THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

TWO ESSAYS ON EQUILIBRIUM ANALYSIS OF CUSTOMERS' QUEUEING BEHAVIOR

HUANG FENGFENG

PhD

The Hong Kong Polytechnic University

2019

The Hong Kong Polytechnic University

Department of Logistics and Maritime Studies

Two Essays on Equilibrium Analysis of Customers' Queueing Behavior

Huang Fengfeng

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

February 2019

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Huang Fengfeng (Name of student)

Abstract

In the past decades, the development of information technology has reshaped how customers obtain product related information, and has revolutionized the ways organizations relate to the marketplace. Meanwhile, progress in behavioral research has deepened our understanding of customer behavior. It shows that customers are not passive receiving end but active decisions makers. Those changes in technology and knowledge have created a new world of opportunities and challenges for all aspects of the enterprise. In this thesis, we research into their impacts on the operations of service facilities.

In the first topic, we consider the new challenges associated with the development of information technology. Social media and word-of-mouth forums are shown to have great influence on customers' purchase decisions, and have made the buyer-generated content non-negligible. We consider a typical situation where a service provider serves two types of customers, sophisticated and naive. Sophisticated customers are well-informed about service-related information and make their joining-or-balking decisions strategically, whereas naive customers do not have such information and rely on online rating information to make such decisions. We demonstrate that under certain conditions a service provider can increase its profitability by simply 'dancing' its price, that is, replacing the static pricing strategy with a cyclic high-low pricing strategy. The success of this strategy relies on two key conditions: the potential market size is large enough so that congestion is a key concern in the service system and the rating provides the average price and average utility information. We also show that when customers are loss averse and the rating serves as a reference point for them, the loss aversion behavior dilutes the effectiveness of cyclic pricing strategy. Finally we show that

the cyclic pricing strategy is never socially optimal.

In the second topic, we incorporate new understandings of customer behavior with their equilibrium queueing decisions in health care service context. We investigate the patient' doctor shopping behavior when they seek diagnostic service. When a patient' belief about her health status is inconsistent with a doctor's diagnosis, cognitive dissonance may arise. The patient may seek more doctoral opinions to mitigate such dissonance without referrals, that is, the patient is engaged in doctor shopping. We derive the patient' optimal "doctor shopping" stopping time by adopting the simple 'one-stage look-ahead' rule and Bayesian updating. We show that a patient stops the doctor shopping process whenever two successive diagnostic results are consistent with each other. Doctor shopping always results in a highly system congestion. Interestingly, we find that the patients' doctor shopping may not necessarily undermine the social welfare. Doctor shopping improves patients' psychological gains. Doctor shopping shall always be tolerated when the accuracy of the diagnosis is not very high. When the accuracy of the diagnosis is high, doctor shopping may be also tolerated if the policy maker cares about the patients' psychological gains.

Publications Arising from the Thesis

Huang F., Guo P., Wang Y. 2018. Dancing Service Price When Customers Queue with Rating Information. Production Operations Management, major revision.

Huang F., Guo P., Wang Y. 2018. Illness Perceptions and Doctor-Shopping Behavior among Patients. Up to Submit to Production Operations Management.

Acknowledgment

First, I would like to express my gratitude to Prof. Guo Pengfei, my Chief Supervisor, and Dr. Wang Yulan, my Co-Supervisor. I am grateful to Prof. Guo for his supervision. His patient guidance and support have helped me overcome many difficulties throughout my PhD study. I am thankful to have Dr. Wang as my Co-Supervisor. Her constructive suggestions have improved the elegance of my research outputs. I feel fortunate to have both of them as my supervisors.

I would also like to thank my fellow research students friends in LMS. Life has never been boring with them, not to mention that we have had numerous incredible food journeys together. Meanwhile, I shall not forget the small hiking group led by Prof. Ji Ping from ISE and my gym and fighting club friends, without whom I would have gained a lot more weight.

My special thanks to my parents and my sister for their unconditional support and trust. I feel lucky to have such cool family, who believe in my value on my own but not on whether I have found a man.

I was told that the journey to get a PhD is painful and tedious. Thank to all of them, my life in Hong Kong is everything but what I was told.

Table of Contents

List of Figures

Chapter 1

Introduction

The advance of information technology in the past decade has fundamentally changed how we communicate and consume. Meanwhile, it has revolutionized the ways organizations relate to the marketplace, creating a new world of possibilities and challenges for all aspects of the enterprise (Aral et al. 2013). For example, the prosperous of the Internet forums, such as word-of-mouth websites like Yelp.com and Dianping.com, and social media, have allowed customers to easily share their consumption experiences through posting reviews and ratings. This buyer-generated contend has become a new peer-to-peer channel for customers to obtain service-related information. According to the latest global report on www.pwc.com,[1.1] 78% of respondents say that social media has influenced their purchase decisions.

Knowing the value of buyer-generated information, some retailers resort to forum manipulation to boost sales. Even though some researchers, such as Mayzlin et al. (2014), Dellarocas (2006), Mayzlin (2006), suggest that forum manipulation does not necessarily lead to less informative ratings and reviews, it raises concerns of the customers and causes ethical and legal issues. The concerns of the customers spurs the emergence of websites like www.fakespot.com, which help customers to spot dodgy product reviews/ratings. From the perspective of online retailing platforms, protecting customer welfare matters not only the images of the platforms, but also their prosperity. Two online retailing giants, US-based

[1.1] The information can be accessed at https://www.pwc.com/gx/en/industries/retail-consumer/global-total-retail/global-key-findings.html.

Amazon and China-based Alibaba, have promised to fight against forum manipulation.

However, we shall demonstrate that a service firm can rip off uninformed customers by utilizing the rating information without raising legal and ethical issues. It can strategically manage the ratings by simply "dancing" its price, that is, replace the static pricing strategy with a cyclic pricing strategy. Towards this end, in Chapter 2, we consider a monopoly service provider serving two types of customers who are heterogeneous in information access, namely, sophisticated and naive customers. Sophisticated customers are aware of service-related information such as service rate and service reward, and they make strategic decisions by taking into consideration the joining-and-balking decisions made by others. Naive customers, however, do not have such service-related information, so they rely on buyer-generated information to make the decisions; they join if price is not higher than the average rating and balk otherwise.

We propose a cyclic pricing strategy for the service provider to strategically manage buyer-generated information and to generate high profit with the service facilities. The cyclic pricing strategy requires that the congestion effect plays a profound role; otherwise, it does not work. One extreme case is the goods market, in which delay is not an issue and how much a customer enjoys a goods item is determined merely by her idiosyncratic features. Moreover, the cyclic pricing strategy is never socially optimal.

Other than the challenges led by the advance of technology, new understandings of the customer behavior is also challenging how the service facilities shall be managed. Conventional service operations research treats customers as the passive receiving end. This thinking has helped us to understand the process of most categories of services we encounter in daily life. However, under certain circumstances, customers are joint decision makers, along with the providers. A typical situation is medical consultation, which is a kind of credence service rendered by experts. Customers, i.e., patients, have little knowledge about their needs, and hence cannot assess the quality of the service they received. On the

other hand, they usually have biased opinions on their own situation, and are constantly found visiting multiple medical practitioners during a single illness episode without referrals, namely, engaging in doctor shopping behavior (Kasteler et al. 1976).

Despite that doctor shopping behavior is widely observed in numerous health care systems, most medical practitioners believe that doctor shopping causes unnecessary and repetitive consultations and tests and shall be avoided on cost ground. In Chapter 3, we investigate the dynamics of doctor shopping behavior and its effect on the health care system. We consider a stream of patients facing with a similar set of symptoms but having different illness perceptions. An individual's illness perception measures her prior belief about on what probability she is sick; it is generally not consistent with the doctor's belief (which is consistent with reality). Both the doctor and the patient update their beliefs according to Bayes rule upon obtaining a diagnostic result. The doctor decides whether to dismiss a patient or send her to further treatment; the patient decides whether to follow the doctor's advice or seek a second opinion.

We show that whether a patient seeks the service and how many visits a patient pays in one illness episode are jointly determined by her own illness perception, the costs associated with each visit and the doctor's diagnosis quality (that is, the degree of accuracy). Whenever two successive diagnostic results are consistent with each other, the patient stops doctor shopping. Moreover, although repetitive diagnoses (due to patients' doctor shopping) help little in improving the objective reward of the diagnostic service, it may patients' psychological gains. A welfare-maximizing social planner, when taking into account patients' psychological gains, shall tolerate a certain level of doctor shopping.

Chapter 2

Dancing Service Price When Customers Queue with Rating Information

2.1 Introduction

When customers seek service, they make their joining-or-balking decisions based on the information they have access to. Customers have varying accesses to information; some are aware of service-related information such as service quality and service rate, whereas others are not. Instead of making blind decisions, customers who are not aware of service-related information often rely on buyer-generated information, such as customer ratings and reviews online, to make their decisions. In fact, buyer-generated information has reshaped customers' habits. Checking ratings/reviews before making consumption decisions has become a ritual for many of today's customers.

The cyclic pricing strategy is often adopted by service providers. For example, in the US, many theme parks offer discounts from March to October, the high tourist season.[2.1] Similarly, many tourist cities of China offers discounts on entrance fees to some tourist destinations from July to September. For example, in 2018, the tourism bureau of Guizhou province, China, published advertisement on several portal websites announcing that a 50% discount would be offered to tourists to several popular tourist destinations in the province.[2.2]

[2.1] See https://www.moneysavingexpert.com/deals/cheap-theme-parks/#longleat.
[2.2] See, for example http://www.sohu.com/a/236603031_395859.

In this work, we will demonstrate that, the cyclic pricing strategy, together with rating-dependent customers, can achieve a higher profit than the static pricing strategy. To illustrate the mechanism behind this, we consider a monopoly service provider serving two types of customers who are heterogeneous in information access, namely, sophisticated and naive customers. Sophisticated customers are aware of service-related information, such as service rate and service reward, and they make strategic decisions by taking into consideration the joining-and-balking decisions made by others. Naive customers are those one-time shoppers, and they do not have such service-related information. They rely on buyer-generated information to make the decisions; they join if price is not higher than the average rating and balk otherwise. For example, local residents often know the food quality (service reward) and the probable waiting times of their nearby restaurants, whereas tourists generally do not know such information, and they have to rely on reviews/ratings posted on websites such as Dianping.com and Yelp.com to decide whether to join.

The proportion of each type of customer, that is, the customer type composition parameter, is known to the service provider. For example, according to a report on Variety.com,[2.3] local visitors account for 39% of the total attendance at Hong Kong Disneyland, while tourists (mainland Chinese and international) comprise the other 61% (41% and 20%, respectively, for mainland Chinese and international visitors). The interaction between the service provider and customers is modeled as an $M/M/1$ queue. The queue length is unobservable to the customers. In the main part, we assume that customers, after obtaining the service, posts a rating equal to her *consumption utility* (service reward less the waiting cost). The *average rating* on consumption utility is then advertised to future arriving customers (we assume that incoming customers do not realize the whole history of rating information). We also extend our study to other information scenarios where the server reveals the average rating on the *net utility* (consumption utility less the price) and the *average price* to incoming customers.

[2.3]The details can be found at http://variety.com/2016/biz/asia/hong-kong-disneyland-in-loss-1201706466/.

We illustrate that the mechanism still works in the second information scenario.

We show that the optimal pricing strategy, if is cyclic, must be of high-low type: sophisticated customers join during the high-price phase and obtain a net utility of zero, while naive customers join during the low-price phase and obtain a negative utility. Recall that a customer's consumption utility is determined by the congestion level of the system. The system is less congested during the high-price phase; hence, the ratings on consumption utility are higher. Ratings from both sophisticated and naive customers are averaged, and the average rating is advertised to incoming customers. The service provider just needs to charge a price equal to the average rating to lure naive customers into joining, which turns out to be higher than their expected consumption utility. In short, the high-price phase uplifts the average rating, which then allows the service provider to rip off naive customers during the low-price phase. Numerical studies show that the profit increment brought about by cyclic pricing is around 5%. Maybe this incremental amount is not regarded as high, but it could be particularly important for entertainment parks like Hong Kong Disneyland that are struggling to break even financially.[2.4]

The intriguing thing about this mechanism of improving profit with cyclic pricing is that it fully assumes *honest feedbacks* from customers who have experienced the service, without any distortion of their ratings. This feature makes it easy to be implemented, without worrying about the ethical issues. It is worth mentioning that the adoption of the cyclic pricing strategy requires *the potential market size to be above a certain threshold value* so that congestion is a significant factor affecting customers' patronizing decision. A customer's consumption utility is affected by the congestion of the system, which, in turn, is determined by effective arrival rates controlled by the prices. When the potential arrival rate is small, or the service capacity is very large, the congestion effect does not play a profound role. Thus, the aforementioned cyclic-pricing mechanism does not work; instead, static pricing is preferred. One extreme case is the goods mar-

[2.4]Please refer to http://www.scmp.com/news/hong-kong/economy/article/2133962/hong-kong-disneyland-falls-further-red-losses-double-2017-hit for the details.

ket, in which delay is not an issue, and in which how much a customer enjoys a goods item is determined merely by her idiosyncratic features. In such a case, our 'dancing price' strategy does not work.

In the extension part, we consider another information scenario where customers post ratings on their net utility, namely consumption utility less the price. The extra condition for cyclic pricing strategy to work is that the average price shall be provided to incoming customers. If this is the case, the incoming customer can still nail down the rating on the consumption utility if their utility function is linear, and hence the above cyclic pricing mechanism still works. However, if customers' utility function is nonlinear, cyclic pricing still works in some situations but its effect is dampened.

We consider a typical situation for non-linear customer utility where customers are loss averse: they treat the rating information as their reference points, and their feeling of loss when their actual experienced quality is below this reference point is stronger than the equal-sized gain when their experienced quality is higher than the reference point. Such behavior results in a kinked utility function, which is nonlinear. We extend our analysis to this case and numerically show that such loss-averse behavior reduces the effectiveness of the cyclic pricing strategy: the more loss-averse customers are, the less effective the cyclic pricing strategy is. This is because that as the server rips off the naive customers by revealing only aggregate rating information, those customers may feel a big loss after finding out that the actual quality is much lower than the rating and hence will likely post a very low score online. Hence, when managers consider the cyclic pricing and using the average rating strategy to boost their profit, they shall take into considerations of the system's *congestion level* and whether *customers are loss averse or not*.

Finally, we examine whether cyclic pricing strategy is socially desired; that is, whether it can maximize the sum of service provider's profit and customers' surplus. In the classic queueing literature with unobservable queues and identical customers, all customer surplus is internalized through pricing and goes to the

service provider. Therefore, welfare- and profit-maximization are equivalent; see Hassin and Haviv (2003). However, in our model, customers differ on information access. We show that the welfare-maximizing pricing strategy is always static and hence a high-low cyclic pricing strategy is never socially optimal. The key rational behind is that cyclic pricing strategy allows the firm to rip off naive customers because when they are lured to join the system by the rating information, they do not consider the negative externality of their joining behavior on others and the system is too congested, causing welfare loss. In fact, we show that, when the cyclic pricing strategy is adopted, naive customers always obtain a negative expected utility during the low-price phase, which renders that the profit of the service provider is higher than the social welfare it offers. Welfare-maximization prefers even workloads across the time periods, whereas high-low pricing brings about uneven workloads. In most cases, the system is heavily congested during the low-price phase and underutilized during the high-price phase. Interestingly, in certain situations where the market size is not large while at least half of the population is sophisticated customers, welfare loss is caused by underutilization during both phases.

The remainder of this chapter is organized as follows. In Section 2.2, we review the related literature. We introduce basic assumptions and the model setup in Section 2.3. In Section 2.4, we analyze the provider's optimal pricing strategies. We compare the pricing strategies and equilibrium outcomes of the pricing strategies in Section 2.5. Moreover, we check the performance of the cyclic pricing strategy under the scenario that customers are reference-dependent in Section 2.6. Section 2.7 concludes this chapter. All the proofs are relegated to the online Appendix.

2.2 Literature Review

Our work belongs to the stream of study on customers' strategic queueing behavior and the provider's information disclosure. Pioneering work on this can be traced back to Naor (1969), who studies the equilibrium joining strategies

of customers under observable queues and proposes regulating the demand rate by imposing fees or tolls. Edleson and Hildebrand (1975) extend the analysis to an unobservable $M/M/1$ queue. Many studies have been conducted along these lines since then; see Hassin and Haviv (2003) and Hassin (2016) for the comprehensive surveys of studies in this field. In the following, we shall review the most closely related works, which we classify into the following three streams of research: information disclosure in queues, customers' bounded rationality in making queueing decisions, and service pricing.

Studies concerning information disclosure in queues can be further classified according to information content: waiting time, service reward, and service rate. A number of papers consider whether queue length information should be released for free, including Hassin (1986), Whitt (1999), Armony and Maglaras (2004), Dobson and Pinker (2006), Guo and Zipkin (2007), Allon et al. (2011) and some others. Another group of studies considers offering customers the choice of paying a fee to inspect queue length, including Hassin and Haviv (1994), Hassin and Roet-Green (2011, 2013). All these papers assume that customers are strategic and that they have the same access to the service information. In our work, however, only a proportion of customers are aware of such information; others simply follow the online rating information.

Our model is closely related to the work of Hu et al. (2017), who consider the effect of customer information heterogeneity on waiting time. In their model, some customers are informed of a real-time delay (informed customers) and others (uninformed customers) make their joining/balking decisions based on past experiences. They find that a certain level of information heterogeneity leads to more efficient outcomes. Similar to Hu et al. (2017), we consider customers' information heterogeneity. In our model, a proportion of customers (sophisticated customers) are aware of service-related information such as service reward and service rate, whereas others (naive customers) do not have such information. Different from Hu et al. (2017), the queue length under our setting is unobservable to both types of customers. We find that a profit-maximizing provider can make

higher profits by taking advantage of information heterogeneity, but it never leads to higher social welfare.

A second stream of literature on information disclosure in queues considers imperfect service reward information; these studies include Veeraraghavan and Debo (2009, 2011), Debo et al. (2012), and Guo et al. (2015). The study by Veeraraghavan and Debo (2009) shows how information externalities due to congestion affect customers' choice between two servers. In their model, customers have private information about service quality and queue length. They find that information externalities lead to cycles during which one server is thriving and the other is not. Veeraraghavan and Debo (2011) study the herding behavior of customers choosing between two congested services with unknown service qualities in which customers observe an imperfect private signal on service qualities and queue lengths before making their choices. They characterize the equilibrium joining behavior and the extent of customer learning from the queue information. Debo et al. (2012) examine customers' queueing strategy when the service reward is known only by a proportion of customers, and find that uninformed customers adopt a "hole" strategy, i.e., they balk when the queue is at a certain length (called the "hole") and behave the same as informed customers otherwise. Guo et al. (2015) examine two competing servers with unknown qualities. They show that under certain conditions, the low-quality server has a higher incentive to reveal queue length. In contrast to these models in which customers are always strategic, only a portion of customers in our model are strategic, with others being naive.

A third stream of work on information disclosure in queues focuses on the information about service rate and arrival rate, including Guo et al. (2011), Debo and Veeraraghavan (2014), Cui and Veeraraghavan (2014), and Afeche and Ata (2011). Guo et al. (2011) and Debo and Veeraraghavan (2014) assume that customers do not know the service rate but have information on its distribution. Cui and Veeraraghavan (2014) consider "blind queues", in which customers only know some vague information on service rate. Afeche and Ata (2011) propose the

"learning-and-earning" problem where customers are classified into patient and impatient customers but the proportion of each type is unknown.

Our work is related to studies on the bounded rationality of customers, including Huang et al. (2013), Huang and Chen (2015), and Li et al. (2016). In these studies, customers make their decisions without fully assessing service quality and waiting time due to limited cognitive ability or a lack of opportunities. Huang et al. (2013) demonstrate that strategically taking advantage of the customer bounded rationality may lead to significant increases in both revenue and social welfare. Li et al. (2016) consider customer-intensive services and find that revenue-maximizing firms do not always exploit customers' bounded rationality and may leave positive net utility to customers under certain circumstances. Huang and Chen (2015) consider anecdotal reasoning customers, who rely on a limited sample to make queueing decisions. They find that with anecdotal reasoning, customers are less price-sensitive, and that revenue and welfare maximization lead to different pricing strategies. Similar to these models, the naive customers in our model are boundedly rational. However, customers in our system consist of both sophisticated and naive customers, whereas most of the above studies just consider a single type of customers.

Our model is also related to the pricing strategy for service facilities. Dewan and Mendelson (1990) first propose a joint optimization of capacity and pricing. Stidham (1992) shows that even for a simple $M/M/1$ system, the joint pricing and capacity problems have multiple local optima. Numerous studies on capacity investment and admission control have been conducted; see Stidham (2009) for a comprehensive review.

In addition, much research has been conducted on vacation queueing systems, in which a service shuts down when no customers are present and resumes when the queue reaches a critical length (Guo and Hassin 2011, 2012, Guo and Li 2013, Guo and Zhang 2013, Wang and Li 2008, Zhang et al. 2013, Economou et al. 2011). Similar to these vacation queueing systems, the system in our study oscillates between high- and low-arrival states when the provider adopts a cyclic

pricing strategy.

Lastly, our model relates to the research on consumer ratings/reviews. Numerous researches have explored how rating and review information can affect a firm's pricing and information disclosure strategies in the retailing business, and what information a potential consumer is likely to gain from this buyer-generated information. Two of these works are closely relevant to our own work. The first is Crapis et al. (2016), which analyzes the social learning mechanism and its effect on a seller's pricing decision. In their model, customers, after consumption, rate the product as either "like" or "dislike"; later-arriving customers observe the rating profile (the proportion of "like" and "dislike") and infer how much they are likely to enjoy the product. They compare two pricing strategies– a static price and one with a single price change– and suggest that pricing policies that account for social learning may increase revenues. The second is Shin and Zeevi (2017), which studies a fluid model and investigates a monopolist's optimal dynamic pricing strategy over a finite horizon. In both models, customers have private information on their preferences, and the demand function evolves in conjunction with the review profile/dynamics. As far as we know, our work is one of the first works on customer ratings in the context of service/queueing.

2.3 Model Setup and Preliminaries

Consider a monopoly service provider (he) whose service times are i.i.d and exponentially distributed with rate μ. Potential customers arrive according to a Poisson process with rate Λ. Once served, a customer (she) receives a service reward R, and incurs a price p and a waiting cost that is proportional to her waiting time in the system (the sum of the waiting time in the queue and the service time) with a unit-time cost c. Those customers who have joined the queue form an arrival process with rate $\lambda(p)$, which is called the effective arrival rate. The service system can hence be modeled as an $M/M/1$ queue, and the waiting time W is an exponential random variable with parameter $\mu - \lambda(p)$, i.e., $W \sim \exp(\mu - \lambda(p))$. Clearly, $E(W) = 1/(\mu - \lambda(p))$.

There are two types of customers: θ proportion of them are *sophisticated* and have the knowledge about the service rate μ and the service reward R; the rest $1-\theta$ proportion are *naive* and do not have such information, representing those one-time customers. For example, local residents generally are aware of service-related information such as the probable waiting times and food quality of a restaurant whereas tourists have no such information. We assume that the customer type composition parameter θ is common knowledge. Denote the potential market size, i.e., the potential arrival rates, of sophisticated and naive customers by Λ_s and Λ_n, respectively. Then,

$$\Lambda_s = \theta\Lambda \text{ and } \Lambda_n = (1-\theta)\Lambda.$$

For the pricing strategy, we consider a general cyclic pricing strategy. The time horizon is divided into periods with length T. Each period is divided into N phases indexed by i. The price in phase i is denoted as p_i and the corresponding phase length is denoted as T_{p_i}. The general cyclic pricing strategy $\boldsymbol{p}$ can be captured by a finite sequence of N combinations of *prices* p_i and the *phase length* that price p_i continues for (T_{p_i}), i.e., $\boldsymbol{p} = \{(p_i, T_{p_i})|i = 1, 2, ...N\}$. Clearly, $T = \sum_{i=1}^{i=N} T_{p_i}$. When $N = 1$, the cyclic pricing strategy degenerates into a static one. We assume that the cycle length T is long enough compared to the expected waiting time such that the time of the transient process to the steady states is negligible in each pricing circle. For example, for typical service providers on yelp.com, such as restaurants, dentists and optometrists, the waiting time per service episode is normally no more than 1 hour, whereas the cyclic length is one month (Yelp.com displays a "Monthly Trend", which shows the average rating of each month). Hence, the transient process within each pricing circle is negligible.

The service provider decides on his pricing policy $\boldsymbol{p}$ to maximize his long-run average profit per unit of time as follows:

$$\max_{\boldsymbol{p}} \quad \Pi = \frac{1}{T}\left[\sum_{i=1}^{N} p_i \lambda(p_i) T_{p_i}\right],$$

where $\lambda(p_i)$ is the effective arrival rate of customers when the service provider charges a price p_i, which includes both naive and sophisticated customers. This

objective function is in principle equivalent to maximizing the "profit rate" as in the queueing literature (see, for example, the comprehensive review of Hassin and Haviv (2003)), where the provider's objective is generally to maximize $p\lambda(p)$, where $\lambda(p)$ is an arrival rate. Define L_{p_i} as follows:

$$L_{p_i} = \frac{T_{p_i}}{T}.$$

In other words, L_{p_i} represents the proportion of time the price p_i is charged by the service provider in a cycle. Then, we can rewrite the provider's long-run average profit as

$$\max_{\boldsymbol{p}} \quad \Pi = \sum_{i=1}^{N} p_i \lambda(p_i) L_{p_i}.$$

2.3.1 Average Rating and Equilibrium Arrival Rates

After the service, each customer, regardless of her type, *honestly rates* her *consumption utility*, $R - cw$, where w is the actual waiting time she has experienced. Ratings accumulate over time and the *average rating on the consumption utility* is advertised to incoming customers by the server. Customers are also informed about the current period price upon arrival. By simply comparing the average rating on the consumption utility with the posted current-period price, naive customers decide to join or not.

As our goal is to illustrate the effect of cyclic pricing in the long run, we directly consider the performance of the system *during the stable states* and ignore the transient process leading up to them. Given a pricing strategy $\boldsymbol{p} = \{(p_i, T_{p_i}) | i = 1, 2, ...N\}$, the average rating converges to a constant number in the long run, and is a function of $\boldsymbol{p}$, which we denote as $\eta(\boldsymbol{p})$. The online Appendix B illustrates such a convergence process if customers adopt exponential smoothing to aggregate all ratings. Here, we shall not expand our discussion on this as it is not our focus. Instead, we directly move on to solve the long-run performance; that is, for a given pricing strategy $\boldsymbol{p}$, the average rating $\eta(\boldsymbol{p})$ and equilibrium arrival rates in each phase $\lambda(p_i)$'s, $i = 1, 2, ...N$, can be obtained through solving multiple equations. Given arrival rates $\lambda(p_i)$'s, we can determine the average rating $\eta(\boldsymbol{p})$;

given the average rating $\eta(\boldsymbol{p})$, we can derive the equilibrium arrival rates in each phase. Below, we illustrate these two steps in details.

Step 1: Determine the average rating $\eta(\boldsymbol{p})$ given arrival rates.

$\eta(\boldsymbol{p})$ is determined by customers' queueing behavior through the whole pricing cycle, and can be written as

$$\eta(\boldsymbol{p}) := \frac{\sum_{i=1}^{i=N} v(p_i)\lambda(p_i)T_{p_i}}{\sum_{i=1}^{i=N} \lambda(p_i)T_{p_i}} = \frac{\sum_{i=1}^{i=N} v(p_i)\lambda(p_i)L_{p_i}}{\sum_{i=1}^{i=N} \lambda(p_i)L_{p_i}}, \tag{2.1}$$

where

$$v(p_i) := R - cE[W(p_i)] = R - \frac{c}{\mu - \lambda(p_i)} \tag{2.2}$$

is the expected consumption utility of a customer who joins at price p_i.

Step 2: Determine arrival rates given average rating $\eta(\boldsymbol{p})$.

A customer arriving during phase i observes the current price p_i and the average rating $\eta(\boldsymbol{p})$. She then makes the joining-or-balking decision to maximize her utility. Denote $\delta_n(p_i)$ and $\delta_s(p_i)$ as the joining decisions of the naive and sophisticated customers, respectively, when faced with the price p_i.

As naive customers do not have information about service-related parameters R and μ, they are incapable of anticipating their expected utility. Instead, they rely on the online rating information to decide whether to join or to balk. A naive customer joins if $p_i \leq \eta(\boldsymbol{p})$ and balks otherwise. Note that the assumption of customer joining when $p_i = \eta(\boldsymbol{p})$ (the break-even case) is not critical, as the service provider can always reduce the price by an infinitesimal amount to lure naive customers to join in practice. The joining decision of a naive customer is hence captured by a binary variable:

$$\delta_n(p_i) = \begin{cases} 1, & \text{if } p_i \leq \eta(\boldsymbol{p}) \\ 0, & \text{if } p_i > \eta(\boldsymbol{p}) \end{cases}. \tag{2.3}$$

As all naive customers have the same information access, they all join ($\delta_n(p_i) = 1$) or all balk ($\delta_n(p_i) = 0$). Note that this scenario is different from the no-information case referred to in some queueing strategy literature (see Guo and Zipkin (2007)). There, the queue is unobservable but customers still know other information about the system, such as potential arrival and service rates; thus,

uninformed customers can still make a strategic queueing decision, that is, by adopting a mixed-strategy in joining. Here, the herding behavior of naive customers results from lacking information. The system capacity is assumed to be large enough such that when naive customers join collectively, their expected consumption utility is non-negative i.e., $R - \frac{c}{\mu-(1-\theta)\Lambda} \geq 0$.

Sophisticated customers are strategic in making their queueing decisions. They choose a joining probability $\delta_s(p_i)$ to maximize their expected utility $R - cE[W(p_i)] - p_i$, in which they also take naive customers' joining decision $\delta_n(p_i)$ into consideration:

$$\mathcal{U}(p_i) = \max\{R - cE[W(p_i)] - p_i, 0\} = \max\left\{R - \frac{c}{\mu - \lambda(p_i)} - p_i, 0\right\}, \quad (2.4)$$

where the effective arrival rate is

$$\lambda(p_i) = \delta_s(p_i)\Lambda_s + \delta_n(p_i)\Lambda_n. \quad (2.5)$$

Clearly, if the potential market size is large enough, $\delta_s(p_i)$ in equilibrium simply solves the equation

$$R - \frac{c}{\mu - \lambda(p_i)} - p_i = 0.$$

We thus have the following proposition regarding the joining/balking decisions of the two types of customers.

Proposition 2.1 *Given that the service provider charges a price $p_i \in \boldsymbol{p}$, the equilibrium arrival rates are as follows.*

1. *If $p_i > \eta(\boldsymbol{p})$, naive customers all balk, i.e., $\delta_n(p_i) = 0$. Only sophisticated customers join, and they join with probability $\delta_s(p_i) = \min\left\{\frac{1}{\Lambda_s}\left(\mu - \frac{c}{R-p_i}\right), 1\right\}$. The effective arrival rate is $\lambda(p_i) = \min\left\{\mu - \frac{c}{R-p_i}, \Lambda_s\right\}$.*

2. *If $p_i \leq \eta(\boldsymbol{p})$, all the naive customers join, i.e., $\delta_n(p_i) = 1$, whereas sophisticated customers' joining probability is determined by the magnitude of $\eta(\boldsymbol{p})$ and V_n:*

I. if $\eta(\boldsymbol{p}) \geq V_n$ and $V_n \leq p_i \leq \eta(\boldsymbol{p})$, all the sophisticated customers balk, i.e., $\delta_s(p_i) = 0$. The effective arrival rate is $\lambda(p_i) = \Lambda_n = (1-\theta)\Lambda$;

II. otherwise, sophisticated customers join with probability

$$\delta_s(p_i) = \min\left\{\frac{1}{\Lambda_s}\left(\mu - \Lambda_n - \frac{c}{R-p_i}\right), 1\right\}.$$

The effective arrival rate is $\lambda(p_i) = \min\left\{\mu - \frac{c}{R-p_i}, \Lambda\right\}$.

Proposition 2.1 implies that naive customers join only when the price is not high ($p_i \leq \eta(\boldsymbol{p})$). Sophisticated customers, however, always choose a positive joining probability, except that the price falls into an intermediate range of $V_n \leq p_i \leq \eta(\boldsymbol{p})$ provided that $\eta(\boldsymbol{p}) \geq V_n$, because in this case naive customers all join, and they gain a non-positive expected utility, as the price charged is higher than their expected consumption utility V_n.

2.3.2 A Benchmark Case: $\theta = 1$

Before proceeding with the detailed analysis, we first introduce a benchmark case where all the customers are sophisticated, i.e., $\theta = 1$. This case has been widely studied in the literature on customers' strategic queueing behavior (see Chapter 3 of Hassin and Haviv (2003)). Here, we briefly review the result of this traditional setting where the queue length is unobservable.

Consider a monopoly service provider, modeled as an $M/M/1$ queue, who decides on his profit-maximizing price p. As customers are all strategic, the effective arrival rate λ must solve $R - p - c/(\mu - \lambda) = 0$. The profit-maximizing provider's problem is to maximize $p\lambda$, subject to $0 \leq \lambda \leq \Lambda$ and $R-p-c/(\mu-\lambda) = 0$. It can be easily shown that the optimal price

$$p^* = \begin{cases} p_b := R - \sqrt{\frac{cR}{\mu}}, & \text{if } \Lambda > \mu - \sqrt{\frac{c\mu}{R}}, \\ R - \frac{c}{\mu - \Lambda}, & \text{otherwise.} \end{cases} \tag{2.6}$$

For ease of notation, denote

$$\lambda_b := \mu - \sqrt{\frac{c\mu}{R}}. \tag{2.7}$$

That is, when all the customers are strategic, if $\Lambda \leq \lambda_b$, it is optimal to admit all the customers into the system; if $\Lambda > \lambda_b$, some customers balk, and the equilibrium effective arrival rate is λ_b. Note that under this benchmark setting, the optimal decision of a profit-maximizing provider is the same as that of a welfare-maximizing provider.

2.4 Optimal Pricing Decision

In this section, we first present the results on welfare maximization, then show the results about profit maximization and finally compare the system performance under these two scenarios. For the purpose of easy notations, let

$$V_s := R - \frac{c}{\mu - \Lambda_s} \text{ and } V_n := R - \frac{c}{\mu - \Lambda_n}.$$

Note that V_s (V_n, respectively) represents the expected consumption utility when *only* the sophisticated (naive, respectively) customers join the service system.

2.4.1 Welfare Maximization

Anticipating the customers' joining decisions stated in Proposition 2.1, the service provider decides his optimal pricing strategy $\boldsymbol{p} = \{(p_i, T_{p_i}) | i = 1, 2, ..., N\}$. For a welfare-maximizing service provider, the price transfer between the customer and the provider is internalized. Maximizing the long-run average social welfare is equivalent to maximizing the expected total consumption utility given as follows:

$$\max_{\boldsymbol{p}} \mathcal{SW} = \sum_{p_i \in \boldsymbol{p}} v(p_i)\lambda(p_i)L_{p_i} \tag{2.8}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} L_{p_i} = 1, L_{p_i} \geq 0,$$

where $\lambda(p_i)$ is obtained from Proposition 2.1, and $v(p_i)$ is given in (2.2).

