0} \begin{pmatrix} 0 \\ \boldsymbol{t}(k)^* \\ \mathbf{0}_{2n\times 1} \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{t}(k)^* \\ \mathbf{0}_{2n\times 1} \end{pmatrix}^T,$$

where K_+ and K_0 are finite. For $\epsilon \in (0,1)$, I define a sequence of random vectors $\tilde{\boldsymbol{u}}^*_\epsilon$ as follows:

$$\begin{cases} \mathbf{P}\left(\tilde{\boldsymbol{u}}^*_\epsilon = \frac{\boldsymbol{t}(k)^*}{\pi(k)^*}\right) = \left(1-\epsilon^2\right)\pi(k)^{*2},\ \forall k \in K_+ \\ \mathbf{P}\left(\tilde{\boldsymbol{u}}^*_\epsilon = \frac{\sqrt{|\mathcal{K}_0|}\boldsymbol{t}(k)^*}{\epsilon}\right) = \epsilon^2 \frac{1}{|\mathcal{K}_0|},\ \forall k \in K_0 \end{cases}$$

where $|K_0|$ denotes the cardinality of the set K_0. $\tilde{\boldsymbol{u}}^*_\epsilon$ is a valid probability distribution because

$$\begin{aligned} \sum_{k\in\mathcal{K}_+} \left(1-\epsilon^2\right)\pi(k)^{*2} + \sum_{k\in\mathcal{K}_0} \epsilon^2 \tfrac{1}{|\mathcal{K}_0|} &= \left(1-\epsilon^2\right) \sum_{k\in\mathcal{K}_+} \pi(k)^{*2} + \epsilon^2 \sum_{k\in\mathcal{K}_0} \tfrac{1}{|\mathcal{K}_0|} \\ &= \left(1-\epsilon^2\right) + \epsilon^2 \\ &= 1 \end{aligned}$$

Moreover, $\tilde{\boldsymbol{u}}^*_\epsilon$ is a valid service time distribution in the limiting sense, i.e., the moment conditions for the service time distribution are satisfied by $\tilde{\boldsymbol{u}}^*_\epsilon$

when $\epsilon \downarrow 0$,

$$\begin{aligned}
\mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^*] &= \sum_{k\in\mathcal{K}_+} \frac{\boldsymbol{t}(k)^*}{\pi(k)^*}\left(1-\epsilon^2\right)\pi(k)^{*2} + \sum_{k\in\mathcal{K}_0} \frac{\sqrt{|\mathcal{K}_0|}\boldsymbol{t}(k)^*}{\epsilon}\epsilon^2\frac{1}{|\mathcal{K}_0|} \\
&= (1-\epsilon^2)\sum_{k\in\mathcal{K}_+}\boldsymbol{t}(k)^*\pi(k)^* + \epsilon\sum_{k\in\mathcal{K}_0}\frac{\boldsymbol{t}(k)^*}{\sqrt{|\mathcal{K}_0|}} \\
&\underset{\epsilon\downarrow 0}{\rightarrow} \sum_{k\in\mathcal{K}_+}\boldsymbol{t}(k)^*\pi(k)^* \\
&= \boldsymbol{\mu}
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^*\tilde{\boldsymbol{u}}_\epsilon^{*T}] &= \sum_{k\in\mathcal{K}_+}\left(\frac{\boldsymbol{t}(k)^*}{\pi(k)^*}\right)\left(\frac{\boldsymbol{t}(k)^*}{\pi(k)^*}\right)^T\left(1-\epsilon^2\right)\pi(k)^{*2} + \sum_{k\in\mathcal{K}_0}\left(\frac{\sqrt{|\mathcal{K}_0|}\boldsymbol{t}(k)^*}{\epsilon}\right)\left(\frac{\sqrt{|\mathcal{K}_0|}\boldsymbol{t}(k)^*}{\epsilon}\right)^T\epsilon^2\frac{1}{|\mathcal{K}_0|} \\
&= (1-\epsilon^2)\sum_{k\in\mathcal{K}_+}\boldsymbol{t}(k)^*\boldsymbol{t}(k)^{*T} + \sum_{k\in\mathcal{K}_0}\boldsymbol{t}(k)^*\boldsymbol{t}(k)^{*T} \\
&\underset{\epsilon\downarrow 0}{\rightarrow} \sum_{k\in\mathcal{K}}\boldsymbol{t}(k)^*\boldsymbol{t}(k)^{*T} \\
&= \Sigma
\end{aligned}$$

