Three Essays in Operations Management

by

Xin Geng

B.Sc. (Mathematics), Zhejiang University, 2008
M.Sc. (Mathematics), The University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Business Administration)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2015

Abstract

There are three topics in operations management presented in this dissertation. Each topic deals with a specific issue encountered by managers from various organizations.

In the context of non-profit operations, we study a two-customer sequential resource allocation problem whose objective function has a max-min form. For finite discrete demand distribution, we give a sufficient and necessary condition under which the optimal solution has monotonicity property. However, this property never holds with unbounded discrete distributions.

Then, we look at a service system with two servers serving arriving single class jobs. Servers care about fairness, and they can endogenously choose capacities in response to the routing policy. We focus on four commonly seen policies and examine the two-server game where the servers' objective functions have a term that reflects fairness. Theoretical results concerning the existence and uniqueness of the Nash equilibrium are proved for some policies. Numerical studies also provide insights on servers' off-equilibrium behaviours and the system efficiency under different policies.

Finally, suppose that a firm has heterogeneous servers who provide service with different quality levels, and that there exists a learning curve of the servers so that the quality can be improved by accumulating experience in serving customers. As customers decide their service procurement based on the quality and system congestion, what pricing scheme should the firm adopt to achieve optimal revenue in the long run? We compare a traditional pricing scheme with a proposed one, and theoretically establish the superiority of the proposed pricing scheme. Based on both theoretical and numerical evidence, we characterize the sensitivity of some parameters with respect to the comparison.

Preface

Chapter 2 and Chapter 3 are co-authored by my supervisors Tim Huh and Mahesh Nagarajan. In both chapters, I made the main contribution, which includes identifying the topics, developing the models, carrying out the analysis and presenting them. A version of Chapter 2 has been published (Geng, Huh, and Nagarajan, 2014a); Chapter 3 is forthcoming (Geng, Huh, and Nagarajan, 2014b). Chapter 4 is authored by myself, but is also greatly benefited from my supervisors' insightful comments. This chapter will be modified and submitted for publication in academic peer reviewed journals.

Table of Contents

List of Tables

List of Figures

Acknowledgements

The research I have been involved for the past five years has been a great treasure for me. This dissertation, although hefty in weight, is merely a small portion of the product of those research activities. The largest and the most precious portion, however, is the issuing of an ongoing professional and personal growth. Much of this growth is credited to the vast resources in the Operations and Logistic Division in Sauder School of Business at UBC. Indeed, I am greatly thankful to many past/present faculties, staffs and graduate students in the division, who have been altogether my friends, teachers and mentors.

I want to extend my gratitude to my dear supervisors, Tim Huh and Mahesh Nagarajan. From the very beginning, before I have any clue, they have designed an academic training for me that eventually turns out to be quite successful. Moreover, they have endured my stupidity and patiently supervised me in the course of all my research projects. I also appreciate all the encouraging words they spoke to me to keep me positive. Their generous financial support, even in my last year, is definitely a priceless contribution that I am deeply indebted to. In all, they have been very inspiring role models to me. I wish that our friendship continues and matures beyond my time at UBC.

I also would like to thank Maurice Queyranne, Hong Chen, Martin Puterman and Steven Shechter for offering the fundamental courses during my first year, which are quite helpful to me. In addition, I thank Harish Krishnan, Yichuan Ding and Hao Zhang for their insightful comments on my research; I thank Danielle van Jaarsveld for her willingness to serve as a dissertation committee member; and I thank Anming Zhang for caring me and always super warmly greeting me in the hall way.

My final yet most important gratitude goes to my loving family. My wife has supported me in every possible ways and sacrificed a lot so that I could pursue the doctoral degree. Her willingness to talk about my research and even discuss ideas with me is truly precious to me. My parents also loved me without holding back. Even though they are thousands miles away from me, I am certain that their hearts are with me and mine with them.

Dedication

To my wife 于湉
and my parents 耿全合, 王亞萍

Chapter 1

Introduction

The area of operations management sees many interesting and challenging problems, which attracts both researchers and practitioners. It is indeed an area where practical issues resort to theoretical resolution and the development of relevant theories is driven by real-life applications. In this dissertation, we focus on three topics in operations management that arise in practice. Each topic deals with a specific issue (resource allocation; employee fairness; pricing strategies) that is encountered by managers from either non-profit or profit-seeking organizations. We apply analytical and numerical tools and methodologies to tackle these issues in order to gain useful managerial insights. Moreover, we try to approach an operations management problem from an interdisciplinary standpoint, where concepts and assumptions from other management areas, e.g. organizational behaviour and marketing, are introduced and utilized.

In the first part of this dissertation, we study a two-customer sequential resource allocation problem in the context of non-profit operations. Different from a profit maximizing goal, the non-profit organization's objective is to make the minimum fill rate (ratio of allocated amount to demand) of all agencies as large as possible. Given that the resource is limited and the agencies' demands are random and arrive sequentially in time, the manager's decision is subject to a trade off: If he satisfies the current demand, then there may not be enough resource for the later demand, which could be large; if he reserve much resource, then the later demand may be small and the left-over resource would be wasted.

For this sequential resource allocation problem with a max-min objective function, we are primarily interested in the structural properties of the optimal allocation policy, which is fully characterized by the allocation to the first agency as a function of its demand realization. Particularly, we ask the following question: If the realized demand is higher, do we always allocate more? For finite discrete demand distribution, we give a sufficient and necessary condition under which the optimal solution has monotonicity property. However, this property never holds with unbounded discrete distributions.

The second part of this dissertation looks at a simple service system with two servers serving arriving jobs (single class). Our interest is in examining the effect of routing policies on servers when they care about fairness among themselves, and when they can endogenously choose capacities in response to the routing policy. Therefore, we study the two-server game where the servers' objective functions have a term explicitly modeling fairness. To mathematically measure fairness, we turn to the equity theory, which was established in organizational behavior and human resource literature. The application of equity theory in our setting epitomizes the

interesting research area at the interface of OM and OBHR.

Focusing on four commonly seen policies that are from one general class, we theoretically prove the existence and uniqueness of the Nash equilibrium under two policies. By numerical methods, we also draw several general conclusions concerning servers' off-equilibrium behaviours under the other two policies. In addition, we obtain some empirical patterns/properties for the game outcome under these policies. Finally, we study, again numerically, how servers' attitudes towards fairness and servers' heterogeneity affect a policy's performance in system efficiency.

Finally, we study the revenue management problem for a service firm facing time-sensitive customers. The firm has heterogeneous servers who provide quality-differentiated services, where the quality can be improved based on accumulated experience. That is, servers learn by doing and improve quality as serving more customers. Risk neutral customers decide whether to procure the service based on the expected quality and congestion cost. In such settings, the traditional pricing scheme, which we call differentiated pricing, posts prices on each servers and customers queue in front of their chosen servers. One unavoidable pitfall of this pricing scheme is that the low quality servers will suffer from slow improvement due to limited customer volume induced. To resolve this issue and take advantage of the servers' learning effect, we propose another pricing scheme, called uniform pricing, and prove that it collects equal or higher revenue compared to differentiated pricing. Furthermore, we scrutinize the impact of servers' learning speed and their heterogeneity on the relative revenue advantage. Interesting and counter-intuitive results are shown by analytical and numerical approaches.

The proposed uniform pricing scheme is essentially a type of probabilistic (opaque) selling strategy, which has recently been extensively studied in the marketing literature. By probabilistic (opaque) selling, the firm hides one or more attributes of the product until the transaction is complete. It has been commonly used in travel industries (hotel and airline) where customers make purchases via an intermediary platform, e.g. Priceline and Hotwire, without prior knowledge of the brand name. In our study, the service attribute that uniform pricing hides from the customers is the quality. Our work is among the first to introduce probabilistic (opaque) selling to time-sensitive service settings; besides, we provide a new reason to advocate this new pricing strategy.

Each of these three parts is self-contained and is developed based on different sets of assumptions. The literature reviews, model formulations and conclusions are not repeated in separate parts. Yet, they all represent some interesting directions in the research body of operations management. The rest of the dissertation is organized so that each part is presented in one chapter, in the same order as in this chapter. All proofs are delayed to Appendix A.

Chapter 2

Sequential Resource Allocation With Constraints: Two-Customer Case

2.1 Background

The sequential resource allocation (SRA) problem has received much attention in literature. In this problem, a supplier has a limited but known quantity of resource available for allocation. Independent random demands arrive sequentially from a number of customers (or agencies), and the supplier needs to sequentially allocate the resource for each customer at a time. When allocating resource to a customer, the supplier sees the realization of the customer's demand, but not the realization of remaining demands (the supplier knows only the distributions of the remaining demands). In other words, the decision of how much to allocate to the current customer is made one at a time. Since the total amount of the resource is limited, the trade-off is usually whether to allocate it to the current customer or save it for future demands.

Two types of objectives are commonly studied in this research area. The first type involves maximizing profit/revenue. The single resource capacity allocation problem in revenue management is a good example; see the first chapter of Talluri and Van Ryzin (2005) for the detailed study on the theoretical properties as well as useful heuristics. One specific such problem studies airline seat allocation to customers with different fare classes. Brumelle and McGill (1993) and Robinson (1995) have provided a structural characterization for this problem, and approximate solutions have been proposed, such as EMSR-a Belobaba (1987) and EMSR-b Belobaba (1989).

The second type of objective in SRA problem does not explicitly model monetary pay-off. In such contexts, government-run companies or non-profit organizations often play the role of supplier, and they aim at enhancing the satisfaction level of the overall society rather than profit. While there are few papers taking such a perspective compared to the more typical profit-maximizing objective, there have been a number of papers advocating the necessity and urgency of studying non-profit logistics in the past couple of decades. In the context of public service such as education, emergent medical care and library supplies, Savas (1978) proposes and studies three measures of performance: effectiveness (how well the service is rendered), efficiency (ratio of service outputs to inputs) and equity (fairness or impartiality of service). In fact, the SRA problem arises naturally in the contexts such as healthcare allocation and food distribution. For example, Campbell et al. (2008) stress effectiveness by optimizing the arrival times of the relief efforts in the aftermath of server disasters; Solak et al. (2012), focusing

on efficiency, minimize the service inputs of non-profit food distribution networks under a constraint of fixed service output.

In particular, Savas (1978) further claims that equity, among others, is of growing importance that deserves more attention. This statement is still valid today, especially for resource allocation problems; Bertsimas et al. (2012) study a class of efficiency-fairness objective functions and indicate applications in several contexts involving allocating resources. However, only a limited number of papers incorporate fairness into the decision making on allocating capacities. For instance, Alkan et al. (1991) explicitly bring up and discuss the fairness in allocation of indivisible goods. They consider a static (instead of sequential) allocation problem and they include two kinds of resources. In the sequential allocation context, Swaminathan (2003) studies the decision making in allocating scarce drugs, attempting to equalize the distribution of drugs amongst clinics without undermining the effectiveness and efficiency. More recently, Lien Lien (2008) models food donation delivery scheme over multiple customers as an SRA problem, and incorporates into the objective both equity and effectiveness. In this paper, we focus our attention to SRA problems with equity (fairness) as the objective, and we refer to our problem the *equity based* SRA problem.

As one may imagine, the term equity is amorphous and its measurement varies according to the context. There are numerous criteria of fairness in allocation problems proposed in many literatures. Although no single criterion is universally accepted in every setting, there are some general theories on justice and fairness that serve as the basis of most fairness measures; see Bertsimas et al. (2012). We discuss three of them used in welfare economics. In welfare economics, the social welfare function (SWF) is a function from the vector space of feasible utility combinations of all the agents to the real numbers. A general form of SWF that incorporates fairness measure is introduced in Atkinson (1970) and studied in many classic textbooks Barr (1993); Mas-Collel et al. (1995). It is parametrized by an real number $\rho \leq 1$. To be specific, let $U = (u_1, \cdots, u_n)$ be a utility vector of n agents, then this SWF has the form

$$W(U) = \left(\sum_i u_i^\rho \right)^{1/\rho} \quad \text{for} \quad \rho \neq 0, \tag{2.1}$$

and

$$W(U) = \sum_i \log u_i \quad \text{for} \quad \rho = 0.$$

Searching for an allocation of utility to maximize $W(U)$ is closely related to the three general theories and fairness criteria. The first one is utilitarianism. The key idea is simply to maximize the summation of utilities, which corresponds to the case $\rho = 1$ in the above general SWF. This criterion is criticised by some scholars (Young, 1995) as having ethical issues. Indeed, although it maximizes average utility, utilitarianism sometimes manipulates the allocation among agents in a way that is contrary to the common sense of fairness. The second general theory originates from game theory. Nash (1950) uses an axiomatic approach to characterize a cooperative two-

player bargaining game. The Nash solution, which is proposed in Nash (1950), is to maximize the product of all the utilities (assuming that all utilities are positive). Hence, the general form of SWF with $\rho = 0$ is exactly this criterion. This criterion is discussed in many literatures. For example, Chevaleyre et al. (2005) call it Nash product and consider it better than utilitarianism in achieving fairness; Kelly et al. (1998) and Bertsimas et al. (2011) name it proportional fairness and analyze it in telecommunication settings and resource allocation problems, respectively. The last general theory is proposed by Rawls (1971); hence the name *Rawlsian justice*. The principle of Rawlsian justice is to make the minimum utility as large as possible. Note that when $\rho \to -\infty$ in (2.1), $W(U)$ equals the minimum utility. Therefore, maximizing the SWF in this case is to apply Rawlsian justice. This max-min criterion has been widely used in data network (Bertsekas et al., 1992) and has initiated applications in bandwidth allocation problems (Bonald and Massoulié, 2001; Luss, 1999) as well as general resource allocation problems (Alkan et al., 1991; Lien et al., 2008).

In general, the different values of ρ indicate the different levels of inequity. The aversion to inequity increases as ρ decreases to negative infinity (Bertsimas et al., 2012; Lan et al., 2010; Mas-Collel et al., 1995). Hence, in the above three general theories, Rawlsian justice retains the most fairness. Moreover, Chevaleyre et al. (2005) define the worst utility as *egalitarian social welfare*, and they claim that it "offers a level of fairness and may be a suitable performance indicator when we have to satisfy the minimum needs of a large number of customers" (p.17). This fits to our setting of non-profit food allocation very well. Consequently, we will use a max-min objective which is in line with the commonly used Rawlsian justice. In other words, our objective function is based on (2.1) with $\rho \to -\infty$. We will consider more general cases in Section 2.3.1 to show that our result also holds for finite negative ρ values; but we will not consider the cases $0 \leq \rho \leq 1$. Furthermore, we model the customer's utility as the ratio of the allocated amount to the demand, which is named *fill rate*. Fill rate captures the proportion of demand satisfied for each customer, and customers tend to compare this measure with one another after allocation is completed. Indeed, both Lien (2008) and Swaminathan (2003) adopt this measurement in their models, and we believe it is a suitable candidate and we employ it in our model as well.

The SRA problem over multiple (more than two) periods is hard to solve in exact form due to the curse of dimensionality. Many related works in literature search for reasonable heuristics by studying the two-period special case. The doctoral thesis Lien (2008) is an example and is closest to our work. Lien formulates the SRA problem with an equity based objective as a dynamic program and develop some structural properties for the two-period case. Then, he proposes a "two-node-decomposition" heuristic which solves the large size problem by decomposing it into several two-period subproblems. Using the structural results discovered, he manages to solve those subproblems. In this paper we explore the two-customer case in great detail. We present an exact expression of the optimal solution when demand follows a discrete distribution. Our purpose here is to study the properties of this special case as it is important for gaining insight

and generating heuristics - our results can be used, for example, in each two-period subproblem of Lien's decomposition heuristic.

Therefore, our work adds to this body of research by providing more theoretical results concerning the basic structure of the optimal solutions. The limitation of two-customer special case notwithstanding, our results provide a reference for further theoretical studies as well as heuristic development. In addition, our contribution differs from that of the theoretical work on profit-based SRA, e.g. Brumelle and McGill (1993); since the objective functions have different forms and properties, their results or methods are not directly applicable in studying equity based SRA.

2.2 Model Formulation

In this section, we formulate the SRA model. A supplier needs to allocate to N customers with a fixed amount s of resource. The customers are sequentially ordered. Each customer's demand is random and will be known to the supplier only after all demands of the previous customers have been realized and allocation decisions to those customers have been made, but before the allocated amount to him is decided. As discussed in the previous section, we aim to maximize the minimal fill rate of all customers to achieve Rawlsian justice. Since the demands are random, our objective is therefore to maximize the expected minimum fill rate over all the customers. In this paper, we focus on the two-customer case only. Studying this special case simplifies the problem while keeping the inherent challenges of sequential decision making. Besides, it is straightforward to extend some of the main results to multiple customers. Hence, we aim at finding structural properties that help understanding the N-customer case.

As a result, the optimization problem has one decision variable: the allocation to customer 1. The second customer will receive what is left. Let x_i $(i = 1, 2)$ be the allocation to customer i. Since $x_2 = s - x_1$, x_1 is the only decision. Let D_i $(i = 1, 2)$ be the random variable representing demand from customer and d_i $(i = 1, 2)$ be the realized demand. Throughout this paper we assume that the two demands are independent (but not necessarily identical), which is commonly assumed in literature. Moreover, we assume that $D_i > 0$ $(i = 1, 2)$ almost surely. Then, the fill rate takes the form of x_i/D_i $(i = 1, 2)$.

Let $a \wedge b$ represents $\min\{a, b\}$. Given an initial supply $s > 0$ and a realized first demand d_1, define

$$R(x_1, d_1) \;=\; E_{D_2}\left(\frac{x_1}{d_1} \wedge \frac{s - x_1}{D_2}\right) \tag{2.2}$$

to be the expectation of the minimum of the two fill rates, where $0 \le x_1 \le d_1$ and the expectation is taken with respect to D_2. It is straightforward to see that function $R(x_1, d_1)$ is jointly concave in x_1 and s. Further, let

$$v(s, d_1) = \max_{0 \le x_1 \le \min\{s, d_1\}} R(x_1, d_1)\ , \tag{2.3}$$

then the optimal expected minimum fill rate is given by

$$u(s) = E_{D_1} v(s, D_1) \ .$$

Since x_1 is decided after D_1 is realized, we need only to focus our attention on the random variable D_2 and how it affects the structure of the optimal decision conditioned on the the realized value of the first demand d_1.

Our interest, therefore, is in solving (2.3). Let $x_1^* = x_1^*(d_1)$ be an optimal solution to (2.3). Note that the constraint $x_1 \leq d_1$ simply says that there is no need to give the first customer more than demanded. As a result, $v(s, d_1)$ cannot exceed 1. Equivalently, one can remove the constraint $x_1 \leq d_1$ and modify the objective function as $E_{D_2}\left(1 \wedge \frac{x_1}{d_1} \wedge \frac{s-x_1}{D_2}\right)$ (see Lien (2008) for an example). Although they are equivalent formulation, we will use the former objective function, which turns out to be easier to work with. To further simplify the problem, we enlarge the feasible region for x_1 in (2.3), and consider the following relaxed problem.

$$\tilde{v}(s, d_1) = \max_{0 \leq x_1 \leq s} R(x_1, d_1) \ . \tag{2.4}$$

Let $\phi(d_1)$ be an optimal solution to (2.4),

$$\phi(d_1) \ \in \arg\max_{0 \leq x_1 \leq s} R(x_1, d_1).$$

Comparing the two problems, the feasible region in (2.4) is not bounded by d_1. Therefore, $x_1^*(d_1)$ is at most d_1 while $\phi(d_1)$ may exceed d_1; similarly, $v(s, d_1)$ is bounded above by 1 while $\tilde{v}(s, d_1)$ may exceed 1. However, the two problems are closely related. In fact, if $\phi(d_1) \leq d_1$, then it is easy to see that the relaxed problem has an optimum that is feasible for the original problem so $x_1^*(d_1) = \phi(d_1)$. Otherwise, if $s \geq \phi(d_1) > d_1$, then by concavity of $R(x_1, d_1)$, the objective function is increasing in x_1 on the interval $[0, \phi(d_1)]$. Hence it is also increasing on $[0, d_1]$, which implies that the optimum of the original problem is achieved at the boundary d_1, i.e., $x_1^*(d_1) = d_1$. To summarize, we always have $v(s, d_1) \leq \tilde{v}(s, d_1)$ and

$$x_1^*(d_1) \ = \ \phi(d_1) \wedge d_1. \tag{2.5}$$

For our SRA problem, we are primarily interested in the optimal allocation policy rather than the optimal value of the objective function. To understand the structure of $x_1^*(d_1)$, it is convenient for us to first study $\phi(d_1)$ in the relaxed problem (2.4) since we can recover the information about the optimal allocation through (2.5). To the best of our knowledge, this method of relaxation has not been used in such equity based SRA problem before. It bears noting that it is possible that, for fixed s, $\phi(d_1)$ has multiple values for a specified d_1. This is because the objective function is not strictly concave. However, we need $\phi(d_1)$ to be a function of d_1 so that we can perform analysis on the structural properties. To circumvent the difficulty, we have to predefine which optimal value $\phi(d_1)$ takes if there are many. One way of doing this

is to assign the smallest value of all the optima to $\phi(d_1)$.

From the formulation of problem (2.4), we make several observations. First, the predetermined distribution of D_2 plays an important role in that it completely determines $\phi(d_1)$ with fixed s and d_1. Second, if we increase the supply s, we not only increase the objective value but also increase the feasible region. It then follows that the optimal value $v(s, d_1)$ is increasing in s. This makes sense in the context. The more the initial supply is, the more we can give to satisfy the two demands, and thus the larger the fill rates. Third, if the realized value d_1 become higher, then $R(x_1, d_1)$ and subsequently $\tilde{v}(s, d_1)$ become lower. Other comparative statics results may not be obvious, and one of the questions that we are interested is: what happens to the optimal allocation as d_1 increases? In other words, if the realized demand d_1 is higher, do we always allocate more to the first customer? This is one of the research questions analysed in this paper, and is discussed in the next section. All detailed proofs of our results are presented in Appendix A.1.

2.3 Structure of Optimal Solution

Since the optimal allocation x_1^* depends on the realization of the first demand d_1 and the initial supply s, we now study the structure of $x_1^*(d_1)$. To simplify analysis and notation, we suppose the supply s is fixed throughout this section, and omit s in the functions defined in Section 2.2; for example, we write $x_1^*(d_1)$ instead of $x_1^*(d_1)$. We then ask how x_1^* depends on d_1. Preliminary intuition drives us to believe that the larger d_1 is, the more we should allocate to it, resulting in larger x_1^*. This is asserted as being true in Lien (2008) and later corrected in Lien et al. (2008). In this section, we provide another example where this is not true and provide sufficient conditions under which this assertion is true.

2.3.1 Example: Non-Monotonicity of $x_1^*(d_1)$

Suppose the second period demand D_2 is such that $D_2 = 1$ or 4 with equal probability. Let the initial supply $s = 1$. We then compute the optimal allocation for two cases, in which $d_1 = 3$ and $d_1 = 5$, respectively. Using (2.2) and some algebra, we have

$$\begin{aligned} R(x_1, 3) &= E_{D_2}\left(\frac{x_1}{3} \wedge \frac{1 - x_1}{D_2}\right) \\ &= \frac{1}{2}\left(\frac{x_1}{3} \wedge (1 - x_1)\right) + \frac{1}{2}\left(\frac{x_1}{3} \wedge \frac{1 - x_1}{4}\right) . \end{aligned}$$

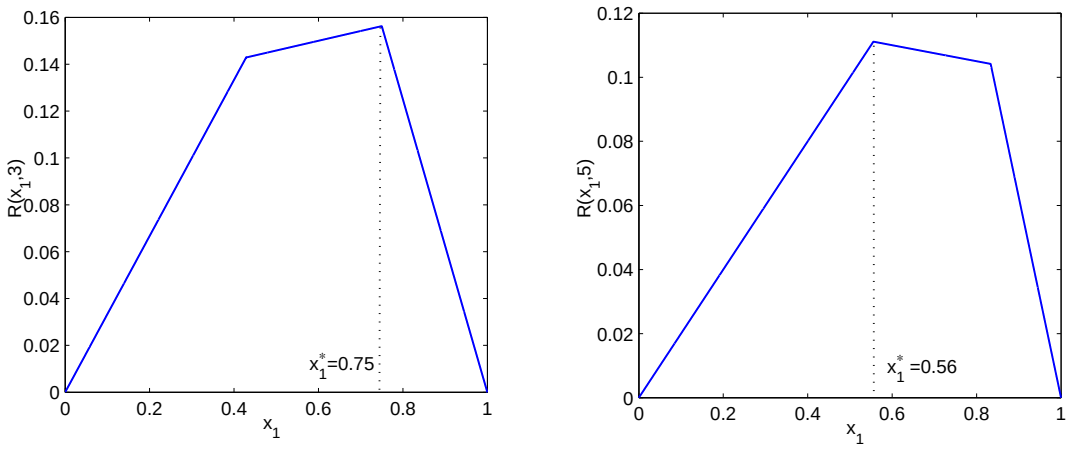
Noting that the two expressions in the above big parentheses are piecewise linear, we can write $R(x_1, 3)$ as

$$R(x_1,3) \;=\; \begin{cases} \dfrac{x_1}{3} & 0 \le x_1 < \dfrac{3}{7} \\ \dfrac{x_1}{24} + \dfrac{1}{8} & \dfrac{3}{7} \le x_1 < \dfrac{3}{4} \\ \dfrac{5}{8}(1 - x_1) & \dfrac{3}{4} \le x_1 < 1 \end{cases}.$$

Therefore, maximizing $R(x_1, 3)$ over the interval $[0, 1]$ yields

$$x_1^*(3) \;=\; \frac{3}{4} \;=\; 0.75.$$

See Figure 2.1 for an illustration.



(a) Objective Function with $d_1 = 3$: $R(x_1, 3)$ (b) Objective Function with $d_1 = 5$: $R(x_1, 5)$

Figure 2.1: Illustration of optimal allocations with different d_1

Likewise we can write $R(x_1, 5)$ as

$$R(x_1,5) = \begin{cases} \dfrac{x_1}{5} & 0 \le x_1 < \dfrac{5}{9} \\ -\dfrac{x_1}{40} + \dfrac{1}{8} & \dfrac{5}{9} \le x_1 < \dfrac{5}{6} \\ \dfrac{3}{8}(1 - x_1) & \dfrac{5}{6} \le x_1 < 1 \end{cases}$$

and obtain

$$x_1^*(5) \;=\; \frac{5}{9} \;\approx\; 0.56.$$

Hence, $x_1^*(5) < x_1^*(3)$. This shows that $x_1^*(d_1)$ may not be increasing in d_1. Interestingly, similar examples can be constructed to show the non-monotonicity using the general form of objective function (2.1) with $\rho < 0$. Some of these examples are as follows.

The general form of the objective function is given by

$$R(x_1, d_1) = E_{D_2}\left(\left(\frac{x_1}{d_1} \wedge 1\right)^{\rho} + \left(\frac{s - x_1}{D_2} \wedge 1\right)^{\rho}\right)^{1/\rho}$$

Consider the function $W(U)$ in (2.1) and the case $n = 2$. By verifying the second order condition, we see $W(U)$ is concave. Besides, it is also increasing in each component. $R(x_1, d_1)$ is the expectation of a composition of $W(U)$ and two concave functions. Hence it is concave in x_1 on $[0, s]$. Therefore, we can define $x_1^*(d_1)$ in the same way as we did in Section 2.2. Now we construct examples which show that $x_1^*(d_1)$ is not necessarily increasing in d_1. Suppose D_2 is uniform on $\{1, 4\}$, and $s = 1$. We then plot the $x_1^*(d_1)$ over $1 \leq d_1 \leq 5$ for three different ρ values ($\rho = -6, -8, -10$). See Figure 2.2. A non-monotonic trend can be clearly seen in

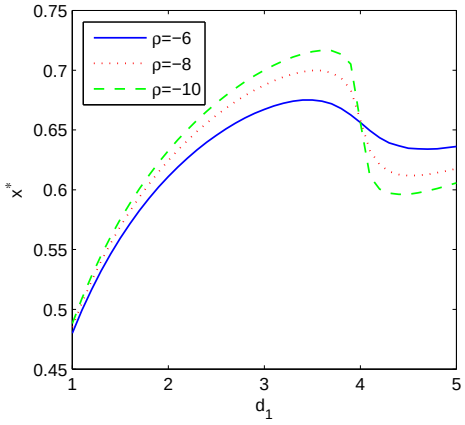


Figure 2.2: Illustration for non-monotonicity of $x_1^*(d_1)$ in d_1 with three different ρ values. Second customer demand D_2 is discrete uniform on $\{1, 4\}$, and $s = 1$.

all three curves, indicating that $x_1^*(d_1)$ is not necessarily increasing in d_1. In fact, as the ρ value decreases the non-monotonicity remains, and the curve becomes more alike the one where $\rho \to -\infty$.

The above examples indicate that the relation between x_1^* and d_1 may be counter-intuitive. This result depends on the distribution of D_2 and the amount of the initial supply. If the resource is limited in amount, then the supplier needs to make a choice between allocating to the first customer and saving for the later demand. As a result, the supplier may increase x_1^* as d_1 gets larger but may also want to reserve more for the second customer as d_1 increases to some level, leading to a decrease of x_1^*. This trade-off is affected by the distribution of second demand. Sometimes it is better to increase x_1^* whereas other times saving more for the second demand is optimal even when d_1 is increasing. Generally speaking, for a distribution whose density function has several "bumps", i.e. points with similarly large densities, it is likely that x_1^* is not overall monotonic in d_1. In the above example, D_2 follows a discrete distribution, which is an extreme case where densities become point masses. However, we want to stress that

demand being discrete is not the inherent reason for non-monotonicity. Examples showing the non-monotonicity of $x_1^*(d_1)$ can also be constructed using a continuous distribution for D_2 by following a similar intuition. Under discrete distributions the analysis is cleaner and we keep this assumption for the remainder of this section. We will investigate conditions under which monotonicity holds.

2.3.2 Bounded Discrete Distribution

This subsection studies the case where D_2 has a bounded discrete distribution. Throughout this subsection, we make the following assumption, which is satisfied by many commonly used distributions such as discrete uniform and binomial.