For ease of notation, let

$$\underline{\Lambda} = \frac{\lambda_b}{\max\{(1-\theta), \theta\}}, \ \overline{\Lambda} = \frac{\lambda_b}{\min\{(1-\theta), \theta\}},$$

and

$$\widehat{\Lambda} = \frac{1}{2\theta(1-\theta)} \left(\mu - \sqrt{\mu \left[\mu - 4\theta(1-\theta) \left(\mu - \frac{c}{R} \right) \right]} \right).$$

It can be easily shown that $\underline{\Lambda}$ and $\widehat{\Lambda}$ increase (decrease, respectively) in θ while $\overline{\Lambda}$ decreases (increases, respectively) in θ when $\theta < \frac{1}{2}$ ($\theta \geq \frac{1}{2}$, respectively). Then, we have the following proposition.

Proposition 2.2 *The welfare-maximizing pricing strategy is static, i.e., $N = 1$.*

1. *When $\underline{\Lambda} < \Lambda < \overline{\Lambda}$ and $\theta < \frac{1}{2}$, the optimal price is as follows:*

   (a) *If $\underline{\Lambda} < \Lambda \leq \widehat{\Lambda}$, $p^*_{sw} = V_n = R - \frac{c}{\mu-(1-\theta)\Lambda}$. All naive customers join while all sophisticated customers balk, that is, $\delta_s = 0$ and $\delta_n = 1$;*

   (b) *If $\widehat{\Lambda} < \Lambda < \overline{\Lambda}$, $p^*_{sw} = V_s = R - \frac{c}{\mu-\theta\Lambda}$. All sophisticated customers join while all naive customers balk, i.e., $\delta_s = 1$ and $\delta_n = 0$.*

2. *Otherwise, the service provider always sets the price equal to that under the benchmark case in which all customers are sophisticated, that is, $p^*_{sw} = p^*$, where p^* is given by (2.6). In particular,*

   (a) *If $\Lambda \leq \lambda_b = \mu - \sqrt{\frac{c\mu}{R}}$, $p^*_{sw} = R - \frac{c}{\mu-\Lambda}$. Both naive and sophisticated customers join, i.e., $\delta_s = \delta_n = 1$;*

   (b) *If $\lambda_b < \Lambda \leq \underline{\Lambda}$, or if $\underline{\Lambda} < \Lambda < \overline{\Lambda}$ and $\frac{1}{2} \leq \theta < 1$, $p^*_{sw} = R - \sqrt{\frac{cR}{\mu}}$. Naive customers all join, i.e., $\delta_n = 1$, whereas sophisticated customers join with probability $\delta_s = \frac{\lambda_b - \Lambda_n}{\Lambda_s}$;*

   (c) *If $\Lambda \geq \overline{\Lambda}$, $p^*_{sw} = R - \sqrt{\frac{cR}{\mu}}$. Naive customers all balk, i.e., $\delta_n = 0$, whereas sophisticated customers join with probability $\delta_s = \frac{\lambda_b}{\Lambda_s}$.*

Proposition 2.2 shows how market conditions affect the market equilibrium under welfare maximization (see Figure 2.1). If the potential market size is either small ($\Lambda \leq \underline{\Lambda}$) or large ($\Lambda \geq \overline{\Lambda}$), the welfare-maximizing provider charges a price that is the same as when all customers are sophisticated; naive customers are served when $\Lambda \leq \underline{\Lambda}$, while only sophisticated customers are served when $\Lambda \geq \overline{\Lambda}$. When the potential market size is in an intermediate range ($\underline{\Lambda} < \Lambda < \overline{\Lambda}$), the customer type composition parameter θ plays a critical role in the provider's admission strategy. If $\theta \geq 0.5$, the provider admits all naive customers into the

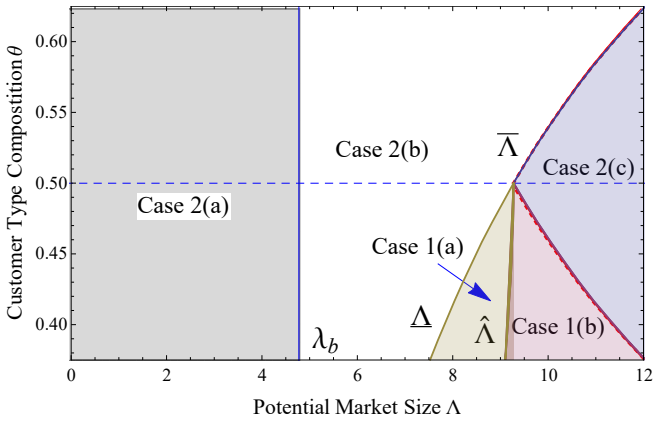
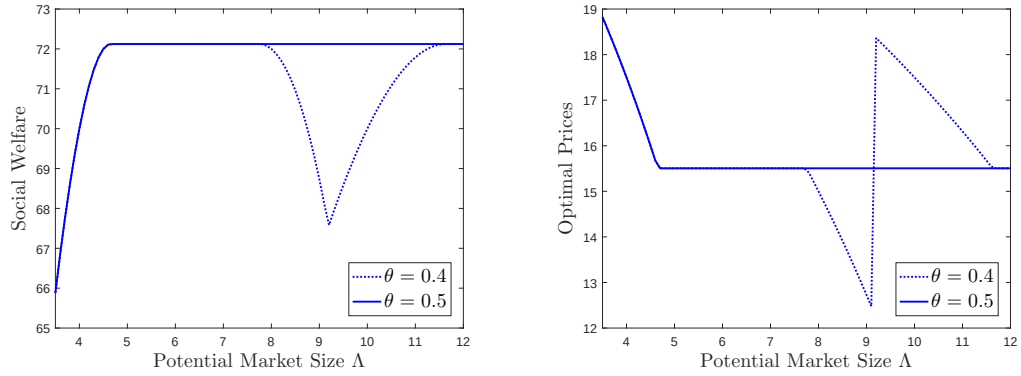


Figure 2.1: Equilibrium Pricing and Customer Joining Behavior under the Welfare Maximization: $R = 40$, $c = 180$, $\mu = 12$

system since the potential market size of naive customers is small, i.e., $\Lambda_n < \lambda_b$. However, when $\theta < 0.5$, the potential market size of naive customers is larger than (whereas that of sophisticated customers is smaller than) what the welfare-maximizing provider desires; that is, $\Lambda_s < \lambda_b < \Lambda_n$. As naive customers join or balk collectively, the welfare-maximizing provider can serve only one type of customers. Specifically, the provider serves only naive customers when $\underline{\Lambda} < \Lambda < \widehat{\Lambda}$ and only sophisticated customers when $\widehat{\Lambda} < \Lambda < \overline{\Lambda}$.

To summarize, Proposition 2.2 implies that the service provider serves all the customers when the potential market size is small ($\Lambda \leq \lambda_b$), and ideally, if possible, serves λ_b customers by taking into consideration the herding behavior of the naive customers.

Figure 2.2 shows how the potential market size Λ and the customer type composition parameter θ affect both social welfare and optimal price. The solid line and the dashed line correspond to the cases $\theta = 0.4$ and $\theta = 0.5$, respectively. Figure 2.2(a) confirms that social welfare increases in the potential market size Λ when Λ is small, and reaches its peak at $\Lambda = \lambda_b$, the socially-desired arrival rate. A further increase of Λ harms social welfare when $\theta < 0.5$ and $\underline{\Lambda} < \Lambda < \overline{\Lambda}$ due to naive customers' herding behavior, which makes the effective arrival rate $\lambda_b = 4.65$ unachievable. Figure 2.2(b) shows that the optimal price decreases in Λ when it is smaller than what is socially desired ($\Lambda \leq \lambda_b = 4.65$), and remains



(a) Social Welfare (b) The Optimal Price

Figure 2.2: Social Welfare and Optimal Price under the Welfare Maximization: $R = 40$, $c = 180$, $\mu = 12$

unchanged when the potential market size Λ is beyond the socially desired λ_b as long as λ_b is achievable. Note that when $\theta = 0.4$ and $\underline{\Lambda} = 7.7 < \Lambda < 11.6 = \overline{\Lambda}$, $\lambda_b = 4.65$ cannot be achieved. Figure 2.2(b) shows that in this range, the optimal price first decreases, then jumps up at $\widehat{\Lambda} = 9.2$ and then decreases again as the market size Λ increases. This is due to the switch from serving naive customers only to serving sophisticated customers only.

2.4.2 Profit Maximization

The optimization problem of the profit-maximizing provider is very similar to that of the welfare-maximizing provider, except that $v(p_i)$'s are replaced by p_i's in the objective function. Thus, his optimization problem is given as follows:

$$\max_{\boldsymbol{p}} \Pi = \sum_{p_i \in \boldsymbol{p}} p_i \lambda(p_i) L_{p_i}$$
$$\text{s.t.} \quad \sum_{i=1}^{N} L_{p_i} = 1, L_{p_i} \geq 0,$$

where $\lambda(p_i)$ is also obtained from Proposition 2.1. We obtain the following result:

Proposition 2.3 *The profit-maximizing pricing strategy satisfies the following properties.*

1. *There exists a threshold $\widetilde{\Lambda}$ on the potential market size above which the cyclic pricing strategy is preferred by the service provider, while below which the*

static pricing strategy makes the service provider better off and he behaves exactly like that under welfare maximization.

2. *The optimal cyclic pricing strategy is high-low cyclic, which can be simply denoted as* $\boldsymbol{p} = \{(p_h, L), (p_l, 1-L)\}$ *where* p_h *and* p_l *represent the high and low price, respectively, and* L *represents the proportion of time the high price remains. Under the optimal cyclic pricing strategy, sophisticated customers join during the high-price phase and naive ones join during the low-price phase.*

3. *The optimal low price satisfies* $p_l^* = \eta(\boldsymbol{p})$*, and the other two decision variables,* (p_h^*, L^*)*, are given as follows.*

   I. *If* $\theta < \frac{1}{2}$ *and* $\widetilde{\Lambda} < \Lambda < \ddot{\Lambda}$ *(*$\ddot{\Lambda}$ *is given by (A.35)),* $p_h^* = V_s$ *and* $L^* = L_b$*, where* L_b *solves (A.34). Here,* $\delta_s = \delta_n = 1$*.*

   II. *If* $\theta < \frac{1}{2}$ *and* $\Lambda \geq \ddot{\Lambda}$*, or if* $\theta \geq \frac{1}{2}$ *and* $\Lambda > \widetilde{\Lambda}$*,* $p_h^* = p_h^0 > V_s$ *and* $L^* = L^0$*, where* (p_h^0, L^0) *solves (A.28) and (A.36) simultaneously. Moreover,* $\delta_s < 1$ *and* $\delta_n = 1$*.*

   The detailed expressions of (A.28), (A.34), (A.35) and (A.36) can be found in the online Appendix. Moreover, $p_h^* > R - \sqrt{\frac{cR}{\mu}}$*.*

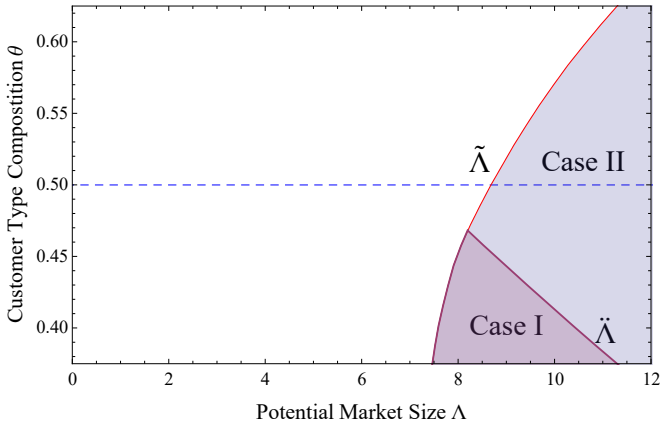


Figure 2.3: Equilibrium Pricing and Customer Joining Behavior under the Profit Maximization: $R = 40$, $c = 180$, $\mu = 12$

Proposition 2.3 shows that only when the potential market size Λ is above a certain value will the cyclic pricing strategy be preferred by the service provider

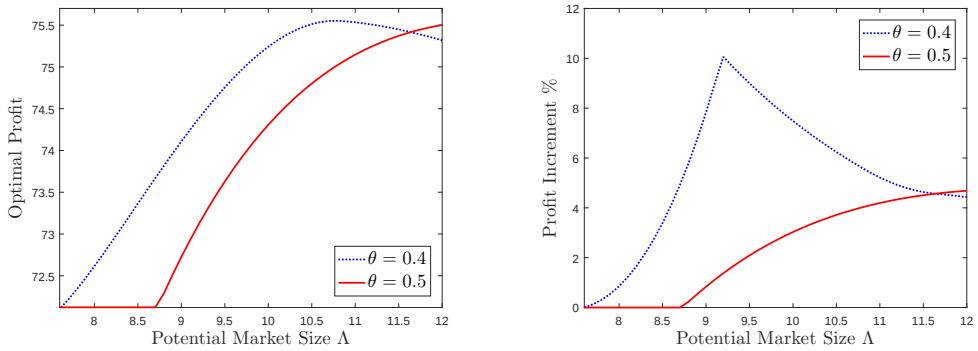
over the static pricing strategy (see Figure 2.3). As illustrated in Figure 2.3, at the left side of the red line $\Lambda = \widetilde{\Lambda}$, the profit-maximizing provider adopts exactly the same static pricing strategy as the welfare-maximizing provider does (see Figure 2.1); while at the right side of the red line, the cyclic pricing strategy is adopted. Under the optimal cyclic pricing, sophisticated customers join only within the high-price phase and they obtain a net utility of zero; naive customers join only within the low-price phase and their expected consumption utility is V_n. Since the optimal low price p_l^* equals the long-run average rating and is higher than V_n, naive customers' net utility, $V_n - p_l^*$, is always negative. Therefore, although naive customers are offered a lower price, the price is not low enough to guarantee a non-negative utility. This overcharging is possible due to higher ratings generated from sophisticated customers, which boost up the average rating. Figure 2.3 also shows that $\widetilde{\Lambda}$ increases with θ, the proportion of sophisticated customers in the market. That is, when the market comprises a larger proportion of sophisticated customers, static pricing is favored.

To thoroughly understand the mechanism underlying the profit gain, we now turn to the comparison between the static and cyclic pricing strategies. Under a static pricing strategy, all customer surplus is internalized through pricing and goes towards the service provider's profit; thus, maximizing profit is the same as maximizing welfare. The classic pricing literature in queueing systems has shown that the objective function of a profit-maximizing service provider is concave in the effective arrival rate, and that λ_b is the best demand level that a monopoly provider should maintain whenever it is achievable. According to Proposition 2.2, the optimal arrival rate λ_b can be achieved under the static pricing strategy when more than half of the customers are sophisticated ($\theta \geq 0.5$). One might believe that the average arrival rate under the cyclic pricing strategy should be equal to this one and, therefore, if the demand rate during the high-price phase is smaller than λ_b, the one in the low-price phase should be larger than λ_b. Does this belief hold water? The following corollary provides a different answer.

Corollary 2.1 *When $\theta \geq 0.5$ and $\widetilde{\Lambda} < \Lambda < \overline{\Lambda}$, the effective arrival rates under*

the optimal cyclic prices are always less than that under the static one, specifically, $\lambda(p_h^*) < \lambda(p_l^*) < \lambda(p_{sw}^*)$.

Corollary 2.1 shows that, under the cyclic pricing strategy, a service provider may deliberately serve a number of customers less than λ_b during both low- and high-price phases and achieve a higher profit. Why not lower both high and low prices to attract more customers? Why does the original price and demand tradeoff not work here? A closer investigation of the cyclic pricing strategy reveals that the server cannot do that. During the low-price phase, naive customers all join and obtain a negative expected utility. Sophisticated customers know that and hence, they will never join then. Therefore, demand cannot be increased by lowering the price unless naive customers can obtain a non-negative expected utility. During the high-price phase, demand can be increased by lowering the price; however, this results in more congestion, which will reduce the ratings on consumption utility and jeopardize the strategy of ripping off naive customers during the low-price phase. Consequently, due to a higher profit margin from ripping off naive customers, the original price-demand tradeoff is twisted, and fewer customers are served in both phases.

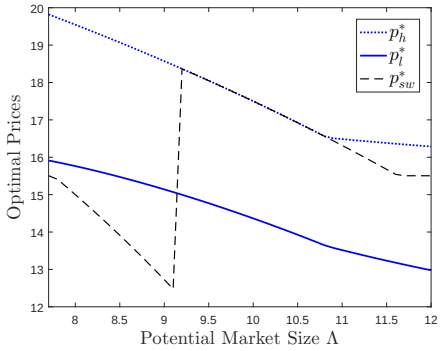


(a) Optimal Profit (b) Profit Increment

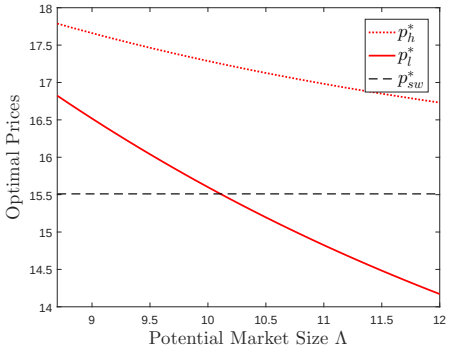
Figure 2.4: Profit Performance between Cyclic Pricing and Static Pricing: $R = 40$, $c = 180$, $\mu = 12$, $\widetilde{\Lambda} = 7.62$ when $\theta = 0.4$ and $\widetilde{\Lambda} = 8.66$ when $\theta = 0.5$

Next, we numerically examine the benefit brought about by the cyclic pricing strategy. Let the customer type composition parameter θ change from 0.3875 to 0.6125 with a step length 0.0125. For each given θ, we vary the potential market

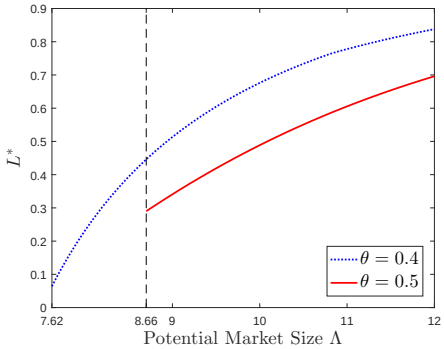
size Λ to find the highest profit increment brought about by the cyclic pricing strategy over the static one. The aggregated numerical results show that the highest increment amount is around 5.19% on average but can reach as high as 11.22%. See Figure 2.4 for an illustration of the results when $\theta = 0.4$ and $\theta = 0.5$, respectively.



(a) Optimal Prices: $\theta = 0.4$



(b) Optimal Prices: $\theta = 0.5$



(c) Proportion of Time Staying at the High Price (L)

Figure 2.5: Optimal Pricing Strategy under Profit Maximization: $R = 40$, $c = 180$, $\mu = 12$, $\widetilde{\Lambda} = 7.62$ when $\theta = 0.4$ and $\widetilde{\Lambda} = 8.66$ when $\theta = 0.5$

We also numerically investigate how market size affects the optimal cyclic pricing strategy; see Figure 2.5. Figures 2.5(a) and 2.5(b) depict the optimal high and low prices when the customer type composition parameter θ equals 0.4 and 0.5, respectively. They show that for a given customer type composition parameter, both optimal prices decrease in the potential market size Λ. For the sake of comparison, we also plot the optimal static price (that is, the welfare-maximizing price) with a black dashed line in both Figures 2.5(a) and 2.5(b). We

find that under both cases, when the market size is not large, both high and low prices adopted in the cyclic pricing strategy are larger than the static one. This observation implies that the conclusion in Corollary 2.1 might be generally true for a small size market and not necessarily just restricted to the case of $\theta \geq 0.5$, but we are unable to prove this for the general case. Figure 2.5(c) shows that L^*, the proportion of time that the high price shall be charged, increases in the potential market size Λ. That is, as Λ increases, the service provider charges the high price for a longer time.

2.5 Comparison between Profit Maximization and Welfare Maximization

So far, we have derived the service provider's optimal pricing strategies under both welfare and profit maximization. Recall that the profit-maximizing provider behaves exactly as a welfare-maximizing provider when the potential market size Λ is below the threshold $\widetilde{\Lambda}$ (see Proposition 2.3). Thus, the system performances under these two objectives are the same when $\Lambda \leq \widetilde{\Lambda}$. Hereafter, we focus on comparing the system performance under these two objectives when $\Lambda > \widetilde{\Lambda}$, that is, when the service provider adopts static pricing under welfare maximization but cyclic pricing under profit maximization. We examine how the cyclic pricing strategy affects the system's welfare. We define *welfare loss* of the system under profit maximization as

$$\frac{\mathcal{SW}^* - \mathcal{SW}_{pm}}{\mathcal{SW}^*},$$

where $\mathcal{SW}_{pm}$ denotes the social welfare under the optimal cyclic pricing strategy. $\mathcal{SW}_{pm}$ is given as follows:

$$\mathcal{SW}_{pm} = v(p_h^*)\lambda(p_h^*)L^* + V_n\Lambda_n(1-L^*) = p_h^*\lambda(p_h^*)L^* + V_n\Lambda_n(1-L^*),$$

where p_h^* and L^* are the optimal high price and the corresponding proportion of time the high price remains under the cyclic pricing strategy, respectively.

Figure 2.6 illustrates the impact of market condition on welfare loss when the customer type composition parameter $\theta = 0.4, 0.5$. An interesting observation

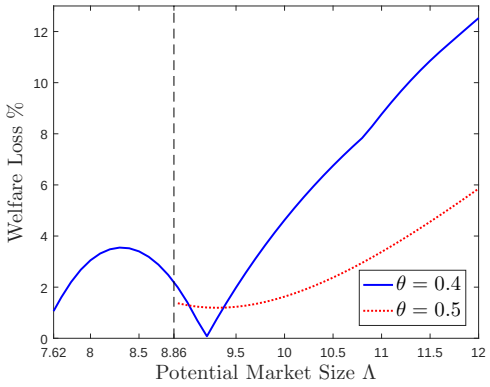


Figure 2.6: Optimal Cyclic Pricing Strategy's Social Efficiency: $R = 40$, $c = 180$, $\mu = 12$, $\widetilde{\Lambda} = 7.62$ when $\theta = 0.4$ and $\widetilde{\Lambda} = 8.66$ when $\theta = 0.5$

is that welfare loss is not monotone in the market size. For the case of $\theta = 0.4$, the optimal pricing strategy is cyclic only when the potential market size $\Lambda > \widetilde{\Lambda} = 7.62$. Starting from $\Lambda = 7.62$, the welfare loss first increases, then decreases and finally increases again in Λ. Specifically, when Λ falls into the range between $\widetilde{\Lambda} = 7.62$ and $\widehat{\Lambda} = 9.19$, the welfare loss first increases, then decreases and reaches 0 at $\Lambda = \widehat{\Lambda} = 9.19$. In this range, the welfare maximization requires that only naive customers be served, and the socially optimal price is $p^*_{sw} = V_n$. Thus, the maximal social welfare is $\mathcal{SW}^* = V_n\Lambda_n$. However, under the cyclic pricing strategy, only sophisticated customers are served at the high price $p^*_h = V_s$ (see Case I of Proposition 2.3), and all naive customers are served at the low price. Therefore, the difference in welfare only occurs during the high-price phase. In short, when $\Lambda \in [7.62, 9.19)$, the welfare loss is caused by underutilization of the system during the high-price phase, due to the fact that the number of sophisticated customers is less than that of naive customers. Such welfare loss can be expressed as

$$\frac{\mathcal{SW}^* - \mathcal{SW}_{pm}}{\mathcal{SW}^*} = \frac{(V_n\Lambda_n - V_s\Lambda_s)}{V_n\Lambda_n} \cdot L^*. \tag{2.9}$$

We can see that welfare loss is the product of two terms, $\frac{V_n\Lambda_n - V_s\Lambda_s}{V_n\Lambda_n}$ and L^*. We now consider the impact of marker size on these two terms. One can check that

$$\frac{V_n\Lambda_n - V_s\Lambda_s}{V_n\Lambda_n} = 1 - \frac{\theta}{1-\theta} \cdot \frac{V_s}{V_n}$$

is decreasing in Λ. As the market size increases, the number of sophisticated customers increases proportionally and the underutilization effect is reduced. Meanwhile, L^* is increasing in Λ, because the proportion of high-price phase must be lengthened to balance the negative reviews from an increased number of naive customers. When Λ is very small, the second effect (the increasing monotonicity of L^*) dominates, and as Λ further increases, the first effect (the diminishing effect of underutilization) dominates. These two effects jointly drive the product term in (2.9) to first increase and then decrease in Λ. Note that when $\Lambda = \widehat{\Lambda} = 9.19$, $V_n\Lambda_n = V_s\Lambda_s$. Hence, the welfare loss stated in (2.9) becomes 0. For the range $\Lambda \geq \widehat{\Lambda} = 9.19$, welfare maximization requires that only sophisticated customers should be served, whereas cyclic pricing admits all naive customers during the low-price phase. This indicates that when $\Lambda > 9.19$, the welfare loss is mainly caused by over-utilization of the system during the low-price phase. As the potential market size increases, welfare loss becomes larger due to the increased over-crowdedness of the system.

When $\theta = 0.5$, the cyclic pricing strategy is adopted by the profit-maximizing provider when the potential market size $\Lambda \geq \widetilde{\Lambda} = 8.66$. Figure 2.6 shows that when $\Lambda \geq 8.66$, welfare loss first decreases and then increases, reaching the minimum at $\Lambda = \overline{\Lambda} = 9.30$. This observation can be explained as follows. According to Proposition 2.2, when $\theta \geq 0.5$, the socially optimal price is $p^*_{sw} = p(\lambda_b)$, where λ_b is the socially desirable effective arrival rate defined in (2.7). Also, according to Case II of proposition 2.3, the optimal high price satisfies $p^*_h > p(\lambda_b)$. Hence, under profit maximization, compared with what is socially desirable, the system is always under-utilized during the high-price phase. In addition, the provider always serves all the naive customers during the low-price phase. When Λ is small such that $\Lambda_n < \lambda_b$, the system is also under-utilized during the low-price phase. This observation is also consistent with the conclusion in Corollary 2.1: cyclic pricing could cause underutilization in both phases. It contradicts our conventional belief that welfare loss caused by cyclic pricing is due to the uneven workloads in the two phases– the underutilization in the high-price phase and

over-utilization in the low-price phase. Thus, when $\Lambda_n < \lambda_b$, increasing Λ reduces the effect of underutilization and thus, mitigates welfare loss. Welfare loss reaches its minimum when $\Lambda_n = \lambda_b$, i.e., $\Lambda = \overline{\Lambda} = 9.30$. When $\Lambda_n > \lambda_b$, the system is over-utilized during the low-price phase. In this range, increasing Λ further increases the level of crowdedness and thus, enlarges the welfare loss.

Recall that welfare is the sum of server's profit and all customers' utilities. The price becomes internal transfer between the two parties and hence does not appear on the welfare expression. Only the amount of served customers and each customer's consumption utility matter in determining the social welfare. Tradeoff between the amount of served customers and consumption utility per customer determines an optimal congestion level. The foregoing analysis shows that welfare loss is mainly caused by the uneven workloads during two phases — over-congestion during the low-price phase and underutilization during the high-price phase. However, in a situation where the market size is not large and sophisticated customers comprise not less than half of the population, cyclic pricing can cause underutilization in both phases, also resulting in welfare loss. In the latter case, two few customers are served, which is also not socially desired.

2.6 Extension: Ratings on Net Utility

In this section, we consider another information scenario where customers rate on the net utility $R - cw - p$ rather than the consumption utility $R - cw$; in other words, they take price into consideration in their rating. If only such rating information is available, incoming customers cannot differentiate whether a low score is caused by high congestion or a high price, and hence naive customers cannot make a correct decision on joining; our cyclic pricing strategy does not work in this pure rating over net utility case. In practice, websites, such as dianping.com, allow a customer to have ratings on *multiple classified items* such as experience, dining and cost-effectiveness, which make it possible for incoming customers to infer the consumption utility from such classified ratings. Below we examine such a case with both the rating over net utility and the average price

revealed to customers.

If customers' utility function is linear, incoming customers can infer the expected consumption utility by simply adding the rating score on the expected net utility with the average price. To see this equivalence, let $\eta'(\boldsymbol{p})$ be the average rating on the net utility. Then,

$$\begin{aligned}\eta'(\boldsymbol{p}) &= \frac{\sum_{i=1}^{i=N}(v(p_i)-p_i)\lambda(p_i)L_{p_i}}{\sum_{i=1}^{i=N}\lambda(p_i)L_{p_i}} \\ &= \frac{\sum_{i=1}^{i=N}v(p_i)\lambda(p_i)L_{p_i}}{\sum_{i=1}^{i=N}\lambda(p_i)L_{p_i}} - \frac{\sum_{i=1}^{i=N}p_i\lambda(p_i)L_{p_i}}{\sum_{i=1}^{i=N}\lambda(p_i)L_{p_i}} \\ &= \eta(\boldsymbol{p})-\bar{p}.\end{aligned}$$

Consequently, incoming customers can still obtain the average rating on the consumption utility by adding the average rating on the net utility $\eta'(\boldsymbol{p})$ with the average price $\bar{p}$, and all the rationales with the scenario of rating over the consumption utility hold here.

One rationale for the cyclic pricing to generate a higher profit is that, even though naive customers are unsatisfied with service and post low scores, the overall ratings can still be uplifted by high ratings posted by sophisticated customers. Clearly, this mechanism requires an underlying assumption that these naive customers are one-time customers, not returned customers. Another situation which could dampen the effect of cyclic pricing is that customers take the average rating as their reference point when joining and compare their experience with this reference point: if the experienced quality is higher than the rating, they feel a gain and otherwise they feel a loss. Furthermore, behavior literature often assume that customers are loss averse, i.e., losses are more painful than equal-sized gains being pleasant; see, e.g., Baron et al. (2015). We now extend our study to such loss-averse customers, with the overall utility of a customer can be expressed as the sum of her original net utility and the gain-loss utility.

Specifically, for two components of the customers utility function, namely the net utility term measured by $R-cw$ and monetary term measured by the price p, customers obtain a gain-loss utility along both dimensions by comparing their actual outcomes with the reference points on the two dimensions. A customer

determines her joining decision based on the information she has access to upon observing that the current price is p_i ($i = h,\ l$ for the cyclic pricing strategy, and the subscript shall be left out for the static pricing strategy). Denote the joining probability as δ_s and δ_n, for sophisticated and naive customers, respectively.

The gain-loss utility function with loss-averse customers can be formulated as follows, following Baron et al. (2015). Denote $\boldsymbol{k} = (k^v,\ k^p)$ as the two dimensional outcomes with k^v measures the consumption utility and k^p measures the price. The actual consumption outcome can be either $(V(p_i), -p_i)$ when a customer joins or $(0, 0)$ when the customer balks. Denote $\boldsymbol{r} = (r^v,\ r^p) \in \{(\widehat{V}(p_i), -p_i),\ (0, 0)\}$ as their reference point. Due to different information accesses, the reference points for sophisticated and naive customers are different; that is, $\widehat{V}(p_i)$ is different for the two types of customers, which will be discussed shortly. The gain-loss utility for an outcome $\boldsymbol{k}$ by comparing with a reference point $\boldsymbol{r}$ is given as

$$v_{gl}(\boldsymbol{k}|\boldsymbol{r}) = (k^v - r^v)^+ + \alpha(k^v - r^v)^- + (k^p - r^p)^+ + \alpha(k^p - r^p)^-, \qquad (2.10)$$

where $x^+ = \max\{x, 0\}$ and $x^- = \min\{x, 0\}$. The parameter $\alpha > 1$ measures the degree of loss-averse behavior. The range $\alpha > 1$ implies that the customer feels losses being more painful than the equal-sized gain being pleasant. The term $(k^v - r^v)^+$ measures the gain due to the consumption utility being larger than the reference point on that; the term $\alpha(k^v - r^v)^-$ measures the loss due to the consumption utility being less than the reference point on that. Similar interpretations hold for the two terms on the price dimension: $(k^p - r^p)^+$ and $\alpha(k^p - r^p)^-$. The overall utility of a consumption outcome $\boldsymbol{k}$ conditional on a reference point $\boldsymbol{r}$ is the sum of actual net utility and the gain-loss utility, i.e.,

$$u(\boldsymbol{k}|\boldsymbol{r}) = k^v + k^p + v_{gl}(\boldsymbol{k}|\boldsymbol{r}).$$

Clearly, this utility function with loss-averse customers is no longer linear in the price and consumption utility. It is kinked at the reference point. Furthermore, as the reference points are endogenous, the overall function form is not even piecewise linear.

For sophisticated customers, they adopt a mixed strategy to join or to balk;

that is, they choose to join in probability δ_s and balk in probability $1-\delta_s$. Since they know service related information R and μ, their reference point on the consumption utility is fully endogenized as their rational expectations of the outcomes. That is, sophisticated customers reference point on the consumption utility, $\widehat{V}(p_i)$, is stochastically equivalent to the actual outcome $V(p_i)$, and they follow the same distribution. The expected utilities of a sophisticated customer, to choose to join and to balk, are respectively given as follows:

$$\begin{aligned} U(\text{join}) &= \delta_s E[u((V(p_i),-p_i)|(\widehat{V}(p_i),-p_i))] + (1-\delta_s)E[u((V(p_i),-p_i)|(0,0))], \\ U(\text{balk}) &= \delta_s E[u((0,0)|(\widehat{V}(p_i),-p_i))] + (1-\delta_s)E[u((0,0)|(0,0))]. \end{aligned}$$

The detailed expressions for $U(\text{join})$ and $U(\text{balk})$ can be derived following Yang et al. (2018) and the equilibrium joining probability can be obtained by solving equation $U(\text{join}) = U(\text{balk})$. For example, note that $V(p_i) = R - cw$ and w is drawn from an exponential waiting time distribution with rate $\mu - \lambda(p_i)$, the expectation

$$E[u((V(p_i),-p_i)|(\widehat{V}(p_i),-p_i))] = R - \frac{c}{\mu-\lambda(p_i)}\frac{\alpha+1}{2} - p_i.$$

For naive customers, as they do not have information on system performances, still they sum over the average rating over the overall utility with the average price to obtain a gain-loss based consumption utility, and they compare this rating score with the current price to determine whether or not to join. After the service, they also post a rating on their overall feeling, i.e., the sum of their net utility and the gain-loss utility. As naive customers have no other information besides ratings when they arrive, we can naturally assume that their reference points are such rating score, i.e., $\widehat{V}(p_i) = \eta'(\boldsymbol{p})$ and $r^p = \bar{p}$. Therefore, a naive customer's *ex post* utility is given as $u((R-cw,-p_i)|(\eta'(\boldsymbol{p}),-\bar{p}))$. As different customers have different ex-post utility, the average one posted by them is the expectation of their ex-post utility, i.e.,

$$\begin{aligned} &E[u((R-cw,-p_i)|(\eta'(\boldsymbol{p}),-\bar{p}))] \\ &= (1+\alpha)\left(R - \frac{c}{\mu-\lambda(p_i)}\right) - 2p_i + \bar{p} - \alpha\eta'(\boldsymbol{p}) \end{aligned}$$

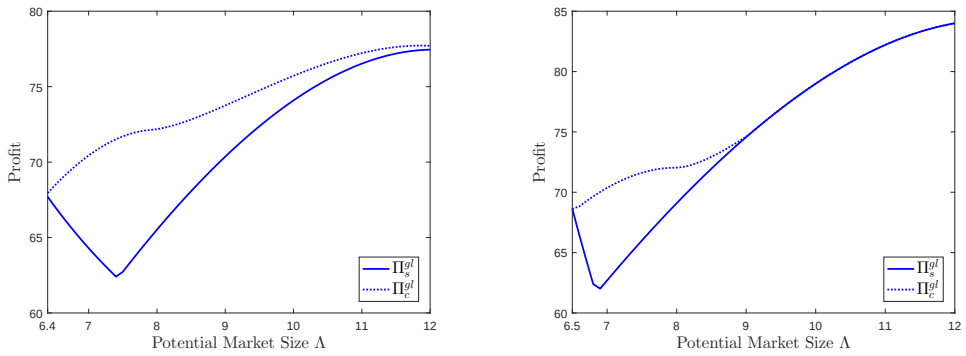
$$+(\alpha-1)\frac{c}{\mu-\lambda(p_i)}e^{-(\mu-\lambda(p_i))\frac{R-\eta'(\boldsymbol{p})}{c}}$$

After analyzing the joining decisions and the ratings of the two types of customers, we move on to numerical studies of the optimal decision of both cyclic and static pricing strategy of the service. Figure 2.7 illustrates the performance of the cyclic pricing strategy in comparison to the static pricing strategy given that $\theta = 0.4$. Figure 2.7(a) shows the case with $\alpha = 1.2$. We observe that the cyclic pricing strategy is more profitable than the static pricing strategy if the potential market size satisfies $\Lambda \geq 6.4$. When $\alpha = 1.5$, which is shown in Figure 2.7(b), the preferred interval is shrank into $(6.50,\ 9.00)$. As α increases to 1.8 and 2.0, the preferred range is further shrank into $(6.80, 8.00)$ and $(7.00, 7.80)$, respectively; see Figures 2.7(c) & 2.7(d), respectively. Clearly, as α increases, the size of the preferred range is reduced and hence cyclic pricing strategy is less likely to be chosen by the provider. We observe similar phenomena by changing parameter θ.