As $\epsilon \downarrow 0$, the random vectors $\tilde{\boldsymbol{u}}_\epsilon^*$ converge almost surely (a.s.)[1] to $\tilde{\boldsymbol{u}}^*$ defined as

$$\mathbf{P}\left(\tilde{\boldsymbol{u}}^* = \frac{\boldsymbol{t}(k)^*}{\pi(k)^*}\right) = \pi(k)^{*2},\ \forall k \in K_+.$$

From the Continuous Mapping Theorem,

$$\tilde{\boldsymbol{u}}_\epsilon^* \rightarrow \tilde{\boldsymbol{u}}^*\, a.\, s. \implies f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*) \rightarrow f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*)\, a.\, s.$$

Furthermore, since the feasible space for $f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)$ is bounded, i.e., every feasible solution $\boldsymbol{y}(\tilde{\boldsymbol{u}}) \leq K\boldsymbol{e}$, for some $0 < K < \infty$, where $\boldsymbol{e}$ is a vector of ones.

[1] Rigorously speaking, the convergenc of $\tilde{\boldsymbol{u}}_\epsilon^*$ to $\tilde{\boldsymbol{u}}^*$ is a weak convergence, i.e., convergence in distribution. However, since it is up to our construction on $\tilde{\boldsymbol{u}}_\epsilon^*$ and $\tilde{\boldsymbol{u}}^*$, from Skorohod's Theorem, I can construct them in the same probability space with the same probability measure and $\tilde{\boldsymbol{u}}_\epsilon^*$ converges to $\tilde{\boldsymbol{u}}^*$ almost surely.

Hence, the second moment of $f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)$ is bounded for all $\epsilon \in (0,1)$, i.e.,

$$\begin{aligned}
\mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)^2\right] &\leq \sum_{k\in\mathcal{K}_+} K^2 \left[\left(\tfrac{\boldsymbol{t}(k)^*}{\pi(k)^*} - \boldsymbol{s}\right)^T \boldsymbol{e}\right]^2 (1-\epsilon^2)\,\pi(k)^{*2} \\
&\quad + \sum_{k\in\mathcal{K}_0} K^2 \left[\left(\tfrac{\sqrt{|\mathcal{K}_0|}\boldsymbol{t}(k)^*}{\epsilon} - \boldsymbol{s}\right)^T \boldsymbol{e}\right]^2 \epsilon^2 \tfrac{1}{|\mathcal{K}_0|} \\
&\leq \sum_{k\in\mathcal{K}_+} K^2 \left[\boldsymbol{t}(k)^{*T}\boldsymbol{e}\right]^2 + \sum_{k\in\mathcal{K}_0} K^2 \left[\boldsymbol{t}(k)^{*T}\boldsymbol{e}\right]^2 \\
&< \infty
\end{aligned}$$

The finiteness of the second moment implies that the sequence $f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)$ is uniformly integrable. Therefore, I have

$$\lim_{\epsilon\downarrow 0} \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)\right] = \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*)\right].$$

For any schedule $\boldsymbol{s}$, define the space of all feasible first and second moments supported on $\mathbb{R}_+^n$ and the corresponding expected objective value as

$$\mathcal{K}(\boldsymbol{s}) = \left\{\lambda\left(1, \hat{\boldsymbol{\mu}}, \hat{\Sigma}, \hat{f}\right) : \lambda \geq 0,\ \hat{f} = \mathbf{E}\left[f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right)\right],\ \text{for some } \tilde{\boldsymbol{u}} \sim \left(\hat{\boldsymbol{\mu}}, \hat{\Sigma}\right)^+\right\}.$$

It can be easily verified that $\mathcal{K}(\boldsymbol{s})$ is a convex cone. Then the closure of $\mathcal{K}(\boldsymbol{s})$ (denoted as $\overline{\mathcal{K}(\boldsymbol{s})}$) would be a closed convex cone. For every $\epsilon \in (0,1)$, I have

$$\left(1, \mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^*], \mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^* \tilde{\boldsymbol{u}}_\epsilon^{*T}], \mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)]\right) \in \overline{\mathcal{K}(\boldsymbol{s})}.$$

Hence, the limit of this sequence of points also lies in the closure, i.e.,

$$\lim_{\epsilon \downarrow 0} \left(1, \mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^*], \mathbf{E}[\tilde{\boldsymbol{u}}_\epsilon^* \tilde{\boldsymbol{u}}_\epsilon^{*T}], \mathbf{E}[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}_\epsilon^*)]\right) \in \overline{\mathcal{K}(\boldsymbol{s})},$$

or equivalently,

$$(1, \boldsymbol{\mu}, \Sigma, \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*)\right]) \in \overline{\mathcal{K}(\boldsymbol{s})}.$$