Assumption 2.3.1. *D_2 has a discrete distribution with finitely many realizations.*

Let the probability mass function of D_2 be given by

$$\mathbf{P}(D_2 = a_k) = p_k, \qquad k = 1, 2, \cdots, n.$$

We assume that

$$0 < a_1 < a_2 < \cdots < a_n. \tag{2.6}$$

Once the distribution is given, we want to write the objective function in an explicit way. First, if x_1 has been allocated to the first customer, then the minimum fill rate conditioned on the realized value $D_2 = a_k$ $(k = 1, 2, \cdots, n)$ is given by

$$f_k(x_1, d_1) \;=\; \left(\frac{x_1}{d_1} \wedge \frac{s - x_1}{D_2}\right)\Big|_{D_2 = a_k} \;=\; \frac{x_1}{d_1} \wedge \frac{s - x_1}{a_k}.$$

As a minimum of two linear functions with different slopes, each f_k is a piecewise linear concave function of x_1. In addition, the break point that connects the two pieces is exactly the x_1 that makes the two expressions for the minimum operator equal, i.e, $x_1/d_1 = (s - x_1)/a_k$. It can be shown that the break point is given by

$$z_k = z_k(d_1) = \frac{s d_1}{d_1 + a_k}\ . \tag{2.7}$$

Note that while each break point z_k is a function of d_1, it is bounded above by s. Thus, z_k can be interpreted as the allocation amount that equalizes the fill rates of both customers in case of $D_2 = a_k$. Now, having defined $\{z_k\}$, we can write f_k explicitly as

$$f_k(x_1, d_1) = \begin{cases} \dfrac{x_1}{d_1} & \text{if } 0 \le x_1 \le z_k \\ \dfrac{s - x_1}{a_k} & \text{if } z_k \le x_1 \le s\ . \end{cases} \tag{2.8}$$

Thus, we obtain that f_k is increasing on $[0, z_k]$ and decreasing on $[z_k, s]$, and $x_1 = z_k = \dfrac{s d_1}{d_1 + a_k}$

maximizes $f_k(x_1, d_1)$ for each k.

We now proceed to the objective function in problem (2.4), by considering D_2 as a random variable. Taking the expectation with respect to D_2, we obtain

$$R(x_1, d_1) \;=\; \sum_{k=1}^{n} p_k f_k(x_1, d_1).$$

Recall that the value of d_1 is realized before x_1 is decided. Since every f_k is a piecewise linear concave function of x_1, their convex combination $R(x_1, d_1)$ is also a piecewise linear concave function in x_1. Moreover, the set of all the break points of $R(x_1, d_1)$ is exactly $\{z_k : k = 1, 2, \cdots, n\}$. Since the sequence of a_k's is increasing in k, it follows from (2.7) that the break points are strictly monotonic with

$$0 < z_n < \cdots < z_1 < s \;. \tag{2.9}$$

It is convenient to define $z_0 = s$ and $z_{n+1} = 0$. Then, we can derive the analytic form of each linear piece by expanding (2.8) for every $k = 1, 2, \cdots, n+1$:

$$R(x_1, d_1) \;=\; \left[\left(\sum_{j=1}^{k-1} p_j \right) \frac{1}{d_1} - \sum_{j=k}^{n} \frac{p_j}{a_j} \right] x_1 + \sum_{j=k}^{n} \frac{s}{a_j}, \tag{2.10}$$

$$\text{for } z_k \le x_1 < z_{k-1}.$$

We are interested in how the optimal solution depends on the value of d_1. To examine the impact of d_1 on the maximizer of $R(x_1, d_1)$, we first find the optimal solution $\phi(d_1)$ of problem (2.4). There may exist multiple optimal solutions, in which case we predefine only one of them in order that $\phi(d_1)$ is uniquely defined as a function d_1; let

$$\phi(d_1) \;=\; \inf \arg \max_{0 \le x_1 \le s} R(x_1, d_1).$$

Due to the concavity and piecewise linearity of $R(x_1, d_1)$, we know that the point whose left derivative is positive and right derivative is non-positive achieves the maximum, and is $\phi(d_1)$. To find such a point, we examine the slope of each linear piece in (2.10) in the order of $k = n+1, n, \cdots, 1$, and look for the first one that is less or equal to zero. Suppose the piece corresponding to $[z_k, z_{k-1})$ is the first one, then $\phi(d_1) = z_k$ by our predefined specification. (Note that $k \neq n+1$ because the leftmost piece has strictly positive slope.) Hence, the optimal solution $\phi(d_1)$ must be one of the n break points $\{z_k\}$, i.e.,

$$\phi(d_1) = z_k(d_1), \qquad \text{for some } k = 1, 2, \cdots, n.$$

Hence, knowing how d_1 affects z_k directly helps us know how d_1 affects $\phi(d_1)$. From the above reasoning, the choice of the optimal z_k depends on the signs of the slopes of the linear

pieces. From (2.10), we see that the slope is closely related to d_1. To reveal how the value of d_1 determines the slopes, we introduce a sequence of thresholds for d_1, each of which sets the corresponding slope in (2.10) to zero. Define

$$\theta_k = \frac{\sum_{j=1}^{k-1} p_j}{\sum_{j=k}^{n} p_j/a_j}, \qquad \text{for } k = 2, 3, \cdots, n. \tag{2.11}$$

It is straightforward to verify that if $d_1 < \theta_k$, then $R(x_1, d_1)$ is increasing in x_1 on $[z_k, z_{k-1})$; if $d_1 > \theta_k$, then $R(x_1, d_1)$ is decreasing in x_1 on $[z_k, z_{k-1})$. If $d_1 = \theta_k$, then $R(x_1, d_1)$ is constant for x_1 in $[z_k, z_{k-1})$.

Note that these threshold points depend only on the distribution of D_2, and therefore are predetermined. From the definition of θ_k's in (2.11), it can be seen that as we increase d_1, the slope of $R(x_1, d_1)$ in the interval $[z_k, z_{k-1})$ changes its sign as d_1 crosses θ_k. This is significant in the following two ways: (i) a small change in d_1, not crossing any of the θ_k thresholds, keeps the sign of $R(x_1, d_1)$'s slope unchanged (either negative or positive) within $[z_k, z_{k-1})$ even though the exact value of z_k and z_{k-1} changes; (ii) when the increasing d_1 crosses θ_k, the sign of $R(x_1, d_1)$'s slope in the interval $[z_k, z_{k-1})$ changes (from positive to negative), and the maximizer of $R(x_1, d_1)$ shifts from z_{k-1} to z_k. We formalize these observations below.

Define $\theta_1 = 0$ and $\theta_{n+1} = \infty$. It follows directly from the definition in (2.11) that

$$0 = \theta_1 < \theta_2 < \cdots < \theta_n < \theta_{n+1} = \infty. \tag{2.12}$$

The following lemma characterizes $\phi(d_1)$, the optimal solution to (2.4) as a function of the realized demand of the first customer. The signs "+" and "−" in the function expression means the right and left limits, respectively.

Lemma 2.3.2. *Given Assumption 2.3.1, the function $\phi(d_1)$ is concave and increases in d_1 for $d_1 \in [\theta_k, \theta_{k+1})$. In addition, for $k = 2, 3, \cdots, n$, $\phi(d_1)$ is discontinuous at each $d_1 = \theta_k$, and we have $\phi(\theta_k-) > \phi(\theta_k+)$.*

To visually illustrate Lemma 2.3.2, we give Figure 2.3 as below. Suppose that $\mathbf{P}(D_2 = i) = 1/3$ for $i = 1, 2, 3$, and $s = 1$. As above lemma shows, that the relaxed solution $\phi(d_1)$ is well-behaved except at some predetermined threshold points. As d_1 increases, $\phi(d_1)$ is both increasing and concave in each of the interval; however, at each threshold, $\phi(d_1)$ makes a downward jump.

As discussed in Section 2.2, the optimal allocation is related to the relaxed solution $\phi(d_1)$ through (2.5), i.e.

$$x_1^*(d_1) = \phi(d_1) \wedge d_1.$$

Then, from Lemma 2.3.2, we immediately have the following corollary.

Corollary 2.3.3. *Given Assumption 2.3.1, the optimal allocation $x_1^*(d_1)$ is a piecewise continuous function of d_1; moreover, it is increasing and concave in d_1 on each piece. However, at*

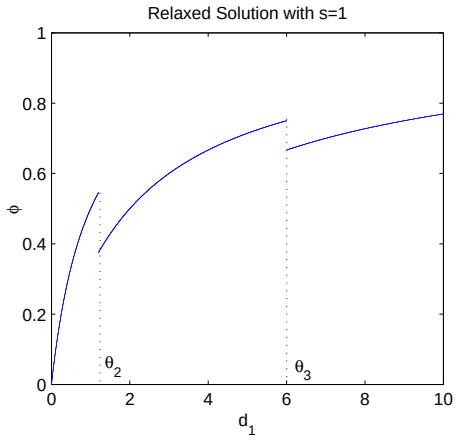


Figure 2.3: Illustration for discontinuities in $\phi(d_1)$ as a function of d_1. Second customer demand D_2 is discrete uniform on $\{1, 2, 3\}$, and $s = 1$.

every discontinuity, if there is any, $x_1^*(d_1+) < x_1^*(d_1-)$.

From the last statement in Corollary 2.3.3, it is shown that the monotonicity of $x_1^*(d_1)$ in d_1 requires no discontinuity in $x_1^*(d_1)$. The question now is when exactly there is no discontinuity, i.e., when $x_1^*(d_1)$ is overall increasing and concave in d_1. From (2.5), x_1^* is the smaller of (i) $\phi(d_1)$ and (ii) the function $x_1^* = d_1$, whose image is the 45° line in a demand-allocation graph (e.g., the dot line in Figure 2.4). This 45° line represents the level of allocation that fully satisfies the first customer; therefore this line will be referred to as *fulfilment line* henceforth. Obviously, the fulfilment line is continuous and increasing. Hence, roughly speaking, we want all the $n - 1$ discontinuous points of $\phi(d_1)$ to lie on or above the fulfilment line, so that they will not be part of x_1^*.

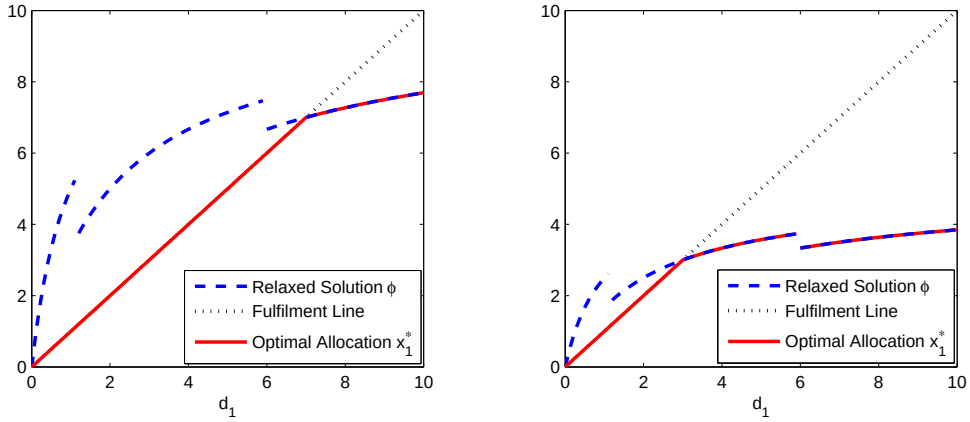
Figure 2.4 shows an example picture that illustrates our point. Again, we use the distribution $\mathbf{P}(D_2 = i) = 1/3$ for $i = 1, 2, 3$. For the left graph (Case A), we use $s = 10$ while for the right graph (Case B), $s = 5$. As noted, the optimal allocation is the lower part of the relaxed solution and fulfilment line. Hence, x_1^* is continuous in Case A but not in Case B.

The following theorem identifies conditions for the monotonicity and concavity of $x_1^*(d_1)$ in d_1.

Theorem 2.3.4. *Under Assumption 2.3.1, a necessary and sufficient condition for* $x_1^*(d_1)$ *to be overall increasing and concave in* d_1 *is that the initial supply* s *satisfies*

$$s \geq \theta_n + a_n. \tag{2.13}$$

This theorem shows that when the initial supply is sufficiently large, the optimal allocation is strictly increasing with the first customer's demand. Consider what would happen to the amount, x_1, allocated to customer 1 as the demand d_1 increases. The supplier balances the impact of the following two actions: (i) increasing x_1 to allocate more to the first customer,



(a) Case A: $s = 10$ (b) Case B: $s = 5$

Figure 2.4: Illustration for the monotonicity and lack of monotonicity. Second customer demand D_2 is discrete uniform on $\{1, 2, 3\}$, and $s = 10$ or $s = 5$.

and (ii) reserving more for the second customer by decreasing x_1. (See our discussion following the definitions of θ_k's in (2.11)). Theorem 2.3.4 simply states that as long as the initial supply is sufficient enough to meet the first demand up to θ_n (the highest level of d_1 for the supplier to consider the effect of (ii)) and the second demand up to a_n (the largest realization of D_2), respectively, the supplier can always allocate more to the first customer as d_1 increases. Conversely, the reason why the allocation x_1 could decrease even as d_1 is increasing is that the initial supply is limited – since the supplier's action makes a more significant impact on customer 2's fill rate.

Alternately, the condition in Theorem 2.3.4, equation (2.13), can be written as following using equation (2.11):

$$s \geq \left(\frac{\sum_{j=1}^{n-1} p_j}{p_n} + 1 \right) a_n.$$

Let $\beta = \frac{\sum_{j=1}^{n-1} p_j}{p_n} + 1$. Then, β is a constant that is completely determined by the distribution of D_2. Hence, the inequality $s \geq \beta a_n$ reveals the interrelationship between the supply and the second distribution in affecting the structure of optimal allocation as a function of the first demand. Furthermore, under many commonly seen distributions, the factor β is likely to be more than 2 because p_n is generally smaller than $1 - p_n$. Therefore, if we let D_1 and D_2 be identically distributed and $\beta > 2$, then the above condition means that the initial supply is sufficient enough to meet both customers' largest demands, which is why the supplier will not have to worry about facing a large demand in the future while he increases x_1.

In summary, Theorem 2.3.4 tells us sufficient initial supply leads to an increasing optimal allocation function. Besides, the distribution of D_2 also plays a key role in that it determines both θ_n and β. Note that condition (2.13) is possible to hold for some s because the realized

demands are bounded; cf. Section 2.3.3.

Recall from the definition of x_1^* in (2.5) that x_1^* is the minimum of two functions, ϕ and the identity mapping. The condition in Theorem 2.3.4 ensures that the points of discontinuity in $\phi(d_1)$ are not the points of discontinuity in $x_1^*(d_1)$, and thus $x_1^*(d_1)$ is overall continuous. Meanwhile, we are interested to see when the function $x_1^*(d_1)$ would be determined by $\phi(d_1)$ instead, i.e., $x_1^*(d_1) = \phi(d_1)$. In this case, $x_1^*(d_1)$ becomes discontinuous at every θ_k $(k = 2, 3, \cdots, n)$. Besides, the optimal allocation stays below the fulfilment line; i.e., the first customer is never fully satisfied. Intuitively, this would happen when the second period demand is much larger than the initial supply. Because the second customer's fill rate will never reach 1, there is no need for the supplier to allocate to the first customer the full demanded amount, regardless how small it is. The above intuition is verified in the following proposition.

Proposition 2.3.5. *Suppose that Assumption 2.3.1 holds. If $s \leq a_1$, then*

$$x_1^*(d_1) = \phi(d_1) < d_1 \quad \text{for } d_1 > 0.$$

If the initial supply is neither too large nor too small compared to the second demand (thus Theorem 2.3.4 and Proposition 2.3.5 are inapplicable), we have to combine equation (2.5) and Lemma 2.3.2 to see where x_1^* is increasing in d_1 and where the discontinuous downward jump occurs.

2.3.3 Unbounded Discrete Distribution

In the earlier discussion, a_n (the maximum possible value of D_2), has played an important role in characterizing the monotonicity condition for the optimal allocation. In fact, a_n being finite is the main reason why inequality (2.13) can hold for some s. This subsection focuses on the case where the second period demand is discrete but unbounded (e.g. Poisson distribution). Again, we use this assumption throughout this subsection.

Assumption 2.3.6. *D_2 is discretely distributed, and $\mathbf{P}(D_2 > M) > 0$ for any $M \geq 0$.*

We will still use the same notations as defined in the previous subsection, with slight modifications to adapt to the unbounded case in this subsection. To be specific, instead of being finitely many, all sequences $\{a_k\}$, $\{z_k\}$ and $\{\theta_k\}$ have infinitely many points. For instance, the probability mass function of D_2 becomes

$$\mathbf{P}(D_2 = a_k) = p_k > 0, \qquad k = 1, 2, \cdots$$

with $\{a_k\}$ strictly increasing to $+\infty$. Besides, θ_k now takes the form that

$$\theta_k = \frac{\sum_{j=1}^{k-1} p_j}{\sum_{j=k}^{\infty} p_j / a_j}.$$

Note that θ_k is well-defined because the series in the denominator converges. Moreover, it can be shown that each θ_k is finite, but $\{\theta_k\}$ is an increasing and unbounded sequence of positive numbers.

Because of the infinitely many break points $\{z_k\}$ of the objective function, the function $\phi(d_1)$ now has infinitely many continuous pieces. Each of the piece is strictly increasing and concave in d_1, and each discontinuous point sees a decrease in function value. In other words, Lemma 2.3.2 basically remains the same under Assumption 2.3.6, except that there are now infinitely many θ_k's. In contrast, for the optimal allocation x_1^*, there is an interesting difference from Theorem 2.3.4. The condition (2.13) in the previous case holds for some s, whereas it can never hold for any s in the unbounded case, simply because s can never be larger than the "largest" demand realization. Therefore, the overall continuity can never happen under Assumption 2.3.6. We present it as the next theorem.

Theorem 2.3.7. *Under Assumption 2.3.6, given any initial supply s, there exists $\bar{d}_1 > 0$ such that $x_1^*(d_1)$ is discontinuous at $d_1 = \bar{d}_1$. Moreover, $x_1^*(\bar{d}_1-) > x_1^*(\bar{d}_1)$.*

The above theorem shows that the unboundedness of D_2 (Assumption 2.3.6) plays an important role. Suppose that the increment from a_{k-1} to a_k remains constant for each k. Then, it can be seen from the proof of Theorem 2.3.7 that $z_k(\theta_k) \to s$ as $d_1 \to +\infty$ for every k, we know that the discontinuity gaps of $\phi(d_1)$ become smaller as d_1 gets larger. However, because the distribution is discrete, there are always positive gaps. This leads to the inevitable discontinuity of the optimal allocation. The discussion under Theorem 2.3.4 shows that if the optimal allocation is continuous in all $d_1 > 0$, then the initial supply must be large enough compared to the second demand. Hence, under unbounded assumption, it is expected that condition (2.13) will never hold because the probability of D_2 being very large is positive and s cannot exceed it.

Meanwhile, Proposition 2.3.5 continues to hold. This is because the result depends on the relationship between the initial supply and the smallest realization of the second demand, a_1. Since a_1 is a fixed finite number, unless $a_1 = 0$, there exists some $s > 0$ such that $s \leq a_1$.

2.4 Sensitivity in Initial Supply

In the following, we briefly discuss how the change of the initial supply affects the optimal allocation. To that end, assume s is not fixed any more. As a result, all the functions defined in previous sections become functions of two variables, namely, the first demand d_1 and the initial supply s. Therefore, the optimal allocation x_1^* depends on both s and d_1 now. In the SRA setting, s is decided before any realization of demands, so the results are mainly for the sake of ex ante sensitivity analysis.

Let $x_1^*(d_1)$ and $\phi(s, d_1)$ be the optimal allocation and the optimal solution to problem (2.4), respectively. Assume that the second period demand is still discrete. We claim that under this assumption, the initial supply s can be factored out so that $\phi(s, d_1)$ is linear in s. To see this,

we consider problem (2.4). With s as another variable, the same deduction in the proof of Lemma 2.3.2 applies and result in (A.1) becomes

$$\phi(s, d_1) = z_k = \frac{sd_1}{d_1 + a_k} \text{ for some } k.$$

Due to its analytical form, we can write $\phi(s, d_1) = \phi(1, d_1)s$. Then, by equation (2.5) we have

$$x_1^*(d_1) = \phi(1, d_1)s \wedge d_1. \tag{2.14}$$

Hence, if the second demand follows a discrete distribution, then the optimal solution to problem (2.4) is separable in d_1 and s, and it is linear in s. The intuition behind the result is: since the resource is divisible, we can scale demand and allocation by the total initial supply s, and the problem becomes sequential allocation of a unit resource. To be more precise regarding $x_1^*(d_1)$ with fixed fix d_1, by (2.14), the optimal allocation is linear in s while being truncated at some level d_1. The truncation represents the case where the supplier has sufficient amount of supply with which the first demand can be fulfilled. But below that level, the allocated amount is linearly increasing in s with the slope decided by the fixed d_1; i.e. $\phi(1, d_1)$. In either case, increasing s will not decrease the optimal allocation to the first customer.

Chapter 3

Fairness Among Servers When Capacity Decisions Are Endogenous

3.1 Introduction

Many service systems are characterized by arriving requests (jobs) processed by heterogeneous servers with different service capacities. In such systems, the decision on how the incoming jobs are routed to the servers is important. These decision rules, referred to as routing policies, are often designed by an *administrator*. The administrator cares about the performance of the system, which is measured through a combination of operational metrics such as flow times, utilization, quality, etc. This drives the design of the routing policy. For instance, if the administrator's goal is to obtain a fast speed of service, then customers are likely to be sent to the fastest server available. This policy is named Fastest Server First (FSF) in the literature.

In a many-server heavy-traffic system, the FSF policy minimizes the flow times (Armony, 2005). However, FSF causes a significant distortion in the distribution of idle time among the servers. In fact, Armony (2005) shows that the FSF policy completely labors the faster servers in large-scale settings. If the servers are employees, then the notion of the more efficient employees being tasked with a higher workload can lead to perceptions of unfairness. This is a well-understood phenomenon in the human resources literature (Huseman et al., 1987), and it is often cited in operation management literature such as Armony and Ward (2010) and Cui et al. (2007).

The human resources literature has documented that such perceptions of unfairness have strong negative consequences like absenteeism and disloyalty, which can be damaging to the organization that houses these service systems. This issue is of particular significance in health-care and call center settings. The research in organizational behavior (Colquitt et al. (2001) and Cohen-Charash and Spector (2001), for example) provides strong evidence that inequitable treatment often leads to significant employee dissatisfaction. Employee dissatisfaction further impedes the system's proper functionality. In particular, Colquitt et al. (2001) compare and test several models from different studies to show the close relationship between *organizational justice* (how employees judge the behavior of the organization and their resulting attitude and behavior) and job satisfaction and commitment. Moreover, Fehr and Schmidt (1999) claim that people "are willing to give up some material pay-off to move to the direction of more equitable outcomes." In addition, Kahneman et al. (1986) point out that the administrator, despite being

profit-maximizing, will have incentive to act in a manner that is perceived as fair if she is aware that the employees care about fairness.

The importance of fairness and its eventual consequence to the performance of the employees, and in turn to the organization, has led some operations managers to explicitly consider fairness when designing their service systems. In certain instances, this may mean that system efficiency is sacrificed to maintain fairness. See Bertsimas et al. (2012) for a study on this efficiency-fairness trade-off and its application to allocation problems in several areas including call center design and healthcare scheduling.

To precisely describe the notion of fairness in our setting, we apply the traditional *equity theory* model proposed by researchers in the organizational justice area. Adams (1965) and Huseman et al. (1987) find that, as a measure of fairness, an employee compares the ratio of "outcome" to "input" with others, where the former term usually refers to the reward and the latter term refers to the effort. Huseman et al. (1987) further suggest that, if the ratio is smaller (bigger) than others, the employee will consider being under-rewarded (over-rewarded). Employees sensitive to fairness require this ratio to be close to equal across all employees. They feel distress when under-rewarded and they experience guilt when over-rewarded. Huseman et al. (1987) also show that, in cases of unfairness, employees take actions to restore equity. In a call center with the FSF policy for instance, this translates to reducing the speed at which calls are answered (Mandelbaum et al., 2010).

Following the equity theory discussed above, we model server's processing rate or capacity as the "input", which is quite natural and straightforward. However, the reward ("outcome") has several plausible candidates in our setting. Although monetary rewards accruing through wages and bonuses seem a reasonable choice, this is not the case in many contexts. Two reasons for this are institutional policies and the legacy of the organizations where these service systems reside in. For example, healthcare is unionized in many areas of the world and thus does not easily allow for differentiated rewards. Moskowitz (1983, p.827) discusses how federal legislation imposes barriers on differentiated rewards in service systems in the non-profit and healthcare areas. Another reason for the absence of monetary awards is that setting up a differentiated and dynamic reward system would impose a significant administrative and informational burden on the organization. For these reasons, we do not consider monetary pay-off in this paper. Instead, we use intrinsic process-related measures that employees care about, such as idle time. More idle time means less active work, which is an outcome of faster service in some cases. Under this reward scheme, equitable treatment translates to idle times being distributed in relation to the capacities of individual servers.

Some recent papers such as Armony and Ward (2010) have addressed the fairness issue by modeling it as an exogenous constraint. However, to the best of our knowledge, endogenizing fairness has not been attempted in the literature. To address this issue, we study a service system where the fairness issue arises internally. The servers choose their capacities in response to the routing policy in order to optimize their utility that includes a fairness component. We

believe that we are the first to *explicitly* model fairness among servers who strategically determine their capacities. Consequently, our work broadens the research area and introduces new applications. Due to the endogeneity of servers' capacities, characterizing the optimal routing policy for the administrator becomes a formidable task, which we do not intend to accomplish in this paper. Instead, we take an important first step to analyze a class of policies and several of the most commonly used ones. In particular, we are interested in how the servers react to the policies and how the policies affect the system efficiency in a decentralized environment in which the servers make decisions independently and simultaneously. Furthermore, our aim is to understand the role that fairness plays in the decision making process.

3.1.1 Literature Review

The notion of fairness arises in various organizations (Greenberg, 1987), and it is often context dependent. In a service system, two kinds of views exist on fairness: (1) fairness among differentiated customer classes and (2) equitable treatments toward servers (employees). A large amount of literature on fair routing policies focuses on the first kind. Avi-Itzhak et al. (2008) and Wierman (2007) provide good reviews on this type of work. The second kind of view (equitable treatment to employees), although important, has received less attention in the operations management literature. Our work is from this perspective.

Armony and Ward (2010) study a large-scale (large arrival rate and many servers) heterogeneous-server system and aim to minimize the steady-state expected waiting time subject to a fairness constraint on workload division. They propose a dynamic threshold routing policy that determines the assignment priorities of servers based on the total number of customers in the system. This policy is shown to be optimal asymptotically as the arrival rate and the number of servers grow to infinity. Given that the idleness proportion of each server is fixed in the Whitt-Halfin regime, the average waiting time for customers is minimized.

Atar et al. (2011) focus on the same type of service system and bring in the notion of servers' fairness. They analyze the performance of a policy that routes customers to the server with the longest cumulative idleness. This policy is "blind" as it requires no information on service or arrival rates. Not surprisingly, the policy attains equalization of cumulative idleness asymptotically. Recently, Reed and Shaki (2014) extend the framework of Atar et al. (2011) to general service time distribution, and also study a blind policy which aims to spread the incoming work amongst servers pools in an equitable way.

Tseytlin (2009) and Mandelbaum et al. (2010) study emergency departments in hospitals. They seek to allocate patients to wards in an equitable manner. They introduce the Randomized-Most-Idle (RMI) policy, which routes customers to a server pool (group of servers) with a probability that equals the ratio of the number of idle servers in that pool to the total number of idle servers in the system. This policy depends on neither cumulative idleness times nor capacities of the pools.

Tezcan (2008) studies the optimal control in a distributed parallel system. Since customers are routed upon arrival in this system, it differs from the setting in our model and in the other studies discussed thus far. One of Tezcan's proposed policies, named Minimum-Expected-Delay-Load-Balancing (MED-LB), is proved to asymptotically achieve the goal of having the average utilization of each server balanced.

All of the above papers employ asymptotic analyses and the theoretical results hold for large-scale systems. In contrast to asymptotic results, some papers look at server's fairness with exact analysis. As an extension of his investigation on the slow server problem (Cabral, 2005), Cabral (2007) studies the question of "who works most". He shows that, in a system with two different servers and a routing policy that sends customers to each server with equal probability when both are idle, the faster server serves more customers but works less hours in the long run. Wu et al. (2007) place emphasis on servers' fairness in the context of a bandwidth allocation problem in telecommunication networks. The fairness measurement used in their paper is not directly applicable to our model.

The asymptotic and the exact studies mentioned above assume that the administrator is the only decision maker and that the servers provide exogenous capacities. This paper differs from the above papers in that we study a decentralized system in which the servers choose their capacities, which is seen in many contexts where employees are not supervised strictly all the time.

Several papers also assume that servers choose their speed endogenously. Gopalakrishnan et al. (2013) study a dispatching and staffing problem in a strategic server context that resembles ours. Like our research, their research is motivated by fairness issues. However, server's trade-off does not include fairness in their model. On the contrary, we model fairness explicitly as part of server's utility. In addition, we look at heterogeneous servers and thus study asymmetric equilibria, which Gopalakrishnan et al. (2013) do not consider. Some other papers that study strategic servers are Kalai et al. (1992), Gilbert and Weng (1998), Cachon and Zhang (2007), Ching et al. (2010) and Choi et al. (2011). We note that the model, objective, and analysis of these papers are different from ours. One significant difference is that fairness is not an issue in these papers.

Our contributions in this paper are as follows.

1. We treat the issue of servers fairness as an endogenous factor. In our model, fairness is a component in players' utilities and affects their actions. Thus, their choice of capacity as a response to the announced policy considers fairness. To our knowledge, we are the first to take this view.

2. Our analysis does not appeal to a heavy traffic analysis. Thus, we do not need large scale systems for our results to hold. In fact, in some cases (e.g. off-peak hours in hospitals when the number of servers is quite small) the many-server heavy-traffic assumption may not apply. As a result, our result is exact and perhaps more appropriate in certain settings.

3. We extend the research on game-theoretic queueing by adding a component of fairness measurement into the server's objective function, which is not seen in this stream of literature. From a technical perspective, this extension makes our research significantly different from that in the literature and more challenging to cope with.