Why is the cyclic pricing strategy less preferred when customers are more loss averse (α is larger)? When customers become more loss averse, naive customers suffer from a larger loss if they find that their actual experienced utility is below the rating and hence will likely post a very low score. The average rating will likely be lowered down, making the cyclic pricing strategy less attractive for the server. Therefore, when a server considers to adopt cyclic pricing strategy, he shall take customers' behavior into consideration: the strategy does not work if customers are very loss averse.

2.7 Conclusion

In this chapter, we consider a typical service situation where customers are heterogeneous in information accesses: some customers know the service-related information, whereas others do not; the latter relies on buyer-generated information to make their queueing decisions. We demonstrate that a cyclic pricing strategy can be used to improve the profitability of a service provider without distorting customers' ratings. Under the optimal high-low cyclic pricing strategy, sophisti-



(a) $\alpha = 1.2$ (b) $\alpha = 1.5$

(c) $\alpha = 1.8$ (d) $\alpha = 2.0$

Figure 2.7: The Profit of the Cyclic Pricing Strategy When Customers are Loss-Averse: $R = 40$, $c = 180$, $\mu = 12$, $\theta = 0.4$

cated customers join at the high price and naive customers join at the low price. During the high-price phase, the system is less congested, and the ratings are relatively high, which boosts the average rating and allows the provider to charge a price higher than naive customers' expected consumption utility during the low-price phase. The interesting part is that, even though naive customers feel unsatisfied and post low scores after consumption, the average rating can still be maintained by getting high rating scores in the high-price phase, which allows the server to obtain a higher profit than that under a static pricing strategy.

The validity of this strategy requires the potential market size to be above a certain threshold value such that congestion is a significant factor in affecting customers' joining decision. We further extend the analysis to the non-linear utility case. Specifically, we consider that naive customers treat historical rating as their reference point and compare their actual experienced utility with this benchmark value. They feel a loss if the experienced utility is below the rating and a gain if otherwise; furthermore, the loss is more painful than an equal-size gain being pleasant. We find that such reference-dependent and loss-aversion behavior dilute the attractiveness of cyclic pricing strategy; the more loss averse customers are, the less attractive the cyclic pricing is.

We also show that although the cyclic pricing strategy can improve profitability, it harms social welfare. Welfare maximization prefers even workloads across periods, and hence prefers the static pricing strategy. However, the system can be either over- or under-utilized under the cyclic pricing strategy, leading to welfare loss.

Our study's main takeaway is to show that, even without manipulating customer ratings, a service provider can still rip off uninformed customers by implementing a cyclic pricing strategy and advertising the average rating to incoming customers. Nevertheless, the limit of this mechanism heavily depends on the congestion of the system and the degree of customers' loss-averse behavior. From the viewpoint of these naive customers, they shall take in the aggregated rating information with *caution*: when the products are service goods, an average rat-

ing in the past does not mean that they can get the same consumption utility if they join. How to protect naive customers' welfare remains to be an interesting research question.

Chapter 3

Modeling Patients' Illness Perception and Equilibrium Analysis of the Doctor Shopping Behavior

3.1 Introduction

Patients often cannot assess the quality of the health care service, a kind of credence service offered by experts (i.e., doctors) (Pac and Veeraraghavan 2010). As patients lack the knowledge or expertise to judge the doctor's diagnosis or treatment plan, they may actively seek opinions from multiple doctors during a single illness episode without referrals; that is, patients engage in *doctor shopping* (Kasteler et al. 1976). Field studies reveal that doctor shopping is quite common. For example, the prevalence of doctor shopping was nearly 40% in the government out-patient departments in Hong Kong (Lo et al. 1994), 18% in Canada (Macpherson et al. 2001), 23% in Japan (Sato et al. 1995) and 48% among high-income families in the United States (Kasteler et al. 1976).

The reasons for doctor shopping include patient dissatisfaction with or distrust of the doctor (Billinghurst and Whitfield 1993, Guo et al 2002, Harris 2003) and a lesser understanding of doctors' explanations and disbelief of diagnosis and treatment (Sato et al. 1995). The patients' doctor shopping incentive is also affected by the factors related with the health care system such as the waiting time, the charged fees and the reputation of the doctor (Billinghurst and Whitfield

1993, Yeung et al. 2004, Leung et al. 2006, Lo et al. 1994). Patients often hold prior beliefs about their own medical conditions; that is, they have their illness perceptions (Petrie et al. 2007). When a doctor's diagnosis contradicts a patient's belief, a patient may encounter cognitive dissonance; he/she may further seeks second opinion from another doctor (Donkin et al. 2006, Hagihara et al. 2005).

Undoubtedly, in a diagnostic healthcare system, patients' doctor shopping results in repeated consultations and examinations, which increases the doctor workload. Most medical practitioners believe that patients' doctor shopping behavior should be controlled (Katon et al. 1992). Does patients' doctor shopping really hurt the system's performance? This requires us to conduct a thorough analysis of the implications of patients' doctor shopping. Specifically, we aim to analytically examine the patient's doctor shopping behavior and its impact on the diagnostic healthcare system by addressing the following research questions that have not been investigated in the existing literature:

- How does the patients' illness perception affect their doctor shopping decision?
- How does the patients' doctor shopping behavior affect the diagnostic healthcare system's performance in terms of the waiting time, the effective arrival rate and the social welfare?
- Understanding the effects brought by the patients' doctor shopping behavior, what measures can the social planner take to prohibit such shopping (seeking multi-doctors' diagnostic opinions) behavior?

To tackle the above questions, we consider a public diagnostic healthcare system facing a stream of delay-sensitive patients. Patients exhibit the similar symptoms but hold their own illness perceptions. A patient's illness perception measures her *belief* about the likelihood she is sick, which, however, is necessarily an indicator of her physical status (Petrie et al. 2007). The doctor, based on the diagnostic result, decides whether to dismiss a patient or refer her for further treatment. The patient then decides whether to take the doctor's advice or seek

a second opinion from another doctor. In each visit, the patient incurs a fixed copayment charge, which, however, does not fully cover the cost of the public diagnostic service. A patient decides whether to further seek doctors' opinions by adopting the *One-Stage Look-Ahead* rule. We show that whether a patient seeks the service and how many visits a patient pays in one illness episode are jointly determined by her own illness perception, the costs associated with each visit and the doctor's diagnosis quality (that is, the degree of accuracy). We show that whenever two successive diagnostic results are consistent with each other, the patient stops doctor shopping. We show that increasing the per-visit copayment fee can be an effective means to mitigate patients' doctor shopping and thus reduce the system congestion. When the copayment charge is high enough, no patients can afford doctor shopping. Although repetitive diagnoses (due to patients' doctor shopping) help little in improving the objective reward of the diagnostic service, it may patients' psychological gains. A welfare-maximizing social planner, when taking into account patients' psychological gains, shall tolerate a certain level of doctor shopping.

The remainder of this chapter is organized as follows. Section 3.2 reviews the related literature. Model setup is discussed in Section 3.3. In Section 3.4, we analyze the patients's doctor shopping behavior by taking into account the cost associated with their visits. In Section 3.5, we investigate the impact of patient doctor shopping on the system performance. Concluding remarks are provided in Section 3.6. All the proofs are relegated to the appendix.

3.2 Literature Review

This study is related to the stream of the research that investigates the accuracy-congestion trade-off in diagnostic systems (Hasija et al. 2005, Wang et al. 2010, de Vericourt and Sun 2009, Alizamir et al. 2013). Hasija et al. (2005) examine the optimal staffing levels and referral rates. They utilize a gatekeeper model that incorporates staffing, customer waiting times and mistreatment costs so as to minimize total costs. Wang et al. (2010) study patient behaviors in a call

center composed of triage nurses, where service is provided via telephone to help patients choose the appropriate care based on their symptoms. They find that the diagnostic accuracy affects the effective arrival to the call center and increasing capacity may actually increase congestion. de Vericourt and Sun (2009) propose several cognitive heuristics adapted to congestion. They find that simple fixed threshold rules appears to be very robust and judgments based only on the most relevant piece of information performs reasonably well. Alizamir et al. (2013) consider a diagnostic process consisting of sequential tests. They study how the service provider's belief about a customer being positive affects its service provision. In this chapter, we also consider diagnostic service. Different from the aforementioned studies, we investigate patients' service seeking decision in which patients actively decide whether she should continue to seek more diagnosis or not. Patients updates their beliefs of being positive upon observing the doctor's diagnosis.

Pac and Veeraraghavan (2010) consider a kind of diagnostic service where the expert identifies the problem and prescribes the proper service to customers. The expert may cheat the customer by prescribing a service more (over-provision) or less (rationing) than she needs. They find that expert cheating is mitigated when the system is constrained by capacity and congestion. Similar to this research, we consider a diagnostic service which customers have no/little knowledge about. However, we do not consider the moral hazard issue. Instead, we consider customers' doctor shopping behavior: customers may seek multiple episodes of service due to their limited knowledge.

Our study is also related to the studies that consider the readmission reduction in the service provision systems (de Vericourt and Zhou 2005, Chan et al. 2011, 2014, Yom-Tov and Mandelbaum 2014, Guo et al. 2016). de Vericourt and Zhou (2005) consider the routing issue in a call center with service failure. The call resolution probability and the service time are service-provider dependent. They investigate the optimal routing policy to minimize the average total time of call resolution. Chan et al. (2011) study the impact of different discharge strate-

gies on the total readmission load under uncertainty in a capacity-constrained intensive care unit via empirical data. They also show that their index policy for discharge is optimal in certain regimes. Chan et al. (2014) consider a state-dependent queuing network where the provider may speed up service in order to temporarily alleviate congestion. They identify scenarios where speedup should not be used. Yom-Tov and Mandelbaum (2014) analyze a time-varying Erlang-R queue that accommodates reentrant customers. Guo et al. (2016) consider the readmission problem in a public healthcare system. They examine how the two payment schemes, fee-for-service and bundled payment, affect the healthcare provider service time decision and thus the readmission rate.

Patients'doctor shopping under our study is similar in spirit to readmission. Both doctor shopping and readmission increase the workload of the service system and exacerbate congestion. However, their fundamental driving forces are different. In our model, a patient doctor-shops (i.e., reenters) the diagnostic service system due to her anxiety from dealing with conflicting information between her belief and the doctor's diagnosis, rather than due to the treatment failure (Guo et al. (2016), Chan et al. (2014), Yom-Tov and Mandelbaum (2014)) or the premature discharge (Chan et al. (2011)). We model the dynamics of how patients updates their beliefs, and explain the inner incentives of the patients engaging in doctor shopping.

Other related studies include Guo et al. (2014) and Qian et al. (2017). Guo et al. (2014) investigate the Downs-Thomson paradox in the healthcare systems, wherein improvements in the public facilities may not reduce congestion. Qian et al. (2017) suggest that the public sector can cooperate with the private sector through some subsidy schemes; the cooperation can help avoid the Downs-Thomson paradox and offer quick remedies to the excessive waiting without requiring large investments.

3.3 Model Setup

Consider a *public* diagnostic service system (he) that faces a stream of delay-sensitive customers (she). A public healthcare system usually consists of multiple stations (i.e., hospitals and clinics) and each station consists of multiple servers (i.e., doctors). These stations and servers are often under the supervision of a common government department. For example, in Hong Kong, the Hospital Authority manages all the public hospitals and clinics, and offers over 80% inpatient service as of 2016 [3.1]. The Hospital Authority allocates resources under "same service, same funding" principle [3.2]. In China, all hospitals, regardless of government-owned or privately-owned, are managed under a 3-grade 10-level system by the National Health and Family Planning Commission; hospitals rated at the same level are generally considered similar in their service quality. In our model, we consider a public diagnostic service system that consists of the facilities rated at the same level. Thus, hereafter, we use a representative service provider/doctor to represent the system.

The doctor provides diagnostic service which determines the patients' need and offers medical advice rather than treatments. The service times are independent and identically distributed exponential random variables with rate μ. A stream of patients exhibiting the similar symptoms arrive to seek the diagnostic service according to a Poisson process with an exogenous rate Λ. Λ represents the potential demand rate of the patients. The doctor, based on the diagnostic results, decides whether to dismiss the patient or recommend the patient to seek further specific treatments. The patients are served based on the First-Come First-Served rule.

The patient decides whether to take the doctor's advise or to seek a second opinion from another doctor by comparing the reward of stopping at the current state with the expectation that can be achieved after paying another visit. Given that a patient's true type is i ($i = 0,\ 1$), she gains a value V_i ($V_i > 0$) if correctly

[3.1] Please refer to http://www.ha.org.hk/haho/ho/stat/HASR15_16.pdf, Section 1

[3.2] http://www.legco.gov.hk/yr13-14/english/panels/hs/papers/hs0120cb2-671-5-e.pdf

identified, and suffers a loss L_i if misidentified. In other words, each patient's decision is an optimal stopping problem, where a decision to terminate the visiting process involves two scenarios: leaving reassured as healthy and leaving to seek cure. Therefore, the optimal stopping set takes forms of

$$S = \{\alpha | 0 < \alpha \leq \underline{\alpha},\ \overline{\alpha} \leq \alpha < 1\}. \tag{3.1}$$

We derive $\underline{\alpha}$ and $\overline{\alpha}$ shortly. We interpret the stopping set twofold: first, if a patient decides to join, she terminates the visiting process and follows the doctor's advice once her illness perception falls into the stopping set. Second, an individual whose initial illness perception is in the stopping set will not join the diagnostic system. For example, those whose illness perception fall below $\underline{\alpha}$ usually do not feel the need to see a doctor, while those whose illness perception lie above $\overline{\alpha}$ will be too assertive about being positive; they need neurologies more often than not.

3.3.1 Diagnostic Quality and Accuracy

Consider that each patient is one of the following two types (denoted by t): positive/ill ($t = 1$) or negative/healthy ($t = 0$). The diagnosis, however, is not perfect; it may produce false outcomes. Let s be the doctor's diagnostic result (signaling type s), where $s = 1$ indicates a positive diagnosis outcome while $s = 0$ indicates a negative outcome. Depending on the patient's true type t, the diagnostic quality can be expressed as

$$Q = \begin{bmatrix} q(s=0|t=0), & q(s=1|t=0) \\ q(s=0|t=1), & q(s=1|t=1) \end{bmatrix}.$$

where $q(s = j|t = i)$ denotes the probability that a patient of type i is identified as type j, $i, j \in \{0, 1\}$. For simplicity, let $q(s = j|t = i) := q_{ij}$. Note that when $i = j$, the diagnostic result is correct. However, when $i \neq j$, the diagnosis is false; the doctor either misidentifies a healthy patient ($t = 0$) to be sick ($s = 1$) or a sick patient ($t = 1$) to be healthy ($s = 0$).

We assume that $0 < q_{10} \leq q_{01} < \frac{1}{2}$, or equivalently, $\frac{1}{2} < q_{00} \leq q_{11} < 1$. This assumption indicates that a false positive diagnosis is equally or more likely to occur than a false negative diagnosis, which has been verified by the studies

in several fields and adopted in the operations management literature; see, e.g., Alizamir et al. (2013). This assumption also ensures that the likelihood that the diagnosis is correct is greater than 50%. We assume that the diagnostic process, even though not perfect, is reliable enough for doctors to base their recommendations on; that is, the doctor would recommend a patient further medical treatments if a positive diagnostic outcome is observed but dismiss the patient if the diagnosis outcome is negative.

Moreover, medical studies define the accuracy of a diagnosis as the percentage of accurate diagnoses among all diagnoses; see, for example, a review of Linnet (1994). Let α_0 denote the prevalence of the disease; it measures the probability of a patient being positive (or the base rate of type 1) among the patients with the set of symptoms. We assume that $\alpha_0 \in (\underline{\alpha}, \overline{\alpha})$. We can express the accuracy of the diagnosis as

$$\text{Accuracy} = q_{11}\alpha_0 + q_{00}(1-\alpha_0).$$

3.3.2 Patients' Illness Perceptions and Decision Rules

Each patient holds an *illness perception* α towards her symptoms; it measures the patient's belief about on what probability that she is positive. Since patients generally perceive the similar (or the same) medical condition differently, and they hold different beliefs on their medical condition from doctors (Petrie et al. 2007), it is reasonable to assume that *a patient's illness perception is not an indicator of her physical status, or her true type.* We assume that each α is independently drawn from a uniform distribution, $U(0,1)$.

A patient's illness perception evolves according to Bayes' rule; a posterior illness perception serves as the prior at the next visit. Denote $g_1(\alpha)$ and $g_0(\alpha)$ as the (updated) illness perception of the patient after receiving a positive ($s=1$) and negative ($s=0$) diagnostic result, respectively, given that her current illness perception is α. Then,

$$g_1(\alpha) = \frac{q_{11}\alpha}{q_{11}\alpha + q_{01}(1-\alpha)}, \tag{3.2}$$

$$g_0(\alpha) = \frac{q_{10}\alpha}{q_{10}\alpha + q_{00}(1-\alpha)}. \tag{3.3}$$

It can be easily shown that $g_1(\alpha)$ and $g_0(\alpha)$ are both increasing in α.

The patient pays a fee f ($f \geq 0$) and incurs a waiting cost that is proportional to her waiting time in the system W (from the time she makes an appointment to the time she gets served) with a unit-time cost c associated with each visit. We assume that the waiting time will not worsen the symptoms. W is an exponential random variable with parameter $\mu - \lambda$, i.e., $W \sim \exp(\mu - \lambda)$, where λ denotes the *effective* arrival rate of the patients. Thus, the expected overall cost associated with each patient visit is

$$C_p := E[f + cW] = f + c/(\mu - \lambda) := f + cw.$$

The effective arrival rate λ is an aggregated result of the patient population's joining-and-balking/stopping-and-continuing decisions. Denoting N as the visiting times of a patient in one illness episode, λ is given as

$$\lambda = \Lambda E[N]. \tag{3.4}$$

We will derive $E[N]$ shortly. We assume that f and c are not very large so as to ensure that some patients indeed join the diagnostic service system; that is, $\lambda > 0$.

Consider a typical patient whose illness perception is α. Her stopping-and-continuing decision is based on her (likely biased) present illness perception α. Let $r(\alpha)$ denote the reward of stopping at state α. If she leaves and identifies herself as negative ($t = 0$), her gain is $(1-\alpha)V_0 - \alpha L_1$; otherwise, it is $\alpha V_1 - (1-\alpha)L_0$. Hence, $r(\alpha)$ is denoted as

$$r(\alpha) = \max\{\alpha V_1 - (1-\alpha)L_0, (1-\alpha)V_0 - \alpha L_1\},$$

which is equivalent to

$$r(\alpha) = \begin{cases} (1-\alpha)V_0 - \alpha L_1 \text{ if } \alpha < \widehat{\alpha} \\ \alpha V_1 - (1-\alpha)L_0 \text{ if } \alpha \geq \widehat{\alpha} \end{cases}, \tag{3.5}$$

where

$$\widehat{\alpha} := \frac{V_0 + L_0}{V_0 + V_1 + L_1 + L_0}. \tag{3.6}$$

Cognitive hierarchy theory suggests that untrained individuals generally make decisions based on limited depth of strategic thinking. For example, Camerer et al. (2004) found that the average number of steps individuals look ahead is 1.5 for many games. Due to the reason of traceability, we assume that the patients' stopping-and-continuing decision is based on the *One-Step Look-Ahead* (OSLA) rule. In other words, a patient continues if and only if there is an advantage in paying one more visit and then stopping; she applies the OSLA rule repeatedly at successive states visited by the random process until it reaches the stopping set S. Let $v(\alpha)$ denote her maximum expected reward net of costs. The optimality equation based on the OSLA rule is simplified as follows:

$$v(\alpha) = E\left[\max\left\{r(\alpha), -(f + cW) + p(s = 1|\alpha)r(g_1(\alpha)) + p(s = 0|\alpha)r(g_0(\alpha))\right\}\right],$$

which leads to that a patient with illness perception α continues if and only if

$$r(\alpha) < -C_p + p(s = 1|\alpha)r(g_1(\alpha)) + p(s = 0|\alpha)r(g_0(\alpha)), \tag{3.7}$$

where $p(s = 1|\alpha)$ and $p(s = 0|\alpha)$ denote her illness perceptions upon obtaining a positive ($s = 1$) and a negative ($s = 0$) diagnostic result, respectively. We have

$$p(s = 1|\alpha) = q_{01}(1 - \alpha) + q_{11}\alpha, \tag{3.8}$$

$$p(s = 0|\alpha) = q_{00}(1 - \alpha) + q_{10}\alpha. \tag{3.9}$$

Obviously, $p(s = 1|\alpha) + p(s = 0|\alpha) = 1$.

3.3.3 Classification of Patients

Since the diagnosis is reliable, if the patient's current illness perception is the same as the doctor's belief, i.e., if $\alpha = \alpha_0$, she would follow the doctor's advice. That is,

$$g_0(\alpha_0) \leq \underline{\alpha} \text{ and } g_1(\alpha_0) \geq \overline{\alpha}.$$

Considering the continuity of $g_1(\alpha)$ and $g_0(\alpha)$, there exists a small interval $[\underline{\alpha_0}, \overline{\alpha_0}]$ around α_0 ($\underline{\alpha_0} \leq \alpha_0 \leq \overline{\alpha_0}$) such that if the patient's current illness perception

satisfies $\alpha \in [\underline{\alpha_0}, \overline{\alpha_0}]$, she will pay one visit and follow the doctor's advice, where $g_1(\underline{\alpha_0}) = \overline{\alpha}$ and $g_0(\overline{\alpha_0}) = \underline{\alpha}$, respectively. It can be easily shown that

$$\underline{\alpha_0} = \frac{\overline{\alpha} q_{01}}{\overline{\alpha} q_{01} + (1-\overline{\alpha}) q_{11}}. \tag{3.10}$$

$$\overline{\alpha_0} = \frac{\underline{\alpha} q_{00}}{\underline{\alpha} q_{00} + (1-\underline{\alpha}) q_{10}}, \tag{3.11}$$

We call a patient whose illness perception falls within interval $[\underline{\alpha_0}, \overline{\alpha_0}]$ a *neutral* patient, whose illness perception falls in $(\overline{\alpha_0}, \overline{\alpha})$ a *pessimistic* patient, and whose illness perception falls in $(\underline{\alpha}, \underline{\alpha_0})$ an *optimistic* patient. Obviously, a pessimistic (optimistic, respectively) patient will leave the system immediately after a positive (negative, respectively) diagnostic result.

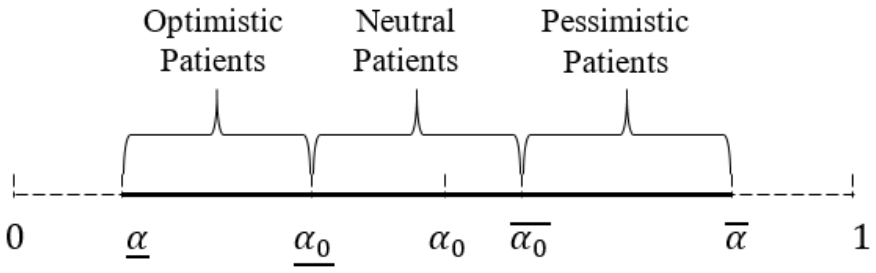


Figure 3.1: Illustration of Illness Perceptions

3.4 Patients' Optimal Stopping Problem

In this section, we first investigate each individual patient's optimal stopping problem, and then the aggregated result of the patients' optimal stopping problem. Now we derive the stopping set S.

Lemma 3.1 *A sufficient condition of a patient paying a visit is that her illness perception satisfies $\alpha < \widehat{\alpha}$ and $g_1(\alpha) > \widehat{\alpha}$, or $\alpha > \widehat{\alpha}$ and $g_0(\alpha) < \widehat{\alpha}$, where $\widehat{\alpha}$ is given by (3.6).*

Based on Lemma 3.1, we obtain the following result.

Proposition 3.1 *The patients' thresholds of continuing/stopping on the illness perception are given as follows:*

$$\underline{\alpha} = \frac{q_{01}(L_0+V_0)+C_p}{q_{11}(L_1+V_1)+q_{01}(L_0+V_0)}; \tag{3.12}$$

$$\overline{\alpha} = \frac{q_{00}(L_0+V_0)-C_p}{q_{00}(L_0+V_0)+q_{10}(L_1+V_1)}. \tag{3.13}$$

If the results of two successive diagnoses are consistent, any patient would terminate the visiting process.

Proposition 3.1 shows that the patients' optimal stopping-and-continuing decision depends on the gain of being correctly identified (i.e., V_1 and V_0), the loss of being misidentified (i.e., L_0 and L_1), the quality of the diagnosis Q, and her expected cost associated with a visit C_p. In most occasions, V_1, V_0, L_0 and L_1 are determined by the properties of the disease, which are unlikely to be changed; the quality of the diagnosis Q is usually determined the available technology and the qualification of the doctors, both of which are unlikely to be improved in the short run. The only factor the policy maker can have some influence on is the patient's expected cost associated with a visit C_p. The higher C_p is, less patients join, and hence less congestion.

The thresholds $\underline{\alpha}$ and $\overline{\alpha}$ determine the stopping set S. After obtaining the thresholds, we investigate the behavior of the joining patients. We consider a worst-case scenario, where an optimistic (pessimistic, respectively) patient firstly obtains a negative (positive, respectively) result, and in the following visits, none of any successive diagnoses consists with each other. We obtain the following result.

Lemma 3.2 *Provided that* $\frac{1}{2} < q_{00} < q_{11} < 1$*, we have the following three cases.*

1. *If*

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \geq \frac{q_{01}}{q_{11}}, \tag{3.14}$$

there is no doctor shopping patients, and all the visiting patients visit once. Otherwise,

2. *If*

$$\frac{q_{10}}{q_{00}} \leq \frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{01}}{q_{11}}; \tag{3.15}$$

there are neutral and optimistic patients, but no pessimistic patients; an optimistic patient visits at most twice and a neural patient visits once;

3. *If*

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{10}}{q_{00}}, \tag{3.16}$$

there are pessimistic, neutral, and optimistic patients. A pessimistic patient ($\overline{\alpha_0} < \alpha < \overline{\alpha}$) *pays no more than* $2n+1$ *visits whereas an optimistic patient* ($\underline{\alpha} < \alpha < \underline{\alpha_0}$) *pays at most* $2n+2$ *visits, where* n *is an integer determined by*

$$n = \min\left\{ j \geq 0 : \frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \left(\frac{q_{10}}{q_{00}}\right)^k \left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^j,\ j \in \boldsymbol{Z} \right\}. \tag{3.17}$$

where α *is the individual's illness perception and*

$$k = \begin{cases} 0 \text{ if } \underline{\alpha} < \alpha < \underline{\alpha_0} \\ 1 \text{ if } \overline{\alpha_0} < \alpha < \overline{\alpha} \end{cases}.$$

Lemma 3.2 offers insights on how a patient's illness perception evolves when she faces contradicting information. It shows that an optimistic patient visits an even number of times, and a pessimistic patient visits an odd number of times under the worst-case scenario. We find that in the groups of pessimistic and optimistic patients, the visit times of a patient under the worst-case scenario is weakly increasing with her illness perception, respectively, and that the most visiting times of the patients is determined by the discrepancy between $\overline{\alpha}$ and $\underline{\alpha}$.

We can obtain from Lemma 3.2 that when $q_{11} > q_{00}$, pessimistic and optimistic patients can be further categorized into several sub-groups according to their visiting times in the worst-case scenario. We can easily obtain through simply algebra transformation of (3.17) that a pessimistic patient whose illness perception falls into interval $(\alpha_{2n-1}, \alpha_{2n+1}]$ pays at most $2n+1$ times, and that

an optimistic patient whose illness perception falls into interval $(\alpha_{2n-2}, \alpha_{2n}]$ pays at most $2n$ times, where

$$\alpha_{2n} = \frac{\underline{\alpha}}{(1-\underline{\alpha})\left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^n + \underline{\alpha}}, \tag{3.18}$$

$$\alpha_{2n+1} = \frac{\underline{\alpha}}{(1-\underline{\alpha})\frac{q_{10}}{q_{00}}\left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^n + \underline{\alpha}}. \tag{3.19}$$

We obtain the following result.

Proposition 3.2 *When (3.14) is satisfied, the expected visiting times of a patient is $E[N] = \overline{\alpha} - \underline{\alpha}$; otherwise, if $q_{00} = q_{11}$,*

$$\begin{aligned} E[N] =& \overline{\alpha} - \underline{\alpha} + \left(\frac{1-q_{11}^2}{1-q_{11}q_{10}}\alpha_0 + \frac{q_{00}(2-q_{00})}{1-q_{00}q_{01}}(1-\alpha_0)\right)(\overline{\alpha} - \overline{\alpha_0}) \\ &+ \left(\frac{q_{11}(2-q_{11})}{1-q_{11}q_{10}}\alpha_0 + \frac{1-q_{00}^2}{1-q_{00}q_{01}}(1-\alpha_0)\right)(\underline{\alpha_0} - \underline{\alpha}); \end{aligned} \tag{3.20}$$

If $q_{00} < q_{11}$, we have the following two cases.

1. *When (3.15) is satisfied,*

$$E[N] = (\overline{\alpha} - \underline{\alpha}) + [\alpha_0 q_{11} + (1-\alpha_0)q_{01}](\underline{\alpha_0} - \underline{\alpha}). \tag{3.21}$$

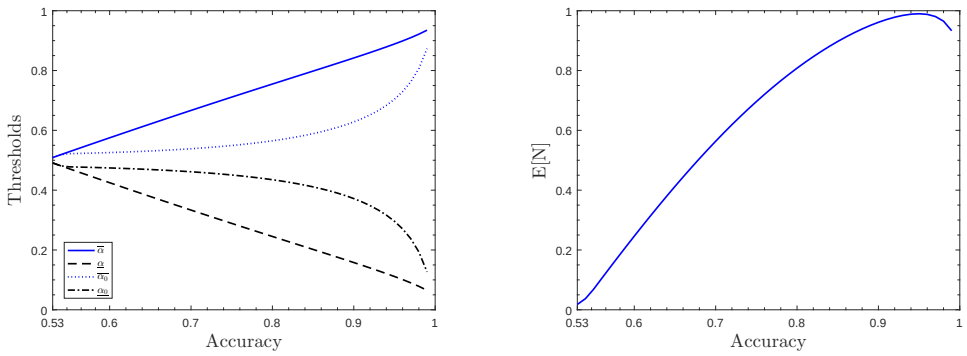
2. *When (3.16) is satisfied,*

$$\begin{aligned} E[N] =& \alpha_0\left(\overline{\alpha} + q_{11}\underline{\alpha_0}\right)\left(\frac{1+q_{10}}{1-q_{11}q_{10}} - \frac{q_{10}(1+q_{11})}{1-q_{11}q_{10}}(q_{11}q_{10})^m\right) \\ &+ \alpha_0(1+q_{11})\sum_{i=0}^{m-1}(q_{11}q_{10})^i\left(\alpha_{2i} - q_{10}\alpha_{2i+1}\right) \\ &+ \alpha_0(1+q_{11})(q_{11}q_{10})^m\alpha_{2m} \\ &+ (1-\alpha_0)\left(\overline{\alpha} + q_{01}\underline{\alpha_0}\right)\left(\frac{1+q_{00}}{1-q_{01}q_{00}} - \frac{q_{00}(1+q_{01})}{1-q_{01}q_{00}}(q_{01}q_{00})^m\right) \\ &+ (1-\alpha_0)(1+q_{01})\sum_{i=0}^{m-1}(q_{01}q_{00})^i\left(\alpha_{2i} - q_{00}\alpha_{2i+1}\right) \\ &+ (1-\alpha_0)(1+q_{01})(q_{01}q_{00})^m\alpha_{2m}, \end{aligned} \tag{3.22}$$

where $m := \max n$ and n is defined by (3.17).

$\underline{\alpha_0}$, $\overline{\alpha_0}$, $\underline{\alpha}$, $\overline{\alpha}$, α_{2i}, and α_{2i+1} are given by (3.10), (3.11), (3.12), (3.13), (3.18), and (3.19), respectively.

Proposition 3.2 shows that $E[N]$ is jointly given by $\underline{\alpha_0}$, $\overline{\alpha_0}$, $\underline{\alpha}$, $\overline{\alpha}$, α_{2i}, α_{2i+1}, and the quality of the diagnosis Q, which is indexed by q_{11} and q_{00}. Note that $\overline{\alpha_0}$, α_{2i}, and α_{2i+1} are determined by $\underline{\alpha_0}$ (see (3.11), (3.18), and (3.19), respectively), and that $\underline{\alpha_0}$ is determined by $\overline{\alpha}$ (see (3.10)). Moreover, Proposition 3.1 shows that $\underline{\alpha}$, and $\overline{\alpha}$ are determined by Q and the cost associated with a visit C_p, whereas C_p consists of a direct charge f and a waiting cost cw. The waiting cost cw is an aggregated result of the patient population's joining-and-balking/stopping-and-continuing decisions; that is, it is determined by the effective arrival rate to the system λ, where λ is given by (3.4). In a word, the joining-and-balking/stopping-and-continuing decisions of the patients are determined by the quality of the diagnosis Q, (or equivalently, by q_{11} and q_{00}) in the end. Figure 3.2 shows how the accuracy of the diagnosis affects the thresholds, including $\underline{\alpha}$, $\overline{\alpha}$, $\underline{\alpha_0}$, and $\overline{\alpha_0}$, and the expected number of visits a patient pays during one illness episode $E[N]$.