Since the point $(1, \boldsymbol{\mu}, \Sigma, Z_P(\boldsymbol{s}))$ lies on the boundary of this closed convex cone, I have $Z_P(\boldsymbol{s}) \geq \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*)\right]$. Thus,

$$\begin{aligned} Z_P(\boldsymbol{s}) &\geq \mathbf{E}\left[f(\boldsymbol{s}, \tilde{\boldsymbol{u}}^*)\right] \\ &\geq \sum_{k \in \mathcal{K}_+} \left[\begin{pmatrix} \frac{\boldsymbol{t}(k)^*}{\pi(k)^*} - \boldsymbol{s} \\ \mathbf{0}_{n \times 1} \end{pmatrix}^T \frac{\boldsymbol{v}(k)^*}{\pi(k)^*} \pi(k)^{*2} \right] \\ &= Y(\boldsymbol{s}) \bullet Z^* \\ &= Z_P'(\boldsymbol{s}) \end{aligned}$$

Therefore, I have completed the proof. ■

Remark 6. It is clear from the above proof that decomposition (3.5) does not give the exact worst case distribution, but merely provides us a way to construct a sequence of distributions that satisfies the moment conditions and approximates the objective value $Z_P(\boldsymbol{s})$ in the limiting sense. In fact, I do not have an explicit characterization of the worst case distribution.

Appendix C. An Example on Strong Conic Duality

Consider a simple two dimensional problem as follows for any $Y \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}$:

$$\begin{aligned} Z_P := \max \quad & Y \bullet Z \\ s.t. \quad & Z_{1,1} = 1 \\ & Z \in \mathcal{D} \end{aligned}$$

where

$$\mathcal{D} := \text{conv}\left\{ \begin{pmatrix} \pi \\ v \end{pmatrix} \begin{pmatrix} \pi \\ v \end{pmatrix}^T : \pi \geq 0, v \geq 0, v = b\pi \right\}.$$

Clearly, $\mathcal{D}$ is not fully dimensional, since $\dim(\mathcal{D}) = 1$. In this case,

$$\{Z : Z_{11} = 1, Z \in \mathcal{D}\} = \left\{ \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix} \right\},$$

so

$$Z_P = Y \bullet \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix}.$$

On the other hand, the dual cone of $\mathcal{D}$ is

$$\mathcal{D}^* = \left\{ W : W \bullet \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix} \geq 0 \right\},$$

and the dual optimum

$$Z_D := \min \left\{ \alpha : W = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix} - Y \in \mathcal{D}^* \right\} = Y \bullet \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix} = Z_P.$$

Appendix D. Proofs of the Propositions in Section 3.6

Before presenting the proofs, I first define the necessary dual variables of (S). Let Z be the dual variables of the copositivity constraint. Note that Z is exactly the conic variable in (C), i.e., $Z \in \mathcal{D}$. Denote

$$\begin{cases} Z_{1,n+1+i} = \hat{y}_i, \ i = 1, 2, \ldots, n \\ Z_{1,2n+1+i} = \hat{z}_{i+1}, \ i = 1, 2, \ldots, n \end{cases}$$

Then from the probabilistic interpretation on the decomposition of Z shown in (3.5), I have

$$\begin{cases} \hat{y}_i = \mathbf{E}\left[y_i \left(\boldsymbol{s}, \tilde{\boldsymbol{u}} \right) \right], \ i = 1, 2, \ldots, n \\ \hat{z}_{i+1} = \mathbf{E}\left[z_{i+1} \left(\boldsymbol{s}, \tilde{\boldsymbol{u}} \right) \right], \ i = 1, 2, \ldots, n \end{cases}$$

where $(\boldsymbol{y}(\boldsymbol{s}, \tilde{\boldsymbol{u}}), \boldsymbol{z}(\boldsymbol{s}, \tilde{\boldsymbol{u}}))$ is the optimal solution to $f\left(\boldsymbol{s}, \tilde{\boldsymbol{u}}\right)$ under the worst case distribution of $\tilde{\boldsymbol{u}}$. Define the dual variables of the constraints in (3.8) by θ and λ_i, where θ corresponds to the total session time limit constraint, whereas λ_i corresponds to the non-negativity constraint for s_i.

The proofs are based on the KKT conditions and the network structure shown in Figure 3.2. I have shown that the Slater's constraints qualifica-

tion is satisfied, so the KKT conditions are both necessary and sufficient in characterizing the optimal solutions.

