



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations

---

2014

## Essays on Service Information, Retrials and Global Supply Chain Sourcing

Shiliang Cui

University of Pennsylvania, shiliang.cui@gmail.com

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Operational Research Commons](#)

---

### Recommended Citation

Cui, Shiliang, "Essays on Service Information, Retrials and Global Supply Chain Sourcing" (2014). *Publicly Accessible Penn Dissertations*. 1247.

<https://repository.upenn.edu/edissertations/1247>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/1247>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Essays on Service Information, Retrials and Global Supply Chain Sourcing

## Abstract

In many service settings, customers have to join the queue without being fully aware of the parameters of the service provider (for e.g., customers at check-out counters may not know the true service rate prior to joining). In such "blind queues", customers typically make their decisions based on the limited information about the service provider's operational parameters from past experiences, reviews, etc. In the first essay, we analyze a firm serving customers who make decisions under arbitrary beliefs about the service parameters. We show, while revealing the service information to customers improves revenues under certain customer beliefs, it may however destroy consumer welfare or social welfare.

When consumers can self-organize the timing of service visits, they may avoid long queues and choose to retry later. In the second essay, we study an observable queue in which consumers make rational join, balk and (costly) retry decisions. Retrial attempts could be costly due to factors such as transportation costs, retrial hassle and visit fees. We characterize the equilibrium under such retrial behavior, and study its welfare effects. With the additional option to retry, consumer welfare could worsen compared to the welfare in a system without retrials. Surprisingly, self-interested consumers retry too little (in equilibrium compared to the socially optimal policy) when the retrial cost is low, and retry too much when the retrial cost is high. We also explore the impact of myopic consumers who may not have the flexibility to retry.

In the third essay, we propose a comprehensive model framework for global sourcing location decision process. For decades, off-shoring of manufacturing to China and other low-cost countries was a no-brainer decision for many U.S. companies. In recent years, however, this trend is being challenged by some companies to re-shore manufacturing back to the U.S., or to near-shore manufacturing to Mexico. Our model framework incorporates perspectives over the entire life cycle of a product, i.e., product design, manufacturing and delivering, and after-sale service support, and we use it to test the validity of various competing theories on global sourcing. We also provide numerical examples to support our findings from the model.

## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

## Graduate Group

Operations & Information Management

## First Advisor

Morris A. Cohen

## Second Advisor

Senthil K. Veeraraghavan

## Keywords

Global Supply Chain, Rational Retrials, Service Information, Service Operations

## Subject Categories

Operational Research

ESSAYS ON SERVICE INFORMATION, RETRIALS  
AND GLOBAL SUPPLY CHAIN SOURCING

Shiliang Cui

A DISSERTATION

in

Operations and Information Management

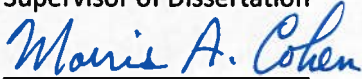
For the Graduate Group in Managerial Science and Applied Economics  
Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the  
Degree of Doctor of Philosophy

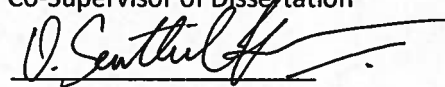
2014

Supervisor of Dissertation



Morris A. Cohen  
Panasonic Professor of Manufacturing & Logistics  
Professor of OPIM

Co-Supervisor of Dissertation



Senthil K. Veeraraghavan  
Associate Professor of OPIM

Graduate Group Chairperson



Eric T. Bradlow  
The K.P. Chao Professor  
Professor of Marketing, Statistics and Education

Dissertation Committee:

Gérard P. Cachon, Fred R. Sullivan Professor of Operations and Information Management  
Noah F. Gans, Anheuser-Busch Professor of Management Science, Professor of OPIM  
Xuanming Su, Associate Professor of OPIM

ESSAYS ON SERVICE INFORMATION, RETRIALS  
AND GLOBAL SUPPLY CHAIN SOURCING

© COPYRIGHT

2014

Shiliang Cui

This work is licensed under the  
Creative Commons Attribution  
NonCommercial-ShareAlike 3.0  
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

*for Jiaqi*

## ACKNOWLEDGEMENT

I could not have completed my Ph.D. studies without the help, support and efforts from a lot of people. First, I would like to express my deepest gratitude to my dissertation co-advisors Dr. Morris Cohen and Dr. Senthil Veeraraghavan for their guidance, support and encouragement throughout the course of my Ph.D. studies. Morris and Senthil have continuously inspired me to learn, to think, to conduct better research, and to become a better person. The guidance I received from them made my Ph.D. purposeful and fulfilling, and I can not imagine having better advisors and mentors than the two.

I would like to thank the chair of my dissertation committee Dr. Noah Gans (formly the departmental Ph.D. coordinator), for guiding me through my professional development at Wharton and for taking his time to carefully read and evaluate my research work. I would also like to thank my committee members and co-authors, Dr. Gérard Cachon and Dr. Xuanming Su for their guidance and patience over the years and for their insightful comments on my work. Finally, my sincere thanks goes to all the faculty, students, staff and alumni of the Operations and Information Management Department.

I am grateful to Dr. Evan Fisher, Dr. Christopher Ruebeck and Dr. Ge Xia, who encouraged me to pursue my Ph.D. degree when I was an undergraduate student, and supported me ever since. I would also like to thank all of my friends for their support and encouragement.

Most importantly, none of this would have been possible without the unconditional love and support of my family. I would like to thank my wife, my parents and my parents-in-law for everything they have done for me. In the meantime, I am extremely proud that my wife, Jiaqi Li, has joined me in pursuing a Ph.D. degree at the University of Pennsylvania.

## ABSTRACT

### ESSAYS ON SERVICE INFORMATION, RETRIALS AND GLOBAL SUPPLY CHAIN SOURCING

Shiliang Cui

Morris Cohen

Senthil Veeraraghavan

In many service settings, customers have to join the queue without being fully aware of the parameters of the service provider (for e.g., customers at check-out counters may not know the true service rate prior to joining). In such “blind queues”, customers typically make their decisions based on the limited information about the service provider’s operational parameters from past experiences, reviews, etc. In the first essay, we analyze a firm serving customers who make decisions under arbitrary beliefs about the service parameters. We show, while revealing the service information to customers improves revenues under certain customer beliefs, it may however destroy consumer welfare or social welfare.

When consumers can self-organize the timing of service visits, they may avoid long queues and choose to retry later. In the second essay, we study an observable queue in which consumers make rational join, balk and (costly) retry decisions. Retrial attempts could be costly due to factors such as transportation costs, retrial hassle and visit fees. We characterize the equilibrium under such retrial behavior, and study its welfare effects. With the additional option to retry, consumer welfare could worsen compared to the welfare in a system without retrials. Surprisingly, self-interested consumers retry too little (in equilibrium compared to the socially optimal policy) when the retrial cost is low, and retry too much when the retrial cost is high. We also explore the impact of myopic consumers who may not have the flexibility to retry.

In the third essay, we propose a comprehensive model framework for global sourcing location decision process. For decades, off-shoring of manufacturing to China and other low-cost countries was a no-brainer decision for many U.S. companies. In recent years, however, this trend is being challenged by some companies to re-shore manufacturing back to the U.S., or to near-shore manufacturing to Mexico. Our model framework incorporates perspectives over the entire life cycle of a product, i.e., product design, manufacturing and delivering, and after-sale service support, and we use it to test the validity of various competing theories on global sourcing. We also provide numerical examples to support our findings from the model.



## TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iv
ABSTRACT . . . . .	v
LIST OF TABLES . . . . .	ix
LIST OF ILLUSTRATIONS . . . . .	xi
CHAPTER 1 : INTRODUCTION . . . . .	1
CHAPTER 2 : BLIND QUEUES: THE IMPACT OF CONSUMER BELIEFS ON REVENUES AND CONGESTION . . . . .	5
2.1 INTRODUCTION . . . . .	5
2.2 MODEL . . . . .	9
2.3 CUSTOMER BELIEFS UNDER THE LACK OF SERVICE INFORMATION	14
2.4 IMPACT OF REVEALING SERVICE INFORMATION . . . . .	26
2.5 APPLYING OUR FINDINGS TO SPECIFIC BELIEF MODELS . . . . .	30
2.6 UNOBSERVABLE QUEUES . . . . .	36
2.7 CONCLUSIONS AND IMPLICATIONS . . . . .	42
CHAPTER 3 : A MODEL OF RATIONAL RETRIALS IN QUEUES . . . . .	46
3.1 INTRODUCTION . . . . .	46
3.2 A MODEL OF RETRIALS: BASE MODEL . . . . .	51
3.3 EQUILIBRIUM STRATEGIES . . . . .	61
3.4 CONSUMER WELFARE ANALYSIS . . . . .	73
3.5 EXTENDED MODEL: TWO CLASSES OF CONSUMERS . . . . .	82
3.6 CONCLUSIONS AND IMPLICATIONS . . . . .	90

CHAPTER 4 : MANUFACTURING SOURCING IN A GLOBAL SUPPLY CHAIN: A LIFE CYCLE ANALYSIS . . . . .	92
4.1 INTRODUCTION . . . . .	92
4.2 LITERATURE REVIEW . . . . .	95
4.3 MODEL . . . . .	98
4.4 EXTENDED MODELS . . . . .	111
4.5 NUMERICAL ILLUSTRATIONS . . . . .	123
4.6 CONCLUSIONS . . . . .	129
APPENDICES . . . . .	131
BIBLIOGRAPHY . . . . .	184

## LIST OF TABLES

TABLE 1 :	Illustration of the possible equilibrium strategy types (with $v = 65$ and $\frac{c}{\mu} = 10$ ).	63
TABLE 2 :	$\mathcal{G}_1$ and $\mathcal{G}_{2'}$ used in the proof of Lemma 3/(ii)	138
TABLE 3 :	$\mathcal{G}_1$ and $\mathcal{G}_{2'}$ used in the proof of Lemma 3/(iii)	142
TABLE 4 :	$\mathcal{G}_1$ and $\mathcal{G}_{2'}$ used in the proof of Theorem 2'/(ii)	145
TABLE 5 :	$\mathcal{G}_1$ and $\mathcal{G}_{2'}$ used in the proof of Theorem 2'/(iii)	146

## LIST OF ILLUSTRATIONS

FIGURE 1 :	As consumers remember more service experiences ( $s \uparrow \infty$ ), their estimates of the service rate become consistent with the true service rate. That is, (i) $\mathbb{E}(\tilde{\mu}_s) \downarrow \mu$ , and (ii) $\text{Var}(\tilde{\mu}_s) \downarrow 0$ . . . . .	33
FIGURE 2 :	Illustration of the join, balk versus retriial decisions in our model.	52
FIGURE 3 :	Illustration of consumer welfare at equilibrium as a function of the retriial cost $\alpha \in (0, \alpha_L]$ . . . . .	75
FIGURE 4 :	Illustration of consumer welfare at equilibrium as a function of the retriial cost $\alpha$ over the entire region. The welfare decreases in $\alpha$ on $(0, \alpha_L]$ , increases on $[\alpha_L, \alpha_H]$ and stays flat on $[\alpha_H, \infty)$ . . . . .	78
FIGURE 5 :	Illustration of equilibrium strategies (bottom) versus the socially optimal policies (middle) as a function of the retriial cost $\alpha$ . Side by side comparison is given at top. . . . .	80
FIGURE 6 :	Illustration of the welfare <i>per strategic consumer</i> at equilibrium as the myopic population increases. Note that $\alpha_1, \alpha_2, \dots, \alpha_{N-1}, \alpha_L$ and $\alpha_H$ all decrease in $\theta$ , and $\alpha_H \rightarrow \alpha_L$ as $\theta \rightarrow 1$ . For any particular $\theta \in [0, 1)$ , the welfare curve remains the down-up-flat shape. . . . .	87
FIGURE 7 :	Illustration of the welfare <i>per myopic consumer</i> as the myopic population increases. . . . .	88
FIGURE 8 :	U.S. manufacturing employment. Data: U.S. Bureau of Labor Statistics, <a href="http://www.bls.gov/iag/tgs/iag31-33.htm#workforce">http://www.bls.gov/iag/tgs/iag31-33.htm#workforce</a> . . . . .	93
FIGURE 9 :	Sequence of events in the two-period model. Decisions are underlined. Features in the extended models in Section 4 are denoted with * . . . . .	100

FIGURE 10 : Comparison of hourly wage received by manufacturing workers in the U.S. and China. Data is retrieved from the website of U.S. Bureau of Labor Statistics, except that the last three data points for the Chinese wage are estimated by extrapolation (not available from the website). . . . .	110
FIGURE 11 : Estimated landed cost for a product that has high labor cost and low shipping cost. . . . .	124
FIGURE 12 : Expected profit for producing a product that has high labor cost and low shipping cost. . . . .	125
FIGURE 13 : Estimated landed cost for a product that has high labor input and high shipping cost. . . . .	126
FIGURE 14 : Expected profit for producing a product that has high labor cost and high shipping cost. . . . .	127
FIGURE 15 : Estimated landed cost and expected profit for producing a product that has low labor cost and high shipping cost. . . . .	128
FIGURE 16 : In these three subfigures, we let $l = 0.7$ and plot $f_1(\rho)$ and $f_2(\rho)$ w.r.t. different values of $r$ . In every case, $f_1(\rho)$ and $f_2(\rho)$ cross the point $(\rho = 1, f(\rho) = \frac{l}{1-l}) = (1, \frac{7}{3})$ . (a) When $r = \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ , $f_2(\rho)$ is a tangent line to the curve $f_1(\rho)$ . (b) When $r = 3 > \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ , they cross above. (c) When $r = \frac{1}{2} < \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ , they cross below. .	154
FIGURE 17 : Illustration of the welfare under socially optimal policies (the upper envelope of $L_1, L_2, \dots, L_N$ and $L$ ). . . . .	169

## CHAPTER 1 : INTRODUCTION

This dissertation contains two essays on modeling consumer behavior in service operations management and one essay on global operations and supply chain management.

The goal of the first two essays is to model consumer decision-making in queueing settings, and study their effect on congestion, revenues and welfare. Much of the Service Operations literature in queueing does not include psychological considerations in consumer decisions. Our research project seeks to fill this gap, by demonstrating such decision-making process can significantly influence queue outcomes. In particular, we consider two distinct settings: First, we consider settings when consumers join queues with biased information – we call them “blind queues”. Second, we consider “retrial queues” based on the observation that consumers often postpone their decisions based on several considerations. Both of these decision-making issues have not been considered in the queueing literature.

Much of the traditional operations literature on queues assumes that the service parameters (service time distributions, etc.) are common knowledge and fully known to consumers when making their decisions. While this is acceptable in the operations literature, in reality, it is likely that only the service firm knows its capacity, and the consumers may not be fully informed of the service capacity. In fact, it is even likely that consumers could be misinformed about a firm’s service capacity. In such blind queues, consumers typically make their decisions based on the limited information and the biases they arrive with.

For instance, a customer might have visited a restaurant or an amusement park only once or twice. It is conceivable that his estimate of the service time will be strongly dependent on the bias formed based from previous service experience. In some cases, consumers might augment their information using feedback from external acquaintances, but even such information is likely to be a smaller sample than what is needed to know the full service distribution (which is often assumed to be known accurately in the literature).

On the other hand, almost all queueing models have focused on consumers' join and balk decisions – with customers not making forward looking decisions. In reality, when the queue is very long, consumers may not be willing to wait, rather they choose to retry later (as opposed to balking). For instance, consider a customer who arrives at the package pick-up service at a post office, or a customer who goes to have his parking permit renewed. Upon seeing the status of the queue, i.e., the number of consumers that are already in the queue, this customer can either decide to join the queue or to leave only to return back at a later more convenient time. In scenarios such as the post office and the parking permit renewal examples, and in many other real-life queues such as discretionary shopping decisions, postponing is a commonly used practice among consumers.

In a blind queue, it is important to understand if the service firm is motivated to reveal its private service information to the customers to alleviate costs of biases, or just want to remain blind. In a retrial queue, it is important to understand if retrials can make the consumers, the firm, and/or the society better off.

We will pursue modeling work based on the classical queueing model by Naor (1969) but extend it in two ways. Naor (1969) studies a single-server system with an observable queue where rational consumers (who know the service parameters) make join or balk decisions. In contrast, we will model consumers with decision biases, who may have arbitrarily different (even misled) beliefs about the service rate in the blind queues, and then allow the consumers to postpone their join or balk decision at later time period with a retrial cost in the retrial queues. This retrial cost can be an external cost incurred by the customer, but not collected by the server (for e.g., costs associated with transportation back and forth, or a retrial “hassle” cost such as rescheduling other activities, which may cause irrational deviations) or an internal fee collected by the server (for e.g., a toll for entering the system, such as visit fees and copays in insurance services).

The third essay will look at emerging issues on global supply chain re-structuring. For decades, a dominant strategy in manufacturing has been to outsource to low labor-cost

countries such as China. This has led to the transfer of manufacturing jobs and development activities out of the U.S. to these low labor-cost countries. Today, however, this trend is being challenged by a movement by some companies to move back their manufacturing to the U.S. (i.e., “re-shoring”), or moving it to Mexico (i.e., by “near-shoring”). At the same time many firms continue to select offshore locations for outsourcing of material inputs and services. Our research aims to study the drivers and impact of these outsourcing, re-shoring and near-shoring decisions.

In a recent review, we found over 50 cases where major U.S. and global companies have announced significant re-structuring decisions to their global supply chain in the past three years. Among them, 19 companies have increases outsourcing by shifting production to an off-shore location. In contrast, 19 other companies have re-shored by bringing production back to their home country. Well-known examples include decisions by Apple to invest \$100 million in producing some of its Mac computers in the U.S., and General Electronics to invest \$1 billion into domestic appliances manufacturing capabilities at Louisville, Kentucky. There is also evidence that some companies are near-shoring by bringing production to a country that is closer to major customers, and that others are investing in manufacturing technology (e.g., the adoption of robots).

A popular theory for the re-shoring phenomenon is called the “Tipping Point Theory”, which comes from the observation of rising wages in low labor-cost countries. Take China for example. Chinese wages are growing at a rate of 15 percent annually, as opposed 2 percent in the U.S. As a result, the landed-cost advantage of producing a product in China is diminishing, and could be eliminated in a few years. The Tipping Point Theory thus argues that when this advantage falls below a critical level, more and more manufacturers will re-shore to the U.S.

There exist competing theories other than the Tipping Point Theory raised by scholars to explain companies’ changes on their global supply chain manufacturing decisions. The perspectives of these theories include cost of ownership, real options, product development



and innovation, information technology and automation, government policies and supply chain risks, etc.

Existing global sourcing literature has mainly focused on the impact of costs on sourcing decisions. We, however, propose a comprehensive model that will incorporate perspectives from product design, manufacturing and after-sale service support in order to test the various theories noted above. To the best of our knowledge, this research will be the first to conduct a life cycle analysis of global sourcing strategy.

The rest of this dissertation contains the three essays:

Chapter 2: Blind Queues: The Impact of Consumer Beliefs on Revenues and Congestion (under the supervision of Prof. Senthil Veeraraghavan).

Chapter 3: A Model of Rational Retrials in Queues (under the supervision of Prof. Xuanming Su & Prof. Senthil Veeraraghavan).

Chapter 4: Manufacturing Sourcing in a Global Supply Chain: A Life Cycle Analysis (under the supervision of Prof. Morris Cohen).

## CHAPTER 2 : BLIND QUEUES: THE IMPACT OF CONSUMER BELIEFS ON REVENUES AND CONGESTION

### 2.1. INTRODUCTION

Almost all the literature on queues assumes that service parameters are common knowledge and fully known to customers when making their decisions. In reality, it is likely that only the service firm knows its capacity, and customers may not be fully informed of the service capacity. It is even possible that customers could be systematically misinformed about a firm's service capacity. Hence, it is important to understand if the firm is motivated to reveal its private information on service rate to the customers, and if the firm does reveal the information, whether the information would increase consumer welfare or firm revenues. While there are papers that have focused on firms announcing (real time) delay information in terms of queue and waiting time information to its customers, those models also typically assume that the firm's service parameters are known to the customers.

We expect that customers that have had limited past interactions with a service provider will not be able to accurately predict its true service rate. For instance, a customer might have visited a restaurant or an amusement park only once or twice, and it is conceivable that her best estimate of the service capacity will be based on the service times she had experienced (in the absence of other inputs). In some cases, customers might augment their information using feedback from external acquaintances, but even such information is likely to be a smaller sample than what is needed to know the full service distribution (which is often assumed to be known accurately in the literature). In line with many real-life services, but in contrast to the existing literature, we allow for the customers to *not know* the service parameters accurately, unless they are informed about it by the firm. We term such queues *blind queues*.

Our approach to the analysis is general, i.e., individual customers can have arbitrarily different beliefs about the service rate. It might be possible that the population is correct

on average but individual customers may be idiosyncratically misinformed. We also consider the possibility that the population as a whole is mis-informed systematically.

In observable queues, when customers arrive with different beliefs about the true service rate, they exhibit different balking behaviors due to their beliefs. For instance, a customer may join the queue when he ought not to (if he overestimates the service rate), and conversely, an impatient misinformed customer may balk from the queue when he should not.

Note that customers' internal beliefs about the service rate are not observable; Typically, the server only observes customers' joining/balking decisions. Hence, we work with the observable balking threshold distribution that results from the original service-rate belief distribution. There is some recent empirical evidence (see Lu et al. (2013) that uses queueing data from a Deli), supporting the approach that customers in observable queues may rely primarily on the length of the queue to make their purchasing/joining decisions. Thus, by understanding the impact of balking thresholds on system performance, our results could be further implemented to models where service values and waiting costs are heterogeneous.

### *2.1.1. Related Literature*

The literature on queueing models with strategic customers dates back to the seminal paper of Naor (1969), who studies a single-server system with an observable queue. In Naor's model, homogeneous customers (who *know* the service parameters) observe the queue length upon arrival before making a decision to join the system. Because of homogeneity, customers have identical balking thresholds.

In our paper, customers are not aware of the true service parameters. We allow them to have arbitrarily distributed heterogeneous beliefs over the service rate. Thus, our work is also closely related to the classical queueing papers with heterogeneous customers (with full information), in addition to those papers that examine the effect of delay announcements. Following Naor (1969), queues with heterogeneous service values and time costs have been studied, as seen in the comprehensive review by Hassin and Haviv (2003).

There is a large volume of literature that examines the provision of fixed or variable *delay* information (i.e. queue length or real-time waiting time, etc.) to arriving customers. In the context of call-centers, there are several papers that study the provision of current delay information. For instance, see Armony and Maglaras (2004a,b) and Jouini et al. (2011). We refer the reader to an excellent review by Akşin et al. (2007) of the call center literature, on the role of delay information on customers' balking behavior. Nevertheless, the service capacity and arrival information is often assumed to be known to all customers in these papers.

Hassin (1986) considers a revenue-maximizing server who may hide queue lengths to improve revenue. Whitt (1999) shows that customers are more likely to be blocked in a system where the delay information is not provided to a system where it is provided. Guo and Zipkin (2007) studies an  $M/M/1$  queue extension with three modes of information: no information, partial information (the queue length) and full information (the exact waiting time). Economou and Kanta (2008) and Guo and Zipkin (2009) study models where some partitioned queue information (such as range of queue-lengths) is available to customers to make their decisions. However, in all the aforementioned papers (including the no-information cases), customers are aware of the service rate parameter.

Thus, there are very few papers that treat service information as a firm's private information about which customers are either not informed or have incorrect beliefs. Besbes et al. (2011) and Debo and Veeraraghavan (2014) analyze customers' equilibrium joining behaviors in queues with limited information on service rate. In Hassin (2007), the true service rate is either fast or slow. While the probabilities of the service rate being fast or slow is known to customers, the server can choose to reveal or not to reveal the realized rate. In Guo et al. (2011), partial distributional information is conveyed to the customers, who then employ the max-entropy distribution in deciding whether to join or balk from the queues. In these papers, individual customers have correct distributional information over the service rate.

In contrast, we do impose any such condition, i.e., individual customers could have incorrect

information. We study the impact of announcing service information in such cases in both observable and unobservable queue settings. To the best of our knowledge, we are not aware of other papers that deal with customer decisions when the service provider has not provided *any* information about its service parameters to the customers.

Finally, our approach complements the perspective in Besbes and Maglaras (2009) and Haviv and Randhawa (2012), where the service firm does not fully know the demand (volume) information. Instead, we study a system where customers do not know a firm's service information. We focus only on the decision whether the firm should reveal the unknown information to the customers.

The main theoretical contributions of our paper can be summarized as follows:

1. Customers may have arbitrary balking thresholds due to their beliefs and decision frameworks. We characterize those beliefs under which the firm will gain revenues from revealing its service information.
2. We show that as customer balking threshold beliefs become less dispersed in the population, the firm improves its revenues. If the customer population systematically underestimates the service capacity, the service provider should always reveal its service rate information.
3. The welfare effects of information revelation are mixed. Typically, system congestion (both queue lengths and wait times) increases with information revelation. Individual consumer welfare thus typically declines with more information, especially when a firm with a high service rate releases its information.
4. We find that social welfare may fall with more information, in both observable and unobservable queue settings, because the improvement in revenues may be insufficient to overcome the consumer welfare loss.
5. Our approach on blind queues is general and does not depend on the origin of the initial

belief distributions, which may emerge from bounded rationality, sampling, learning from past experiences, etc. We show that sampling from finite data creates consumer optimism, but will lead to true-learning asymptotically. We show that Quantal response beliefs (bounded rational errors) can be biased, but are not consistent with learning by sampling.

The paper is structured as follows. Section 2.2 introduces the model and characterizes the system performance in terms of belief distributions. In Section 2.3, we analyze customer populations with different beliefs. In Section 2.4, we investigate the impact of the revelation of service information on revenues, congestion, and welfare. In Section 2.5, we incorporate our belief structures into different cognitive models in the literature. In Section 2.6, we extend the model and results to unobservable queues. All technical proofs are deferred to the appendix.