(a) Thresholds (b) Expected Visiting Times

Figure 3.2: Thresholds and Expected Visiting Times of the Patients: $V_0 = V_1 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $q_{11} = q_{00}$, $\alpha_0 = 0.50$, $f = 0$

Figure 3.2(a) shows that as the accuracy of the diagnosis increases, $\underline{\alpha}$ decreases, and $\overline{\alpha}$ increases, indicating that more patients join the diagnostic system. Meanwhile, $\underline{\alpha_0}$ decreases, and $\overline{\alpha_0}$ increases (except the region of $\underline{\alpha} = \underline{\alpha_0}$ and $\overline{\alpha} = \overline{\alpha_0}$), indicating that more patients tend to believe in the diagnosis. Also noted that the discrepancy between $\underline{\alpha}$ and $\underline{\alpha_0}$, as well as that between $\overline{\alpha}$ and $\overline{\alpha_0}$, widens as the accuracy of the diagnosis increases in the interval of $0.55 <$ Accuracy < 0.91, which indicates that more patients tend to doctor shop.

As it continues increase to Accuracy > 0.91, $\underline{\alpha_0}$ increases ($\overline{\alpha_0}$ decreases, respectively) faster than $\underline{\alpha}$ ($\overline{\alpha}$, respectively). That is, when the accuracy of the diagnosis is high, a further increase in diagnosis quality makes the patients less likely to doctor shop.

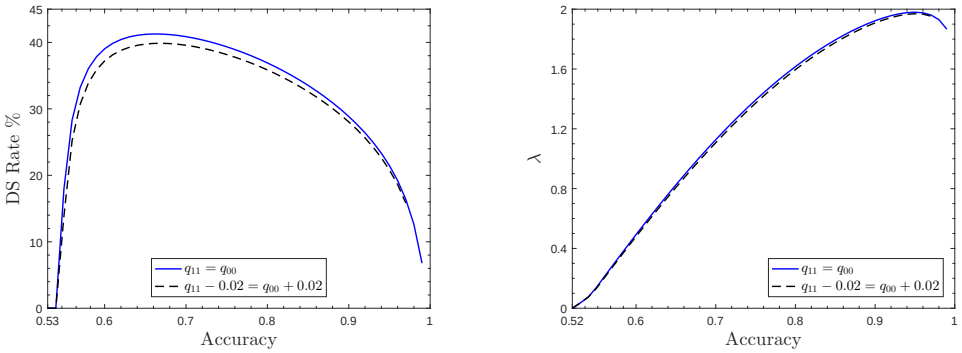
Figure 3.2(b) shows that as the quality of the diagnosis increases, i.e., as the accuracy of the diagnosis increases, the expected visiting time of the patients first increases, and then decreases; it reaches the maximum around Accuracy $= 0.91$. In the interval of $0.53 <$ Accuracy < 0.91, the increases in the accuracy of the diagnosis leads to patients more likely to join and doctor shop, and hence $E[N]$ increases. A further increase in the accuracy of the diagnosis when $0.91 <$ Accuracy < 1, even though the probability of a patient joining increases, that of a patient doctor shopping dramatically decreases, which leads to lower $E[N]$.

In the aforementioned works which give statistics about doctor shopping behavior, including Macpherson et al. (2001), Sato et al. (1995), Lo et al. (1994), Kasteler et al. (1976), doctor shopping rate is given as the proportion of patients who visit more than one doctor in one illness episode without referrals among all surveyed patients. Following them, doctor shopping rate can be expressed as

$$\text{DS Rate} = \frac{\lambda - (\overline{\alpha} - \underline{\alpha})\Lambda}{\lambda}.$$

Figure 3.3 shows the accuracy of the diagnosis affects the doctor shopping rate and the effective arrival rate. We can see from Figure 3.3(a) that as the accuracy of the diagnosis increases, the doctor shopping rate of the patients first steeply increases, next slowly decreases, and then steeply decreases. Figure 3.3(a) also compares the two scenarios, i.e., the symmetric diagnosis with $q_{11} = q_{00}$ and the asymmetric diagnosis $q_{11} > q_{00}$. It shows that under the asymmetric diagnosis the doctor shopping rate is slightly lower than under the symmetric diagnosis. Because of this, the asymmetric diagnosis also leads to slightly lower effective arrival rate, as shown in Figure 3.3(b).

In most occasions, the quality of the diagnosis Q is usually determined the available technology and the qualification of the medical practitioners; V_1, V_0, L_0 and L_1 are determined by the properties of the disease. All of them are unlikely



(a) Doctor Shopping Rate (b) Effective Arrival Rate

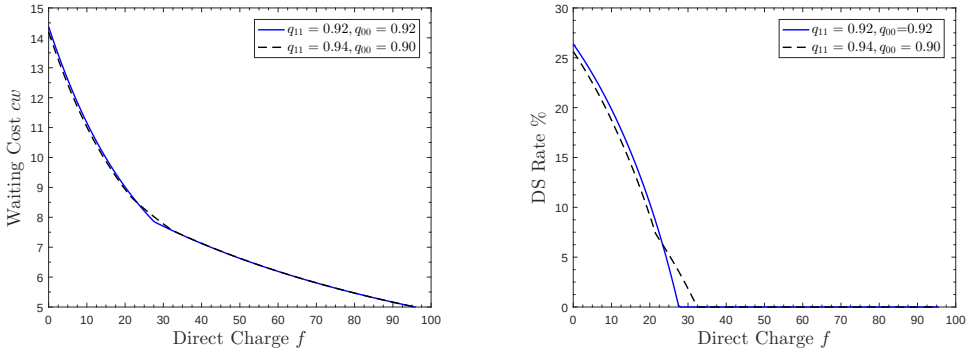
Figure 3.3: Doctor Shopping Rate and Effective Arrival Rate: $V_0 = V_1 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $\alpha_0 = 0.50$, $f = 0$

to be changed in the short run. The only manageable factor is the direct charge f, which regulates the effective arrival rate to the system via affecting the patient's expected cost associated with a visit C_p. The higher C_p is, the higher $\underline{\alpha}$ is and the lower $\overline{\alpha}$ is; that is, lower effective arrival rate, and hence less congestion. The following result demonstrates how the direct charge f affects the cost associated with a visit C_p and the waiting cost cw.

Proposition 3.3 *A lower direct charge f leads to decreased the total cost of the patients associated with a visit C_p, which induces larger coverage of the system and higher doctor shopping rate of the patients.*

A higher direct charge f is associated with a decreased waiting cost cw. Proposition 3.3 indicates that the effect of increasing the direct charge f dominates that of decreased waiting cost cw. Due to this, a higher f leads to increased total cost of the patients associated with a visit C_p. Figure 3.4(a) illustrates how cw changes corresponding to f: as f increases, cw decreases due to less arrival; the decreasing slope is less than -1, verifying that the effect of increasing the direct charge f dominates that of decreased waiting cost cw. Moreover, the direct charge f affects the workload of the system twofolds. On the one hand, higher f leads to higher C_p, which results in lower coverage of the system (lower $\overline{\alpha}$ and higher $\underline{\alpha}$; see Proposition 3.1). On the other hand, a larger range of patients joins the system indicates the joining patients visit more times in each illness episode on

average; see Lemma 3.2. In other words, a higher f is associated with a lower doctor shopping rate; see Figure 3.4(b). Hence, a higher f follows decreased waiting cost of the patients and decreased waste caused by repetitive diagnoses.



(a) Waiting Cost (b) Doctor Shopping Rate

Figure 3.4: The Effect of The Direct Charge on Waiting Cost and Doctor Shopping Rate: $V_0 = V_1 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $\alpha_0 = 0.50$

Moreover, Figure 3.4(a) shows that with other conditions unchanged, the waiting cost under the symmetric-error scenario, i.e., $q_{11} = q_{00}$, is slightly higher than that under the asymmetric-error scenario, i.e., $q_{11} > q_{00}$, when $0 \leq f \leq 22.5$, and it is slightly lower than that under the asymmetric-error scenario when $22.5 < f \leq 32.5$. This is because the doctor shopping rate under the symmetric-error scenario is higher when $0 \leq f \leq 22.5$ and is lower when $22.5 < f \leq 32.5$. Under the symmetric-error scenario, doctor shopping patients disappear when $f = 29.5$, whereas under the asymmetric-error scenario, there are three types of patients among all visiting patients when $0 \leq f \leq 22.5$, two types of patients, i.e., optimistic and neutral, when $22.5 < f \leq 32.5$, and merely neutral patients when $f > 32.5$. It leads to that the doctor shopping rate decreases slower under the asymmetric-error scenario when $22.5 < f \leq 32.5$ than that under the symmetric-error scenario. It results in a higher arrival rate under the asymmetric-error scenario, and hence a higher waiting cost.

3.5 The Effect of Doctor Shopping Behavior

Even though doctor shopping behavior of the patients is common and prevalent in many health care system, in some countries/regions, such as the United Kingdom, Singapore, and South Korea, there are well-functioned gatekeeper systems, and referrals are required to see specialists in the public sector, which effectively prevent doctor shopping behavior of the patients. In this section, we investigate how doctor shopping of the patients affect the health care system and how the policy maker shall respond to it.

The policy makers' vision on the role of their health care system is a reflection of the nation's values, politics, and economy. NHS (England), the England health care supervision agent, "strongly believe in health and high quality care for all"[3.3], indicating that the English system aims for universal coverage, whereas Hospital Authority (HA), the Hong Kong supervision agent, describes her version as "committing ourselves to the health of our community" [3.4], which indicates that the Hong Kong system stresses on maximizing social welfare.

Based on Proposition 3.1, we infer that in order to achieve universal coverage, it requires a lowest possible C_p, patient cost associated with a visit. It partially explains why the English system does not impose any direct charge and requires referrals. In this section, we investigate how a welfare-maximizing policy maker like HA shall respond to doctor shopping.

3.5.1 Doctor Shopping Behavior on the Overall Reward and Congestion

In order to understand the effects of the patients' doctor shopping behavior on the health care system, we shall understand a benchmark case, i.e., the performance of the system when doctor shopping is prohibited. Denote R^u_{ds} and R^u as the service reward from the perspective of an unbiased observer when doctor shopping is allowed and prohibited, respectively. Denote R^p_{ds} and R^p as the per-

[3.3] https://www.england.nhs.uk/about/about-nhs-england/

[3.4] http://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=10009&Lang=ENG&Dimension=100&Parent_ID=10004

ceived reward of the patients when doctor shopping is allowed and prohibited, respectively. Hereafter, we refer to R^u and R^u_{ds} as objective rewards, and R^p and R^p_{ds} as perceived rewards.

Under the scenario that doctor shopping is prohibited, each visiting patient has to terminate the visiting process after the first visit. The visiting patient, regardless of her individual illness perception, is diagnosed as positive and negative with probabilities of $p(s=1|\alpha_0)$ and $p(s=0|\alpha_0)$, respectively, where $p(s=1|\alpha)$ and $p(s=0|\alpha)$ are given by (3.8) and (3.9), respectively. Each patient gains a reward of $r(g_1(\alpha_0))$ ($r(g_1(\alpha))$, respectively) and $r(g_0(\alpha_0))$ ($r(g_0(\alpha))$, respectively) after the visit from the perspective of the unbiased observer (from the perspective of the patient with illness perception α, respectively), provided that she obtains a positive and negative result, respectively. Therefore,

$$R^u = [p(s=1|\alpha_0)r(g_1(\alpha_0)) + p(s=0|\alpha_0)r(g_0(\alpha_0))](\overline{\alpha} - \underline{\alpha}). \tag{3.23}$$

$$R^p = \int_{\underline{\alpha}}^{\overline{\alpha}} [p(s=1|\alpha_0)r(g_1(\alpha)) + p(s=0|\alpha_0)r(g_0(\alpha))]d\alpha. \tag{3.24}$$

Following the same line of thoughts, we can obtain R^u_{ds} and R^p_{ds}. However, note that the derivation of R^u_{ds} and R^p_{ds} requires an explicit knowledge on each individual's optimal stopping problem; the detailed derivation of R^u_{ds} and R^p_{ds} is shown in Appendix B.

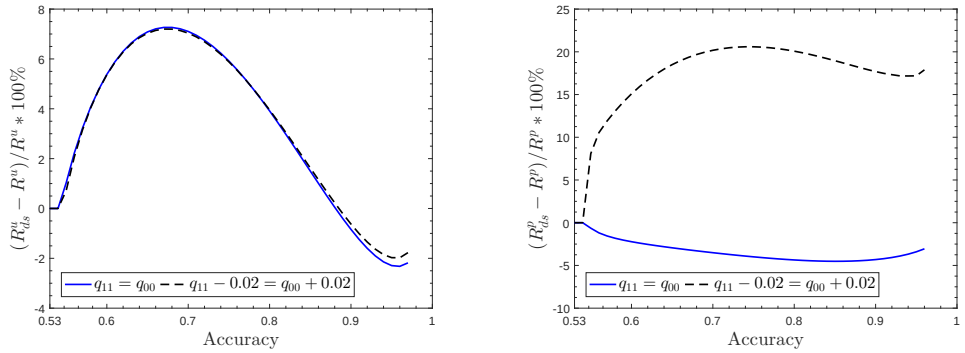
Before we investigate how a welfare-maximizing policy maker shall respond to doctor shopping, we need to understand how doctor shopping affect the health care system, like the service rewards, the congestion of the system, and patients' welfare. We start with how doctor shopping behavior affects the reward of the system. We assume that only joining patients brings reward to the system whereas balking ones do not. By looking into the perspective of an unbiased observer, we obtain the following result.

Proposition 3.4 *Given that the patients' cost associated with a visit C_p stays constant, when $q_{11} = q_{00}$, doctor shopping always improves the objective reward, and when $q_{11} > q_{00}$ and $\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \geq \frac{q_{10}}{q_{00}}$, doctor shopping improves the objective*

reward if

$$\alpha_0 < \frac{q_{00}q_{01}(V_0 + L_0)}{q_{00}q_{01}(V_0 + L_0) + q_{11}q_{10}(V_1 + L_1)}. \tag{3.25}$$

Proposition 3.4 compares how doctor shopping affects the objective reward when the patients' cost associated with a visit C_p is given. It shows that allowing doctor shopping leads to lower objective reward under certain circumstances. For general cases, C_p is a result of the aggregated result of the patients' joining decisions; we conduct numerical experiment to show how doctor shopping behavior affects both the objective and the perceived rewards when patients incur no direct charge, which is illustrated by Figure 3.5.



(a) Objective Reward (b) Perceived Reward

Figure 3.5: Doctor Shopping on Rewards: $V_1 = V_0 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $f = 0$

Figure 3.5(a) shows that when the accuracy of the diagnosis falls into (0.53, 0.88), doctor shopping leads to higher objective reward, and when the accuracy is higher than 0.88, doctor shopping behavior would lead to decreased objective reward. The increment in the objective reward led by doctor shopping behavior reaches the maximum, which is 7.4%, when the accuracy is 0.68. We can also see that when the accuracy of the diagnosis is high (Accuracy = 0.88), the decrement in the objective reward led by doctor shopping behavior is lower under the asymmetric-error scenario than that under the symmetric-error scenario.

Figure 3.5(b) shows that how the perceived reward of the patients are affected by the accuracy of the diagnosis. It shows that allowing doctor shopping always

leads to increased the perceived reward under the asymmetric-error scenario, and decreased the perceived reward under the symmetric-error scenario. Under the asymmetric-error scenario, where $q_{11} > q_{00}$, patients becomes more optimistic when they receive inconsistent diagnostic results (see B.4), leading that any patients would terminate the visiting process after a finite number of visits. Here, the increment in perceived reward led by doctor shopping reaches as high as 21%. However, inconsistent diagnostic results under the symmetric-error scenario do not help patients change mind (see (B.4)), but only lead to increased congestion, which induces higher patients cost and hence more balking patients. Therefore, doctor shopping behavior of the patients leads to decreased the perceived reward under the symmetric-error scenario.

Next, we show how shopping behavior affects the congestion of the system; see Figure 3.6. Let w_n to denote the expected waiting time of the patient in each visit under the scenario that doctor shopping behavior is prohibited. It shows that comparing with the scenario that doctor shopping behavior is prohibited, doctor shopping behavior leads to increased congestion in the system, and the increment reaches as high as 46%. As the accuracy of the diagnosis increases, the increment led by doctor shopping behavior increases, and it reaches the maximum at Accuracy $= 0.88$. As the accuracy of the diagnosis further increases, the increment led by doctor shopping behavior then decreases. Figure 3.6 also shows that the increment led by doctor shopping behavior under the asymmetric-error scenario is always no higher than that under the symmetric-error scenario. It can be explained by Figure 3.3, which shows that given the same accuracy, both doctor shopping rate and the effective arrival rate to the system are slightly lower under the asymmetric-error scenario than those under the symmetric-error scenario.

Comparing Figure 3.5 and 3.6, we can see that doctor shopping behavior leads to slightly increased rewards (to at most 7.4% on objective reward and 20.8% on perceived reward), but hugely increased congestion (to as much as 46%). Therefore, we can easily infer that when there is no direct charge incurred by

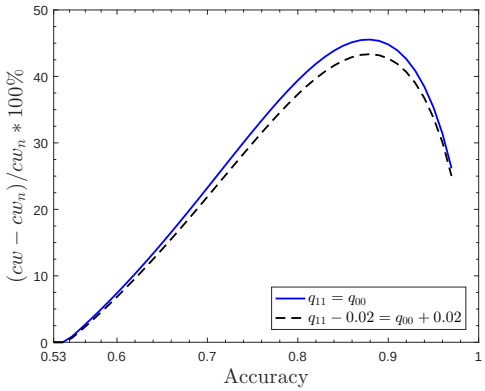


Figure 3.6: Doctor Shopping on Congestion: $V_1 = V_0 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $\alpha_0 = 0.50$, $f = 0$

the patients, the increased congestion dominates the increased reward, resulting decreased social welfare. However, a policy maker is usually able to mitigate the congestion effect by imposing fees. Next, we explore the policy maker's optimal decision of imposing fees and how it is affected by the accuracy of the diagnosis.

3.5.2 The Policy Maker's Optimal Decision

In this section, we investigate the relations between patients' doctor shopping behavior and social welfare, and examine its effect on public health policy. We consider a welfare-maximizing policy maker, who takes both the objective reward of the system and perceived reward of the patients into consideration. It is similar to the settings in in health economics studies where doctors are imperfect agent of their patients, that is, they consider the patients' needs as well as other concerns; see, for example, a comprehensive review of Chandra et al. (2011). Here the policy maker lays different weights on the objective reward and the patients' perceived reward. The objective of the policy maker is to maximize social welfare via controlling the direct charge f, and is given as follows:

$$\begin{aligned} SW &= \max_f \ [\theta R^u + (1-\theta) R^p]\Lambda - cw_n\lambda_n, \\ SW_{ds} &= \max_f \ [\theta R^u_{ds} + (1-\theta) R^p_{ds}]\Lambda - cw\lambda, \end{aligned}$$

where $0 \leq \theta \leq 1$. Higher θ indicates that the policy maker attaches more importance to the objective quality and less to the patients' perceived quality.

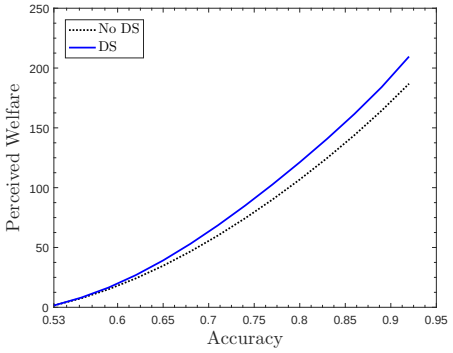
Specifically, if $\theta = 0$, the policy maker fully represents the patients, and if $\theta = 1$, he merely represents the unbiased observer (doctors). The transfer of the direct charge f is endogenized. Our numerical experiment shows that the optimal direct charge f^* shall always be 0.

We show that as far as the patients concern, their perceived welfare is highly dependent on the probabilities of the errors. Under the asymmetric-error scenario, i.e., if $q_{11} > q_{00}$, doctor shopping leads to higher perceived welfare; see Figure 3.7(a). Under the symmetric-error scenario, i.e., if $q_{11} = q_{00}$, allowing doctor shopping leads to decreased perceived welfare of the patients; see Figure 3.7(b). Moreover, comparing Figure 3.7(a) & 3.7(b), we can see that given the same accuracy, the perceived welfare of the patients is lower under the symmetric-error scenario. The underlying reason can be explained by Figure 3.5(b).

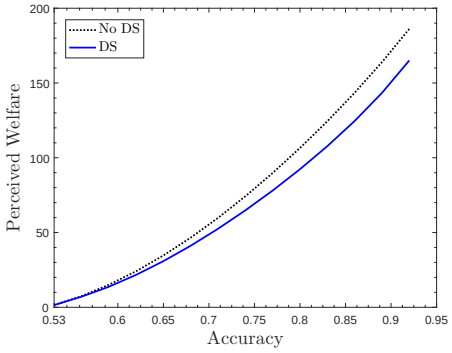
From the perspective of an unbiased observer (like a doctor), when the accuracy of the diagnosis is not high (≤ 0.75), it does not make any difference in terms of social welfare regardless of whether doctor shopping is allowed or not. When the accuracy is high, however, doctor shopping shall be prohibited for leading to welfare decrement; see Figure 3.7(c). Moreover, Figure 3.7(d) shows whether a policy maker would allow doctor shopping behavior under different θ and accuracies; doctor shopping will be tolerated beneath the dashed line. We can see that as the diagnosis becomes more accurate, the policy maker is less likely to tolerate doctor shopping behavior of the patients.

3.6 Conclusion

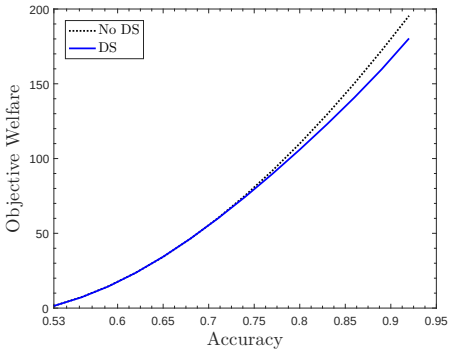
We research into the problem of providing diagnostic service with the prevalence of doctor shopping behavior. Patients perceive the illness differently from doctors and differently from one another. They are active decision makers and determine whether to follow a doctor's advice based on their individual illness perception. Whether a patient will join the public system and how many times to visit in one illness episode are jointly determined by her illness perception, the quality of diagnosis, and the cost associated with a visit she will incur. Whenever two



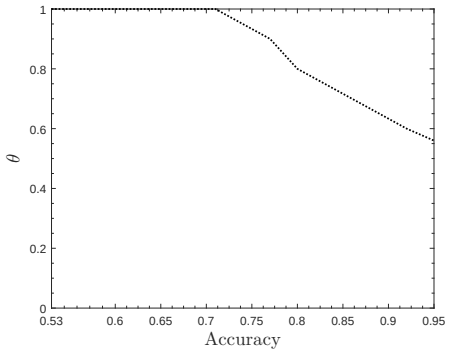
(a) Perceived Welfare: $\theta = 0$, $q_{11} - q_{00} = 0.04$



(b) Perceived Welfare: $\theta = 0$, $q_{11} = q_{00}$



(c) Objective Welfare: $\theta = 1$, $q_{11} - q_{00} = 0.04$



(d) Doctor Shopping or not

Figure 3.7: Doctor Shopping on Social Welfare and Whether it shall be Allowed: $V_1 = V_0 = 160$, $L_1 = L_0 = 80$, $\mu = 3$, $\Lambda = 2$, $c = 15$, $\alpha_0 = 0.50$

successive diagnostic results are consistent, the patient terminates the visiting process.

Existing research on doctor shopping behavior is overwhelmingly carried out from the perspective of medical practitioners and believes doctor shopping shall be avoided since it exaggerates congestion, but help little, or even adversely, in improving the objective reward. Meanwhile, doctor shopping increases the psychological gains of the patients. Therefore, a certain level of doctor shopping rate shall be tolerated under most circumstances from the perspective of welfare maximization. Our model captures the dynamic updating process of patients' illness perceptions and provides useful managerial insights and suggestions for policy makers to make appropriate decisions on health care service.

Chapter 4

Conclusion and Future Work

We deal with two topics in this thesis. First, the advance of information technology in the past decade has reshaped our consumption habits. Strategically managing buyer-generated information, such as ratings and reviews online, is an important part of today's business. In the first model, we consider a typical service situation where customers are heterogeneous in information accesses: some customers know the service-related information, whereas others do not; the latter relies on buyer-generated information to make their queueing decisions. We demonstrate that a service firm can strategically manage the ratings and improve the profitability by simply "dancing" its price, that is, replace the static pricing strategy with a cyclic pricing strategy.

The optimal cyclic pricing strategy is high-low cyclic. Under the optimal high-low cyclic pricing strategy, sophisticated customers join at the high price, and naive customers join at the low price. The system is less congested during the high-price phase, and the ratings are relatively high. The higher ratings boost the average rating and allow the provider to charge a price higher than naive customers' expected consumption utility during the low-price phase. Even though naive customers feel unsatisfied and post low scores after consumption, the average rating can still be maintained by getting high rating scores in the high-price phase. Hence, the cyclic pricing strategy allows the server to obtain a higher profit than that under a static pricing strategy.

The validity of this strategy requires the potential market size to be above a certain threshold value such that congestion is a significant factor in affecting

customers' joining decision. We further consider reference-dependent customers and find that the loss-aversion behavior dilute the attractiveness of cyclic pricing strategy; the more loss averse customers are, the less attractive the cyclic pricing is.

There are other issues calling for further study. First, we consider the stable states of the system, but the transient process to reach such stable states is left out. The transient process of the system is particularly important for short-life-cycle service products, and studying such transient process remains as an interesting future research question. Second, in our work, naive customers rely on the average rating information. In practices, they might use anecdotal reasoning, that is, they draw on a sample of the ratings and reviews and make their joining-or-balking decisions based on this limited sample. So a pricing strategy based on customer anecdotal reasoning could be an interesting research question. Third, learning over the service rate is not critical in our study as incoming customers only care about the rating on the consumption utility. It would be an interesting extension of our current study into the healthcare setting where the service quality is related with the service rate. In that setting, learning over the service rate would also be very critical. We leave it as a future research question

In the second topic, we look into the issue of how new understandings of the customer behavior is challenging how the service facilities shall be managed. We consider healthcare context in particular. We investigate the dynamics of doctor shopping behavior and its effect on the health care system. Customers, i.e., patients, are not the passive receiving end, but are joint decision makers along with the providers. In fact, patients perceive the illness differently from doctors and differently from one another. They are active decision makers and determine whether to follow a doctor's advice based on their individual illness perception. We research into the problem of providing diagnostic service with the prevalence of doctor shopping behavior.

We find that whether a patient will join the public system and how many times to visit in one illness episode are jointly determined by her illness percep-

tion, the quality of diagnosis, and the cost associated with a visit she will incur. Moreover, whenever two successive diagnostic results are consistent, the patient terminates the visiting process. Existing research on doctor shopping behavior is overwhelmingly carried out from the perspective of medical practitioners. It is believed that doctor shopping shall be avoided since it exaggerates congestion, but help little, or even adversely, in improving the objective reward. However, from the perspective of the patients, doctor shopping increases the psychological gains. In fact, a certain level of doctor shopping rate shall be tolerated under most circumstances from the perspective of welfare maximization.

Our model captures the dynamic updating process of patients' illness perceptions and provides useful managerial insights and suggestions for policy makers to make appropriate decisions on health care service. It can be extended in two ways. First, we consider a representative policy maker. Some health care systems adopt a tiered structure, and the service qualities are different between tiers. Investigating the doctor shopping behavior between different tiers might offer useful insights. Second, even though current for-profit health care facilities take up a very small share in the health care sector, it is steadily increasing in almost every country. Future work can also explore the competition/interaction between the for-profit and not-for-profit facilities, and its effect on the optimal stopping decisions and doctor shopping behaviors of the patients.

Appendix A

Proofs and Supplement for Chapter 2

A.1 Proofs of Propositions and Corollaries

A.1.1 Proof of Proposition 2.1

Recall that the overall effective arrival rate at price p_i is given by (2.5).

First, if $p_i > \eta(\boldsymbol{p})$, all naive customers balk according to (2.3), i.e., $\delta_n(p_i) = 0$. Sophisticated customers maximize their expected utility according to (2.4), which is simplified as $\mathcal{U} = \max\left\{R - \frac{c}{\mu - \delta_s(p_i)\Lambda_s} - p_i, 0\right\}$. Then, $\delta_s(p_i) = \min\left\{\frac{1}{\Lambda_s}\left(\mu - \frac{c}{R-p_i}\right), 1\right\}$. Consequently, $\lambda(p_i) = \min\left\{\mu - \frac{c}{R-p_i}, \Lambda_s\right\}$.

Next, consider $p_i \leq \eta(\boldsymbol{p})$. According to (2.3), naive customers all join, i.e., $\delta_n(p_i) = 1$. The pricing is classified into two cases:

1. If $V_n < p_i \leq \eta(\boldsymbol{p})$, $R - \frac{c}{\mu - \Lambda_n} - p_i = V_n - p_i < 0$. Sophisticated customers know that joining leads to a negative consumption utility, and therefore, according to (2.4), they all balk, i.e., $\delta_s(p_i) = 0$. Consequently, $\lambda(p_i) = \Lambda_n = (1-\theta)\Lambda$.

2. If $p_i < \min\{V_n, \eta(\boldsymbol{p})\}$, sophisticated customers decide $\delta_s(p_i) > 0$ to maximize the expected utility: $\mathcal{U} = \max\left\{R - \frac{c}{\mu - (\delta_s(p_i)\Lambda_s + \Lambda_n)} - p_i, 0\right\}$. They keep joining until $\mathcal{U} = 0$ or until all customers have joined. That is, $\delta_s(p_i) = \min\left\{\frac{1}{\Lambda_s}\left(\mu - \Lambda_n - \frac{c}{R-p_i}\right), 1\right\}$, and consequently, $\lambda(p_i) = \min\left\{\mu - \frac{c}{R-p_i}, \Lambda\right\}$

A.1.2 Proof of Proposition 2.2

When the queue is unobservable, a welfare-maximizing provider is not worse off by leaving customers an expected utility of zero (Hassin and Haviv 2003). We can simply let $v(p_i) = p_i$ without reducing social welfare. Hence, the objective function (2.8) can be written as

$$\max_{\boldsymbol{p}} \mathcal{SW} = \sum_{p_i \in \boldsymbol{p}} v(p_i)\lambda(p_i)L_{p_i} = \sum_{p_i \in \boldsymbol{p}} p_i \left(\mu - \frac{c}{R - p_i}\right) L_{p_i}.$$

One can show that $p_i \left(\mu - \frac{c}{R-p_i}\right) L_{p_i}$ is concave in p_i. Hence, there exists a p_k maximizing the term $p_i \left(\mu - \frac{c}{R-p_i}\right)$. Clearly, the optimal pricing strategy is to set $L_{p_k} = 1$ because $p_k \left(\mu - \frac{c}{R-p_k}\right) \geq \sum_{p_i \in \boldsymbol{p}} p_i \left(\mu - \frac{c}{R-p_i}\right) L_{p_i}$. In short, the optimal pricing strategy is static.

After showing that the optimal welfare-maximizing pricing strategy is static, we next obtain the optimal price. Now, we can drop the subscript and simplify the objective function of the welfare-maximizing provider (2.8) as follows:

$$\max_{p} \mathcal{SW} = p \left(\mu - \frac{c}{R - p}\right).$$

There exists a one-to-one mapping between p and $\lambda(p)$. Since it is more straightforward to use $\lambda(p)$ as the variable, we can rewrite the above objective function as

$$\max_{\lambda(p)} \mathcal{SW} = \lambda(p) \left(R - \frac{c}{\mu - \lambda(p)}\right).$$

Based on Proposition 2.1, we can see that the provider inherently decides whether he wants to serve naive customer in the long run. If he does not, $\lambda(p) < \Lambda_n$; otherwise, $\lambda(p) \geq \Lambda_n$. In the following analysis, we can solve the above optimization problem under these two constraints separately. We then compare the results to obtain the optimal pricing strategy.