Proof of Proposition 5

Proof. Assume in the optimal solution, $s_i^* = 0$ and $s_{i+1}^* > 0$ for some $i \in \{1, 2, \ldots, n-1\}$. Then the cost on arc $(i+1, i)$ in the network is $\tilde{c}_i = \tilde{u}_i - s_i^* = \tilde{u}_i \geq 0$. Due to the nature of maximal cost flow problem, any flow entering node $i+1$ will choose arc $(i+1, i)$ instead of arc (i, s) whose cost is zero in any situations. Then I have $z_{i+1}(\tilde{\boldsymbol{u}}) = 0$ for any realization of $\tilde{\boldsymbol{u}}$, and consequently $\mathbf{E}[z_{i+1}] = 0$, i.e., in the optimal solution to problem (S), $\hat{z}_{i+1}^* = 0$.

Recall that ρ_i is the cost of the waiting time of the i^{th} patient in the sequence. From the following KKT conditions:

$$\begin{cases} \lambda_i^* s_i^* = 0 \\ \lambda_{i+1}^* s_{i+1}^* = 0 \\ \lambda_{i+1}^* \geq 0 \\ \hat{y}_i^* = \theta^* - \lambda_i^* \\ \hat{y}_{i+1}^* = \theta^* - \lambda_{i+1}^* \\ \rho_{i+1} + \hat{y}_{i+1}^* = \hat{y}_i^* + z_{i+1}^* \end{cases}$$

I get

$$\begin{aligned} & s_{i+1}^* > 0 \\ \Longrightarrow \quad & \lambda_{i+1}^* = 0 \\ \Longrightarrow \quad & \hat{y}_{i+1}^* = \theta^* - \lambda_{i+1}^* = \theta^* \\ \Longrightarrow \quad & \hat{y}_i^* = \rho_{i+1} + \hat{y}_{i+1}^* - \hat{z}_{i+1}^* = \rho_{i+1} + \theta^*. \end{aligned}$$

Since $\hat{y}_i^* = \theta^* - \lambda_i^*$, I have

$$\begin{aligned} & \rho_{i+1} + \theta^* = \theta^* - \lambda_i^* \\ \Longrightarrow \quad & \lambda_i^* = -\rho_{i+1} < 0, \end{aligned}$$

which contradicts $\lambda_i^* \geq 0$. Hence, the result follows. ■

Proof of Proposition 6

Proof. By a similar proof as in Proposition 3, all the negative time slots should be scheduled at the end of the session. Hence, I only need to prove that there is only one such slot, which is s_{n+1}^*.

Assume in an optimal schedule, denoted by $\boldsymbol{s}^{(1)}$, there are at least two nonpositive time slots, i.e., $s_{n-1}^{(1)} < 0$ and $s_n^{(1)} < 0$. Consider a new schedule, $s^{(2)}$ defined as

$$\begin{cases} s_i^{(2)} = s_i^{(1)}, \ \forall i = 1, 2, \ldots, n-2 \\ s_{n-1}^{(2)} = 0 \\ s_n^{(2)} = s_{n-1}^{(1)} + s_n^{(1)} \end{cases}$$

Let $TC^{(1)}(\tilde{\boldsymbol{u}})$ and $TC^{(2)}(\tilde{\boldsymbol{u}})$ be the total waiting time cost for the schedule $\boldsymbol{s}^{(1)}$ and $\boldsymbol{s}^{(2)}$, respectively. Note for any service time distribution, $\tilde{u}_{n-1}^{(k)} - s_{n-1}^{(k)} \geq 0$, and $\tilde{u}_n^{(k)} - s_n^{(k)} \geq 0$, $k = 1, 2$. Then considering the input of the

last two nodes, i.e., ρ_n entering node n and ρ_{n+1} entering node $n+1$, I get

$$\begin{aligned} TC^{(1)}(\tilde{\boldsymbol{u}}) - TC^{(2)}(\tilde{\boldsymbol{u}}) &= \rho_n(\tilde{u}_{n-1} - s_{n-1}^{(1)}) + \rho_{n+1}(\tilde{u}_n - s_n^{(1)} + \tilde{u}_{n-1} - s_{n-1}^{(1)}) \\ &\quad -[\rho_n(\tilde{u}_{n-1} - s_{n-1}^{(2)}) + \rho_{n+1}(\tilde{u}_n - s_n^{(2)} + \tilde{u}_{n-1} - s_{n-1}^{(2)})] \\ &= -\rho_n s_{n-1}^{(1)} \\ &> 0 \end{aligned}$$

with probability 1.