## 2.2. MODEL

We focus on a single-server queueing system. (All results extend to a multi-server system.) Customers arrive to the queue according to a Poisson process at rate  $\lambda > 0$  per unit time. The service time is exponentially distributed with service rate  $\mu > \lambda$ . Let  $\rho \triangleq \lambda/\mu < 1$  denote the traffic intensity. Arriving customers line up at the server if the server is busy, and the queue discipline is first-come first-served (FCFS). Every arriving customer observes the number of the customers already waiting in the system before making an irrevocable join or balk decision (i.e, there is no renegeing). On joining, all customers incur a linear waiting cost of  $c$  per unit time when they wait. The server provides a service of value  $v$ . Thus, all customers are homogeneous in their service valuation and in their waiting costs. The firm charges an *exogenous* price  $p$  for its service.

Upon arrival, the customers decide whether to join the queue based on the net value they expect to receive from the service (i.e.,  $v - p$ ) and their expected waiting cost. Suppose the customers knew the true service rate  $\mu$ . Then, a customer arriving when there are  $n$

customers already in the system, would join the queue when  $v - p - (n + 1)c/\mu \geq 0$ , or balk otherwise. This is the model described in Naor (1969).

**Model of Customer (mis-)Information:** In contrast to the existing literature, we relax the assumption that customers are aware of the service time distribution or the service rate. We posit that customers typically will not have complete information on the true distribution. For instance, customers with limited or idiosyncratic past interactions with the server, may have widely varying service rate beliefs.

In this paper, we use the superscript  $\tilde{\cdot}$  (tilde) to describe customer beliefs. Customers may have heterogeneous beliefs regarding the service rate, and their beliefs could differ arbitrarily from the true service rate  $\mu$ . We denote customer service rate beliefs by  $\tilde{\mu} \in (0, \infty)$  with some cumulative distribution function (cdf)  $G_{\tilde{\mu}}$  across the entire population. Note that every customer has a deterministic belief. The beliefs form a random distribution because customers with different beliefs arrive to the system randomly.

If the mean of the random variable,  $\tilde{\mu}$ , is equal to the true  $\mu$ , i.e., the service rate belief of the population is “correct” on average, we describe the service rate beliefs of the population as being *consistent*. If the mean of  $\tilde{\mu}$  is not the true  $\mu$ , we address the service rate beliefs as being *biased*. Specifically, if the population mean is greater (less) than  $\mu$ , the service rate beliefs are *optimistic (pessimistic)*, i.e., the population is on average optimistic (pessimistic) on the speed of the server.

Upon arrival, a customer with belief  $\tilde{\mu}$  who observes  $n$  customers currently waiting in the system (including the person who is under the service, if any) makes the following decision:

$$\left\{ \begin{array}{ll} v - p \geq \frac{(n+1)c}{\tilde{\mu}} : & \text{The customer joins the queue;} \\ \text{otherwise:} & \text{The customer balks from the queue.} \end{array} \right.$$

Throughout the paper, we will assume  $v - p \geq \frac{c}{\tilde{\mu}}$  a.s. to eliminate trivial outcomes and ensure that customers will join an empty queue.

**Balking Threshold Beliefs:** Define  $\tilde{N} \triangleq \lfloor \frac{\tilde{\mu}(v-p)}{c} \rfloor$ , i.e.,  $\tilde{N}$  is an integer such that  $\tilde{N} \leq \frac{\tilde{\mu}(v-p)}{c} < \tilde{N} + 1$ . Intuitively,  $\tilde{N}$  describes the *balking threshold belief* for a customer with service rate belief  $\tilde{\mu}$ : the customer who arrives to see  $n$  customers waiting in the system will join if  $n+1 \leq \tilde{N}$  and balks, otherwise. Each customer has a deterministic balking threshold as a result of his (internally-held) service rate belief, and will make a deterministic decision upon seeing the queue length.

Let  $F_{\tilde{N}}$  be the cdf that characterizes the random variable  $\tilde{N}$ . The balking thresholds are random because customers appear at random at the queue. Note that customers' beliefs about the service rate may be drawn from the continuous distribution  $G_{\tilde{\mu}}$ , whereas the balking threshold beliefs are drawn from a discrete distribution  $F_{\tilde{N}}(n) = \Pr[\tilde{N} \leq n]$ . Since  $v - p \geq \frac{c}{\tilde{\mu}}$ , we have  $\tilde{N} \in \{1, 2, \dots\}$ . In essence, we translate the (uncountable) customer beliefs regarding the service rate to actions dictated by beliefs about balking thresholds (which is countable). For notational convenience, we suppress the subscript  $\tilde{N}$  in  $F_{\tilde{N}}$  and denote  $F_{\tilde{N}}$  simply as  $F$  wherever unambiguous.

Note that the server only observes customers' joining/balking decisions, but not their service rate beliefs. Hence, we will directly analyze the balking threshold beliefs  $\tilde{N}$ . Our terminology on biases in beliefs (pessimism, optimism and consistency) also appropriately applies to balking threshold beliefs. Biases in service rate beliefs typically follow the same direction as the biases in the corresponding balking threshold beliefs, but not always, because of the floor function in the mapping from  $\tilde{\mu}$  to  $\tilde{N}$ .

**System Evolution under Threshold Beliefs:** We have a population comprised of customers who are heterogeneous in their joining behavior due to varying individual balking threshold beliefs. Since  $\tilde{N} \in \{1, 2, \dots\}$ , we have a queuing system with state-dependent arrivals - a system whose buffer size equal to the maximum balking threshold (possibly infinity, in which case we have an  $M/M/1$  system). In contrast, note that when customers fully know  $\mu$ , we get the classical  $M/M/1/N$  system with state-independent arrivals that emerges in Naor (1969) where  $N \triangleq \lfloor \frac{\mu(v-p)}{c} \rfloor$ .



Let the state of system be denoted by  $i$  where  $i$  is the number of customers in the system (including the customer at the server). Since  $\lambda < \mu$ , this queueing system is recurrent, and long-run steady state probabilities exist. Let  $\pi_i$  denote the long-run probability that the system is in state  $i$ . Now consider state  $i$ : among all arrivals, only those customers who have the balking threshold greater than or equal to  $i + 1$  will join the queue. Thus, the effective joining probability at state  $i$  is given by  $\Pr[\tilde{N} \geq i + 1] = \Pr[\tilde{N} > i] = \bar{F}(i)$  (by letting  $\bar{F}_{\tilde{N}}(\cdot) = 1 - F_{\tilde{N}}(\cdot)$ ). The effective arrival rate at any state  $i$  is  $\lambda\bar{F}(i)$ .

From the steady state rate balance equations, we have  $\pi_{i+1} = \rho\bar{F}(i)\pi_i$  for  $i \in \{0, 1, 2, \dots\}$ , which gives  $\pi_i = \rho^i\pi_0 \prod_{n=0}^{i-1} \bar{F}(n)$  for  $i \in \{1, 2, 3, \dots\}$ . Since  $\rho < 1$ , it follows from  $\sum_{i=0}^{\infty} \pi_i = 1$  that

$$\pi_0 = 1 \left/ \left( 1 + \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right) \right. . \quad (2.1)$$

The average number of customers in the system, denoted by  $L$ , is given by

$$\begin{aligned} L &= \sum_{i=0}^{\infty} i\pi_i = \sum_{i=1}^{\infty} i\pi_i = \pi_0 \sum_{i=1}^{\infty} i\rho^i \prod_{n=0}^{i-1} \bar{F}(n) \\ &= \sum_{i=1}^{\infty} i\rho^i \prod_{n=0}^{i-1} \bar{F}(n) \left/ \left( 1 + \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right) \right. . \end{aligned} \quad (2.2)$$

By convention, we set all empty products to 1. For instance,  $\prod_{n=0}^{-1} \bar{F}(n) = 1$ . Then,

$$L = \sum_{i=0}^{\infty} i\rho^i \prod_{n=0}^{i-1} \bar{F}(n) \left/ \left( \sum_{i=0}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right) \right. . \quad (2.3)$$

The long-run revenue rate at the server, denoted by  $R$ , is given by  $p\mu(1 - \pi_0)$ , e.g., see Larsen (1998). It follows that the long-run effective arrival rate at the system, denoted by  $\lambda_{eff}$ , is:

$$\lambda_{eff} = \mu(1 - \pi_0) < \lambda \quad (2.4)$$

Also note that,  $\pi_{i+1} = \rho \bar{F}(i) \pi_i$  for  $i \in \{0, 1, 2, \dots\}$ . Summing up over  $i \geq 0$ , we get,

$$\sum_{i=0}^{\infty} \pi_{i+1} = \sum_{i=0}^{\infty} \rho \bar{F}(i) \pi_i \Leftrightarrow \sum_{i=1}^{\infty} \pi_i = \rho \sum_{i=0}^{\infty} \bar{F}(i) \pi_i \Leftrightarrow (1 - \pi_0) = \rho \sum_{i=0}^{\infty} \bar{F}(i) \pi_i. \quad (2.5)$$

Alternatively,  $\lambda_{eff}$  is given by  $\sum_{i=0}^{\infty} \pi_i \lambda \bar{F}(i)$ , and we have

$$\lambda_{eff} = \lambda \sum_{i=0}^{\infty} \bar{F}(i) \pi_i = \mu \cdot \rho \sum_{i=0}^{\infty} \bar{F}(i) \pi_i = \mu(1 - \pi_0) \quad (\text{from Condition (2.5)}). \quad (2.6)$$

Finally, let  $W$  denote the average time a customer spends in the system, i.e, his waiting time in the queue plus his service time. By Little's Law,

$$\begin{aligned} W &= \frac{L}{\lambda_{eff}} = \frac{\pi_0 \sum_{i=1}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}(n)}{\mu(1 - \pi_0)} \quad (\text{from conditions (2.2) and (2.6)}) \\ &= \frac{1}{\mu} \frac{\pi_0}{1 - \pi_0} \sum_{i=1}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}(n). \end{aligned} \quad (2.7)$$

Recall from (2.1) that  $\pi_0 = 1 / \left( 1 + \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right)$ , hence,  $\frac{\pi_0}{1 - \pi_0} = 1 / \left( \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right)$ . Plugging in (2.7), we have

$$W = \sum_{i=1}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}(n) / \left( \mu \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right). \quad (2.8)$$

Thus, as long as we can characterize the balking threshold beliefs  $\tilde{N}$ , we can derive the performance measures for the queueing system through its cdf  $F$ . This allows us to compare any two systems with populations that differ arbitrarily in their beliefs. To this end, in the next section, we set up a sequence of systems with customer beliefs that are stochastically ordered in some sense.

## 2.3. CUSTOMER BELIEFS UNDER THE LACK OF SERVICE INFORMATION

When  $\mu$  is fully known to customers, the balking threshold distribution is a one-point distribution (i.e., all customers have identical balking threshold). In contrast, balking thresholds are distributed arbitrarily when customers are not fully informed. Recall that when there is optimistic (pessimistic) bias, the average balking threshold is higher (lower) than the true threshold. In §2.3.1, we consider balking threshold belief distributions that have bias. In §2.3.2, we compare populations with consistent beliefs, where the average balking threshold beliefs are accurate, but there is arbitrary variability on the individual thresholds. Our analysis in these sections assists us in pinning down the performance differences among queueing systems that differ in their customer beliefs.

We will focus on the system in which the customers' balking threshold beliefs are distributed over a finite interval in §2.3.1 and §2.3.2, i.e., a system in which there is no customer in the population who has infinite patience to always join the queue. However, all of our results extend to multiple servers (see §2.3.3) and beliefs with infinite support (see §2.3.4).

### 2.3.1. Population with Biased Beliefs

We first compare systems under beliefs  $\tilde{N}$  and  $\tilde{N}'$  that differ in their mean.

**Definition 1** First Order Stochastic Dominance (FOSD): (*Quirk and Saposnik, 1962*) Let  $\mathcal{F}$  and  $\mathcal{G}$  be the cdf's of random variables  $X$  and  $Y$ .  $X$  is said to be smaller than  $Y$  with respect to the first-order stochastic order (written  $X \leq_{st} Y$ ) if  $\mathcal{F}(t) \geq \mathcal{G}(t)$  for all real  $t$ , or equivalently, if  $\bar{\mathcal{F}}(t) \leq \bar{\mathcal{G}}(t)$  for all real  $t$ .

FOSD is also termed as the *usual stochastic order* by Müller and Stoyan (2002), and frequently called *the stochastic order*. It is well known that variables ordered by FOSD have different means (e.g., see Theorem 1.2.9/(a) in Müller and Stoyan (2002)). Through the FOSD relation, we can compare two threshold beliefs with respect to their 'biases'. Essentially, pessimistic threshold beliefs are stochastically dominated by more optimistic beliefs.

We use  $\lambda_{eff,\tilde{N}}, R_{\tilde{N}}, L_{\tilde{N}}, W_{\tilde{N}}$  to denote the long-run effective arrival (rate), the long-run revenue (rate) at the firm, the average number of customers and the average time a customer spends in the system when the balking threshold beliefs of the population are characterized by  $\tilde{N}$ . The following Theorem 1 compares the performance metrics of two systems under beliefs ordered by FOSD.

**Theorem 1** *If  $\tilde{N} \leq_{st} \tilde{N}'$ , then (i)  $\lambda_{eff,\tilde{N}} \leq \lambda_{eff,\tilde{N}'}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}'}$ ); (ii)  $L_{\tilde{N}} \leq L_{\tilde{N}'}$ ; (iii)  $W_{\tilde{N}} \leq W_{\tilde{N}'}$ .*

Proof of Theorem 1 follows directly by comparing the stochastics of two queues and by taking expectations of the distribution of the corresponding performance metrics (e.g., see Theorem 1 in Bhaskaran (1986)). Also see Berger and Whitt (1992) which compares two queues each with a single balking threshold. In our model, there are multiple balking thresholds in the same queue depending on the individual customer beliefs.

Theorem 1 states that if customers become more patient (higher balking thresholds), the firm serves more customers and receives higher revenue, at the same time incurring a higher system congestion. Note that the results from Theorem 1 are distribution-free, i.e., the performance metrics of queues can be ordered for any balking threshold belief distribution, as long as the underlying distributions can be (first-order) stochastically ordered. Also, note that the ordering of performance metrics is invariant to the true service rate of the system.

### *2.3.2. Population with Consistent Beliefs*

In this section, we consider mean-preserving spreads to examine balking threshold belief distributions that have the same mean as the “true” balking threshold, but differ in how the individual balking thresholds are distributed. If all customers knew the service rate exactly, then every customer would use the same “true” balking threshold  $N$ , which is a special consistent belief distribution. We use the notion of Single Mean Preserving Spread to order belief distributions that are consistent.

**Definition 2** Single Mean Preserving Spread (SMPS): *(Rothschild and Stiglitz, 1970)* Let  $\mathcal{F}$  and  $\mathcal{G}$  be the cdf's of two discrete random variables  $X$  and  $Y$  whose common support is a sequence of real numbers  $a_1 < a_2 < \dots < a_n$ . Suppose the probability mass functions  $f$  and  $g$  describe  $X$  and  $Y$  completely:  $\Pr(X = a_i) = f_i$  and  $\Pr(Y = a_i) = g_i$  where  $\sum_{i=1}^n f_i = \sum_{i=1}^n g_i = 1$ . Suppose  $f_i = g_i$  for all but four  $i$ , say  $i_1, i_2, i_3$  and  $i_4$  where  $i_k < i_{k+1}$ . Define  $\gamma_{i_k} = g_{i_k} - f_{i_k}$ . Then we say that  $Y$  differs from  $X$  by a single Mean Preserving Spread (written  $\mathcal{F} \leq_{SMPS} \mathcal{G}$ ) if  $\gamma_{i_1} = -\gamma_{i_2} \geq 0$ ,  $\gamma_{i_4} = -\gamma_{i_3} \geq 0$  and  $\sum_{k=1}^4 a_{i_k} \gamma_{i_k} = 0$ .

The notion of mean preserving spread (MPS) is often employed to model risk order of two random variables that may have the same mean but different variability. If two distributions  $\mathcal{F}$  and  $\mathcal{G}$  describe the returns of two risky investments, and  $\mathcal{F} \leq_{MPS} \mathcal{G}$ , then the distribution  $\mathcal{F}$  is considered less risky. SMPS in Definition 2 is a stricter condition than MPS:  $\mathcal{F} \leq_{SMPS} \mathcal{G} \Rightarrow \mathcal{F} \leq_{MPS} \mathcal{G}$ .

Consider consistent balking threshold beliefs  $\tilde{N}$ , i.e.,  $\mathbb{E}[\tilde{N}]$  equals the balking threshold  $N$  (when the service parameters are fully known to the customers). We seek to compare the performance metrics under beliefs  $\tilde{N}$  to the metrics if the true parameters of the system were known. To that end, we create a sequence of random variables ordered by SMPS that begin at an initial belief distribution. Using a fairly general but intuitive construction technique, we will show that the sequence (generated using our construction) will terminate at a specific “final” distribution within a finite number of steps, regardless of the initial distribution. We then characterize an ordering of the performance metrics for the entire sequence. This construction not only allows us to compare the performance under the initial belief distribution to the canonical system with fully informed customers, but it also facilitates a comparison between any two arbitrary (consistent) balking threshold belief distributions.

Let our initial beliefs be characterized by some random variable  $\tilde{N}_0$ . In Construction 1, we create a sequence of random variables  $\{\tilde{N}_K\}_{K \geq 0}$  (the  $K$ -th term in the sequence is distributed with the cdf  $F_K$ ), and discuss the properties of the sequence. The cdf  $F_K$  has

support over some finite sequence of natural numbers  $a_{K_1} < a_{K_2} < \dots < a_{K_n}$ . We denote by  $f_K$  its probability mass function (pmf) such that  $f_K(a_{K_i}) > 0$  for  $i \in \{1, 2, \dots, n\}$  and  $\sum_{i=1}^n f_K(a_{K_i}) = 1$ .

Consider the transformation of  $\tilde{N}_K$  to  $\tilde{N}_{K+1}$  in the following Construction 1. The succeeding random variable in the sequence,  $\tilde{N}_{K+1}$ , is constructed from the preceding random variable  $\tilde{N}_K$  by taking an equal probability mass from both ends of the distribution  $F_K$  and adding those weights to the “middle” of the support.

**Construction 1**

$$\text{When } a_{K_n} - 1 > a_{K_1} + 1, \left\{ \begin{array}{l} f_{K+1}(a_{K_1}) = f_K(a_{K_1}) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1}(a_{K_1} + 1) = f_K(a_{K_1} + 1) + \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1}(x) = f_K(x) \quad \forall x \in \{a_{K_1} + 2, a_{K_1} + 3, \dots, a_{K_n} - 2\} \\ f_{K+1}(a_{K_n} - 1) = f_K(a_{K_n} - 1) + \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1}(a_{K_n}) = f_K(a_{K_n}) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1} = 0 \text{ otherwise.} \end{array} \right.$$

$$\text{When } a_{K_n} - 1 = a_{K_1} + 1, \left\{ \begin{array}{l} f_{K+1}(a_{K_1}) = f_K(a_{K_1}) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1}(a_{K_1} + 1) = f_K(a_{K_1} + 1) + 2 \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1}(a_{K_n}) = f_K(a_{K_n}) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ f_{K+1} = 0 \text{ otherwise.} \end{array} \right.$$

Stop the sequence when  $\tilde{N}_T$  is such that  $a_{T_n} - 1 < a_{T_1} + 1$ .

We illustrate Construction 1 with an example. Consider  $\tilde{N}_K \in \{3, 4, 5, 7, 9\}$ , with pmf  $f_K(x) = \{0.1, 0.3, 0.2, 0.1, 0.3\}$  for  $x = \{3, 4, 5, 7, 9\}$  respectively. To form  $\tilde{N}_{K+1}$ , Construction 1 requires  $0.1 = \min\{0.1, 0.3\}$  of the probability mass at the “ends” of the support to be re-allocated towards the “middle”, i.e., from “3” to “4”, and also from “9” to “8”. This results in  $\tilde{N}_{K+1} \in \{4, 5, 7, 8, 9\}$ , with pmf  $f_{K+1}(x) = \{0.4, 0.2, 0.1, 0.1, 0.2\}$  for  $x = \{4, 5, 7, 8, 9\}$ . In the next step, applying the transformation from Construction 1 would lead to  $\tilde{N}_{K+2} \in \{4, 5, 7, 8\}$ , with pmf  $f_{K+2}(x) = \{0.2, 0.4, 0.1, 0.3\}$  for  $x = \{4, 5, 7, 8\}$ .

With Construction 1, we create the sequence  $\{\tilde{N}_0, \tilde{N}_1, \dots, \tilde{N}_T\}$ , beginning from an initial belief distribution  $\tilde{N}_0$ , with the corresponding cdf’s  $\{F_0, F_1, \dots, F_T\}$ . Thus, we have a sequence of random variables that describe the customer threshold beliefs that are ordered in some sense. In the following Lemma 1, we show that  $T$  is finite and the distributions are mean-preserving spreads.

**Lemma 1** *Consider a sequence  $\{\tilde{N}_K\}$  from Construction 1. (i) The sequence terminates at some finite  $K = T$ . (ii)  $F_K \leq_{SMPS} F_{K-1}$  for  $K \in \{1, 2, \dots, T\}$ .*

Lemma 1 states that all the distributions along the sequence built through Construction 1 have the same mean (i.e., they obey the mean preserving property). As long as the first balking threshold belief distribution is consistent, all the belief distributions in Construction 1 will be consistent. Furthermore, every succeeding distribution in the sequence is dominated (under the SMPS criterion) by the preceding distribution, i.e., every distribution in the sequence is followed by another distribution that has a lower “spread”.

We now show in Lemma 2 that, for all initial belief distributions that have the same mean, the sequence *always* terminates at the same distribution  $\tilde{N}_T$ . Depending on the parameters of the initial distribution (support etc.), the number of steps taken to reach  $\tilde{N}_T$  may differ. Thus,  $T$  depends on the initial distribution, but  $\tilde{N}_T$  does not.

**Lemma 2** *Given any  $\tilde{N}_0$ , the sequence  $\{F_K\}$  terminates at the same  $F_T$  with the random variable  $\tilde{N}_T \in \{\lfloor \mathbb{E}(\tilde{N}_0) \rfloor, \lceil \mathbb{E}(\tilde{N}_0) \rceil\}$  such that  $\mathbb{E}(\tilde{N}_T) = \mathbb{E}(\tilde{N}_0)$ .*

Now that we have a sequence of random variables ordered SMPS by Construction 1, we can compare the performance metrics of the queueing system under different balking threshold beliefs along the sequence. Using the notation introduced earlier, we denote  $\lambda_{eff, \tilde{N}_K}$ ,  $R_{\tilde{N}_K}$ ,  $L_{\tilde{N}_K}$  and  $W_{\tilde{N}_K}$  the effective arrival, the firm revenue, the average queue length, and the average waiting time, corresponding to  $\tilde{N}_K$  in the sequence of belief distributions  $\{\tilde{N}_K\}_{K \geq 0}$ .

**Lemma 3** *Let  $\{\tilde{N}_K\}$  be any sequence from Construction 1. (i)  $\lambda_{eff, \tilde{N}_K} < \lambda_{eff, \tilde{N}_{K+1}}$  ( $R_{\tilde{N}_K} < R_{\tilde{N}_{K+1}}$ ) for all  $\rho$ ; (ii) If  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right)$ , then  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$ ; and (iii) If  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-2}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-2} \right)$ , then  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$ .*

Recall that Construction 1 builds a sequence of belief distributions with decreasing spreads, while maintaining consistency (i.e., identical means). It follows from Lemma 3/(i) that, regardless of the traffic in the system, the revenues at the firm improve as the customers' belief distributions become less spread out (or narrower). This result is *not* due to Jensen's inequality.<sup>1</sup>

Revenue improvements along the sequence emerge from the following two mechanisms: (i) Customer beliefs are gradually altered along the sequence in the construction which changes the long-run probabilities for *all* states. (ii) Along the construction path, the balking threshold increases for some customers, and decreases for some others. We prove that the throughput/revenue from increased joining of customers with improved balking thresholds compensates for the decreased joining of those customers with reduced balking thresholds. This is proven for *any* prior belief distribution.

Using similar proof arguments, Lemma 3/(ii) and (iii) provide *distribution-free* sufficient conditions for  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  and  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$ , respectively. It is possible to derive

---

<sup>1</sup>Consider a belief  $\tilde{N}_0$  with cdf  $F_0$ , pmf  $f_0$  and some integer mean  $\mathbb{E}(\tilde{N}_0)$ . Jensen's inequality would imply  $R_{\mathbb{E}(\tilde{N}_0)} > \sum_{n=0}^{\infty} f_0(n)R_n$  where  $R_{\mathbb{E}(\tilde{N}_0)}$  and  $R_n$  are revenues when *all* customers use the balking threshold  $\mathbb{E}(\tilde{N}_0)$  and  $n$  respectively. Lemma 3 states that  $R_{\mathbb{E}(\tilde{N}_0)} > R_{\tilde{N}_0} = p\mu(1 - \pi_0)$  where  $\pi_0 = 1 / \left( 1 + \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_0(n) \right)$  (from Equation (2.1)) which is not related to  $\sum_{n=0}^{\infty} f_0(n)R_n$ .



stronger distribution-specific conditions for each inequality. Unlike revenues, which always increase as beliefs become less spread out, expected queue lengths and/or expected waiting times can *increase* or *decrease*. We provide numerical examples below to support this observation.

**Numerical Illustration:** We explore the performance metrics as the balking threshold belief  $\tilde{N}_K$  is transformed into  $\tilde{N}_{K+1}$  according to Construction 1. Since  $\lambda_{eff, \tilde{N}_K} < \lambda_{eff, \tilde{N}_{K+1}}$  (from Lemma 3/(i)), following Little's Law, it is impossible to have  $L_{\tilde{N}_K} > L_{\tilde{N}_{K+1}}$  and  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  at the same time. All three other cases are possible: (i)  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  and  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$ , (ii)  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  and  $W_{\tilde{N}_K} > W_{\tilde{N}_{K+1}}$ , and, (iii)  $L_{\tilde{N}_K} > L_{\tilde{N}_{K+1}}$  and  $W_{\tilde{N}_K} > W_{\tilde{N}_{K+1}}$ .