We first consider the case of $\lambda(p) < \Lambda_n$. It can be easily obtained that

$$\lambda(p) = \begin{cases} \Lambda_s, & \text{if } \Lambda \leq \frac{\lambda_b}{\theta} \\ \lambda_b, & \text{if } \Lambda > \frac{\lambda_b}{\theta} \end{cases}, \text{ i.e., } p = \begin{cases} V_s, & \text{if } \Lambda \leq \frac{\lambda_b}{\theta} \\ p_b, & \text{if } \Lambda > \frac{\lambda_b}{\theta} \end{cases}.$$

The corresponding joining probabilities of the customers are as follows:

$$(\delta_s(p), \delta_n(p)) = \begin{cases} (1,\ 0), & \text{if } \Lambda \le \frac{\lambda_b}{\theta} \\ \left(\frac{\lambda_b}{\Lambda_s},\ 0\right), & \text{if } \Lambda > \frac{\lambda_b}{\theta} \end{cases}.$$

The social welfare is

$$\mathcal{SW}|_{\lambda(p)<\Lambda_n} = \begin{cases} V_s\Lambda_s, & \text{if } \Lambda \le \frac{\lambda_b}{\theta} \\ p_b\lambda_b, & \text{if } \Lambda > \frac{\lambda_b}{\theta} \end{cases}. \tag{A.1}$$

Next, we consider the case of $\lambda(p) \ge \Lambda_n$. We obtain

$$\lambda(p) = \begin{cases} \Lambda, & \text{if } 0 < \Lambda \le \lambda_b \\ \lambda_b, & \text{if } \lambda_b < \Lambda < \frac{\lambda_b}{1-\theta} \\ \Lambda_n, & \text{if } \Lambda \ge \frac{\lambda_b}{1-\theta} \end{cases}, \text{ i.e., } p = \begin{cases} R - \frac{c}{\mu-\Lambda}, & \text{if } 0 < \Lambda \le \lambda_b \\ p_b, & \text{if } \lambda_b < \Lambda < \frac{\lambda_b}{1-\theta} \\ V_n, & \text{if } \Lambda \ge \frac{\lambda_b}{1-\theta} \end{cases}.$$

The corresponding joining probabilities of the customers are

$$(\delta_s(p), \delta_n(p)) = \begin{cases} (1,\ 1), & \text{if } 0 < \Lambda \le \lambda_b \\ \left(\frac{\lambda_b-\Lambda_n}{\Lambda_s},\ 1\right), & \text{if } \lambda_b < \Lambda < \frac{\lambda_b}{1-\theta} \\ (0,\ 1), & \text{if } \Lambda \ge \frac{\lambda_b}{1-\theta} \end{cases}.$$

The corresponding social welfare is

$$\mathcal{SW}|_{\lambda(p)\ge\Lambda_n} = \begin{cases} \Lambda\left(R - \frac{c}{\mu-\Lambda}\right), & \text{if } 0 < \Lambda \le \lambda_b \\ p_b\lambda_b, & \text{if } \lambda_b < \Lambda < \frac{\lambda_b}{1-\theta} \\ V_n\Lambda_n, & \text{if } \Lambda \ge \frac{\lambda_b}{1-\theta} \end{cases}. \tag{A.2}$$

Last, the provider compares (A.1) and (A.2) to obtain the optimal pricing strategy. We start with the case $\theta < \frac{1}{2}$:

$$\mathcal{SW}|_{p\ge V_n} - \mathcal{SW}|_{p<V_n} = \begin{cases} V_s\Lambda_s - \Lambda\left(R - \frac{c}{\mu-\Lambda}\right) < 0, & \text{if } 0 < \Lambda \le \lambda_b \\ V_s\Lambda_s - p_b\lambda_b < 0, & \text{if } \lambda_b < \Lambda \le \underline{\Lambda} = \frac{\lambda_b}{1-\theta} \\ V_s\Lambda_s - V_n\Lambda_n, & \text{if } \underline{\Lambda} < \Lambda \le \overline{\Lambda} = \frac{\lambda_b}{\theta} \\ p_b\lambda_b - V_n\Lambda_n > 0, & \text{if } \Lambda > \overline{\Lambda} \end{cases}$$

Since both $\mathcal{SW}|_{p\ge V_n}$ and $\mathcal{SW}|_{p<V_n}$ are continuous, $\mathcal{SW}|_{p\ge V_n} - \mathcal{SW}|_{p<V_n}$ is continuous. It is worth mentioning the following two points here: at $\Lambda = \underline{\Lambda}$, $\Lambda_s V_s - \Lambda_n V_n = \Lambda_s V_s - \lambda_b p_b < 0$; while at $\Lambda = \overline{\Lambda}$, $\Lambda_s V_s - \Lambda_n V_n = \lambda_b p_b - \Lambda_n V_n > 0$. We then look into the interval $\underline{\Lambda} < \Lambda < \overline{\Lambda}$:

$$\Lambda_s V_s - \Lambda_n V_n = (2\theta - 1)\Lambda\left(R - \frac{c\mu}{(\mu - \theta\Lambda)(\mu - (1-\theta)\Lambda)}\right).$$

Define $g_{st}(\Lambda)$ as follows:

$$g_{st}(\Lambda) := R - \frac{c\mu}{(\mu - \theta\Lambda)(\mu - (1-\theta)\Lambda)}. \tag{A.3}$$

Function $g_{st}(\Lambda)$ is continuous in Λ. As $(2\theta - 1)\Lambda < 0$, the sign of $\Lambda_s V_s - \Lambda_n V_n$ is opposite to that of $g_{st}(\Lambda)$. Therefore, $g_{st}(\underline{\Lambda}) > 0$ and $g_{st}(\overline{\Lambda}) < 0$. Moreover,

$$\frac{dg_{st}(\Lambda)}{d\Lambda} = -\frac{c\mu(\mu - 2(1-\theta)\theta\Lambda)}{(\mu - \theta\Lambda)^2(\mu - (1-\theta)\Lambda)^2} < 0; \tag{A.4}$$

that is, $g_{st}(\Lambda)$ is monotonically decreasing in Λ. Therefore, $g_{st}(\Lambda)$ crosses 0 once at $\Lambda = \widehat{\Lambda}$, where $\widehat{\Lambda} \in (\overline{\Lambda}, \underline{\Lambda})$. We can easily obtain that

$$\widehat{\Lambda} = \frac{1}{2\theta(1-\theta)}\left(\mu - \sqrt{\mu\left[\mu - 4\theta(1-\theta)\left(\mu - \frac{c}{R}\right)\right]}\right).$$

Thus, when $\theta < \frac{1}{2}$, the optimal price and the corresponding social welfare are, respectively,

$$p^*_{sw} = \begin{cases} R - \frac{c}{\mu - \Lambda}, & \text{if } 0 < \Lambda \le \lambda_b \\ p_b, & \text{if } \lambda_b < \Lambda \le \underline{\Lambda} \\ V_n, & \text{if } \underline{\Lambda} < \Lambda \le \widehat{\Lambda} \\ V_s, & \text{if } \widehat{\Lambda} < \Lambda \le \overline{\Lambda} \\ p_b, & \text{if } \Lambda > \overline{\Lambda} \end{cases}$$

and

$$\mathcal{SW}^* = \begin{cases} \Lambda\left(R - \frac{c}{\mu - \Lambda}\right), & \text{if } 0 < \Lambda \le \lambda_b \\ p_b\lambda_b, & \text{if } \lambda_b < \Lambda \le \underline{\Lambda} \\ V_n\Lambda_n, & \text{if } \underline{\Lambda} < \Lambda \le \widehat{\Lambda} \\ V_s\Lambda_s, & \text{if } \widehat{\Lambda} < \Lambda \le \overline{\Lambda} \\ p_b\lambda_b, & \text{if } \Lambda > \overline{\Lambda} \end{cases}. \tag{A.5}$$

For the case of $\theta \ge \frac{1}{2}$,

$$\mathcal{SW}|_{p \ge V_n} - \mathcal{SW}|_{p < V_n} = \begin{cases} V_s\Lambda_s - \Lambda\left(R - \frac{c}{\mu - \Lambda}\right) < 0, & \text{if } 0 < \Lambda \le \lambda_b \\ V_s\Lambda_s - p_b\lambda_b < 0, & \text{if } \lambda_b < \Lambda \le \underline{\Lambda} = \frac{\lambda_b}{\theta} \\ 0, & \text{if } \underline{\Lambda} < \Lambda \le \overline{\Lambda} = \frac{\lambda_b}{1-\theta} \\ p_b\lambda_b - V_n\Lambda_n > 0, & \text{if } \Lambda > \overline{\Lambda} \end{cases}.$$

Then, it can be easily derived that the optimal price and social welfare are, respectively,

$$p^*_{sw} = \begin{cases} R - \frac{c}{\mu - \Lambda}, & \text{if } 0 < \Lambda \le \lambda_b \\ p_b, & \text{if } \Lambda > \lambda_b \end{cases}$$

and

$$\mathcal{SW}^* = \begin{cases} \Lambda\left(R - \frac{c}{\mu-\Lambda}\right), & \text{if } 0 < \Lambda \leq \lambda_b \\ p_b\lambda_b, & \text{if } \Lambda > \lambda_b \end{cases}. \tag{A.6}$$

A.1.3 Proof of Proposition 2.3

The Optimal Cyclic Pricing Strategy is High-Low Cyclic

A profit-maximizing provider does not leave any positive utility to the customers, and he charges a price as high as possible as long as the customers' joining decisions remain unaffected; that is, $v(p_i) \leq p_i$, where "=" holds whenever sophisticated customers join. According to Proposition 2.1, we consider the following two regions based on the magnitudes of V_n and $\eta(\boldsymbol{p})$ and derive the local optima in each region.

Region 1: $\eta(\boldsymbol{p}) \leq V_n$**.** According to Proposition 2.1, the effective arrival rate is as follows:

$$\lambda(p_i) = \begin{cases} \min\{\mu - \frac{c}{R-p_i}, \Lambda_s\} & \text{if } \eta(\boldsymbol{p}) \leq V_n < p_i \text{ or } \eta(\boldsymbol{p}) < p_i \leq V_n; \\ \min\{\mu - \frac{c}{R-p_i}, \Lambda\} & \text{if } p_i \leq \eta(\boldsymbol{p}) < V_n. \end{cases}$$

Since sophisticated customers join with positive probability at any price p_i in this region, the equilibrium arrival rate must satisfy $v(p_i) = p_i$. Hence, the profit-maximizing provider's optimization problem is the same as that of the welfare-maximizing provider, which indicates that the (local) optimal pricing strategy is static in this region.

Region 2: $\eta(\boldsymbol{p}) > V_n$**.** By Proposition 2.1, the effective arrival rate at each price p_i is

$$\lambda(p_i) = \begin{cases} \min\left\{\mu - \frac{c}{R-p_i}, \Lambda_s\right\}, & \text{if } p_i > \eta(\boldsymbol{p}); \\ \Lambda_n = (1-\theta)\Lambda, & \text{if } V_n < p_i \leq \eta(\boldsymbol{p}); \\ \min\left\{\mu - \frac{c}{R-p_i}, \Lambda\right\}, & \text{if } p_i \leq V_n. \end{cases}$$

Since sophisticated customers join with positive probability when $p_i > \eta(\boldsymbol{p})$ or $p_i \leq V_n$, we have $v(p_i) = p_i$ in these two price sets. Hence, in the price sets of $p_i > \eta(\boldsymbol{p})$ (labelled 'set A') and $p_i \leq V_n$ (labelled 'set C'), $\lambda(p_i) = \mu - \frac{c}{R-p_i}$ and thus, $p_i\left(\mu - \frac{c}{R-p_i}\right)$ is concave. There exists a unique optimal price in each

of these two sets, which we denote as p_A and p_C, respectively. Moreover, in the price set of $V_n < p_i \leq \eta(\boldsymbol{p})$ (labelled 'set B'), the arrival rate is constant as Λ_n, and hence there also exists a unique optimal price in this set, which we denote as p_B. Then, the server's profit-maximizing problem can be simplified as

$$\begin{aligned}
\max_{p_i,\ L_{p_i}, i=A,B,C} \Pi &= p_A\lambda(p_A)L_{p_A} + p_B\Lambda_n L_{p_B} + p_C\lambda(p_C)L_{p_C} \\
\text{s.t.} \quad & p_A > \eta(\boldsymbol{p}), \\
& V_n < p_B \leq \eta(\boldsymbol{p}), \\
& p_C \leq V_n, \\
& L_{p_A} + L_{p_B} + L_{p_C} = 1,\ L_{p_i} \geq 0,
\end{aligned}$$

where

$$\eta(\boldsymbol{p}) = \frac{p_A\lambda(p_A)L_{p_A} + V_n\Lambda_n L_{p_B} + p_C\lambda(p_C)L_{p_C}}{\lambda(p_A)L_{p_A} + \Lambda_n L_{p_B} + \lambda(p_C)L_{p_C}}. \tag{A.7}$$

The Lagrangian function is as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{p}, \boldsymbol{\alpha}) = {} & p_A\lambda(p_A)L_{p_A} + p_B\Lambda_n L_{p_B} + p_C\lambda(p_C)L_{p_C} \\
& + \alpha_1(p_A - \eta(\boldsymbol{p})) - \alpha_2(V_n - p_B) - \alpha_3(p_B - \eta(\boldsymbol{p})) - \alpha_4(p_C - V_n) \\
& - \alpha_5(L_{p_A} + L_{p_B} + L_{p_C} - 1) + \alpha_6 L_{p_A} + \alpha_7 L_{p_B} + \alpha_8 L_{p_C}.
\end{aligned}$$

We then obtain the following Kuhn-Tucker conditions:

$$\frac{\partial \mathcal{L}}{\partial p_A} = \frac{d(p_A\lambda(p_A))}{dp_A}L_{p_A} + \alpha_1\left(1 - \frac{\partial \eta(\boldsymbol{p})}{\partial p_A}\right) + \alpha_3\frac{\partial \eta(\boldsymbol{p})}{\partial p_A} = 0; \tag{A.8}$$

$$\frac{\partial \mathcal{L}}{\partial p_B} = \Lambda_n L_{p_B} - \alpha_1\frac{\partial \eta(\boldsymbol{p})}{\partial p_B} + \alpha_2 - \alpha_3\left(1 - \frac{\partial \eta(\boldsymbol{p})}{\partial p_B}\right) = 0; \tag{A.9}$$

$$\frac{\partial \mathcal{L}}{\partial p_C} = \frac{d(p_C\lambda(p_C))}{dp_C}L_{p_C} - \alpha_1\frac{\partial \eta(\boldsymbol{p})}{\partial p_C} + \alpha_3\frac{\partial \eta(\boldsymbol{p})}{\partial p_C} - \alpha_4 = 0; \tag{A.10}$$

$$\frac{\partial \mathcal{L}}{\partial L_{p_A}} = p_A\lambda(p_A) - (\alpha_1 - \alpha_3)\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_A}} - \alpha_5 + \alpha_6 = 0; \tag{A.11}$$

$$\frac{\partial \mathcal{L}}{\partial L_{p_B}} = p_B\Lambda_n - (\alpha_1 - \alpha_3)\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_B}} - \alpha_5 + \alpha_7 = 0; \tag{A.12}$$

$$\frac{\partial \mathcal{L}}{\partial L_{p_C}} = p_C\lambda(p_C) - (\alpha_1 - \alpha_3)\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_C}} - \alpha_5 + \alpha_8 = 0; \tag{A.13}$$

$$\alpha_1(p_A - \eta(\boldsymbol{p})) = 0;\ \alpha_2(V_n - p_B) = 0;$$
$$\alpha_3(p_B - \eta(\boldsymbol{p})) = 0;\ \alpha_4(p_C - V_n) = 0;$$
$$p_A - \eta(\boldsymbol{p}) > 0;\ p_B - V_n > 0;\ p_B - \eta(\boldsymbol{p}) \leq 0;\ p_C - V_n \leq 0;$$
$$\alpha_5(L_{p_A} + L_{p_B} + L_{p_C} - 1) = 0;\ \alpha_6 L_{p_A} = 0;$$
$$\alpha_7 L_{p_B} = 0;\ \alpha_8 L_{p_C} = 0; \alpha_i \geq 0\ (i = 1, 2...8)$$
$$L_{p_A} + L_{p_B} + L_{p_C} - 1 = 0;$$
$$0 \leq L_{p_A} \leq 1;\ 0 \leq L_{p_B} \leq 1;\ 0 \leq L_{p_C} \leq 1;$$

where

$$\overline{\lambda} = \lambda(p_A)L_{p_A} + \Lambda_n L_{p_B} + \lambda(p_C)L_{p_C}$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial p_A} = \frac{1}{\overline{\lambda}}\left(\frac{d(p_A\lambda(p_A))}{d\lambda(p_A)} - \eta(\boldsymbol{p})\right)\frac{d\lambda(p_A)}{dp_A}L_{p_A}; \tag{A.14}$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial p_B} = 0;$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial p_C} = \frac{1}{\overline{\lambda}}\left(\frac{d(p_C\lambda(p_C))}{d\lambda(p_C)} - \eta(\boldsymbol{p})\right)\frac{d\lambda(p_C)}{dp_C}L_{p_C}; \tag{A.15}$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_A}} = \frac{\lambda(p_A)}{\overline{\lambda}}(p_A - \eta(\boldsymbol{p})); \tag{A.16}$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_B}} = \frac{\Lambda_n}{\overline{\lambda}}(V_n - \eta(\boldsymbol{p})); \tag{A.17}$$
$$\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_C}} = \frac{\lambda(p_C)}{\overline{\lambda}}(p_C - \eta(\boldsymbol{p})). \tag{A.18}$$

Note that $p_A - \eta(\boldsymbol{p}) > 0$ and $p_B - V_n > 0$ imply that $\alpha_1 = \alpha_2 = 0$. As $\frac{\partial \eta(\boldsymbol{p})}{\partial p_B} = 0$ and $\alpha_1 = \alpha_2 = 0$, (A.9) is simplified as $\frac{\partial \mathcal{L}}{\partial p_B} = \Lambda_n L_{p_B} - \alpha_3 = 0$, according to which the pricing strategy is classified into two cases: (1) $L_{p_B} = 0$ and $\alpha_3 = 0$; (2) $L_{p_B} > 0$ and $\alpha_3 > 0$.

Case 1: $L_{p_B} = 0$ and $\alpha_3 = 0$. In this case, p_B becomes irrelevant. The provider allocates the pricing circle between p_A and p_C, and sophisticated customers join at both prices, indicating that $v(p_A) = p_A$ and $v(p_C) = p_C$. Hence, maximizing profit becomes the same as maximizing welfare; thus, the optimal pricing strategy is static.

Case 2: $L_{p_B} > 0$ and $\alpha_3 > 0$. Since $\alpha_3(p_B - \eta(\boldsymbol{p})) = 0$, we have $p_B^* = \eta(\boldsymbol{p})$.

Then, (A.8) and (A.10) are simplified as

$$\frac{\partial \mathcal{L}}{\partial p_A} = \left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) \frac{d(p_A \lambda(p_A))}{dp_A} L_{p_A} - \frac{\alpha_3}{\overline{\lambda}} \frac{d\lambda(p_A)}{dp_A} \eta(\boldsymbol{p}) L_{p_A} = 0,$$
$$\frac{\partial \mathcal{L}}{\partial p_C} = \left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) \frac{d(p_C \lambda(p_C))}{dp_C} L_{p_C} - \frac{\alpha_3}{\overline{\lambda}} \frac{d\lambda(p_C)}{dp_C} \eta(\boldsymbol{p}) L_{p_C} - \alpha_4 = 0,$$

by inserting (A.14) and (A.15), respectively. Then, we have

$$\begin{aligned} & L_{p_C} \frac{\partial \mathcal{L}}{\partial p_A} - L_{p_A} \frac{\partial \mathcal{L}}{\partial p_C} \\ &= \left[\left(R + \frac{\alpha_3 (R - \eta(\boldsymbol{p}))}{\overline{\lambda}} \right) \left(\frac{c}{(R - p_C)^2} - \frac{c}{(R - p_A)^2} \right) L_{p_C} + \alpha_4 \right] L_{p_A} = 0. \end{aligned} \tag{A.19}$$

As $p_C < p_A$, $\frac{c}{(R-p_C)^2} < \frac{c}{(R-p_A)^2}$. Thus, (A.19) implies one of the following two cases: (2-i) $L_{p_A} = 0$; (2-ii) $L_{p_A} > 0$ and $\alpha_4 > 0$.

Case 2-i: If $L_{p_A} = 0$, the provider decides how to allocate the pricing circle between p_B and p_C. Recall that $p_C \leq V_n$. Then from (A.7), we have $\eta(\boldsymbol{p}) \leq V_n$, i.e., the aforementioned Region 1 case. Thus, according to the discussion there, the optimal pricing shall be static pricing.

Case 2-ii: If $L_{p_A} > 0$ and $\alpha_4 > 0$, by $\alpha_4(p_C - V_n) = 0$, we have $p_C = V_n$. Based on (A.16) and (A.18), taking the difference between (A.11) and (A.13) yields

$$\begin{aligned} & p_A \lambda(p_A) - p_C \lambda(p_C) + \alpha_3 \left(\frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_A}} - \frac{\partial \eta(\boldsymbol{p})}{\partial L_{p_C}} \right) + \alpha_6 - \alpha_8 \\ &= \left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) (p_A \lambda(p_A) - p_C \lambda(p_C)) - \frac{\alpha_3 (\lambda(p_A) - \lambda(p_C)) \eta(\boldsymbol{p})}{\overline{\lambda}} + \alpha_6 - \alpha_8 \end{aligned}$$

Next, we consider the sign of $\left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) (p_A \lambda(p_A) - p_C \lambda(p_C))$, based on which we further have three sub-cases:

1. $\left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) (p_A \lambda(p_A) - p_C \lambda(p_C)) < 0$. Then, $\alpha_6 > 0$. By $\alpha_6 L_{p_A} = 0$, it indicates $L_{p_A} = 0$. This is the same as Case 2-i stated above. That is, the optimal pricing shall be static pricing.

2. $\left(1 + \frac{\alpha_3}{\overline{\lambda}}\right) (p_A \lambda(p_A) - p_C \lambda(p_C)) > 0$. Then, $\alpha_8 > 0$. By $\alpha_8 L_{p_C} = 0$, it indicates $L_{p_C} = 0$. Thus, the two prices of the cyclic pricing strategy satisfy $p_A > \eta(\boldsymbol{p})$ and $p_B = \eta(\boldsymbol{p})$.

3. $\left(1+\frac{\alpha_3}{\lambda}\right)(p_A\lambda(p_A)-p_C\lambda(p_C))=0$. Then, $\alpha_6=\alpha_8$. We further consider the difference between (A.12) and (A.13). Based on (A.17) and (A.18), and considering $p_C=V_n$, we then have

$$\begin{aligned}&p_B\Lambda_n-p_C\lambda(p_C)+\alpha_3\left(\frac{\partial\eta(\boldsymbol{p})}{\partial L_{p_B}}-\frac{\partial\eta(\boldsymbol{p})}{\partial L_{p_C}}\right)+\alpha_7-\alpha_8\\&=(p_B-V_n)\Lambda_n\left(1+\frac{\alpha_3}{\lambda}\right)+\alpha_7-\alpha_8=0.\end{aligned}$$

As $p_B>V_n$, the above equation indicates $\alpha_8>0$. Since here $\alpha_6=\alpha_8$, this implies $\alpha_6>0$. By $\alpha_6 L_{p_A}=0$ and $\alpha_8 L_{p_C}=0$, we then have $L_{p_A}=0$ and $L_{p_C}=0$. Therefore, $L_{p_B}=1$; that is, the optimal pricing strategy becomes static pricing.

In summary, there are at most two prices in the optimal pricing strategy, namely, the optimal pricing strategy is either static or high-low cyclic. Provided that it is high-low cyclic, hereafter we can simplify the notations by denoting the higher price as p_h and the lower price as p_l. The two prices satisfy $p_h>\eta(\boldsymbol{p})$ and $p_l=\eta(\boldsymbol{p})$. Otherwise, the optimal pricing strategy always reduces to static pricing.

The Optimal Profit-Maximizing Pricing Strategy

If the profit-maximizing provider adopts a static pricing strategy, i.e., $N=1$, since $p=v(p)$, his optimization problem shall be equivalent to that of the welfare-maximizing provider. Thus, the optimal profit-maximizing static price shall be the same as that under welfare maximization, which is given in Proposition 2.2. In the following analysis, we focus on obtaining the optimal cyclic pricing strategy.

We drop the subscript of L_{p_i} $(i=h,\ l)$ and use L and $1-L$ to denote the proportion of time remaining at p_h and p_l, respectively. Then,

$$\boldsymbol{p}=\{(p_h,L),(p_l,1-L)\}.$$

Since we have shown above that under the optimal cyclic pricing strategy, $p_l=\eta(\boldsymbol{p})$, the provider's decision under a cyclic pricing strategy becomes deciding p_h

and L. Therefore, his optimization problem under the cyclic pricing strategy can be simplified as follows:

$$\begin{aligned}
\max_{p_h,L}\ & \Pi_{cy} = p_h\left(\mu - \frac{c}{R-p_h}\right)L + \Lambda_n\eta(\boldsymbol{p})(1-L), \\
\text{s.t.}\quad & p_h > \eta(\boldsymbol{p}), && \text{(A.20)}\\
& \mu - \frac{c}{R-p_h} \le \Lambda_s, && \text{(A.21)}\\
& 0 < L < 1,
\end{aligned}$$

where

$$\begin{aligned}
\eta(\boldsymbol{p}) &= \frac{v(p_h)\lambda(p_h)L + V_n\Lambda_n(1-L)}{\lambda(p_h)L + \Lambda_n(1-L)} \\
&= \frac{p_h\left(\mu - \frac{c}{R-p_h}\right)L + V_n\Lambda_n(1-L)}{\left(\mu - \frac{c}{R-p_h}\right)L + \Lambda_n(1-L)}. && \text{(A.22)}
\end{aligned}$$

Since there exists a one-to-one mapping between $\lambda(p_h) = \mu - \frac{c}{R-p_h}$ and p_h, deciding $\lambda(p_h)$ is equivalent to deciding p_h. Hereafter, we use $\lambda(p_h)$ as a direct decision variable since it is more straightforward. Moreover, (A.20) and (A.21) can be simplified as:

$$\begin{cases} \lambda(p_h) \le \Lambda_s & \text{if } \theta < \frac{1}{2} \\ \lambda(p_h) < \Lambda_n & \text{if } \theta \ge \frac{1}{2} \end{cases}.$$

Scenario $\theta < \frac{1}{2}$. The Lagrangian function can be written as

$$\begin{aligned}
\mathcal{L}(p_h, L, \alpha_1, \alpha_2, \alpha_3) &= p_h\lambda(p_h)L + \Lambda_n\eta(\boldsymbol{p})(1-L) \\
&\quad -\alpha_1(\lambda(p_h) - \Lambda_s) + \alpha_2 L - \alpha_3(L-1).
\end{aligned}$$

The Kuhn-Tucker conditions are as follows:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\lambda(p_h)} &= \frac{d(\lambda(p_h)p_h)}{d\lambda(p_h)}L + \frac{\partial\eta(\boldsymbol{p})}{\partial\lambda(p_h)}\Lambda_n(1-L) - \alpha_1 \\
&= \frac{\partial\Pi_{cy}}{\partial\lambda(p_h)} - \alpha_1 = 0; && \text{(A.23)}\\
\frac{\partial\mathcal{L}}{\partial L} &= \lambda(p_h)p_h - \Lambda_n\eta(\boldsymbol{p}) + \frac{\partial\eta(\boldsymbol{p})}{\partial L}\Lambda_n(1-L) + \alpha_2 - \alpha_3 = 0; && \text{(A.24)}\\
\lambda(p_h) &\le \Lambda_s;\ \alpha_1(\lambda(p_h) - \Lambda_s) = 0; && \text{(A.25)}\\
0 < L &< 1;\ \alpha_2 L = 0;\ \alpha_3(L-1) = 0; \alpha_1, \alpha_2, \alpha_3 \ge 0,
\end{aligned}$$

where

$$\frac{\partial \eta(\boldsymbol{p})}{\partial \lambda(p_h)} = \frac{L}{\Lambda_n(1-L)+\lambda(p_h)L}\left(\frac{d(\lambda(p_h)p_h)}{d\lambda(p_h)} - \eta(\boldsymbol{p})\right); \tag{A.26}$$

$$\begin{aligned}\frac{\partial \eta(\boldsymbol{p})}{\partial L} &= \frac{\lambda(p_h)(p_h-\eta(\boldsymbol{p}))+\Lambda_n(\eta(\boldsymbol{p})-V_n)}{\Lambda_n(1-L)+\lambda(p_h)L} \\ &= \frac{\lambda(p_h)\Lambda_n(p_h-V_n)}{(\Lambda_n(1-L)+\lambda(p_h)L)^2}.\end{aligned} \tag{A.27}$$

A solution $(\lambda(p_h^*), L^*)$ that maximizes the provider's profit shall satisfy all above conditions. Note that in order to make a cyclic pricing strategy viable, we shall have $0 < L < 1$; that is, it requires $\alpha_2 = \alpha_3 = 0$. Otherwise, cyclic pricing degenerates into static pricing.

We start with the requirements of $\alpha_2 = \alpha_3 = 0$, under which (A.24) is simplified as:

$$\frac{\partial \mathcal{L}}{\partial L} = \frac{\partial \Pi_{cy}}{\partial L} = \lambda(p_h)p_h - \Lambda_n\eta(\boldsymbol{p}) + \frac{\partial \eta(\boldsymbol{p})}{\partial L}\Lambda_n(1-L) = 0. \tag{A.28}$$

Moreover,

$$\begin{aligned}\frac{\partial^2 \eta(\boldsymbol{p})}{\partial L^2} &= 2\frac{\Lambda_n - \lambda(p_h)}{\Lambda_n(1-L)+\lambda(p_h)L}\frac{\partial \eta(\boldsymbol{p})}{\partial L}; \\ \frac{\partial^2 \mathcal{L}}{\partial L^2} &= \frac{\partial^2 \Pi_{cy}}{\partial L^2} = -2\Lambda_n\frac{\partial \eta(\boldsymbol{p})}{\partial L} + \frac{\partial^2 \eta(\boldsymbol{p})}{\partial L^2}\Lambda_n(1-L) \\ &= -\frac{2\Lambda_n\lambda(p_h)}{\Lambda_n(1-L)+\lambda(p_h)L}\frac{\partial \eta(\boldsymbol{p})}{\partial L} < 0,\end{aligned} \tag{A.29}$$

since $\frac{\partial \eta(\boldsymbol{p})}{\partial L} > 0$ (see (A.27)) whenever $\theta < \frac{1}{2}$, as $\lambda(p_h) \le \Lambda_s < \Lambda_n$ implies $p_h > V_n$. This implies that $\frac{\partial \Pi_{cy}}{\partial L}$ ($\frac{\partial \mathcal{L}}{\partial L}$) decreases in L. Moreover, by (A.22), $\lim_{L\to 1}\eta(\boldsymbol{p}) = p_h$, $\lim_{L\to 0}\eta(\boldsymbol{p}) = V_n$. Moreover, by (A.27),

$$\lim_{L\to 1}\frac{\partial \eta(\boldsymbol{p})}{\partial L} = \frac{\Lambda_n}{\lambda(p_h)}(p_h - V_n), \quad \lim_{L\to 0}\frac{\partial \eta(\boldsymbol{p})}{\partial L} = \frac{\lambda(p_h)}{\Lambda_n}(p_h - V_n).$$

If we define the following function

$$\phi(L,\Lambda) := \frac{\partial \Pi_{cy}}{\partial L} = \lambda(p_h)p_h - \Lambda_n\eta(\boldsymbol{p}) + \frac{\partial \eta(\boldsymbol{p})}{\partial L}\Lambda_n(1-L),$$

then $\phi(L,\Lambda)$ decreases in L, and

$$\begin{aligned}\phi(0,\Lambda) &= \lambda(p_h)(2p_h - V_n) - V_n\Lambda_n; \\ \phi(1,\Lambda) &= p_h(\lambda(p_h) - \Lambda_n).\end{aligned} \tag{A.30}$$

Thus, to ensure that equation (A.28) has a solution satisfying $0 < L < 1$ (the cyclic pricing requirement), $\phi(0, \Lambda) > 0$ and $\phi(1, \Lambda) < 0$ are required. Note that $\phi(1, \Lambda) < 0$ always holds. We only need to determine the market condition under which $\phi(0, \Lambda) > 0$ holds.

Next, consider the requirement on α_1. By (A.23), we have $\alpha_1 = \frac{\partial \Pi_{cy}}{\partial \lambda(p_h)}$, and hence,

$$\begin{aligned}
\frac{\partial \alpha_1}{\partial \lambda(p_h)} &= \frac{\partial^2 \Pi_{cy}}{\partial \lambda(p_h)^2} \\
&= \frac{L}{\Lambda_n(1-L) + \lambda(p_h)L} \Bigg[(2\Lambda_n(1-L) + \lambda(p_h)L) \frac{d^2(p_h\lambda(p_h))}{d\lambda(p_h)^2} \\
&\qquad - \frac{2\Lambda_n(1-L)L}{\Lambda_n(1-L) + \lambda(p_h)L} \left(\frac{d(p_h\lambda(p_h))}{d\lambda(p_h)} - \eta(\boldsymbol{p}) \right) \Bigg] \\
&< \frac{2\Lambda_n(1-L)L}{(\Lambda_n(1-L) + \lambda(p_h)L)^2} \Bigg[(\Lambda_n(1-L) + \lambda(p_h)L) \frac{d^2(p_h\lambda(p_h))}{d\lambda(p_h)^2} \\
&\qquad - L \left(\frac{d(p_h\lambda(p_h))}{d\lambda(p_h)} - \eta(\boldsymbol{p}) \right) \Bigg] \\
&< \frac{2\Lambda_n(1-L)L^2}{(\Lambda_n(1-L) + \lambda(p_h)L)^2} \left(\lambda(p_h) \frac{d^2(p_h\lambda(p_h))}{d\lambda(p_h)^2} - \frac{d(p_h\lambda(p_h))}{d\lambda(p_h)} + \eta(\boldsymbol{p}) \right) \\
&< \frac{2\Lambda_n(1-L)L^2}{(\Lambda_n(1-L) + \lambda(p_h)L)^2} \left(\lambda(p_h) \frac{d^2(p_h\lambda(p_h))}{d\lambda(p_h)^2} - \frac{d(p_h\lambda(p_h))}{d\lambda(p_h)} + p_h) \right) \\
&= -\frac{2\Lambda_n(1-L)L^2}{(\Lambda_n(1-L) + \lambda(p_h)L)^2} \frac{c\lambda(p_h)(\mu + \lambda(p_h))}{(\mu - \lambda(p_h))^3} < 0.
\end{aligned} \tag{A.31}$$

The first and second "<" follow $\frac{d^2(p_h\lambda(p_h))}{d\lambda(p_h)^2} = -\frac{2c\mu}{(\mu-\lambda(p_h))^3} < 0$, and the third follows $\eta(\boldsymbol{p}) < p_h$. Thus, α_1 decreases in $\lambda(p_h)$.

Whenever the cyclic pricing strategy is feasible, by (A.23) and (A.26), we have

$$\begin{aligned}
&\frac{d(p_h\lambda(p_h))}{d\lambda(p_h)}|_{(p_h^*,\ L^*)} \\
&= \left(\frac{\Lambda_n(1-L)\eta(\boldsymbol{p})}{\lambda(p_h)L + 2\Lambda_n(1-L)} + \frac{\Lambda_n(1-L) + \lambda(p_h)L}{\lambda(p_h)L + 2\Lambda_n(1-L)} \frac{\alpha_1}{L} \right) |_{(p_h^*,\ L^*)}.
\end{aligned} \tag{A.32}$$

As $\alpha_1 \geq 0$ and $0 < L < 1$, $\frac{d(p_h\lambda(p_h))}{d\lambda(p_h)}|_{(p_h^*,\ L^*)} > 0$. Recall that $\frac{d(p\lambda(p))}{d\lambda(p)}|_{\lambda(p)=\lambda(p_b)} = 0$, where $p_b = R - \sqrt{\frac{cR}{\mu}}$. Therefore, $\lambda(p_h^*) < \lambda(p_b)$. This implies that $p_h^* > p_b$.

We consider the following three cases according to whether $\lambda(p_h) \leq \Lambda_s$ is binding or not.

Case 1: $\alpha_1 > 0$. It indicates $\lambda(p_h) = \Lambda_s$, or equivalently, $p_h^* = V_s$. Then, through

some simple algebra, (A.30)can be simplified as:

$$\phi(0,\Lambda) = (2\theta-1)\Lambda\left(R - \frac{c(\mu+\theta\Lambda)}{(\mu-\theta\Lambda)(\mu-(1-\theta)\Lambda)}\right).$$

Define

$$g_{cy}(\Lambda) := R - \frac{c(\mu+\theta\Lambda)}{(\mu-\theta\Lambda)(\mu-(1-\theta)\Lambda)}. \tag{A.33}$$

It is easy to show $g_{cy}(\Lambda)$ is continuous and decreasing in Λ whenever $V_n \geq 0$. Let $g_{cy}(\dot{\Lambda}) = 0$. We obtain

$$\dot{\Lambda} = \frac{1}{2\theta(1-\theta)}\left(\mu + \frac{c\theta}{R} - \sqrt{\left(\mu+\frac{c\theta}{R}\right)^2 - 4\theta(1-\theta)\mu\left(\mu-\frac{c}{R}\right)}\right)$$

on its domain. Since $g_{cy}(\Lambda)$ is decreasing and $g_{cy}(\dot{\Lambda}) = 0$, $g_{cy}(\Lambda) < 0$ when $\Lambda > \dot{\Lambda}$. Also, considering that $(2\theta-1)\Lambda < 0$ ($\theta < \frac{1}{2}$), the sign of $\phi(0,\Lambda)$ is opposite to that of $g_{cy}(\Lambda)$. Therefore, $\phi(0,\Lambda) > 0$ when $\Lambda > \dot{\Lambda}$. In other words, only if $\Lambda > \dot{\Lambda}$ will there be a feasible cyclic pricing strategy.