3.2 Model

3.2.1 Basic Assumptions

We study a two-server queueing system with Poisson arrivals and exponential service times. A homogeneous class of customers arrive at rate λ and the capacities of server 1 and server 2 are μ_1 and μ_2 (chosen endogenously, as will be clear later). We treat λ as exogenous in this paper, and thus normalize it to 1. Later, it will become apparent that our model and results apply to arbitrary λ. We require that the servers' minimum capacity is 1/2. This assumption makes sense in that employees are expected to work no less than a minimum level of efficiency (Gopalakrishnan et al., 2013). To control the system, the administrator announces a routing policy that satisfies certain requirements. We adopt the three assumptions in Armony and Ward (2010) regarding the feasible policies: (1) we do not allow pre-emption; (2) when there is a waiting customer, no server is allowed to be idle (this non-idling assumption is widely used in the literature); and (3) we only consider policies that are independent of future information (i.e. non-anticipatory). Furthermore, the system applies a single-queue formation, meaning that the arrived customers are routed only when at least one server is available. The use of a single queue instead of parallel queues is not only operationally sensible, but it also enhances customer satisfaction based on perceptions of the system (Larson, 1987). Finally, we assume that there is no abandonment.

We consider randomized routing policies. Let ϕ be the probability that the customer who arrives at an empty system is sent to server 1. Under our assumptions, the routing policy affects the system only when both servers are idle. It can be fully expressed by the routing probability ϕ. Therefore, the routing probability is henceforth referred to as the policy. We further assume that the policy does not depend on time (stationary policy). In fact, all our analysis assumes a long run stable system. After a policy is announced, a non-cooperative game unfolds between the two servers who simultaneously choose μ_1 and μ_2. We assume that changes in capacities only affect process flow measures and have no other effects (such as on the quality of service).

3.2.2 Individual Server's Objective

In this subsection, we introduce the server's objective function and the formulation of the decentralized problems, in which servers separately and simultaneously decide their capacities. We model the server's objective using two aspects. They are referred to as self-focused and comparison-based disutility. The first aspect pertains to a server's own performance measures

encompassing idleness and effort. The second aspect is borrowed from the organizational behavior literature (Festinger, 1954), and it reflects the dissatisfaction towards unfairness due to comparison. We will use the sum of the inverse-idleness function and the capacity cost function to capture self-focused disutility. Also, we will use an "unfairness" function to describe comparison-based disutility.

Self-focused Disutility.

First, consider the long run average fraction of idle time for each server, denoted by τ_i $(i = 1, 2)$. We define the inverse-idleness function by

$$g_i(\mu_1, \mu_2) = \frac{1}{\tau_i}, \quad i = 1, 2. \tag{3.1}$$

Recall that the minimum capacity is $1/2$, then the only case which results in an unstable system is $\mu_1 = \mu_2 = 1/2$. In such a case, $\tau_1 = \tau_2 = 0$ and g_i is defined as infinity. Hence, we see that, although choosing such a low rate is allowed, overloading the system is never preferred by the servers. This resonates with practice where systems are rarely overloaded in the long run. Since neither server prefers the outcome of both choosing the minimum capacity, we ignore this case. Therefore, from here onwards, we only consider a stable system. In general, g_i is decided by the capacities and the routing policy; however, we simply write it as a function of capacities. Later it will be clear that the policies we are interested in can be expressed through μ_1 and μ_2.

Second, as mentioned, the two servers are possibly heterogeneous in their capability in providing service. In other words, it requires different effort levels for them to provide the same capacity. We capture this by having different unit costs of capacity. Without loss of generality, let server 2 be the costlier one. Assume that server 1's unit cost is t and his capacity cost is $t\mu_1$. In addition, assume that server 2's unit cost is $t + \Delta t$ and his capacity cost is $(t + \Delta t)\mu_2$. Suppose $t > 0$ and $\Delta t \geq 0$. Note that, when $\Delta t = 0$, we have the case of homogeneous servers. Our linear cost assumption is a simplification of the convex cost function used, for example, by Kalai et al. (1992), who model servers' cost as a strictly convex function. However, we will see later that our main results remain valid even if we assume strict convexity.

Together, the inverse-idleness and the capacity cost contribute to server's self-focused disutility: $g_1 + t\mu_1$ for server 1 and $g_2 + (t + \Delta t)\mu_2$ for server 2.

Comparison-based Disutility.

Fairness is indeed amorphous and depends on the context. Although people's judgement on fairness may have some connection with their general expectations, research has shown that it is based more strongly on social comparison (Austin et al., 1980). Therefore, we model the dissatisfaction caused by unfairness between servers as their comparison-based disutility. Following the discussion in Section 3.1, we apply the equity model studied by Adams (1965) and Huseman et al. (1987). Specifically, servers perceive fairness by comparing the outcome/input ratio with one another. Inequity exists if and only if the ratios are unequal, incurring the comparison-based disutility. As discussed earlier, we use the chosen capacity as input and the

average fraction of idle time τ_i as outcome. A similar treatment can be seen in literature. For example, Mandelbaum et al. (2010) use "utilization", which is $1 - \tau_i$ in our setting, as a factor for fairness measurement.

Since the input and the outcome are identified, we apply the equity model directly. Hence, servers consider it fair if and only if

$$\frac{\tau_1}{\mu_1} = \frac{\tau_2}{\mu_2}. \tag{3.2}$$

If the above equality becomes "<" (for example, $\mu_1 > \mu_2$ and $\tau_1 < \tau_2$), then the faster server (server 1) is exposed to a higher workload instead of being rewarded with more rest time. We say that server 1 is under-rewarded in this case. Similarly, he is over-rewarded if the equality becomes ">" in (3.2). Both cases are perceived as unfair. According to the study in Huseman et al. (1987), servers have an incentive to restore fairness whether over-rewarded or under-rewarded. However, inequity that is to their disadvantage generally produces a higher such incentive, as opposed to inequity that is to their advantage (see Fehr and Schmidt (1999) and Cui et al. (2007), for example). As a result, to simplify our model, we assume that the under-rewarded server incurs disutility, whereas the over-rewarded server is unaffected.

From the above discussion, server 1's magnitude of inequity can be described as $\left(\frac{\tau_2}{\mu_2} - \frac{\tau_1}{\mu_1}\right)^+$. We now define the unfairness function based on this quantity. First, consider the following two scenarios: (1) Server 1 inputs capacity $\mu_1^{(1)}$ and experiences outcome/input difference $d > 0$. (2) Server 1 inputs capacity $\mu_1^{(2)}$ and experiences outcome/input difference of the same amount d. Apparently, if $\mu_1^{(1)} > \mu_1^{(2)}$, then server 1 must experience more dissatisfaction in scenario (1) than he does in scenario (2). Studies in Brockner et al. (1992) provides supporting evidence. Therefore, we multiply the server's capacity with the magnitude of inequity to model this feature. Second, as part of server's disutility, it is reasonable to expect the function to have increasing margins. As the difference becomes larger, the server's dissatisfaction increases at a greater rate. To model this, we take the quadratic form of the above expression. Therefore, the unfairness function is defined by

$$f_1 = \left(\mu_1 \left(\frac{\tau_2}{\mu_2} - \frac{\tau_1}{\mu_1}\right)^+\right)^2 \tag{3.3}$$

for server 1, and

$$f_2 = \left(\mu_2 \left(\frac{\tau_1}{\mu_1} - \frac{\tau_2}{\mu_2}\right)^+\right)^2 \tag{3.4}$$

for server 2.

After a routing policy is announced, the servers compete in capacities to minimize their individual disutility functions with the policy being common knowledge. Throughout this paper, we focus only on the pure strategies. Server's total disutility is a weighed combination of the self-focused disutility (weight 1) and comparison-based disutility. Let $\alpha_f^{(1)}$ and $\alpha_f^{(2)}$ be the positive weights that server 1 and 2 associate to the comparison-based disutility, respectively.

In general, we assume that the two weights are equal when $\Delta t = 0$, but otherwise arbitrary. The servers' problems are as follows. Server 1, given server 2's chosen capacity $\mu_2 \geq 1/2$, solves

$$\text{(S1)} \qquad \min_{\mu_1 \geq 1/2} F_1 = \alpha_f^{(1)} f_1 + g_1 + t\mu_1.$$

Given $\mu_1 \geq 1/2$, server 2 solves

$$\text{(S2)} \qquad \min_{\mu_2 \geq 1/2} F_2 = \alpha_f^{(2)} f_2 + g_2 + (t + \Delta t)\mu_2.$$

In this two-player static game, besides the usual trade-off of idle time and capacity cost, servers also need to consider fairness between themselves. Note that problems (S1) and (S2) both have the announced policy embedded in the objective functions. As a result, the game structure largely relies on the routing policy that is applied.

3.2.3 Routing Policies

We have made several assumptions in Section 3.2.1 concerning the policy ϕ. Without further specification, however, it is intractable to devise a feasible policy for the administrator to guarantee the optimal system efficiency. The problem remains difficult even if we restrain our attention to a reasonably small class of policies. As mentioned earlier, our goal in this paper is not to solve for optimal policies, but to analyze a number of practical policies. We aim to understand how these policies affect servers' choices and the fairness among them. This is an important stepping stone to the search for the best-performing policies within that class. In this section, we characterize the policy class we are interested in and list the policies that we will analyze in detail.

Let $\mathbb{R}$ be the set of real numbers. For every $r \in \mathbb{R}$, define a policy

$$\phi_r = \frac{\mu_1^r}{\mu_1^r + \mu_2^r}.$$

In addition, define $\phi_\infty = \lim_{r\to\infty} \phi_r$ (∞ can be either positive or negative infinity). We shall focus on the policy class defined by

$$\mathcal{P} = \{\phi_r \quad | \quad r \in \mathbb{R} \cup \{\pm\infty\}\}.$$

Note that every policy in $\mathcal{P}$ is determined only through μ_1 and μ_2. Moreover, the policies are related to the comparison between the two capacities. Specifically, if $r > 0$, then the faster server gets more customers. If $r < 0$, then the slower server gets more customers. When $r = 0$, the customers are equally routed to the two servers regardless of their capacities. For this reason, policies with small r are considered to be more fairness-oriented (the faster server is rewarded with less work). Furthermore, all of the policies in $\mathcal{P}$ are continuous with respect to μ_1 and μ_2 except when $r \in \{\pm\infty\}$. The two special cases with discontinuity will be carefully

treated later in this paper. In fact, the class $\mathcal{P}$ is also studied by Gopalakrishnan et al. (2013) and is named "rate-based" policies. We follow them and call $\mathcal{P}$ the capacity-based policy class.

In this paper, we will mainly study four specific policies. All of them are commonly used in practice because they are easy to implement. First, we discuss the two discontinuous policies in $\mathcal{P}$. As will be observed, the discontinuities of both policies are at $\mu_1 = \mu_2$. When $r = +\infty$, we see by definition that

$$\phi_{+\infty} = \begin{cases} 1 & \mu_1 > \mu_2 \\ \frac{1}{2} & \mu_1 = \mu_2 \\ 0 & \mu_1 < \mu_2 \end{cases}.$$

This is the *Faster Server First* (FSF) policy, which we mentioned in Section 3.1. Several papers have discussed the FSF policy, including Armony and Ward (2010). This policy is popular due to certain obvious operational benefits. Intuitively, it is the best policy if there are no issues of fairness. However, it has negative consequences if servers care about fairness. On the other end of spectrum is the case where $r = -\infty$. Again, by definition we have

$$\phi_{-\infty} = \begin{cases} 0 & \mu_1 > \mu_2 \\ \frac{1}{2} & \mu_1 = \mu_2 \\ 1 & \mu_1 < \mu_2 \end{cases}.$$

This policy is named the *Slower Server First* (SSF) policy and is a natural remedy to the unfairness caused by FSF. Despite its effective treatment to unfairness among servers, previous works on servers fairness in a queueing system seem to have little interest in SSF. The main reason for this is that most of the previous models assume exogenous capacities (Armony and Ward, 2010; Mandelbaum et al., 2010). Unsurprisingly, SSF makes least use of the total capacity, and thus it leads to low system efficiency. However, in a model where capacities are endogenously determined by servers, SSF induces a strong incentive for them to provide high capacities. Therefore, for models that assume strategic servers such as Gopalakrishnan et al. (2013), it makes sense to study SSF closely.

Second, we list two continuous policies from $\mathcal{P}$ that we will pay close attention to. The definition of equity, (3.2), indicates that to maintain fairness, the idle time of the server with lower capacity must be small. One can achieve this by sending more customers to the slower server than the FSF policy. As noted, the SSF policy does this but fails to admit continuous disutility functions. Hence, we propose a policy that is in the same spirit as SSF but much smoother. Let $r = -1$ and then

$$\phi_{-1} = \frac{1/\mu_1}{1/\mu_2 + 1/\mu_1} = \frac{\mu_2}{\mu_1 + \mu_2}.$$

This policy also sends more customers to the slower server to incentivize higher capacity. Since this policy routes the customers to each server with the probability proportionate to the inverse

capacities, we shall refer to it henceforth as the *proportional* policy (Prop). Finally, we include the simplest policy in $\mathcal{P}$

$$\phi_0 = \frac{1}{2}.$$

We refer to this policy as the *Half-Half* (HH) policy. Being a constant, HH is the easiest to implement and analyze. Many previous works with either strategic or non-strategic servers apply this policy, e.g. Cabral (2007); Kalai et al. (1992). Moreover, sharing the customers equally between servers bears a natural sense of equity, especially when the heterogeneity between servers is unclear.

In the rest of the paper, we write ϕ_{FSF}, ϕ_{SSF}, ϕ_{Prop} and ϕ_{HH} instead of $\phi_{+\infty}$, $\phi_{-\infty}$, ϕ_{-1} and ϕ_0, respectively, and let $\tilde{\mathcal{P}} = \{\phi_{\text{FSF}}, \phi_{\text{SSF}}, \phi_{\text{Prop}}, \phi_{\text{HH}}\}$.

3.3 Two-Server Game: Analysis and Results

In this section, we study the optimization problems posed to the two servers and provide some results regarding the structural properties of their optimal decisions. We do not consider mixed strategies but only look at pure strategies. Denote the announced policy by $\bar{\phi}$. We require that $\bar{\phi} \in \mathcal{P}$. Taking $\bar{\phi}$ as common knowledge, both servers simultaneously choose capacities to solve problems (S1) and (S2), respectively. We focus on server 1's objective functions in detail because parallel results for server 2 follow without difficulty by exchanging the roles of μ_1 and μ_2. To clarify notations, recall that g_i is the inverse-idleness function for server i ($i = 1, 2$) defined in (3.1). In the following discussion, we write it as $g_i(\mu_i|\mu_j, \bar{\phi})$. The first argument μ_i is server i's action, which is the variable of the function, and μ_j and $\bar{\phi}$ are his opponent's action and the announced policy, which are fixed to him. A similar notation is used for the unfairness function f_i. In addition, the $'+'$ and $'-'$ signs beside the variable mean right and left limit, respectively. All proofs are in the Appendix A.2.

To begin with, we look at a benchmark case where $\Delta t = 0$ (hence $\alpha_f^{(1)} = \alpha_f^{(2)}$). This means that the two servers are homogeneous. We study this case because it is a good approximation in circumstances where the heterogeneity is very small or even negligible. In this case, a symmetric Nash equilibrium where the two servers choose the same capacity is expected. This is indeed true as long as the implemented policy is not FSF or SSF. The following theorem characterizes the symmetric equilibrium for the continuous policies.

Theorem 3.3.1. *Suppose $\Delta t = 0$ and t is fixed. Let $\mathcal{P}_c = \mathcal{P} \setminus \{\phi_{FSF}, \phi_{SSF}\}$. For every $\bar{\phi} \in \mathcal{P}_c$, there exists a unique symmetric Nash equilibrium (μ^*, μ^*). Moreover,*

(i) for any small $\epsilon > 0$, there exists $\bar{r} > 0$ such that for any ϕ_r with $r > \bar{r}$, we have $\mu^* < 1/2 + \epsilon$; and

(ii) for any big $M > 0$, there exists $\underline{r} < 0$ such that for any ϕ_r with $r < \underline{r}$, we have $\mu^* > M$.

The above theorem only deals with symmetric equilibrium. Other asymmetric equilibria may exist, but we do not consider them. One important implication of Theorem 3.3.1 is as

follows. Although a symmetric equilibrium for every continuous policy always exists, not all of them are favorable. Examples of unfavorable equilibrium include small μ^* (close to 1/2), which makes g_i large and big μ^* (approaching infinity), which results in large capacity cost.

Theorem 3.3.1 shows that we can find a sequence of policies $\{\phi_r\}$ which approaches ϕ_{FSF}, such that the corresponding equilibrium sequence converges to 1/2. Similarly, for policies approaching SSF, the equilibrium sequence increases to $+\infty$. For policies that have large $|r|$, they are continuous, but are "almost" FSF and SSF. Theorems 10 and 11 in Gopalakrishnan et al. (2013) also characterize the properties of these policies. Because the capacity cannot exceed a fixed large number by their assumption, they claim that policies that are "almost" SSF do not induce symmetric equilibrium. Although we cannot deduce that FSF and SSF do not admit symmetric equilibrium directly from Theorem 3.3.1, we do have such result as a corollary of Propositions 3.3.2 and 3.3.3, which will be introduced shortly. We emphasize again that the above discussion only concerns symmetric equilibrium. As will be shown in Section 3.4, when $\Delta t = 0$, the FSF policy may induce Nash equilibria that are not symmetric.

Now, we move to the general case where servers are not necessarily homogeneous. In the following discussion, we restrict our attention to the four policies $\bar{\phi} \in \tilde{\mathcal{P}}$. We will study these policies from two aspects. First, we examine how servers behave in the two-server game under those policies. This game may or may not have an equilibrium. Second, supposing that we have equilibria under the policies, we will compare them along the measure of system efficiency, which will be discussed in Section 3.4. Naturally, the policy that leads to higher efficiency is preferred by the administrator.

As an important preparation for the equilibrium analysis, we give some basic results that depict the shape of the objective functions. Two propositions characterize server 1's inverse-idleness function and unfairness function, respectively, under the four policies.

Proposition 3.3.2. *Suppose $\mu_2 > 1/2$ is fixed.*

(i) If $\bar{\phi} \in \{\phi_{Prop}, \phi_{HH}\}$, then $g_1(\mu_1|\mu_2, \bar{\phi})$ is continuous, strictly decreasing and convex for $\mu_1 > 1/2$.

(ii) If $\bar{\phi} \in \{\phi_{FSF}, \phi_{SSF}\}$, then $g_1(\mu_1|\mu_2, \bar{\phi})$ has two continuous pieces on $1/2 < \mu_1 < \mu_2$ and $\mu_1 > \mu_2$ respectively. On both pieces, $g_1(\mu_1|\mu_2, \bar{\phi})$ is strictly decreasing and convex. Furthermore,

$$g_1(\mu_2 - |\mu_2, \phi_{FSF}) < g_1(\mu_2|\mu_2, \phi_{FSF}) < g_1(\mu_2 + |\mu_2, \phi_{FSF}) \tag{3.5}$$

$$g_1(\mu_2 - |\mu_2, \phi_{SSF}) > g_1(\mu_2|\mu_2, \phi_{SSF}) > g_1(\mu_2 + |\mu_2, \phi_{SSF}). \tag{3.6}$$

Proposition 3.3.3. *Let $\mu_2 > 1/2$ be fixed.*

(i) If $\bar{\phi} \in \{\phi_{FSF}, \phi_{Prop}, \phi_{HH}\}$, then $f_1(\mu_1|\mu_2, \bar{\phi}) = 0$ for $1/2 < \mu_1 \leq \mu_2$, and $f_1(\mu_1|\mu_2, \bar{\phi})$ is strictly convex and increases to infinity on $\mu_1 > \mu_2$.

(ii) If $\bar{\phi} = \phi_{SSF}$, then there exists $1/2 \leq a < \mu_2$ and $b > \mu_2$ such that $f_1(\mu_1|\mu_2, \phi_{SSF})$ is continuous, strictly increasing and convex on $a \leq \mu_1 < \mu_2$ and $\mu_1 \geq b$, respectively. Moreover, $f_1(\mu_1|\mu_2, \phi_{SSF}) = 0$ for both $1/2 \leq \mu_1 \leq a$ and $\mu_2 \leq \mu_1 \leq b$.

Note that we exclude the case $\mu_2 = 1/2$. In fact, the only part in the propositions that needs to change to accommodate this case is Proposition 3.3.2 (ii). Just one decreasing convex piece will exist for g_1, i.e. $\mu_1 > \mu_2$. All other results remain valid.

Proposition 3.3.2 implies that the inverse-idleness function g_1 is continuous only if the policy is continuous. Under FSF or SSF, g_i has a discontinuity at $\mu_1 = \mu_2$ because the policy is discontinuous there. As a result, when $\bar{\phi} \in \{\phi_{\text{FSF}}, \phi_{\text{SSF}}\}$ and the two servers have the same capacity, server 1 can reduce a strictly positive amount of disutility by deviating a little from μ_2. Proposition 3.3.3 shows the distinguishing feature of the SSF policy. For the other three policies, the unfairness function f_1 equals zero if $1/2 < \mu_1 \leq \mu_2$, and it becomes positive immediately after μ_1 surpasses μ_2. However, under the SSF policy, there is a positive piece of f_1 when $\mu_1 < \mu_2$, and f_1 remains zero after μ_1 exceeds μ_2 until the difference is sufficiently large. This characteristic shows that SSF strongly incentivizes the slow server.

Recall that the overall objective function F_i is a linear combination of g_i, f_i, and a linear cost function. An immediate and useful implication is that F_i also has the corresponding convexity properties. A further implication concerning FSF and SSF is that

$$F_1(\mu_2 - |\mu_2, \phi_{\text{FSF}}) < F_1(\mu_2|\mu_2, \phi_{\text{FSF}}) < F_1(\mu_2 + |\mu_2, \phi_{\text{FSF}})$$

$$F_1(\mu_2 - |\mu_2, \phi_{\text{SSF}}) > F_1(\mu_2|\mu_2, \phi_{\text{SSF}}) > F_1(\mu_2 + |\mu_2, \phi_{\text{SSF}}).$$

Since the above inequalities are strict, μ_2 can never be the value of μ_1 minimizing F_1. Thus, a unilateral deviation from any symmetric strategy profile will always exist. Therefore, as a corollary of the above propositions, we conclude that FSF and SSF do not induce symmetric equilibrium. The jump in F_1 for the two discontinuous policies also poses difficulty in finding an asymmetric equilibrium. We can derive sufficient conditions for the existence of equilibrium, but they are implicit and far from illustrative. Instead, we will mainly treat it numerically in Section 3.4. We will see that the sufficient conditions are liberal for FSF but somewhat stringent for SSF. That is, a large number of instances display an equilibrium under FSF while no instance with $\Delta t < t$ displays equilibrium under SSF.

Fortunately, Prop and HH result in continuous objective functions with desirable properties. Under Prop or HH, the two-server game is resolved by solving two convex optimization problems simultaneously. The theorem below summarizes our main findings.

Theorem 3.3.4. *Suppose $\Delta t > 0$ and $\bar{\phi} \in \{\phi_{Prop}, \phi_{HH}\}$. Then there exists a unique Nash equilibrium $(\overline{\mu}_1, \overline{\mu}_2)$, and $\overline{\mu}_1 > \overline{\mu}_2$.*

We will apply the result in Nikaidô and Isoda (1955) to show the existence of the equilibrium, and we will turn to Theorem 7 in Cachon and Netessine (2006) for the proof of uniqueness. We

prove $\overline{\mu}_1 > \overline{\mu}_2$ by contradiction. This result clearly shows the benefit of using Prop or HH as the routing policy. These policies both induce a unique equilibrium in which the cheaper server provides higher capacity. It is also noteworthy that Theorem 3.3.4 is valid for all heterogeneous servers cases ($\Delta t > 0$), regardless of the unit cost and the weights of fairness.

Three points summarize our main results in this section. First, we showed that all continuous capacity-based policies admit a unique symmetric equilibrium when the servers are homogeneous. However, this common capacity is either too small or too large if the policy is close to FSF or SSF. Second, we outlined the shape of server's objective function for different policies, and thus laid a solid foundation for further analyses. Third, we established the existence and uniqueness of the Nash equilibrium under the Prop and HH policies.

3.4 Simulations and Discussions

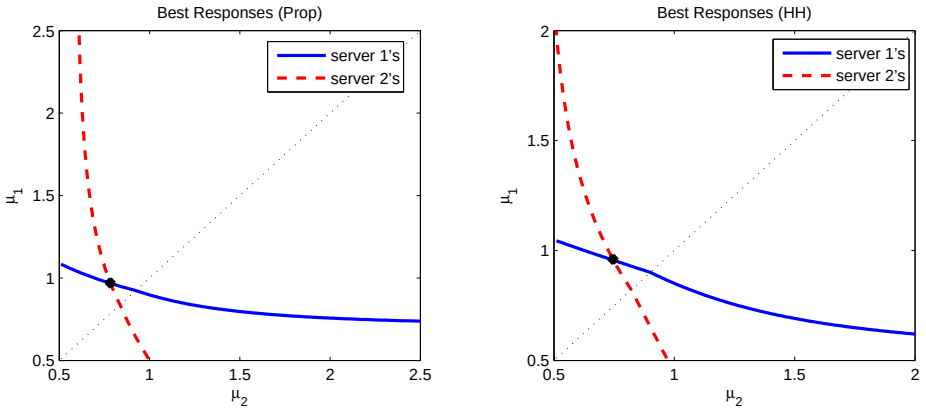
In this section, we further study the four policies in $\tilde{\mathcal{P}}$ by simulations. Our goals are to understand the servers' behaviors under each policy, to compare the policies' performances with respect to system efficiency, and to obtain further insights. Our approach consists of three parts. (1) In Section 3.4.1, we examine the existence and uniqueness of equilibrium by plotting the best response functions under each policy. A unique equilibrium exists for Prop and HH (as predicted in Theorem 3.3.4). For FSF and SSF, the non-existence of Nash equilibrium is illustrated. In this case, we characterize the servers' off-equilibrium behaviors by an interval of capacities. (2) Section 3.4.2 explicitly compares the outcome of the game under different policies. We also obtain some empirical patterns regarding the equilibrium properties for FSF and SSF. (3) In Section 3.4.3, we focus on equilibrium outcomes and compare the four policies based on system efficiency. Particularly, we study how servers' attitudes towards fairness and servers' heterogeneity affect a policy's performance.

To simplify the simulation, we reduce the freedom of parameters by setting $\alpha_f^{(2)} = (t + \Delta t)\alpha_f^{(1)}/t$. Here we assume that the costlier server cares about fairness more and the weights are proportional to capacity costs. This is not a crucial assumption as our main results and observations under the opposite assumption ($\alpha_f^{(2)} = (t + \Delta t)\alpha_f^{(1)}/t$) are the same. Arbitrary $(\alpha_f^{(1)}, \alpha_f^{(2)})$ cases can also be treated in a similar manner. Hence, we simulate with three free variables: $\alpha_f^{(1)} \in [0, 10]$, $t \in (0, 10]$, and $\Delta t \in [0, 10]$. The discretization step is 0.5. Algorithms are implemented using Matlab and are run on a standard PC.

3.4.1 Best Response Functions and Off-Equilibrium Behaviors

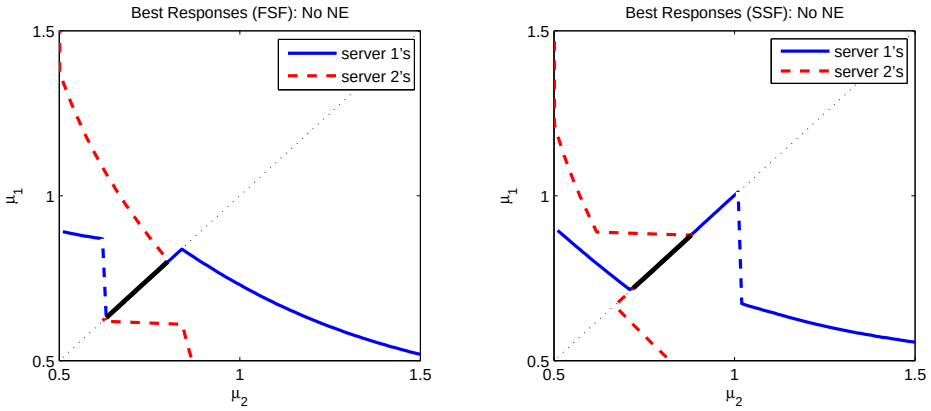
The convexity properties shown in Propositions 3.3.2 and 3.3.3 grant the use of binary search algorithm to get the best response functions for each server. Figure 3.1 (Prop and HH) and Figure 3.2 (FSF and SSF) illustrate their typical shapes.

An immediate observation from Figure 3.1 is that the best response functions for Prop and HH are continuous and intersect exactly once. Besides, the intersection lies above the



(a) $\alpha_f^{(1)} = 4$, $t = 2$, $\Delta t = 1$ (b) $\alpha_f^{(1)} = 4$, $t = 2$, $\Delta t = 1$

Figure 3.1: Best response functions (Prop and HH).



(a) $\alpha_f^{(1)} = 10$, $t = 2$, $\Delta t = 0.5$ (b) $\alpha_f^{(1)} = 4$, $t = 5$, $\Delta t = 3$

Figure 3.2: Best response functions (FSF and SSF). Linear segments on 45° line are not part of the functions.

$\mu_1 = \mu_2$ line, which means that server 1 provides higher capacity at equilibrium. Furthermore, Figure 3.1 also confirms our theoretical result that the equilibrium is unique. According to Cachon and Netessine (2006), in the two-player case, an equivalent condition to Theorem 7 in their paper is that the multiplication of slopes of best response functions should not exceed one at equilibrium. Figure 3.1 clearly shows that the two best response functions both have derivatives with absolute values of less than 1 at the intersection. Therefore the condition holds and we have the uniqueness of equilibrium.

Because of the discontinuity of FSF and SSF, the best response curves under these policies have jumps and equilibrium properties (whether it exists and, if yes, how many) are unclear. Although FSF and SSF induce equilibrium in many instances, the instance shown in Figure 3.2 does not have any equilibrium and we try to understand servers' off-equilibrium behavior.

Although it seems that, by Figure 3.2, there is a continuum of symmetric equilibria for both FSF and SSF, it is not true. We draw the linear segments as part of best response curves in

the graphs to represent that the functions are not defined there. Take server 1 for example, by inequalities (3.5) and (3.6), we know that it is actually the left limit of μ_2 (μ_2-) that achieve the infimum, not minimum, of server 1's objective function under FSF, and the right limit of μ_2 (μ_2+) under SSF. In the simulation, we plot them as $\mu_1 = \mu_2 - \epsilon$ and $\mu_1 = \mu_2 + \epsilon$, respectively, with very small $\epsilon > 0$. The same logic and technique applies to server 2. Therefore, servers' best response function does not actually exist along the 45 degree line.