For example, consider the random variable  $\tilde{N}_K \in \{3, 4, 5\}$ , such that its pmf  $f_K(x) = \{0.2, 0.6, 0.2\}$  for  $x = \{3, 4, 5\}$  respectively. Following Construction 1, we have  $\tilde{N}_{K+1}$  such that  $Pr(\tilde{N}_{K+1} = 4) = 1$ , which is also the last step. Let  $\mu = 1$  in all cases below.

**Case (i):** When  $\rho = 0.4$ ,

$$L_{\tilde{N}_K} = 0.609 < L_{\tilde{N}_{K+1}} = 0.615 \text{ and } W_{\tilde{N}_K} = 1.551 < W_{\tilde{N}_{K+1}} = 1.562,$$

**Case (ii):** When  $\rho = 0.825$ ,

$$L_{\tilde{N}_K} = 1.617 < L_{\tilde{N}_{K+1}} = 1.621 \text{ and } W_{\tilde{N}_K} = 2.265 > W_{\tilde{N}_{K+1}} = 2.262.$$

**Case (iii):** When  $\rho = 0.9$ ,

$$L_{\tilde{N}_K} = 1.793 > L_{\tilde{N}_{K+1}} = 1.790 \text{ and } W_{\tilde{N}_K} = 2.380 > W_{\tilde{N}_{K+1}} = 2.368.$$

Having illustrated the comparative statics for the sequence of beliefs in Lemma 3, we can now compare the performance metrics of (any) initial belief with the terminal belief distribution. This is captured in Theorem 2. It turns out that, when customers' balking threshold beliefs become more accurate, the firm always improves its revenues. On the other hand, customers have to wait longer on average, if the traffic is lighter than some threshold level.

**Theorem 2** *Let  $\tilde{N}$  be any balking threshold belief and  $\tilde{N}_T$  be the last term from Construction 1 initiated at  $\tilde{N}_0 = \tilde{N}$ . Then, (i)  $\lambda_{eff,\tilde{N}} < \lambda_{eff,\tilde{N}_T}$  ( $R_{\tilde{N}} < R_{\tilde{N}_T}$ ) for all  $\rho$ ; (ii)  $\exists \rho_L$  s.t.  $L_{\tilde{N}} < L_{\tilde{N}_T} \forall \rho \leq \rho_L$ ; and (iii)  $\exists \rho_W$  s.t.  $W_{\tilde{N}} < W_{\tilde{N}_T}$  if  $\rho \leq \rho_W$ .*

Theorem 2 indicates that revenue at the server always improves when balking thresholds become less spread-out. Essentially, some customers with high balking threshold become less patient, and some others with low balking thresholds become more patient. Due to the PASTA property, the system occupancy when customers arrive is more likely to be low (recall  $\rho < 1$ ). As a result, there is increased joining of customers leading to higher revenues.

The effect on congestion and waiting times depends on the traffic  $\rho$ . When the traffic intensity  $\rho$  is small, beliefs that are *less* spread-out can increase congestion ( $L$  or  $W$ ). This could be understood through externalities imposed by joining/balking customers as beliefs become less spread out.

At low queue-lengths, more customers join (as beliefs become less spread out) causing increased negative externalities for future arrivals. At high queue-lengths, fewer customers join decreasing the negative externalities at those states. The net effect of the negative externalities imposed depends on the likelihood of the low queues lengths to high queue lengths.

When  $\rho$  is small, the queue typically resides at low states and visits higher states less often. Thus when  $\rho$  is small, the increased negative externalities imposed by customers joining at low states exceed any benefit from reduced negative externalities at high states. As a result, the expected wait times and queue lengths are higher when the beliefs are less spread out.

When  $\rho$  is large, the higher states are relatively more likely to be visited as opposed to when  $\rho$  is small. Thus, it is possible that the benefits accrued at higher states can overcome the negative externalities imposed at lower states. For instance, see Numerical Illustration cases (ii) and (iii). Thus, in low traffic, as balking threshold beliefs become more accurate, customers wait longer and suffer higher disutility. This result is intriguing because for a

given population of consumers, a faster server announcing its true service rate is more likely to result in welfare loss due to increased congestion.

**Lower Bounds:** We can use the analytical properties of the bounds along the sequence to derive *distribution-free* bounds (with respect to  $\tilde{N}$ ) on  $\rho_L$  and  $\rho_W$  that hold for any arbitrary customer belief. The lower bounds  $\underline{\rho}_L$  and  $\underline{\rho}_W$  are such that  $\rho_L \geq \underline{\rho}_L = 0.5$  and  $\rho_W \geq \underline{\rho}_W = 0.414$  respectively. We defer the details of the derivation to the appendix.

We now extend our theoretical findings to the case of a firm with multiple servers (in §2.3.3) and consumer beliefs that have infinite support (in §2.3.4).

### 2.3.3. Beliefs with Multiple Server Queues

We begin by characterizing the evolution of the queue when there are  $s$  identical servers each with service rate parameter  $\mu$  ( $M/M/s$  model). Assume that  $s\mu > \lambda$  so the traffic  $\rho = \lambda/s\mu < 1$ . All other aspects of the model are the same as in the single-server setting.

Let  $\tilde{N} \in \{1, 2, 3, \dots\}$ , whose cdf is  $F$ , describe consumers' balking beliefs. As in the single server case, we assume that every consumer will join the system on arrival if one of the servers is idle. Thus we associate a consumer  $j$ 's balking threshold belief  $\tilde{N}_j \in \{1, 2, 3, \dots\}$  in the  $M/M/s$  system in the following way: Consumer  $j$  with  $\tilde{N}_j$ , will join the  $M/M/s$  system upon arrival, if and only if she observes less than  $\tilde{N}_j + s - 1$  consumers already in the system.<sup>2</sup> This specification ensures that no-one balks when a server is idle, and is consistent with the single-server model when  $s = 1$ .

Let  $\{0, 1, 2, \dots\}$  be the states of the  $M/M/s$  system (number of consumers in the system), and  $\{\pi_i : i = 0, 1, 2, \dots\}$  be the corresponding steady-state probabilities. From the rate

---

<sup>2</sup>For example, consider consumer  $j$  with the strictest balking threshold, i.e.,  $\tilde{N}_j = 1$ . This consumer will join the system if and only if she observes less than  $s$  ( $= \tilde{N}_j + s - 1$ ) consumers in the system, i.e., at least one of the servers is idle.

balance equations, we have:

$$\pi_i = \begin{cases} \frac{\rho^i}{i!} \pi_0 & \text{for } i = 1, 2, \dots, s-1, s. \\ \frac{\rho^i}{s!} \prod_{n=0}^{i-s} \bar{F}(n) \pi_0 & \text{for } i = s, s+1, s+2, \dots \end{cases} \quad (2.9)$$

Note that when  $i = s$  the two cases in Equation (2.9) provide the same result, i.e.,  $\frac{\rho^i}{i!} \pi_0 = \frac{\rho^i}{s!} \prod_{n=0}^{i-s} \bar{F}(n) \pi_0$ , because  $\bar{F}(0) = 1$ . Let  $a \wedge b \triangleq \min\{a, b\}$  and let empty products, if any, be equal to 1, then from (2.9) we have

$$\pi_i = \frac{\rho^i}{(i \wedge s)!} \prod_{n=0}^{i-s} \bar{F}(n) \pi_0 \text{ for } i = 1, 2, 3, \dots \quad (2.10)$$

From (2.10), we derive expressions for performance metrics for the  $M/M/s$  system under the consumer beliefs  $\tilde{N}$  with cdf  $F$ :

$$\pi_0 = 1 \left/ \left( \sum_{i=0}^{s-1} \frac{\rho^i}{i!} + \frac{\rho^{s-1}}{s!} \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n) \right) \right. \quad (2.11)$$

$$R_{\tilde{N}} = p \cdot \mu [s - (s\pi_0 + (s-1)\pi_1 + \dots + 2\pi_{s-2} + 1\pi_{s-1})] \quad (2.12)$$

$$L_{\tilde{N}} = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} \frac{i\rho^i}{(i \wedge s)!} \prod_{n=0}^{i-s} \bar{F}(n) \left/ \sum_{i=0}^{\infty} \frac{\rho^i}{(i \wedge s)!} \prod_{n=0}^{i-s} \bar{F}(n) \right. \quad (2.13)$$

$$W_{\tilde{N}} = \frac{L_{\tilde{N}}}{\lambda_e} = \sum_{i=1}^{\infty} \frac{i\rho^i}{(i \wedge s)!} \prod_{n=0}^{i-s} \bar{F}(n) \left/ \mu \sum_{i=1}^{\infty} (i \wedge s) \frac{\rho^i}{(i \wedge s)!} \prod_{n=0}^{i-s} \bar{F}(n) \right. \quad (2.14)$$

Note that all expressions from (2.11) to (2.14) coincide with the corresponding expressions for the  $M/M/1$  system when  $s = 1$ . We now recover the results of Theorem 1 and Theorem 2 for consumer beliefs in multi-server queues. The proofs can be found in the appendix.

**Theorem 1'** Consider consumer beliefs  $\tilde{N}$  and  $\tilde{N}'$  at an  $M/M/s$  queue. If  $\tilde{N} \leq_{st} \tilde{N}'$ , then (i)  $\lambda_{eff, \tilde{N}} \leq \lambda_{eff, \tilde{N}'}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}'}$ ), (ii)  $L_{\tilde{N}} \leq L_{\tilde{N}'}$  and (iii)  $W_{\tilde{N}} \leq W_{\tilde{N}'}$ .

**Theorem 2'** Let  $\tilde{N}$  be any balking threshold beliefs distribution for the  $M/M/s$  queue. Let  $\tilde{N}_T$  be the last term from Construction 1 initiated at  $\tilde{N}_0 = \tilde{N}$ . Then,  $\lambda_{eff, \tilde{N}} \leq \lambda_{eff, \tilde{N}_T}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}_T}$ ); (ii)  $\exists \rho_L$  s.t.  $L_{\tilde{N}} < L_{\tilde{N}_T} \forall \rho \leq \rho_L$ ; and (iii)  $\exists \rho_W$  s.t.  $W_{\tilde{N}} < W_{\tilde{N}_T}$  if  $\rho \leq \rho_W$ .

#### 2.3.4. Beliefs over an Infinite Support

We further relax the assumption that the balking threshold beliefs have finite support. Recall that Theorem 1 in §2.3.1 states that when consumers have two sets of balking threshold beliefs  $\tilde{N}$  and  $\tilde{N}'$  (which have finite supports) such that  $\tilde{N} \leq_{st} \tilde{N}'$ , then  $\lambda_{eff, \tilde{N}} \leq \lambda_{eff, \tilde{N}'}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}'}$ ),  $L_{\tilde{N}} \leq L_{\tilde{N}'}$  and  $W_{\tilde{N}} \leq W_{\tilde{N}'}$ . It is clear that the approach used in the proof of Theorem 1 continues to hold when  $\tilde{N}$  and  $\tilde{N}'$  have infinite support. Thus we have the following theorem.

**Theorem 1''** Consider consumer beliefs  $\tilde{N}$  and  $\tilde{N}'$  that may be distributed on an infinite support. If  $\tilde{N} \leq_{st} \tilde{N}'$ , then (i)  $\lambda_{eff, \tilde{N}} \leq \lambda_{eff, \tilde{N}'}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}'}$ ), (ii)  $L_{\tilde{N}} \leq L_{\tilde{N}'}$  and (iii)  $W_{\tilde{N}} \leq W_{\tilde{N}'}$ .

However, to extend Theorem 2 to the infinite support case, we need additional preparatory groundwork. Recall that for consistent beliefs  $\tilde{N}$  with finite support, Theorem 2 states that  $R_{\tilde{N}} < R_{\tilde{N}_T}$  for all  $\rho$ ,  $L_{\tilde{N}} < L_{\tilde{N}_T}$  and  $W_{\tilde{N}} < W_{\tilde{N}_T}$  for small  $\rho$ , where  $\tilde{N}_T$  is the terminal distribution in the sequence formed by Construction 1 initiated from  $\tilde{N}$ . Lemma 2 shows that  $\tilde{N}_T \in \{\lfloor \mathbb{E}(\tilde{N}) \rfloor, \lceil \mathbb{E}(\tilde{N}) \rceil\}$  such that  $\mathbb{E}(\tilde{N}_T) = \mathbb{E}(\tilde{N})$ .

Now we relax the assumption on the finite support and allow  $\tilde{N}$  to take values from a countable set within  $\{1, 2, 3, \dots\}$  with finite mean  $\mathbb{E}(\tilde{N})$ . Our approach is to develop a new sequence of random variables  $\tilde{N}^K$ , in Construction 2, each with finite support and mean  $\mathbb{E}(\tilde{N})$ . We then show that the sequence  $\{\tilde{N}^K\}$  converges to  $\tilde{N}$  in probability, i.e.,  $\tilde{N}^K \xrightarrow{p} \tilde{N}$ .

---

**Construction 2** Let  $\tilde{N} \in \{1, 2, 3, \dots\}$  with  $\mathbb{E}(\tilde{N}) < \infty$ . For each  $K$ , let  $\tilde{N}^K$  be a random variable that takes values only in  $\{1, 2, \dots, K-1, K\} \cup \{\lfloor \mathbb{E}[\tilde{N} | \tilde{N} > K] \rfloor, \lceil \mathbb{E}[\tilde{N} | \tilde{N} > K] \rceil\}$

such that when  $\mathbb{E}[\tilde{N}|\tilde{N} > K]$  is an integer,

$$\begin{cases} \Pr(\tilde{N}^K = n) = \Pr(\tilde{N} = n) \text{ for } n \in \{1, 2, \dots, K-1, K\} \\ \Pr(\tilde{N}^K = \mathbb{E}[\tilde{N}|\tilde{N} > K]) = \Pr(\tilde{N} > K) \end{cases}$$


---


$$\text{Otherwise, } \begin{cases} \Pr(\tilde{N}^K = n) = \Pr(\tilde{N} = n) \text{ for } n \in \{1, 2, \dots, K-1, K\} \\ \Pr(\tilde{N}^K = \lfloor \mathbb{E}[\tilde{N}|\tilde{N} > K] \rfloor) = (\lceil \mathbb{E}[\tilde{N}|\tilde{N} > K] \rceil - \mathbb{E}[\tilde{N}|\tilde{N} > K]) \Pr(\tilde{N} > K) \\ \Pr(\tilde{N}^K = \lceil \mathbb{E}[\tilde{N}|\tilde{N} > K] \rceil) = (\mathbb{E}[\tilde{N}|\tilde{N} > K] - \lfloor \mathbb{E}[\tilde{N}|\tilde{N} > K] \rfloor) \Pr(\tilde{N} > K) \end{cases}$$

Construction 2 replaces the tail of the distribution of  $\tilde{N}$  (the portion where  $\tilde{N} > K$ ), with a single or two finite probability mass points that take on the corresponding conditional mean  $\mathbb{E}[\tilde{N}|\tilde{N} > K]$ . It then can be easily verified that  $\mathbb{E}(\tilde{N}^K) = \mathbb{E}(\tilde{N})$ . Thus Construction 2 provides a mean-preserving transformation, i.e., the consistency of beliefs is preserved.

Let  $\{\tilde{N}_K\}_{K=1,2,3,\dots}$  be built from  $\tilde{N}$  using Construction 2. Note that  $\Pr(\tilde{N} \neq \tilde{N}^K) \leq \Pr(\tilde{N} > K)$  and  $\lim_{K \rightarrow \infty} \Pr(\tilde{N} > K) = 0$ . So  $\{\tilde{N}^K\}$  converges to  $\tilde{N}$  in probability which also implies the convergence in distribution. It immediately follows that  $\lim_{K \rightarrow \infty} R_{\tilde{N}^K} = R_{\tilde{N}}$ ,  $\lim_{K \rightarrow \infty} L_{\tilde{N}^K} = L_{\tilde{N}}$ ,  $\lim_{K \rightarrow \infty} W_{\tilde{N}^K} = W_{\tilde{N}}$ . For each  $K$ ,  $\tilde{N}^K$  is a random variable with mean  $\mathbb{E}(\tilde{N})$  and a finite support. By Theorem 2 (for the finite support case), we have (for each  $K$ ),  $R_{\tilde{N}^K} < R_{\tilde{N}_T}$  for all  $\rho$ ,  $L_{\tilde{N}^K} < L_{\tilde{N}_T}$  and  $W_{\tilde{N}^K} < W_{\tilde{N}_T}$  for small  $\rho$ . By letting  $K \rightarrow \infty$ , we can extend Theorem 2 to the case when  $\tilde{N}$  takes an infinite support.

**Theorem 2''** *Let  $\tilde{N}$  be any balking threshold beliefs distribution that may have an infinite support. Let  $\tilde{N}_T \in \{\lfloor \mathbb{E}(\tilde{N}) \rfloor, \lceil \mathbb{E}(\tilde{N}) \rceil\}$  such that  $\mathbb{E}(\tilde{N}_T) = \mathbb{E}(\tilde{N})$ . Then, (i)  $\lambda_{eff, \tilde{N}} \leq \lambda_{eff, \tilde{N}_T}$  ( $R_{\tilde{N}} \leq R_{\tilde{N}_T}$ ); (ii)  $\exists \rho_L$  s.t.  $L_{\tilde{N}} \leq L_{\tilde{N}_T} \forall \rho \leq \rho_L$ ; and (iii)  $\exists \rho_W$  s.t.  $W_{\tilde{N}} \leq W_{\tilde{N}_T}$  if  $\rho \leq \rho_W$ .*

Thus, we recover the conclusions from Theorems 1 and 2, when applying our results to beliefs that have infinite supports. The lower bounds derived on  $\rho_L$  and  $\rho_W$  in §2.3.2 also

continue to hold.

## 2.4. IMPACT OF REVEALING SERVICE INFORMATION

In the previous section, we compared performance metrics when customers had different balking threshold distributions. In this section, we calibrate the impact of a firm revealing its true service rate  $\mu$ , by comparing revenues and system congestion under two balking threshold distributions corresponding to the firm revealing or not revealing its service information.

When customers are uninformed, they may have arbitrary beliefs about the service rate, and balking thresholds may be distributed according to some  $\tilde{N}$ . When the firm chooses to inform its customers of the true service rate, customers will follow identical balking thresholds. In the  $M/M/1$  queue, the balking threshold beliefs are given by  $\tilde{N} = \lfloor \frac{\tilde{\mu}(v-p)}{c} \rfloor$  and true balking threshold is given by  $N = \lfloor \frac{\mu(v-p)}{c} \rfloor$ . Since the native beliefs could be arbitrary, it is unclear when a firm should reveal its service rate. We examine this question in the following Proposition.

**Proposition 1** *When  $\mathbb{E}(\tilde{N}) \leq N$ , the firm benefits from revealing service information ( $R \uparrow$ ). In addition, when traffic  $\rho$  is small, the average queue length and the average customer waiting time in the system both increase on revelation ( $L, W \uparrow$ ).*

From the Proposition, we find that, when customer balking beliefs are pessimistic or even consistent, i.e.,  $\mathbb{E}(\tilde{N}) \leq N$ , it is always in the firm's interest to reveal its service rate, since the firm sees more revenue as the announcement is made. When  $\rho$  is small, system congestion (average queue length and average customer waiting time) increases on customers knowing the true information.

Note that in Proposition 1 we have stated customer pessimism in terms of balking thresholds rather than internally held service-rate beliefs. This is mainly because, when the service firm investigates whether to reveal service information to its customers, it cannot directly observe

customers' service rate beliefs. Instead, the firm can infer the balking distribution from customers' joining/balking decisions. However, a pessimistic service rate belief distribution usually leads to a pessimistic balking threshold distribution.<sup>3</sup>

Moreover, these results are distribution-free and also parameter-free, i.e., it is sufficient for the firm to only know that the beliefs are pessimistic or consistent (with respect to the balking thresholds), before the decision to reveal true information is made. Under such cases, the exact distribution of beliefs does not influence the decision to reveal information.

To intuit this result, we consider a population with beliefs  $\tilde{N}$ . Let  $\{\tilde{N}_0, \tilde{N}_1, \dots, \tilde{N}_T\}$  be a sequence of balking threshold beliefs from Construction 1 starting with  $\tilde{N}_0 = \tilde{N}$ . Theorem 2 states that  $R \uparrow$  and  $L, W \uparrow$  (for small  $\rho$ ) when customers adopt  $\tilde{N}_T$  instead of  $\tilde{N}$ . Now suppose the beliefs are pessimistic or consistent, i.e.,  $\mathbb{E}(\tilde{N}) \leq N$ , it then follows that  $\tilde{N}_T \leq_{st} N$ . So by Theorem 1, upon revealing  $\mu$ , we have  $R \uparrow$  and  $L, W \uparrow$ . Thus, combining Theorems 1 and 2, when  $\mathbb{E}(\tilde{N}) \leq N$ , we have  $R \uparrow$  and  $L, W \uparrow$  (for small  $\rho$ ).

Now suppose that customers have optimism bias ( $N < \mathbb{E}(\tilde{N})$ ). Again, we construct the sequence  $\{\tilde{N}_0, \tilde{N}_1, \dots, \tilde{N}_T\}$  using Construction 1 starting with  $\tilde{N}_0 = \tilde{N}$ . We have  $R_{\tilde{N}} < R_{\tilde{N}_1} < \dots < R_{\tilde{N}_T}$  by Theorem 2. On the other hand,  $N < \mathbb{E}(\tilde{N})$  implies  $N <_{st} \tilde{N}_T$ , so by Theorem 1 we have  $R_N < R_{\tilde{N}_T}$ . Depending on  $\tilde{N}$ , we may have  $R_N < R_{\tilde{N}}$  or  $R_N > R_{\tilde{N}}$ . Recall that  $\{\tilde{N}_0, \tilde{N}_1, \dots, \tilde{N}_T\}$  is a sequence of beliefs that have progressively lower spreads. So we conclude that when customers population is optimistic, the firm may still reveal its service information as long as it observes high variance in customers' balking behaviors. We provide numerical examples to support this observation in the following section.

#### 2.4.1. Revenue and Welfare Effects of Service Information Revelation under Bias

We now examine specific cases where our findings will apply by studying  $M/M/1$  queues under different beliefs in the population. In all cases, we set the service value  $v = \$8$ , price  $p = \$2$ , and customer linear waiting cost at  $c = \$4/\text{hr}$ . customers are not aware of

---

<sup>3</sup>The exceptions are created when the floor function is involved to transfer  $\tilde{\mu}$  into the corresponding  $\tilde{N}$  shifts the mean threshold by one.



the provider's true service rate,  $\mu$  and their beliefs are uniformly distributed over  $[2, 8]$  per hour, i.e.,  $\tilde{\mu} \sim U[2, 8]$  with mean  $\mathbb{E}(\tilde{\mu}) = 5$ . As a result, customers' balking threshold beliefs,  $\tilde{N} = \lfloor \frac{\tilde{\mu}(v-p)}{c} \rfloor = \lfloor 3\tilde{\mu}/2 \rfloor$ , is a discrete uniform distribution taking values  $\{3, 4, \dots, 11\}$  with mean  $\mathbb{E}(\tilde{N}) = 7$ .

We examine three scenarios where the true threshold  $N$  associated with the true service rate  $\mu$  is (i) greater than, (ii) equal to or (iii) less than  $\mathbb{E}(\tilde{N})$ . These three instances correspond to pessimism, consistency and optimism in beliefs. For each case, we examine the firm's revenue, the average queue length and the average customer waiting time, as well as consumer welfare and social welfare.

The first line in each table that follows corresponds to the situation in which the firm hides the service rate information from its customers (customers adopt balking threshold beliefs  $\tilde{N}$ ); the last line of each table corresponds to the situation in which the firm reveals the service rate information to its customers (customers thus adopt balking threshold beliefs  $N$ ). All rows in between the first and the last rows communicate the terms in the sequence in Construction 1. The percentage change in a parameter (compared to the original beliefs  $\tilde{N}$ , first line) is noted in parenthesis.

**Pessimistic Beliefs:** The arrival rate is  $\lambda = 5/\text{hr}$  and the true service rate is  $\mu = 6/\text{hr}$ . Note that  $\mu = 6 > \mathbb{E}(\tilde{\mu}) = 5$  and  $N = 9 > \mathbb{E}(\tilde{N}) = 7$ . Therefore, customers have pessimistic beliefs.

Beliefs	Firm Revenue	Avg. Queue Length	Avg. Waiting Time	Consumer Welfare	Social Welfare
Uninformed $\tilde{N} = \tilde{N}_0$	8.75	1.86	0.42	18.81	27.56
Construction: $\tilde{N}_1$	8.87 (+1.41%)	1.94 (+4.32%)	0.44 (+2.87%)	18.86 (+0.26%)	27.73 (+0.62%)
$\tilde{N}_2$	9.03 (+3.23%)	2.08 (+11.94%)	0.46 (+8.44%)	18.77 (-0.21%)	27.80 (+0.88%)
$\tilde{N}_3$	9.17 (+4.88%)	2.24 (+20.47%)	0.49 (+14.86%)	18.57 (-1.28%)	27.75 (+0.68%)
$\tilde{N}_4 = \tilde{N}_T$	9.23 (+5.47%)	2.29 (+23.35%)	0.50 (+16.95%)	18.52 (-1.58%)	27.74 (+0.66%)
Informed $N$	9.52 (+8.83%)	2.84 (+52.78%)	0.60 (+40.38%)	17.21 (-8.52%)	26.73 (-3.01%)

In this case, revealing the true service rate increases the firm's revenue by 8.83% but also increases the average queue length by 52.78% and the average waiting time by 40.38%. The firm thus benefits from revealing its service rate information (in line with Proposition 1), but the increased benefit is not sufficient to overcome the loss in consumer welfare (-8.52%).

As a result, overall social welfare drops by 3.01%.

**Consistent Beliefs:** Let  $\lambda = 4/\text{hr}$ ,  $\mu = 5/\text{hr}$ . Note that  $\mu = \mathbb{E}(\tilde{\mu}) = 5$  and  $N = \mathbb{E}(\tilde{N}) = 7$ . Therefore beliefs are consistent in the population. Nevertheless, the individual customer beliefs vary uniformly from 2 to 8.