Denote the optimal pricing decisions provided $\alpha_1 > 0$ and $\Lambda > \dot{\Lambda}$ as $(p_h^*, L^*) = (V_s, L_b)$. Then L_b solves $\frac{\partial \Pi_{cy}}{\partial L}|_{p_h^*=V_s} = 0$ where $\Lambda > \dot{\Lambda}$; specifically,

$$V_s\Lambda_s + \frac{\Lambda_n(1-L_b)(V_s\Lambda_s - V_n\Lambda_n)}{\Lambda_n(1-L_b)+\Lambda_s L_b} - \frac{\Lambda_s\Lambda_n(V_s\Lambda_s L_b + V_n\Lambda_n(1-L_b))}{(\Lambda_n(1-L_b)+\Lambda_s L_b)^2} = 0. \tag{A.34}$$

Case 2: $\alpha_1 = 0$ and $\lambda(p_h) = \Lambda_s = \theta\Lambda$. This captures the "binding but irrelevant" case, under which the optimal solution is $(p_h^*,\ L^*) = (V_s,\ L_b)$, where L_b is determined by (A.34). Plugging $(p_h^*,\ L^*) = (V_s,\ L_b)$ into $\alpha_1 = \frac{\partial \Pi_{cy}}{\partial \lambda(p_h)}$ and let $\alpha_1|_{\Lambda=\ddot{\Lambda}} = 0$, we get

$$\left(R - \frac{c\mu}{(\mu-\Lambda_s)^2} - \frac{\Lambda_n(1-L_b)}{2\Lambda_n(1-L_b)+\Lambda_s L_b}\frac{V_s\Lambda_s L_b + V_n\Lambda_n(1-L_b)}{\Lambda_n(1-L_b)+\Lambda_s L_b}\right)|_{\Lambda=\ddot{\Lambda}} = 0 \tag{A.35}$$

Similar to the aforementioned analysis, we can show that the cyclic pricing strategy is feasible only when $\Lambda > \dot{\Lambda}$. Recall that α_1 decreases in $\lambda(p_h)$. As $\lambda(p_h) = \theta\Lambda$, α_1 decreases in Λ. In the above case 1, $\alpha_1|_{\lambda(p_h)=\theta\dot{\Lambda}} > 0$ and here $\alpha_1|_{\lambda(p_h)=\theta\ddot{\Lambda}} = 0$. These imply that $\dot{\Lambda} < \ddot{\Lambda}$. The optimal solution is indeed feasible. Moreover, since $\lambda(p_h^*) = \theta\ddot{\Lambda} < \lambda(p_b) = \lambda_b$, $\ddot{\Lambda} < \frac{\lambda_b}{\theta}$. As $\overline{\Lambda} = \frac{\lambda_b}{\theta}$ if $\theta < \frac{1}{2}$, we can see $\ddot{\Lambda} < \overline{\Lambda}$.

Case 3: $\alpha_1 = 0$ and $\lambda(p_h) < \Lambda_s$. This captures the unbinding case, where $\lambda(p_h) = \mu - \frac{c}{R-p_h} < \Lambda_s$. As $\alpha_1|_{\lambda(p_h)=\theta\ddot{\Lambda}} = 0$ and $\alpha_1|_{\lambda(p_h)=\mu-\frac{c}{R-p_h}<\theta\Lambda} = 0$, we get

that $\Lambda > \ddot{\Lambda}$ is required. The optimal solution, if existing, shall be obtained in the interior, solving (A.28) and the following equation simultaneously:

$$\frac{\partial \mathcal{L}}{\partial \lambda(p_h)} = \frac{d(\lambda(p_h)p_h)}{d\lambda(p_h)} L + \frac{\partial \eta(\boldsymbol{p})}{\partial \lambda(p_h)} \Lambda_n (1 - L) = 0. \tag{A.36}$$

Denote the interior solution as (p_h^0, L^0).

Last, we check the Kuhn-Tucker conditions on L. We derive from (A.30) that

$$\frac{\partial \phi(0, \Lambda)}{\partial \lambda(p_h)} = 2\frac{d(\lambda(p_h)p_h)}{d\lambda(p_h)} - V_n.$$

Under the unbinding case, by (A.32) and $\alpha_1 = 0$, the effective arrival rate at the high price always satisfies

$$\frac{d(p_h \lambda(p_h))}{d\lambda(p_h)} = \frac{\Lambda_n (1 - L)\eta(\boldsymbol{p})}{\lambda(p_h)L + 2\Lambda_n(1 - L)} > \frac{V_n}{2}.$$

Therefore, $\frac{\partial \phi(0,\Lambda)}{\partial \lambda(p_h)} > 0$. Besides, from the above analysis, we have that $g_{cy}(\Lambda)$ decreases in Λ and $g_{cy}(\dot{\Lambda}) = 0$. As $\ddot{\Lambda} > \dot{\Lambda}$, $g_{cy}(\ddot{\Lambda}) < g_{cy}(\dot{\Lambda}) = 0$. Consequently, it can be shown that $\phi(0, \ddot{\Lambda}) > \phi(0, \dot{\Lambda}) = 0$. Since $\phi(0, \Lambda)$ is increasing in $\lambda(p_h)$ and $\lambda(p_h)$ weakly increases in Λ, we have $\phi(0, \Lambda) > 0$ when $\Lambda > \ddot{\Lambda}$. This ensures that indeed there exists an optimal $L^0 \in (0, 1)$. Thus, the interior solution exists when $\Lambda > \ddot{\Lambda}$.

Scenario $\theta \geq \frac{1}{2}$. The Kuhn-Tucker conditions remain the same as those of Scenario $\theta < \frac{1}{2}$ except that the constraints in (A.25) change to

$$\lambda(p_h) - \Lambda_n < 0; \ \alpha_1(\lambda(p_h) - \Lambda_n) = 0.$$

Hence, it requires $\alpha_1 = 0$. Below we consider the following two cases: *binding but irrelevant* and *unbinding*.

Under binding but irrelevant case, $\alpha_1 = 0$ and $\lambda(p_h) = \Lambda_n$. (A.24) then can be simplified as:

$$\frac{\partial \mathcal{L}}{\partial L} = \alpha_2 - \alpha_3 = 0,$$

which requires $\alpha_2 = \alpha_3$. As $0 < L < 1$, $\alpha_2 L = 0$ and $\alpha_3(1 - L) = 0$ are required, we then have $\alpha_2 = \alpha_3 = 0$. The Kuhn-Tucker conditions on L are all

satisfied. Next, consider the Kuhn-Tucker conditions on p_h (A.23). By (A.23) and $\lambda(p_h) = \Lambda_n$,

$$\alpha_1 = L\left((2-L)\left(R - \frac{c\mu}{(\mu-\Lambda_n)^2}\right) - V_n(1-L)\right) := L\psi(\Lambda, L).$$

As $0 < L < 1$, solving $\alpha_1 = 0$ is equivalent to solving $\psi(\Lambda = \breve{\Lambda}, L) = 0$, from which we get

$$\breve{\Lambda} = \frac{1}{1-\theta}\left(\mu + \frac{c(1-L)}{R} - \sqrt{\left(\frac{c(1-L)}{2R}\right)^2 + \frac{c\mu(2-L)}{R}}\right).$$

We can also show that

$$\begin{aligned}
\frac{\partial\psi(\Lambda, L)}{\partial L} &= \frac{c\Lambda_n}{(\mu-\Lambda_n)^2} > 0; \\
\frac{\partial\psi(\Lambda, L)}{\partial \Lambda} &= -\frac{c\ (1-\theta)}{(\mu-\Lambda_n)^2}\left((2-L)\frac{2\mu}{\mu-\Lambda_n} - (1-L)\right) < 0.
\end{aligned}$$

Then, by the implicit function theorem, we have:

$$\frac{d\breve{\Lambda}}{dL} = -\frac{\frac{\partial\psi(\Lambda,L)}{\partial L}}{\frac{\partial\psi(\Lambda,L)}{\partial \Lambda}}\bigg|_{\Lambda=\breve{\Lambda}} > 0.$$

That is, $\breve{\Lambda}$ increases in L. Furthermore, when $L \to 1$, $\breve{\Lambda} \to \frac{1}{1-\theta}\left(\mu - \sqrt{\frac{c\mu}{R}}\right) = \overline{\Lambda}$, as $\theta \geq \frac{1}{2}$. Therefore, $\breve{\Lambda} < \overline{\Lambda}$.

Next, consider the unbinding case, where $\alpha_1 = 0$ and $\lambda(p_h) < \Lambda_n$. Similar to the analysis of Case 3 under Scenario $\theta < \frac{1}{2}$, here we can show that to ensure that $\lambda(p_h) < \Lambda_n$ holds, $\Lambda > \breve{\Lambda}$ is required. Besides, since $\lambda(p_h) = \Lambda_n$ when $\Lambda = \breve{\Lambda}$ as shown above, it can be easily obtained that $\phi(0,\ \breve{\Lambda}) = 0$ (see (A.30)). When $\alpha_1 = 0$, recall that as shown in Case 3 of Scenario $\theta < \frac{1}{2}$, $\frac{\partial\phi(0,\Lambda)}{\partial\lambda(p_h)} > 0$ and $\lambda(p_h)$ weakly increases in Λ. Thus, $\phi(0, \Lambda) > 0$ when $\Lambda > \breve{\Lambda}$. This ensures that indeed there exists an optimal $L^0 \in (0, 1)$. All the optimization conditions are satisfied. Thus, the interior solution does exist when $\Lambda > \breve{\Lambda}$.

To summarize, we show that the cyclic pricing strategy is feasible, and

1. if $\theta < \frac{1}{2}$ and $\dot{\Lambda} < \Lambda \leq \ddot{\Lambda}$ ($\ddot{\Lambda} < \overline{\Lambda}$), it is determined by a corner solution $(p_h^*, L^*) = (V_s, L_b)$, where L_b is determined by (A.34) and $p_h^* > p_b$.

2. if $\theta < \frac{1}{2}$ and $\Lambda > \ddot{\Lambda}$, or if $\theta \geq \frac{1}{2}$ and $\Lambda > \breve{\Lambda}$, it is determined by an interior solution $(p_h^*, L^*) = (p_h^0, L^0)$, where p_h^0 and L^0 are determined by (A.28) and (A.36) simultaneously.

Then, it can be easily shown that the interior solution (p_h^0, L^0) is obtained whenever $\Lambda \geq \overline{\Lambda}$, regardless of the magnitude of θ as $\ddot{\Lambda} < \overline{\Lambda}$ and $\breve{\Lambda} < \overline{\Lambda}$. Note that the above solution offers a necessary condition for the optimal cyclic pricing strategy. Considering $\frac{\partial^2 \Pi_{cy}}{\partial \lambda(p_h(\Lambda))^2} < 0$ and $\frac{\partial^2 \Pi_{cy}}{\partial L^2} < 0$ (see (A.29) and (A.31), respectively), the solution we obtain from the above Lagrangian method is unique. Also, note that the static pricing strategy is actually a boundary case of the cyclic pricing strategy with $L = 0$ or 1. Following the technique used by Chen and Wan (2003) and Yang et al. (2018), if the cyclic pricing strategy we obtain from the above Kuhn-Tucker conditions outperforms the boundary cases (i.e., the static pricing strategy), it indicates that it is indeed optimal. Below, we identify market conditions under which the cyclic pricing strategy we obtain from the above Lagrangian method outperforms the static pricing strategy and hence is indeed optimal.

Define $\Pi_{cy}^*(\Lambda)$ as the profit under the cyclic pricing strategy, i.e.,

$$\Pi_{cy}^*(\Lambda) = \Pi_{cy}(p_h^*(\Lambda), L^*(\Lambda)).$$

Function $\Pi_{cy}^*(\Lambda)$ is continuous in Λ. Taking the first- and second-order derivatives of $\Pi_{cy}^*(\Lambda)$ with respect to Λ, we obtain

$$\begin{aligned}\frac{d\Pi_{cy}^*(\Lambda)}{d\Lambda} = & \frac{\partial \Pi_{cy}(p_h^*(\Lambda), L^*(\Lambda))}{\partial L^*(\Lambda)} \frac{dL^*(\Lambda)}{d\Lambda} + \frac{\partial \Pi_{cy}(p_h^*(\Lambda), L^*(\Lambda))}{\partial \lambda(p_h^*(\Lambda))} \frac{d\lambda(p_h^*(\Lambda))}{d\Lambda} \\ & + \frac{\partial \Pi_{cy}(p_h^*(\Lambda), L^*(\Lambda))}{\partial \Lambda_n} \frac{d\Lambda_n}{d\Lambda},\end{aligned}$$

$$\begin{aligned}\frac{d^2\Pi^*_{cy}(\Lambda)}{d\Lambda^2} =& \frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial L^*(\Lambda)^2}\left(\frac{dL^*(\Lambda)}{d\Lambda}\right)^2 \\ &+ \frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial L^*(\Lambda)}\frac{d^2L^*(\Lambda)}{d\Lambda^2} \\ &+ \frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\lambda(p_h^*(\Lambda))^2}\left(\frac{d\lambda(p_h^*(\Lambda))}{d\Lambda}\right)^2 \\ &+ \frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\lambda(p_h^*(\Lambda))}\frac{d^2\lambda(p_h^*(\Lambda))}{d\Lambda^2} \\ &+ \frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\Lambda_n^2}\left(\frac{d\Lambda_n}{d\Lambda}\right)^2 + \frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\Lambda_n}\frac{d^2\Lambda_n}{d\Lambda^2}.\end{aligned} \tag{A.37}$$

Obviously, $\Lambda_n = (1-\theta)\Lambda$, and hence $\frac{d^2\Lambda_n}{d\Lambda^2} = 0$. Moreover,

$$\frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\Lambda_n} = \left(\eta(\boldsymbol{p})(1-L) + \Lambda_n(1-L)\frac{\partial\eta(\boldsymbol{p})}{\partial\Lambda_n}\right)\Bigg|_{(p_h^*,L^*)},$$

$$\frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\Lambda_n^2} = \left(2(1-L)\frac{\partial\eta(\boldsymbol{p})}{\partial\Lambda_n} + \Lambda_n(1-L)\frac{\partial^2\eta(\boldsymbol{p})}{\partial\Lambda_n^2}\right)\Bigg|_{(p_h^*,L^*)} < 0 \tag{A.38}$$

where

$$\begin{aligned}\frac{\partial\eta(\boldsymbol{p})}{\partial\Lambda_n} &= \frac{1-L}{\Lambda_n(1-L)+\lambda(p_h)L}\left(R - \frac{c\mu}{(\mu-\Lambda_n)^2} - \eta(\boldsymbol{p})\right) < 0, \\ \frac{\partial^2\eta(\boldsymbol{p})}{\partial\Lambda_n^2} &= \frac{1-L}{\Lambda_n(1-L)+\lambda(p_h)L}\left(-\frac{2c\mu}{(\mu-\Lambda_n)^3} - 2\frac{\partial\eta(\boldsymbol{p})}{\partial\Lambda_n}\right).\end{aligned} \tag{A.39}$$

The "<" in (A.39) and (A.38) follows from the fact that $R - \frac{c\mu}{(\mu-\Lambda_n)^2} < R - \frac{c}{\mu-\Lambda_n} = V_n$ and the requirement of $V_n < \eta(\boldsymbol{p})$ for the cyclic pricing strategy to be feasible. In addition, we have

$$\frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial L^*(\Lambda)} = 0 \text{ and } \frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial L^*(\Lambda)^2} < 0;$$

see (A.28) and (A.29). When a corner solution is obtained, $\lambda(p_h^*(\Lambda)) = \theta\Lambda$, indicating $\frac{d^2\lambda(p_h^*(\Lambda))}{d\Lambda^2} = 0$, whereas when an interior solution is obtained,

$$\frac{\partial\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\lambda(p_h^*(\Lambda))} = 0.$$

Moreover, from (A.31), we have

$$\frac{\partial^2\Pi_{cy}(p_h^*(\Lambda),L^*(\Lambda))}{\partial\lambda(p_h^*(\Lambda))^2} < 0.$$

Therefore, we simplify (A.37) as:

$$\begin{aligned}\frac{d^2\Pi^*_{cy}(\Lambda)}{d\Lambda^2} &= \frac{\partial^2\Pi_{cy}(p^*_h(\Lambda), L^*(\Lambda))}{\partial L^*(\Lambda)^2}\left(\frac{dL^*(\Lambda)}{d\Lambda}\right)^2 \\ &+ \frac{\partial^2\Pi_{cy}(p^*_h(\Lambda), L^*(\Lambda))}{\partial \lambda(p^*_h(\Lambda))^2}\left(\frac{d\lambda(p^*_h(\Lambda))}{d\Lambda}\right)^2 \\ &+ \frac{\partial^2\Pi_{cy}(p^*_h(\Lambda), L^*(\Lambda))}{\partial \Lambda_n^2}\left(\frac{d\Lambda_n}{d\Lambda}\right)^2 < 0.\end{aligned} \tag{A.40}$$

That is, $\Pi^*_{cy}(\Lambda)$ is concave in Λ.

We have obtained that the interior solution (p^0_h, L^0) is obtained whenever $\Lambda \geq \overline{\Lambda}$, regardless of the magnitude of θ. Moreover, recall that we have obtained in Proposition 2.2 that when $\Lambda \geq \overline{\Lambda}$, $\Pi^*_{st}(\Lambda) = p_b\lambda_b$ regardless of the magnitude of θ. Therefore, when $\Lambda \geq \overline{\Lambda}$, we have

$$\Pi^*_{cy}(\Lambda) = \Pi_{cy}(p^0_h(\Lambda), L^0(\Lambda)) > \max_{p_h} \Pi_{cy}(p_h(\Lambda), 1) = p_b\lambda_b = \Pi^*_{st}(\Lambda). \tag{A.41}$$

That is, the cyclic pricing strategy is always preferred when $\Lambda \geq \overline{\Lambda}$.

When $\Lambda < \overline{\Lambda}$, we consider the following two cases according to the magnitude of θ.

Case 1: $\theta < \frac{1}{2}$ and $\dot{\Lambda} < \Lambda < \overline{\Lambda}$

Comparing $g_{cy}(\Lambda)$ defined in (A.33) with $g_{st}(\Lambda)$ defined in (A.3), we can easily obtain that $g_{cy}(\Lambda) < g_{st}(\Lambda)$. Thus, $g_{cy}(\dot{\Lambda}) = 0$ indicates $g_{st}(\dot{\Lambda}) > 0$. Recall that $g_{st}(\Lambda)$ decreases in Λ (see (A.4)) and $g_{st}(\widehat{\Lambda}) = 0$. Thus, $\dot{\Lambda} < \widehat{\Lambda}$. Moreover, it can be easily shown that $\dot{\Lambda} > \lambda_b$. That is, $\lambda_b < \dot{\Lambda} \leq \widehat{\Lambda}$.

From (A.5), we can show that in the interval $(\lambda_b,\ \overline{\Lambda})$, the profit under the optimal static pricing strategy $\Pi^*_{st}(\Lambda)$, equal to the corresponding social welfare, first (weakly) decreases and then increases in Λ. Recall that $\Pi^*_{cy}(\Lambda)$ is concave in Λ; see (A.40). Hence, $\Pi^*_{st}(\Lambda)$ and $\Pi^*_{cy}(\Lambda)$ cross each other at most twice when $\dot{\Lambda} < \Lambda \leq \overline{\Lambda}$. Recall that $\phi(0, \dot{\Lambda}) = 0$ when $\theta < \frac{1}{2}$, which implies $L^* = 0$ at $\Lambda = \dot{\Lambda}$. Thus, based on (A.5) and by $\lambda_b < \dot{\Lambda} \leq \widehat{\Lambda}$, we have:

$$\begin{aligned}\lim_{\Lambda\to\dot{\Lambda}^-} \Pi^*_{cy}(\Lambda) &= \lim_{\Lambda\to\dot{\Lambda}^-} \Pi_{cy}(p^*_h(\Lambda), L^*(\Lambda)) \\ &= \Pi_{cy}(V_s,\ 0) = V_n\Lambda_n \leq \lim_{\Lambda\to\dot{\Lambda}^-} \Pi^*_{st}(\Lambda) = \Pi^*_{st}(\dot{\Lambda}).\end{aligned}$$

We also show in (A.41) that when $\Lambda \geq \overline{\Lambda}$, $\Pi^*_{cy}(\Lambda) > \Pi_{st}(\Lambda)$. Thus, considering the continuity of $\Pi_{cy}(\Lambda)$, we obtain

$$\lim_{\Lambda \to \overline{\Lambda}^+} \Pi^*_{cy}(\Lambda) = \Pi^*_{cy}(\overline{\Lambda}) > \lim_{\Lambda \to \overline{\Lambda}^+} \Pi^*_{st}(\Lambda) = \Pi^*_{st}(\overline{\Lambda}).$$

Hence, we can see that $\Pi^*_{cy}(\Lambda)$ crosses $\Pi^*_{st}(\Lambda)$ once from below. Therefore, when $\theta < \frac{1}{2}$, there exists a $\dot{\Lambda} < \widetilde{\Lambda} < \overline{\Lambda}$ such that $\Pi_{cy}(\Lambda) \geq \Pi_{st}(\Lambda)$ when $\widetilde{\Lambda} < \Lambda < \overline{\Lambda}$ and $\Pi_{cy}(\Lambda) < \Pi_{st}(\Lambda)$ when $\dot{\Lambda} < \Lambda < \widetilde{\Lambda}$.

Case 2: $\theta \geq \frac{1}{2}$ and $\breve{\Lambda} < \Lambda < \overline{\Lambda}$

When $\theta \geq \frac{1}{2}$, the optimal static profit is given by (A.6), which is weakly increasing and concave in Λ. Since $\Pi^*_{cy}(\Lambda)$ is also concave in Λ; see (A.40), $\Pi^*_{st}(\Lambda)$ and $\Pi^*_{cy}(\Lambda)$ cross each other at most twice. Recall from the analysis of Scenario $\theta \geq \frac{1}{2}$ above, $\lambda(p^*_h(\Lambda)) = \Lambda_n$ when $\Lambda = \breve{\Lambda}$. Thus,

$$\lim_{\Lambda \to \breve{\Lambda}^-} \Pi^*_{cy}(\breve{\Lambda}) = \max_L \Pi_{cy}(p^*_h(\breve{\Lambda}), L) = \Lambda_n V_n \leq \lim_{\Lambda \to \breve{\Lambda}^-} \Pi^*_{st}(\Lambda) = \Pi^*_{st}(\breve{\Lambda}).$$

At the same time, by the continuity of $\Pi_{cy}(\Lambda)$ and using (A.41), we have:

$$\lim_{\Lambda \to \overline{\Lambda}^+} \Pi^*_{cy}(\Lambda) = \Pi^*_{cy}(\overline{\Lambda}) > \lim_{\Lambda \to \overline{\Lambda}^+} \Pi^*_{st}(\Lambda) = \Pi^*_{st}(\overline{\Lambda}).$$

Therefore, when $\theta \geq \frac{1}{2}$, there exists $\breve{\Lambda} < \widetilde{\Lambda} < \overline{\Lambda}$ such that $\Pi_{cy}(\Lambda) \geq \Pi_{st}(\Lambda)$ when $\widetilde{\Lambda} \leq \Lambda < \overline{\Lambda}$ and $\Pi_{cy}(\Lambda) < \Pi_{st}(\Lambda)$ when $\breve{\Lambda} < \Lambda < \widetilde{\Lambda}$.

Proposition 2.3 is thus completely proved.

A.1.4 Proof of Corollary 2.1

When $\theta \geq 0.5$, we have $\widetilde{\Lambda} < \overline{\Lambda}$ (see Proof of Proposition 2.3). Clearly, $\lambda(p^*_h) < \lambda(p^*_l)$. And the effective arrival rate at the low price phase satisfies $\lambda(p^*_l) = \Lambda_n = (1-\theta)\Lambda < (1-\theta)\overline{\Lambda} = \lambda_b$.

A.2 Supplement

A.2.1 Discussion on the Average Rating and How it is Formed

In this model, we directly assume that incoming customers are informed about the average rating of the system, which is static in the long run. Here, we relax this

assumption and illustrate that the average rating can still be achieved through a convergence process when customers adopt exponential smoothing to aggregate the recent review data with historical data. We have the following assumptions about customers:

1. A customer observes all the rating information up to her arrival time and adopts an exponential smoothing method to compute the "average rating".

2. All customers, regardless of their types, understand the randomness embedded in the service process; thus, naive customers will not consider joining a service until a considerable number of ratings have been accumulated so that the fluctuation in the service process is averaged out.

Customers adopt an exponential smoothing method in calculating their expected rating. Consider a customer arriving at t. She divides all the historical ratings up to t into two parts. The first part consists of ratings within the time period $[t-T, t]$, where T is the same as the aforementioned pricing cycle length, and the second part includes ratings in the time period $[0, t-T]$. In the example of Yelp.com, the pricing cycle length is one month.

In this continuous-time model, at any time t, the average rating for the current time period $[t-T, t]$ is always the ratings of an intact pricing circle; see Figure A.1 for the illustration.

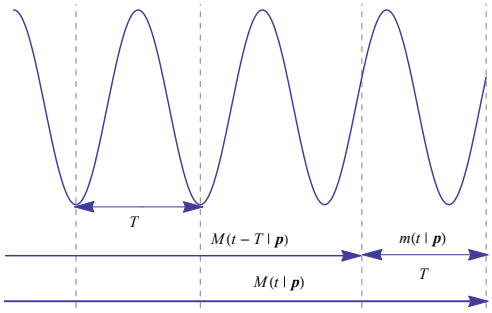


Figure A.1: An Illustration of Average Rating Computation

Denote the average rating during the time period $[t-T, t]$ by $m(t|\boldsymbol{p})$ and the average rating up to time t by $M(t|\boldsymbol{p})$. Then, we have

$$M(t|\boldsymbol{p}) = \alpha m(t|\boldsymbol{p}) + (1-\alpha)M(t-T|\boldsymbol{p}), \tag{A.42}$$

where α $(0 < \alpha < 1)$ is the smoothing factor and represents the weight the customer puts on the recent rating information.

Now we can show that the long-run average rating $M(t|\boldsymbol{p})$ converges to the average rating in a pricing circle $\eta(\boldsymbol{p})$. For a customer arriving at any time t, the first part of the average rating she observes $m(t|\boldsymbol{p})$ equals $\eta(\boldsymbol{p})$, the expected average rating in each pricing circle. According to (A.42),

$$\begin{aligned} M(t|\boldsymbol{p}) &= \alpha\eta(\boldsymbol{p}) + (1-\alpha)M(t-T|\boldsymbol{p}) \\ &= \alpha\eta(\boldsymbol{p}) + (1-\alpha)(\alpha\eta(\boldsymbol{p}) + (1-\alpha)M(t-2T|\boldsymbol{p})) \\ &= ... \\ &= \frac{\alpha\eta(\boldsymbol{p})(1-(1-\alpha)^n)}{1-(1-\alpha)} + (1-\alpha)^n M(t-nT|\boldsymbol{p}) \\ &= \eta(\boldsymbol{p}) + (1-\alpha)^n[M(t-nT|\boldsymbol{p}) - \eta(\boldsymbol{p})]. \end{aligned} \tag{A.43}$$

As $t \to \infty$, $n \to \infty$, and thus, $(1-\alpha)^n \to 0$. Then, $\lim_{t\to\infty} M(t|\boldsymbol{p}) = \eta(\boldsymbol{p})$. Hence, the long-run average rating $M(t|\boldsymbol{p})$ at any given time t converges to $\eta(\boldsymbol{p})$, where $\eta(\boldsymbol{p})$ is determined by $\boldsymbol{p}$ and is given by (2.1).

A.2.2 Algebra in Extensions

Each realized waiting time w is independently drawn from W, an exponential random variable with parameter $\mu - \lambda(p_i)$, i.e., $W \sim \exp(\mu - \lambda(p_i))$. Denote its CDF as $F(\cdot)$.

First, we show the following integration:

$$\begin{aligned} \int_0^\infty (r-w)^- dF(r) &= \int_0^w (r-w)dF(r) \\ &= \frac{1}{\mu-\lambda(p_i)} - w - \frac{1}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))w} \end{aligned} \tag{A.44}$$

$$\begin{aligned} \int_0^\infty (r-w)^+ dF(r) &= \int_w^\infty (r-w)dF(r) \\ &= \int_0^\infty (r-w)dF(r) - \int_0^w (r-w)dF(r) \\ &= \frac{1}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))w} \end{aligned} \tag{A.45}$$

Similarly, we have

$$\int_0^\infty (R-cw)^- dF(w) = R - \frac{c}{\mu-\lambda(p_i)} + \frac{c}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))\frac{R}{c}} \quad \text{(A.46)}$$

$$\int_0^\infty (R-cw)^+ dF(w) = -\frac{c}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))\frac{R}{c}} \quad \text{(A.47)}$$

Let

$$\begin{aligned} X(w) &:= v_{gl}((V(p_i), -p_i)|(\widehat{V}(p_i), -p_i)) \\ &= \int_0^\infty [(R-cw)-(R-cr)]^+ dF(r) \\ &\quad + \alpha \int_0^\infty [(R-cw)-(R-cr)]^- dF(r) \\ &= \int_0^\infty c(r-w)^+ dF(r) + \alpha \int_0^\infty c(r-w)^- dF(r) \\ &= \alpha \left(\frac{c}{\mu-\lambda(p_i)} - cw \right) - (\alpha-1)\frac{c}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))w} \end{aligned} \quad \text{(A.48)}$$

The last "=" follows (A.44) and (A.45). We further obtain

$$\begin{aligned} E[X(w)] &= \int_0^\infty X(w) dF(w) \\ &= -(\alpha-1)\frac{c}{\mu-\lambda(p_i)} \int_0^\infty e^{-(\mu-\lambda(p_i))w} dF(w) \\ &= -\frac{\alpha-1}{2}\frac{c}{\mu-\lambda(p_i)}. \end{aligned} \quad \text{(A.49)}$$

Since $u(\boldsymbol{k}|\boldsymbol{r}) = k^v + k^p + v_{gl}(\boldsymbol{k}|\boldsymbol{r})$, we can re-write $U(\text{join})$ as

$$\begin{aligned} U(\text{join}) &= E[V(p_i) - p_i] + \delta_s E[v_{gl}((V(p_i), -p_i)|(\widehat{V}(p_i), -p_i))] \\ &\quad + (1-\delta_s) E[v_{gl}((V(p_i), -p_i)|(0,0))] \\ &= E[V(p_i) - p_i] + \delta_s E[X(w)] - (1-\delta_s)\alpha p_i \\ &\quad + (1-\delta_s)\left(\int_0^\infty (R-cw)^+ dF(w) + \alpha \int_0^\infty (R-cw)^- dF(w) \right). \end{aligned}$$

By (A.46), (A.47), and (A.49), we obtain

$$\begin{aligned} U(\text{join}) &= R - \frac{c}{\mu-\lambda(p_i)} - p_i - \delta_s \frac{\alpha-1}{2}\frac{c}{\mu-\lambda(p_i)} \\ &\quad + (1-\delta_s)\left[\alpha \left(R - \frac{c}{\mu-\lambda(p_i)} \right) + (\alpha-1)\frac{c}{\mu-\lambda(p_i)} e^{-(\mu-\lambda(p_i))\frac{R}{c}} - \alpha p_i \right]. \end{aligned}$$

Similarly, we can simplify $U(\text{balk})$ as

$$\begin{aligned} U(\text{balk}) &= \delta_s E[v_{gl}((0,0)|(\widehat{V}(p_i), -p_i))] \\ &= \delta_s \left(\int_0^\infty [-(R-cw)]^+ dF(w) + \alpha \int_0^\infty [-(R-cw)]^- dF(w) + p_i \right) \\ &= -\delta_s \left(\int_0^\infty (R-cw)^- dF(w) + \alpha \int_0^\infty (R-cw)^+ dF(w) \right) + \delta_s p_i \\ &= -\delta_s \left(R - \frac{c}{\mu - \lambda(p_i)} - (\alpha - 1) \frac{c}{\mu - \lambda(p_i)} e^{-(\mu - \lambda(p_i))\frac{R}{c}} \right) + \delta_s p_i. \end{aligned}$$

The last "=" follows (A.46) and (A.47). Let $g_{gl}(\delta_s) := U(\text{join}) - U(\text{balk})$. We obtain that

$$\begin{aligned} g_{gl}(\delta_s) = (1 + \delta_s + \alpha - \alpha\delta_s) \left(R - \frac{c}{\mu - \lambda(p_i)} - p_i \right) \\ - (\alpha - 1) \frac{c}{\mu - \lambda(p_i)} \left(\frac{\delta_s}{2} + (1 - 2\delta_s) e^{-(\mu - \lambda(p_i))\frac{R}{c}} \right). \end{aligned} \tag{A.50}$$

Calculations for customer ratings.

For sophisticated customers,

$$\begin{aligned} E[u((R-cw, -p_i)|(\widehat{V}(p_i), -p_i))] &= E[R - cw - p_i + X(w)] \\ &= R - \frac{c}{\mu - \lambda(p_i)} \frac{\alpha + 1}{2} - p_i. \end{aligned}$$

It follows (A.48) and (A.49).

For naive customers,

$$\begin{aligned} & E[u((R-cw, -p_i)|(\eta'(\boldsymbol{p}), -\bar{p}))] \\ = \; & E[R - cw - p_i + v_{gl}((R-cw, -p_i)|(\eta'(\boldsymbol{p}), -\bar{p}))] \\ = \; & R - \frac{c}{\mu - \lambda(p_i)} - p_i + (-p_i + \bar{p})^+ + \alpha(-p_i + \bar{p})^- \\ & + \int_0^\infty (R - cw - \eta'(\boldsymbol{p}))^+ dF(w) + \alpha \int_0^\infty (R - cw - \eta'(\boldsymbol{p}))^- dF(w) \\ = \; & (1+\alpha) \left(R - \frac{c}{\mu - \lambda(p_i)} \right) - 2p_i + \bar{p} - \alpha\eta'(\boldsymbol{p}) \\ & + (\alpha - 1) \frac{c}{\mu - \lambda(p_i)} e^{-(\mu - \lambda(p_i))\frac{R - \eta'(\boldsymbol{p})}{c}}. \end{aligned}$$

Appendix B

Proofs and Supplement for Chapter 3

B.1 Proofs of Propositions and Lemmas

B.1.1 Proof of Lemma 3.1

We start with a patient whose illness perception satisfies $\alpha < \widehat{\alpha}$. Her illness perceptions after a negative and positive diagnostic result are $g_0(\alpha)$ and $g_1(\alpha)$, respectively (see (3.2) and (3.3)). Obviously, $g_0(\alpha) < \alpha < \widehat{\alpha}$. According to (3.5), we have

$$r(\alpha) = (1-\alpha)V_0 - \alpha L_1; \tag{B.1}$$

$$r(g_0(\alpha)) = (1-g_0(\alpha))V_0 - g_0(\alpha)L_1. \tag{B.2}$$

We assume that her illness perception α satisfies $g_1(\alpha) \leq \widehat{\alpha}$. Based on (3.5), her reward of stopping after a positive diagnosis is

$$r(g_1(\alpha)) = (1-g_1(\alpha))V_0 - g_1(\alpha)L_1. \tag{B.3}$$

Inserting (B.1), (B.2), and (B.3) into the continuing condition (3.7), it is simplifies as

$$\frac{C_p}{V_0+L_1} \leq p(s=1|\alpha)(g_1(\alpha)-\alpha) - p(s=0|\alpha)(\alpha - g_0(\alpha)).$$

Moreover, recall that the patient's beliefs on obtaining positive diagnostic result (i.e., signal) and negative diagnostic result are $p(s=1|\alpha)$ and $p(s=0|\alpha)$,

given in (3.8) and (3.9), respectively. We obtain

$$\begin{aligned}
&p(s=1|\alpha)(g_1(\alpha)-\alpha) - p(s=0|\alpha)(\alpha - g_0(\alpha)) \\
&= [q_{01}(1-\alpha)+q_{11}\alpha]\left(\frac{q_{11}\alpha}{q_{11}\alpha+q_{01}(1-\alpha)} - \alpha\right) \\
&\quad - [q_{00}(1-\alpha)+q_{10}\alpha]\left(\alpha - \frac{q_{10}\alpha}{q_{10}\alpha+q_{00}(1-\alpha)}\right) \\
&= 0.
\end{aligned}$$

Since $\frac{C_p}{V_0+L_1} > 0$, the continuing condition (3.7) is violated, which indicates that given that a patient currently identifies herself as negative, i.e., $\alpha < \widehat{\alpha}$, she would not pay one more visit if $g_1(\alpha) \leq \widehat{\alpha}$. Similarly, we can show that for a patient whose illness perception satisfies $\alpha > \widehat{\alpha}$, she would not consider one more visit if $g_0(\alpha) \geq \widehat{\alpha}$.