However, this linear segment is the key to understand servers' off-equilibrium behaviors. Take Figure 3.2 (a) (FSF) for example. The best response iteration leads the strategy profile (μ_1, μ_2) to eventually land on the $\mu_1 = \mu_2 - \epsilon$ segment. This means that server 1 undercuts server 2 by a small quantity. Then the same strategy is taken by server 2 ($\mu_2 = \mu_1 - \epsilon$). As a result, the servers decrease the capacities in turn until the inverse-idleness becomes sufficiently large. Next, one of the servers increases his capacity, and another round of undercutting begins. The thick solid line shown in Figure 3.2 (a) indicates the undercutting range. Since the line segment lies in the 45 degree line, we can express the endpoints of the line segment as $(\underline{\mu}, \underline{\mu})$ and $(\bar{\mu}, \bar{\mu})$, and we simply write $[\underline{\mu}, \bar{\mu}]$ to denote the servers' off-equilibrium behavior. Similarly, for SSF, we also have such an interval. However, instead of undercutting, the servers increase capacities in turn within the range of the interval, and they decrease capacities when the cost is too high to bear.

3.4.2 Comparison of Equilibria

Table 3.1 shows what we obtained for different sets of parameters and policies. These are some of the representative instances from our extensive simulations. The first kind of entry is presented with parentheses and represents the equilibrium capacity decisions $(\overline{\mu}_1, \overline{\mu}_2)$. The second kind of entry is marked with square brackets and reports the interval $[\underline{\mu}, \bar{\mu}]$ of the off-equilibrium range.

A few noteworthy observations emerge from our simulation. First, the servers' possible off-equilibrium capacities under FSF are very small compared to the equilibrium capacities under Prop or HH. Indeed, we observe from Table 3.1 that $\bar{\mu}$ under FSF is smaller than $\overline{\mu}_1$ under Prop and HH. This is a major disadvantage of FSF when accounting for endogenous capacity. On the contrary, the off-equilibrium capacities under SSF are high enough to compare with Prop and HH. In fact, for the same parameters, every $\bar{\mu}$ under SSF is higher than $\overline{\mu}_1$ under Prop and HH; and in some case $\underline{\mu}$ is higher than $\overline{\mu}_2$. Hence, by effectively incentivizing the slow server, SSF induces, even being off-equilibrium, possibly large capacities for both servers.

Second, although the equilibrium $(\overline{\mu}_1, \overline{\mu}_2)$ induced by SSF is relatively high compared to that induced by Prop and HH (especially for server 2), SSF cannot induce equilibrium in quite a few instances. This means that the first order condition under SSF is stringent. During our simulation, we find that Δt must be at least as large as t for an equilibrium to exist under SSF (i.e. server 2's capacity cost has to be at least twice the capacity cost of server 1). Another observation is that, given fixed t and Δt, if an equilibrium exists for $\alpha_f^{(1)}$, then one also exists for all larger weights of fairness. In other words, when more weight is given to fairness, it is

Δt		$t=1$			$t=5$		
		$\alpha_f^{(1)}=1$	$\alpha_f^{(1)}=5$	$\alpha_f^{(1)}=10$	$\alpha_f^{(1)}=1$	$\alpha_f^{(1)}=5$	$\alpha_f^{(1)}=10$
0	FSF	[0.70, 0.98]	[0.66, 0.98]	[0.64, 0.98]	(0.90,0.57), (0.57,0.90)	[0.60, 0.71]	[0.60, 0.71]
	SSF	[0.98, 1.94]	[0.98, 1.94]	[0.98, 1.94]	[0.71, 1.01]	[0.71, 1.02]	[0.71, 1.02]
	Prop	(1.13, 1.13)	(1.13, 1.13)	(1.13, 1.13)	(0.75, 0.75)	(0.75, 0.75)	(0.75, 0.75)
	HH	(1.08, 1.08)	(1.08, 1.08)	(1.08, 1.08)	(0.74, 0.74)	(0.74, 0.74)	(0.74, 0.74)
0.5	FSF	[0.70, 0.89]	[0.66, 0.89]	[0.64, 0.89]	(0.92,0.54), (0.59,0.87)	(0.84, 0.59)	[0.60, 0.70]
	SSF	[0.98, 1.61]	[0.98, 1.61]	[0.98, 1.61]	[0.71, 0.98]	[0.71, 0.98]	[0.71, 0.99]
	Prop	(1.16, 0.96)	(1.12, 0.97)	(1.10, 0.97)	(0.77, 0.72)	(0.77, 0.72)	(0.77, 0.72)
	HH	(1.14, 0.89)	(1.08, 0.92)	(1.04, 0.93)	(0.77, 0.71)	(0.77, 0.71)	(0.76, 0.71)
3	FSF	(1.17, 0.50)	(0.97, 0.57)	(0.90, 0.62)	(0.94, 0.50)	(0.89, 0.50)	(0.84, 0.53)
	SSF	(1.21, 0.68)	(1.20, 0.68)	(1.19, 0.68)	[0.71, 0.88]	[0.71, 0.88]	[0.71, 0.89]
	Prop	(1.26, 0.63)	(1.14, 0.65)	(1.07, 0.67)	(0.84, 0.61)	(0.84, 0.61)	(0.83, 0.62)
	HH	(1.25, 0.56)	(1.08, 0.61)	(1.01, 0.65)	(0.87, 0.57)	(0.85, 0.58)	(0.84, 0.59)
6	FSF	(1.17, 0.50)	(0.98, 0.50)	(0.91, 0.50)	(0.94, 0.50)	(0.89, 0.50)	(0.85, 0.50)
	SSF	(1.33, 0.50)	(1.23, 0.53)	(1.18, 0.54)	[0.71, 0.82]	(0.82, 0.59)	(0.82, 0.59)
	Prop	(1.32, 0.51)	(1.17, 0.52)	(1.09, 0.54)	(0.91, 0.52)	(0.90, 0.53)	(0.88, 0.53)
	HH	(1.26, 0.51)	(1.09, 0.51)	(1.02, 0.52)	(0.92, 0.51)	(0.90, 0.51)	(0.88, 0.51)

Table 3.1: Possible game outcomes. Entries with square brackets represent the interval while those with parentheses represent the equilibrium.

more likely that an equilibrium will exist. Table 3.1 illustrates these empirical patterns.

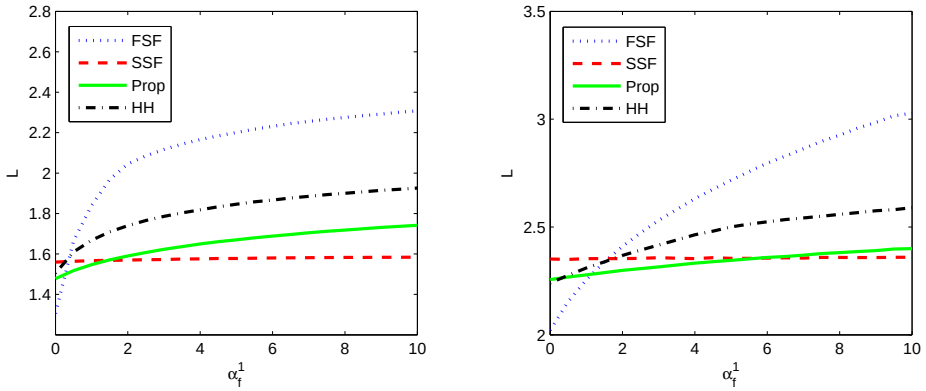
Third, while the possibility of an arbitrary number of equilibria has not been ruled out, we have observed only the two equilibria case under the FSF policy. Besides, it only happens in a small number of instances. The key condition for FSF to have two equilibria is Δt being small (i.e. the heterogeneity between two servers is negligible). For example, see the cases when $\Delta t = 0$ and 0.5 in Table 3.1. Except for these cases with two equilibria under FSF, all other instances have at most one equilibrium for FSF or other policies. Interestingly, when there is only one equilibrium $(\overline{\mu}_1, \overline{\mu}_2)$, we always observe $\overline{\mu}_1 \geq \overline{\mu}_2$, regardless of the policy, which implies an intuitive property that the cheaper server provides a higher capacity. For Prop and HH, we proved this theoretically; for FSF and SSF, however, this is observed empirically.

3.4.3 Comparison in Policy Performance

In the following, we will only focus on equilibrium outcomes and compare the policies. The performance measure is system efficiency, which we assume to be the average number of sojourn customers in the system: $L = L(\mu_1, \mu_2, \bar{\phi})$. The administrator prefers this quantity to be small. Note that if there are two equilibria, we compute L for both and use the smaller one for comparison. We do not plot L when no equilibrium exists.

In the experiment, we vary two parameters in our model: (1) how much servers care about fairness and (2) the heterogeneity between servers. The first is characterized by the weight $\alpha_f^{(1)}$ while the second by Δt. In the simulations, we plot L against these two parameters. Hence, we gain insights into how servers' attitude towards fairness and servers' heterogeneity affect a policy's performance. For illustration purposes, we only show the representative instances from the extensive simulations we conducted.

Figure 3.3 contains two L versus $\alpha_f^{(1)}$ plots with different cost parameters (specified in the caption). The plots clearly show that $L(\phi_{\text{FSF}})$ has the biggest increase along $\alpha_f^{(1)}$. Curves

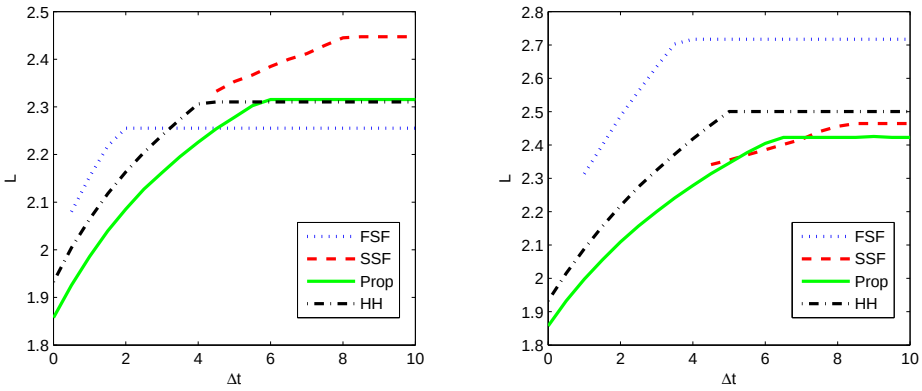


(a) Low cost: $t = 1$, $\Delta t = 3$ (b) High cost: $t = 3$, $\Delta t = 5$

Figure 3.3: Performance of policies against fairness weight $\alpha_f^{(1)}$ with different costs.

representing Prop and HH also have observable increases, but are pretty stable. For SSF, $L(\phi_{\text{SSF}})$ is almost flat in the plots. This indicates that, if a policy sends less customers to the slower server, then it is more sensitive to $\alpha_f^{(1)}$. Indeed, such a policy is intuitively unfair and is affected greatly as servers care about fairness more. Besides, higher costs make the sensitivity of $L(\phi_{\text{FSF}})$ more obvious, as it is steeper in (b) than in (a). Furthermore, when $\alpha_f^{(1)}$ is small, FSF induces the least congestion among all policies. This is obvious when $\alpha_f^{(1)} = 0$ (i.e. there is no fairness issue because servers do not care about it). In fact, even when $\alpha_f^{(1)}$ is positive, $L(\phi_{\text{FSF}})$ may still be the smallest. In addition, a higher costs case admits more such positive $\alpha_f^{(1)}$ values. In general, when $\alpha_f^{(1)}$ is small and the cost is high, FSF can perform better in terms of system efficiency than other policies. This finding appears to be in contrast to Armony and Ward (2010), who have shown that FSF is completely unsuitable if servers care about fairness (i.e. $\alpha_f^{(1)} > 0$ in our model). This difference is due to the key assumption about capacity endogeneity. The capacities, and therefore the cost functions, are exogenous in their model.

Next, we study how heterogeneity affects a policy's performance. Figure 3.4 gives two L versus Δt plots with the same $t = 3$ but different weights of fairness. Notice that curves are drawn only when an equilibrium exists. Since all curves are continuous, the plots indicate that, if an equilibrium exists for some Δt, then one also exists for all larger Δt. As discussed earlier, it is empirically true that the SSF policy induces equilibrium only when Δt is bigger than t. In



(a) Small weight: $\alpha_f^{(1)} = 1$, $t = 3$ (b) Large weight: $\alpha_f^{(1)} = 5$, $t = 3$

Figure 3.4: Performance of policies against extra cost Δt.

this sense, SSF's performance is sensitive to Δt. Indeed, curves under SSF begin at $\Delta t = 4.5$ in the plots. For FSF, although its curves also begin at $\Delta t > 0$, we see that the starting point Δt is much smaller. Furthermore, for each policy, the L curve sees an increase at first, and then becomes quite stable over Δt. Interestingly, in both plots (a) and (b), as Δt increases, $L(\phi_{\text{FSF}})$ is always the first to become flat. This shows how insensitive FSF is to the heterogeneity of servers.

We give a brief summary to conclude this section. When FSF is adopted, the total expected number of customers in waiting is quite sensitive to the weight of fairness, but somewhat insensitive to the cost parameter. While FSF sometimes induces less congestion than other policies, in many cases it does not outperform others policies, and the performance gaps to other policies are substantial. For this reason, we do not recommend FSF unless the manager has a clear understanding of how much each server values fairness. On the other hand, under SSF, the number of customers waiting is insensitive to $\alpha_f^{(1)}$ but sensitive to Δt. In fact, it is so susceptible to the non-existence of equilibrium that the system may be unstable in many cases. Therefore, we do not recommend SSF as long as the two servers are not highly differentiated. We recommend Prop since it is considerably stable with respect to system parameters considered here, and always admits a unique equilibrium, making the system predictable and stable. We note that Prop outperforms HH in almost all cases.

3.5 Summary and Extensions

This paper studies a service system with a single arrival class and two servers. Our model has two distinguishing features. First, the routing policy not only decides the workload for both servers, but also affects servers' perception of the fairness between them. Second, capacities are endogenous as the two servers independently and simultaneously choose them after the policy is announced. Some recent papers such as Armony and Ward (2010) have also noted the fairness

issue, but most of them treat the problem with fixed capacity assumption. Only a few papers such as Gopalakrishnan et al. (2013) take the game-theoretic view as we do, but our work distinguishes itself by the explicit modeling of the unfairness term. Besides, Gopalakrishnan et al. (2013) focus on the homogeneous servers case while we model servers' heterogeneity by an extra cost term Δt.

We analyze the two-server game under a class of routing policies. More specifically, we examine four commonly seen policies. Theoretically, we prove some properties of server's objective function, and establish the existence and uniqueness of the Nash equilibrium for the Prop and HH policies. Using simulation, we further study the policies from two perspectives: (1) equilibrium/off-equilibrium game outcomes and (2) the system efficiency performance. As an empirical guidance to managerial application, we recommend the use of the Prop policy for its stability and reliability.

Our work is a first step towards a greater understanding of endogenizing fairness among servers. We list two direct extensions here. First, the N-server ($N > 2$) case awaits exploration. Although the disutility functions can be extended to the general case, doing so greatly curtails the tractability of the problem. Second, recall that the set of policies that we have considered are four specific policies. It is not clear what an optimal policy would be in general. To answer that, a study on the Stackelberg game between the administrator and the servers is in order.

Chapter 4

Uniform Pricing in Service Systems With Experience Based Quality Improvement

4.1 Introduction

How should a revenue-maximizing firm price its delay-sensitive service that is provided by multiple servers differing in their quality? Customers seeking such service usually must endure waits in addition to paying the nominal price. Consequently, when deciding whether to use the service, customers weigh the service value (typically correlated to the service quality) against the total cost incurred, which includes the paid price and the waiting cost due to congestion. A natural and commonly seen solution to the firm's problem is to price based on the service provider's quality so that the customers' surplus is fully exploited. To be more specific, in the case where the firm owns quality-differentiated servers, a price is posted for each server and customers make rational decisions on whether to procure the service and which server to queue up to if they do. Given such customers' behavior, the firm decide the set of prices for the servers to maximize the revenue. We call this pricing scheme *differentiated pricing*. Indeed, there has been a large body of literature studying this type of pricing scheme in queueing systems since Naor (1969); see Hassin and Haviv (2003) for an excellent review.

Although such scheme is used by many quality-differentiated service firms, we argue that there is a potential pitfall associated with it. Under differentiated pricing, it is plausible that the low quality server can only serve a small customer volume. For example, a majority of customers prefer senior doctors to residents in the teaching hospital, experienced hair stylists to their apprentices, famous top-rated chiropractors to the unknown ones, etc. Even if the price is set to be very low, the induced demand may still be limited. However, in many contexts such as the above examples, servers can learn by doing and improve the service quality by accumulating more experience (Gaynor et al., 2005). The plausibility of this experience-based improvement hypothesis rests on the idea of "practice makes perfect". As a server satisfies a large volume of customers' demand, he has to apply a correspondingly large amounts of care, effort and scrutiny to the service processes. Hence, he learns and makes corrective changes that result in a higher quality service in the future. However, due to differentiated pricing, the low quality server will improve very slowly, seriously impairing the firm's long-run revenue. In fact,

continuous improvement in service quality by training and servers' learning is considered as a prerequisite to firm competitiveness in service management. Letting the low quality servers serve more customers is without doubt an effective way of training and helping them improve. Hence, when a firm can and should take advantage of the experience-based quality improvement, it is an interesting question to examine whether applying differentiated pricing is a good option and whether it induces sufficient learning for the low-quality server. Moreover, to identify a better pricing scheme that benefits the firm more is of great interest.

Motivated by a nascent academic research on probabilistic (opaque) selling of products in the marketing literature, we propose another pricing scheme, *uniform pricing*, to manage the revenue of a quality-differentiated service firm. In contrast to differentiated pricing, the firm now charges a uniform price to all customers who decide to procure the service. However, the customers who enter the system are subject to a random routing to the servers with different quality levels. Once the randomness is unravelled, the customers are sent to different servers and queue separately. In this case, the firm sets a proper price and decides a routing policy, which are both publicly known to the customers. The customers, on the other hand, make *ex ante* decisions based on the expected service quality and waiting cost. In other words, the firm offers a synthetic service product that essentially amounts to a lottery among the service of all quality levels. This pricing scheme resembles the probabilistic (opaque) selling strategy introduced and studied by Fay and Xie (2008) and Jiang (2007) in retailing and travel industries. The main characteristic of such pricing strategy is for the seller to hide some attribute of a set of products and only reveal it to buyers after the payment transaction is complete. In our setting, the hidden attribute is the service quality or, equivalently, which server to queue up to. This attribute is revealed to a customer immediately after payment as the firm maintains paralleled queues.

The primary goal of this paper is to investigate the following question. In presence of experience-based quality improvement, which pricing scheme, differentiated or uniform pricing, is more beneficial to the firm's revenue management? By analytical modelling, we demonstrate that uniform pricing is indeed superior to differentiated pricing because it generates larger long-term total revenue for the firm. The underlying reason is that low quality servers improve faster under uniform pricing, and thus contribute higher revenue sooner. Our result shows that the pricing lever alone is not enough for effective service management. In fact, uniform pricing uses other types of levers such as operations lever (routing) and information lever (offering a lottery on service quality) to more effectively manage the service system. Both differentiated and uniform pricing schemes have exactly the same number of controlled variables. The former has the prices for the servers and the latter has a uniform price along with the routing probabilities. However, in term of the types of levers, uniform pricing utilizes more, resulting in a better performance. Moreover, our analysis requires very mild conditions. Relaxation of several model assumptions does not hurt the validity of our main result. Further to the main question, we also study the impact to the relative revenue advantage of the uniform pricing from two factors:

servers' learning speed and their initial heterogeneity. On the one hand, in the long run, the benefit of uniform pricing is higher when servers learn slowly compared to when they learn fast. On the other hand, depending on the servers' quality, the relative revenue difference between uniform and differentiated pricing could be either increasing or decreasing in servers' heterogeneity.

The next section reviews the background literature and positions our contributions. Section 2.2 details our base model and assumptions. Section 4.4 and 4.5 compare the two pricing schemes in the static and dynamic models, respectively. Section 4.6 explores how certain parameters affect the revenue difference, with a focus on the impact of servers' learning speed and their initial heterogeneity. Section 4.7 summarizes and concludes.

4.2 Related Literature and Positioning

Our study is primarily related to three streams of literature: economics of queues, learning curve theory, and probabilistic (opaque) selling.

Economics of Queues. We adopt the natural framework provided by queueing theory for modeling delay-sensitive service management. In our analysis, we closely follow the classic models where a monopolist charges rational customers prices for using its service where congestion is unavoidable; e.g. Naor (1969), Edelson and Hilderbrand (1975), Knudsen (1972), Mendelson (1985) and Chen and Frank (2004). In particular, models assuming that queue length is unobservable to customers are more closely related to ours. See Chapter 3 in Hassin and Haviv (2003) for a comprehensive review. Among various models along this line, not many papers have examined the issue of service quality in the context of pricing the queueing system. Anand et al. (2011) study the dependence of service quality on service duration and the resulting customers' equilibrium behavior as well as the firm's pricing decision. Assuming the similar relationship between quality and speed, Tong and Rajagopalan (2014) investigate different pricing schemes for a service provider to maximize revenue. In different contexts, other notable works that explore the issue of service quality in delay-sensitive service systems include Kostami and Rajagopalan (2013), Allon and Federgruen (2007) and so forth. Most of these papers focus on the quality that is correlated with the service capacity. Our model, however, does not require a interrelationship between service quality and capacity. More importantly, none of the previous literature along this line considers experience-based quality improvement, whereas we manage to fill this gap. These differences distinguish our work in important ways from the research of others.

Learning Curve Theory. By including learning model in our study, we want to analytically understand the process in a service operations environment for the purpose of comparing firm's pricing schemes based on their revenue performance. We do not aim to contribute to the theory of learning per se. However, we do want to be consistent with the established theory in the literature. Learning-by-doing (learning curve) has been studied mainly in manufacturing

for competitive settings (Cabral and Riordan, 1994; Dasgupta and Stiglitz, 1988; Fudenberg and Tirole, 1983) and non-competitive settings (Adler and Clark, 1991; Lapré et al., 2000). There are also some papers studying the learning effect in service contexts, e.g. healthcare (Edmondson et al., 2003) and call center setting (Gans and Zhou, 2002; Gans et al., 2010; Ryder et al., 2008).

The most common model in the literature that characterizes the rates at which learning occurs is the log-linear "learning curve" (Yelle, 1979). This type of particular functional forms has been widely used in manufacturing, which states that production cost for a product unit decreases as cumulative volume increases (Spence, 1981). In fact, there is empirical evidence that it also applies to service environment (Gustafson, 1982). Hence, we extend this model to service operations and employ a learning curve where service quality is increasing in servers' cumulative experience and asymptotically approaches to some upper limit. In this sense, Misra et al. (2004) is very close to our work. They also study experience-based learning, but in a salesforce design context. In particular, they include the decisions of both pricing and staffing, whereas the latter decision is absent in our model. Nevertheless, the customers' choice (arriving demand) is exogenous in Misra et al. (2004), but in our model, it is affected by the charged price. Endogenizing the demand rates undoubtedly incorporates an important practical factor to the model, and therefore deserves a careful study. Furthermore, unlike Misra et al. (2004) who consider differentiated pricing only, our main focus is how to avoid the potential pitfall of differentiated pricing. Finally, while some of the previous literature has also studied the counterpart effect of learning, i.e. forgetting (McCreery and Krajewski, 1999; Shafer et al., 2001), we do not model such effect. As mentioned, our focus is on the comparison of the two pricing schemes. Hence, we try to limit the sources of possible impact. Besides, including forgetting effect in our model does not provide more interesting results.

Probabilistic (Opaque) Selling. Our work also belongs to a growing body of literature that studies probabilistic (opaque) selling, whereby a seller hides some attribute of a product until after the buyers purchase it. The practice of Priceline and Hotwire exemplify such selling strategy. The earliest works in this stream include Fay and Xie (2008), Jiang (2007) and Jerath et al. (2010), which lead many followers to pursue answers to the following question: When and why is probabilistic (opaque) selling attractive to firms? Researchers from both marketing and operations management have provided several plausible explanations why firms should adopt this novel selling strategy under certain conditions. According to their findings, probabilistic (opaque) selling helps price discriminate heterogeneous customers and yields higher revenue (Fay and Xie, 2008; Jiang, 2007); it reduces mismatches between demand and supply (Gallego and Phillips, 2004; Jerath et al., 2010); it softens price competition for the more lucrative market segment (Shapiro and Shi, 2008); it exploits consumer's bounded rationality (Huang and Yu, 2014); it facilitates efficient inventory management (Elmachtoub, 2014; Fay and Xie, 2015); it profitably disposes excess capacity in a vertically differentiated market (Zhang et al., 2014). Our work contributes to this stream of literature by providing another reason to advocate

probabilistic (opaque) selling - when quality improvement is experience-based, it induces quality improvement of the low quality servers such that the long run revenue can be improved.

While the proposed uniform pricing is a form of probabilistic (opaque) selling, we recognize two important features that have been largely neglected in the previous research. First, probabilistic (opaque) selling has never been applied to a delay-sensitive service setting. This paper is the first attempt to introduce such a new pricing scheme to congestion-susceptible environment. It is both interesting and practically useful to identify new realm where probabilistic (opaque) selling is viable and even preferable. Second, with an exception of Zhang et al. (2014), only horizontally differentiated market has been investigated. In our setting, however, the differentiation is vertical. Moreover, the differentiating attribute, service quality, is not exogenous as often assumed in horizontal differentiation models. Rather, it is endogenously determined by the pricing strategy via servers' learning and improvement process.

4.2.1 Our Contributions

There are two main contributions from this paper. (1) Our study provides insights on how service systems should be managed in the presence of experience-based service quality improvement. For example, we show that the firm should adopt the proposed uniform pricing scheme rather than the regular commonly seen pricing scheme. The uniform pricing takes advantage of pricing, operations and information levers to effectively manage the system so that the service quality can be improved faster. (2) We add an important new dimension to the studies on probabilistic (opaque) selling. For the first time, this new pricing strategy is investigated in the context of service system with congestions. Besides, we are among the first few research that focus on the vertical differentiation of the product.

4.3 Basic Model and Assumptions

In this section, we describe the basic model and make some assumptions that follow through the main analysis. The basic model is the classical queueing model in a single server case (Edelson and Hilderbrand, 1975; Naor, 1969). The analysis of our main model can be reduced to that of this one. To be specific, a market of potential customers arrive at a monopolist to procure service, which is provided by a single server. Customers are homogeneous in the valuation towards the service. This service valuation, a, is a deterministic function of the server's quality s in providing the service; i.e. $a = a(s)$. Before proceeding to the formal model, we make two important assumptions here concerning the service quality.

(1) Generally speaking, the function $a(s)$ is increasing and concave (Misra et al., 2004). However, to simplify our analysis without losing insights, we assume $a(s) = s$. Considering a strictly concave valuation-quality function does not harm the validity of our results. In fact, if $a(s)$ is strictly concave in s, then uniform pricing could perform better than differentiated pricing even with fewer degree of freedom (to be explained later). To get a concise exposition,

we henceforth employ a to represent both valuation and quality, and interchangeably use the two terms.

(2) We do not require the quality to depend on service capacity. In fact, in the main analysis, we assume that service quality is independent on capacity. Besides, the service duration is not affected by quality either. We are aware that in certain contexts these two measures are correlated; e.g. Gans (2002) (congestion is a serious and costly problem) and Anand et al. (2011) (service is customer-intensive). Our analysis and results can be easily modified according to their specific relationship. However, there are many situations where the service time is not a major concern and service quality is perceived from other angles such as the server's ability, special skills and established reputation (hairstylists, chiropractors, etc). This kind of quality also plays an important role in many service systems. Moreover, the improvement on such kind of quality is closely based on server's accumulated experience. Therefore, while we acknowledge the existence of various other kinds of quality, we focus on capacity-independent service quality.

Now, we detail the the basic model by describing the customers' behavior and the firm's decision.

4.3.1 Customers' Behavior

The firm decides a price p and charges it for admission, operates at an exponential service time with mean μ, and adopts the first-come-first-serve (FCFS) priority rule. The arrival of the potential rational customers is according to a Poisson process with rate Λ. They do not know the exact queue length, but they know all the other parameters from which they can postulate the expected queue length. Hence, the customers weigh over the valuation and the total price (sum of the nominal price and the waiting cost) of the service. We focus on mixed strategy equilibrium concerning customers' queue-joining behavior. Since the customers are homogeneous, the equilibrium is symmetric where they all form a common probability with which to join the queue (Hassin and Haviv, 2003). Customers who do not join never come back and the demands are lost. As a result, the proportion of customers who join the system also forms a Poisson arrival. Let the arrival rate be $\lambda \in [0, \Lambda]$, then the average total time spent in the system is given by $W(\lambda) = 1/(\mu - \lambda)$. We assume that customers incur a congestion cost that is linear in the total waiting time with a margin c. Therefore, the proportion of joined customers is found by solving

$$a = p + cW(\lambda). \tag{4.1}$$

Typically, three market outcomes are possible - full ($\lambda = \Lambda$), partial ($0 < \lambda < \Lambda$), or zero ($\lambda = 0$) market coverage. We will closely look at the partial coverage case. In other words, the system will get too crowded that not all the customers will choose to join, i.e. $\mu < \Lambda$, and the valuation is high enough that at least one customer will join, i.e. $a > c/\mu$. The term c/μ is used because customer's waiting time only consists of the service duration, which has expected value $1/\mu$.

4.3.2 Firm's Decision

Knowing the customers' behavior, the firm chooses a price to maximize its revenue, which is simply $p\lambda$. Since in our model the congestion cost is paid by customers, the firm does not incur any cost associated to delay. That is, the firm solves the optimization problem $\max p\lambda$. The price and the served demand are related in equilibrium by (4.1), and

$$\lambda = \mu - \frac{c}{a-p}.$$

Hence, the firm's problem is

$$\text{(B')} \quad \max_{p\geq 0} p\left(\mu - \frac{c}{a-p}\right).$$

Instead of using price as a decision variable, we can alternatively formulate an equivalent optimization problem using the served demand λ as decision variable:

$$\text{(B)} \quad \max \lambda\left(a - \frac{c}{\mu-\lambda}\right)$$

$$\text{s.t.} \qquad 0 \leq \lambda \leq \mu - \frac{c}{a}.$$

The constraint in (B) is due to the fact that the corresponding price cannot be negative. The second formulation is valid because of our partial market coverage assumption. Under such assumption, the constraint $\lambda \leq \Lambda$ becomes redundant. This formulation, however, is less intuitive as the firm actually decides λ implicitly via pricing. We will work on formulation (B) as it turns out to be convenient for our analysis later. Both problems (B') and (B) are convex optimizations and are easy to solve. In addition, the partial market coverage assumption ensures an interior optimal solution. We find that the optimal price is

$$p^* = a - \sqrt{\frac{ca}{\mu}}.$$

Under this price, the corresponding served demand (market share) is

$$\lambda^* = \mu - \sqrt{\frac{c\mu}{a}}$$

and the maximized revenue is

$$r^* = p^*\lambda^* = (\sqrt{\mu a} - \sqrt{c})^2.$$

This basic model has been intensively studied in literature. For more detailed steps of the above computation, refer to Hassin and Haviv (2003).