Belief Type	Firm Revenue	Avg. Queue Length	Avg. Waiting Time	Consumer Welfare	Social Welfare
Uninformed $\tilde{N} = N_0$	7.08 (+0.00%)	1.74 (+0.00%)	0.49 (+0.00%)	14.29 (+0.00%)	21.37 (+0.00%)
Construction: $\tilde{N}_1$	7.18 (+1.43%)	1.82 (+4.37%)	0.51 (+2.89%)	14.29 (+0.00%)	21.47 (+0.47%)
$\tilde{N}_2$	7.31 (+3.24%)	1.95 (+11.89%)	0.53 (+8.37%)	14.15 (-0.97%)	21.46 (+0.43%)
$\tilde{N}_3$	7.43 (+4.86%)	2.09 (+20.19%)	0.56 (+14.62%)	13.92 (-2.61%)	21.34 (-0.13%)
$\tilde{N}_4 = \tilde{N}_T$	7.47 (+5.46%)	2.14 (+23.12%)	0.57 (+16.75%)	13.84 (-3.15%)	21.31 (-0.30%)
Informed $N = N_T$	7.47 (+5.46%)	2.14 (+23.12%)	0.57 (+16.75%)	13.84 (-3.15%)	21.31 (-0.30%)

In the consistent beliefs case, revealing the true service rate still improves revenues (by 5.46%) in line with Proposition 1. On the other hand, the average queue length and the average waiting time both increase significantly (by 23.12% and 16.75% respectively). The firm benefits from revealing the service rate, almost fully at the expense of consumer welfare (-3.15%), but the overall social welfare does not fall significantly (-0.30%) due to the increase in throughput (i.e., number of customers served).

**Optimistic Beliefs:** Let  $\lambda = 3/\text{hr}$  and  $\mu = 4/\text{hr}$ . Note  $\mu = 4 < \mathbb{E}(\tilde{\mu}) = 5$  and  $N = 6 < \mathbb{E}(\tilde{N}) = 7$ . Hence, population beliefs are optimistic.

Beliefs	Firm Revenue	Avg. Queue Length	Avg. Waiting Time	Consumer Welfare	Social Welfare
Uninformed $\tilde{N} = N_0$	5.40 (+0.00%)	1.57 (+0.00%)	0.58 (+0.00%)	9.93 (+0.00%)	15.33 (+0.00%)
Construction: $\tilde{N}_1$	5.48 (+1.45%)	1.64 (+4.39%)	0.60 (+2.90%)	9.89 (-0.41%)	15.37 (+0.25%)
$\tilde{N}_2$	5.58 (+3.22%)	1.75 (+11.65%)	0.63 (+8.17%)	9.72 (-2.11%)	15.30 (-0.23%)
$\tilde{N}_3$	5.66 (+4.76%)	1.88 (+19.48%)	0.66 (+14.04%)	9.48 (-4.54%)	15.14 (-1.26%)
$\tilde{N}_4 = \tilde{N}_T$	5.69 (+5.34%)	1.92 (+22.37%)	0.68 (+16.16%)	9.39 (-5.43%)	15.08 (-1.64%)
Informed $N$	5.57 (+3.03%)	1.70 (+8.31%)	0.61 (+5.13%)	9.90 (-0.31%)	15.46 (-0.86%)

Although customers are optimistic about the service rate, revealing the true service rate would still increase firm's revenue by 3.03%. Examining the second column of the table (firm's revenue column) reveals what we have discussed for the optimism bias case: When customers' optimistic balking threshold beliefs are more dispersed (as in the example the original belief  $\tilde{N}$ ), it is beneficial for the firm to reveal service rate.

In this example, we see that, if customers' balking beliefs is characterized by  $\tilde{N}$  or  $\tilde{N}_1$ , then the firm increases its revenue by revealing the true service rate. As in the previous cases, the revenue accrual comes at the expense of increased queue lengths and waiting times for

customers. On the contrary, if customer beliefs are less dispersed, (for e.g., if the beliefs were  $\tilde{N}_3$ ), the firm does not gain from revealing its service rate. In this case, customers are better off in both expected queue lengths and wait times.

To summarize, while the revenues improve with more information, the welfare effects are mixed. Typically the firm benefits from revealing service information, to the detriment of welfare. The gains in revenues are usually, but not always, lower than the loss in consumer welfare. Thus typically, social welfare is reduced, as a consequence of more information in the system. However, it is also possible that both the firm revenues and consumer welfare improve upon service information revelation (even with consistent beliefs). This can occur when the traffic is very high and customers' prior beliefs are almost deterministic. One such example is given by Case (iii) of the numerical illustration in §2.3.2 where  $\lambda_{eff} \uparrow$  and  $L \downarrow$ .

## 2.5. APPLYING OUR FINDINGS TO SPECIFIC BELIEF MODELS

While our results hold for any general belief structure, it is helpful to evaluate what our findings imply under some specific belief considerations that have been examined in the literature. In this context, it is germane to consider the following issue: If consumers arrive to a queue endowed with some pre-existing beliefs, how do these different beliefs form? In the following section, we consider some behavioral/operational antecedents to belief structures, show our analysis apply to those cases and derive conclusions from those applications.

### 2.5.1. *Quantal Response Errors*

Quantal response models are used to model deviations from optimal consumer decisions in the absence of full information. For instance, in queues, consumers may make “errors” in their estimate of the true service rate due to cognitive limitations following Quantal Choice Theory (Luce, 1959), which argues that decision makers do not always choose the “correct” alternative, but better alternatives (i.e., alternatives with smaller errors) are chosen with a higher probability than the alternatives that are worse. Quantal choice approach has

been employed to model bounded rationality in the newsvendor contexts by Su (2008), and subsequently in queueing settings by Huang et al. (2013).

If consumer population made i.i.d. belief draws from a distribution that align with Quantal Choice Theory, a large fraction of the population will have small errors in their beliefs about the true service rate, and a diminished fraction of customers make arbitrarily large errors in their beliefs. Furthermore, the mode of such a belief distribution will coincide with the true service parameter.

For a consumer  $j$ , let the belief on the true service rate ( $\mu$ ) be  $\tilde{\mu}_j$ . We use  $[|\tilde{\mu}_j - \mu| + 1]^{-1} \in (0, 1]$  to indicate the accuracy of her belief.<sup>4</sup> Assuming i.i.d. customers, we could use a logit model – the mostly commonly used Quantal response distribution – for  $\tilde{\mu}$ , to model the accuracy of consumer beliefs. Then, the pdf for the belief distribution  $\tilde{\mu}$  is given by:

$$f_{\tilde{\mu}}(x) = \frac{\exp\{[\beta(|x - \mu| + 1)]^{-1}\}}{\int_{x=\underline{\mu}}^{x=\bar{\mu}} \exp\{[\beta(|x - \mu| + 1)]^{-1}\} dx}$$

where  $\beta$  is a cognitive parameter that measures “distance” from perfect rationality. As  $\beta \rightarrow \infty$ ,  $\tilde{\mu} \sim U[\underline{\mu}, \bar{\mu}]$  (consumers are totally uninformed and make ‘random’ errors), and when  $\beta \rightarrow 0$ ,  $Pr(\tilde{\mu} = \mu) = 1$  (consumers are fully informed and we recover Naor’s model in this context). When the belief distribution is symmetric, i.e.,  $\mu = (\underline{\mu} + \bar{\mu})/2$ , we note that as  $\beta$  decreases from  $\infty$  to 0, the underlying consumer beliefs undergo the transformation described in Construction 1.

Quantal response choices only require that the zero-error choice is chosen with the highest probability, and therefore are *not necessarily* consistent beliefs. For example, consumer beliefs can be optimistic (if  $\mu < (\underline{\mu} + \bar{\mu})/2$ ) or pessimistic (if  $\mu > (\underline{\mu} + \bar{\mu})/2$ ). Quantal response beliefs are examples of beliefs where the results of our paper apply. Although Quantal response beliefs can explain some deviations from the optimal/true choice, they do not inform how these beliefs form. We examine some specific causes (e.g., past experiences)

---

<sup>4</sup>Other measures of accuracy could be employed without altering our conclusions.

in the following sections.

### 2.5.2. Learning by Sampling Past Experiences

In many service instances, consumers have limited and infrequent interactions with the service provider. In such cases, consumers could use their past service experience as samples to learn more about the service rate. This sampling helps consumers to arrive at their beliefs and eventually make their decisions. Suppose that all consumers in a sufficiently large population, use only their past service experience to estimate the service rate. Specifically, let us examine a case in which all consumers have visited the server  $s$  times, ( $s \geq 1$ ) or only remember the past  $s$  service time experiences. We assume that consumers are homogeneous in  $s$  in this section, but relax the assumption in §2.5.3.

Consider a consumer with the following service time samples  $\{\tau_1, \tau_2, \dots, \tau_s\}$ . A rational consumer who knows the service distribution, but not the exact parameters, will use the observed samples to arrive at an estimate that maximizes the likelihood of observing those  $s$  samples. Simply, a rational consumer would use Maximum Likelihood Estimator (MLE) for calculating the parameters of the service distribution. Suppose the service times are i.i.d. exponential, then it is well known that the MLE for the service rate is given by

$$\hat{\mu}(\tau_1, \tau_2, \dots, \tau_s) = s \left/ \sum_{i=1}^s \tau_i \right. . \quad (2.15)$$

This is the point estimate for the service rate for the consumer with samples  $\{\tau_1, \tau_2, \dots, \tau_s\}$ . Thus, consumers will have different beliefs (estimates) based on their individual samples. Define  $\tilde{\mu}_s$  to be the random variable associated with the belief distribution, when all consumers use  $s$  samples to arrive at their beliefs through MLE. We note that  $\sum_{i=1}^s \tau_i$  in equation (2.15) has an Erlang distribution with shape parameter  $s$  and rate parameter  $\mu$ . It follows that over the population, the individual consumer beliefs  $\tilde{\mu}_s$  are distributed Inverse-Gamma with shape parameter  $s$  and scale parameter  $s\mu$ , i.e.,  $\tilde{\mu}_s \sim \text{Inv-Gamma}(s, s\mu)$ . The pdf for

$\tilde{\mu}_s$  is given by

$$f_{\tilde{\mu}_s}(x) = \frac{(s\mu)^s}{\Gamma(s)} x^{-s-1} \exp\left(-\frac{s\mu}{x}\right) \text{ for } x > 0 \quad (2.16)$$

where  $\Gamma$  denotes the upper incomplete gamma function. Furthermore, if the service times are independent, the estimate remains unchanged if these samples are collected on a single visit or over multiple visits.

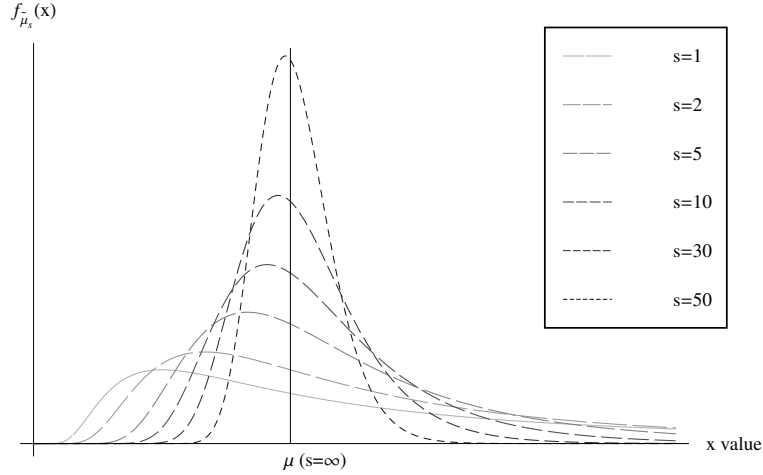


Figure 1: As consumers remember more service experiences ( $s \uparrow \infty$ ), their estimates of the service rate become consistent with the true service rate. That is, (i)  $\mathbb{E}(\tilde{\mu}_s) \downarrow \mu$ , and (ii)  $\text{Var}(\tilde{\mu}_s) \downarrow 0$ .

In Figure 1, we illustrate  $\tilde{\mu}_s$  for different sampling sizes  $s$ . Since  $\mathbb{E}(\tilde{\mu}_s) = \frac{s}{s-1}\mu$  and  $\text{Var}(\tilde{\mu}_s) = \frac{s^2}{(s-1)^2(s-2)}\mu$ , we note that for any finite  $s$ , the population mean is higher than the true  $\mu$ . Thus the population is optimistically biased. Further, as  $s$  increases, consumer beliefs get less noisy (i.e.,  $\text{Var}(\tilde{\mu}_s) \rightarrow 0$  as  $s \uparrow \infty$ ), as reflected in the distributions getting less spread-out in Figure 1. Eventually, as the number of samples approaches infinity in (2.16), the distribution of  $\tilde{\mu}_s$  approaches a one-point distribution at  $\mu$ , i.e.,  $Pr(\tilde{\mu} = \mu) = 1$ . As consumers learn service rate through sampling, they remain optimistic but diminishingly so, as they collect more samples. Thus MLE derived through sampling is *biased* but *asymptotically consistent*.

Finally, for any  $s$ , the mode of the belief distribution  $\tilde{\mu}_s$  does not coincide with the true

service rate  $\mu$  ( $s = \infty$  line). Therefore, the beliefs that emerge from learning through sampling, are not Quantal choice beliefs. Hence, sampling distributions are another distinct example of our belief distribution results.

**Learning through Waiting Times.** Note that for a given  $s$ , as long as the service times are i.i.d., it does not matter whether a consumer’s MLE is built from her own service times, or from her observations of service times for other consumers. In some cases, consumers may not be able to observe all other service times, due to limited cognitive attention to the sequence of events, or due to system environment. As an example, consider ticket queues (see Xu et al. (2007)) where a consumer learns only her own wait time and ticket number. Another similar setting occurs at emergency room queues, where consumers learn their own total wait time but not the exact service times of everyone in the queue.

Suppose that a consumer only knows her total wait time  $W_s$  and the number of customers during the wait,  $s$  (for e.g., queue length), but does not observe the individual service times  $(\tau_1, \tau_2, \dots, \tau_s)$  during the wait. Her service time belief is based on the observed waiting time, i.e.,  $\hat{\tau}|W_s = W_s/s$  and the corresponding service rate belief is  $\hat{\mu}|W_s = [\hat{\tau}|W_s]^{-1} = s/W_s$ . This estimate  $\hat{\mu}|W_s$  is identical to the MLE in equation (2.15) for a consumer who observed individual service times which sum up to  $W_s$ . The overall belief distribution using the waiting time estimate  $W_s$  is thus identical to the MLE case with  $s$  samples.

### 2.5.3. Consumer Heterogeneity in Learning

So far, we limited consumer sampling to be homogeneous across the entire population. However, it is possible that consumers differ in  $s$ . Typically, there is some underlying distribution of  $s$  in the population, based on how consumers accumulate information or are exposed to it. For instance, consumers may consider external reviews or auxiliary information from other consumers. We denote this sampling heterogeneity by a discrete random variable  $S$  taking values in  $\{1, 2, \dots\}$  with pdf  $f_S$ . When  $S$  is a one-point distribution, we retain homogeneity in learning. With (2.16) in hand, we can write the continuous *unconditional* distribution of

beliefs denoted by  $\tilde{\mu}$  in the population, through the following pdf:

$$f_{\tilde{\mu}}(x) = \sum_{s \in S} \frac{(s\mu)^s}{\Gamma(s)} x^{-s-1} \exp(-\frac{s\mu}{x}) f_S(s) \text{ for } x > 0. \quad (2.17)$$

We now consider two different ways to model the learning heterogeneity,  $f_S(s)$ .

**Sampling through Poisson Arrivals.** One natural distribution of sampling in the population could be based on what consumers observe when they arrive to a queue. Applying PASTA property (Wolff, 1982), when consumers arrive according to a Poisson process to the server, their sampling distribution follows the steady state distribution of the queuing system. In an  $M/M/1$  queue, a customer who arrives at the state  $s - 1$  observes  $s$  samples in total. With traffic intensity  $\rho = \lambda/\mu$ , the sampling distribution would be  $f_S(s) = (1 - \rho)\rho^{s-1}$ ,  $s \geq 1$ . Clearly, if the queue is a system with limited buffer size, then the sampling distribution is on finite support, and therefore the population becomes more optimistic, following our previous discussion.

**Limited Recall/External Reviews.** Consumers may come across information (for e.g., reviews on bulletin boards or websites), but may not recall all observed information due to cognitive limitations. When subjected to an increasing amount of information, a consumer may remember only a limited amount of information, and forget an amount that is proportional to the information she is exposed to. We examine such a sampling system, using the model of limited memories due to Nelson (1974). We assume that consumers confront reviews according to a Poisson Process with rate  $r$  and they also forget a review at a rate that is directly proportional to the number of reviews a person remembers. Let  $a$  be the constant of the proportionality. We denote the state  $n \in \{0, 1, 2, \dots\}$  as the number of reviews that a particular customer remembers in the long run. The fraction of customers who remember  $n$  reviews is given by steady-state probability distribution  $\{\pi_n : n = 0, 1, 2, \dots\}$ . We then have  $\pi_n = (r/a)^n e^{-r/a} / n!$  for  $n \in \{0, 1, 2, \dots\}$ .

Suppose that consumers form their beliefs only after reading at least one review or service



experience. Then, the fraction of consumers who remember  $s$  reviews among the population is given by

$$f_S(s) = \frac{\pi_s}{1 - \pi_0} = \frac{(r/a)^n}{n!} \frac{e^{-r/a}}{1 - e^{-r/a}} \text{ for } s = 1, 2, \dots$$

The fraction  $f_S(s)$  use  $s$  reviews to arrive at their belief  $\tilde{\mu}_s$ . We can then use  $f_S$  to build the distribution of consumer beliefs in the queue through equation (2.17).

Regardless of how we model the sampling heterogeneity  $f_S$ , we conclude that consumer population will tend to be optimistic, even though all reviews/service experiences are assumed to be accurate. This is because, using MLE, each consumer class is optimistic. Thus, learning through sampling can lead to optimism bias even in heterogenous populations. We briefly examine how pessimistic beliefs can emerge.

From a behavioral point of view, consumers may have *availability bias* (Tversky and Kahneman, 1973) when processing information. Under availability bias, consumers remember *unusual* experiences saliently, in forming their beliefs. In the case of exponential servers, when consumers have finite sample sizes, short service times are much likely to be present in their sample than long service times. Thus, consumers recall longer-than-usual service times more vividly. As a result, availability bias could lead to tempered optimism or even pessimism in the population beliefs, i.e.,  $\mathbb{E}(\tilde{\mu}) < \mu$ . Another cause of pessimistic bias, can be due to Prospect Theory (Kahneman and Tversky, 1979), where longer (worse) service times affect updating more significantly. See Gaur and Park (2007) for such asymmetric consumer learning in inventory context.

## 2.6. UNOBSERVABLE QUEUES

In this section, we analyze settings where consumers cannot observe the queue lengths upon arrival. Recall that in the observable model, the customers arrive with some service rate belief ( $\tilde{\mu}$ ). When arriving customers observe the state of the queue  $n$ , they immediately form beliefs about their expected waiting time  $(n+1)/\tilde{\mu}$ , from which they can make joining or balking decisions. Each customer's decision is *fully* determined by his belief over the

service rate and by the queue length he sees. In other words, his decision does not depend in any way on other customers' decisions, i.e., there is no gaming or coordination among customers.

However, in unobservable settings, a customer is not only unaware of the queue length information, but also has no additional information about the system to make his decision. His decisions (and hence the impact on revenues and welfare) depend on the nature of beliefs he arrives with.

If the customer arrives with some beliefs on the expected wait time (or, if his expected wait time is revealed to him by the firm on his arrival), he can make his best decision without requiring any additional information regarding others' decisions (just as he would in the observable setting). We analyze the effect of such beliefs in §2.6.1.

On the other hand, if he arrives only with service rate beliefs, he cannot make join/balk decisions without additional information or beliefs. In the observable setting, seeing the queue length on arrival provides the necessary additional information to the customer to make his decision. Absent such additional information, he would need to convert his belief on service rate into an estimate on the expected waiting time. This estimation requires guessing/infering information about other customers' decisions or beliefs, because their decisions affect his expected wait time and consequently his optimal decision. Such estimation involves gaming and equilibrating, which we examine in §2.6.2.

### *2.6.1. Unobservable Settings with Waiting Time Beliefs*

We examine the same single-server system analyzed earlier in the paper, except that queue lengths are now not visible. Consider a customer who arrives with belief  $\tilde{w}$  about the expected waiting time in system to complete his service. The belief  $\tilde{w}$  (perhaps based on his past experience, etc.) can be arbitrarily different from the true expected waiting time.

Absent other information, he makes the following rational decision:

$$\begin{cases} v - p \geq c\tilde{w} : & \text{The customer joins the queue;} \\ \text{otherwise:} & \text{The customer balks from the queue.} \end{cases}$$

As in the observable model,  $v - p$  here represents the value of service net of price, and  $c$  denotes the cost of waiting per unit of time.

Let us denote  $\tilde{W}$ , with cdf  $F_{\tilde{W}}$ , the distribution of the population beliefs on the expected waiting time. The beliefs that are deterministic for each customer form a random distribution because customers arrive to the system randomly. As before, we do not restrict the belief distribution. Specifically, the no-information case in Guo and Zipkin (2007) in which all customers arrive with the same deterministic belief on the waiting time (equaling the long-run expected waiting time) is a special case of our analysis.

The effective joining rate under  $\tilde{W}$  is given by

$$\lambda_{eff,\tilde{W}} = \lambda \Pr\{v - p \geq c\tilde{W}\} = \lambda F_{\tilde{W}}\left(\frac{v - p}{c}\right), \quad (2.18)$$

and we can write waiting time metrics in terms of  $\lambda_{eff,\tilde{W}}$ . The average customer waiting time is  $1/(\mu - \lambda_{eff,\tilde{W}})$ ; and the average congestion in the system is  $\lambda_{eff,\tilde{W}}/(\mu - \lambda_{eff,\tilde{W}})$ . Note that the latter is not observed by customers but would be observed by the firm.

We now ask the central question of the paper about revealing information. Note in the observable setting, when the true service rate is revealed, the firm is effectively giving out the expected waiting time information for each customer based on the queue-length they arrive at. The corresponding question in this model is as follows: When would a firm reveal real-time expected waiting time information to its customers on their arrival? For example, in practice, when calling the customer services line of Macy's or AT&T, or using the live-agent kiosks at Hertz airport locations, information such as "Your current expected waiting time is 13 minutes" is revealed.

The effect of revealing wait-times here is similar to the effect of revealing the service rate in the observable setting. On one hand, some customers who have overestimated the expected waiting time will join the queue after hearing the server's announcement. On the other hand, some customers that had underestimated the waiting time might balk. Therefore, the firm's decision to reveal depends on customers' belief distribution  $\tilde{W}$ .

Specifically, when the firm announces the real-time expected waiting time information to each arrival who then joins or balks, the underlying unobservable queue coincides with an *observable*  $M/M/1/N$  system where every consumer knows the true service rate  $\mu$ . (However, customers do not observe the state of the system or  $\mu$ ). The effective joining rate is given by Naor (1969) as:

$$\lambda(1 - \rho^N)/(1 - \rho^{N+1}) \tag{2.19}$$

In this case the average congestion in the system is given by

$$\frac{\rho}{1 - \rho} - \frac{(N + 1)\rho^{N+1}}{1 - \rho^{N+1}},$$

and the average customer waiting time follows immediately from Little's Law.

From (2.18) and (2.19), we find that it is better for the firm to reveal the waiting time information if

$$F_{\tilde{W}}\left(\frac{v - p}{c}\right) < \frac{1 - \rho^N}{1 - \rho^{N+1}} \tag{2.20}$$

but not otherwise. The LHS of the condition (2.20) is fixed for a population with a belief distribution. For the population arriving at rate  $\lambda$ , it can be shown that RHS is increasing in  $\mu$ . As a result, a firm with a faster service rate is more likely to reveal its waiting time information.

However, when a fast server (i.e., small  $\rho$ ) reveals waiting time information, system congestion and the resulting average waiting time also increases. In fact, just as in the observable queue settings, we find that there exist some threshold  $\rho_L$  (and  $\rho_W$  respectively) such

that when  $\rho \leq \rho_L$  ( $\rho_W$ ), the system congestion (the average waiting time) will increase upon waiting time information revelation. The impact on overall consumer welfare is typically negative due to increased congestion, despite the increase in the number of customers served. Hence, social welfare can fall on the provision of information. A similar conclusion is reached by Plambeck and Wang (2012) when customers make time-inconsistent decisions (hyperbolic discounting).

### 2.6.2. Unobservable Settings with Service Rate Beliefs

We now assume that the customers arriving to an unobservable single-server queue have beliefs about the service rate instead. Thus, similar to the model under the observable setting, customers do not know the true service rate  $\mu$  unless they are informed by the service firm. The key technical issue in this case is that a customer has to make a join/balk decision not only based on his own service-rate belief, but also on what he *believes* about other customers' service-rate beliefs. These decisions require each customer to evaluate the best response to others' strategies.

Prior research has modeled such scenarios by imposing that all customers know the true service rate  $\mu$ . For an extensive collection of such papers, see Chapter 3 of Hassin and Haviv (2003). We extend this research stream to scenarios where the service rate is unknown.

Assume a customer  $j$  believes that the service time is exponentially distributed with some service rate  $\tilde{\mu}_j$ . (An exponential distribution is assumed for analytical simplicity and can be extended to a general distribution.) We use random variable  $\tilde{\mu}$  with cdf  $G_{\tilde{\mu}}$  to denote the heterogeneity in service-rate beliefs, as in the observable model. Customer  $j$  (with belief  $\tilde{\mu}_j$ ) in making his decision, thinks that all the other customers (should) have the same service-rate belief, namely  $\tilde{\mu}_j$ .<sup>5</sup> Hence, the distribution of each customer's beliefs over others' deterministic beliefs follows the same distribution as the native distribution  $G_{\tilde{\mu}}$ .

As is done in the classical unobservable-queue models, customers know the arrival rate  $\lambda$ .

---

<sup>5</sup>We could impose other characterizations on beliefs over other consumers' beliefs, but this characterization appears natural and does not impose additional informational requirements.