B.1.2 Proof of Proposition 3.1

We consider the scenario of $\alpha < \widehat{\alpha}$ and $g_1(\alpha) > \widehat{\alpha}$, where

$$\begin{aligned}
&r(\alpha) - p(s=0|\alpha)r(g_0(\alpha)) - p(s=1|\alpha)r(g_1(\alpha)) + C_p \\
&= (1-\alpha)V_0 - \alpha L_1 - p(s=0|\alpha)[(1-g_0(\alpha))V_0 - g_0(\alpha)L_1] \\
&\quad - p(s=1|\alpha)[g_1(\alpha)V_1 - (1-g_1(\alpha))L_0] + C_p \\
&= q_{01}(1-\alpha)(V_0+L_0) - q_{11}\alpha(V_1+L_1) + C_p.
\end{aligned}$$

By the continuing condition (3.7), we can see that the patients continues if and only if

$$q_{01}(1-\alpha)(V_0+L_0) - q_{11}\alpha(V_1+L_1) + C_p < 0,$$

i.e.,

$$\alpha \geq \frac{q_{01}(L_0+V_0)+C_p}{q_{11}(L_1+V_1)+q_{01}(L_0+V_0)},$$

which indicates the lower bound is

$$\underline{\alpha} = \frac{q_{01}(L_0+V_0)+C_p}{q_{11}(L_1+V_1)+q_{01}(L_0+V_0)},$$

which is (3.12). Similarly, we look into the case of $\alpha > \widehat{\alpha}$ and $g_0(\alpha) < \widehat{\alpha}$, where

$$\begin{aligned}
&r(\alpha) - p(s=1|\alpha)r(g_1(\alpha)) - p(s=0|\alpha)r(g_0(\alpha)) + C_p \\
&= \alpha V_1 - (1-\alpha)L_0 - p(s=1|\alpha)[g_1(\alpha)V_1 - (1-g_1(\alpha))L_0] \\
&\quad - p(s=0|\alpha)[(1-g_0(\alpha))V_0 - g_0(\alpha)L_1] + C_p \\
&= -q_{00}(1-\alpha)(V_1+L_1) + q_{10}\alpha(V_0+L_0) + C_p.
\end{aligned}$$

By the continuing condition (3.7), the patients continues if and only if

$$-q_{00}(1-\alpha)(V_1+L_1) + q_{10}\alpha(V_0+L_0) + C_p < 0,$$

which is

$$\alpha \leq \frac{q_{00}(L_0+V_0) - C_p}{q_{00}(L_0+V_0) + q_{10}(L_1+V_1)}.$$

Thus, we obtain the upper threshold as

$$\overline{\alpha} = \frac{q_{00}(L_0+V_0) - C_p}{q_{00}(L_0+V_0) + q_{10}(L_1+V_1)},$$

which is (3.13).

Recall that the illness perceptions of the patient after obtaining a positive and negative diagnostic result are given by (3.2) and (3.3), respectively. It can be easily shown that if $q_{00} = q_{11}$, $g_1(g_0(\alpha)) = g_0(g_1(\alpha)) = \alpha$, and if $q_{00} < q_{11}$, $g_1(g_0(\alpha)) = g_0(g_1(\alpha)) < \alpha$. That is,

$$g_1(g_0(\alpha)) = g_0(g_1(\alpha)) \leq \alpha \tag{B.4}$$

for any α.

First, we consider the special case of $\alpha = \overline{\alpha}$, the most pessimistic patient who will visit the doctor. By (B.4), we have $g_1(g_0(\overline{\alpha})) \leq \overline{\alpha}$, which indicates $g_0(\overline{\alpha}) \leq \underline{\alpha_0}$; considering that $g_0(\alpha)$ is increasing in α, we have the following relation:

$$g_0(\alpha) < g_0(\overline{\alpha}) \leq \underline{\alpha_0}, \tag{B.5}$$

from which we have

$$g_0(g_0(\alpha)) < g_0(g_0(\overline{\alpha})) \leq g_0(\underline{\alpha_0}) = \underline{\alpha}.$$

That is, any pessimistic patient will cease the visiting process after obtaining two negative results successively.

Next, consider a scenario where an optimistic patient obtains two positive results successively. Her illness perception becomes

$$g_1(g_1(\alpha)) = \frac{q_{11}^2\alpha}{q_{11}^2\alpha + q_{01}^2(1-\alpha)}. \tag{B.6}$$

Moreover,

$$\begin{aligned} g_1(g_1(\underline{\alpha})) - \overline{\alpha} &= \frac{q_{11}^2\underline{\alpha}(1-\overline{\alpha}) - q_{01}^2\overline{\alpha}(1-\underline{\alpha})}{q_{11}^2\underline{\alpha} + q_{01}^2(1-\underline{\alpha})} \\ &= \frac{q_{11}^2\overline{\alpha}(1-\underline{\alpha})}{q_{11}^2\underline{\alpha} + q_{01}^2(1-\underline{\alpha})}\left(\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} - \frac{q_{01}^2}{q_{11}^2}\right). \end{aligned}$$

Using the expressions of $\underline{\alpha}$ and $\overline{\alpha}$, given by (3.12) and (3.13), respectively, we obtain

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} = \frac{(q_{01}(L_0+V_0)+C_p)(q_{10}(L_1+V_1)+C_p)}{(q_{00}(L_0+V_0)-C_p)(q_{11}(L_1+V_1)-C_p)} > \frac{q_{01}q_{10}}{q_{00}q_{11}}. \tag{B.7}$$

We then have

$$g_1(g_1(\underline{\alpha})) - \overline{\alpha} > \frac{q_{11}^2\overline{\alpha}(1-\underline{\alpha})}{q_{11}^2\underline{\alpha} + q_{01}^2(1-\underline{\alpha})}\frac{q_{01}}{q_{11}}\left(\frac{q_{10}}{q_{00}} - \frac{q_{01}}{q_{11}}\right).$$

As $q_{11} \geq q_{00}$, $\frac{q_{10}}{q_{00}} - \frac{q_{01}}{q_{11}} \geq 0$, which indicates $g_1(g_1(\underline{\alpha})) - \overline{\alpha} \geq 0$. Therefore, the most pessimistic patient will cease the visiting process after obtaining two positive results successively. Again, considering that $g_1(\alpha)$ is increasing in α, we have

$$g_1(g_1(\alpha)) > g_1(g_1(\underline{\alpha})) \geq \overline{\alpha}. \tag{B.8}$$

Proposition 3.1 is proved.

B.1.3 Proof of Lemma 3.2

$\underline{\alpha_0}$ and $\overline{\alpha_0}$ are determined by $\overline{\alpha}$ and $\underline{\alpha}$, respectively (see (3.10) and (3.11)), and

$$\overline{\alpha} - \overline{\alpha_0} = \frac{\overline{\alpha}(1-\underline{\alpha})q_{10} - \underline{\alpha}(1-\overline{\alpha})q_{00}}{\underline{\alpha}q_{00} + (1-\underline{\alpha})q_{10}}, \tag{B.9}$$

$$\underline{\alpha_0} - \underline{\alpha} = \frac{\overline{\alpha}(1-\underline{\alpha})q_{01} - \underline{\alpha}(1-\overline{\alpha})q_{11}}{\overline{\alpha}q_{01} + (1-\overline{\alpha})q_{11}}. \tag{B.10}$$

If there were doctor-shopping patients in the population, at least one of the following conditions is required: (1) $\overline{\alpha_0} < \overline{\alpha}$ and (2) $\underline{\alpha_0} > \underline{\alpha}$, which are equivalent to

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{10}}{q_{00}} \text{ and } \frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{01}}{q_{11}},$$

respectively. We then have the following three cases according to the magnitudes of $\frac{q_{10}}{q_{00}}$, $\frac{q_{01}}{q_{11}}$, and $\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})}$. Since $q_{11} > q_{00}$, $\frac{q_{10}}{q_{00}} < \frac{q_{01}}{q_{11}}$. We have the following three cases.

Case 1:

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \geq \frac{q_{01}}{q_{11}},$$

which is (3.14). Obviously, $\overline{\alpha_0} > \overline{\alpha}$ and $\underline{\alpha_0} \leq \underline{\alpha}$; everyone is neutral, and hence each visits once. We look into Case 2 and Case 3 in the following proof.

Case 2:

$$\frac{q_{10}}{q_{00}} \leq \frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{01}}{q_{11}},$$

which is (3.15). In this case, $\overline{\alpha_0} \geq \overline{\alpha}$ and $\underline{\alpha_0} > \underline{\alpha}$; there are neutral and optimistic patients, but no pessimistic ones.

Case 3:

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \min\left\{\frac{q_{10}}{q_{00}}, \frac{q_{01}}{q_{11}}\right\} = \frac{q_{10}}{q_{00}},$$

which is (3.16). Here $\overline{\alpha_0} < \overline{\alpha}$ and $\underline{\alpha_0} > \underline{\alpha}$; there are three types of patients, namely, pessimistic, neutral, and optimistic.

Then, we look into Case 2 illustrated by (3.15), where there are two types patients, neutral and optimistic. Obviously, an optimistic patient would leave after the first visit if she obtains a negative result. If she obtains a positive result at the first visit, (1) by (B.8), we have $g_1(\alpha) > \underline{\alpha_0}$; (2) by $\underline{\alpha} < \alpha < \underline{\alpha_0}$, $g_1(\alpha) < \overline{\alpha}$. Since $\overline{\alpha} < \overline{\alpha_0}$ in Case 2, we have the following relation

$$\underline{\alpha_0} < g_1(\alpha) < \overline{\alpha_0};$$

that is, the non-neutral patient becomes neutral if she obtains a positive result, and hence she would leave after another visit.

Last, we look into Case 3, where there are three types of patients, illustrated by (3.16). We start with the worst-case scenario for a pessimistic patient (i.e., $\overline{\alpha_0} < \alpha < \overline{\alpha}$), where she first obtains a negative result, and in the following visits, none of any successive diagnoses consists with one another. Recall that a pessimistic patient would become optimistic after obtaining a negative diagnostic result, i.e., $\underline{\alpha} < g_0(\alpha) < \underline{\alpha_0}$ (see (B.5)). By (B.8), it can be easily concluded that an optimistic patient would become neutral or pessimistic upon obtaining a positive diagnosis; that is, $\underline{\alpha_0} < g_1(g_0(\alpha)) < \overline{\alpha}$. Obviously, if $\underline{\alpha_0} < g_1(g_0(\alpha)) < \overline{\alpha_0}$, she would leave after a third visit; otherwise, her illness perception would swing back and forth for multiple times until it falls into interval $(\underline{\alpha_0}, \overline{\alpha_0})$, and then pay one last visit. The illness perception evolves as follows:

$$\alpha... \to g_0 \circ (g_1 \circ g_0)^{i-1}(\alpha) \to (g_1 \circ g_0)^i(\alpha)...$$

Since $g_1(g_0(\alpha)) = g_0(g_1(\alpha)) < \alpha$, we have

$$(g_1 \circ g_0)^{i-1}(\alpha) > (g_1 \circ g_0)^i(\alpha),$$
$$g_0 \circ (g_1 \circ g_0)^{i-1}(\alpha) > g_0 \circ (g_1 \circ g_0)^i(\alpha),$$

where $i = 1, 2....$ We then infer that after some $2n$th ($n = 1...$) visit she becomes neutral instead of pessimistic under this worst-case scenario. Obviously, n is determined by the smallest integer satisfying the following relations:

$$(g_1 \circ g_0)^n(\alpha) \leq \overline{\alpha_0}. \tag{B.11}$$

In other words, she needs $2n+1$ visits in total before she eventually leaves in the worst-case scenario. It is easy to verify that

$$(g_1 \circ g_0)^n(\alpha) = (g_0 \circ g_1)^n(\alpha) = \frac{(q_{11}q_{10})^n\alpha}{(q_{11}q_{10})^n\alpha + (q_{00}q_{01})^n(1-\alpha)}, \tag{B.12}$$

Through some simple algebra, (B.11) can be rewritten as

$$\frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \frac{q_{10}}{q_{00}}\left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^n; \tag{B.13}$$

that is,

$$n = \min\left\{j \geq 1 : \frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \frac{q_{10}}{q_{00}}\left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^j, j \in \boldsymbol{Z}\right\}. \tag{B.14}$$

Obviously, $\frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})}$ decreases in α. Meanwhile, since $\frac{1}{2} < q_{00} < q_{11} < 1$, $\frac{q_{11}q_{10}}{q_{00}q_{01}} < 1$, and hence $\left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^j$ decreases in j. Hence, $n \in \boldsymbol{Z}$ is non-decreasing in α. It indicates that any pessimistic patient would never pay more visits than the most pessimistic patient in the worst-case scenario.

As for the optimistic patients, i.e., $\alpha \in (\underline{\alpha}, \underline{\alpha_0})$, we have shown that $g_1(\alpha) \in (\underline{\alpha_0}, \overline{\alpha})$ (see (B.8)). first, when $g_1(\alpha) \in (\overline{\alpha_0}, \overline{\alpha})$, we just substitute $g_1(\alpha)$ for α in (B.12), and we obtain

$$\frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^n.$$

In other word, under the worst-case scenario, given that an optimistic patient's illness perception satisfies $g_1(\alpha) \in (\overline{\alpha_0}, \overline{\alpha})$, she becomes pessimistic after the first visit, and then she needs $2n$ more visits to becomes neutral; she needs one more visit to leave after becoming neutral. That is, she needs $2n+2$ visits in total, where $n = 1, 2...$ Moreover, when $g_1(\alpha) \in (\underline{\alpha_0}, \overline{\alpha_0})$, it is easy to verify that $\frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^0$; it also satisfies $2n+2$ with $n = 0$. Hence, for the optimistic patients, n is determined by

$$n = \min\left\{j \geq 0 : \frac{\underline{\alpha}(1-\alpha)}{\alpha(1-\underline{\alpha})} \geq \left(\frac{q_{11}q_{10}}{q_{00}q_{01}}\right)^j, j \in \boldsymbol{Z}\right\}. \tag{B.15}$$

Similar to the analysis following (B.14), we can show that any optimistic patient would never pay more visits than the most pessimistic patient in the optimistic patient population in the worst-case scenario. Combining (B.14) and (B.15), we obtain (3.17).

B.1.4 Proof of Proposition 3.2

Recall that whenever two successive diagnostic results are consistent (not restrict to this), the patient leaves the system; see Proposition 3.1. Since whenever (3.14) is satisfied, every joining patient is neutral and visits once (see Proof of Lemma 3.2), and hence, $E[N] = \overline{\alpha} - \underline{\alpha}$. When (3.14) is not satisfied, we have two scenarios, that is, the symmetric-error scenario with $q_{00} = q_{11}$ and the asymmetric-error scenario, i.e., $q_{00} < q_{11}$.

We first look into the symmetric-error scenario, i.e., $q_{00} = q_{11}$, where $\frac{q_{10}}{q_{00}} = \frac{q_{01}}{q_{11}}$. If

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \geq \frac{q_{10}}{q_{00}} = \frac{q_{01}}{q_{11}},$$

$\overline{\alpha_0} \geq \overline{\alpha}$ and $\underline{\alpha_0} \leq \underline{\alpha}$, which indicates every joining patient is neutral. If

$$\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} < \frac{q_{10}}{q_{00}} = \frac{q_{01}}{q_{11}},$$

$\overline{\alpha_0} < \overline{\alpha}$ and $\underline{\alpha_0} > \underline{\alpha}$; there are three types of patients, namely, pessimistic, neutral, and optimistic. Moreover, $g_1(g_0(\alpha)) = g_0(g_1(\alpha)) = \alpha$, which indicates that under the worst-case scenario, a non-neutral patient's illness perception will transits back and forth between two states, α and $g_1(\alpha)$ if she is optimistic, and α and $g_0(\alpha)$ if she is pessimistic. Considering the results we obtain in Proposition 3.1, we can see that when $q_{01} = q_{10}$, the pessimistic and optimistic patients leave the system if and only if two successive diagnoses are consistent with each other.

Consider a patient whose illness perception falls into interval $(\overline{\alpha_0}, \overline{\alpha}]$.

1. If $N = 2i+1$ $(i = 0, 1, 2, ...)$, she first obtains a negative result and becomes optimistic; then she obtains a positive result and becomes pessimistic. This pattern repeats i times ($i = 0$ represent the scenario that she obtains a positive diagnosis at the first visit and leaves), i.e., she becomes optimistic for i times and pessimistic for i times. Before the last visit, she is pessimistic; she obtains a positive result at this last visit and leaves. That is,

$$P(N = 2i+1 | \alpha \in (\overline{\alpha_0}, \overline{\alpha}]) = q_{11}(q_{11}q_{10})^i\alpha_0 + q_{01}(q_{00}q_{01})^i(1-\alpha_0);$$

2. If $N = 2i+2$ $(i = 0, 1, 2, ...)$, her illness perception transits back and forth between α and $g_0(\alpha)$ for i times, and then she obtains two negative results successively and leaves. That is,

$$P(N = 2i+2 | \alpha \in (\overline{\alpha_0}, \overline{\alpha}]) = (q_{10})^2(q_{11}q_{10})^i\alpha_0 + (q_{00})^2(q_{00}q_{01})^i(1-\alpha_0).$$

That is, for $i = 0, 1, 2, ...$,

$$\begin{cases} P(N = 2i+1 | \alpha \in (\overline{\alpha_0}, \overline{\alpha}]) = q_{11}(q_{11}q_{10})^i\alpha_0 + q_{01}(q_{00}q_{01})^i(1-\alpha_0), \\ P(N = 2i+2 | \alpha \in (\overline{\alpha_0}, \overline{\alpha}]) = (q_{10})^2(q_{11}q_{10})^i\alpha_0 + (q_{00})^2(q_{00}q_{01})^i(1-\alpha_0). \end{cases} \tag{B.16}$$

Similarly, we obtain the probability mass of the patient whose illness perception falls into interval $[\underline{\alpha}, \underline{\alpha_0})$ as follows:

$$\begin{cases} P(N = 2i+1 | \alpha \in [\underline{\alpha}, \underline{\alpha_0})) = q_{10}(q_{11}q_{10})^i \alpha_0 + q_{00}(q_{00}q_{01})^i (1-\alpha_0), \\ P(N = 2i+2 | \alpha \in [\underline{\alpha}, \underline{\alpha_0})) = (q_{11})^2 (q_{11}q_{10})^i \alpha_0 + (q_{01})^2 (q_{00}q_{01})^i (1-\alpha_0). \end{cases} \tag{B.17}$$

By (B.16) and (B.17), we obtain

$$E[N | \alpha \in (\overline{\alpha_0}, \overline{\alpha}]] = \frac{2-q_{11}}{1-q_{11}q_{10}}\alpha_0 + \frac{1+q_{00}}{1-q_{00}q_{01}}(1-\alpha_0), \tag{B.18}$$

$$E[N | \alpha \in [\underline{\alpha}, \underline{\alpha_0})] = \frac{1+q_{11}}{1-q_{11}q_{10}}\alpha_0 + \frac{2-q_{00}}{1-q_{00}q_{01}}(1-\alpha_0), \tag{B.19}$$

respectively. We further obtain

$$\begin{aligned} E[N] &= E[N | \alpha \in [\overline{\alpha_0}, \overline{\alpha_0})](\overline{\alpha} - \overline{\alpha_0}) + (\overline{\alpha_0} - \underline{\alpha_0}) + E[N | \alpha \in [\underline{\alpha}, \underline{\alpha_0})](\underline{\alpha_0} - \underline{\alpha}) \\ &= (\overline{\alpha} - \underline{\alpha}) + (E[N | \alpha \in [\overline{\alpha_0}, \overline{\alpha_0})] - 1)(\overline{\alpha} - \overline{\alpha_0}) \\ &\quad + \left(E[N | \alpha \in [\underline{\alpha}, \underline{\alpha_0})] - 1\right)(\underline{\alpha_0} - \underline{\alpha}). \end{aligned}$$

Inserting (B.18) and (B.19) into the above equation, we obtain (3.20)

Next, we look into the asymmetric-error scenario, i.e., $q_{00} < q_{11}$. The case illustrated (3.15) is easy, where the joining patients are either optimistic or neutral. Here an optimistic patient pays at most two visits, and the second visit is needed only if she obtains a positive diagnosis (the probability of which is $\alpha_0 q_{11} + (1-\alpha_0)q_{01}$) at the first visit. Therefore, the expected visiting times of a patient $E[N]$ as follows (3.21).

Now we focus on the case illustrated (3.16). We start with the scenario that the patient is **pessimistic** and her illness perception falls into interval $(\alpha_{2n-1}, \alpha_{2n+1}]$, where $n \geq 1$.

1. If $N = 2i+1 < 2n+1$ $(i = 0, 1, ...n-1)$, she first obtains a negative result and becomes optimistic; then she obtains a positive result and becomes pessimistic. This pattern repeats i times, i.e., she becomes optimistic for i times and pessimistic for i times. Before the last visit, she is pessimistic; she obtains a positive result at this last visit and leaves. That is,

$$P(N = 2i+1 | \alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) = q_{11}(q_{11}q_{10})^i \alpha_0 + q_{01}(q_{00}q_{01})^i (1-\alpha_0);$$

2. If $N = 2n+1$, this is the worst case. Same as the case of $N = 2i+1 < 2n+1$, her mind swings back and forth between optimistic and pessimistic for n times; different from the case of $N = 2i+1 < 2n+1$, she becomes neutral after the swinging, and follows whatever advice the last doctor gives. That is,

$$P(N = 2n+1|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) = (q_{11}q_{10})^n\alpha_0 + (q_{00}q_{01})^n(1-\alpha_0);$$

3. If $N = 2i+2 < 2n+1$ $(i = 0, 1, ...n-1)$, her mind swings back and forth between optimistic and pessimistic for $i-1$ times, and then she obtains two negative results successively and leaves. That is,

$$P(N = 2i+2|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) = (q_{10})^2(q_{11}q_{10})^i\alpha_0 + (q_{00})^2(q_{00}q_{01})^i(1-\alpha_0).$$

To sum up, for $i = 0, 1, ...n-1$,

$$\begin{cases} P(N = 2i+1|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) \\ \qquad = q_{11}(q_{11}q_{10})^i\alpha_0 + q_{01}(q_{00}q_{01})^i(1-\alpha_0), \\ P(N = 2n+1|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) \\ \qquad = (q_{11}q_{10})^n\alpha_0 + (q_{00}q_{01})^n(1-\alpha_0), \\ P(N = 2i+2|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]) \\ \qquad = (q_{10})^2(q_{11}q_{10})^i\alpha_0 + (q_{00})^2(q_{00}q_{01})^i(1-\alpha_0). \end{cases} \tag{B.20}$$

Similarly, we obtain when $\alpha \in (\alpha_{2n-2}, \alpha_{2n}]$, i.e., the patient is **optimistic**,

$$\begin{cases} P(N = 2i+1|\alpha \in (\alpha_{2n-2}, \alpha_{2n}]) \\ \qquad = q_{10}(q_{11}q_{10})^i\alpha_0 + q_{00}(q_{00}q_{01})^i(1-\alpha_0), \\ P(N = 2i+2|\alpha \in (\alpha_{2n-2}, \alpha_{2n}]) \\ \qquad = (q_{11})^2(q_{11}q_{10})^i\alpha_0 + (q_{01})^2(q_{00}q_{01})^i(1-\alpha_0), \\ P(N = 2n+2|(\alpha_{2n-2}, \alpha_{2n}]) \\ \qquad = q_{11}(q_{11}q_{10})^n\alpha_0 + q_{01}(q_{00}q_{01})^n(1-\alpha_0), \end{cases} \tag{B.21}$$

where $i = 0, 1, ...n-1$. By (B.20) and (B.21), we obtain

$$\begin{aligned} E[N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]] =& \frac{\alpha_0}{1-q_{11}q_{10}}\left(1 + q_{10} - q_{10}(1+q_{11})(q_{11}q_{10})^n\right) \\ &+ \frac{1-\alpha_0}{1-q_{01}q_{00}}\left(1 + q_{00} - q_{00}(1+q_{01})(q_{01}q_{00})^n\right), \end{aligned} \tag{B.22}$$

and

$$\begin{aligned} E[N|\alpha \in (\alpha_{2n-2}, \alpha_{2n}]] =& \frac{\alpha_0(1+q_{11})}{1-q_{11}q_{10}}\left(1 - (q_{11}q_{10})^n\right) \\ &+ \frac{(1-\alpha_0)(1+q_{01})}{1-q_{01}q_{00}}\left(1 - (q_{01}q_{00})^n\right), \end{aligned} \tag{B.23}$$

respectively, where $n \geq 1$. Moreover, since $m = \max n$, we obtain m as follows through some easy algebra transformation:

$$m = \left\{ j \geq 0 : \frac{q_{10}}{q_{00}} \left(\frac{q_{11}q_{10}}{q_{00}q_{01}} \right)^j \leq \frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \leq \frac{q_{01}}{q_{11}} \left(\frac{q_{11}q_{10}}{q_{00}q_{01}} \right)^j, \ j \in \boldsymbol{Z} \right\}. \quad \text{(B.24)}$$

It can be easily shown that m is weakly increasing in $\underline{\alpha}$ and weakly decreasing in $\overline{\alpha}$. Remember that when $\alpha \in (\underline{\alpha_0}, \overline{\alpha_0})$, the patient visits only once. We have

$$\begin{aligned} E[N] = & \sum_{n=1}^{m-1} E[N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]](\alpha_{2n+1} - \alpha_{2n-1}) \\ & + E[N|\alpha \in (\alpha_{2m-1}, \alpha_{2m+1}]](\overline{\alpha} - \alpha_{2m-1}) + \overline{\alpha_0} - \underline{\alpha_0} \\ & + \sum_{n=1}^{m} E[N|\alpha \in (\alpha_{2n-2}, \alpha_{2n}]](\alpha_{2n} - \alpha_{2n-2}) \\ & + E[N|\alpha \in (\alpha_{2m}, \alpha_{2m+2}]](\underline{\alpha_0} - \alpha_{2m}), \end{aligned}$$

based on which we further obtain (3.22).

B.1.5 Proof of Proposition 3.3

We obtain from (3.12) and (3.13)

$$\frac{d\underline{\alpha}}{dC_p} = \frac{1}{q_{11}(L_1 + V_1) + q_{01}(L_0 + V_0)}, \quad \text{(B.25)}$$

$$\frac{d\overline{\alpha}}{dC_p} = -\frac{1}{q_{00}(L_0 + V_0) + q_{10}(L_1 + V_1)}, \quad \text{(B.26)}$$

respectively; that is, if the waiting time decreases, $\underline{\alpha}$ decreases and $\overline{\alpha}$ increases. In other words, more patients join if the waiting time is reduced. Moreover, from (3.10) and (3.11), we have, respectively,

$$\frac{d\underline{\alpha_0}}{d\overline{\alpha}} = \frac{q_{11}q_{01}}{(\overline{\alpha}q_{01} + (1-\overline{\alpha})q_{11})^2} > 0, \quad \text{(B.27)}$$

$$\frac{d\overline{\alpha_0}}{d\underline{\alpha}} = \frac{q_{00}q_{10}}{(\underline{\alpha}q_{00} + (1-\underline{\alpha})q_{10})^2} > 0. \quad \text{(B.28)}$$

We have obtained that for the case illustrated by (3.14), there is no doctor shopping patients, and $E[N] = \overline{\alpha} - \underline{\alpha}$. Here,

$$\frac{dE[N]}{dC_p} = \left(\frac{d\overline{\alpha}}{dC_p} - \frac{d\underline{\alpha}}{dC_p} \right) < 0, \quad \text{(B.29)}$$

When (3.14) is not satisfied and $q_{00} = q_{11}$, $E[N]$ is given by (3.20), where

$$\begin{aligned}\frac{1-q_{11}^2}{1-q_{11}q_{10}}\alpha_0 + \frac{q_{00}(2-q_{00})}{1-q_{00}q_{01}}(1-\alpha_0) &> 0,\\ \frac{q_{11}(2-q_{11})}{1-q_{11}q_{10}}\alpha_0 + \frac{1-q_{00}^2}{1-q_{00}q_{01}}(1-\alpha_0) &> 0.\end{aligned}$$

We have

$$\begin{aligned}\frac{dE[N]}{dC_p} =& \left(\frac{d\overline{\alpha}}{dC_p} - \frac{d\underline{\alpha}}{dC_p}\right)\\ &+ \left(\frac{1-q_{11}^2}{1-q_{11}q_{10}}\alpha_0 + \frac{q_{00}(2-q_{00})}{1-q_{00}q_{01}}(1-\alpha_0)\right)\left(\frac{d\overline{\alpha}}{dC_p} - \frac{d\overline{\alpha_0}}{d\underline{\alpha}}\frac{d\underline{\alpha}}{dC_p}\right)\\ &+ \left(\frac{q_{11}(2-q_{11})}{1-q_{11}q_{10}}\alpha_0 + \frac{1-q_{00}^2}{1-q_{00}q_{01}}(1-\alpha_0)\right)\left(\frac{d\underline{\alpha_0}}{d\overline{\alpha}}\frac{d\overline{\alpha}}{dC_p} - \frac{d\underline{\alpha}}{dC_p}\right)\\ <& 0.\end{aligned} \tag{B.30}$$

For the asymmetric-error scenario, i.e., $q_{00} < q_{11}$, there are two cases. One is illustrated by (3.15), where $E[N]$ is given by (3.21). It can be easily shown that

$$\frac{dE[N]}{dC_p} = \left(\frac{d\overline{\alpha}}{dC_p} - \frac{d\underline{\alpha}}{dC_p}\right) + [\alpha_0 q_{11} + (1-\alpha_0)q_{01}]\left(\frac{d\underline{\alpha_0}}{d\overline{\alpha}}\frac{d\overline{\alpha}}{dC_p} - \frac{d\underline{\alpha}}{dC_p}\right) < 0. \tag{B.31}$$

Last, we check the asymmetric-error scenario is illustrated by (3.16). From (3.18) and (3.19), we obtain, respectively,

$$\frac{\partial \alpha_{2n}}{\partial \underline{\alpha}} = \frac{K_1}{[(1-\underline{\alpha})K_1 + \underline{\alpha}]^2}, \tag{B.32}$$

$$\frac{\partial \alpha_{2n+1}}{\partial \underline{\alpha}} = \frac{K_2}{[(1-\underline{\alpha})K_2 + \underline{\alpha}]^2}. \tag{B.33}$$

From (B.22), we can see that the expected visiting time of a patient whose illness perception falls into interval $\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]$ is determined by n, where α_{2n-1} is determined by (3.19), which is determined by $\underline{\alpha}$. Obviously, for $n' > n$, $E[N|\alpha \in (\alpha_{2n'-1}, \alpha_{2n'+1}]] > E[N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]]$.

We next check how $E[N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]]$ changes in respond to the changes of $\underline{\alpha}$. Suppose that the lower bound increases from $\underline{\alpha}$ to $\underline{\alpha}'$. Denote the new interval within which a patient visit at most $2n+1$ times as $(\alpha'_{2n-1}, \alpha'_{2n+1}]$. Since $\frac{d\alpha_{2n+1}}{d\underline{\alpha}} > 0$ (see (B.33)), $\alpha'_{2n-1} > \alpha_{2n-1}$ and $\alpha'_{2n+1} > \alpha_{2n+1}$. First, suppose that C_p decreases a value such that m determined by (B.24) stays the same. Then, $\alpha_{2n-1} < \alpha'_{2n-1} < \alpha_{2n+1} < \alpha'_{2n+1}$. Now, we can see that upon the reduced waiting

time, patients in interval $(\alpha'_{2n-1}, \alpha_{2n+1}]$ visit $2n+1$ times, where as patients in interval $[\alpha_{2n-1}, \alpha'_{2n-1}]$ visit $2(n-1)+1$ times. Obviously, the expected visiting time in interval $(\alpha_{2n-1}, \alpha_{2n+1}]$ decreases as lower bound increases from $\underline{\alpha}$ to $\underline{\alpha}'$; namely,

$$\frac{\partial}{\partial \underline{\alpha}}\{E[N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}]](\alpha_{2n+1} - \alpha_{2n-1})\} < 0. \tag{B.34}$$

Similarly, we can show that the above relation also holds when m increases. Similar to the above analysis of (B.34), by (B.23) and (B.32), it can also be shown that

$$\frac{\partial}{\partial \underline{\alpha}}\{E[N|\alpha \in (\alpha_{2n-2}, \alpha_{2n}]](\alpha_{2n} - \alpha_{2n-2})\} < 0. \tag{B.35}$$

By (B.34), (B.35), and (B.25), we can see that

$$\frac{\partial E[N]}{\partial \underline{\alpha}} < 0. \tag{B.36}$$

Moreover, By (3.22), we have

$$\begin{aligned}\frac{\partial E[N]}{\partial \overline{\alpha}} =& \alpha_0 \left(1 + q_{11}\frac{d\underline{\alpha}_0}{d\overline{\alpha}}\right)\left(\frac{1+q_{10}}{1-q_{11}q_{10}} - \frac{q_{10}(1+q_{11})}{1-q_{11}q_{10}}(q_{11}q_{10})^m\right) \\ &+ (1-\alpha_0)\left(1 + q_{01}\frac{d\underline{\alpha}_0}{d\overline{\alpha}}\right)\left(\frac{1+q_{00}}{1-q_{01}q_{00}} - \frac{q_{00}(1+q_{01})}{1-q_{01}q_{00}}(q_{01}q_{00})^m\right),\end{aligned}$$

where $\frac{d\underline{\alpha}_0}{d\overline{\alpha}} > 0$ and is given by (B.27). Obviously, $\frac{\partial E[N]}{\partial \overline{\alpha}} > 0$. Moreover, by (B.36) and (B.26), we have

$$\frac{dE[N]}{dC_p} = \frac{\partial E[N]}{\partial \overline{\alpha}}\frac{d\overline{\alpha}}{dC_p} + \frac{\partial E[N]}{\partial \underline{\alpha}}\frac{d\underline{\alpha}}{dC_p} < 0; \tag{B.37}$$

that is, as the waiting time is reduced, averagely the patients pay more visits in each illness episode.