Thus, $\boldsymbol{s}^{(1)}$ should never be optimal, and I reach a contradiction. ■

Proof of Proposition 7

Proof. The proof only makes use of part of the KKT conditions, i.e.,

$$\begin{cases} -\hat{y}_i^* + \theta^* - \lambda_i^* = 0, \forall i = 1, 2, \ldots, n \\ \lambda_i^* s_i^* = 0, \forall i = 1, 2, \ldots, n \\ \theta^* \geq 0 \end{cases} \tag{6.1}$$

When $s_i^* > 0$, $\forall i \in I \subseteq \{1, 2, \ldots, n\}$, from the second set of constraints in equation (6.1), I get

$$\lambda_i^* = 0, \forall i \in I \subseteq \{1, 2, \ldots, n\}.$$

Hence,

$$\hat{y}_i^* = \theta^* \geq 0, \forall i \in I \subseteq \{1, 2, \ldots, n\},$$

i.e.,

$$\mathbf{E}\left[y_i(\boldsymbol{s}^*, \tilde{\boldsymbol{u}})\right] = \theta^* \geq 0,\ \forall i \in I \subseteq \{1, 2, \ldots, n\}.$$

Defining the constant $K := \theta \geq 0$, I get the desired result. ■

Appendix E: Proof to Theorem 3

Assume that there are n patients to be scheduled. For each patient J_j, $j = 1, 2, \cdots, n$, let p_j denote the allocated appointment interval and q_j the corresponding actual consultation time.

Let $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$ and $\mathcal{Q} = \{q_1, q_2, \cdots, q_n\}$. For any given sequence σ and $\mathcal{P}$, let $C_j(\mathcal{P}, \sigma)$ denote the *scheduled* completion time for patient J_j. Therefore, the appointment time given to patient J_j under σ is:

$$a_j(\mathcal{P}, \sigma) = C_j(\mathcal{P}, \sigma) - p_j. \tag{6.2}$$

Let $C_j(\mathcal{P}, \mathcal{Q}, \sigma)$ denote the *actual* completion time for patient J_j. The waiting time for this patient is

$$w_j = C_j(\mathcal{P}, \mathcal{Q}, \sigma) - q_j - a_j(\mathcal{P}, \sigma) = (C_j(\mathcal{P}, \mathcal{Q}, \sigma) - q_j) - (C_j(\mathcal{P}, \sigma) - p_j). \tag{6.3}$$

The objective of the appointment sequencing problem now is to find a sequence σ such that the total patient's waiting time $\sum_{j=1}^{n} W_j$ is minimized. Interestingly, this problem is related to the following well-known NP-complete problem (cf. Gary & Johnson (1979)).

Numerical 3-Dimensional Matching (N3DM)

Instance: Given three disjoint sets W, X, Y, each containing m elements, a size $s(a) \in Z^+$ for each element $a \in W \cup X \cup Y$, and a bound $B \in Z^+$.

Question: Can $W \cup X \cup Y$ be partitioned into m disjoint sets $A_1, A_2, \cdots, A_m$ such that each A_i contains exactly one element from each of W, X and Y and such that, for $i = 1, 2, \cdots, m$, $\sum_{a \in A_i} s(a) = B$?

Proof. I show that given an instance of N3DM, I can construct an instance of the appointment scheduling problem as follows. Suppose there are $3m+1$ jobs such that $p_j = 5M$ for $j = 1, 2, \cdots, 3m+1$; and

$$\begin{cases} q_i = 6M + s(i) & \text{if } \ i \in W \\ q_j = 4M + s(j) & \text{if } \ j \in X \\ q_k = 5M - B + s(k) & \text{if } \ k \in Y \\ q_{3m+1} = (m^2 + 5)M \end{cases},$$

where M is a number much larger than B,e,g., $M > m^2 B$. The question is: can I find a sequence σ to the appointment scheduling problem such that,

$$\sum_{j=1}^{3m+1} w_j \leq Z = mM + 2\sum_{i \in W} s(i) + \sum_{j \in X} s(j)?$$