This allows us to focus on the heterogeneity in beliefs about the service rate alone. Clearly, every customer's decision to join depends on his beliefs regarding others' beliefs. Due to our assumption, a customer with service-rate belief  $\tilde{\mu}$  makes the join/balk decision as if the unobservable queue were  $M/M/1$  system with arrival rate  $\lambda$  and known service rate  $\tilde{\mu}$ . Following Hassin and Haviv (2003), a customer with belief  $\tilde{\mu}$  joins the queue with some probability  $q(\tilde{\mu})$  and balks with probability  $1 - q(\tilde{\mu})$  in equilibrium where

$$\begin{cases} q(\tilde{\mu}) = \frac{\tilde{\mu} - \frac{c}{v-p}}{\lambda} & \text{if } (\frac{c}{v-p} \leq) \tilde{\mu} < \lambda + \frac{c}{v-p} \\ q(\tilde{\mu}) = 1 & \text{if } \tilde{\mu} \geq \lambda + \frac{c}{v-p}. \end{cases} \quad (2.21)$$

To explore the issue of revealing service-rate information, we note that, when customers are fully informed of the true service rate  $\mu$ , all customers will have the identical joining strategy in equilibrium:

$$\begin{cases} q(\mu) = \frac{\mu - \frac{c}{v-p}}{\lambda} & \text{if } (\frac{c}{v-p} \leq) \mu < \lambda + \frac{c}{v-p} \\ q(\mu) = 1 & \text{if } \mu \geq \lambda + \frac{c}{v-p} \end{cases} \quad (2.22)$$

We denote  $\lambda_{eff}^U$  and  $W^U$  the long-run (effective) joining rate and each customer's true expected waiting time in the system. (The superscript  $U$  denotes the unobservable setting.) The revenue (rate) at the server is given by the product of price charged by the server and the long-run effective arrival rate at the system:  $R^U \triangleq p \cdot \lambda_{eff}^U$ .

**Theorem 1<sup>U</sup>** *If  $\tilde{\mu} \leq_{st} \tilde{\mu}'$ , then (i)  $\lambda_{eff, \tilde{\mu}}^U \leq \lambda_{eff, \tilde{\mu}'}^U$  ( $R_{\tilde{\mu}}^U \leq R_{\tilde{\mu}'}^U$ ), and (ii)  $W_{\tilde{\mu}}^U \leq W_{\tilde{\mu}'}^U$ .*

Theorem 1<sup>U</sup> states that, when the service-rate beliefs are more optimistic, more customers join the queue. More joining customers lead to increased firm revenue as well as higher waiting times. This result is consistent with the first-order result in the observable setting (i.e., Theorem 1).

Next, let  $\lambda_{eff, \mathbb{E}(\tilde{\mu})}^U$ ,  $R_{\mathbb{E}(\tilde{\mu})}^U$ ,  $W_{\mathbb{E}(\tilde{\mu})}^U$  denote the metrics of interest in an unobservable queue

when the customers' service rate beliefs are identical to the mean of the distribution, i.e.,  $\mathbb{E}(\tilde{\mu})$ . This comparison extends our results on mean-preserving spread under the observable queue (i.e., Theorem 2) to the unobservable case.

**Theorem 2<sup>U</sup>** *For any  $\tilde{\mu}$ , (i)  $\lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\mathbb{E}(\tilde{\mu})}^U$  ( $R_{\tilde{\mu}}^U \leq R_{\mathbb{E}(\tilde{\mu})}^U$ ), and (ii)  $W_{\tilde{\mu}}^U \leq W_{\mathbb{E}(\tilde{\mu})}^U$ .*

Theorem 2<sup>U</sup> states that, when the distribution of beliefs in the customer population is less spread out, the firm receives more profit. Also note that average waiting time in the system increases regardless of the traffic level  $\rho$ , which differs from Theorem 2 (of the observable setting). This is due to the fact that, unlike in the observable setting, the effective joining rate in the unobservable setting is independent of the underlying queue-lengths.

We are now ready to state the Proposition that compares performance measures under the native customer beliefs to the case in which the true service rate is revealed.

**Proposition 1<sup>U</sup>** *If  $\mathbb{E}(\tilde{\mu}) \leq \mu$ , then the revelation of the service rate information increases both throughput and revenues. However, service rate revelation also increases the average waiting time.*

As in the observable setting, if customers have pessimistic or consistent beliefs, the revelation of the service-rate information increases the revenues of the firm. However, system congestion - which remains unobserved by the customers, but visible to the firm - also increases. For the same reasons discussed in the observable case and in §2.6.1, consumer welfare and social welfare may increase or decrease with the revelation of service-rate information.

## 2.7. CONCLUSIONS AND IMPLICATIONS

Customers often join queues with very limited information. Most of the literature has assumed that service parameters (typically,  $\mu$ ) that influence the joining behavior as common knowledge. However, customers cannot always fully characterize these service parameters; sometimes even the calculation of mean service time may require repeated sampling or collection of data. In such blind queues, not much is known on how revenues and welfare are

impacted, when a firm reveals its service information (specifically, service capacity). Our paper seeks to fill this gap.

Our approach to solving the information problem is distribution-free. We begin with *any* belief distribution that customers may have on the service rate in the uncountable space, and reduce it to balking threshold beliefs in the countable space. Using our general but intuitive approach, we calibrate the impact of information revelation on the performance of the queueing system, without any restrictions on the distribution of the initial customer beliefs. We can apply the results from our general model on specific belief structures, such as Quantal-response based bounded rationality (Luce (1959), Su (2008), Huang et al. (2013)), learning through sampling either from past experiences (Xu et al., 2007), anecdotal reasoning (Chen and Huang, 2013) and other cognitive biases to characterize their effects on revenues and consumer welfare.

We find that if beliefs are pessimistic or consistent, the service provider always improve revenues by revealing its service parameters. However, the impact of service rate information on congestion and welfare is mixed. Even though a firm's revenues improve on announcing its service rate, the impact on the congestion levels (such as the average queue lengths, or average wait times) are typically negative. As a result, consumer welfare worsens with more information, despite the increased market coverage. In fact, consumer welfare loss can exceed revenue improvements at the firm. As a result, the social welfare can fall with provision of correct information.

Hence, intriguingly, with informational uncertainty, the social welfare typically improves, regardless of the visibility of the actual queue, compared to the case in which customers have full information. When left to their own devices, with full information, more customers join the queue than what is socially optimal. In Naor (1969), tolls/taxes are levied to control the joining population which improves welfare. Likewise, we find that the lack of information acts as an *information tax* that deters admission, which could lead to improved welfare.



The impact of additional information changes with the service capacity, given a market size. Consumer welfare likely worsens in the case when a fast server reveals its service rate, compared to the case when a slower server reveals its rate. Thus, customers are worse off if a *faster* firm with more service capacity reveals its service information. This result is not due to observability of the queues: Even if waiting times were announced in unobservable queues, the welfare loss is likely to occur when the server is fast.

Several future research directions related to queue information appear promising. One question is the timing of announcement decisions, and the type of announcements that could be made. For instance, Allon et al. (2013) studies the impact on announcing (dynamic) delay information on customer joining behavior. There are further informational considerations that are rich for exploration. It is possible that the firm can strategically shade or alter the information it provides, e.g., see Ang et al. (2014a). On the other hand, customers may also dynamically update their beliefs as they wait in the queue, and abandon the queue.

Finally, our findings are under the limitation that  $\lambda < \mu$ . The question on how true revealing information to misinformed population beliefs, in overloaded service systems is open. We believe that much of our results of the first-order stochastics continue to hold, but it appears that sequencing beliefs under higher stochastic orders is more complex.

Our findings have several implications for queue management policies in practice. In primary healthcare settings where the access to service providers is important, revealing the capacity information can lead to an increased customer access to the queue (i.e., more customers will visit the service provider). Nevertheless, customers will observe longer queues on average, and also suffer a higher disutility in waiting time on average.

Thus when there is a significant impetus on treating admitted patients quickly, as in some emergency room settings, revealing the service information may lead to increased crowding and worsen the average wait times. Furthermore, this effect is exacerbated for a facility that has ample service capacity. So, the decision to reveal the service information in aforemen-

tioned settings depends critically on the tradeoffs between improved access and increased congestion.

## CHAPTER 3 : A MODEL OF RATIONAL RETRIALS IN QUEUES

### 3.1. INTRODUCTION

In many service settings such as post offices or ATM machines, consumers usually have to join a queue before they can obtain the service. It is well-known that consumers do not enjoy waiting in queues. In real life, when the queues are long, consumers may not be willing to wait, rather they choose to retry later (as opposed to balking). For instance, consider a customer who arrives at the package pick-up service at a post office. Upon seeing the state of the queue, i.e., the number of consumers that are already in the queue, this customer can either decide to join the queue or to leave only to return at a later more convenient time. However, the existing queueing literature on modeling strategic consumer decisions has focused on join and balk decisions – with customers not making retry decisions.

In the package pick-up example described above and in many other real-life queues such as lottery kiosks and DMV centers, retry decisions are common practices by consumers. In this paper, we build a model to allow the consumers to retry for the service in the future, with some retrial cost, upon seeing the status of the queue. This retrial hassle cost can be associated with the additional transportation back and forth, the disutility of receiving the service at a future date, the hassle of re-planning and rescheduling, or visit fees for entering the system, etc.

Given the wide prevalence of scheduling tasks, it is important to understand how consumers make join and balk, versus retry decisions. Since the amount of retrial hassle can be different under various circumstances, it is necessary to investigate consumers' decision-making as a function of the retrial hassle cost. From a consumer-welfare point of view, a customer who decides to come back to the system later will generate externalities to other consumers. However, it is not clear if the net effect of the overall externalities is positive or negative.

To the best of our knowledge, strategic consumer *retry decisions* in service operations set-

tings has remained almost unaddressed, even though such actions have been acknowledged in both popular press and academic literature. In the paper, we characterize consumer decisions among options to join, balk or retry, and study the impact of retrials on welfare. Our main findings include: (i) With the additional option to retry, consumer welfare can be however worse at individual equilibrium; (ii) More surprisingly, compared to the socially optimal outcome, self-interested consumers do not retry enough when the retrial hassle cost is small. And they retry too much when the retrial hassle cost is high.

We acknowledge that in some service settings, a portion of the consumers cannot afford to delay their workload or balk. Examples include patients arriving at a hospital with critical conditions, or drivers at a gas station whose car has run out of gas. We term these emergency arrivals the “myopic” consumers, as opposed to others who strategically choose to join, balk or retry. We then study the impact of the myopics on strategic consumers’ decision-makings and welfare.

Finally, we note that although our model is based on an observable setting, the results we find can be implemented to unobservable queue settings where queue or wait-time information is provided. For example, using our model we can study consumer behaviors to reschedule when they call customer services phone lines and hear information such as “Your approximate waiting time is 10 minutes”, or “You are currently the 8<sup>th</sup> customer waiting in the line”.

The remainder of this paper is structured as follows. We conclude this section with a review of related literature. Section 2 describes the model. Section 3 investigates strategic consumers’ behaviors as a function of the retrial hassle cost. Section 4 studies welfare, and then compares individual equilibrium to the socially optimal outcome. Section 5 exploits the impact of myopic consumers. Section 6 gives concluding remarks. All technical proofs are deferred to the appendix.

### 3.1.1. *Related Literature*

Mandelbaum and Yechiali (1983) considers a single strategic customer who can join the queue, balk or wait outside the queue to retry, upon seeing the current state of the system, while all other customers join the queue unconditionally. In this paper, we allow for *all* consumers to have the option to retry, therefore the future arrivals to the system are endogenous determined by consumer decisions, and each consumer's decision could also depend on the decisions of other people.

While Kulkarni (1983a,b), Elcan (1994) & Hassin and Haviv (1996) have studied socially optimal and equilibrium retrial frequency decisions of consumers who are forced to retry upon seeing a busy system, our paper reverses the focus to examine consumers' strategic retry versus join and balk decisions upon arrival at a busy server, and does not study the retrial frequency decisions.

Besides Mandelbaum and Yechiali (1983), there seems to be no other previous research on the top of modeling strategic consumer retry decisions, which our paper does. However, two well-established research streams are relevant. These are papers on "modeling strategic consumer decision-making without retry decisions" under the service operations literature, and papers on "retrial queues with non-decision-making consumers" under the network and call center literature.

The first literature on queueing models with strategic consumers dates back to the seminal work by Naor (1969), who studies a single-server system with an observable queue. In Naor's model, homogeneous customers observe the queue length upon arrival before making a decision to join the system or to balk. Our paper directly extends Naor's model by allowing for customers to have the third option to retry.

Following Naor (1969), strategic consumer behaviors have been studied in the context of heterogeneous service values (e.g., Larsen (1998), Miller and Buckman (1987)), time costs (e.g., Afèche and Mendelson (2004)) and many others (e.g., see the survey in Mendelson and

Whang (1990) and the comprehensive review by Hassin and Haviv (2003)). Nevertheless, in all these papers above strategic consumers only make state-dependent join or balk, but not retry decisions.

Second, retrial queues have been employed in network models with non-decision-making consumers, i.e., consumers or other objects are specified by the system on when and how likely they should retry in order to optimize the system efficiency. In contrast, consumers make their own strategic retry decisions in our model. The network models are also called the *orbital models*, as the consumers waiting to try again are said to be in orbit. Because of the intractable nature of the orbital models, analytic results are generally difficult to obtain. Hence, there has been a significant focus on numerical and approximation methods (e.g., Reed and Yechiali (2013)). As a special case, Kulkarni and Choi (1990), Aissani (1994) & Artalejo (1997) consider retrial queues due to unreliable servers. We refer interested readers to Yang and Templeton (1987), Falin (1990); Falin and Templeton (1997) & Artalejo (1999, 2010) for surveys, a monograph and bibliographies on work related to orbital models.

Besides the network literature, ‘retrials’ have also been recognized as an important factor in the call center literature. For example, Whitt (2002) describes some research directions related to stochastic models of call centers and points out that the possibility of postponing some work such as call-back options is worth more careful study. Hoffman and Harris (1986) & Aguir et al. (2004) consider consumer abandonments and retrials in a call center setting to estimate real arrivals (i.e., first-time consumers vs. retrial consumers). Mandelbaum et al. (2002) provide approximations of system metrics of a multi-server system with abandonment and retrials. de Véricourt and Zhou (2005) consider retrials generated by quality problems but not system capacity, that is, a caller whose call has not been resolved will call the service center again with the same concern. Aguir et al. (2008) investigate the impact of disregarding retrials in the staffing of a call center. For other papers in this literature, we refer interested readers to the surveys in Gans et al. (2003) and Akşin et al. (2007). Nevertheless, these papers above typically assume that balking or abandoning consumers

retry with a constant probability.

Artalejo (1995), Artalejo and Lopez-Herrero (2000) & Shin and Choo (2009) consider queues where the retrial probabilities dependent on the number of consumers in the orbit but controlled by the system. In our model, consumers' retry decisions also vary with the state of the system but are endogenously determined by the equilibrium conditions.

Parlaktürk and Kumar (2004) consider self-consumer routing in queueing networks. They study the behavior of the system in Nash equilibrium. Our work is similar with this paper in that any consumer needs to take into consideration future arrivals when making his decisions. Also related is the paper by Hassin and Roet-Green (2011) who consider an unobservable queue where consumers can join, balk or defer their decisions by first 'inspecting' the queue to obtain the information on the queue length.

The closest work to our model contains a set of papers by Armony and Maglaras (2004a,b). In these two papers, the authors study the call center context where customers upon calling and hearing the waiting signal can choose to join the queue, balk or leave their numbers for a customer representative to call back. There is a guaranteed delay before which the call back would take place. Consumers would make their decisions based on this guaranteed call-back delay information with either real-time (Armony and Maglaras, 2004a) or steady-state (Armony and Maglaras, 2004b) waiting-time information provided by the system. The differences between our paper and Armony and Maglaras (2004a,b) are two-fold: First, the retry decisions are driven by consumers in our paper but the call-back decisions by the server in theirs. Second, we focus on the direct analysis of the underlying Markov chain while they provide an asymptotic analysis in heavy-traffic regimes, in a more complex setting.

Kostami and Ward (2009) study an amusement park setting where there is a regular waiting line versus an off-line waiting option (e.g., the FASTPASS system at Disneyland). Similar to Armony and Maglaras (2004a,b), consumers can choose their line to join based on the waiting time information provided by the system, but the authors assume that some con-

sumers waiting in the off-line queue will not return. The focus of this paper is the service provider's capacity allocation rather than consumer decisions.

Finally, Akşin et al. (2013) studies consumers' endogenous abandonment behavior from call center data using a structural estimation approach. They assume that callers waiting in the line make abandon or continue-to-wait decisions at the beginning of discrete time periods. In comparison, the consumers in our model make join and balk decisions, or can defer their decision-making to the following period by exercising the option to retry. Both papers consider costs incurred in the past periods irrelevant for decision-making of the forward-looking consumers, i.e., the consumers do not suffer from sunk cost fallacy.

To summarize, we propose a new model for rational decisions of consumer retrials in queues. We study the welfare effects of retrial decisions as our focus. The main contribution of the paper is to fill in the blank in the existing service operations literature which has typically either omitted retrial decisions or specified it exogenously.

### 3.2. A MODEL OF RETRIALS: BASE MODEL

To introduce our base model, we first consider an observable  $M/M/1$  queue as in Naor (1969). Consumers arrive to the queue with a single server according to a Poisson process. In Naor's model, consumers observe the state of the queue and make join or balk decisions, based on the service value and the cost of waiting in the queue. They join if the queue length is below a certain threshold (typically addressed as the balking threshold), and balk from the queue otherwise. In Naor's model, balking consumers do not return.

#### *3.2.1. Consumer Arrivals and Service Provision*

We consider a model in which all arriving consumers can choose to retry later, in addition to the options to join and balk. For this purpose, we consider an infinite horizon model discretized into time periods. In each period, all arriving customers observe the length of the queue, and make a decision whether to join the queue immediately, or to balk from the



queue (and not return), or decide to come back at retrying the queue in the next period. “Period” is symbolic for the time between retrial attempts, and its length differs by service settings. For instance, a consumer who finds a long line at the post office in the morning will likely choose to retry much later, either in the afternoon or during the next day. A period is equivalent to half a day or a day in this context.

At any given period, the arriving population would consist of consumers who have arrived at the queue for the first time this period (who shall be henceforth referred to as *new* consumers), and those consumers arrived in the past periods, and are retrying the service queue again (referred to simply as *old* consumers). It is likely that a fraction of consumer population has retried multiple times in the past. A schematic representation of retrials is provided in Figure 2 below. Note that retrial is simply a “deferral” action. Eventually, every consumer has to join the queue or balk from it.

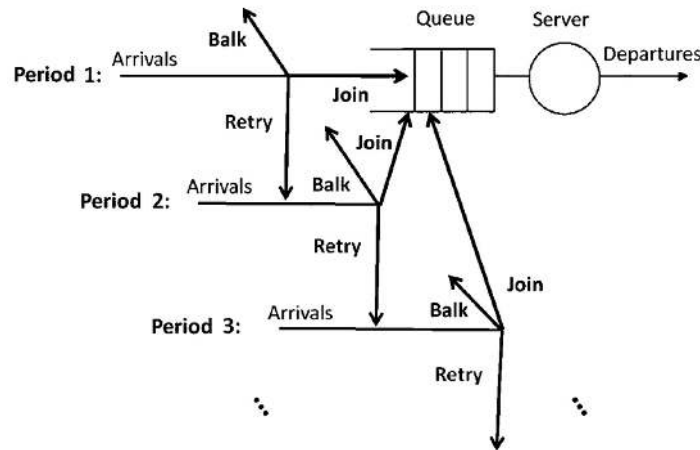


Figure 2: Illustration of the join, balk versus retrial decisions in our model.

We assume that the new consumers arrive to the queue according to a Poisson process with rate  $\lambda$ . In addition to the new consumers, there could be other old consumers re-attempting to join the queue, having decided to retry in the past. The service provider processes consumers waiting in the queue at a rate of  $\mu$  per unit time and the service times are independent and exponentially distributed. We assume, as is done typically, that the values of  $\lambda$  and  $\mu$  are common knowledge.

Upon joining the queue, all consumers are served according to *first come, first served* (FCFS) discipline. We assume that the new arrivals do not exceed the system capacity, i.e.,  $\lambda < \mu$ . Thus, the workload at the server due to new arrivals,  $l \triangleq \lambda/\mu$ , is bounded by 1. However, the total arrivals to the system in any period might exceed the system capacity, since the arrivals also include the consumers retrying from past attempts. Note that while the new arrivals are exogenous in the model, the size of the old thus the total arrivals is endogenously determined, based on the consumer decisions (to retry) in the past.

### 3.2.2. Consumer Actions

Every consumer on arrival to the system observes the queue length, which we refer to the state of the queue  $n$ . On observing  $n$ , the consumer has three choices of actions: Join the queue, Retry or Balk. We define the action set  $\mathcal{A} \triangleq \{J, R, B\}$ . Every *rational* consumer chooses an action that maximizes his long-run risk neutral expected utility. We explore the different actions below.

**Joining:** The joining consumers wait in the queue and incur a cost of  $c$  per unit of time, when they are waiting in line. After service completion, they receive a value  $v$  (net of price) from the service. We do not consider pricing decisions in the model. To receive service, therefore, when there are  $n$  people ahead in the queue, when a customer joins, he receives an expected net value of  $v - \frac{c}{\mu}(n + 1)$ , on the completion of his service. We assume that there is no renegeing or abandonment.

For ease of exposition, we assume that (i)  $v > \frac{c}{\mu}$ , i.e., consumers will join the server if it is idle (otherwise they should not have come to the server at the first place); (ii)  $v$  is not a multiple of  $\frac{c}{\mu}$ . In other words, the payoff from a join decision,  $v - \frac{c}{\mu}(n + 1) \forall n \in \mathbb{N}_0$ , is either positive or negative. When the retry option is not allowed (i.e., in the model of Naor (1969)), this specification implies that there exists a unique balking threshold  $N$  such that  $v - \frac{c}{\mu}(N + 1) < 0 < v - \frac{c}{\mu}N$ .

Formally, we define this *Naor's balking threshold* as

$$N \triangleq \lceil v/\frac{c}{\mu} \rceil - 1 = \lfloor v/\frac{c}{\mu} \rfloor. \quad (3.1)$$

**Balking:** We assume that consumers who balk do not receive the service nor do they incur any waiting cost. Therefore, without loss of generality, we normalize the payoff for the balking consumers to zero.

**Retrials:** In addition to joining the queue or balking from it, a consumer may also decide to retry: He leaves the queue without waiting any further but will return to the server during the following period. When he returns in the following period, this consumer may again decide to join, retry, or balk from the queue, on observing the queue. We can extend the model to randomly distributed retrial intervals – a consumer may retry after a random number of periods, or after a random time length. The discussion is deferred to the appendix for the sake of brevity.

When a consumer chooses to retry, he suffers no waiting cost even as he waits in an off-line queue, like in Cachon and Feldman (2011). However, retrial attempts are costly. Each retrial attempt creates some “hassle” cost to the consumer. We denote this hassle cost as “ $\alpha$ ” which is incurred on *every* retrial attempt. This retrial cost could come from the transportation cost from consumer’s work/home to the service center, or a toll to re-enter the service system, or the penalty cost for rescheduling the visit, or the opportunity cost for the time spent on the trip back and forth. We assume that consumers are homogeneous in their waiting cost  $c$ , their retrial cost  $\alpha$ , and in their value from the service  $v$ . In Section 3.5, we will consider heterogeneous consumer classes.

Consumers can retry as often as they want but have to pay for the “hassle” cost  $\alpha$  every time they retry. They are forward-looking consumers. In each period, they compare the expected payoffs from join, balk and retry decisions, and then choose a state-dependent action that maximizes their expected payoff. We assume that the retrial cost is sunk, i.e.,

retrial costs incurred before arrival to a service occasion are irrelevant to decisions to be made in that period or onward.

Finally, we assume that periods are much longer than a service cycle. This assumption is identical to the frequency of repeat users in queues in prior research. For example, consider the subscription buyers who use a service multiple times in Cachon and Feldman (2011), and the multiple trip users who engage in hyperbolic discounting in Plambeck and Wang (2013). Consistent with these papers, in our model, consumers when retry and come back in the following period expect to see a queue length, that is independent from the current state of the queue.<sup>1</sup> Thus, the state they observe in the next period will be a draw (independent of the current observed states) from the distribution of queue lengths. Consider again the parcel pick-up example - since a single service duration (typically, few minutes for a customer to get his parcel) is relatively short compared to the time between retrials (half a day, or a day). Hence, up on his return to the queue the customer who is retrying would see a *fresh* realization from the distribution of the queue length probabilities.

### 3.2.3. Consumer Strategies

Let  $\mathbb{N}_0 \triangleq \{0, 1, 2, \dots\}$  denote the state space of the queue, i.e., the queue length. Then, a consumer's strategy on a particular service occasion specifies a probability distribution over the action space  $\mathcal{A}$  for any state  $n \in \mathbb{N}_0$ . That is,  $\sigma$  is a strategy if

$$\sigma : \mathbb{N}_0 \rightarrow [0, 1] \times [0, 1] \times [0, 1]$$

such that we write  $\sigma(n) \triangleq [\sigma_J(n) \ \sigma_R(n) \ \sigma_B(n)]^T \ \forall n \in \mathbb{N}_0$ , and

$$\sigma_J(n) + \sigma_R(n) + \sigma_B(n) = 1. \tag{3.2}$$

---

<sup>1</sup>It is well known that a queue converges to its steady-state characteristics, independent of the system's initial state, in a roughly exponential manner, e.g., see Morse (1955), Abate and Whitt (1988), etc. Therefore, the auto-correlation between queue lengths at two time points in a stationary  $M/M/1$  queue converges exponentially to zero as the time interval in-between increases. When a period is much longer than a service cycle, queue lengths observed at two time points in two different periods can be assumed to be independent of each other. Also see Odoni and Roth (1983) for an upper bound on the time it takes for the queue length in a  $M/M/1$  queue to come to steady state.

In words, a consumer who follows strategy  $\sigma$  arriving at the system to observe state  $n \in \mathbb{N}_0$  will join the queue with probability  $\sigma_J(n)$ , retry with probability  $\sigma_R(n)$  and balk with probability  $\sigma_B(n)$ .