Now we have shown that $\frac{dE[N]}{dC_p} < 0$ for all of the scenarios we categorized in Lemma 3.2, respectively in (B.29), (B.30), (B.31), and (B.37), and hence we have

$$\frac{d\lambda}{dC_p} = \frac{dE[N]}{dC_p}\Lambda < 0. \tag{B.38}$$

Now we look into the relations between C_p and the direct charge f. We have

$$\frac{dC_p}{df} = 1 + c\frac{dw}{df},$$

where

$$\frac{dw}{df} = \frac{1}{(\mu-\lambda)^2}\frac{d\lambda}{dC_p}\left(1+c\frac{dw}{df}\right) = w^2\frac{d\lambda}{dC_p}\left(1+c\frac{dw}{df}\right).$$

We then obtain the following equations:

$$\frac{dw}{df} = \frac{w^2\frac{d\lambda}{dC_p}}{1-cw^2\frac{d\lambda}{dC_p}},\ \frac{dC_p}{df} = \frac{1}{1-cw^2\frac{d\lambda}{dC_p}}. \tag{B.39}$$

By (B.38), we can see that

$$-1 < c\frac{dw}{df} < 0,\ 0 < \frac{dC_p}{df} < 1.$$

Proposition 3.3 is proved.

B.1.6 Proof of Proposition 3.4

By Lemma 3.1, it can be easily shown that $g_1(\alpha_0) > \widehat{\alpha}$ and $g_0(\alpha_0) < \widehat{\alpha}$, where $g_1(\cdot)$, $g_0(\cdot)$, and $\widehat{\alpha}$ are given by (3.2), (3.3), and (3.6), respectively. Moreover, by (3.5), we obtain

$$r(g_1(\alpha_0)) = g_1(\alpha_0)V_1 - (1-g_1(\alpha_0))L_0,$$
$$r(g_0(\alpha_0)) = (1-g_0(\alpha_0))V_0 - g_0(\alpha_0)L_1.$$

Therefore, (3.23) is simplified as

$$R^u = [q_{11}\alpha_0 V_1 - q_{01}(1-\alpha_0)L_0 + q_{00}(1-\alpha_0)V_0 - q_{10}\alpha_0 L_1](\overline{\alpha}-\underline{\alpha}).$$

Let

$$R := q_{11}\alpha_0 V_1 - q_{01}(1-\alpha_0)L_0 + q_{00}(1-\alpha_0)V_0 - q_{10}\alpha_0 L_1;$$

R is a constant. Define a function as follows: $G(x,\ y) := R(y-x)$. Obviously, $R^u = G(\underline{\alpha},\ \overline{\alpha})$.

Since when (3.14) is satisfied, no patient doctor shops, and hence $R^u = R^u_{ds}$. In the following analysis, we focus on the condition that (3.14) is not satisfied, where there are doctor-shopping patients.

First, we consider the symmetric-error scenario, i.e., $q_{11} = q_{00}$. We take a pessimistic patient as an example. Denote $v(N|\alpha \in (\overline{\alpha_0}, \overline{\alpha}))$ and $v_0(N|\alpha \in$

$(\overline{\alpha_0}, \overline{\alpha}))$ as the perceived and objective reward of the patients starting from α and reaching to the stopping set after N visits, respectively. The probability mass function is given in (B.16). We can easily show that the corresponding $v_0(N|\alpha \in (\overline{\alpha_0}, \overline{\alpha}))$ is given as

$$\begin{cases} v_0(N = 2i+1|\alpha \in (\overline{\alpha_0}, \overline{\alpha})) \\ \qquad\qquad = g_1 \circ (g_1 \circ g_0)^i(\alpha_0)V_1 - (1 - g_1 \circ (g_1 \circ g_0)^i(\alpha_0))L_0 \\ v_0(N = 2i+2|\alpha \in (\overline{\alpha_0}, \overline{\alpha})) \\ \qquad\qquad = (1 - g_0 \circ g_0 \circ (g_1 \circ g_0)^i(\alpha_0))V_0 - g_0 \circ g_0 \circ (g_1 \circ g_0)^i(\alpha_0)L_1 \end{cases},$$

where $i \geq 0$. Since $g_1 \circ g_0(\alpha) = \alpha$, we can simplify the above equation as

$$\begin{cases} v_0(N = 2i+1|\alpha \in (\overline{\alpha_0}, \overline{\alpha})) = g_1(\alpha_0)V_1 - (1 - g_1(\alpha_0))L_0 \\ v_0(N = 2i+2|\alpha \in (\overline{\alpha_0}, \overline{\alpha})) = (1 - g_0 \circ g_0(\alpha_0))V_0 - g_0 \circ g_0(\alpha_0)L_1 \end{cases}. \quad \text{(B.40)}$$

Similarly, we obtain the objective reward of an optimistic patient, $v_0(N|\alpha \in (\underline{\alpha}, \underline{\alpha_0}))$, where

$$\begin{cases} v_0(N = 2i+1|\alpha \in (\underline{\alpha}, \underline{\alpha_0})) = (1 - g_0(\alpha_0))V_0 - g_0(\alpha_0)L_1 \\ v_0(N = 2i+2|\alpha \in (\underline{\alpha}, \underline{\alpha_0})) = g_1 \circ g_1(\alpha_0)V_1 - (1 - g_1 \circ g_1(\alpha_0))L_0 \end{cases}, \quad \text{(B.41)}$$

where $i \geq 0$. The corresponding probability are given by (B.16) and (B.17), respectively. By (B.16), (B.17), (B.40), and (B.41), we have

$$\begin{aligned} R^u_{ds} &= G(\underline{\alpha_0}, \overline{\alpha_0}) + \sum_{N=0}^{\infty} v_0(N|\alpha \in (\overline{\alpha_0}, \overline{\alpha}))P(N|\alpha \in (\overline{\alpha_0}, \overline{\alpha}))(\overline{\alpha} - \overline{\alpha_0}) \\ &\quad + \sum_{N=0}^{\infty} v_0(N|\alpha \in (\underline{\alpha}, \underline{\alpha_0}))P(N|\alpha \in \alpha \in (\underline{\alpha}, \underline{\alpha_0}))(\underline{\alpha_0} - \underline{\alpha}) \\ &= G(\underline{\alpha_0}, \overline{\alpha_0}) \\ &\quad + (\overline{\alpha} - \overline{\alpha_0}) \left(g_1(\alpha_0)V_1 - (1 - g_1(\alpha_0))L_0\right) \left(\frac{q_{11}\alpha_0}{1 - q_{11}q_{10}} + \frac{q_{01}(1 - \alpha_0)}{1 - q_{00}q_{01}}\right) \\ &\quad + (\overline{\alpha} - \overline{\alpha_0}) \left((1 - g_0 \circ g_0(\alpha_0))V_0 - g_0 \circ g_0(\alpha_0)L_1\right) \left(\frac{q^2_{10}\alpha_0}{1 - q_{11}q_{10}} + \frac{q^2_{00}(1 - \alpha_0)}{1 - q_{00}q_{01}}\right) \\ &\quad + (\underline{\alpha_0} - \underline{\alpha}) \left((1 - g_0(\alpha_0))V_0 - g_0(\alpha_0)L_1\right) \left(\frac{q_{10}\alpha_0}{1 - q_{11}q_{10}} + \frac{q_{00}(1 - \alpha_0)}{1 - q_{00}q_{01}}\right) \\ &\quad + (\underline{\alpha_0} - \underline{\alpha}) \left(g_1 \circ g_1(\alpha_0)V_1 - (1 - g_1 \circ g_1(\alpha_0))L_0\right) \left(\frac{q^2_{11}\alpha_0}{1 - q_{11}q_{10}} + \frac{q^2_{01}(1 - \alpha_0)}{1 - q_{00}q_{01}}\right). \end{aligned}$$

and $g_1(\cdot)$, $g_0(\cdot)$, and $g_1 \circ g_1(\cdot)$ are given in (3.2), (3.3), and (B.6), respectively. Moreover,

$$g_0 \circ g_0(\alpha) = \frac{q^2_{10}\alpha}{q^2_{10}\alpha + q^2_{00}(1 - \alpha)}.$$

The above R^u_{ds} can be written as

$$\begin{aligned}
R^u_{ds} &= G(\underline{\alpha_0}, \overline{\alpha_0}) \\
&\quad +(\overline{\alpha} - \overline{\alpha_0})\frac{q_{11}\alpha_0 V_1 - q_{01}(1-\alpha_0)L_0}{q_{11}\alpha_0 + q_{01}(1-\alpha_0)}\left(\frac{q_{11}\alpha_0}{1-q_{11}q_{10}} + \frac{q_{01}(1-\alpha_0)}{1-q_{00}q_{01}}\right) \\
&\quad +(\overline{\alpha} - \overline{\alpha_0})\frac{q^2_{00}(1-\alpha_0)V_0 - q^2_{10}\alpha_0 L_1}{q^2_{10}\alpha_0 + q^2_{00}(1-\alpha_0)}\left(\frac{q^2_{10}\alpha_0}{1-q_{11}q_{10}} + \frac{q^2_{00}(1-\alpha_0)}{1-q_{00}q_{01}}\right) \\
&\quad +(\underline{\alpha_0} - \underline{\alpha})\frac{q_{00}(1-\alpha_0)V_0 - q_{10}\alpha_0 L_1}{q_{10}\alpha_0 + q_{00}(1-\alpha_0)}\left(\frac{q_{10}\alpha_0}{1-q_{11}q_{10}} + \frac{q_{00}(1-\alpha_0)}{1-q_{00}q_{01}}\right) \\
&\quad +(\underline{\alpha_0} - \underline{\alpha})\frac{q^2_{11}\alpha_0 V_1 - q^2_{01}(1-\alpha_0)L_0}{q^2_{11}\alpha_0 + q^2_{01}(1-\alpha_0)}\left(\frac{q^2_{11}\alpha_0}{1-q_{11}q_{10}} + \frac{q^2_{01}(1-\alpha_0)}{1-q_{00}q_{01}}\right)
\end{aligned}$$

By $q_{11} = q_{00}$,

$$\begin{aligned}
R^u_{ds} =& G(\underline{\alpha_0}, \overline{\alpha_0}) \\
&+ \frac{\overline{\alpha} - \overline{\alpha_0}}{1-q_{11}q_{10}}[q_{11}\alpha_0 V_1 - q_{01}(1-\alpha_0)L_0 + q^2_{00}(1-\alpha_0)V_0 - q^2_{10}\alpha_0 L_1] \\
&+ \frac{\underline{\alpha_0} - \underline{\alpha}}{1-q_{11}q_{10}}[q_{00}(1-\alpha_0)V_0 - q_{10}\alpha_0 L_1 + q^2_{11}\alpha_0 V_1 - q^2_{01}(1-\alpha_0)L_0]
\end{aligned} \tag{B.42}$$

Taking difference between R^u, given by (B.42), and R^u_{ds}, given by (3.23), we obtain

$$\begin{aligned}
R^u_{ds} - R^u &= \frac{q_{11}q_{10}}{1-q_{11}q_{10}}(\overline{\alpha} - \overline{\alpha_0})[q_{11}\alpha_0(V_1+L_1) - q_{10}(1-\alpha_0)(V_0+L_0)] \\
&\quad + \frac{q_{11}q_{10}}{1-q_{11}q_{10}}(\underline{\alpha_0} - \underline{\alpha})[q_{11}(1-\alpha_0)(V_0+L_0) - q_{10}\alpha_0(V_1+L_1)] \\
&= \frac{q_{11}q_{10}[\overline{\alpha}(1-\underline{\alpha})q_{10} - \underline{\alpha}(1-\overline{\alpha})q_{11}]}{1-q_{11}q_{10}}X
\end{aligned}$$

where it follows (B.9) and (B.10) from the first "=" to the second "=", and

$$\begin{aligned}
X &= \frac{q_{11}\alpha_0(V_1+L_1) - q_{10}(1-\alpha_0)(V_0+L_0)}{\underline{\alpha}q_{11} + (1-\underline{\alpha})q_{10}} \\
&\quad + \frac{q_{11}(1-\alpha_0)(V_0+L_0) - q_{10}\alpha_0(V_1+L_1)}{\overline{\alpha}q_{10} + (1-\overline{\alpha})q_{11}}.
\end{aligned}$$

By (3.12) and (3.13), we have

$$\begin{aligned}
X &= \frac{q_{11}\alpha_0(V_1+L_1) - q_{10}(1-\alpha_0)(V_0+L_0)}{q_{11}q_{10}(L_0+V_0+L_1+V_1) + (q_{11}-q_{10})C_p}(q_{11}(L_1+V_1) + q_{10}(L_0+V_0)) \\
&\quad + \frac{q_{11}(1-\alpha_0)(V_0+L_0) - q_{10}\alpha_0(V_1+L_1)}{q_{11}q_{10}(L_0+V_0+L_1+V_1) + (q_{11}-q_{10})C_p}(q_{11}(L_0+V_0) + q_{10}(L_1+V_1)). \\
&= \frac{(q_{11}-q_{10})(\alpha_0(V_1+L_1)^2 + (1-\alpha_0)(V_0+L_0)^2)}{q_{11}q_{10}(L_0+V_0+L_1+V_1) + (q_{11}-q_{10})C_p} > 0.
\end{aligned}$$

The sign of $R^u_{ds} - R^u$ is the same as that of X, and hence, $R^u_{ds} - R^u > 0$. That is, doctor shopping improves objective reward under this scenario.

When $q_{11} > q_{00}$ and $\frac{\underline{\alpha}(1-\overline{\alpha})}{\overline{\alpha}(1-\underline{\alpha})} \geq \frac{q_{10}}{q_{00}}$, the patients whose illness perceptions fall into interval $[\underline{\alpha_0}, \overline{\alpha})$ visit only once, whereas optimistic patients whose illness perceptions fall into interval $(\underline{\alpha}, \underline{\alpha_0})$ pay a second visit if they obtain positive results at the first visit, and they terminate the visiting process if they obtain negative results. We obtain

$$\begin{aligned} R^u_{ds} =& G(\underline{\alpha_0}, \overline{\alpha}) + p(s_1 = 0|\alpha_0) r(g_0(\alpha_0)) \Lambda(\underline{\alpha_0} - \underline{\alpha}) \\ &+ p(s_1 = 1, s_2 = 1|\alpha_0) r(g_1 \circ g_1(\alpha_0)) \Lambda(\underline{\alpha_0} - \underline{\alpha}) \\ &+ p(s_1 = 1, s_2 = 0|\alpha_0) r(g_1 \circ g_0(\alpha_0)) \Lambda(\underline{\alpha_0} - \underline{\alpha}) \end{aligned}$$

where

$$\begin{aligned} p(s_1 = 1, s_2 = 1|\alpha_0) =& q_{11}^2 \alpha_0 + q_{01}^2 (1 - \alpha_0), \\ p(s_1 = 1, s_2 = 0|\alpha_0) =& q_{11} q_{10} \alpha_0 + q_{00} q_{01} (1 - \alpha_0) \end{aligned}$$

By Lemma 3.2, an optimistic patient would leave being reassured as not sick after a positive and a negative result at the first and the second visit, respectively, and hence, the reward she brings to the system is

$$r(g_1 \circ g_0(\alpha_0)) = (1 - g_1 \circ g_0(\alpha_0)) V_0 - g_1 \circ g_0(\alpha_0) L_1;$$

moreover,

$$r(g_1 \circ g_1(\alpha_0)) = g_1 \circ g_1(\alpha_0) V_1 - (1 - g_1 \circ g_1(\alpha_0)) L_0.$$

Remember that $g_1 \circ g_0(\cdot)$ and $g_1 \circ g_1(\cdot)$ are given by (B.12), with $n = 1$, and (B.8), respectively. Taking difference between R^u and R^u_{ds}, we obtain

$$\begin{aligned} & R^u_{ds} - R^u \\ &= [p(s_1 = 1, s_2 = 1|\alpha_0) r(g_1 \circ g_1(\alpha_0)) + p(s_1 = 1, s_2 = 0|\alpha_0) r(g_1 \circ g_0(\alpha_0)) \\ &\quad - p(s_1 = 1|\alpha_0) r(g_1(\alpha_0))](\underline{\alpha_0} - \underline{\alpha}) \\ &= [(q_{11}^2 \alpha_0 V_1 - q_{01}^2 (1 - \alpha_0) L_0) + (q_{00} q_{01} (1 - \alpha_0) V_0 - q_{11} q_{10} \alpha_0 L_1) \\ &\quad - (q_{11} \alpha_0 V_1 - q_{01} (1 - \alpha_0) L_0)](\underline{\alpha_0} - \underline{\alpha}) \\ &= [q_{00} q_{01} (1 - \alpha_0)(V_0 + L_0) - q_{11} q_{10} \alpha_0 (V_1 + L_1)](\underline{\alpha_0} - \underline{\alpha}). \end{aligned}$$

Therefore, if $q_{00} q_{01} (1 - \alpha_0)(V_0 + L_0) - q_{11} q_{10} \alpha_0 (V_1 + L_1) > 0$, which is equivalent to (3.25), $R^u_{ds} - R^u > 0$.

B.2 Supplement: Derivation of R^u_{ds} and R^p_{ds}

Following the thought of Proof of Proposition 3.3, we categorize patients according to the maximum visits they need until they leave the system eventually under the worst-case scenario. Since the patients apply OSLA rule repeatedly until there is no advantage of paying another visit, we denote $v(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}))$ as the *perceived* reward of the pessimistic patients starting from α (where $\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})$) and reaching to the stopping set after N visits; the *objective* reward is hence $v_0(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}))$. Similar for the optimistic patients.

First, consider the objective reward R^u_{ds}. Specially, we consider a pessimistic patient whose illness perception satisfies $\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})$, where $n = 1, ...m$, and m is given in (B.24). By using the results of Proof of Proposition 3.2, we obtain

$$\begin{cases} v_0(N = 2i+1|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) \\ \qquad = g_1 \circ (g_1 \circ g_0)^i(\alpha_0)V_1 - (1 - g_1 \circ (g_1 \circ g_0)^i(\alpha_0))L_0 \\ v_0(N = 2n+2|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) \\ \qquad = (1 - g_0 \circ g_0 \circ (g_1 \circ g_0)^i(\alpha_0))V_0 - g_0 \circ g_0 \circ (g_1 \circ g_0)^i(\alpha_0)L_1 \\ v_0(N = 2n+1, +|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) \\ \qquad = g_1 \circ (g_1 \circ g_0)^n(\alpha_0)V_1 - (1 - g_1 \circ (g_1 \circ g_0)^n(\alpha_0))L_0 \\ v_0(N = 2n+1, -|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) \\ \qquad = (1 - g_0 \circ (g_1 \circ g_0)^n(\alpha_0))V_0 - g_0 \circ (g_1 \circ g_0)^n(\alpha_0)L_1 \end{cases}, \tag{B.43}$$

where $0 \leq i < n$, and the third and fourth rows of (B.43) are the rewards of the patient when she leaves as positive and negative after $2n+1$ visits, the worst-case scenario, respectively. Moreover, the probabilities of the patient visiting $N = 2i, 2i+1, 2n+1$ times are given in (B.20).

Similarly, we obtain the reward of the optimistic patients, which is

$$\begin{cases} v_0(N = 2i+1|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) \\ \qquad = (1 - g_0 \circ (g_1 \circ g_0)^i(\alpha_0))V_0 - g_0 \circ (g_1 \circ g_0)^i(\alpha_0)L_1 \\ v_0(N = 2i+2|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) \\ \qquad = g_1 \circ g_1 \circ (g_1 \circ g_0)^i(\alpha_0)V_1 - (1 - g_1 \circ g_1 \circ (g_1 \circ g_0)^i(\alpha_0))L_0 \\ v_0(N = 2n+2, +|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) \\ \qquad = g_1 \circ g_1 \circ (g_1 \circ g_0)^n(\alpha_0)V_1 - (1 - g_1 \circ g_1 \circ (g_1 \circ g_0)^n(\alpha_0))L_0 \\ v_0(N = 2n+2, -|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) \\ \qquad = (1 - (g_1 \circ g_0)^{n+1}(\alpha_0))V_0 - (g_1 \circ g_0)^{n+1}(\alpha_0)L_1 \end{cases}, \tag{B.44}$$

where $0 \leq i < n$, and the third and fourth rows of (B.44) are the rewards of the patient when she leaves as positive and negative after $2n+2$ visits, the worst-case scenario, respectively. The probabilities of each visiting times of the optimistic patients are given in (B.21). Moreover,

$$\begin{aligned} R^u_{ds} &= G(\underline{\alpha_0}, \overline{\alpha_0}) \\ &+ \sum_{n=1}^{m} \sum_{N=1}^{2n+1} v_0(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) * P(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}))(\alpha_{2n+1} - \alpha_{2n-1}) \\ &+ \sum_{n=1}^{m+1} \sum_{N=1}^{2n} v_0(N|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) * P(N|\alpha \in (\alpha_{2n-2}, \alpha_{2n}))(\alpha_{2n} - \alpha_{2n-2}). \end{aligned}$$

Next, we consider R^p_{ds}. By substituting α_0 with α in (B.43) and (B.44), we obtain $v(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1}))$ and $v(N|\alpha \in (\alpha_{2n-2}, \alpha_{2n}))$, respectively . Moreover,

$$\begin{aligned} R^p_{ds} &= F(\underline{\alpha_0}, \overline{\alpha_0}) \\ &+ \sum_{n=1}^{m} \int_{\alpha_{2n-1}}^{\alpha_{2n+1}} \left[\sum_{N=1}^{2n+1} v(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) * P(N|\alpha \in (\alpha_{2n-1}, \alpha_{2n+1})) \right] d\alpha \\ &+ \sum_{n=1}^{m+1} \int_{\alpha_{2n-2}}^{\alpha_{2n}} \left[\sum_{N=1}^{2n+2} v(N|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) * P(N|\alpha \in (\alpha_{2n-2}, \alpha_{2n})) \right] d\alpha. \end{aligned}$$

where

$$\begin{aligned} F(x,\ y) =& \frac{q_{01}(1-\alpha_0) + q_{11}\alpha_0}{q_{11} - q_{01}} (q_{11}V_1 + q_{01}L_0)(y - x) \\ &- \frac{q_{01}(1-\alpha_0) + q_{11}\alpha_0}{q_{11} - q_{01}} \frac{q_{11}q_{01}(V_1 + L_0)}{q_{11} - q_{01}} \ln\left(\frac{q_{01} + (q_{11} - q_{01})y}{q_{01} + (q_{11} - q_{01})x} \right) \\ &+ \frac{q_{00}(1-\alpha_0) + q_{10}\alpha_0}{q_{00} - q_{10}} (q_{00}V_0 + q_{10}L_1)(y - x) \\ &+ \frac{q_{00}(1-\alpha_0) + q_{10}\alpha_0}{q_{00} - q_{10}} \frac{q_{00}q_{10}(V_0 + L_1)}{q_{00} - q_{10}} \ln\left(\frac{q_{00} - (q_{00} - q_{10})y}{q_{00} - (q_{00} - q_{10})x} \right) \end{aligned}.$$

Through simple algebra transformation of (3.24), we can obtain that $R^p = F(\underline{\alpha},\ \overline{\alpha})$. Moreover, when $q_{11} = q_{00}$,

$$\begin{aligned} R^p_{ds} &= F(\underline{\alpha_0}, \overline{\alpha_0}) + \left(\frac{q_{11}\alpha_0}{1 - q_{11}q_{10}} + \frac{q_{01}(1-\alpha_0)}{1 - q_{00}q_{01}} \right) \int_{\overline{\alpha_0}}^{\overline{\alpha}} \frac{q_{11}\alpha V_1 - q_{01}(1-\alpha)L_0}{q_{11}\alpha + q_{01}(1-\alpha)} d\alpha \\ &+ \left(\frac{q_{10}^2\alpha_0}{1 - q_{11}q_{10}} + \frac{q_{00}^2(1-\alpha_0)}{1 - q_{00}q_{01}} \right) \int_{\overline{\alpha_0}}^{\overline{\alpha}} \frac{q_{00}^2(1-\alpha)V_0 - q_{10}^2\alpha L_1}{q_{10}^2\alpha + q_{00}^2(1-\alpha)} d\alpha \end{aligned}$$

$$+\left(\frac{q_{10}\alpha_0}{1-q_{11}q_{10}}+\frac{q_{00}(1-\alpha_0)}{1-q_{00}q_{01}}\right)\int_{\underline{\alpha}}^{\underline{\alpha_0}}\frac{q_{00}(1-\alpha)V_0-q_{10}\alpha L_1}{q_{10}\alpha+q_{00}(1-\alpha)}d\alpha$$
$$+\left(\frac{q_{11}^2\alpha_0}{1-q_{11}q_{10}}+\frac{q_{01}^2(1-\alpha_0)}{1-q_{00}q_{01}}\right)\int_{\underline{\alpha}}^{\underline{\alpha_0}}\frac{q_{11}^2\alpha V_1-q_{01}^2(1-\alpha)L_0}{q_{11}^2\alpha+q_{01}^2(1-\alpha)}d\alpha.$$

References

Afeche, P., B. Ata. 2013. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management*, 15(2), 292-304.

Alizamir, S., de Vericourt, F., Sun, P. 2013. Diagnostic accuracy under congestion. *Management Science*, 59(1), 157-171.

Allon G., A. Bassamboo, I. Gurvich. 2011. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research*, 59, 1382-1394.

Aral, S., Dellarocas, C., Godes, D. 2013. Introduction to the special issuesocial media and business transformation: a framework for research. *Information Systems Research*, 24(1), 3-13.

Armony M., C. Maglaras. 2004. Contact center with a call-back option and real-time delay information. *Operations Research*, 52, 527–545.

Baron, O., Hu, M., Najafi-Asadolahi, S., Qian, Q. 2015. Newsvendor selling to loss-averse consumers with stochastic reference points. *Manufacturing & Service Operations Management*, 17(4), 456-469.

Billinghurst, B., Whitfield, M. 1993. Why do patients change their general practitioner? A postal questionnaire study of patients in Avon. *British Journal of General Practice*. 43 (373), 336338.

Camerer, C. F., Ho, T. H., Chong, J. K. 2004. A cognitive hierarchy model of games. The Quarterly Journal of Economics, 119(3), 861-898.

Chan, C. W., Farias, V. F., Bambos, N., Escobar, G. J. 2011. Maximizing throughput of hospital intensive care units with patient readmissions. Work-

ing Paper.

Chan, C. W., Yom-Tov, G., Escobar, G. 2014. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2), 462-482.

Chandra, A., Cutler, D., Song, Z. 2011. Who ordered that? The economics of treatment choices in medical care. *In Handbook of health economics* (Vol. 2, pp. 397-432). Elsevier.

Chen, H., Wan, Y. W. 2003. Price competition of make-to-order firms. *IIE Transactions*, 35(9), 817-832.

Crapis, D., Ifrach, B., Maglaras, C., Scarsini, M. 2016. Monopoly pricing in the presence of social learning. *Management Science*, 63(11), 3586-3608.

Cui, S., S. K. Veeraraghavan. 2014. Blind queues: The impact of consumer beliefs on revenues and congestion. Available at SSRN 2196817.

de Vericourt, F., Sun, P. 2009. Judgement accuracy under congestion in service systems. Fuqua School of Business. Working Paper. Duke University.

de Vericourt, F., Zhou, Y. P. 2005. Managing response time in a call-routing problem with service failure. *Operations Research*, 53(6), 968-981.

Debo, L., C. Parlur, U. Rajan. 2012. Signaling quality via queues. *Management Science*, 58(5), 876-891.

Debo, L., S. Veeraraghavan. 2014. Equilibrium in queues under unknown service times and service value. *Operations Research*, 62(1), 38-57.

Dewan, S., H. Mendelson. 1990. User delay costs and internal pricing for a service facility. *Management Science*, 36(12), 1502-1517.

Dellarocas, C. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10), 1577-1593.

Dobson G., E. Pinker. 2006. The value of sharing lead time information. *IIE Transactions*,38, 171-183.

Donkin, L., Ellis, C. J., Powell, R., Broadbent, E., Gamble, G., Petrie, K. J. 2006. Illness perceptions predict reassurance following a negative exercise stress testing result. *Psychology and Health*, 21(4), 421-430.

Economou, A., A. Gomez-Corral, S. Kanta. 2011. Optimal balking strategies in single-server queues with general service and vacation times. *Performance Evaluation*. 68(10), 967-982.

Edleson, N., D. Hildebrand. 1975. Congestion tolls for Poisson queueing process. *Econometrica*, 43, 81–92.

George, J. M., Harrison, J. M. 2001. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5), 720-731.

Guo, Y., Kuroki, T., Yamamoto, S., Koizumi, S., 2002. Illness behavior and patient satisfaction as correlates of self-referral in Japan. *Family Practice*. 19, 326332.

Guo, P., Tang, C. S., Wang, Y., Zhao, M. The Impact of Reimbursement Policy on Patient Welfare, Readmission Rate and Waiting Time in a Public Healthcare System: Fee-for-Service vs. Bundled Payment. Working paper. The Hong Kong Polytechnic University, Hong Kong.

Guo, P., Lindsey, R., Zhang, Z. G. 2014. On the Downs-Thomson paradox in a self-financing two-tier queuing system. *Manufacturing & Service Operations Management*, 16(2), 315-322.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53, 962-970.

Guo, P., R. Hassin. 2011. Strategic behavior and social optimization in Markovian vacation queues. *Operations Research*, 59(4), 986-997.

Guo, P., R. Hassin. 2012. Strategic behavior and social optimization in Markovian vacation queues: the case of heterogeneous customers. *European Journal of Operational Research*, 222(2), 278-286.

Guo, P., Q. Li. 2013. Strategic behavior and social optimization in partially-observable Markovian vacation queues. *Operations Research Letters*, 41(3), 277-284.

Guo, P., W. Sun, Y. Wang. 2011. Equilibrium and optimal strategies to join a

queue with partial information on service times. *European Journal of Operational Research*, 214(2), 284-297.

Guo, P., Z. Zhang. 2013. Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management*, 15(1), 118-131.

Guo, P., M. Haviv, Y. Wang. 2015. Equilibrium queueing strategies when service quality is unknown to some customers. Working paper. The Hong Kong Polytechnic University, Hong Kong.

Hagihara, A., Tarumi, K., Odamaki, M., Nobutomo, K. 2005. A signal detection approach to patientdoctor communication and doctorshopping behaviour among Japanese patients. *Journal of evaluation in clinical practice*, 11(6), 556-567.

Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica*, 54, 1185-1195.

Hassin, R. 2016. *Rational Queueing*. CRC Press.

Hassin, R., M. Haviv. 1994. Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Stochastic Models*, 10, 415–435.

Hassin, R., Haviv, M. 2003. To queue or not to queue: Equilibrium behavior in queueing systems. Kulwer Academic Publisher, Nowerll, MA.

Hassin, R., R. Roet-Green. 2011. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly. Working paper. Tel Aviv University, Israel.

Hassin, R., R. Roet-Green. 2013. Cascade strategies in a partially-observable queueing system with parallel servers. Working paper. Tel Aviv University, Israel.

Harris, K.M., 2003. How do patients choose physicians? Evidence from a national survey of enrollees in employment-related health plans. *Health Service Research*. 38 (2), 711732.

Hasija, S., Pinker, E. J., Shumsky, R. A. 2005. Staffing and routing in a two-tier call centre. *International Journal of Operational Research*, 1(1-2), 8-29.

Kasteler, J., Kane, R. L., Olsen, D. M., Thetford, C. 1976. Issues underlying prevalence of "doctor-shopping" behavior. *Journal of Health and Social Behavior*, 328-339.

Katon, W., Von Korff, M., Lin, E., Bush, T., Russo, J., Lipscomb, P., Wagner, E. 1992. A randomized trial of psychiatric consultation with distressed high utilizers. *General hospital psychiatry*, 14(2), 86-98.

Hu, M., Li, Y., Wang, J. 2017. Efficient ignorance: Information heterogeneity in a queue. *Management Science*, 64(6), 2650-2671.

Huang, T., G. Allon, A. Bassamboo. 2013. Bounded rationality in service systems. *Manufacturing & Service Operations Management*, 15(2), 263-279.

Huang, T., Y. Chen. 2015. Service systems with experience based anecdotal reasoning customers. *Production and Operations Management*, 24(5), 778-790.

Leung, G. M., Yeung, R. Y. T., Wong, I. O. L., Castan-Cameo, S., Johnston, J. M. 2006. Time costs of waiting, doctor-shopping and privatepublic sector imbalance: Microdata evidence from Hong Kong. *Health Policy*, 76(1), 1-12.

Li, X., P. Guo, Z. Lian. 2016. Quality-speed competition in customer-intensive cervices with boundedly rational customers. *Production and Operations Management*, *Forthcoming*.

Linnet, K. 1988. A review on the methodology for assessing diagnostic tests. Clinical chemistry, 34(7), 1379-1386.

Lo, A., A. Hedley, G. Pei, S. Ong, L. Ho, R. Fielding, L. Daniel. 1994. Doctor-shopping in Hong Kong: implications for quality of care. *International Journal for Quality in Health Care*, 6(4), 371-381.

Macpherson, A., M. Kramer, F. Ducharme, H. Yang, F. Blanger. 2001. Doctor shopping before and after a visit to a paediatric emergency department. *Paediatrics & Child Health*. 6(6), 341.

Mayzlin, D. 2006. Promotional chat on the Internet. *Marketing Science*, 25(2), 155-163.

Mayzlin, D., Dover, Y., Chevalier, J. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8), 2421-55.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, 15-24.

Pac, M. F., Veeraraghavan, S. 2010. Strategic diagnosis and pricing in expert services. Working paper, Wharton School, University of Pennsylvania, Philadelphia.

Petrie, K. J., Jago, L. A., Devcich, D. A. 2007. The role of illness perceptions in patients with medical conditions. *Current Opinion in Psychiatry*, 20(2), 163-167.

Qian, Q., Guo, P., Lindsey, R. 2017. Comparison of Subsidy Schemes for Reducing Waiting Times in Healthcare Systems. *Production and Operations Management*, 26(11), 2033-2049.

Sato T, M. Takeichi, M. Shiraham, T. Fukui, J. Gude. 1995. Doctor shopping patients and users of alternative medicine among Japanese primary care patients. *General Hospital Psychiatry*, 17:115-25.

Shin, D., Zeevi, A. 2017. Dynamic pricing and learning with online product reviews. Working paper, Columbia University.

Sun, M. 2012. How does the variance of product ratings matter?. *Management Science*, 58(4), 696-707.

Stidham, S. 1992. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science*, 38(8), 1121–1139.

Stidham, S. 2009. *Optimal Design of Queueing Systems*. CRC Press, London.

Veeraraghavan, S., L. Debo. 2009. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management*, 11(4), 543-562.

Veeraraghavan, S. K., L. Debo. 2011. Herding in queues with waiting costs: Rationality and regret. *Manufacturing & Service Operations Management*, 13(3), 329-346.

Wang, J., J. Li. 2008. A repairable M/G/1 retrial queue with Bernoulli vacation and two-phase service. *Quality Technology and Quantitative Management*, 5(2), 179-192.

Wang, X., Debo, L. G., Scheller-Wolf, A., Smith, S. F. 2010. Design and analysis of diagnostic service centers. *Management Science*, 56(11), 1873-1890.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Science*, 45, 192-207.

Yang, L., Guo, P., Wang, Y. 2018. Service Pricing with Loss-Averse Customers. *Operations Research*, *Forthcoming*.

Yeung, R. Y., Leung, G. M., McGhee, S. M., Johnston, J. M. 2004. Waiting time and doctor shopping in a mixed medical economy. *Health Economics*, 13(11), 1137-1144.

Yom-Tov, G. B., Mandelbaum, A. 2014. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2), 283-299.

Zhang, F., J. Wang, B. Liu. 2013. Equilibrium balking strategies in Markovian queues with working vacations. *Applied Mathematical Modelling*, 37(16), 8264-8282.