For notational convenience, I call jobs corresponding to set W, X, Y, W-type, X-type and Y-type respectively. To start with, suppose that the N3DM problem has a feasible solution. I then can obtain a sequence containing the following $m+1$ consecutive blocks, where the jobs in Block i, $i = 1, \cdots, m, m+1$

correspond to $a \in A_i$, and within each A_i jobs are sequenced in W, X, Y order. The $m+1^{\text{th}}$ block contains only J_{3m+1}. By definition, in the above sequence, $C_j(\mathcal{P}, \sigma) = 5jM$ for $j = 1, 2, \cdots, 3m+1$. Also, in the actual processing, there is no idle time for the machine. In each block, the waiting time is zero for the W-type job, $M + s(i)$ for X-type job and $s(i) + s(j)$ for Y-type job, where $s(i)$ corresponds to the W-type job processed in the first position and $s(j)$ corresponds to the X-type job processed in the second position. Also, the waiting time is zero for J_{3m+1}. Hence, the total waiting time equals to

$$mM + 2\sum_{i \in W} s(i) + \sum_{j \in X} s(j) = Z.$$

Then, suppose that there exists a sequence σ to the appointment scheduling problem such that $\sum_{j=1}^{3m+1} w_j \leq Z = mM + 2\sum_{i \in W} s(i) + \sum_{j \in X} s(j)$. I first claim that in such a sequence, J_{3m+1} must be processed in the last position. Otherwise, after J_{3m+1} there are still other patients left and the waiting time is at least $m^2 M$, which is larger than Z. This is a contradiction to our assumption. Note that the waiting time of any patient that is scheduled immediately after a W-type patient is at least $M + s(i)$. Therefore, the total waiting time for those patient is at least $mM + \sum_{i \in X} s(i)$. Now I show that X-type patient must be scheduled immediately after a W-type patient. If not, what follows W-type patient must be W-type or Y-type patient. Firstly, if it is W-type patient who is scheduled immediately after, then the total waiting time should be at least $(m+1)M + \sum_{i \in X} s(i)$, which is bigger than Z. Otherwise, if Y-type patient would be put immediately after W-type patient, then total waiting time for all patients should be at

least $(m+1)M-B$, which is also bigger than Z and a contradiction. In such a case, in each block, the sequence is W-type, X-type, Y-type and the total waiting time of all the patients is $mM + 2\sum_{i\in W} s(i) + \sum_{j\in X} s(j) = Z$. In the actual processing, there will not be any idle time for the machine. These three consecutive jobs W, X, Y form a block with total processing time $15M$, to ensure that a W-type job in the next block can start without waiting. Namely. in each block I have three jobs and $s(i) + s(j) + s(k) = B$. ■

Appendix II: Proofs in Section 4.5

Proof of Lemma 2

Consider any two processes τ_1, τ_2 that satisfy conditions (1)–(3). Take an arbitrary t_0 with $0 \leq t_o \leq t$. For any given $\hat{W}_{t-r+1}, \ldots, \hat{W}_{t_0}$, define

$$\xi_t(W_1, W_2, \ldots) = \begin{cases} 1 & \text{if } W_t = \hat{W}_{t_0}, \ldots, W_{t-t_0+1} = \hat{W}_1 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

If $t_0 = 0$, let $\xi_t = 1$ for all t. The process ξ_t is $\mathcal{F}^t$-measurable. Let $\nu_n = \sum_{t \leq n} \xi_t$. Dawid's calibration theorem states that a coherent Bayesian is a well-calibrated one, i.e., the long-run average proportion that an event happens converges to the estimated probability (Dawid (1982)). According to this theorem, for τ_1- almost every sequence (generated by τ_1 almost surely), if ν_n diverges to infinity, then

$$\lim_{n \to +\infty} \frac{1}{\nu_n} \sum_{t \leq n} \xi_t \left[I\{W_{t+1} = \hat{W}_{t_0+1}\} - \tau_1(W_{t+1} = \hat{W}_{t_0+1} \mid W_{t-r+1}, \ldots, W_t) \right] = 0 \, .$$

Similarly, for τ_2-almost every sequence, if ν_n diverges to infinity, then

$$\lim_{n \to +\infty} \frac{1}{\nu_n} \sum_{t \leq n} \xi_t \left[I\{W_{t+1} = \hat{W}_{t_0+1}\} - \tau_2(W_{t+1} = \hat{W}_{t_0+1} \mid W_{t-r+1}, \ldots, W_t) \right] = 0 \, .$$

The conditions I impose make τ_1 a recurrent process. τ_1 gives positive probability to any finite history, and for τ_1-almost every sequence, ν_n diverges to infinity. The same is true for τ_2. Therefore, if there exists a measurable set