4.3.3 Assumptions in Multi-Server Case

We now consider the monopolist firm with multiple servers who are heterogeneous in service quality. There are some previous works on pricing in multi-server environment (Bradford, 1996). We adapt those models to our setting, particularly for the ease of analysing and comparing differentiated and uniform pricing schemes. Let $N = \{1, 2, \cdots, n\}$ be the index set of all the servers, and without loss of generosity suppose $0 < a_1 \leq a_2 \leq \cdots \leq a_n < a_{max}$, where a_{max} is the maximum service quality possible. Although servers differ in quality, their capacity is the same μ. This is to apply our assumption that service quality is capacity-independent. All customers have the same valuations toward using the service provided by these servers. Moreover, the waiting cost function has exactly the same margin c for all customers using any server. Depending on the pricing schemes, the resulting models are a bit different in term of routing. Under differentiated pricing, a price is posted for each server. The customers need to decide whether to procure the service, and if yes, which server to queue in front of. Hence, the joining customers form N separate demand streams. Again, we assume that customers apply mixed strategy. Therefore, the firm actually maintains N *paralleled* M/M/1 queues and operates in the same way as it does in the single server case for each server.

Under uniform pricing, the firm announces a single price and decides the routing policy. Like the price, the routing policy is also known to all customers. However, we focus on static (not dependent on the queue length) random routings. As in the probabilistic (opaque) selling of physical products, customers only know the distribution but not the exact server routed to. In other words, the firm is essentially offering a lottery. We assume that the customers are risk-neutral and they consider the expected service value and expected waiting cost to make a decision. An important assumption regarding the implementation of uniform pricing is that firm maintains paralleled queues. That is, the firm randomly routes the customer immediately after the price is paid, and the customer queues in front of the assigned server. In contrast to this assumption, the firm could use a non-idling routing policy and have the customers wait in one line. To the customers, which server will eventually provide the service is still random in this case, but the randomness is not necessarily unravelled right after they enter the system. These two alternatives (paralleled queues and single queue) are in the same spirit as the "early commitment" and "late commitment" discussed in Fay and Xie (2015).

In our model, we adopt paralleled queues assumption ("early commitment") for two reasons. First, in order to implement a non-idling routing policy, the firm needs to monitor the system continuously in time, which is a large administrative investment. There will also be an informational burden to the customers because the policy will be more complex (e.g. it may depends on which servers are idle). Hence, routing the joining customers immediately is easier for both the firm and the customers. Second, "late commitment" gives the firm a huge operational advantage as it results in a risk pooling system. This is in itself an attractive feature. However, since differentiated pricing cannot utilize risk pooling, we assume "early-commitment" to maintain a fair comparison between these two pricing schemes.

Finally, we keep our focus on the partial market coverage outcome:

$$\frac{c}{\mu} < a_1, \quad \Lambda > n\mu. \tag{4.2}$$

The first inequality ensures that at least one customer will use the server with lowest quality under differentiated pricing. If, for example, $c/\mu > a_k$ for some $k > 1$, then no customer will choose to queue in front of servers $1, 2, \cdots, k$. As a result, there are only $N - k$ effective servers. This is potentially another disadvantage associated with differentiated pricing that can be mitigated by uniform pricing's dictated routing. However, we do not consider this simpler case. The second inequality in (4.2) means that the firm does not have enough capacity to serve the whole market. We make this assumption to ease our analysis by eliminating the boundary solution where the whole market is captured (see next section). Doing this simplifies the analysis without restricting the validity of our result.

4.4 Static Model Without Learning Effect

In this section, we analyse the differentiated and uniform pricing schemes in a static model, where servers' quality does not change over time. We will compare the two pricing schemes with respect to the generated revenue as well as the served customer volumes. To capture the essence of relevant insights, it is enough to look at a firm with just two servers. Generalization to the N-server case is easy but does not add more useful results. Therefore, let us focus on a monopolist with two servers whose services are valued as a_1 and a_2, respectively, where $0 < a_1 \leq a_2 < a_{max}$.

4.4.1 Differentiated Pricing

Under differentiated pricing, the firm sets prices p_1 and p_2 for the two servers. The customers choose whether to join a queue, and which one to join if they are joining. Resultantly, the customers form a common mixed strategy that assigns three probabilities to the options of not joining, joining server 1 and joining server 2. As mentioned, the mixed strategy equilibrium generates two streams of customers, each of which can be seen as a single server case. Formally, we establish the following lemma.

Lemma 4.4.1. *Under the partial market coverage conditions* (4.2)*, the firm's problem can be formulated as*

$$\begin{aligned} \textit{(S-Diff)} \ \max_{\mathbf{x}\in \mathrm{R}^2} r_D(\mathbf{x}) &= \sum_{i=1,2} x_i \left(a_i - \frac{c}{\mu - x_i} \right) \\ s.t. \qquad 0 \leq x_i &\leq \mu - \frac{c}{a_i}, \quad \forall i = 1, 2. \end{aligned} \tag{4.3}$$

For $i = 1, 2$, the optimal price is

$$p_i^* = a_i - \sqrt{\frac{ca_i}{\mu}},$$

the market share is

$$\lambda_i^* = x_i^* = \mu - \sqrt{\frac{c\mu}{a_i}},$$

and the generated revenue is

$$r_i^* = (\sqrt{\mu a_i} - \sqrt{c})^2;$$

thus the total revenue is $r_D^*(\mathbf{x}) = \sum_i r_i^*$.

To intuitively explain the reason why the system can be seen as two separate M/M/1 queues, note that with a common probability distribution over server 1, server 2 and leaving, no customer will take a unilateral deviation. Indeed, joining one queue with larger probability increases the congestion level and thus results in negative customer surplus, whereas the customer surplus remains zero if the not joining option has a higher probability.

4.4.2 Uniform Pricing

The firm could alternatively charge a uniform price to all customers that decide to use the service. In addition, the firm also decides a static random routing policy, which is characterized by parameters β_i, i.e. the probability that the customer is routed to server i, for $i = 1, 2$. Since $\beta_1 + \beta_2 = 1$, β_1 alone can represent the policy. The random routing is realized immediately after the customers pay the price and enter the system. As a customer is routed to one server, jockeying between queues is not allowed. As mentioned, we assume that customers are risk-neutral and hence the valuation on the service is the expectation of the service valuations from both servers. Similarly, the waiting cost is also the expected cost. As a result, given the price and the routing policy, the customers have to make ex ante decisions whether to procure the service. In this sense, the uniform pricing is in nature a probabilistic (opaque) selling strategy, where the firm offers an *opaque service*.

Let the uniform price be p_o, and the valuation to the opaque service be $\bar{a} = \beta_1 a_1 + \beta_2 a_2$. Suppose at equilibrium the customers who join the system cover a market of size $0 < \lambda_o < \Lambda$ (again, we apply assumption (4.2)), then we formulate the firm's problem in the following lemma.

Lemma 4.4.2. *Define* $\mathbf{y} = (y_1, y_2) \in \mathrm{R}^2$ *as our decision variable. The firm's problem can be formulated as*

$$\textit{(S-Unif')} \quad \max_{0 \le \mathbf{y} \le \mu} r_U(\mathbf{y}) = \sum_{i=1,2} y_i \left(a_i - \frac{c}{\mu - y_i} \right)$$

$$s.t. \qquad \sum_{i=1,2} y_i a_i \ge \sum_{i=1,2} \frac{c y_i}{\mu - y_i}. \tag{4.4}$$

Given a solution $\mathbf{y}$*, we can recover the routing policy, the uniform price and the market share*

by the following transformation.

$$\begin{cases} \beta_i = \dfrac{y_i}{y_1 + y_2} \quad i = 1, 2; \\ p_o = \displaystyle\sum_{i=1,2} (\beta_i a_i - \frac{c\beta_i}{\mu - y_i}); \\ \lambda_o = y_1 + y_2. \end{cases}$$

In this lemma, we follow the spirit of formulation (B) and do not use price as the decision variable. Rather, we optimize over the two induced demand streams. This is because that the price and the total served demand are entangled in one equation and it is hard to represent one in term of the other. This is why we prefer formulation (B) to (B').

4.4.3 Comparison

This subsection compares the revenue generated by the differentiated and uniform pricing in the static model and comments on the comparison. First of all, we take a closer look at the problem (S-Unif') and solve for its optimal solution. One important observation is that the constraint (4.4) is redundant.

Lemma 4.4.3. *Problem (S-Unif') is equivalent to*

$$\textit{(S-Unif)} \quad \max_{0 \le \mathbf{y} \le \mu} r_U(\mathbf{y}) = \sum_{i=1,2} y_i \left(a_i - \frac{c}{\mu - y_i} \right).$$

Due to the above lemma, we just need to compare problems (S-Diff) and (S-Unif). They have exactly the same objective function, but the feasible set of (S-Unif) is clearly larger than that of (S-Diff). Consequently, the optimal revenue under uniform pricing is no worse than that under differentiated pricing. That is,

$$r_U^* \ge r_D^*. \tag{4.5}$$

In fact, because both problems are convex programs, and the optimal solution to (S-Diff), $\mathbf{x}^*$, is an interior point, we deduce that $\mathbf{x}^*$ also maximizes $r_U(\mathbf{y})$. Another more direct way to see this is simply by solving (S-Unif). To be specific, we observe that $r_U(\mathbf{y})$ is strictly concave in $\mathbf{y}$ and is separable with respect to its components. By solving the first-order conditions, we have $\mathbf{y}^* = \mathbf{x}^*$ and the equality in (4.5) holds. Here we stress the importance of the assumption (4.2), especially that the low quality server attracts at least one customer ($a_1 > c/\mu$). Had this assumption been removed, $\mathbf{x}^*$ would not be an interior point any more and r_U^* would be strictly larger. With assumption (4.2), the two pricing schemes yield to the same revenue in a static maximization problem. We write this formally as our first result.

Proposition 4.4.4. *If the firm only considers a static pricing problem, then differentiated and uniform pricing are equivalent at optimality. They induce the same market coverage, assign the*

same amount of customer volume to the servers, and generate the same revenue for the firm.

We make three remarks with respect to the comparison. (1) both pricing schemes leave the firm the same number of decision variables: two prices for differentiated pricing and a price and a routing probability for uniform pricing. Hence, in term of the degrees of freedom, the two schemes are truly alternative and the comparison is fair. However, uniform pricing utilizes two types of management levers, which are pricing lever (the uniform price) and operations lever (routing policy). In contrast, differentiated pricing only applies pricing lever. Although the firm can use price to affect customer's choice, which in turn determines the distribution of customer volumes over the two servers, the impact is indirect compared to directly routing the customers. (2) The intuition behind Proposition 4.4.4 is the mutual mimicking of the two pricing schemes. For any customer volumes differentiated pricing induces for the two servers, uniform pricing can induce exactly the same size and distribution of the market share. The way is to use the pricing and routing levers mentioned above. The converse is also true for differentiated pricing. (3) Although in the one-period model the equivalence of the two pricing schemes seems quite intuitive based on the last remark, this intuition does not work once we consider multiple-period problem. Specifically, the inequality (4.5) could be strict once we consider the learning effect, which is discussed in the next section.

4.5 Dynamic Model With Learning Effect

In the presence of experience-based service quality improvement, how does the long term total revenue generated by uniform pricing compare to that of differentiated pricing? We answer this question in this section by considering a multiple-period model where the firm dynamically prices the service using either differentiated or uniform pricing. The dynamic in this system is the quality improvement due to servers' learning as they serve more customer volume. The learning curve is assumed to be increasing and concave with an asymptotic upper bound. Hence, more customers served by a server in the current period means a larger quality improvement in the next period. We assume that the time for the server to learn and improve quality is small compared to a single period. Therefore, the quality improvement is accomplished before the new pricing season starts. The firm's goal is to maximize the total revenue. The trade-off is between a higher current revenue (with smaller customer volume) and improving the service quality sooner to generate higher revenue in the future.

Our results show that uniform pricing is advantageous in that it generates more total revenue than differentiated pricing does. To give a foretaste of why uniform pricing is better, let us consider the following example. Consider a single-period model again, and let $\beta_1 = \beta_2 = 1/2$. This means that the routing policy is to send customers to servers with equal probability (pure randomly). We believe that this policy is one of the simplest to implement, and it avoids much administrative and informational burden to both the firm and the customers. As mentioned in the previous section, uniform pricing utilizes both pricing lever and operations lever to manage

the system. As such, dictating a half-half routing is to directly affect the operational measures in the system. Denote the customer volume at optimality by $\bar{\lambda}_i$ for server 1 and 2, respectively. The expected quality $\bar{a} = (a_1 + a_2)/2$. Then,

$$\bar{\lambda}_i = \mu - \sqrt{\frac{c\mu}{\bar{a}}} \quad i = 1, 2.$$

Compare to the customer volumes under differentiated pricing, we find that $\bar{\lambda}_1 + \bar{\lambda}_2 > \lambda_1^* + \lambda_2^*$ and $\bar{\lambda}_1 > \lambda_1^*$ (by monotonicity and concavity of the above function). Therefore, although uniform pricing with half-half routing dilutes the revenue, the low quality server (server 1) gets more customer volume to serve. This is advantageous in the far-sighted sense because the low quality server will become much better in the future, helping the firm gain more revenue. Moreover, the customers that high quality server gets is less ($\bar{\lambda}_2 < \lambda_2^*$); but this will not offset the benefit of training the low quality server because the learning curve is assumed to be strictly concave. In other words, the higher quality servers have no more room for large improvement, so why not send some of their customers to train lower quality servers? This shows the power of operations lever that uniform pricing can utilize. In the following, we formally build the argument that uniform pricing is preferred when learning effect exists.

4.5.1 Model Formulation

First of all, let us describe the function that characterizes the learning effect. Write this quality improvement function $L = L(a, z)$, where a is the server's starting quality and z represents the accumulative experience. Much of the literature on learning curves in the manufacturing contexts uses units of output as a measure of experience (Fudenberg and Tirole, 1983; Spence, 1981). In the service settings such as ours, a plausible candidate to measure experience is the customer volume. Hence, suppose that z is given by the total number of customers served from the beginning. As mentioned, the function $L(a, z)$ is assumed to be increasing and concave in z. Furthermore, it approaches to an upper bound as z becomes larger. Denote this upper bound by a_{max}, which is the maximum service quality a server could possibly achieve. Hence, based on these assumptions, a proper improvement function has the form

$$L(a, z) = a_{max} - (a_{max} - a)e^{-\delta z}, \tag{4.6}$$

where $\delta > 0$ is the learning speed. The larger δ is, the faster the server can learn and thus improve his quality sooner. In general, it is possible that the learning structure differs from server to server (Gans et al., 2010). However, we assume that the servers have a common learning curve with the same a_{max} and δ, because our results will not change without this assumption (to be explained later).

Consider a finite horizon dynamic model with time period $t = 1, 2, \cdots, T$. We add another subscript t to all the relevant notations to represent time. Then, based on the above discussion,

we have

$$a_{i,t+1} = L(a_{i1}, \sum_{k=1}^{t} \lambda_{ik}), \quad i = 1, 2;\ t = 1, \cdots, T.$$

In the above, λ_{ik} is the customer volume served by server i in period k. Due to the log-linear functional form of the learning curve, we have a memoryless property for the quality improvement. The service quality in period $t+1$ is equivalently determined by the quality and the experience gained in period t, i.e.

$$a_{i,t+1} = L(a_{it}, \lambda_{it}), \quad i = 1, 2;\ t = 1, \cdots, T.$$

Given the dynamics of learning and quality improvement, the firm sets prices at each period to maximize the total revenue along the planning horizon. Note that we could consider the discounted total revenue, but there is no difference in nature as it is a finite horizon problem. The firm uses either differentiated or uniform pricing. We assume that there is no switching between these two schemes during the entire planning horizon. To formulate the problems, we take the similar approach as in the single period model. That is, the decision variables are the customer volumes served by the servers.

Differentiated Pricing. Let the $2 \times T$ matrix $\mathbf{X} = (x_{it})$ be the decision variable, where x_{it} represent the customer volume of server i at time period t. Then, the firm's problem is

$$\text{(M-Diff)}\ \max R_D(\mathbf{X}) = \sum_{t=1}^{T} \sum_{i=1,2} x_{it} \left(a_{it} - \frac{c}{\mu - x_{it}} \right)$$

$$\text{s.t.} \qquad 0 \le x_{it} \le \mu - \frac{c}{a_{it}}, \quad \forall i, t; \tag{4.7}$$

$$a_{i,t+1} = L(a_{it}, x_{it}), i = 1, 2;\ t = 1, 2, \cdots, T-1.$$

Uniform Pricing. In each time period t, let p_{ot} be the uniform price and β_{it} be the routing probabilities. Our decision variable is the $2 \times T$ matrix $\mathbf{Y} = (y_{it})$, where y_{it} is the customer volume of server i at time period t. Hence, the total customer volume is given by

$$\lambda_{ot} = \sum_{i=1,2} y_{it} \beta_{it}.$$

With a straightforward generalization from the single period model, the firm' problem is

$$\text{(M-Unif)}\ \max R_U(\mathbf{Y}) = \sum_{t=1}^{T} \sum_{i=1,2} y_{it} \left(a_{it} - \frac{c}{\mu - y_{it}} \right)$$

$$\text{s.t.} \qquad \sum_{i=1,2} a_{it} y_{it} \ge \sum_{i=1,2} \frac{c y_{it}}{\mu - y_{it}}, \forall t; \quad 0 \le y_{it} \le \mu, \forall i, t; \tag{4.8}$$

$$a_{i,t+1} = L(a_{it}, y_{it}), i = 1, 2;\ t = 1, 2, \cdots, T-1.$$

Given a solution $\mathbf{Y}$, the price at period t is recovered by

$$\begin{cases} \beta_{it} = \dfrac{y_{it}}{y_{1t} + y_{2t}} & i = 1, 2; \\ p_{ot} = \displaystyle\sum_{i=1,2} (a_{it}\beta_{it} - \frac{c\beta_{it}}{\mu - y_{it}}). \end{cases}$$

Note that problem (M-Unif) is essentially

$$\max R_U(\mathbf{Y}) = \sum_{t=1}^{T} p_{ot}\lambda_{ot}.$$

Once we take into account the learning effect, the firm will have incentive to induce large demand in the early periods. Based on the customers' behavior, lower price always leads to larger demand. Hence, the constraint (4.8) is to make sure that p_{ot} is non-negative for every time period t. As a result, unlike in the single period model, the constraint (4.8) is not redundant and cannot be dropped for free.

4.5.2 Comparison

Similar to the observation in the single period model, here we find that the objective functions in (M-Diff) and (M-Unif) are exactly the same. In addition, condition (4.7) implies condition (4.8). To see this, suppose $\mathbf{X}$ satisfies (4.7), then automatically

$$0 \le x_{it} \le \mu, \, \forall i, t.$$

In addition, rewrite (4.7) as

$$a_{it} \ge \frac{c}{\mu - x_{it}}, \tag{4.9}$$

and note $x_{it} \ge 0$. Finally, multiplying both sides of (4.9) by x_{it} and summing over i shows that $\mathbf{X}$ also satisfies condition (4.8). Therefore, the feasible set of problem (M-Unif) is larger than that of (M-Diff), resulting in our main result.

Theorem 4.5.1. *If the firm considers the learning and quality improvement effect over multiple periods, then the total revenue generated by differentiated pricing is no higher than that generated by uniform pricing. More precisely,*

$$R_U(\mathbf{X}^*) \ge R_D(\mathbf{Y}^*). \tag{4.10}$$

The intuition behind Theorem 4.5.1 is that no matter what size of customer volumes differentiated pricing gets, uniform pricing can mimic by inducing the same total demand and distributes it to the servers using proper routing probabilities. However, different from Proposition 4.4.4, the converse of the previous statement is not true. In particular, for the low quality server, uniform pricing can generate a large customer volume that differentiated pricing is not

able to. Indeed, if $a_{1t} < a_{2t}$, then $a_{1t} < \bar{a}_t$ for any routing policy where $\beta_{1t} > 0$. Hence, a low quality a_{1t} places much limitation on the largest demand differentiated pricing can possibly induce for server 1; uniform pricing mitigates this situation by only showing customers the expected quality. Therefore, the inequality in (4.10) could be strict. Take the following three-period model for an example. The initial service values are $a_{11} = 1, a_{21} = 8$. The learning curve has parameters $a_{max} = 24$ and $\delta = 1$. Then, solving (M-Diff), we have the total maximized revenue 37.33; solving (M-Unif) gives total revenue 44.68, which is about 20% larger. Note that in the first period, the optimal uniform pricing scheme induces 0.76 customer volume for server 1, while the optimal differentiated pricing only induces 0.2 customer volume for server 1, which is exactly the upper bound for x_{11} in the problem (M-Diff). Consequently, some of the inequalities in (4.7) will be violated by the optimal $\mathbf{Y}^*$, resulting a strictly larger objective function.

As mentioned previously, the strict superiority of uniform pricing in the dynamic model proves that using pricing lever alone could be an inefficient service management. Both pricing schemes optimize over the same number of variables ($2T$). However, uniform pricing gives up some degree of freedom from the pricing lever but makes up by yielding to routing, an important operations lever. Moreover, by random routing, uniform pricing also has some control over customers expectation towards the service quality. By making the observed servers' quality "opaque", uniform pricing actually also plays with the information lever to manage the queueing system. In contrast, all the variables in differentiated pricing are prices. In term of the type of lever, differentiated pricing resorts to only one type. Therefore, an important insight of our analysis is that the combination of various types of management levers can be more powerful than just applying one type.

4.5.3 Discussions

In this section, we comment on our model from two viewpoints: formulation and robustness.

Dynamic Programming Formulation. In the above, the problems (M-Diff) and (M-Unif) are formulated as non-linear optimizations. The concavity of the objective function is not clear. Alternatively, we can also formulate the optimization problems using deterministic dynamic programming. Let us write variables in column vector (whose components correspond to the two servers), and the single subscript represents time. The state space is service quality $0 \leq \mathbf{a}_t \leq a_{max}$. The action space is the induced customer volume $0 \leq \mathbf{z}_t \leq \mu$. The dynamic is the quality improvement, which remains the same as described:

$$\mathbf{a}_{t+1} = L(\mathbf{a}_t, \mathbf{z}_t).$$

With a little abuse of notation, here $L(\cdot, \cdot)$ is computed component-wise. Let

$$r(\mathbf{a}_t, \mathbf{z}_t) = \sum_{i=1,2} z_{it} \left(a_{it} - \frac{c}{\mu - z_{it}} \right)$$

be the immediate reward of taking action $\mathbf{z}_t$ at period t when the service quality is $\mathbf{a}_t$. In addition, let $R_t(\mathbf{a}_t)$ be the total optimal revenue from period t onward when the current service quality is $\mathbf{a}_t$; set $R_{T+1} = 0$. Then, the optimality equation is given by

$$\text{(DP) } R_t(\mathbf{a}_t) = \max_{\mathbf{z} \in A_t} \{r(\mathbf{a}_t, \mathbf{z}_t) + R_{t+1}(\mathbf{a}_{t+1})\}. \tag{4.11}$$

The feasible action set A_t is different in the two pricing schemes. Specifically, under differentiated pricing, A_t is given by a product set

$$A_t^D = \Pi_{i=1}^n \left\{ 0 \leq z_{it} \leq \mu - \frac{c}{a_{it}} \right\},$$

whereas under uniform pricing,

$$A_t^U = \left\{ \sum_{i=1,2} a_{it} z_{it} \geq \sum_{i=1,2} \frac{c z_{it}}{(\mu - z_{it})}, 0 \leq z_{it} \leq \mu \right\}.$$

As shown previously, it is not hard to establish that $A_t^D \subset A_t^U$. Resultantly, Theorem 4.5.1 holds true through this formulation too.

In contrast to the first formulation, we are able to derive some second-order properties for the objective functions in (DP).

Proposition 4.5.2. *Let $J_t(\mathbf{a}_t, \mathbf{z}_t) = r(\mathbf{a}_t, \mathbf{z}_t) + R_{t+1}(\mathbf{a}_{t+1})$, then*

(a) $J_t(\mathbf{a}_t, \mathbf{z}_t)$ is concave in $\mathbf{z}_t$ for all t.

(b) For each t, $R_t(\mathbf{a}_t)$ is concave in $\mathbf{a}_t$ and increasing with respect to a_{it} $(i = 1, 2)$.

Robustness to Model Assumptions. Our model is indeed established on many simplifying assumptions. However, most of these assumptions are made because not only they ensure succinct analysis but also relaxing them yields little additional insights. In other words, our main results remain valid if we relax those assumptions.

First, the servers can be heterogeneous in many dimensions besides quality. They can be different in the service capacity μ, the unit congestion cost c, the maximum achievable quality a_{max} and the learning speed δ. Clearly, applying these changes results in the same conclusion concerning the comparison of the two pricing schemes as they will still have the same objective function.

Second, functional structure changes have little impact on our results. For example, the congestion cost can be increasing and strictly convex. Then Theorem 4.5.1 still holds; and the proof of Proposition 4.5.2 follows through by applying the composition rules of increasing convex functions. We can also relax the assumption that equalizes service quality to customers' valuation on the service. That is, instead of $a(s) = s$, we assume that $a(s)$ is increasing and strictly concave (Misra et al., 2004). Then, this change, again, only mildly affects the proof of

Proposition 4.5.2. The optimal revenue at period t, R_t, is still increasing and jointly concave in service quality. To note another interesting impact of this assumption change, the function $a(s)$ being strictly concave actually makes uniform pricing more preferable. As mentioned, uniform pricing offers the customers an service with "opaque" quality, and the customers make decisions based on the expected quality. Hence, the valuation to an opaque service, $a(\bar{s})$, will be higher than the expectation of observed valuations, $\overline{a(s)}$, due to concavity. In fact, with the simple half-half routing policy (described in the beginning of this section), uniform pricing can perform better than differentiated pricing if $a(s)$ is strictly concave[1].

Third, the assumption that service quality and service capacity are independent with each other could be unsettling to some managers. Indeed, in some situations higher quality means larger capacity (Gans, 2002), whereas in some other situations longer service time leads to a perception of higher quality (Anand et al., 2011). In both cases, capacity can be written as $\mu = \mu(a)$. No matter how the function $\mu(a)$ looks like, once the problems are formulated, the objective functions under the two pricing schemes always coincide. Therefore Theorem 4.5.1 holds. As for Proposition 4.5.2, result (b) holds if $\mu(a)$ is increasing and concave in a; otherwise it may fail to be true. To see this, we only need the function

$$v(a) := a - \frac{c}{\mu(a) - x}$$

to be increasing and concave in a. Monotonicity follows immediately as $\mu(a)$ is increasing. Note that

$$v''(a) = \frac{2(\mu'(a))^2 - \mu''(a)(\mu(a) - x)}{(\mu(a) - x)^3} \geq 0,$$

as desired. The condition that service capacity is increasing and concave in server's quality is met in many situations, and therefore our results are still valid. Even when $\mu(a)$ is decreasing in a, the fact that uniform pricing is still advantageous to differentiated pricing is unchanged.

To sum up, the model are robust to most of the assumptions, changing of which keeps the objective functions under both pricing schemes identical. However, there are assumptions that do affect our analysis significantly. For example, if the customers are risk averse rather than risk neutral, then uniform pricing will suffer from a loss due to its "opacity". As a result, it is not clear which pricing scheme is better.

4.6 How Do Parameters Affect the Revenue Difference?

In the previous section, we established that, compared to differentiated pricing, uniform pricing generates equal or higher total revenue. This section further investigates how the revenue difference changes as some of the parameters change. First of all, we specify the way we

[1]The construction of $a(s)$ is available upon request.

measure the difference. Define

$$\Delta = \frac{R_U^* - R_D^*}{R_D^*} \times 100\%,$$

i.e. the relative revenue advantage of uniform pricing compared to differentiated pricing. It is the percentage increase in revenue if the firm uses uniform pricing instead of differentiated pricing. We will use Δ as the measure of the comparison between the two pricing schemes. Apparently, $\Delta \geq 0$ and larger Δ means that uniform pricing possesses more advantage.

Basically, we are interested in the impact to the revenue advantage Δ from three factors. To be specific, we look at the change in Δ with respect to the change of the planning horizon (T), the servers' learning speed (δ) and the initial heterogeneity of the two servers ($a_{21} - a_{11}$). These parameters are closely related to the servers' learning process (length and speed) as well as the "opacity" rendered by the uniform pricing, which are the main characteristics that account for the revenue advantage. Hence, a study on how they affect the comparison is both interesting and useful. Rather, other parameters such as congestion cost c, capacity μ and maximum quality level a_{max} are fixed. Since we only look at the servers' heterogeneity at the beginning, we drop the time period subscript of quality levels for the sake of a simpler exposition. That is, we write a_1 and a_2 instead of a_{11} and a_{21} in this section. Therefore, we are interested in the revenue advantage as a function of several parameters in the problem, i.e.

$$\Delta = \Delta(T, \delta, a_1, a_2).$$

Depending on the context, which will be made clear as we develop our analysis, we omit all the fixed parameters in the bracket and just write the varying ones. For most of the following analysis, we compute Δ numerically. Algorithms are implemented in Matlab and are run on a standard PC. In some cases, however, we also develop analytical results based on simplifying assumptions.

Our results consists of three parts. First, we study the asymptotic behaviors of Δ with respect to T and δ. We show that the revenue advantage vanishes as the planning horizon or the learning speed is growing large. Second, we find that whether larger learning speed leads to more revenue advantage depends on the length of the planning horizon. Third, we answer the question how heterogeneity ($a_2 - a_1$) affect Δ. Although changing either a_2 or a_1 can equivalently change their difference, we show an interesting finding where the two ways of changing drives the revenue advantage in the opposite directions.