A strategy  $\sigma$  is a pure strategy if  $\sigma_a(n)$  is an integer  $\forall a \in \mathcal{A}, \forall n \in \mathbb{N}_0$ . Otherwise,  $\sigma$  is a mixed strategy. For example, the balking threshold strategy being used in Naor (1969) is a pure strategy where  $\sigma_J(n) = 1$  for  $n < N$ , and  $\sigma_B(n) = 1$  for  $n \geq N$ .

Also, in Naor (1969), consumers only have the options to join or to balk. The expected payoff from a join decision  $v - \frac{c}{\mu}(n+1)$  or a balk decision 0 is perfectly determined by the current state of the queue  $n$ , but not influenced by decisions made by future customers. In contrast, in our model consumers are allowed to retry, and they choose an action including retrials to maximize their long-run expected payoff. Since the (long-run) expected payoff from a retry decision would depend on the decisions or strategies chosen by other consumers, we shall pursue an equilibrium analysis.<sup>2</sup>

As all consumers in our model are ex-ante symmetric, we consider symmetric equilibria under which all consumers adopt identical strategy on every service occasion. (We will relax this consideration in Section 3.5.) Nevertheless, consumers may arrive at different states of queue lengths, and end up choosing differing actions. We define the equilibrium strategy below.

**Definition 3** *A (mixed) strategy  $\sigma$  is a symmetric equilibrium strategy if when all consumers adopt  $\sigma$ , and no consumer can strictly improve his expected payoff by unilaterally deviating from  $\sigma$  along any equilibrium path that occurs with a positive probability.*

Since our game is an infinite player game on states that evolve according to a Markovian process, the appropriate equilibrium solution concept is *Markov Perfect Equilibrium* due to Maskin and Tirole (2001) which specifies equilibrium actions for all states that are

---

<sup>2</sup>For example, if a consumer were to make a retry decision during the current period, his long-run expected payoff from this decision would depend on the arrivals in the future periods. And future arrivals are endogenously determined by retrial decisions made by consumers who show up in the current or subsequent periods.

positive recurrent. A Markov Perfect Equilibrium is a Nash Equilibrium. Essentially, given a symmetric equilibrium strategy  $\sigma$  in our model, all the states above  $\underline{n} \triangleq \min\{n \in \mathbb{N}_0 : \sigma_B(n) + \sigma_R(n) = 1\}$  are eventually transient, and have zero probability measure.

The long-run effective workload of the system when the population adopts some strategy  $\sigma$  is bounded above by  $l$  which is less than 1. Thus, there exists a stationary probability distribution of the underlying birth and death process. For a given strategy  $\sigma$ , let us define  $\pi_n^\sigma$  be the long-run probability that the queueing system is in state  $n \in \mathbb{N}_0$ , and use  $\pi_J^\sigma, \pi_R^\sigma$  and  $\pi_B^\sigma$ , to represent the unconditional probabilities that a customer arriving to the queue will choose to join, retry or balk respectively, when the population adopts the strategy  $\sigma$ . Then, the rate for total arrivals in any period is given by

$$\lambda_{total}^\sigma \triangleq \lambda + \lambda\pi_R^\sigma + \lambda(\pi_R^\sigma)^2 + \dots = \frac{\lambda}{1 - \pi_R^\sigma}. \quad (3.3)$$

To understand (3.3), we illustrate the quantity of this period's total arrival rate. Total arrivals in this period include new arrivals (i.e.,  $\lambda$ ) plus the old arrivals who came to the service provider for the first time in the previous period, but decided to retry according to  $\sigma$  (i.e.,  $\lambda\pi_R^\sigma$  for 1-period old consumers), plus the old arrivals who came the service provider two periods back and have retried two periods in a row according to  $\sigma$  (i.e.,  $\lambda(\pi_R^\sigma)^2$  for 2-period old consumers), and so on.

When a period is sufficiently long, we assume that the arrival times for  $k$ -period old customers, who can strategically choose a time point to return in the period, are governed by some renewal process  $\forall k = 1, 2, \dots, \infty$ , according to Lariviere and Van Mieghem (2004). The total arrival process, a superposition of infinitely many renewal processes, is then of the Poisson type (see Feller (1971), pp. 370-371, and Albin (1982)), so we should model the total arrivals including new and all old consumers as a Poisson process with rate  $\lambda_{total}^\sigma$ .

Using PASTA property, we then have

$$\pi_J^\sigma = \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_J(n); \quad \pi_R^\sigma = \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_R(n); \quad \pi_B^\sigma = \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_B(n). \quad (3.4)$$

And it is clear from (3.2) and (3.4) that for any strategy  $\sigma$ ,

$$\pi_J^\sigma + \pi_R^\sigma + \pi_B^\sigma = 1 \quad (3.5)$$

Finally, we denote the traffic intensity of the system under  $\sigma$  by

$$\rho^\sigma \triangleq \frac{\lambda_{total}^\sigma}{\mu} = \frac{\lambda/(1 - \pi_R^\sigma)}{\mu} = \frac{l}{1 - \pi_R^\sigma}. \quad (3.6)$$

Note that although the new workload  $l < 1$ , the total traffic intensity  $\rho$  can be less than, equal to, or greater than 1, which depends on the underlying strategy  $\sigma$  being adopted by the population.

#### 3.2.4. Consumer Best Response

Now we can consider the best response strategy of a consumer  $j$ . Fix the strategy of all other consumers  $i \neq j$  at  $\sigma$  on every service occasion.

The consumer  $j$  can retry repeatedly but has to pay for the “hassle” cost  $\alpha$  every time he retries. Since consumers in our model do not suffer from sunk cost fallacy, when the consumer  $j$  makes a retry versus a join or balk decision upon a particular service occasion, only the future retrial and expected waiting costs are in consideration. Therefore, if a strategy is best for him for one service occasion in response to the population strategy  $\sigma$ , it will be so for all service occasions.

Thus, let us suppose that the consumer  $j$  adopts some strategy  $\sigma^j$  on every service occasion. We can now consider the conditional payoffs of this consumer arriving to the queue and observing a state  $n$ . As described earlier, that the (expected) payoff for consumer  $j$  joining

at a state  $n$ , is the value of the service net of waiting costs, i.e.,  $v - \frac{c}{\mu}(n+1)$ . On the other hand, the payoff for consumer  $j$  balking at any state  $n$  is 0.

Now consider the payoffs for the consumer  $j$  from choosing to retry at state  $n$  (and return during the subsequent period with a retrial cost  $\alpha$ ). In the next period, he may join the queue, balk from it, or retry again, based on the state of the queue he observes. Since all other consumers follow strategy  $\sigma$ , by PASTA property, the consumer  $j$ 's (unconditional) probability of joining, retrying and balking in the next period according to the strategy  $\sigma'$  are, respectively,

$$\pi_J^{\sigma, \sigma^j} \triangleq \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_J^j(n); \quad \pi_R^{\sigma, \sigma^j} \triangleq \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_R^j(n); \quad \pi_B^{\sigma, \sigma^j} \triangleq \sum_{n \in \mathbb{N}_0} \pi_n^\sigma \cdot \sigma_B^j(n). \quad (3.7)$$

If the consumer  $j$  retries again next period, the same decision process plays out, and he faces the same steady-state probabilities in the period after, and so on. Note that the customer  $j$  eventually balks or joins the queue.

Let  $W^\sigma$  denote the expected waiting time for consumer  $j$  conditional on joining the queue in a period when the system is under  $\sigma$ . Then his expected long-run payoff from the retry decision at state  $n$  in the current period is given by

$$\begin{aligned} & \pi_J^{\sigma, \sigma'} \cdot (v - cW^\sigma - \alpha) + \pi_B^{\sigma, \sigma'} \cdot (0 - \alpha) \\ & + \pi_R^{\sigma, \sigma'} \cdot [\pi_J^{\sigma, \sigma'} \cdot (v - cW^\sigma - 2\alpha) + \pi_B^{\sigma, \sigma'} \cdot (0 - 2\alpha)] \\ & + (\pi_R^{\sigma, \sigma'})^2 \cdot [\pi_J^{\sigma, \sigma'} \cdot (v - cW^\sigma - 3\alpha) + \pi_B^{\sigma, \sigma'} \cdot (0 - 3\alpha)] + \dots \\ = & \pi_J^{\sigma, \sigma'} (v - cW^\sigma) + \pi_B^{\sigma, \sigma'} \cdot 0 - (1 - \pi_R^{\sigma, \sigma'})\alpha \\ & + \pi_R^{\sigma, \sigma'} \pi_J^{\sigma, \sigma'} (v - cW^\sigma) + \pi_R^{\sigma, \sigma'} \pi_B^{\sigma, \sigma'} \cdot 0 - \pi_R^{\sigma, \sigma'} (1 - \pi_R^{\sigma, \sigma'}) 2\alpha \\ & + (\pi_R^{\sigma, \sigma'})^2 \pi_J^{\sigma, \sigma'} (v - cW^\sigma) + (\pi_R^{\sigma, \sigma'})^2 \pi_B^{\sigma, \sigma'} \cdot 0 - (\pi_R^{\sigma, \sigma'})^2 (1 - \pi_R^{\sigma, \sigma'}) 3\alpha + \dots \\ = & \frac{\pi_J^{\sigma, \sigma'}}{1 - \pi_R^{\sigma, \sigma'}} (v - cW^\sigma) + \frac{\pi_B^{\sigma, \sigma'}}{1 - \pi_R^{\sigma, \sigma'}} \cdot 0 - \frac{1}{1 - \pi_R^{\sigma, \sigma'}} \cdot \alpha. \end{aligned} \quad (3.8)$$

$$= \frac{\pi_J^{\sigma, \sigma'}}{1 - \pi_R^{\sigma, \sigma'}} (v - cW^\sigma) - \frac{1}{1 - \pi_R^{\sigma, \sigma'}} \cdot \alpha. \quad (3.9)$$



It is clear that the expressions (3.8) and (3.9) are equivalent. However, from (3.8), we have a better interpretation of the underlying quantities. Recall that retrial is only a deferral option, and eventually consumer  $j$  completes his “mission” by either joining the queue or balking. The first term in (3.8) can be interpreted as the probability that consumer  $j$  completes mission by joining the queue (at some period) according to  $\sigma^j$ , times the conditional expected payoff that he will receive upon joining. Similarly, the second term in (3.8) can be interpreted as the probability that consumer  $j$  ends his mission by balking (at some point), times the conditional expected payoff upon balking. And the last term is interpreted as the expected number of periods it takes for consumer  $j$  to finish (by either joining or balking), times the retrial cost. Therefore, we can describe the expected long-run payoff from a retry decision for consumer  $j$  at state  $n$ , given by the expression (3.8), as the total expected end-of-mission payoff less the total expected retrial cost incurred during the process.

On the other hand, we observe that the retrial payoff of the consumer  $j$  at state  $n$  given by (3.8) or (3.9) does not actually depend on  $n$ , and this is because the current state of the queue is independent of the queue length realization in the next period should consumer  $j$  choose to retry. But unlike the joining or the balking payoff, the retrial payoff depends on the underlying population strategy  $\sigma$ . In other words, actions of other consumers may make the retry option more or less attractive to consumer  $j$ .

Suppose  $\sigma^j$  is the best strategy of the consumer  $j$  in response to all other consumers adopting  $\sigma$ , then for every state  $n \in \mathbb{N}_0$ ,  $\sigma^j$  must specify a decision or decisions that maximize the expected payoff for consumer  $j$  among i) the balking payoff 0, ii) the joining payoff  $v - \frac{c}{\mu}(n+1)$ , and iii) the retrial payoff given by (3.9).

At a symmetric equilibrium, we set  $\sigma^j$  to  $\sigma$ , and  $\sigma$  must belong in the best response strategy set in response to itself. It follows from (3.4) and (3.7) that when  $\sigma^j = \sigma$ , we have  $\pi_J^{\sigma, \sigma^j} = \pi_J^\sigma$ ,  $\pi_R^{\sigma, \sigma^j} = \pi_R^\sigma$  and  $\pi_B^{\sigma, \sigma^j} = \pi_B^\sigma$ , and (3.9) is updated. We then have the following proposition, that characterizes the conditions for a strategy to be an equilibrium.

**Proposition 2**  $\sigma$  is an equilibrium strategy, if and only if, for all  $n \in \mathbb{N}_0$ ,

$$\sigma_B(n) > 0 \Rightarrow 0 \geq \max\left\{v - \frac{c}{\mu}(n+1), \frac{\pi_J^\sigma}{1 - \pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1 - \pi_R^\sigma}\right\}; \quad (\text{BALK})$$

$$\sigma_J(n) > 0 \Rightarrow v - \frac{c}{\mu}(n+1) \geq \max\left\{0, \frac{\pi_J^\sigma}{1 - \pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1 - \pi_R^\sigma}\right\}; \quad (\text{JOIN})$$

$$\sigma_R(n) > 0 \Rightarrow \frac{\pi_J^\sigma}{1 - \pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1 - \pi_R^\sigma} \geq \max\left\{0, v - \frac{c}{\mu}(n+1)\right\}. \quad (\text{RETRY})$$

Now that we have introduced the model and the equilibrium concept, we are ready to study the strategic consumers' equilibrium behaviors and their impact on consumer welfare. In Sections 3 and 4, we assume that the population consists of all strategic consumers. Then, in Section 5, we extend our findings on one class of strategic consumers to two classes consisting of both strategic and myopic consumers.

### 3.3. EQUILIBRIUM STRATEGIES

Let the population be entirely made up of *strategic* consumers, i.e., every new and old consumer makes rational state-dependent join, balk and retry decisions, upon arrival to the system.

In (3.9), we showed that the payoff from a retry decision depends on the strategy being adopted by the population. However, since the queue length one sees in the current period is of no use in predicting what the queue realization will be during the following period, the payoff from a retry decision does not depend on the state at which the decision is being made in the current period.<sup>3</sup> It infers that when the population adopts any fixed strategy, the expected payoff from any retry decision is a fixed value.

Now suppose the system reaches an equilibrium when the population adopts some strategy  $\sigma$ . Proposition 2 indicates that for every state  $n \in \mathbb{N}_0$ ,  $\sigma$  must specify the payoff-maximizing

---

<sup>3</sup>For example, suppose that the population adopts a strategy which specifies that one should retry at seeing state  $n = 3$  or  $n = 8$  upon arrival. Then the payoff one expects to receive from a retry decision made at state  $n = 3$  or made at state  $n = 8$  will be identical, because when this consumer returns, he will observe the same stationary process and make future decisions according to the same strategy.

decision(s). Note that the joining payoff  $v - \frac{c}{\mu}(n + 1)$  is linearly decreasing in state  $n \in \mathbb{N}_0$ , the balking payoff 0 is constant, and the retrial payoff is also constant for all states  $n \in \mathbb{N}_0$ . Therefore, the equilibrium strategy  $\sigma$  must be of threshold type where consumers join the queue up to some threshold queue length where the joining payoff drops below the balking or the retrial payoff. A more careful analysis reveals that the equilibrium strategy  $\sigma$  must be in one of the following four types.<sup>4</sup>

Type (i). A threshold *retry strategy* with some threshold  $n$ , denoted by “ $JnR$ ” (join up to  $n$  then retry): Consumers join the queue at states  $\{0, 1, 2, \dots, n - 1\}$  and retry at state  $n$ .

Type (ii). A threshold *join/retry strategy*, denoted by “ $J_{R(1-\beta)}^{J(\beta)}nR$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, n - 2\}$ , mix join and retry decisions at state  $n - 1$ , and retry at state  $n$ .  $\beta \in [0, 1]$  denotes the probability of a join decision being made at state  $n - 1$ , and  $1 - \beta$  a retry decision.

Type (iii). A threshold *balk strategy* with some threshold  $n$ , denoted by “ $JnB$ ” (join up to  $n$  then balk): Consumers join the queue at states  $\{0, 1, 2, \dots, n - 1\}$  and balk at state  $n$ .<sup>5</sup>

Type (iv). A threshold *retry/balk strategy* denoted by “ $Jn_{B(\gamma)}^{R(1-\gamma)}$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, n - 1\}$ , and mix retry and balk decisions at state  $n$ .  $\gamma \in [0, 1]$  denotes the probability of a balk decision being made at state  $n$ , and  $1 - \gamma$  a retry decision.

Note that we have specified the four types of strategies above at only states 0, 1, 2 up to some  $n$ , and this is because, when the entire population adopts any of these strategies, the resulting queueing system is  $M/M/1/n$  (i.e., no state higher than  $n$  will be ever reached).

We illustrate all the possible types of an equilibrium strategy in Table 1 above. The balking and joining payoffs depend only on the parameters and the underlying state of the system, but not the strategy being adopted by the population. In the second and third columns of Table 1, we display the balking and joining payoffs as a function of the state using  $v = 65$

---

<sup>4</sup>By our assumption made before, the joining payoff can only assume strictly positive or negative values. This assumption eliminates some trivial equilibrium strategy types that mix between join and balk decisions.

<sup>5</sup>This is the type of strategies studied in Naor (1969).

State	Action Payoffs	Balk under any strategy	Join under any strategy	Retry under some $\sigma^1$	Retry under some $\sigma^2$	Retry under some $\sigma^3$	Retry under some $\sigma^4$
n=0		0	55	40	25	-8	0
n=1		0	45	40	25	-8	0
n=2		0	35	40	25	-8	0
n=3		0	25	40	25	-8	0
n=4		0	15	40	25	-8	0
n=5		0	5	40	25	-8	0
n=6		0	-5	40	25	-8	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Illustration of the possible equilibrium strategy types (with  $v = 65$  and  $\frac{c}{\mu} = 10$ ).

and  $\frac{c}{\mu} = 10$ . Then, four *imaginary* strategies,  $\sigma^1$ ,  $\sigma^2$ ,  $\sigma^3$  and  $\sigma^4$ , are being considered in the fourth through the eighth column, such that when the population adopts them, the retrial payoffs, given by (3.9), are equal to 40, 25,  $-8$  and 0, respectively. By Proposition 2, we can conclude that  $\sigma^1$  must be a retry strategy type (i.e.,  $J2R$ ) to possibly be an equilibrium strategy, while  $\sigma^2$  has to be a join/retry strategy type (i.e.,  $J_R^J4R$ ),  $\sigma^3$  a balk strategy type (i.e.,  $J6B$ ) and  $\sigma^4$  a retry/balk strategy type (i.e.,  $J6_B^R$ ).

The following theorem further reduces the number of possible equilibrium strategies by imposing bounds on the threshold queue lengths. Specifically, it shows that the threshold of a retry or a join/retry strategy must be smaller than or equal to Naor’s threshold,  $N$ . And, the threshold of a balk or a retry/balk threshold coincides with Naor’s threshold.

**Theorem 3** *If a strategy  $JnR$  or  $J_R^JnR$  is an equilibrium, then we must have  $1 \leq n \leq N$ . On the other hand, if a strategy  $JnB$  or  $Jn_B^R$  is an equilibrium, then we must have  $n = N$ .*

According to Theorem 3, an equilibrium strategy can only be  $JnR$  for  $n \leq N$ ,  $J_R^JnR$  for  $n \leq N$ ,  $JNB$ , or  $JN_B^R$ . Yet knowing the equilibrium candidates does not guarantee the existence of any equilibrium. In the following subsections, we shall identify all the equilibrium strategies.

Note that a retry or a join/retry strategy vaguely refers to “joining short queues and retrying with longer ones”, while a balk or a retry/balk can be interpreted as “joining the queue if there is a positive payoff”. For a given system (i.e., given the parameters  $\lambda, \mu, v, c$ ), it is

intriguing to find out under what conditions (of the exogenous retrial hassle cost  $\alpha$ ) can each type survive an equilibrium. To proceed, we focus on three regions of the retrial cost when it is high, low or moderate.

### 3.3.1. High Retrial Hassle

When the retry option is too costly compared to the net gain of the service (e.g.,  $\alpha \gg v$ ), then retrials should no longer be considered by rational consumers. (The expected payoff from retrials is always negative, and the retry decision is thus dominated by balking). In fact, the retrial cost  $\alpha$  can be even less than  $v$  when retrials are already not worthwhile, because a consumer also incurs some waiting cost when he returns to join the queue in the future.

Define  $\alpha_H \triangleq (1 - \pi_N^{JNB})(v - cW^{JNB})$  where  $\pi_N^{JNB}$  is the steady-state probability of state  $N$ , and  $W^{JNB}$  is the expected waiting time for a consumer conditional on joining the queue, when the population adopts the balk strategy  $JNB$  (or Naor's strategy).  $W^{JNB}$  can be written as

$$W^{JNB} = \frac{\pi_0^{JNB}}{\pi_J^{JNB}} \frac{1}{\mu} + \frac{\pi_1^{JNB}}{\pi_J^{JNB}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1}^{JNB}}{\pi_J^{JNB}} \frac{N}{\mu} = \frac{\pi_0^{JNB}}{1 - \pi_N^{JNB}} \frac{1}{\mu} + \frac{\pi_1^{JNB}}{1 - \pi_N^{JNB}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1}^{JNB}}{1 - \pi_N^{JNB}} \frac{N}{\mu}. \quad (3.10)$$

The following proposition shows that Naor's strategy,  $JNB$ , is a (unique) equilibrium for the system if the retrial cost exceeds  $\alpha_H$ . Note that  $\alpha_H$  is internally determined and can be calculated for any given system, i.e., for given  $\lambda, \mu, v, c$ .

**Proposition 3** *When the retrial cost is sufficiently high such that  $\alpha \geq \alpha_H$ , Naor's strategy, i.e.,  $JNB$ , is an equilibrium. When  $\alpha < \alpha_H$ , the strategy  $JNB$  cannot be an equilibrium.*

It will be verified later in the paper that all other possible equilibrium types cannot be equilibrium strategies on the region  $\alpha \geq \alpha_H$ . Therefore, by Proposition 3, we know that when it is too costly to retry, retrials will not be considered by rational consumers and Naor's system automatically emerges.

### 3.3.2. Low Retrial Hassle

Next we consider settings in which the retrial hassle is *low* compared to the net value of the service. It turns out that in this case, consumers do not balk when the queue is long (i.e., they do not want to leave the service value on the table). Rather, consumers can choose to come back in the future because they can afford to pay for the retrial cost.

We show in this section that there exists some unique threshold  $\alpha_L$  on the retrial cost for a given system, such that  $\alpha_L < \alpha_H$  and when  $\alpha \in (0, \alpha_L]$ , the system equilibrium is given by either a retry strategy (i.e.,  $JnR$ -type) or a join/retry strategy (i.e.,  $J_R^J nR$ -type). However, the Pareto-dominant equilibrium is given by the former type.

We first look for any equilibrium retry strategy (i.e.,  $JnR$ -type), if it exists. Under the strategy  $JnR$ , any consumer who sees a state smaller than  $n$  will join the queue. Otherwise, he retries during the following period. The underlying queueing system when the population adopts  $JnR$  is thus  $M/M/1/n$ . We have  $\pi_J^{JnR} = \sum_{i=0}^{n-1} \pi_i^{JnR}$ ,  $\pi_R^{JnR} = \pi_n^{JnR}$ , and  $\pi_B^{JnR} = 0$ .

The total arrival rate defined in (3.3) for the system under  $JnR$  is thus given by

$$\lambda_{total}^{JnR} = \frac{\lambda}{1 - \pi_R^{JnR}} = \frac{\lambda}{1 - \pi_n^{JnR}}. \quad (3.11)$$

And the traffic intensity defined in (3.6) for the system under  $JnR$  is

$$\rho^{JnR} = \frac{l}{1 - \pi_R^{JnR}} = \frac{l}{1 - \pi_n^{JnR}}. \quad (3.12)$$

We call equation (3.12) the *stability condition* on  $\rho$  and  $\pi_n$  for strategy  $JnR$  which ensures that the steady-state probabilities are consistent. In fact, it is simply a constraint on the variables  $\rho$  and  $n$ . Note that the traffic intensity  $\rho^{JnR}$  depends on the retrial threshold  $n$  through (3.12), but the system's true workload,  $\rho^{JnR} \pi_J^{JnR} = \rho^{JnR} (1 - \pi_n^{JnR}) = \frac{l}{1 - \pi_n^{JnR}} (1 - \pi_n^{JnR})$ , is always equal to  $l$ , i.e., the effective joining rate is  $\lambda$ . This is because under  $JnR$ , we

have  $\pi_B^{JnR} = 0$ . Every consumer never balks, and he or she receives the service eventually (through a certain number of retrials). Since there is no loss of consumers, the long-run effective joining rate must equal to the new arrival rate, namely  $\lambda$ .

Under the strategy  $JnR$ , the total arrival rate is greater than the new arrival rate due to old consumers who are retrying (i.e.,  $\frac{\lambda}{1-\pi_n^{JnR}} > \lambda$ ), but among them only a portion of the consumers (an amount that equals  $\lambda$ ) join the queue. The rest of them (a amount that is equal to  $\frac{\lambda}{1-\pi_n^{JnR}} - \lambda$  or  $\frac{\lambda}{1-\pi_n^{JnR}} \cdot \pi_n^{JnR}$ ) will observe state  $n$  upon arrival to the system and choose to retry accordingly.