W of output sequences that have positive probability under both probability measures, then there must exist output sequences such that

$$\lim_{n\to+\infty}\frac{1}{\nu_n}\sum_{t\leq n}\xi_t\left[\tau_1(W_{t+1}=\hat{W}_{t_0+1}\mid W_{t-r+1},\ldots,W_t)-\tau_2(W_{t+1}=\hat{W}_{t_0+1}\mid W_{t-r+1},\ldots,W_t)\right]=0\,,$$

or, after simplification according to conditions (1)–(3),

$$\tau_1(\hat{W}_{t_0+1}\mid \hat{W}_{t-r+1},\ldots,\hat{W}_{t_0})=\tau_2(\hat{W}_{t_0+1}\mid \hat{W}_{t-r+1},\ldots,\hat{W}_{t_0})\ .$$

As the equality must be true for arbitrary t_0 and arbitrary sequences $\hat{W}_1,\ldots,\hat{W}_{t_0+1}$, it must be the case that $\tau_1=\tau_2$. The argument carries over directly to the conditional distributions.

Proof of Proposition 8

I want to show that with ρ-probability one on the output sequence, $q_t(\rho)\to 1$. By Bayes rule,

$$q_t(\rho)=q_{t-1}(\rho)\cdot\frac{\rho(W_t\mid W_{t-r+1},\ldots,W_{t-1})}{\sum_{\tau\in\mathcal{T}}q_t(\tau)\cdot\tau(W_t\mid W_{t-r+1},\ldots,W_{t-1})}.$$

Therefore, it suffices to show that, with ρ-probability one on the output sequences $W_1,W_2,\ldots$, for all $\tau\in\mathcal{T}$, $\tau\neq\rho$,

$$\lim_{t\to\infty}\frac{\tau(W_t\mid W_{t-r+1},\ldots,W_{t-1})}{\rho(W_t\mid W_{t-r+1},\ldots,W_{t-1})}=0. \tag{6.4}$$

Let $\rho_t = \rho(W_t \mid W_{t-r+1}, \ldots, W_{t-1})$ and $\tau_t = \tau(W_t \mid W_{t-r+1}, \ldots, W_{t-1})$ be the respective conditonal distributions of ρ and τ for the histories of length t. Consider the stochastic process $z_t = 2\tau_t/(\rho_t + \tau_t)$.

Note that $z_t = E[\tau \mid \mathcal{F}_t]$ where the expectation is taken under the probability measure $\frac{1}{2}(\tau + \rho)$. By Lévy's martingale convergence theorem (See Shiryaev (1995), Chapter 7, Section 4, Theorem 3), z_t converges $\frac{1}{2}(\tau+\rho)$-almost everywhere to z, a Radon-Nikodym derivative of τ with respect to $\frac{1}{2}(\tau + \rho)$. As ρ is absolutely continuous with respect to $\frac{1}{2}(\tau + \rho)$, z_t also converges ρ-almost everywhere to z. By lemma 2, as τ and ρ are orthogonal, τ is zero for every ρ-almost output sequence. Therefore, for every ρ-almost output sequences,

$$\lim_{t\to+\infty} z_t = \lim_{t\to+\infty} 2\frac{\tau_t}{\tau_t + \rho_t} = 0,$$

implying

$$\lim_{t\to+\infty} \frac{\tau_t}{\rho_t} = 0.$$

Proof of Lemma 3

Taking the first order derivative of $f(x)$ yields

$$f'(x) = \log\frac{\alpha_1}{\alpha_2} + \frac{\alpha_2 - 1}{1 + (\alpha_2 - 1)x} - \frac{\alpha_1 - 1}{1 + (\alpha_1 - 1)x}.$$

Thus, I can compute $f'(0) = \log\frac{\alpha_1}{\alpha_2} + \alpha_2 - \alpha_1$ and $f'(1) = \log\frac{\alpha_1}{\alpha_2} + \frac{\alpha_2 - \alpha_1}{\alpha_1\alpha_2}$. Assumption 2 implies that $f'(0+) < 0$ and $f'(1-) < 0$.

Since $f(0) = f(1) = 0$, I can easily derive that $f(0+) < 0$ and $f(1-) > 0$.

The continuity of $f(x)$ on $[0,1]$ implies that there exists at least one x^* such that $f(x^*)=0$.

Rewrite $f'(x)$ as follows:

$$f'(x)=\log\frac{\alpha_1}{\alpha_2}+\frac{\alpha_1-\alpha_2}{(\alpha_1-1)(1-\alpha_2)}\left[x^2-\frac{\alpha_1+\alpha_2-2}{(\alpha_1-1)(1-\alpha_2)}x-\frac{1}{(\alpha_1-1)(1-\alpha_2)}\right]^{-1}.$$

When Assumption 1 holds, $f'(x)$ increases in $[0,\frac{\alpha_1+\alpha_2-2}{2(\alpha_1-1)(1-\alpha_2)}]$ and decreases in $[\frac{\alpha_1+\alpha_2-2}{2(\alpha_1-1)(1-\alpha_2)},1]$. This property ensure that x^* is unique and also $f(x)$ is negative when x is below x^* and $f(x)$ is positive when x is above x^*.