4.6.1 Asymptotic Behavior

In this subsection, we analytically characterize the asymptotic behavior of the revenue advantage Δ as the learning process becomes either very long or very fast.

Proposition 4.6.1. *Fix a_1 and a_2.*

(a). For every $\delta > 0$, $\Delta(T) \to 0$ as $T \to \infty$.

(b). For every $T = 1, 2, \cdots$, $\Delta(\delta) \to 0$ *as* $\delta \to \infty$.

Proposition 4.6.1 simply states that the revenue advantage enjoyed by uniform pricing is vanishing as the planning horizon or the servers' learning speed approaches to infinity. Therefore, the situation where the firm should use uniform pricing and take the advantage is when the total time period is not too large and when servers improve the quality not too fast.

We make a couple of remarks regrading this result. (1). It is obvious that $\Delta = 0$ if $\delta = 0$, i.e. there is no learning effect; and by Proposition 4.4.4, $\Delta = 0$ when $T = 1$. Hence, there is no monotone relationship overall for the two parameters T and δ. (2). Although $\Delta \to 0$ when either T or δ approaches to infinity, the underlying reasons for this same asymptotic behavior are different. When T becomes large, the absolute difference $R_U^* - R_D^*$ approaches to a constant and R_D^* grows unbounded. In contrast, when δ gets large, the gap between R_U^* and R_D^* is closing, so $R_U^* - R_D^*$ approaches to zero. Consequently, had we considered the *discounted* total revenue as the objective, Δ would approach to a *positive* constant as $T \to \infty$, because R_D^* would be bounded by a finite number. However, result (b) remains the same.

4.6.2 Impact of Learning Speed

According to Proposition 4.6.1 (b), the revenue advantage is eventually dropping to zero as δ approaches infinity. Nevertheless, let us suppose that the servers' learning speed is not very large in this subsection. Does increasing δ increase Δ? In other words, does it benefit the firm more when using uniform pricing instead of differentiated pricing if the servers learn fast? Since uniform pricing takes the advantage of the learning effect and the quality improvement process, the intuitive answer to the above questions is yes. Surprisingly, we find that this only holds true when the planning horizon is short. If T is large (not necessarily very large), however, then Δ could be negatively correlated with the learning speed even though δ is far from being very large. We will first analytically show that, when T is small, $\Delta(\delta)$ is non-decreasing based on an approximation of the learning curve and some simplifying assumptions. Then, we present some numerical results that demonstrate the dependence of the shape of $\Delta(\delta)$ on T.

Monotonicity in a Simplified Case. Although the log-linear learning curve (4.6) is mostly seen in the literature, there are works that employ a linear approximation (Dasgupta and Stiglitz, 1988; Jin et al., 2004). To develop tractable analytical result, we also approximate servers' learning by the following piecewise linear curve. Let $\mathbf{z}_t$ be the custom volume for the server at time period t, then

$$\mathbf{a}_{t+1} = \min\{\mathbf{a}_{max}, \mathbf{a}_t + \delta \mathbf{z}_t\}. \tag{4.12}$$

Moreover, for this simplified case, we make several additional assumptions. Define

$$\tilde{\delta} := \frac{a_{max} - a_1}{\mu}.$$

We assume that (1) $\delta \leq \tilde{\delta}$, i.e. the learning in the first period is not enough for the low quality

server to achieve maximum quality in the second period, (2) $a_2 = a_{max}$ so only server 1 needs to learn, and (3) $T = 2$.

Proposition 4.6.2. *Suppose the above assumptions are satisfied and* (4.12) *holds, then* $\Delta(\delta)$ *is non-decreasing in* δ.

In Proposition 4.6.2, we are looking at a short planning horizon. Note that $\tilde{\delta}$ is relatively very large (see the numerical study below), yet having a faster learner always gives uniform pricing more revenue advantage as long as $\delta \leq \tilde{\delta}$. Indeed, short planning horizon and server's learning fast help the uniform pricing exploit the advantage of stimulating quick and large quality improvement by routing. However, it turns out that the validity of this insight depends very much on T. As the planning horizon gets just a few periods longer, $\Delta'(\delta)$ becomes negative even for relatively small δ. We show this numerically.

Numerical Study. For *all* the numerical studies in this paper, we fix servers' capacity, the unit waiting cost, and the maximum quality to be

$$\mu = 1, \quad c = 0.8, \text{and} \quad a_{max} = 24.$$

For this subsection only, we suppose that servers' initial quality levels are exogenous; $a_1 = 1$ and $a_2 = 7$. We have tested that varying a_1 and a_2 does not affect our main result. We plot $\Delta(\delta)$ over the interval $0.5 \leq \delta \leq 1.5$ for several different values of T. See Figure 4.1 below. From the

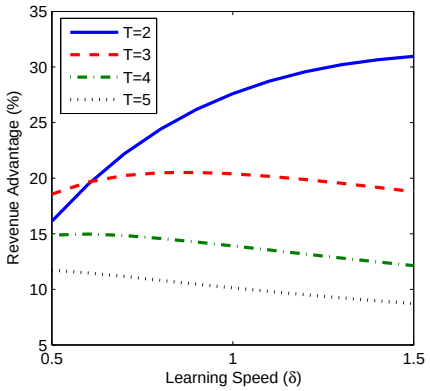


Figure 4.1: The impact of servers' learning speed on the revenue advantage: Dependence on T. Fix $a_1 = 1$ and $a_2 = 7$.

plot, we observe that for any δ_0, $\Delta'(\delta_0)$ is decreasing in T. For example, for $\delta_0 = 0.8$, $\Delta'(\delta_0)$ is positive when $T = 2$, almost zero when $T = 3$, and negative when $T = 4$ or 5. This shows that, even if servers are not learning in a very high speed, the sign of $\Delta'(\delta)$ is quite sensitive to how long the planning horizon is. Hence, counter-intuitively, when the servers learn *slowly* and T is not small, the firm could incur more potential revenue loss if it uses differentiated pricing rather than uniform pricing. This is an interesting result because it illustrates the advantage of uniform pricing in a situation where one would not expect it to possess. For example, when

$T = 5$, the best revenue advantage is achieved when the learning speed is minimum ($\delta = 0.5$). Moreover, our numerical study also confirms the analytical result in Proposition 4.6.2. For $T = 2$, the revenue advantage is increasing in δ; and for $T = 3$, there is a threshold $\bar{\delta} \approx 0.75$ that Δ is increasing if $\delta \leq \bar{\delta}$.

4.6.3 Impact of Heterogeneity

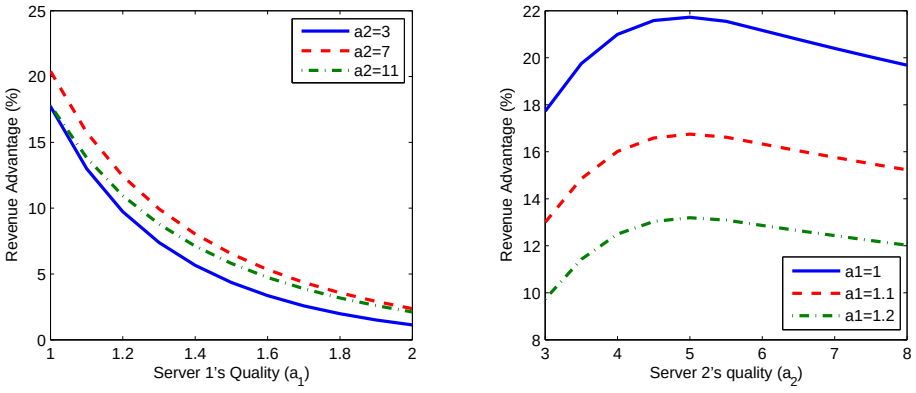
By numerical studies, this subsection investigates how servers' initial heterogeneity affects the revenue advantage of uniform pricing. We fix the planning horizon $T = 3$ and learning speed $\delta = 1$. Note that the cases presented below are the representative instances from our extensive computations. The results are valid with other different values of T and δ.

Define the initial heterogeneity of the two servers as

$$h = a_2 - a_1.$$

Since the uniform pricing poses a certain degree of "opacity" to the customers, who are assumed to be risk neutral, it seems plausible that larger heterogeneity leads to more revenue advantage. However, note that the heterogeneity h can be changed by changing either a_2 or a_1, we show that the two ways of varying heterogeneity lead to significantly different impacts on revenue advantage Δ.

For the two plots in Figure 4.2, we vary one quality with three fixed other quality respectively. In Figure 4.2 (a), $\Delta'(a_1) < 0$ for all three curves, which means that Δ is increasing



(a) Varying a_1 with fixed a_2 (b) Varying a_2 with fixed a_1

Figure 4.2: The impact of servers' heterogeneity on the revenue advantage. Fix $T = 3$ and $\delta = 1$.

with h. This corresponds to our intuition: by narrowing down the gap between the quality levels of the two servers, the uniform pricing enjoys less advantage. However, in Figure 4.2 (b), we observe the non-monotonicity of Δ in h, which results in the following interesting finding:

There exists a pair of initial quality $(\tilde{a}_1, \tilde{a}_2)$ such that

$$\frac{\partial}{\partial a_1}\Delta(\tilde{a}_1, \tilde{a}_2) < 0; \quad \frac{\partial}{\partial a_2}\Delta(\tilde{a}_1, \tilde{a}_2) < 0. \tag{4.13}$$

That is, although increasing $\tilde{a}_2$ can increase h by the same amount as decreasing $\tilde{a}_1$, the corresponding driving forces on Δ can be in the opposite directions; the former decreases Δ whereas the latter increases it. From Figure 4.2, it is easy to verify that the quality pair where $\tilde{a}_1 = 1.2$ and $\tilde{a}_2 = 7$ satisfies (4.13).

Therefore, simply noting the relative difference in quality (h) is not enough to address the full insight of the impact on revenue advantage from servers' heterogeneity. It also largely depends on the absolute value of the servers' quality. To be more precise, the servers' *potential of improvement* matters. Basically, given a fixed initial quality difference, there is more revenue advantage when *both* servers are at the early stage of learning, i.e. a_1 and a_2 are small and there are more room for their improvement by learning. Take Figure 4.2 (b) for example. If server 2's quality $a_2 = 8$, then Δ is larger when a_2 decreases to 6, which means that server 2 is at an earlier stage of learning. Of course, this claim cannot be valid once a_2 is too small because, after all, the relative measure h also affects Δ in its own right. Alternatively, we present Figure 4.3 to show how servers' being at earlier stage of learning benefits uniform pricing's revenue advantage. For

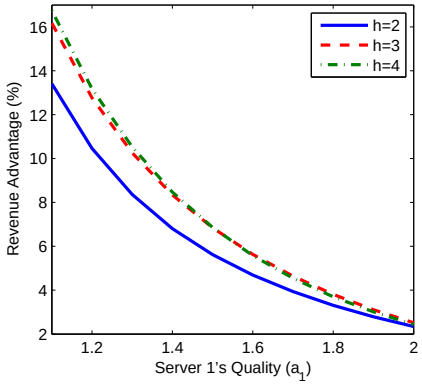


Figure 4.3: The importance of servers' potential of improvement. Fix $T = 3$ and $\delta = 1$.

three fixed values of heterogeneity h, we vary the quality level a_1 and plot the corresponding Δ. Clearly, all three curves are decreasing, which shows that the same amount of heterogeneity provides less revenue advantage as servers climb up the learning curve, i.e. as their learning potential decreases. Moreover, servers' potential of improvement also affects the sign of $\Delta'(h)$. Generally, $\Delta'(h) > 0$ if a_1 is small whereas $\Delta'(h)$ may be negative when a_1 becomes large. In Figure 4.3, for example, the dash curve ($h = 3$) is below the dot-dash curve ($h = 4$) when $a_1 = 1.2$; however, when $a_1 = 1.8$, their positions are reversed.

To conclude this section, we give a brief summary. Since uniform pricing generates equal or higher total revenue than differentiated pricing does in the multiple period problem, this section

tries to understand how the revenue difference between these two pricing schemes is affected by certain parameters. Focusing on the percentage increase of the revenue, we investigate its asymptotic behavior as the planning horizon becomes long or servers learn and improve at very high speed. Under both cases, the revenue advantage vanishes. We also particularly study the impact of servers' learning speed and their heterogeneity in the beginning. Depending on how long the planning horizon is, the revenue advantage and the learning speed could be either positively or negatively correlated. Interestingly, when the planning horizon is long, the uniform pricing is even more advantageous if the servers learn slowly compared to if they learn fast. The impact of the servers' heterogeneity cannot be fully characterized just by the relative difference of their quality levels. Rather, servers' absolute quality, i.e. their potential of improvement, also decides which direction a change on heterogeneity drives the revenue advantage. These results provides some guidelines for the firm to estimate the benefit from using uniform pricing in the environment with varying factors such as servers' learning speed and heterogeneity.

4.7 Conclusion

For firms who own several servers to provide quality-differentiated time-sensitive services, the traditional pricing strategy is to base the price on the customers' valuation, which is correlated to the service quality. However, if the quality improvement process depends on servers' accumulated experience, then this differentiated pricing can only induce a limited customer volume for the low quality servers, which hinders their learning and improvement process. Consequently, the revenue will also be curtailed. This paper proposes another pricing scheme, uniform pricing, as an efficient tool for the revenue management of a quality-differentiated service firm. This pricing scheme resembles the probabilistic (opaque) selling strategy studied in the marketing literature (Fay and Xie, 2008; Jiang, 2007) as it hides from the customers the ex post service quality and presents a lottery to the customers. Hence, it affects their choices via an information lever. More importantly, the uniform pricing takes the advantage of a powerful operations lever, i.e. routing. As a result, when comparing the revenue under the two pricing schemes, although they have the same number of decision variables, uniform pricing utilizes more types of levers than differentiated pricing does. This turns out to be the main reason why uniform pricing is better.

The two pricing schemes are equivalent in a single period model. However, the advantage of uniform pricing manifests itself in the multiple-period pricing problem. We show theoretically that the total revenue under uniform pricing is equal or higher than that under differentiated pricing. Furthermore, this result holds true under considerably many alternative assumptions; but some key assumptions are crucial and cannot be changed. By numerical studies, we also investigate the impact of several parameters to the relative revenue advantage of uniform pricing. Specifically, we look at the impact of the length of the planning horizon, servers' learning speed and their heterogeneity. Our studies yield to some interesting results that are against

intuition. There are some possible extensions to our work. We list two of them. First, there are other ways to model customers' choice besides what we use in the model. Specifically, if the congestion cost is incurred to the firm rather than the customers, then we will need another choice model to characterize customers' decisions on service procurement. Second, this paper considers the monopolist case, which sheds light on the firm's pricing decisions without external forces. However, it would be interesting to consider the pricing schemes in a competitive environment where firms interact.

Bibliography

J. S. Adams. Inequity in social exchange. In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 2, pages 267–299. Academic Press, New York, 1965.

P. S. Adler and K. B. Clark. Behind the learning curve: A sketch of the learning process. *Management Science*, 37(3):267–281, 1991.

A. Alkan, G. Demange, and D. Gale. Fair allocation of indivisible goods and criteria of justice. *Econometrica: Journal of the Econometric Society*, 59(4):1023–1039, 1991.

G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55(1): 37–55, 2007.

K. S. Anand, M. F. Pac, and S. Veeraraghavan. Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science*, 57(1):40–56, 2011.

M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3):287–329, 2005.

M. Armony and A. R. Ward. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637, 2010.

R. Atar, Y. Y. Shaki, and A. Shwartz. A blind policy for equalizing cumulative idleness. *Queueing Systems*, 67(4):275–293, 2011.

A. B. Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

W. Austin, N. C McGinn, and C. Susmilch. Internal standards revisited: Effects of social comparisons and expectancies on judgments of fairness and satisfaction. *Journal of Experimental Social Psychology*, 16(5):426–441, 1980.

B. Avi-Itzhak, H. Levy, and D. Raz. Quantifying fairness in queuing systems: Principles, approaches, and applicability. *Probability in the Engineering and Informational Sciences*, 22 (4):495–517, 2008.

N. A. Barr. *The economics of the welfare state*. Stanford university press, 1993.

P. P. Belobaba. Air travel demand and airline seat inventory management. Technical report, Cambridge, MA: Flight Transportation Laboratory, Massachusetts Institute of Technology, 1987.

P. P. Belobaba. OR practice - application of a probabilistic decision model to airline seat inventory control. *Operations Research*, 37(2):183–197, 1989.

D. P. Bertsekas, R. G. Gallager, and P. Humblet. *Data networks*, volume 2. Prentice-Hall International, 1992.

D. Bertsimas, V. F. Farias, and N. Trichakis. The price of fairness. *Operations Research*, 59 (1):17–31, 2011.

D. Bertsimas, V. F. Farias, and N. Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250, 2012.

T. Bonald and L. Massoulié. Impact of fairness on internet performance. In *SIGMETRICS Performance Evaluation Review*, volume 29(1), pages 82–91. ACM, 2001.

R. M. Bradford. Pricing, routing, and incentive compatibility in multiserver queues. *European Journal of Operational Research*, 89(2):226–236, 1996.

J. Brockner, T. R. Tyler, and R. Cooper-Schneider. The influence of prior commitment to an institution on reactions to perceived unfairness: The higher they are, the harder they fall. *Administrative Science Quarterly*, 37(2):241–261, 1992.

S. L. Brumelle and J. I. McGill. Airline seat allocation with multiple nested fare classes. *Operations Research*, 41(1):127–137, 1993.

F. B. Cabral. The slow server problem for uninformed customers. *Queueing Systems*, 50(4): 353–370, 2005.

F. B. Cabral. Queues with heterogeneous servers and uninformed customers: who works the most? Working paper, 2007.

L. Cabral and M. H. Riordan. The learning curve, market dominance, and predatory pricing. *Econometrica*, 62(5):1115–1140, 1994.

G. P. Cachon and S. Netessine. Game theory in supply chain analysis. *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*, 2006.

G. P. Cachon and F. Zhang. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science*, 53(3):408–420, 2007.

A. M. Campbell, D. Vandenbussche, and W. Hermann. Routing for relief efforts. *Transportation Science*, 42(2):127–145, 2008.

H. Chen and M. Frank. Monopoly pricing when customers queue. *IIE Transactions*, 36(6): 569–581, 2004.

Y. Chevaleyre, P.E. Dunne, U. Endriss, J. Lang, M. Lemaitre, N. Maudet, J. Padget, S. Phelps, J.A. Rodriguez-Aguilar, and P. Sousa. Issues in multiagent resource allocation. *Informatica*, 30(1), 2005.

W. K. Ching, S. M. Choi, and M. Huang. Optimal service capacity in a multiple-server queueing system: A game theory approach. *Journal of Industrial and Management Optimization*, 6 (1):73–102, 2010.

S. M. Choi, X. Huang, W. K. Ching, and M. Huang. Incentive effects of multiple-server queueing networks: The principal-agent perspective. *East Asian Journal on Applied Mathematics*, 1 (4):379–402, 2011.

Y. Cohen-Charash and P. E. Spector. The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 86(2):278–321, 2001.

J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. L. H. Porter, and K. Y. Ng. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, 86(3):425–445, 2001.

T. H. Cui, J. S. Raju, and Z. J. Zhang. Fairness and channel coordination. *Management Science*, 53(8):1303–1314, 2007.

P. Dasgupta and J. Stiglitz. Learning-by-doing, market structure and industrial and trade policies. *Oxford Economic Papers*, 40(2):246–268, 1988.

N. M. Edelson and D. K. Hilderbrand. Congestion tolls for poisson queuing processes. *Econometrica*, 43(1):81–92, 1975.

A. C. Edmondson, A. B. Winslow, R. M.J. Bohmer, and G. P. Pisano. Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. *Decision Sciences*, 34(2):197–224, 2003.

A. N. Elmachtoub. *New approaches for integrating revenue and supply chain management*. PhD thesis, Massachusetts Institute of Technology, 2014.

S. Fay and J. Xie. Probabilistic goods: A creative way of selling products and services. *Marketing Science*, 27(4):674–690, 2008.

S. Fay and J. Xie. Timing of product allocation: Using probabilistic selling to enhance inventory management. *Management Science*, 61(2):474–484, 2015.

E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

L. Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

D. Fudenberg and J. Tirole. Learning-by-doing and market performance. *The Bell Journal of Economics*, 14(2):522–530, 1983.

G. Gallego and R. Phillips. Revenue management of flexible products. *Manufacturing & Service Operations Management*, 6(4):321–337, 2004.

N. Gans. Customer loyalty and supplier quality competition. *Management Science*, 48(2): 207–221, 2002.

N. Gans and Y. Zhou. Managing learning and turnover in employee staffing. *Operations Research*, 50(6):991–1006, 2002.

N. Gans, N. Liu, A. Mandelbaum, H. Shen, and H. Ye. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, volume 6, pages 99–123. Institute of Mathematical Statistics, 2010.

M. Gaynor, H. Seider, and W. B. Vogt. The volume-outcome effect, scale economies, and learning-by-doing. *American Economic Review*, 95(2):243–247, 2005.

X. Geng, W. T. Huh, and M. Nagarajan. Sequential resource allocation with constraints: Two-customer case. *Operations Research Letters*, 42(1):70–75, 2014a.

X. Geng, W. T. Huh, and M. Nagarajan. Fairness among servers when capacity decisions are endogenous. *Production and Operations Management*, 2014b.

S. M. Gilbert and Z. K. Weng. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Science*, 44(12):1662–1669, 1998.

R. Gopalakrishnan, S. Doroudi, A. R. Ward, and A. Wierman. Routing and staffing when servers are strategic. Working paper, 2013.

J. Greenberg. A taxonomy of organizational justice theories. *Academy of Management Review*, 12(1):9–22, 1987.

H. W. Gustafson. Force-loss cost analysis. In W. H. Mobley, editor, *Employee turnover: causes, consequences, and control*. Addison-Wesley, 1982.

R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer, 2003.

D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research: Stochastic Optimization*, volume 2. Courier Corporation, 2003.

T. Huang and Y. Yu. Sell probabilistic goods? a behavioral explanation for opaque selling. *Marketing Science*, 33(5):743–759, 2014.

R. C. Huseman, J. D. Hatfield, and E. W. Miles. A new perspective on equity theory: The equity sensitivity construct. *Academy of Management Review*, 12(2):222–234, 1987.

K. Jerath, S. Netessine, and S. K. Veeraraghavan. Revenue management with strategic customers: Last-minute selling and opaque selling. *Management Science*, 56(3):430–448, 2010.

Y. Jiang. Price discrimination with opaque products. *Journal of Revenue & Pricing Management*, pages 118–134, 2007. Forthcoming.

J. Y. Jin, J. Perote-Pena, and M. Troege. Learning by doing, spillovers and shakeouts. *Journal of Evolutionary Economics*, 14(1):85–98, 2004.

D. Kahneman, J. L. Knetsch, and R. H. Thaler. Fairness and the assumptions of economics. *Journal of Business*, 59(4):285–300, 1986.

E. Kalai, M. I. Kamien, and M. Rubinovitch. Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163, 1992.

F. P. Kelly, A. K. Maulloo, and D. K. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3): 237–252, 1998.

N. C. Knudsen. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica*, 40(3):515–528, 1972.

V. Kostami and S. Rajagopalan. Speed-quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management*, 16(1):104–118, 2013.

T. Lan, D. Kao, M. Chiang, and A. Sabharwal. *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

M. A. Lapré, A. S. Mukherjee, and L. N. Van Wassenhove. Behind the learning curve: Linking learning activities to waste reduction. *Management Science*, 46(5):597–611, 2000.

R. C. Larson. Perspectives on queues: social justice and the psychology of queueing. *Operations Research*, 35(6):895–905, 1987.

R. W. Lien. *Design and control principles for distribution systems: Studies in commercial & nonprofit operations*. PhD thesis, Northwestern University, 2008.

R. W. Lien, S. M. R. Iravani, and K. R. Smilowitz. Sequential resource allocation for nonprofit operations. Working paper, 2008.

H. Luss. On equitable resource allocation problems: A lexicographic minimax approach. *Operations Research*, 47(3):361–378, 1999.

A. Mandelbaum, P. Momčilović, and Y. Tseytlin. On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. Working paper, 2010.

A. Mas-Collel, M. D. Whinston, and J. Green. *Microeconomic Theory*. Oxford University Press, Oxford, 1995.

J. K. McCreery and L. J. Krajewski. Improving performance using workforce flexibility in an assembly environment with learning and forgetting effects. *International Journal of Production Research*, 37(9):2031–2058, 1999.

H. Mendelson. Pricing computer services: queueing effects. *Communications of the ACM*, 28 (3):312–321, 1985.

S. Misra, E. J. Pinker, and R. A. Shumsky. Salesforce design with experience-based learning. *IIE Transactions*, 36(10):941–952, 2004.

S. Moskowitz. Pay equity and american nurses: A legal analysis. *St. Louis University Law Journal*, 27:801, 1983.

P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

J. F Nash. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

H. Nikaidô and K. Isoda. Note on noncooperative convex games. *Pacific Journal of Mathematics*, 5(1):807–815, 1955.

John Rawls. *A theory of justice*. Harvard University Press, Cambridge, MA, 1971.

J. Reed and Y. Shaki. A fair policy for the $g/gi/n$ queue with multiple server pools. Mathematics of Operations Research, Forthcoming, 2014.

L. W. Robinson. Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research*, 43(2):252–263, 1995.

M. Rubinovitch. The slow server problem. *Journal of Applied Probability*, 22(1):205–213, 1985.

G. S. Ryder, K. G. Ross, and J. T. Musacchio. Optimal service policies under learning effects. *International Journal of Services and Operations Management*, 4(6):631–651, 2008.

E. S. Savas. On equity in providing public services. *Management Science*, 24(8):800–808, 1978.

S. M. Shafer, D. A. Nembhard, and M. V. Uzumeri. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science*, 47(12):1639–1653, 2001.

D. Shapiro and X. Shi. Market segmentation: The role of opaque travel agencies. *Journal of Economics & Management Strategy*, 17(4):803–837, 2008.

S. Solak, C. Scherrer, and A. Ghoniem. The stop-and-drop problem in nonprofit food distribution networks. *Annals of Operations Research*, pages 1–20, 2012. ISSN 0254-5330.

M. A. Spence. The learning curve and competition. *The Bell Journal of Economics*, 12(1): 49–70, 1981.

J. M. Swaminathan. Decision support for allocating scarce drugs. *Interfaces*, 33(2):1–11, 2003.

K. T. Talluri and G. Van Ryzin. *The theory and practice of revenue management*. Springer Verlag, 2005.

T. Tezcan. Optimal control of distributed parallel server systems under the halfin and whitt regime. *Mathematics of Operations Research*, 33(1):51–90, 2008.

C. Tong and S. Rajagopalan. Pricing and operational performance in discretionary services. *Production and Operations Management*, 23(4):689–703, 2014.

Y. Tseytlin. *Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments*. PhD thesis, Technion-Israel Institute of Technology, 2009.

A. Wierman. Fairness and classifications. *ACM SIGMETRICS Performance Evaluation Review*, 34(4):4–12, 2007.

H. M. Wu, C. C. Wu, and W. Lin. Improving inter-server fairness in active queue management. *Communications Letters, IEEE*, 11(11):910–912, 2007.

L. E. Yelle. The learning curve: Historical review and comprehensive survey. *Decision Sciences*, 10(2):302–328, 1979.

H. P. Young. *Equity: in theory and practice*. Princeton University Press, Princeton, NJ, 1995.

Z. Zhang, K. Joseph, and R. Subramaniam. Probabilistic selling in quality-differentiated markets. *Management Science*, 2014.

Appendix A

Proofs

A.1 Proofs of Results in Chapter 2

Proof of Lemma 2.3.2. Since $d_1 > 0$, there exists an integer k such that $\theta_k \leq d_1 < \theta_{k+1}$. Then the left derivative of $R(x_1, d_1)$ at z_k is

$$\left(\sum_{j=1}^{k} p_j\right) \frac{1}{d_1} - \sum_{k+1}^{n} \frac{p_j}{a_j} > \left(\sum_{j=1}^{k} p_j\right) \frac{1}{\theta_{k+1}} - \sum_{k+1}^{n} \frac{p_j}{a_j} = 0,$$

and the right derivative of $R(x_1, d_1)$ at z_k is

$$\left(\sum_{j=1}^{k-1} p_j\right) \frac{1}{d_1} - \sum_{j=k}^{n} \frac{p_j}{a_j} \leq \left(\sum_{j=1}^{k-1} p_j\right) \frac{1}{\theta_k} - \sum_{j=k}^{n} \frac{p_j}{a_j} = 0.$$

Therefore, z_k has positive left derivative and non-positive right derivative, and thus is the maximizer. This shows that

$$\phi(d_1) = z_k(d_1) = \frac{sd_1}{d_1 + a_k} \quad \text{for } \theta_k \leq d_1 < \theta_{k+1} \ . \tag{A.1}$$

By straightforward calculus, $\phi(d_1)$ is concave and increasing in d_1 when d_1 lies in the interval $[\theta_k, \theta_{k+1})$. However, discontinuities occur at θ_k, where $k = 2, 3, \cdots, n$. To see this, since $a_k > a_{k-1}$, we have

$$\phi(\theta_k+) = z_k(\theta_k) = \frac{s\theta_k}{\theta_k + a_k} < \frac{s\theta_k}{\theta_k + a_{k-1}} = z_{k-1}(\theta_k) = \phi(\theta_k-),$$

for $k = 2, 3, \cdots, n$. This concludes the lemma. □

Proof of Corollary 2.3.3. Every statement except the last one is a direct consequence of Lemma 2.3.2 since the concavity and monotonicity properties are preserved by the minimum operator in (2.5). If there exists a discontinuous point, then it must be θ_k for some $k = 2, 3, \cdots, n$. We consider the following cases based on where θ_k lies with respect to $\phi(\theta_k+)$ and $\phi(\theta_k-)$. Note that $\phi(\theta_k+) < \phi(\theta_k-)$ by Lemma 2.3.2. If $\phi(\theta_k+) \geq \theta_k$, then $\phi(d_1) = d_1$ by (2.5) when d_1 is in a small neighbourhood of θ_k, in which case ϕ is continuous at θ_k. If $\phi(\theta_k-) \leq \theta_k$ or $\phi(\theta_k+) < \theta_k < \phi(\theta_k-)$, then $\phi(\theta_k+)$ is strictly lower than both $\phi(\theta_k-)$ and θ_k, implying $x_1^*(\theta_k+) < x_1^*(\theta_k-)$. □

Proof of Theorem 2.3.4. We first show that the condition is necessary. Suppose $x_1^*(d_1)$ is overall increasing and concave in d_1. As mentioned above, in this case the fulfilment line must stay below $\phi(d_1)$ at every threshold point θ_k; otherwise the monotonicity would be violated by the downward jump at θ_k.