From (3.9) and fact that  $\pi_J^{JnR} = 1 - \pi_R^{JnR}$ ,  $\pi_R^{JnR} = \pi_n^{JnR}$ , the retrial payoff under  $JnR$  is

$$v - cW^{JnR} - \frac{\alpha}{1 - \pi_n^{JnR}} \quad (3.13)$$

where the expected waiting time for a customer conditional on joining can be written as

$$W^{JnR} = \frac{\pi_0^{JnR}}{1 - \pi_n^{JnR}} \frac{1}{\mu} + \frac{\pi_1^{JnR}}{1 - \pi_n^{JnR}} \frac{2}{\mu} + \dots + \frac{\pi_{n-1}^{JnR}}{1 - \pi_n^{JnR}} \frac{n}{\mu}.$$

If the retry strategy  $JnR$  indeed leads to an system equilibrium, then besides the stability condition from equation (3.12), it also needs to satisfy the condition in Proposition 2 which ensures that consumers would not want to deviate from it at any state, on any service occasion. In this context, we require that the retry payoff is greater than the joining or balking payoff at state  $n$ , but is less than or equal to the joining payoff at state  $n - 1$ , i.e.,

$$\max\{0, v - \frac{c}{\mu}(n + 1)\} \leq v - cW^{JnR} - \frac{\alpha}{1 - \pi_n^{JnR}} \leq v - \frac{c}{\mu}n \quad (3.14)$$

From Theorem 3, we know that  $n \leq N$ . We can then transfer (3.14) into

$$\frac{c}{\mu}(n + 1) \geq cW^{JnR} + \frac{\alpha}{1 - \pi_n^{JnR}} \geq \frac{c}{\mu}n \text{ with } n \leq N \quad (3.15)$$

which we call the *indifference condition* of the equilibrium. To be more precise here, when  $n = N$ , we require the strategy  $JNR$  satisfies  $v \geq cW^{JNR} + \frac{\alpha}{1-\pi_N^{JNR}} \geq \frac{c}{\mu}N$  instead of  $\frac{c}{\mu}(N+1) \geq cW^{JNR} + \frac{\alpha}{1-\pi_N^{JNR}} \geq \frac{c}{\mu}N$  because of the max operator in (3.14).

We shall look for all possible  $n$  that satisfies both the stability and the indifference conditions, and then pick out the integer solutions (because the retrial threshold  $n$  is an integer). Each integer solution for  $n$  then represents a legitimate equilibrium retry strategy in  $JnR$ . It turns out that

**Lemma 4** *Fix a system (i.e.,  $\lambda, \mu, v, c$ ). For any integer  $n \in \{1, 2, \dots, N\}$ , there exists a unique retrial cost  $\alpha_n$  with which  $n$  satisfies (3.12) and (3.15) at the same time. Moreover,  $\alpha_n$  increases in  $n$ .*

Therefore, with retrial cost  $\alpha_n : n \in \{1, 2, \dots, N\}$ , the integer  $n$  is a solution of equations (3.12) and (3.15). Thus,  $\alpha_n$  induces an equilibrium retry strategy for the system, namely  $JnR$ . However, for  $\alpha_N$ , it does not ensure that  $JNR$  is an equilibrium, because  $0 > v - \frac{c}{\mu}(N+1) = v - cW^{JNR} - \frac{\alpha_N}{1-\pi_N^{JNR}}$  and the indifference condition in (3.14) is thus violated. (This issue was mentioned after condition (3.15).) To fix this boundary condition, let  $\alpha_L : \alpha_{N-1} < \alpha_L < \alpha_N$  be the particular retrial cost that induces a zero retrial payoff when the population adopts  $JNR$ , i.e.,  $0 = v - cW^{JNR} - \frac{\alpha_L}{1-\pi_N^{JNR}}$ , or  $\alpha_L \triangleq (1-\pi_N^{JNR})(v - cW^{JNR})$ . It can be shown that  $\alpha_L < \alpha_H$ , and a rigorous proof is provided in the appendix. We are now at a good position to characterize *all* equilibrium retry strategies of the system in the following lemma.

**Lemma 5** *For a given system, there exist a sequence of thresholds on the retrial cost,  $0 < \alpha_1 < \alpha_2 < \dots < \alpha_{N-1} < \alpha_L$ , such that (i) when the retrial cost  $\alpha \leq \alpha_1$ , the retry strategy  $J1R$  is an equilibrium; (ii) when  $\alpha \in [\alpha_n - \frac{cl}{\mu\rho^{JnR}}, \alpha_n]$  for  $n \in \{2, \dots, N-1\}$ , the retry strategy  $JnR$  is an equilibrium; (iii) when  $\alpha \in [\alpha_L - (v - \frac{c}{\mu}N)\frac{l}{\rho^{JNR}}, \alpha_L]$ , the retry strategy  $JNR$  is an equilibrium.*

Lemma 5 provides us with some important insights. It states that there exist  $N$  intervals on



the low retrial cost region  $\alpha \in (0, \alpha_L]$  such that the retry strategies with increasing thresholds  $J1R, J2R, \dots, JNR$  form equilibrium strategies. Vaguely speaking (i.e., by focusing on only the right end of each interval), when the retrial cost increases consumers would retry at only seeing longer queues. We will return to this point after Theorem 4.

However, Lemma 5 does not provide information on whether these  $N$  intervals overlap or there are gaps between the intervals. Therefore, it is possible that under some retrial cost  $\alpha \leq \alpha_L$ , there does not exist an equilibrium retry strategy because  $\alpha$  is not contained in any of the  $N$  intervals, while some other retrial cost may induce more than one equilibrium. This can happen for example, if  $\alpha_n - \alpha_{n-1} \leq \frac{cd}{\mu\rho_n^{JnR}}$ , then  $\alpha_{n-1}$  will induce both equilibrium retry strategies  $J(n-1)R$  and  $JnR$ . We demonstrate below, via looking at equilibrium join/retry strategies, that this is indeed the case for every  $n \in \{2, 3, \dots, N\}$ . Therefore, there is no gap between the  $N$  intervals stated in Lemma 5. Given the presence of multiple equilibria, we will then identify the Pereto-dominant equilibrium strategy for every  $\alpha \leq \alpha_L$ .

We now search for system equilibria under a join/retry strategy in the form of  $J_{R(1-\beta)}^{J(\beta)} nR$ , i.e., all consumers join the queue at states  $\{0, 1, 2, \dots, n-2\}$ , join the queue with probability  $\beta$  and retry with probability  $1 - \beta$  at state  $n - 1$ , and then retry at state  $n$ .

To tackle down such a strategy, we realize that it is an intermediate phase between two retry strategies. Consider  $J_{R(1-\beta)}^{J(\beta)} nR$ : when  $\beta$  goes to 0, it coincides with the retry strategy with threshold  $n - 1$ , namely  $J(n - 1)R$ ; and when  $\beta$  goes to 1, it coincides with the retry strategy with threshold  $n$ , namely  $JnR$ . And when  $\beta \in (0, 1)$ , we have a non-degenerate join/retry mixed strategy.

We first show in the following lemma that there exists a unique retrial cost that induces  $J_{R(1-\beta)}^{J(\beta)} nR$  as an equilibrium strategy, for fixed  $n$ .

**Lemma 6** *Fix  $1 \leq n \leq N$ . There exists a unique retrial cost  $\alpha$ , for any particular  $\beta \in (0, 1)$ , such that  $J_{R(1-\beta)}^{J(\beta)} nR$  is an equilibrium strategy. Therefore, we can write  $\alpha$  which induces the equilibrium as a function as  $\beta$ . Moreover,  $\alpha$  is continuous in  $\beta$ .*

Lemma 6 tells us that  $\alpha$  is continuous in  $\beta$ . Lemma 7 further tells us that  $\alpha$  decreases in  $\beta$ .

**Lemma 7** Fix  $n \in \{2, \dots, N-1, N\}$ . (i)  $\alpha_n - \frac{cl}{\mu\rho^{JnR}} < \alpha_{n-1}$ . (ii) For any  $\beta \in [0, 1]$ , there exists a unique  $\alpha \in [\alpha_n - \frac{cl}{\mu\rho^{JnR}}, \alpha_{n-1}]$  such that the strategy  $J_{R(1-\beta)}^{J(\beta)}nR$  is an equilibrium. (iii) Moreover,  $\alpha$  decreases in  $\beta$ .

A direct consequence of Lemma 7 is that the  $N$  intervals of the retrial cost constructed in Lemma 5, on which  $J1R, J2R, \dots, JNR$  form equilibria, actually overlap with each other. Therefore, for any retrial cost  $\alpha \leq \alpha_L$ , it induces at least one retry strategy (according to Lemma 5). On the other hand, for  $n \in \{2, 3, \dots, N\}$ , both retry strategies  $J(n-1)R$  and  $J(n)R$  are equilibrium strategies on the overlapped region  $[\alpha_n - \frac{cl}{\mu\rho^{JnR}}, \alpha_{n-1}]$ , plus there exist equilibrium join/retry strategies given by Lemma 7. Fortunately, we can establish the uniqueness of the Pareto-dominant equilibrium strategy for every  $\alpha \in (0, \alpha_L]$  in the following theorem, i.e., the equilibrium strategy that gives the highest consumer welfare as a whole or for each person.

**Theorem 4** For any retrial cost  $\alpha \leq \alpha_L$ , we have one or more symmetric equilibria that are of the retry or join/retry types. However, there exists a unique Pareto-dominant equilibrium: For  $\alpha$  that falls in one of the  $N$  intervals denoted by  $\{I_n\}_{n=1,2,\dots,N}$  where  $I_1 \triangleq (0, \alpha_1]$ ,  $I_2 \triangleq (\alpha_1, \alpha_2]$ ,  $\dots$ ,  $I_{N-1} \triangleq (\alpha_{N-2}, \alpha_{N-1}]$  and  $I_N \triangleq (\alpha_{N-1}, \alpha_L]$ , the corresponding retry strategy  $\{JnR\}_{n=1,2,\dots,N}$  is the Pareto-dominant equilibrium.

Theorem 4 indicates that the Pareto-dominant equilibrium strategies on  $\alpha \leq \alpha_L$  are of the retry type  $JnR$ . And as the retrial cost  $\alpha$  rises within the region  $(0, \alpha_L]$ , the threshold  $n$  for the equilibrium strategies also increases, from 1 to  $N$ . Let us apply the result to a concrete example.

Say, a customer would like to pick up a parcel that is not time urgent, and he can always go to the post office that holds the parcel after work. Imagine the first scenario that this customer literally walks past the post office every evening on his way home, then the additional cost for him to retry will be almost neglectable, i.e.,  $\alpha \rightarrow 0$ . Knowing that there

will be no line at the post office at some point (because the work load  $l < 1$ ), this consumer should adopt the *J1R* strategy. That is, he will only join the queue if the server is idle. Even there is only one consumer that is ahead of him in the system, he should not want to join the queue, because retrials are free in this case, and why not come back to an idle server. Now imagine another scenario where getting to the post office requires some detour. Then, Theorem 4 predicts that the more detour or the more hassle there is to retry, the longer queue this customer upon arrival is willing to tolerate/join, as opposed to retrying.

From the same result, we find that any join/retry strategy is an intermediate phase of two retry strategies and is eliminated by the Pareto-dominance criteria. With this refinement, we can discuss the properties of the Pareto-dominant equilibrium strategies in the following corollary.

**Corollary 1** *Suppose consumers follow the Pareto-dominant equilibrium strategies. As the retrial cost  $\alpha$  increases on  $(0, \alpha_L]$ , we observe there are less retrial consumers and traffic intensity in the system.*

The corollary stated above shows the impact of the retrial cost on the queue outcomes through the Pareto-dominant equilibrium strategy it induces. When it is less costly to retry (i.e., lower  $\alpha$ ), we have a lower threshold  $n$  for the equilibrium retry strategy. And a larger proportion of the consumer population will make a retry decision (i.e., more retrial probability  $\pi_R$  for each consumer and more total traffic  $\rho$  for the system). On the other hand, when it is more costly to retry (i.e., higher  $\alpha$ ), consumers are more reluctant to make retry decisions (higher  $n$ , lower  $\pi_R$  and lower  $\rho$ ).

### 3.3.3. Moderate Retrial Hassle

So far, we have showed that retry strategies (including join/retry strategies) are considered by rational consumers when the retrials are affordable (i.e., the retrial hassle cost  $\alpha \leq \alpha_L$ ), and the balk strategy is being considered when retrials are costly (i.e.,  $\alpha \geq \alpha_H$ ). An interesting question arising then is what can be an equilibrium strategy when the retrial

hassle is moderate, i.e., the retrial cost is between  $\alpha_L$  and  $\alpha_H$ .

In what follows in this section, we propose that the retry/balk strategies (i.e.,  $JN_B^R$ -type) will form equilibria for any retrial cost  $\alpha \in (\alpha_L, \alpha_H)$ . In other words, when the retrial hassle is moderate, at equilibrium consumers join the queue at states  $\{0, 1, 2, \dots, N-1\}$ , then mix retry and balk decisions at state  $N$ . Equivalently, consumers are mixing the retry strategy  $JNR$  and the balk strategy  $JNB$ . To intuit the idea that this mixed type of strategy can form an equilibrium when  $\alpha \in (\alpha_L, \alpha_H)$ , let us consider the retrial cost at  $\alpha_L + \Delta\alpha$  where  $\Delta\alpha$  is infinitesimal.

Recall that when the retrial cost is  $\alpha_L$ , the population adopts a retry strategy with threshold  $N$  (i.e.,  $JNR$ ) at equilibrium, and the expected payoff from a retry decision (at any state) is actually equal to zero. At  $\alpha_L + \Delta\alpha$ , if everybody still sticks with the same strategy, then the expected payoff from the retry decision becomes  $-\Delta\alpha$ , which makes retrials less favorable than the balking option. Thus, every individual has an incentive to balk rather than retrying at state  $N$ , and  $JNR$  can no longer be an equilibrium strategy.

On the other hand, when the retrial cost  $\alpha_L$  increases to  $\alpha_L + \Delta\alpha$ , if everybody switches to the balk strategy (i.e.,  $JNB$ ), then the expected waiting cost (compared to that under the equilibrium strategy  $JNR$  when the retrial cost is  $\alpha_L$ ) would suddenly drop because the retrial population all disappears, more than enough to cover the infinitesimal increased retrial fee ( $\Delta\alpha$ ) to make the retrial option to have a positive return. As a result, every consumer has an incentive to retry at seeing state  $N$  rather than balking, so the balk strategy  $JNB$  is also not an equilibrium strategy at  $\alpha_L + \Delta\alpha$ .

Nevertheless, an equilibrium will be reached for  $\alpha : \alpha_L < \alpha < \alpha_H$  if some consumers adopt the retry strategy while others adopt the balk strategy such that a retry decision at state  $N$  (and at all other states) generates an expected payoff of zero. This leads to the following lemma. Note that, when the population adopts the retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$ , the

steady-state joining, retrial and balking probabilities are given by

$$\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = 1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \text{ and } \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} = \gamma\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}. \quad (3.16)$$

**Lemma 8** *The retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  is an equilibrium strategy for the unique retrial cost*

$$\alpha = (1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \quad (3.17)$$

where  $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$  is the steady-state probability on state  $N$  or the steady-state not-joining probability (balking or retrial) under  $JN_{B(\gamma)}^{R(1-\gamma)}$ , and  $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$  is the expected waiting time for a consumer conditional on joining the system under  $JN_{B(\gamma)}^{R(1-\gamma)}$ .  $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$  can be written as

$$W^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{1}{\mu} + \frac{\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{N}{\mu}.$$

Since there exists a unique retrial cost  $\alpha$  that induces the equilibrium strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  for every  $\gamma \in [0, 1]$ , we can write  $\alpha$  as a function of  $\gamma$ . Moreover, since  $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$  and  $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$  are both continuous in  $\gamma$ , by (3.17) we see that  $\alpha(\gamma)$  is a continuous function in  $\gamma$ . It then can be shown that  $\frac{\partial \alpha}{\partial \gamma} > 0$  (see the proof of Theorem 5 in the appendix). We can then conclude that  $\gamma$ , for which  $JN_{B(\gamma)}^{R(1-\gamma)}$  forms an equilibrium strategy for some  $\alpha$ , is also a function of and increases in  $\alpha$ . This gives the basis of the following theorem.

**Theorem 5** *For any retrial cost  $\alpha \in (\alpha_L, \alpha_H)$ , there exists a unique equilibrium strategy for the system which is the retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$ . Moreover, when  $\alpha$  increases from  $\alpha_L$  to  $\alpha_H$ , the corresponding  $\gamma$  in the equilibrium strategy increases from 0 to 1.*

Theorem 5 fills in the last missing piece of our equilibrium analysis. It states that (i) the unique equilibrium strategy for any retrial cost  $\alpha \in (\alpha_L, \alpha_H)$  is given by a retry/balk strategy. (ii) when  $\alpha \rightarrow \alpha_L$ , consumers adopt the retry strategy  $JNR$  (or  $JN_{B(0)}^{R(1)}$ ) at the equilibrium; when  $\alpha \rightarrow \alpha_H$ , consumers adopt the balking strategy  $JNB$  (or  $JN_{B(1)}^{R(0)}$ ); (iii)

when  $\alpha \in (\alpha_L, \alpha_H)$ , consumers adopt the retry strategy  $JNR$  with probability  $1 - \gamma$  and the balk strategy  $JNB$  with probability  $\gamma$  at the equilibrium, but the likelihood of using the balk strategy (i.e.,  $\gamma$ ) increases in the retrial cost.

### 3.4. CONSUMER WELFARE ANALYSIS

Throughout Section 3.3, we have examined all possible equilibrium strategies for our service system with rational retrials. We briefly summarize these equilibrium results below.

(i) For any low retrial hassle cost  $\alpha \leq \alpha_L$ , multiple retry and join/retry strategies survive the equilibrium. In this case, arriving consumers would join the queue if it is short and retry if it is long. (The retrial cost is low so that consumers would not consider a balk decision to leave the service value on the table when the queue is long.) We also identified the Pareto-dominate equilibrium for each retrial cost  $\alpha$ . As  $\alpha$  increases from 0 to  $\alpha_L$ , the Pareto-dominate equilibria are given by retry strategies,  $JnR$ , where the threshold  $n$  increases from 1 to  $N$ . In other words, when retrials become more costly, consumers would retry only upon seeing longer queues.

(ii) For any moderate retrial hassle cost  $\alpha \in (\alpha_L, \alpha_H)$ , a unique equilibrium retry/balk strategy  $JN_B^R$  exists. That is, when the queue is short, consumers still join the queue for service. However, when the queue is long, they choose between the retry and the balk options. Specifically, as the retrial cost increases in the region, consumers decide to retry less frequently and balk more frequently (when seeing the long queue).

(iii) For any high retrial hassle cost  $\alpha \geq \alpha_H$ , the balking strategy  $JNB$  (or Naor's strategy) gives the unique equilibrium. Simply speaking, when there is too much retrial hassle (such that any retrial attempt becomes very costly), rational consumers no longer consider the retry option. Therefore, the retrial system reduces to Naor's model where only join and balk decisions are allowed. As a result, arriving consumers join the queue when it is short and balk when it is too long.

In this section, we turn to study the consumer welfare under the (Pareto-dominant) equilibrium strategies. Since consumers follow different types of equilibrium strategies on the regions  $\alpha \leq \alpha_L$ ,  $\alpha \in (\alpha_L, \alpha_H)$  and  $\alpha \geq \alpha_H$ , respectively, we will analyze the consumer welfare *region by region*.

Recall that when  $\alpha \leq \alpha_L$ , the Pareto-dominant equilibrium strategy is given by  $JnR$  for some  $n$ . Under  $JnR$ , the total arrival rate is  $\frac{\lambda}{1-\pi_n^{JnR}}$  and the effective joining rate is  $\lambda$ . The overall consumer welfare is thus given by

$$\lambda v - \lambda c W^{JnR} - \lambda \left( \frac{1}{1 - \pi_n^{JnR}} - 1 \right) \alpha \quad (3.18)$$

where the first term  $\lambda v$  is the revenue (rate), the second term  $\lambda c W^{JnR}$  is the total waiting cost (rate), and the last term  $\lambda \left( \frac{1}{1 - \pi_n^{JnR}} - 1 \right) \alpha$  indicates the total retrial cost (rate) because  $\lambda \left( \frac{1}{1 - \pi_n^{JnR}} - 1 \right)$  is the amount of consumers (i.e., total arrival rate less the effective joining rate) who are paying for the retrial fee.

Let us denote  $L^{JnR}$  the average number of consumers in the system when the population adopts  $JnR$ . Applying Little's Law to the population that join the queue, we have

$$\lambda W^{JnR} = L^{JnR}. \quad (3.19)$$

(For completeness, we state a rigorous proof for (3.19) in the appendix.) Therefore, the consumer welfare from (3.18) is also equal to

$$\lambda v - c L^{JnR} - \lambda \left( \frac{1}{1 - \pi_n^{JnR}} - 1 \right) \alpha, \quad (3.20)$$

We know from Theorem 4 that as  $\alpha$  increases on  $(0, \alpha_L]$ , the threshold of the equilibrium retry strategy,  $n$ , increases from 1 to  $N$ , and from Corollary 1 that on average less consumers make retry decisions and more consumers make join decisions at equilibrium. As a result, the system congestion goes up as  $\alpha$  increases on  $(0, \alpha_L]$  and the extra congestion hurts the

consumer welfare. To understand why congestion goes up with less retrying/more joining activities, one can think of retrials as a smoothing mechanism. Because consumers retry only when they see long queues, their retrial activities reduce the steady-state probabilities on higher states of the system and increase those on lower states. In other words, retrials can generate positive externalities to other consumers.

On the other hand, when  $\alpha$  increases on  $(0, \alpha_L]$ , on average less consumers are paying for the retrial costs but each pays more at the equilibrium. Therefore, it is not clear *ex-ante* how the overall consumer welfare is affected by the amount of the retrial cost  $\alpha$ . It turns out that congestion drives consumer welfare to drop in  $\alpha$  on the region  $(0, \alpha_L]$ . Result is shown in the following theorem.

**Theorem 6** *When the retrial cost  $\alpha \leq \alpha_L$ , consumer welfare is a decreasing piecewise linear function of the retrial cost  $\alpha$ , with jumps (discontinuity) at the thresholds  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}$ . The maximum welfare,  $\lambda v - c_l$ , is achieved when  $\alpha \rightarrow 0$ , and the minimum welfare,  $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi^{JNR}} - 1)\alpha_L$ , is achieved when  $\alpha = \alpha_L$ . Moreover, the slope for each linear piece becomes flatter as we move along the intervals in the order of  $(0, \alpha_1], [\alpha_1, \alpha_2], \dots, [\alpha_{N-1}, \alpha_L]$ . See Figure 3.*

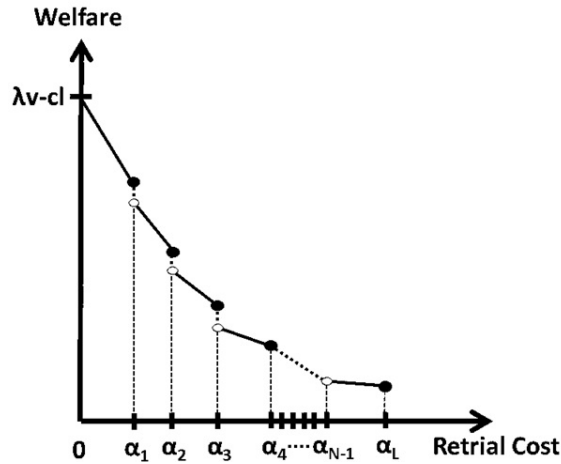


Figure 3: Illustration of consumer welfare at equilibrium as a function of the retrial cost  $\alpha \in (0, \alpha_L]$ .

Theorem 6 states that when the retrial cost increases on  $(0, \alpha_L]$ , consumer welfare de-



creases. The reason we have downward-sloping linear curves on each of the intervals  $\{(0, \alpha_1], [\alpha_1, \alpha_2], \dots, [\alpha_{N-1}, \alpha_L]\}$  is due to the facts that (i) on each individual interval, the same equilibrium holds, i.e., we observe the same system including the same amount of joining, retrial and balking activities, (ii) as  $\alpha$  increases within any individual interval, retrial consumers incur more total retrial cost in proportion. Mathematically, fixing the  $n$ -th individual interval for any  $n \in \{1, 2, \dots, N\}$ , both  $L^{JnR}$  and  $\lambda(\frac{1}{1-\pi_n^{JnR}} - 1)$  stay the same as  $\alpha$  increases. So the total consumer welfare, given by (3.20), decreases linearly in  $\alpha$ . Moreover, the slope gets flatter as  $\alpha$  moves away from left intervals to the right, because the magnitude of the slope, which is described by  $\lambda(\frac{1}{1-\pi_n^{JnR}} - 1) = \rho^{JnR}\mu - \lambda$  from (3.20), decreases in  $n$ . In essence, less consumers make retry decisions as  $\alpha$  increases, so the marginal effect of increasing  $\alpha$  on the total consumer welfare becomes diminishing.

Although consumer welfare decreases over the region  $\alpha \leq \alpha_L$  (i.e., less retrial hassle corresponds to higher consumer welfare), we show in the following Theorem 7 that it actually rises on  $\alpha \in [\alpha_L, \alpha_H]$ . To set up the result, recall that when  $\alpha \in [\alpha_L, \alpha_H]$ , the equilibrium is formed under the retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  for some unique  $\gamma \in [0, 1]$ . Using (3.3) and (3.16), the total arrival rate at the equilibrium is given by

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} = \frac{\lambda}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}. \quad (3.21)$$

Therefore, by (3.3) and (3.21), the effective joining rate at the equilibrium is

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \cdot \pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda(1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \quad (3.22)$$

and the amount of consumers who are paying for the retrial cost is equal to

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \cdot \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda(1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}. \quad (3.23)$$

The consumer welfare (on  $\alpha \in [\alpha_L, \alpha_H]$ ) is given by total payoff less total retrial cost, so by (3.22) and (3.23), it is equal to

$$\frac{\lambda(1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha(\gamma) \quad (3.24)$$

where  $\alpha(\gamma)$  is the particular retrial cost that induces the equilibrium strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$ .