Proof of Proposition 9

Firstly, define a critical value

$$\beta=\frac{-E\left[f(x(t+1,r))\mid x(t+1,r)<x^*\right]}{E\left[f(x(t+1,r))\mid x(t+1,r)\geq x^*\right]-E\left[f(x(t+1,r))\mid x(t+1,r)<x^*\right]}$$

Rewriting the Bayes' Rule as follows:

$$q_{t+1}(\tau_1)=\left(1+\frac{q_t(\tau_2)}{q_t(\tau_1)}\frac{\tau_2(W_{t+1}\mid W_{t-r+1},\ldots,W_t)}{\tau_1(W_{t+1}\mid W_{t-r+1},\ldots,W_t)}\right)^{-1}$$

Since $\frac{q_t(\tau_2)}{q_t(\tau_1)}=\frac{q_{t-1}(\tau_2)}{q_{t-1}(\tau_1)}\cdot\frac{\tau_2(W_t|W_{t-r+1},\ldots,W_{t-1})}{\tau_1(W_t|W_{t-r+1},\ldots,W_{t-1})}$, taking the recursive form into the formula yields

$$q_{t+1}(\tau_1)=\left(1+\frac{q_0(\tau_2)}{q_0(\tau_1)}\prod_{j=r}^{t}\frac{\tau_2(W_{j+1}\mid W_{j-r+1},\ldots,W_j)}{\tau_1(W_{j+1}\mid W_{j-r+1},\ldots,W_j)}\right)^{-1}.$$

To prove that $\lim_{t\to+\infty} q_{t+1}(\tau_1) = 1$ a.e., it is equivalent to show

$$\prod_{t=r}^{+\infty} \frac{\tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\tau_2(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)} = +\infty \text{ a.e..}$$

Take log on both sides,

$$\sum_{t=r}^{+\infty} \log \frac{\tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\tau_2(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)} = +\infty \text{ a.e..}$$

Referring to the definition of τ_1 and τ_2, I know that $\tau_i(\cdot \mid W_{t-r+1}, \ldots, W_t)$, $i = 1, 2$ is bounded. Taking expectations on both side of the above equation,

$$\sum_{t=r}^{+\infty} E\left[\log \frac{\tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\tau_2(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}\right] = +\infty, \text{a.e..} \tag{6.5}$$

Recall that $x(t+1, r)$ is the conational probability that the current and previous outcomes is related. Then, it is easy to show that

$$\begin{aligned} & E\left[\log \frac{\tau_1(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}{\tau_2(W_{t+1} \mid W_{t-r+1}, \ldots, W_t)}\right] \\ = & \log \frac{1 + (\alpha_2 - 1)x(t+1, r)}{1 + (\alpha_1 - 1)x(t+1, r)} + x(t+1, r) \log \frac{\alpha_1}{\alpha_2} \\ = & f(x(t+1, r)) \end{aligned}$$

Take expectations with respect to $W_{t-r+1}, \ldots, W_t$ on both sides of Equation 6.5,

$$\sum_{t=r}^{+\infty} E\left[f(x(t+1,r))\right] = +\infty, \tag{6.6}$$

It is easy to compute the following:

$$\begin{aligned} & E\left[f(x(t+1,r))\right] \\ = \; & E\left[f(x(t+1,r)) \mid x(t+1,r) \geq x^*\right] \rho(x(t+1,r) \geq x^*) \\ + \; & E\left[f(x(t+1,r)) \mid x(t+1,r) < x^*\right] \rho(x(t+1,r) < x^*) \end{aligned}$$

Lemma 2 suggests that $E\left[f(x(t+1,r)) \mid x(t+1,r) \geq x^*\right] \geq 0$ and $\left[f(x(t+1,r)) \mid x(t+1,r) < x^*\right] < 0$. Thus, $\beta \in (0,1]$. $E\left[f(x(t+1,r))\right] > 0$ is equivalent to $\rho(x(t+1,r) \geq x^*) > \beta$. Therefore, I can show that when $\rho(x(t+1,r) \geq x^*) > \beta,\ \forall t \geq r$, Equation 6.6 holds and I prove result 1) in this proposition.

Similar arguments can be extend to prove result 2).