From Lemma 2.3.2, $\phi(\theta_k) = \phi(\theta_k+) < \phi(\theta_k-)$. Thus, we must have $\phi(d_1) \geq d_1$ at every $d_1 = \theta_k$ $(k = 2, 3, \cdots, n)$. Therefore we have the necessary condition

$$\phi(\theta_k) \geq \theta_k, \quad \text{for every } k = 2, 3, \cdots, n.$$

Using equation (A.1) and (2.7), we have $\phi(\theta_k) = z_k(\theta_k) = \dfrac{s\theta_k}{\theta_k + a_k}$. Then we can rewrite the necessary condition as

$$\frac{s\theta_k}{\theta_k + a_k} \geq \theta_k, \quad \text{for every } k = 2, 3, \cdots, n,$$

which reduces to

$$s \geq a_k + \theta_k, \quad \text{for every } k = 2, 3, \cdots, n.$$

The above inequalities are equivalent to the desired result

$$s \geq \max_{2 \leq k \leq n} \{a_k + \theta_k\} = a_n + \theta_n,$$

where the last equality is due to the monotonicity of $\{a_k\}$ and $\{\theta_k\}$ given in (2.6) and (2.12), respectively.

Now we show the sufficiency. Suppose the inequality (2.13) holds. Then the deductions in the above proof are trivially reversible from the last step up to the statement that $\phi(d_1) \geq d_1$ at every $d_1 = \theta_k$ $(k = 2, 3, \cdots, n)$. This means that the fulfilment line is below the points $(\theta_k, \phi(\theta_k))$, for each $k = 2, 3, \cdots, n$.

We now conclude the sufficiency proof by showing that $x_1^*(d_1)$ is overall increasing and concave. Define

$$\tilde{\theta} = \sup_{d_1} \{\phi(d_1) \geq d_1\}.$$

First we claim that $\phi(\tilde{\theta}) = \tilde{\theta}$. If $\phi(\tilde{\theta}) < \tilde{\theta}$, then by its definition, we know that $\tilde{\theta}$ is a point of discontinuity, which must be one of θ_k's. This contradicts the fact that $\phi(\theta_k) \geq \theta_k$. If $\phi(\tilde{\theta}) > \tilde{\theta}$, then let $\delta = \phi(\tilde{\theta}) - \tilde{\theta} > 0$. Since by (A.1) $\phi(d_1)$ is right continuous, we know that there exists an $\epsilon > 0$, such that $|\phi(\tilde{\theta} + \epsilon) - \phi(\tilde{\theta})| + \epsilon \leq \delta$. Hence, $\phi(\tilde{\theta} + \epsilon) \geq \phi(\tilde{\theta}) + \epsilon - \delta = \tilde{\theta} + \epsilon$, which contradicts the definition of $\tilde{\theta}$. Therefore only the case of equality holds, proving the claim.

Now, since $\tilde{\theta} \geq \theta_k$ holds for every $k = 2, 3, \cdots, n$, in particular, we have $\tilde{\theta} \geq \theta_n$. From Lemma 2.3.2, it follows that $\phi(\theta_k-) > \phi(\theta_k) \geq \theta_k$, for $k = 2, 3, \cdots, n$. Hence, by the concavity of $\phi(d_1)$ on each of the intervals $[\theta_k, \theta_{k+1})$ for $k = 1, 2, \cdots, n-1$, and $[\theta_n, \tilde{\theta}]$, we deduce that $\phi(d_1) \geq d_1$ for all $0 \leq d_1 \leq \tilde{\theta}$. Therefore $x_1^*(d_1) = d_1$ on $[0, \tilde{\theta}]$, which is increasing and linear. On the other hand, from the definition of $\tilde{\theta}$, $\phi(d_1) < d_1$ for $d_1 > \tilde{\theta}$; so $x_1^*(d_1) = \phi(d_1)$ when

$d_1 > \tilde{\theta} \geq \theta_n$, which is increasing and concave according to Lemma 2.3.2.

Finally, $x_1^*(d_1)$ is continuous at $d_1 = \tilde{\theta}$ by noting that $\phi(\tilde{\theta}) = \tilde{\theta}$. Besides, since $\phi(d_1)$ goes below the fulfilment line after $d_1 = \tilde{\theta}$, we have $\phi'(\tilde{\theta}+) < 1 = \phi'(\tilde{\theta}-)$. Therefore, $x_1^*(d_1)$ is overall increasing and concave. □

Proof of Proposition 2.3.5. For any $d_1 > 0$, there is some k such that $\phi(d_1) = z_k(d_1)$ by (A.1). Hence, if $s \leq a_1$, then by differentiating equation (A.1),

$$\phi'(d_1+) = \frac{sa_k}{(d_1 + a_k)^2} < \frac{sa_k}{a_k^2} < \frac{s}{a_1} \leq 1,$$

where the last inequality follows from the ordering of $\{a_k\}$ in (2.6). In particular,

$$\phi'(0+) = \left.\frac{sa_1}{(d_1 + a_1)^2}\right|_{d_1=0} = \frac{s}{a_1} \leq 1.$$

Besides, $\phi(0) = 0$. Therefore, combining the result in Lemma 2.3.2, we conclude that $\phi(d_1)$ intersects with the fulfilment line only at $d_1 = 0$; since $\phi(d_1) < d_1$ holds for $d_1 > 0$. By (2.5), we conclude the result. □

Proof of Theorem 2.3.7. Since a_k and θ_k both strictly increase to infinity (see the discussion in the second paragraph of this subsection), we know that for any $s > 0$, there exists an integer $k > 0$ such that

$$s < a_{k-1} + \theta_k. \tag{A.2}$$

Then by equation (A.1), we have

$$\phi(\theta_k) = z_k(\theta_k) = \frac{s\theta_k}{\theta_k + a_k} < \frac{s\theta_k}{\theta_k + a_{k-1}} = z_{k-1}(\theta_k-) = \phi(\theta_k-).$$

Apply inequality (A.2) to deduce

$$\phi(\theta_k) < \phi(\theta_k-) = \frac{s\theta_k}{\theta_k + a_{k-1}} < \theta_k.$$

Hence by equation (2.5), $x_1^*(\theta_k) = \phi(\theta_k) \wedge \theta_k = \phi(\theta_k)$ while $x_1^*(\theta_k-) = \phi(\theta_k-) \wedge \theta_k = \phi(\theta_k-)$. Since $\phi(\theta_k) < \phi(\theta_k-)$, it follows that $x_1^*(d_1)$ is discontinuous at $d_1 = \theta_k$. Let $\bar{d_1} = \theta_k$, then it proves the theorem. □

A.2 Proofs of Results in Chapter 3

Before the proof of all the results in this section, we provide some preliminary results. Consider a stable M/M/2 queueing system, then it follows that there exists a unique stationary distribution. Let π_i be the long run (stationary) probability that there are i customers in the system, including those waiting and being served. In addition, let π_{10} be the stationary probability of server 1 being busy and server 2 idle; and similarly let π_{01} be that of server 2 being busy and server

1 idle. We modify the approach in Rubinovitch (1985) to get the stationary distribution and other related quantities. By balancing the flow-in and flow-out of each state, we have

$$\begin{aligned}(\lambda+\mu_1)\pi_{10} &= \lambda\pi_0\phi + \mu_2\pi_2\\ (\lambda+\mu_2)\pi_{01} &= \lambda\pi_0(1-\phi) + \mu_1\pi_2\\ \lambda\pi_0 &= \pi_{10}\mu_1 + \pi_{01}\mu_2\\ \lambda\pi_n &= (\mu_1+\mu_2)\pi_{n+1}, \quad n \geq 1.\end{aligned}$$

Following the traditional notation, let $\rho = \lambda/(\mu_1+\mu_2) < 1$ be the traffic coefficient. Then we have

$$\pi_n = \rho^{n-1}\pi_1 = \rho^{n-1}(\pi_{10}+\pi_{01}), \quad n \geq 1.$$

Solving the above system of equations gives

$$\pi_{10} = \frac{\lambda\pi_0}{2\mu_1}\gamma_1 \quad \text{and} \quad \pi_{01} = \frac{\lambda\pi_0}{2\mu_2}\gamma_2,$$

where

$$\gamma_1 = \frac{\rho+\phi}{\rho+1/2} \quad \text{and} \quad \gamma_2 = \frac{\rho+(1-\phi)}{\rho+1/2}.$$

Furthermore, since $\pi_0+\pi_1+\cdots = 1$, we have

$$\pi_0 = \frac{(1-\rho)(1+2\rho)}{(k+\frac{1}{k})\rho^2 + ((1-\phi)k+\phi/k)\rho + 2\rho + 1}, \tag{A.3}$$

where $k = \mu_1/\mu_2$ is the ratio of the two service rates. Now we compute the long run average number of customers both waiting and being served:

$$L = \sum_{n=1}^{\infty} n\pi_k = \sum_{n=1}^{\infty} n\rho^{n-1}\pi_1 = \frac{\pi_1}{(1-\rho)^2}.$$

Note that $\pi_1 = \pi_{10}+\pi_{01}$, so we write

$$L = \frac{\pi_{10}+\pi_{01}}{(1-\rho)^2} = \frac{\lambda\pi_0}{2(1-\rho)^2}\left(\frac{\gamma_1}{\mu_1}+\frac{\gamma_2}{\mu_2}\right).$$

More specifically, substituting equation (A.3) gives L in term of service rates and routing policy.

$$L = \frac{1}{1-\rho} - \frac{1+2\rho}{(k+\frac{1}{k})\rho^2 + ((1-\phi)k+\phi/k)\rho + 1 + 2\rho}. \tag{A.4}$$

Finally, the average idle time for server 1, denoted by τ_1, consists of the portion that no customer is in the system and that the only customer in system is at server 2. Therefore,

$$\tau_1 = \pi_0 + \pi_{01} = \pi_0\left(1+\frac{\lambda}{2\mu_2}\gamma_2\right). \tag{A.5}$$

Similarly for server 2,

$$\tau_2 = \pi_0 + \pi_{10} = \pi_0 \left(1 + \frac{\lambda}{2\mu_1}\gamma_1\right). \tag{A.6}$$

To simplify the notations, once the policy $\bar{\phi}$ is made clear in the context, we will write $g_1(\mu_1)$ instead of $g_1(\mu_1|\mu_2, \bar{\phi})$. The simplification applies to other functions in all proofs.

Proof of Theorem 3.3.1. We examine the first order condition to find the Nash equilibrium.

Since we only consider symmetric equilibrium, the policy $\phi_r = 1/2$ at equilibrium. Hence, the idle time is the same for both servers and the unfairness function will be zero by definition. Suppose the equilibrium capacity is $y > 1/2$, by symmetry, we just need to analyse the condition $g_1' + t = 0$ when $\mu_1 = \mu_2 = y$.

Recall the definition of the inverse-idleness function in (3.1). Substituting (A.5) and (A.3), we have

$$g_1(\mu_1) = \frac{(k + \frac{1}{k})\rho^2 + ((1-\bar{\phi})k + \frac{\bar{\phi}}{k})\rho + 1 + 2\rho}{(1-\rho)(1+2\rho)} \frac{1}{\left(1 + \frac{1}{\mu_2}\frac{\rho+1-\bar{\phi}}{1+2\rho}\right)}. \tag{A.7}$$

Substitute policy into (A.7), take derivative and replace μ_1 and μ_2 with y, we can write the first order condition as

$$(2r-4)y^2 - (4+r)y + t(2y+1)(y+1)(2y-1)^2 = 0.$$

Define the left hand side by a function $H(y)$. For the remainder of the proof, we show that for any finite r, $H(y)$ has a unique zero point that is larger than $1/2$.

First, note that $H(1/2) = -3 < 0$,

$$H'(y) = 4(r-2)y - (4+r) + t(2y-1)(16y^2 + 14y + 1),$$

and $H'(1/2) = r - 8$. Moreover, $H(+\infty) = +\infty$ and $H'(+\infty) = +\infty$ because the leading (highest order) terms have positive coefficients. Furthermore, the second order derivative is

$$H''(y) = 4(r-2) + 12t(2y+1)(4y-1).$$

We now have two cases. If $r \geq 2$, then $H''(y) > 0$ for all $y > 1/2$. Since $H(1/2) < 0$, regardless of the monotonicity of H, it can intersect with $y = 0$ only once. Let that intersection be μ^*, then $\mu^* > 1/2$ and the uniqueness is proved. In addition, if $r > 8$, then $H(y)$ is increasing at $y = 1/2$. The intersection point is closer to $1/2$ as r increases because $H''(y)$ increases with r. Therefore, for any small $\epsilon > 0$, we can find a large $\bar{r} > 8$, such that for all $r > \bar{r}$, $\mu^* < 1/2 + \epsilon$.

If $r < 2$, then $H''(y)$ could be negative. Further analysis on H'' indicates that there is at most one inflection point, say $\tilde{y} > 1/2$. If such $\tilde{y}$ does not exist, then we are back to the case where H is convex. If $\tilde{y}$ exists, then H is concave on $[1/2, \tilde{y}]$ and convex on $[\tilde{y}, +\infty)$. Since $H'(1/2) = r - 8 < 0$, we deduce that $H(\tilde{y}) < 0$. Hence, $H(y)$ must have a unique zero point

larger than $\tilde{y}$. Call this point μ^* and we finish the proof for uniqueness. To conclude part (ii), note that the inflection point $\tilde{y}$ increases as r decreases. Therefore, for any large $M > 0$, we can find a $\underline{r} < 0$, such that $\tilde{y} > M$ for all $r < \underline{r}$; then $\mu^* > \tilde{y} > M$. □

Proof of Proposition 3.3.2. (i). The continuity result is obvious. We only prove monotonicity and convexity. Suppose $\bar{\phi} = \phi_{\text{Prop}} = \frac{\mu_2}{\mu_1+\mu_2}$ and μ_2 is fixed. Then by substituting in (A.7), we have

$$\begin{aligned} g_1(\mu_1) &= \frac{(k+\frac{1}{k})\rho^2 + (1+(k+\frac{1}{k}))\rho + 1}{(1-\rho)(1+2\rho)} \frac{\mu_2(1+2\rho)}{\mu_2(1+2\rho)+\rho+\rho\mu_1} \\ &= \frac{\mu_2\mu_1(\mu_1+\mu_2)+\mu_1^2+\mu_2^2}{\mu_1(\mu_2+1)(\mu_1+\mu_2-1)}. \end{aligned}$$

We take the derivatives and obtain

$$g_1'(\mu_1) = -\frac{\mu_1^2+\mu_2^2(2\mu_1-1)+\mu_2^3}{\mu_1^2(\mu_2+1)(\mu_1+\mu_2-1)^2} < 0,$$

and

$$g_1''(\mu_1) = \frac{2\left(\mu_1^3+3\mu_2^2\mu_1(\mu_1+\mu_2-1)+\mu_2^2(\mu_2-1)^2\right)}{\mu_1^3(\mu_2+1)(\mu_1+\mu_2-1)^3} > 0.$$

The inequalities are due to the assumption that $\mu_1 \geq 1/2$ and $\mu_2 > 1/2$. Hence g_1 is strictly decreasing and convex in μ_1.

For the case $\bar{\phi} = \phi_{\text{HH}} = 1/2$, we perform the same computations and get

$$g_1(\mu_1) = \frac{2\mu_2\mu_1(\mu_1+\mu_2)+\mu_1^2+\mu_2^2}{\mu_1(2\mu_2+1)(\mu_1+\mu_2-1)}$$

$$g_1'(\mu_1) = \frac{(\mu_1+\mu_2)(\mu_2\mu_1+\mu_1+\mu_2^2-\mu_2)}{(2\mu_2+1)\mu_1^2(\mu_1+\mu_2-1)^2}$$

$$g_1''(\mu_1) = \frac{2\left(\mu_1^3+3\mu_2^2\mu_1(\mu_1+\mu_2-1)+\mu_2^2(\mu_2-1)^2\right)}{\mu_1^3(2\mu_2+1)(\mu_1+\mu_2-1)^3} > 0.$$

Note that $\mu_1, \mu_2 > 1/2$, so

$$\mu_2\mu_1+\mu_1+\mu_2^2-\mu_2 > \mu_2^2-\frac{1}{2}\mu_2+\frac{1}{2} > 0.$$

Therefore, if the policy is HH, g_1 is also strictly decreasing and convex.

(ii). By substituting in (A.7) directly,

$$g_1(\mu_1|\mu_2,\phi_{\text{FSF}}) = \begin{cases} \frac{(k+\frac{1}{k})\rho^2+k\rho+1+2\rho}{(1-\rho)(1+2\rho)} \frac{1}{1+\frac{1}{\mu_2}\frac{1+\rho}{1+2\rho}} & \text{if} \quad 1/2 < \mu_1 < \mu_2 \\ \frac{1}{1-\rho} & \text{if} \quad \mu_1 = \mu_2 \\ \frac{(k+\frac{1}{k})\rho^2+\rho/k+1+2\rho}{(1-\rho)(1+2\rho)} \frac{1}{1+\frac{1}{\mu_2}\frac{\rho}{1+2\rho}} & \text{if} \quad \mu_1 > \mu_2 \end{cases}.$$

So the function is continuous except at $\mu_1 = \mu_2$. Note that SSF operates in an opposite way to FSF, therefore

$$g_1(\mu_1|\mu_2, \phi_{\text{SSF}}) = \begin{cases} \frac{(k+\frac{1}{k})\rho^2+\rho/k+1+2\rho}{(1-\rho)(1+2\rho)} \frac{1}{1+\frac{1}{\mu_2}\frac{\rho}{1+2\rho}} & \text{if} \quad 1/2 < \mu_1 < \mu_2 \\ \frac{1}{1-\rho} & \text{if} \quad \mu_1 = \mu_2 \\ \frac{(k+\frac{1}{k})\rho^2+k\rho+1+2\rho}{(1-\rho)(1+2\rho)} \frac{1}{1+\frac{1}{\mu_2}\frac{1+\rho}{1+2\rho}} & \text{if} \quad \mu_1 > \mu_2 \end{cases}.$$

Hence, we can just focus on proving results for FSF. Results for SSF can be obtained for free by swapping the two continuous pieces. For example, since

$$g_1(\mu_2 - |\mu_2, \phi_{\text{FSF}}) = \frac{1+\rho}{(1-\rho)(1+\rho+\frac{\rho}{1+2\rho})} < \frac{1}{1-\rho} = g_1(\mu_2|\mu_2, \phi_{\text{FSF}})$$

$$g_1(\mu_2 + |\mu_2, \phi_{\text{FSF}}) = \frac{1+\rho}{(1-\rho)(1+\rho-\frac{\rho}{1+2\rho})} > \frac{1}{1-\rho} = g_1(\mu_2|\mu_2, \phi_{\text{FSF}}),$$

we prove the inequalities in (3.5). Immediately, we also prove the inequalities in (3.6). Therefore, in the following, we just focus on proving monotonicity and convexity under FSF policy.

For $\mu_1 \in (1/2, \mu_2)$, we take the derivative and get

$$g_1'(\mu_1) = -\frac{1}{(\mu_2\mu_1+\mu_2^2+3\mu_2+\mu_1+1)^2(\mu_1+\mu_2-1)^2\mu_1^2} \left(\mu_1^2(\mu_1+1)^2+\mu_2^3(\mu_2+1)^2+ \right.$$
$$\left. \mu_2\mu_1^2(\mu_1+5)(\mu_1+1)+4\mu_2^4\mu_1+3\mu_2^3(2\mu_1^2+2\mu_1-1)+\mu_2^2(4\mu_1^3+9\mu_1^2-1)\right) < 0.$$

Similarly, for $\mu_1 > \mu_2 > 1/2$, we have

$$g_1'(\mu_1) = -\frac{w_g+\mu_2^2(2\mu_2+1)(2\mu_1-1)}{(\mu_2\mu_1+1+\mu_2^2+2\mu_2)^2(\mu_1+\mu_2-1)^2\mu_1^2},$$

where

$$w_g = \mu_1^3\mu_2\left(\mu_2\mu_1+4\mu_2^2+2+2\mu_2\right)+\mu_1^2\left(6\mu_2^2+6\mu_2^3+6\mu_2^4+3\mu_2+1\right)+2\mu_1\mu_2^4\left(3+2\mu_2\right)+\mu_2^5\left(\mu_2+2\right).$$

Hence, $g_1' < 0$. We have finish the proof for monotonicity.

To show convexity, a more tedious algebraic manipulation is needed. We delay the details to Section B (a). □

Proof of Proposition 3.3.3. By the definition of the unfairness function, we can write it as a composition of two functions u and v,

$$f_1 = u \circ v = u(v), \tag{A.8}$$

where $u(v)$ is defined by $u(v) = v^2$ $(v \geq 0)$, and

$$v(\mu_1) = \pi_0 \left(\frac{\mu_1}{\mu_2} - 1 + \frac{2\bar{\phi} - 1}{\mu_2(1+2\rho)} \right)^+. \tag{A.9}$$

(i). We first look at $\bar{\phi} \in \{\phi_{\text{FSF}}, \phi_{\text{Prop}}, \phi_{\text{HH}}\}$. Suppose that $\bar{\phi} = \phi_{\text{FSF}}$, then it is easy to check that

$$\frac{\mu_1}{\mu_2} - 1 + \frac{2\phi_{\text{FSF}} - 1}{\mu_2(1+2\rho)} > 0$$

if and only if $\mu_1 > \mu_2$. For $1/2 < \mu_1 \leq \mu_2$, this part becomes zero, so will f_1 be. We now only consider $\mu_1 > \mu_2$, then

$$\frac{d}{d\mu_1} \left(\frac{\mu_1}{\mu_2} - 1 + \frac{2\phi_{\text{FSF}} - 1}{\mu_2(1+2\rho)} \right) = \frac{1}{\mu_2} + \frac{2\rho^2}{\mu_2(1+2\rho)^2} > 0.$$

Besides,

$$\begin{aligned} \frac{d\pi_0}{d\mu_1} &= \frac{\rho^3(\mu_2+1)\left((\mu_1+\mu_2)^3 + \mu_2(\mu_1+\mu_2) + 2(\mu_1-\mu_2)\right)}{\mu_1^2\mu_2\left((k+\frac{1}{k})\rho^2 + \rho/k + 2\rho + 1\right)^2} \\ &> 0. \end{aligned} \tag{A.10}$$

Therefore, $v(\mu_1)$ is strictly increasing, and so is f_1.

Suppose that $\bar{\phi} = \phi_{\text{Prop}}$. Then

$$\begin{aligned} v(\mu_1) &= \frac{(1-\rho)(1+2\rho)}{(1+\rho)(1+(k+\frac{1}{k})\rho)} \left(\frac{\mu_1}{\mu_2} - 1 + \frac{(\mu_2-\mu_1)\rho}{\mu_2(1+2\rho)} \right)^+ \\ &= \frac{1-\rho}{1+(k+\frac{1}{k})\rho} (\frac{\mu_1}{\mu_2} - 1)^+. \end{aligned}$$

Hence, $f_1(\mu_1|\mu_2, \phi_{\text{Prop}}) = 0$ for $1/2 < \mu_1 \leq \mu_2$. Now we consider $\mu_1 > \mu_2$. Since

$$\frac{d}{d\mu_1} \left(\frac{1-\rho}{1+(k+\frac{1}{k})\rho} \right) = \frac{\rho^2}{(1+(k+\frac{1}{k})\rho)^2} \left\{ \frac{1}{k} \left(1 + \frac{1}{k} \right) (1-\rho) + \frac{1}{k} + \frac{1}{\mu_2} \right\} > 0,$$

we direclty conclude that $v(\mu_1)$ is increasing; so $f_1(\mu_1)$ must also be.

Suppose that $\bar{\phi} = \phi_{\text{HH}}$. By equation (A.9), we have

$$v(\mu_1) = \pi_0 \left(\frac{\mu_1}{\mu_2} - 1 \right)^+.$$

Clearly, this function is positive only when $\mu_1 > \mu_2$. If so, we compute that

$$\frac{d\pi_0}{d\mu_1} = \frac{2\mu_2((1+\mu_2)\mu_1^2 + (2\mu_1 - 1)\mu_2^2 + \mu_2^3)}{(2\mu_2\mu_1^2 + \mu_1^2 + 2\mu_2^2\mu_1 + \mu_2^2)^2} > 0.$$

Hence, we claim that $f_1(\mu_1)$ is increasing on $\mu_1 > \mu_2$.

(ii). We then suppose that $\bar{\phi} = \phi_{\text{SSF}}$ and $\mu_2 > 1/2$ is fixed. Again, we analyse (A.9). In particular, we find the range of μ_1 on which $q := \frac{\mu_1}{\mu_2} - 1 + \frac{2\phi_{\text{SSF}}-1}{\mu_2(1+2\rho)}$ is positive. If $1/2 < \mu_1 < \mu_2$, then there exists a number

$$a := \max\left\{\frac{1}{2}, \frac{\sqrt{9+4(\mu_2^2+\mu_2)}-3}{2}\right\} \geq \frac{1}{2}$$

such that $q > 0$ if $a < \mu_1 < \mu_2$. Moreover, it is easy to check that $a < \mu_2$. Similarly, if $\mu_1 > \mu_2$, we check that there exists

$$b := \frac{\sqrt{1+4(\mu_2^2+3\mu_2)}-1}{2} > \mu_2$$

such that $q > 0$ if $\mu_1 > b$. Hence, f_1 is zero on $1/2 \leq \mu_1 \leq a$ and $\mu_2 \leq \mu_1 \leq b$; and it is continuous at $\mu_1 = a$ and $\mu_1 = b$. Now, we focus on the positive parts and prove the monotonicity. Since $q > 0$, we examine that its derivative

$$\frac{dq}{d\mu_1} = \frac{1}{\mu_2}\left(1 - \frac{2\rho^2}{(1+2\rho)^2}\right) > 0.$$

If $a < \mu_1 < \mu_2$, then π_0 has the same expression as in (A.10). Note that since $1/2 < \mu_1 < \mu_2$ now, we observe that

$$\begin{aligned}(\mu_1+\mu_2)^3 + \mu_2(\mu_1+\mu_2) + 2(\mu_1-\mu_2) &> \mu_2(3\mu_1^2+3\mu_1\mu_2+\mu_1-2)\\ &> 0.\end{aligned}$$

Hence, again we have $\pi_0' > 0$. If $\mu_1 > b > \mu_2$, then

$$\begin{aligned}\frac{d\pi_0}{d\mu_1} &= \frac{\rho^3\left((2\mu_2+4)\mu_1^2 + 3{\mu_2}^2\mu_1 + \mu_2\mu_1 + \mu_2^2(1+\mu_2) + 2(\mu_1-\mu_2)\right)}{\mu_1^2\mu_2\left((k+\frac{1}{k})\rho^2 + \rho k + 2\rho + 1\right)^2}\\ &> 0.\end{aligned}$$

Therefore, we prove the monotonicity for f_1 on both positive pieces. The proof of convexity needs much more computations, so we do it in Section B (b). □

Proof of Theorem 3.3.4. By Propositions 3.3.2 and 3.3.3, the objective functions of the servers are strictly convex and differentiable. Then according to Theorem 3.1 in Nikaidô and Isoda (1955), it is left to show that the two servers' effective strategies are constrained in convex and compact sets. Since $t > 0$, servers' capacity costs are unbounded. However, the inverse-idleness function is bounded below by 1 (maximum fraction of idleness is 1). Therefore, there exists an $N < +\infty$, such that choosing any capacity greater than N is a strictly dominated strategy for both servers. In other words, the effective strategies for both servers are in compact and convex sets; and the existence of Nash equilibrium follows.

Due to Theorem 7 in Cachon and Netessine (2006) (the condition for two-player case discussed in their paper), the uniqueness is proved if we can show that the Hessian matrix of utility functions (we now treat F_i as a function of both capacities) satisfies:

$$\begin{vmatrix} \frac{\partial^2 F_1}{\partial \mu_1^2} & \frac{\partial^2 F_1}{\partial \mu_1 \partial \mu_2} \\ \frac{\partial^2 F_2}{\partial \mu_2 \partial \mu_1} & \frac{\partial^2 F_2}{\partial \mu_2^2} \end{vmatrix} > 0, \forall \mu_1, \mu_2 \text{ s.t. } \frac{\partial F_1}{\partial \mu_1} = 0, \frac{\partial F_2}{\partial \mu_2} = 0. \tag{A.11}$$

Note this condition is required at equilibrium. Hence, this condition is usually hard to apply. Fortunately, once adapted to our problem, we can prove (A.11) by showing the following sufficient condition:

$$\begin{cases} \frac{\partial^2 g_1}{\partial \mu_1^2} \frac{\partial^2 g_2}{\partial \mu_2^2} > \frac{\partial^2 g_1}{\partial \mu_1 \partial \mu_2} \frac{\partial^2 g_2}{\partial \mu_2 \partial \mu_1} \\ \frac{\partial^2 f_1}{\partial \mu_1^2} > \frac{\partial^2 f_1}{\partial \mu_1 \partial \mu_2} \\ \frac{\partial^2 g_2}{\partial \mu_2^2} > \frac{\partial^2 g_2}{\partial \mu_2 \partial \mu_1}. \end{cases} \tag{A.12}$$

Observe that parameters t and Δt are irrelevant under second order operations. Moreover, condition (A.12) also makes the parameters $\alpha_f^{(1)}$ and $\alpha_f^{(2)}$ irrelevant in our proof. Hence, our result holds regardless of the values of the parameters. Therefore, using the symmetric relation of μ_1 and μ_2 in the g_i and f_i functions, we only need to prove (A.12) for $\mu_1 \geq \mu_2$. The results for the case $\mu_1 \leq \mu_2$ will follow immediately. We finish the proof by verifying (A.12), which we will do in Section B (c).

Finally, we claim that for both Prop and HH, $\overline{\mu}_1 > \overline{\mu}_2$ holds. Suppose it is not true, that is, $\overline{\mu}_1 \leq \overline{\mu}_2$. Then by the sufficient and necessary condition, we have

$$\begin{aligned} g_1'(\overline{\mu}_1) + t &= 0 \\ g_2'(\overline{\mu}_2) + \alpha_f^{(2)} f_2'(\overline{\mu}_2) + t + \Delta t &= 0. \end{aligned}$$

By directly calculations, we derive that if $\bar{\phi} = \phi_{\text{Prop}}$,

$$g_1'(\overline{\mu}_1) - g_2'(\overline{\mu}_2) = \frac{(\mu_1 - \mu_2)A}{{\mu_2}^2 {\mu_1}^2 (\mu_1 + 1)(\mu_2 + 1)(\mu_1 + \mu_2 - 1)^2},$$

where

$$A = \mu_1^4(\mu_2 + 1) + \mu_2^4(\mu_1 + 1) + \mu_2^2(\mu_1 - 1/2)(3\mu_1^2 + 2\mu_1 + 2\mu_2) + \mu_1^2(\mu_2 - 1/2)(3\mu_2^2 + 2\mu_2 + 2\mu_1).$$

Since $\overline{\mu}_2 \geq \overline{\mu}_1 > 1/2$, $g_1'(\overline{\mu}_1) \leq g_2'(\overline{\mu}_2)$. Similarly, if $\bar{\phi} = \phi_{\text{HH}}$,

$$g_1'(\overline{\mu}_1) - g_2'(\overline{\mu}_2) = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)(2\mu_1^3(\mu_2 + 1) + B)}{\mu_2^2 \mu_1^2 (2\mu_1 + 1)(2\mu_2 + 1)(\mu_1 + \mu_2 - 1)^2},$$

where

$$B = \mu_1^2 \left(4\mu_2^2 - 1\right) + 2\,\mu_2^3 \mu_1 + \mu_2^3 - \mu_2^2.$$

Because $\overline{\mu}_2 \geq \overline{\mu}_1 > 1/2$, we see that $B > 0$, and $g_1'(\overline{\mu}_1) \leq g_2'(\overline{\mu}_2)$.

Recall that $\Delta t > 0$ by assumption. Therefore, in both cases we have

$$0 = g_1'(\overline{\mu}_1) + t \leq g_2'(\overline{\mu}_2) + t < g_2'(\overline{\mu}_2) + \alpha_f^{(2)} f_2'(\overline{\mu}_2) + t + \Delta t = 0,$$

which is a contradiction. □

A.3 Proofs of Results in Chapter 4

Proof of of Lemma 4.4.1. To establish formulation (S-Diff), let us first consider costumers' behavior. Given two prices p_1 and p_2, the demand streams for the two servers must lead to the same customer surplus at equilibrium. Hence,

$$a_i - p_i - \frac{c}{\mu - \lambda_i} = constant, \text{ for } i = 1, 2,$$

where $0 < \lambda_i < \Lambda$ is the customer rate for server i, or the market share of server i. Then, we consider the firm's decision on maximizing the total revenue $p_1\lambda_1 + p_2\lambda_2$. We deduce that the firm will exploit the customer's surplus at equilibrium; hence the constant in the above equality is zero. Therefore, firm's pricing must satisfies

$$a_i = p_i + \frac{c}{\mu - \lambda_i}, \text{ for } i = 1, 2. \tag{A.13}$$

Let $x_i = \lambda_i$, and so the vector $\mathbf{x} = (x_1, x_2) \in \mathrm{R}^2$ is the decision variable. Then, by (A.13), the prices are $p_i = a_1 - c/(\mu - x_1) \geq 0$. Hence, the firm's problem is

$$\max r_D(\mathbf{x}) = \sum_{i=1,2} x_i \left(a_i - \frac{c}{\mu - x_i} \right)$$

$$\text{s.t.} \qquad 0 \leq x_i \leq \mu - \frac{c}{a_i}, \quad x_1 + x_2 \leq \Lambda \forall i = 1, 2. \tag{A.14}$$

Since the partial market coverage assumption holds, we see that the last constraint $x_1 + x_2 \leq \Lambda$ is redundant. In fact, the solution will not be at the boundary where the whole market is covered; i.e. $x_1 + x_2$ is strictly smaller than Λ at optimality. Therefore, we have the formulation (S-Diff).

One immediate observation is that (S-Diff) is a convex program and the objective function is separable in x_1 and x_2. Because condition (4.2) holds, the solution is an internal point. Moreover, solving each x_i is to solve the same problem as in the single server case, which directly leads us to the desired results. □

Proof of of Lemma 4.4.2. Since the firm is offering a lottery to the customers, *ex ante* the customers can see the firm as a virtual single-server system. Hence, just like in the basic model,

we have

$$\bar{a} = p_o + c\bar{W}(\lambda_o), \tag{A.15}$$

where

$$\bar{W} = \sum_{i=1,2} \frac{\beta_i}{\mu - \lambda_o \beta_i}$$

is the expected waiting time. Based on the customers' equilibrium behavior (A.15), the firm seeks a maximized revenue:

$$\max_{p_o \geq 0, 0 \leq \beta_1 \leq 1} \lambda_o p_o.$$

As can be observed, it is difficult to solve (A.15) and represent the customer volume λ_o in term of the uniform price p_o. Therefore, we let $\mathbf{y} = (y_1, y_2) \in \mathrm{R}^2$ be the decision variable, where $y_i = \lambda_o \beta_i$ is the induced customer rate for server i. Hence, at equilibrium, the firm's revenue can be written as

$$\begin{aligned} \lambda_o p_o &= \lambda_o \left(\bar{a} - c \sum_{i=1,2} \frac{\beta_i}{\mu - \lambda_o \beta_i} \right) \\ &= \sum_{i=1,2} y_i \left(a_i - \frac{c}{\mu - y_i} \right). \end{aligned}$$

Moreover, the constraint that $p_o \geq 0$ is translated by (A.15) to

$$\bar{a} - c \sum_{i=1,2} \frac{\beta_i}{\mu - \lambda_o \beta_i} \geq 0,$$

which is equivalent to (by multiplying on both sides $\lambda_o \geq 0$)

$$\sum_{i=1,2} \left(y_i a_i - \frac{c y_i}{\mu - y_i} \right) \geq 0.$$

Finally, note that y_i must be smaller than μ to ensure a stable system. Therefore, we have established formulation (S-Unif').

To recover the decision variables, it is direct to note that $y_1 + y_2 = \lambda_o$ and $\beta_i = y_i / \lambda_o$. Applying (A.15), we have $p_o = \beta_1 a_1 + \beta_2 a_2 - c \sum_i \beta_i (\mu - \lambda_o \beta_i)^{-1}$, which reduces to the desired result. □

Proof of Proposition 4.5.2. We will prove it by induction on the time period t. When $t = T$, since $R_{T+1} = 0$, we directly check the derivatives and see that

$$J_T(\mathbf{a}_T, \mathbf{z}_T) = \sum_{i=1,2} z_{iT} \left(a_{iT} - \frac{c}{\mu - z_{iT}} \right)$$

is separable in term of i, the server index, and concave in each z_{iT}. Hence, result (a) holds for

$t = T$. Result (b) holds true vacuously for $R_{T+1}(\mathbf{a}_{T+1})$.

Suppose that the results hold for $t+1$. By induction assumption, R_{t+1} is increasing and concave. By directly differentiating L with respect to the second component (recall its analytical form in (4.6)), we see that $L(a, z)$ is increasing and concave in z. Moreover, it is clear that $r(\mathbf{a}_t, \mathbf{z}_t)$ is concave in $\mathbf{z}_t$. Since $J_t(\mathbf{a}_t, \mathbf{z}_t) = r(\mathbf{a}_t, \mathbf{z}_t) + R_{t+1}(\mathbf{a}_{t+1})$ and $\mathbf{a}_{t+1} = L(\mathbf{a}_t, \mathbf{z}_t)$, we therefore deduce that $J_t(\mathbf{a}_t, \mathbf{z}_t)$ is concave in $\mathbf{z}_t$. To see (b), note that J_t is linear in $\mathbf{a}_t$ and increasing in each of its component. Besides, the feasible sets A_t (in both differentiated and uniform pricing) becomes larger when a_{it} gets larger, we see that $R_t(\mathbf{a}_t)$ is increasing in a_{it}. Furthermore, $L(a, z)$ is increasing and concave in a, so $R_{t+1}(L(\mathbf{a}_t, \mathbf{z}_t))$ is also concave in $\mathbf{a}_t$. Since the concavity is preserved under maximization (Proposition B4 in Heyman and Sobel (2003)), claim (b) also holds for case t. The induction concludes that the results hold for all time period $t = 1, 2, \cdots, T$. □

Proof of of Lemma 4.4.3. Note that (4.4) in fact requires the objective function to be non-negative. Since $\mathbf{y} = 0$ is feasible and the resulting objective function equals 0, the optimal objective function should be non-negative automatically. □

Proof of Proposition 4.6.1. Define

$$f(x, a) = x\left(a - \frac{c}{\mu - x}\right), \quad \frac{c}{\mu} \le a \le a_{max}; \quad 0 \le x \le \mu.$$

Then f is strictly concave in x and is maximized by

$$x^*(a) = \mu - \sqrt{\frac{c\mu}{a}}, \tag{A.16}$$

where $x^*(a)$ is increasing in a. Let

$$g(a) = \max_x f(x, a) = (\sqrt{\mu a} - \sqrt{c})^2.$$

Clearly, $g(a)$ is increasing and

$$g'(a) = \mu - \sqrt{\frac{c\mu}{a}} \le \mu. \tag{A.17}$$

According to our formulations (M-Diff) and (M-Unif), we have the following bounds. (1) R_U^* is dominated by the total revenue of the case where $a_1 = a_2 = a_{max}$; i.e. both servers are at the maximum quality level in the beginning. (2) R_D^* is no smaller than the total revenue obtained by myopic heuristic, which simply solves (M-Diff) as T "one-shot" sub-problems. In the above cases, the dynamic from learning process is simplified and therefore we have

$$R_U^* \le \sum_{t=1}^{T} \sum_{i=1,2} g(a_{max})$$

and

$$R_D^* \geq \sum_{t=1}^{T} \sum_{i=1,2} g(a_{it}), \quad \text{where} \quad a_{max} - a_{i,t+1} = (a_{max} - a_{it})e^{-\delta x^*(a_{it})}.$$

Moreover, by (A.16), $x^*(a_{it}) \geq x^*(a_1)$; let $b := x^*(a_1) > 0$ be a fixed positive number. Then we have

$$\frac{a_{max} - a_{i,t+1}}{a_{max} - a_{it}} \leq e^{-\delta b}. \tag{A.18}$$

Further, by intermediate theorem, there exists a series $\{\xi_{it}\}$ such that

$$\begin{aligned} g(a_{max}) - g(a_{it}) &= g'(\xi_{it})(a_{max} - a_{it}) \\ &\leq \mu(a_{max} - a_{it}) \quad \text{by (A.17)} \\ &\leq \mu e^{-\delta bt}(a_{max} - a_1) \quad \text{by (A.18) and } a_1 < a_2. \end{aligned}$$

Therefore,

$$\begin{aligned} R_U^* - R_D^* &\leq \sum_{t=1}^{T} \sum_{i=1,2} (g(a_{max}) - g(a_{it})) \\ &\leq \sum_{t=1}^{T} \sum_{i=1,2} \mu e^{-\delta bt}(a_{max} - a_1) \\ &= 2\mu(a_{max} - a_1)\frac{e^{-\delta b}(1 - e^{-\delta bT})}{1 - e^{-\delta b}} \\ &< 2\mu(a_{max} - a_1)\frac{e^{-\delta b}}{1 - e^{-\delta b}}. \end{aligned} \tag{A.19}$$

Finally, note that

$$R_D^* \geq \sum_{t=1}^{T} \sum_{i=1,2} g(a_1) = 2Tg(a_1).$$

Hence, as $T \to \infty$, $R_D^* \to \infty$ whereas $R_U^* - R_D^*$ is bounded by a finite constant. This proves part (a) of the proposition.

As for part (b), consider the bound in (A.19) with a fixed T. As $\delta \to \infty$, we have

$$\frac{e^{-\delta b}(1 - e^{-\delta bT})}{1 - e^{-\delta b}} \to 0,$$

which gives the desired result. □

Proof of Proposition 4.6.2. By the assumptions and the approximation (4.12), the problem is simplified. In particular, we can just consider server 1. For server 2, the optimal revenue is fixed; let π^* be server 2's optimal revenue in one period. Then let z_1 and z_2 be server 1's

customer volume in period 1 and 2. We consider the function

$$\pi(z_1, z_2) = z_1(a_1 - \frac{c}{\mu - z_1}) + z_2(a_1 + \delta z_1 - \frac{c}{\mu - z_2}),$$

and the optimization problem (U): $\max\{\pi(z_1, z_2) | 0 \leq z_1, z_2 \leq \mu\}$. Apply first order condition, we have

$$\frac{\partial \pi}{\partial z_1} = a_1 + \delta z_2 - \frac{c\mu}{(\mu - z_1)^2} = 0$$

and

$$\frac{\partial \pi}{\partial z_2} = a_1 + \delta z_1 - \frac{c\mu}{(\mu - z_2)^2} = 0.$$

Let $l(s) := a_1 + \delta s$ and $r(s) := c\mu/(\mu - s)^2$. We directly check that for $0 \leq s \leq \mu$, $l(s)$ intersects with $r(s)$ exactly once; let the intersection be $s = z$. Hence, problem (U) has solution $z_1 = z_2 = z$. Note that since the slope of $l(s)$ is δ, z can be written as a function of δ; and $z(\delta)$ is increasing in δ. We claim that, as a function of δ, $\pi(z, z)$ is also increasing in δ. To see this, note that z is the zero of the first order condition, and thus

$$\frac{d}{d\delta}\pi(z, z) = 2\left(\delta z + a_1 - \frac{c\mu}{(\mu - z)^2}\right) z'(\delta) + z^2 = z^2 > 0.$$

Now consider the two pricing schemes. For differentiated pricing, we solve problem (U) under constraint

$$z_1 \leq \bar{x} := \mu - \frac{c}{a_1}. \tag{A.20}$$

For uniform pricing, however, because of its opacity, we solve (U) under a relatively relaxed constraint, i.e.

$$z_1 \leq \bar{y}, \tag{A.21}$$

where $\bar{y}$ is uniquely determined by solving the following equation for y:

$$y(a_1 - \frac{c}{\mu - y}) + \pi^* = 0.$$

Hence, $\bar{y} > \bar{x}$. Define δ_1 and δ_2 such as $z(\delta_1) = \bar{x}$ and $z(\delta_2) = \bar{y}$. Then $\delta_2 > \delta_1$.

Finally, we characterize Δ as a function of δ. When $0 < \delta \leq \delta_1$, the constraints (A.20) and (A.21) are both unbinding. Hence $R_U^* = R_D^*$, and $\Delta = 0$. When $\delta_1 < \delta \leq \delta_2$, constraint (A.20) starts to be binding but (A.21) is still not. Therefore, R_D^* remains fixed while R_U^* is increasing (because $\pi(z, z)$ is increasing in δ, as shown above); so Δ increases in δ. When $\delta_2 < \delta \leq \tilde{\delta}$, both constraints are binding, so both R_U^* and R_D^* are unchanging. Therefore, Δ remains a constant. We conclude that $\Delta(\delta)$ is non-decreasing over $0 < \delta \leq \tilde{\delta}$. □

Appendix B

Detailed Computations

The following appendix gives all the detailed computations that we promised in Chapter 3. These computations are tedious and elementary. Hence for the sake of brevity in the proof part, we delay all of them to this point. All the computations are attributed to Maple (version 14). We basically calculate functions' derivatives and determine their signs in a brute force way. Throughout this section, we use $x = \mu_1$ and $y = \mu_2$ for clearer display.

(a). Proposition 3.3.2: The convexity of inverse-idleness function $g_1(x|y, \phi_{\mathbf{FSF}})$.

Suppose $1/2 < x < y$. The second derivative

$$g_1''(x) = \frac{2(P_1 + y^2 P_2)}{(yx + y^2 + 3y + x + 1)^3 (x + y - 1)^3 x^3},$$

where

$$P_1 = y^8 + (6x + 4)y^7 + x(15x + 21)y^6 + 2x^3(x + 4)(x + 1)^2 y + x^3(x + 1)^3 > 0$$

and P_2 has the form of

$$P_2 = b_3 y^3 + b_2 y^2 + b_1 y + b_0.$$

To be specific, $b_3 = 20x^3 - 10 + 6x + 45x^2 > 40x^2 - 10 > 0$, $b_2 = 54x^3 + 27x^2 - 24x + 15x^4 > 48x^2 - 24x > 0$, $b_1 = -9x^2 + 39x^4 + 4 + 49x^3 + 6x^5 - 9x > 36x^4 - 9x^2 + 36x^3 - 9x > 0$, and $b_0 = 23x^3 + x^6 + 15x^5 + 1 + 39x^4 - 3x^2 > 6x^3 - 3x^2 > 0$. Hence, $g_1''(x) > 0$.

Suppose now $x > y > 1/2$. Similar computation gives

$$g_1''(x) = \frac{2(x^3 Q_1 + Q_2)}{(yx + y^2 + 1 + 2y)^3 (x + y - 1)^3 x^3}.$$

We have

$$Q_1 = y^3x^3+3y^2(1+y+2y^2)x^2+3y(6y^2+1+3y+5y^4+5y^3)x+(20y^6+30y^5+34y^4+25y^3+1+5y+13y^2) > 0$$

and Q_2 takes the form

$$Q_2 = ax^2 + bx + c,$$

where $a = 6y^3+15y^4+30y^6+27y^5+3y^2+15y^7 > 0$, $b = -9y^3+9y^6-12y^4-6y^5+15y^7+6y^8-3y^2$

and $c = -5y^6 + y^7 - 5y^5 + 3y^8 + y^4 + y^9 + y^2 + 3y^3$. Further examination gives

$$b^2 - 4ac = -3y^4((1-y)^2 + 8y^3 + 8y^4)(y-1)^2(y+1)^6 < 0.$$

Hence, $Q_2 > 0$ and thus $g_1''(x) > 0$. This completes the proof.

(b). Proposition 3.3.3: The convexity of unfairness function $f_1(x|y, \bar{\phi})$.

First, let us examine the case $\bar{\phi} = \phi_{\text{SSF}}$. For $x < y$, we have

$$f_1''(x) = \frac{2(y^7 A_1 + y^3 A_2 + A_3)}{(x^2 + y^2 + 3y^2x + y^3 + 2yx^2 + x^3y + 2x^2y^2 + y^3x)^3},$$

where $A_1 = y^2+(6x+3)y+2+15x^2+15x > 0$, $A_3 = (6x^4+9x^5+9x^2+11x^3)y^2+(x^3+2x^6)y+x^6 > 0$. Furthermore, $A_2 = b_3y^3 + b_2y^2 + b_1y + b_0$ where $b_3 = 30x^2 - 4 + 20x^3 + 15x > 16x^2 - 4 > 0$, $b_2 = 30x^3 - 7 + 6x + 15x^4 + 42x^2 > 28x^2 - 7 > 0$, $b_1 = 15x^4 + 52x^3 + 39x^2 - 3x - 3 + 6x^5 > 12x^3 - 3x + 12x^2 - 3 > 0$, and $b_0 = -3x + 36x^3 + 30x^4 + 3x^5 + 21x^2 + x^6 > 6x^2 - 3x > 0$. Hence, we conclude that $f_1''(x) > 0$.

For $x > y$, we use the same technique (with a bit abuse of notations):

$$f_1''(x) = \frac{2y(y^5 B_1 + y^2 B_2 + B_3)}{(x^2 + y^2 + 3x^2y + x^3 + 2y^2x + x^3y + 2x^2y^2 + y^3x)^3},$$

where $B_1 = y^2 + (6x+4)y + 21x + 15x^2 > 0$, $B_3 = (x^6 + 15x^5 + 39x^4 + 25x^3 + 3x^2)y + x^3(x^3 + 9x^2 + 9x + 3) > 0$. Moreover, $B_2 = b_2y^2 + b_1y + b_0$. In this case, $b_2 = 20x^3 + 45x^2 + 6x - 8 > 32x^2 - 8 > 0$, $b_1 = 27x^2 + 15x^4 - 1 + 56x^3 - 21x > 4x^2 - 1 + 23x^2 + 56x^3 - 21x > 0$, and $b_0 = 51x^3 + 6x^5 - 9x + 42x^4 - 3x^2 > 12x^4 - 3x^2 + 36x^3 - 9x > 0$. Hence, we conclude that $f_1''(x) > 0$. We thus finish proof of part (ii).

Consider the case $\bar{\phi} = \phi_{\text{FSF}}$. Note that the expression of f_1 function is the same as $f_1(x|y, \phi_{\text{SSF}})$ when $x < y$. The above proof is valid regardless of the order of x and y. Hence, for $x > y$, $f_1''(x|y, \phi_{\text{FSF}}) > 0$.

We now are left with Prop and HH policies. If $\bar{\phi} = \phi_{\text{HH}}$, we look at $v(x)$ and it suffices to prove $v''(x) > 0$ for $x > y$. Note that

$$v''(x) = \frac{4yw(x)}{(x^2 + y^2 + 2xy(x+y))^3},$$

where

$$w(x) = (1 + 4y + 2y^2)x^3 + (6y^3 + 6y^2 + 3y)x^2 + (-3y^2 + 6y^4)x - y^3 + 2y^5 - 2y^4.$$

Note that $w(y) = 8y^4 + 16y^5 > 0$, $w'(y) = 6y^2(1+2y)^2 > 0$, and

$$w''(x) = (12y^2 + 6 + 24y)x + 12y^2 + 6y + 12y^3 > 0.$$

Hence, $w(x) > 0$ for all $x > y$. This shows that $v''(x) > 0$.

If $\bar{\phi} = \phi_{\text{Prop}}$, $v(x)$ is not convex. As a result, we need to analyse the original function f_1. For $x > y$,

$$f_1''(x) = \frac{2y^2 h(x)}{(x^2 + y^2 + xy(x+y))^4}.$$

We repeatedly factorize $h(x)$ and get

$$h(x) = (((q(x+y-1) + r_4)(x-y) + r_3)(x-y) + r_2)(x+y-1) + r_1.$$

In the above,

$$\begin{aligned} q &= \left(1 + y^2 + 2\,y\right) x^4 + \left(4\,y^3 + 4\,y + 2 + 6\,y^2\right) x^3 + \left(4\,y + 13\,y^2 + 14\,y^3 + 3 + 8\,y^4\right) x^2 \\ &\quad + \left(4 + 12\,y^5 + 18\,y^3 + 18\,y^2 + 32\,y^4\right) x + 42\,y^4 + 38\,y^5 + 21\,y^2 - 6\,y^3 - 6\,y + 5 + 16\,y^6 \\ &> 12y^4 - 6y^3 + 12y^2 - 6y > 0, \end{aligned}$$

$$r_1 = (y-1)^2(2y-1)^2(y^2 - y + 1)^2 > 0,$$

$$r_2 = \left(2\,y^4 + y^3 + 2\,y^2 - 2\,y + 1\right)\left(y^3 + 2\,y^2 + 1\right)(2\,y-1)^2 > 0,$$

$$\begin{aligned} r_3 &= 1 - 4y + 8y^2 + 84y^7 + 44y^8 - 10y^3 + 8y^4 + 25y^6 - 20y^5 \\ &= 44(y^2 + 2.6y + 1.9)(y^2 + 0.5y + 0.38)(y^2 - 0.39y + 0.182)(y^2 - 0.8y + 0.18) > 0, \end{aligned}$$

$$\begin{aligned} r_4 &= 6 - 14\,y + 32\,y^2 + 20\,y^7 - 32\,y^3 + 38\,y^4 + 50\,y^6 + 20\,y^5 \\ &= 20(y + 2.54)(y^2 + 0.94y + 1.32)(y^2 - 0.15y + 0.32)(y^2 - 0.84y + 0.28) > 0. \end{aligned}$$

Along with $x > y$, this shows that $h(x) > 0$ and therefore the $f_1'' > 0$, which complete the proof.

(c). Theorem 3.3.4: Validation of equation (A.12)**.**

Assume $x \geq y$. There are three inequalities in (A.12) to verify for two policies. Denote the difference of left hand side and right hand side in the three inequalities by V_1, V_2 and V_3 respective. We need to prove that they are positive.

(1). $V_1 = \frac{\partial^2 g_1}{\partial x^2}\frac{\partial^2 g_2}{\partial y^2} - \frac{\partial^2 g_1}{\partial x \partial y}\frac{\partial^2 g_2}{\partial y \partial x} > 0$.

First, assume $\bar{\phi} = \phi_{\text{Prop}}$. Direct computation gives

$$V_1 = \frac{r(x)}{x^3 y^3 (1+x)^2 (1+y)^2 (x+y-1)^5},$$

where $r(x)$ is a polynomial of x with coefficients being polynomials of y. Instead of write it explicitly, we give its derivatives w.r.t. x and the evaluations at the point $x = y$. Given the derivative and all the initial conditions, one can always trace back the original function by

integration. Note that

$$r^{(5)}(x) = 240(y+1)(18y^3 - 3y^2 + 36y^2x + 24yx + 3y + 42x^2 - 2) > 6y^3 - 3y^2 + 8x^2 - 2 > 0,$$

$$r^{(4)}(x)|_{x=y} = 24y(257y^2 + 614y^3 + 396y^4 + 8y - 22) > 88y^2 - 22 > 0,$$

$$r^{(3)}(x)|_{x=y} = 18y^2(95y^2 + 322y^3 + 252y^4 + 3y - 16) > 64y^2 - 16 > 0,$$

$$r''(x)|_{x=y} = 2y^3(177y^2 + 938y^3 + 792y^4 - 3y - 52) > 2y^3(12y^3 - 3y + 416y^3 - 52) > 0,$$

$$r'(x)|_{x=y} = 3y^4(21y^2 + 176y^3 + 144y^4 - 3y - 10) > 3y^4(12y^2 - 3y + 80y^3 - 10) > 0,$$

$$r(x)|_{x=y} = 3y^5(4y^2 + 2y + 1)(8y^2 + 7y - 4) > 9y^5/2 > 0.$$

By the above equations, we prove that $r(x)$ is always positive for any $x \geq y$ and $y > 1/2$. Hence $V_1 > 0$. This is exactly the same technique that we apply throughout this part; we even bear a bit abuse of notations to use $r(x)$ again and again. Henceforth, we display the results without detail explanations.

Now assume $\bar{\phi} = \phi_{\mathrm{HH}}$. Then we reach the conclusion $V_1 > 0$ by noting the following:

$$V_1 = \frac{r(x)}{x^3y^3(1+2x)^2(1+2y)^2(x+y-1)^5},$$

where $r^{(7)}(x) = 40320(2y+1)(y+1) > 0$, $r^{(6)}(x)|_{x=y} = 42480y + 167040y^2 + 138240y^3 - 2280 > 0$, and

$$r^{(5)}(x)|_{x=y} = 20280y^2 + 111360y^3 + 117120y^4 - 3720y - 480 > 0,$$

$$r^{(4)}(x)|_{x=y} = 24y(230y^2 + 2000y^3 + 2720y^4 - 95y - 22) > 0,$$

$$r^{(3)}(x)|_{x=y} = 12y^2(60y^2 + 1280y^3 + 2240y^4 - 72y - 25) > 0,$$

$$r''(x)|_{x=y} = 8y^3(-16y^2 + 512y^3 + 1088y^4 - 28y - 15) > 0,$$

$$r'(x)|_{x=y} = 8y^4(-14y^2 + 128y^3 + 288y^4 - 6y - 5) > 0,$$

$$r(x)|_{x=y} = 16y^5(4y^2 + y - 1)(8y^2 + 2y + 1) > 0.$$

(2). $V_2 = \frac{\partial^2 f_1}{\partial x^2} - \frac{\partial^2 f_1}{\partial x \partial y} > 0$.
First, assume $\bar{\phi} = \phi_{\mathrm{Prop}}$, then

$$V_2 = \frac{2(x+y-1)r(x)}{(x^2+y^2+xy(x+y))^4},$$

where

$$\begin{aligned} r^{(4)}(x) &= -120y^2x - 264y^2 + 360yx^2 + 264y^3 + 192y^4 + 1680x^3y \\ &\quad + 3360x^4y + 2880x^2y^2 + 2160y^3x + 2160x^2y^3 + 240y^4x + \\ &\quad 5040x^3y^2 + 840x^3 + 3360x^4 + 360x^2 + 480yx \\ &> 120xy(x-y) + 264(x^2-y^2) > 0. \end{aligned}$$

Moreover, $r^{(3)}(y) = 12y^4(44 + 274y + 231y^2) > 0$, $r''(y) = 4y^4(38y - 24 + 211y^2 + 146y^3) > 0$, $r'(y) = 8y^5(13y-5)(y+1)^2 > 0$, and $r(y) = 8y^6(2y-1)(y+1)^2 > 0$. Hence $V_2 > 0$.

Second, assume $\bar{\phi} = \phi_{\text{HH}}$, then

$$V_2 = \frac{8(x+y-1)r(x)}{(x^2+y^2+2xy(x+y))^4},$$

where

$$\begin{aligned} r^{(5)}(x) &= 480y + 720x + 2160yx - 840y^2 + 10080yx^2 \\ &\quad + 8640y^2x + 3360y^3 + 53760x^3y + 60480x^2y^2 + \\ &\quad 17280y^3x + 960y^4 + 2520x^2 + 26880x^3 \\ &> 840(x^2-y^2) > 0. \end{aligned}$$

Also note that $r^{(4)}(y) = 192y^2(3 + 5y + 94y^2 + 225y^3) > 0$, $r^{(3)}(y) = 48y^5(115 + 231y) > 0$, $r''(y) = 16y^4(-6 - 7y + 85y^2 + 146y^3) > 0$, $r'(y) = 8y^5(13y-5)(2y+1)^2 > 0$, and $r(y) = 8y^6(2y-1)(2y+1)^2 > 0$, as desired.

(3). $V_3 = \frac{\partial^2 g_2}{\partial y^2} - \frac{\partial^2 g_2}{\partial y \partial x} > 0$.

Assume $\bar{\phi} = \phi_{\text{Prop}}$, then

$$V_3 = \frac{r(x)}{y^3(1+x)^2(x+y-1)^2},$$

where $r''(x) = 24x^2 + 24xy + 8y + 4y^2 - 4 > 0$, $r'(y) = 6y(4y^2 + 2y - 1) > 0$, and $r(y) = y^2(8y^2 + 7y - 4) > 0$. Thus $V_3 > 0$.

If $\bar{\phi} = \phi_{\text{HH}}$, then

$$V_3 = \frac{(x+y)r(x)}{y^3(1+2x)^2(x+y-1)^2}.$$

Since $r''(x) = 24x + 8y - 4 > 0$, $r'(y) = 20y^2 + y - 2 > 0$, and $r(y) = 2y(4y^2 + y - 1) > 0$, we conclude $V_3 > 0$.