Note that when  $\gamma = 0$ ,  $JN_{B(0)}^{R(1)} = JNR$  and the consumer welfare in (3.24) reduces to that in (3.18). Similarly, when  $\gamma = 1$ ,  $JN_{B(1)}^{R(0)} = JNB$  and the consumer welfare in (3.24) reduces to

$$\lambda(1 - \pi_N^{JNB})(v - cW_N^{JNB}) = \lambda v(1 - \pi_N^{JNB}) - cL^{JNB} \quad (3.25)$$

which gives the consumer welfare in Naor (1969).<sup>6</sup>

**Theorem 7** *When the retrial cost  $\alpha \in [\alpha_L, \alpha_H]$ , the consumer welfare at equilibrium, given in (3.24), is equal to  $\lambda\alpha(\gamma)$ . That is, consumer welfare at the equilibrium increases linearly in the retrial cost  $\alpha$  with slope  $\lambda$ , from value  $\lambda v - cL^{JNR} - \lambda(\frac{1}{1 - \pi_N^{JNR}} - 1)\alpha_L = \lambda\alpha_L$  when  $\alpha = \alpha_L$  to value  $\lambda v(1 - \pi_N^{JNB}) - cL^{JNB} = \lambda\alpha_H$  when  $\alpha = \alpha_H$ .*

Theorem 7 seems counter-intuitive at first by claiming that consumer welfare as a whole increases on the region  $\alpha \in [\alpha_L, \alpha_H]$  despite more retrial cost, but it can be explained as follows. As the retrial cost  $\alpha$  increases over  $[\alpha_L, \alpha_H]$ , consumers who arrive at state  $N$  gradually adopt the balk strategy more often over the retry strategy at the equilibrium (i.e.,  $\gamma$  increases for the underlying equilibrium strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$ ). For these consumers, their payoffs remain the same, in fact remain 0. But they still have the intensives to balk more frequently and retry less frequently because otherwise their payoffs would be negative. However, retrying at state  $N$  causes negative externalities (i.e., congestion) to other consumers while balking at  $N$  does not, thus as  $\alpha$  increases on  $[\alpha_L, \alpha_H]$ , there are less congestion in the system. As a consequence, the overall welfare goes up.

---

<sup>6</sup>In Naor's system, the effective joining rate is  $\lambda(1 - \pi_N^{JNB})$  therefore by applying Little's Law on the effective-joining population, we have  $\lambda(1 - \pi_N^{JNB})W^{JNB} = L^{JNB}$ .

We note that when  $\alpha < \alpha_L$ , consumer welfare will decrease in the retrial cost  $\alpha$  because of more joining and less retrials as  $\alpha$  increases, and retrials (compared to joining) reduces congestion. But when  $\alpha \in [\alpha_L, \alpha_H]$ , consumer welfare will increase in the retrial cost  $\alpha$  as a result of more balking and less retrials as  $\alpha$  increases, and this time retrials (compared to balking) induce congestion. It is important to recognize that a consumer's retry decision can impose both positive and negative externalities to others in the system.

Finally, when  $\alpha \geq \alpha_H$ , consumers follow the balk strategy *JNB* at the equilibrium. There are no more retrial activities so the welfare is independent and a constant function of the retrial cost. The level of the consumer welfare will be the same as that in Naor (1969). Figure 4 below plots the consumer welfare at the equilibrium as a function of the retrial cost  $\alpha$  over the entire region.



Figure 4: Illustration of consumer welfare at equilibrium as a function of the retrial cost  $\alpha$  over the entire region. The welfare decreases in  $\alpha$  on  $(0, \alpha_L]$ , increases on  $[\alpha_L, \alpha_H]$  and stays flat on  $[\alpha_H, \infty)$ .

It is clear from the discussion in this section so far and from Figure 4 that the shape of the equilibrium welfare curve is down-up-flat as a function of the retrial cost  $\alpha$ . However, what is lacking is the comparison between the values of the consumer welfare when  $\alpha \rightarrow 0$  versus when  $\alpha = \alpha_H$ . We present such result in the next theorem.

**Theorem 8** *The consumer welfare is higher when  $\alpha \rightarrow 0$  compared to that at  $\alpha = \alpha_H$ .*

Let us process the information provided by Theorem 8. First, by Theorem 6, we know that the consumer welfare as  $\alpha \rightarrow 0$  is equal to “ $\lambda v - c l = \lambda(v - \frac{c}{\mu})$ ”. This is the largest possible consumer welfare for the population, as  $\lambda$  is the maximum effective joining rate, and  $v - \frac{c}{\mu}$  is the largest possible welfare any individual consumer can get out from the server. Essentially, when  $\alpha \rightarrow 0$ , the population adopts *J1R*. Every consumer retries repeatedly (with paying zero retrial cost) until he sees an idle queue and joins without any wait in the queue. It is thus clear that the welfare when retrials are “free” (i.e.,  $\alpha \rightarrow 0$ ) exceeds that when they are not free (i.e.,  $\alpha > 0$  or  $\alpha \geq \alpha_H$ ).

A counter-intuitive result then follows from Theorem 8 and the down-up-flat shape of the equilibrium welfare curve: *With the additional option to retry (comparing our model to Naor’s model), consumer welfare could however worsen at equilibrium.* In Figure 4, this result is reflected on the interval roughly between  $\alpha = \alpha_3$  and  $\alpha = \alpha_H$ .

This phenomenon can be explained by an argument similarly to Naor (1969): Consumers are self-interested by heart. When they are presented with the additional option to retry, they will take advantage of it as long as exercising the option can increase their individual payoffs. However, at times, these extra gains in individual welfare cannot compensate for the negative externalities (i.e., congestion costs) imposed to other consumers. Therefore, at these times the overall consumer welfare would actually improve if consumers do not have the privileges in retrying.

Naturally, the next question to be asked is what would the socially optimal policy be.

#### *3.4.1. Socially Optimal Policies*

In queueing systems, self-interested consumers usually form equilibrium that deviates from the socially optimal outcome. For example, Naor (1969) shows that consumers (who have the options to join or balk) over-congest the system when left to their devices. He then

suggests that tolls or taxes can be levied to control the joining population.

We have established equilibrium strategies and welfare results for self-interested consumers with the option to retry in this paper so far, and shall now consider socially optimal policies (i.e., first best solutions) within the model framework. Socially optimal policies are highly state dependent. Like Naor (1969), we will focus on policies of the threshold type. These types of policies have been used for queue controls in the literature for settings that even generalize Naor (1969), e.g., see Yechiali (1971, 1972) and other related papers surveyed in Stidham (1985). Specifically, we consider all the retry and balk strategies, i.e., the class of strategies  $\{s : s = JnR \text{ or } s = JnB \text{ for some } n \geq 1\}$ . We characterize the socially optimal policies (within this class) in the following theorem.

**Theorem 9** *For a given system  $\lambda, \mu, c, v$  (thus  $\alpha_L$  is also determined), there exists a deterministic point  $\alpha' \in (0, \alpha_L]$  such that the optimal policy is a retry strategy (i.e.,  $JnR$ -type) with increasing thresholds for  $\alpha \leq \alpha'$ , and a balk strategy with threshold  $N'$  (i.e.,  $JN'B$ ) for  $\alpha > \alpha'$  where  $N' < N$  is the socially optimal balking threshold defined in Naor (1969). Moreover, for any fixed  $\alpha \leq \alpha'$ , the threshold of the retry strategy at the social optimum is smaller than or equal to that at the consumer individual equilibrium. Also see Figure 5 for an illustration.*

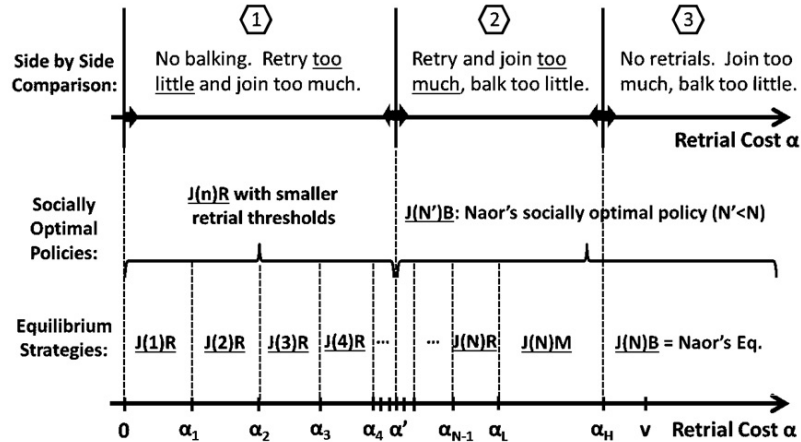


Figure 5: Illustration of equilibrium strategies (bottom) versus the socially optimal policies (middle) as a function of the retrial cost  $\alpha$ . Side by side comparison is given at top.

Theorem 9 indicates that (i) when the retrial cost is low (i.e., region 1 in Figure 5), the retrial thresholds for the equilibrium strategies are smaller than those for the socially optimal policies, and thus consumers retry too little and join too much at the self equilibrium compared to the socially optimum; (ii) when the retrial cost rises (i.e., region 2 in Figure 5), consumers retry and join too much, but balk too little on their own; and (iii) eventually when the retrial cost is high (i.e., region 3 in Figure 5), consumers join too much and balk too little.

Result (iii), where consumers over-join the system, certainly is not new. Recall that when the retrial cost is high, consumers on their own disregard the option to retry, and our model becomes the one in Naor (1969). And Naor (1969) points out in his paper that self-interested consumers join the system too much as opposed to balking because that ignore the negative externalities (i.e., congestion costs to other consumers) in making the decisions.

On the other hand, results (i) and (iii) are new and counter-intuitive. For example, when the retrial cost is low, one should anticipate that consumers, when left to their own devices, would retry a lot because they can afford it. However, our result indicates that they are not retrying enough compared to the socially optimal outcome. On the other hand, when the retrial cost rises, one would anticipate that consumers are discouraged from retrying. But we find that they still retry too much compared to social optimum when everyone were altruistic.

Fortunately, we can explain consumers' deviation from the social optimum above by the same principle that Naor (1969) has observed, i.e., self-selecting consumers suffer from externalities. Recall earlier in the paper, we have identified both the positive and negative externalities of a retrial activity. When the retrial cost is low, there are no balking at both self-equilibrium and social optimum. But, a consumer who retry at a long queue (compared to joining) reduces the congestion costs imposed on other consumers. And self-interested consumers over-join the system and do not retry enough, because they ignore these positive externalities from retrials. In contrast, as the retrial cost rises, consumers retry and join too

much, but balk too little at equilibrium because they fully ignore the negative externalities from joining and retrials. If everyone were acted to maximize the overall consumer welfare in this case, some of those consumers would balk.

### 3.5. EXTENDED MODEL: TWO CLASSES OF CONSUMERS

In many service settings, a portion of the consumer population cannot delay or give up the workload. Examples include patients visiting a hospital who are in critical conditions, or consumers arriving at DMV centers who are renewing driving licenses at the very last minute. In this section, we extend the basic model to two classes of consumers, to include these who arrive at deadlines.

We assume that among the new arrival rate  $\lambda$ , the portion  $(1-\theta)\cdot\lambda$ , are strategic consumers like in the basic model who will make rational decisions to join the queue, balk, or to retry later. Rest of the population,  $\theta\cdot\lambda$ , are non-strategic or *myopic* consumers who will join the queue unconditionally. We assume that  $\theta \in [0, 1)$  to ensure there exist some strategic consumers whose equilibrium strategies are of our interest. On the other hand, the basic model is a special case of the two-class model with  $\theta = 0$ . With two class of consumers, we can now study the impact of myopic consumers on the decision-making and welfare of the strategic consumers, as well as the impact of the strategic consumes on the myopics.

#### 3.5.1. *Equilibrium Strategies*

First we shall examine how strategic consumers make retry, versus join and balk decisions in the presence of the myopics. For any fixed mixture of strategic and myopic consumers in the population, (i.e., fixed  $\theta$ ), due to the monotonicity of the balking, joining and retrial payoffs with respect to the state of the queue, only four types of equilibrium strategies can survive the equilibrium for strategic consumers like in the basic model.

These four types of equilibrium candidates are also almost identical to the ones described before except now they need to specify actions for every state of the system. (All states are

now positive recurrent due to the existence of myopic consumers.) Simply speaking, the equilibrium candidates for the extended model are the ones for the basic model with the action on the last state recurring forever, so we will keep the same notations and names wherever unambiguous:

Type (i). A threshold retry strategy with some threshold  $n \leq N$ , denoted by “ $JnR$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, n-1\}$  and retry at not only state  $n$  but also all the states  $\{n+1, n+2, \dots\}$ .

Type (ii). A threshold join/retry strategy with some threshold  $n \leq N$ , denoted by “ $J_{R(1-\beta)}^{J(\beta)}nR$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, n-2\}$ , mix join and retry decisions at state  $n-1$  and retry at states  $\{n, n+1, n+2, \dots\}$ .  $\beta \in [0, 1]$  still denotes the probability of a join decision being made at state  $n-1$ , and  $1-\beta$  a retry decision.

Type (iii). The threshold balk strategy (Naor’s strategy), denoted by “ $JNB$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, N-1\}$  and balk at states  $\{N, N+1, N+2, \dots\}$ .

Type (iv). Finally, a threshold retry/balk strategy denoted by “ $JN_B^R$ ”: Consumers join the queue at states  $\{0, 1, 2, \dots, N-1\}$  and mix retry and balk decisions at states  $\{N, N+1, N+2, \dots\}$ .

Note that a retry/balk strategy  $\sigma$  in this case can actually have different retrial versus balking probabilities at states  $\{N, N+1, N+2, \dots\}$ . Let  $\gamma_N^\sigma, \gamma_{N+1}^\sigma, \gamma_{N+2}^\sigma, \dots$  denote the corresponding balking probabilities at these states, we will define  $JN_{B(\gamma)}^{R(1-\gamma)}$  for any  $\gamma \in [0, 1]$  as an equivalence class of retry/balk strategies whose ratio of the steady-state balking probability over the steady-state non-joining probability is equal to  $\gamma$ , in other words,

$$JN_{B(\gamma)}^{R(1-\gamma)} \triangleq \{\text{Retry/balk strategy } \sigma : \frac{\pi_B^\sigma}{\pi_B^\sigma + \pi_R^\sigma} = \frac{\gamma_N^\sigma \pi_N^\sigma + \gamma_{N+1}^\sigma \pi_{N+1}^\sigma + \dots}{\pi_N^\sigma + \pi_{N+1}^\sigma + \dots} = \frac{\sum_{i=N}^{\infty} \gamma_i^\sigma \pi_i^\sigma}{\sum_{i=N}^{\infty} \pi_i^\sigma} = \gamma\}.$$

We show in the next lemma that all the strategies in the same equivalence class correspond to the same underlying queueing system. In other words, one cannot differentiate strategies from an equivalence class  $JN_{B(\gamma)}^{R(1-\gamma)}$  by observing the queueing system and its evolution.



**Lemma 9** Fix  $\theta \in [0, 1)$ . When the strategic population,  $(1-\theta)\cdot\lambda$ , adopts any two retry/balk strategies from a given  $JN_{B(\gamma)}^{R(1-\gamma)}$  class defined above, the underlying queueing systems are identical.

As a result of Lemma 9, we will treat each equivalent class  $JN_{B(\gamma)}^{R(1-\gamma)}$  as one strategy. It turns out that with the presence of myopic consumers, the equilibrium strategies adopted by the strategic class share the same structure as before.

For the two-class model, define

$$\alpha_L \triangleq (1 - \pi_R^{JNR})(v - cW^{JNR}) = (1 - \sum_{i=N}^{\infty} \pi_i^{JNR})(v - cW^{JNR}) = (1 - \frac{\pi_N^{JNR}}{1 - \theta l})(v - cW^{JNR}),$$

$$\alpha_H \triangleq (1 - \pi_B^{JNB})(v - cW^{JNB}) = (1 - \sum_{i=N}^{\infty} \pi_i^{JNB})(v - cW^{JNB}) = (1 - \frac{\pi_N^{JNB}}{1 - \theta l})(v - cW^{JNB}),$$

where  $W^\sigma$  is the expected waiting time for a *strategic consumer* conditional on joining the queue, when the *strategic* population adopts  $\sigma$ .

We recover all the results in Sections 3.3 and 3.4 for two classes of consumers:

**Lemma 5'** Fix  $\theta \in [0, 1)$ . There exist a sequence of retrial cost thresholds  $0 < \alpha_1 < \alpha_2 < \dots < \alpha_{N-1} < \alpha_L$  such that (i) when the retrial cost  $\alpha \leq \alpha_1$ , the retry strategy  $J1R$  is an equilibrium. (ii) when  $\alpha \in [\alpha_n - \frac{c}{\mu}(1 - \pi_R^{JnR}), \alpha_n]$  for  $n \in \{2, \dots, N-1\}$ , the retry strategy  $JnR$  is an equilibrium. (iii) when  $\alpha \in [\alpha_L - (v - \frac{c}{\mu}N)(1 - \pi_R^{JNR}), \alpha_L]$ , the retry strategy  $JNR$  is an equilibrium.

**Lemma 7' & Theorem 4'** Fix  $\theta \in [0, 1)$ . (i) For  $n \in \{2, \dots, N-1, N\}$ ,  $\alpha_n - \frac{c}{\mu}(1 - \pi_R^{JnR}) < \alpha_{n-1}$ , and there exists a unique  $\alpha \in [\alpha_n - \frac{c}{\mu}(1 - \pi_R^{JnR}), \alpha_{n-1}]$  such that the join/retry strategy  $J_{R(1-\beta)}^{J(\beta)}nR$  is an equilibrium for any  $\beta \in [0, 1]$ . (ii) The Pareto-dominant equilibrium exists and is unique for any  $\alpha \leq \alpha_L$ : For  $\alpha$  that falls in one of the  $N$  intervals  $\{(0, \alpha_1], (\alpha_1, \alpha_2], \dots, (\alpha_{N-2}, \alpha_{N-1}], (\alpha_{N-1}, \alpha_L]\}$ , the retry strategies  $J1R, J2R, \dots, J(N-1)R, JNR$  represent the Pareto-dominant equilibria, respectively.

**Proposition 3' & Theorem 5'** Fix  $\theta \in [0, 1)$ . (i)  $\alpha_L < \alpha_H$ . (ii) The balk strategy  $JNB$  is the unique equilibrium strategy for  $\alpha \geq \alpha_H$ . (iii) For any retrial cost  $\alpha \in (\alpha_L, \alpha_H)$ ,  $JN_{B(\gamma)}^{R(1-\gamma)}$  is a unique equilibrium strategy (class) for the system. When  $\alpha$  increases from  $\alpha_L$  to  $\alpha_H$ , the corresponding  $\gamma$  in the equilibrium strategy increases from 0 to 1. The bijection between  $\alpha$  and  $\gamma$  is given by

$$\alpha = \left(1 - \sum_{i=N}^{\infty} \pi_i^{JN_{B(\gamma)}^{R(1-\gamma)}}\right)(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) = \left(1 - \frac{\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \theta l}\right)(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}).$$

In essence, strategic consumers are committed to the same pattern of equilibrium strategies with or without presence of the myopic consumers, i.e., a consumer who plays strategically is always going to join the queue if it is short and retry or balk if it is long. In strategic consumers' minds, the myopic population,  $\theta \cdot \lambda$ , is treated as if a given *environment*. A change in the environment (i.e., a change in  $\theta$ ) results in changes in the values of the retrial cost thresholds, but not the structure of the equilibrium strategies. Henceforth, we use environment  $\theta$  to denote a population  $\lambda$  where  $(1 - \theta) \cdot \lambda$  are the strategic consumers and  $\theta \cdot \lambda$  are the myopics.

### 3.5.2. Conditional and Overall Welfare

When there was only one class of strategic consumers in the basic model, we studied the total consumer welfare. Recall that it first decreases, then increases and finally stays constant as a function of the retrial cost  $\alpha$ . Scaling the consumer welfare curve by a factor of  $\frac{1}{\lambda}$ , we know that the consumer welfare per consumer in the basic model also follows the same down-up-flat pattern.

Now that we have two classes of consumers (with some environment  $\theta$ ), we can study the following three welfare quantities: (i) consumer welfare per strategic consumer; (ii) consumer welfare per myopic consumer; and (iii) consumer welfare per consumer in the population. Moreover, we can examine how the three welfare quantities change in  $\theta$ .

**Welfare per Strategic Consumer.** It turns out that not only do strategic consumers adopt the same structure of equilibrium strategies with or without the presence of myopic consumers, the consumer welfare per strategic consumer also bears the same shape as before.

**Theorem 6’, Theorem 7’ & Proposition 8’** Fix  $\theta \in [0, 1)$ . (i) When the retrial cost  $\alpha \leq \alpha_L$ , the consumer welfare per strategic consumer decreases in  $\alpha$ , from  $v - \frac{c}{\mu}$  to  $\alpha_L$ . (ii) When the retrial cost  $\alpha \in [\alpha_L, \alpha_H]$ , the consumer welfare per strategic consumer increases linearly in  $\alpha$ , from  $\alpha_L$  to  $\alpha_H$ . (iii) When the retrial cost  $\alpha \geq \alpha_H$ , the consumer welfare per strategic consumer remains constant at the value of  $\alpha_H$ .

Therefore, for any given  $\theta \in [0, 1)$ , the function that describes the welfare per strategic consumer in the retrial cost  $\alpha$  has the same down-up-flat shape as before. But  $\alpha_L$ ,  $\alpha_H$  and all other retrial cost thresholds are all functions of  $\theta$  and change as  $\theta$  does. We demonstrate in the following lemma that all these values monotonically decrease in  $\theta$ .

**Lemma 10** (i) The values of  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}, \alpha_L, \alpha_H$  all decrease in  $\theta$ . (ii) The unique retrial cost in the region  $[\alpha_L, \alpha_H]$  that induces the equilibrium retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  for each  $\gamma \in [0, 1]$ , denoted by  $\alpha(\gamma)$ , also decreases in  $\theta$ . Note that  $\alpha(0) = \alpha_L$  and  $\alpha(1) = \alpha_H$ . (iii)  $\alpha_H - \alpha_L \rightarrow 0$  as  $\theta \rightarrow 1$ . On the other hand, we still have  $\alpha_1 < \alpha_2 < \dots < \alpha_{N-1} < \alpha_L$  as  $\theta \rightarrow 1$ .

There is an important message from Lemma 10: By focusing on the values of the retrial cost thresholds or the “turning points” of the welfare curve, we see that the presence of more myopic consumers makes retrials more costly for strategic consumers (i.e., consumers are making the same decisions as before only with smaller retrial hassle). We will explain the intuition after presenting the next theorem, which describes how the whole welfare curve shift in  $\theta$ .

Let  $\mathcal{U}_{\alpha, \theta}^*$  denote the welfare *per strategic consumer* under the Pareto-dominant equilibrium strategy when the environment (i.e., the ratio of the myopic consumers in the population) is  $\theta$  and the retrial cost is  $\alpha$ . We show that, fixing any retrial cost, consumer welfare per

strategic consumer decreases when there are more myopic consumers in the population.

**Theorem 10** *If  $0 \leq \theta_1 < \theta_2 < 1$ , then  $U_{\alpha, \theta_1}^* \geq U_{\alpha, \theta_2}^*$  for all retrial cost  $\alpha \in (0, \infty)$ . That is, the presence of myopic consumers reduces the welfare of each strategic consumers. Also see Figure 6.*

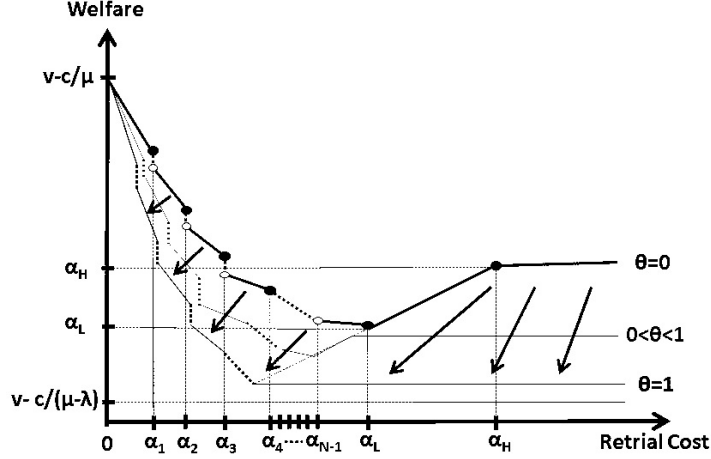


Figure 6: Illustration of the welfare *per strategic consumer* at equilibrium as the myopic population increases. Note that  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}, \alpha_L$  and  $\alpha_H$  all decrease in  $\theta$ , and  $\alpha_H \rightarrow \alpha_L$  as  $\theta \rightarrow 1$ . For any particular  $\theta \in [0, 1)$ , the welfare curve remains the down-up-flat shape.

Theorem 10 provides us with insights on the impact of myopic consumers. It indicates that the presence of myopic consumers makes every strategic consumer worse off. Let us imagine a case where one strategic consumer turns himself into myopic (i.e., consider the environment  $\theta \rightarrow \theta + \Delta\theta$  where  $\Delta\theta$  is infinitesimal). Conditional on the state under which this consumer arrives to the system, say  $n$ : (i) If this consumer's decision were to join the queue (if he were a strategic consumer), then making him a myopic consumer does not affect his or anybody else's welfare, because a myopic consumer joins the queue. (ii) If this consumer's decision were to balk, then forcing him join the queue not only makes this consumer worse off (receiving a negative payoff versus a zero payoff), but his join decision also causes negative externalities to all other consumers in the system. (iii) If this consumer's decision were to retry, then forcing him join the queue reduces his payoff because the retry decision would be more rewarding. On the other hand, both retry and join decisions cause negative externalities to other consumers. However, a retry decision is

not as bad as a joint decision because it can also generate positive externalities.

Therefore, overall speaking, when myopic consumers replace strategic consumers, the system becomes more congested. This explains the result of Lemma 10 that with the presence of myopic consumers, it were as if the case that the retrial hassle for every strategic consumer has increased.

Going forward, we find that the existence of the myopic consumers not only reduces strategic consumers' welfare, but also their own welfare. In other words, the presence of strategic consumers increases the welfare of the myopics.

**Welfare per Myopic Consumer.** We now denote by  $\mathcal{V}_{\alpha,\theta}^*$  the welfare *per myopic consumer* when the environment is  $\theta$  and the retrial cost is  $\alpha$ , (and when the strategic population adopts the equilibrium strategies). As seen in the following theorem, the welfare per myopic consumer under any  $\theta$  exhibits the down-up-flat welfare pattern, and decreases curve-wise in  $\theta$ .

**Theorem 11** Fix a given  $\theta \in [0, 1)$ . Welfare per myopic consumer is a decreasing step function over  $\alpha \leq \alpha_L$ . It rises over  $\alpha \in [\alpha_L, \alpha_H]$ , and then remains constant over  $\alpha \geq \alpha_H$ . Moreover, if  $0 \leq \theta_1 < \theta_2 < 1$ ,  $\mathcal{V}_{\alpha,\theta_1}^* \leq \mathcal{V}_{\alpha,\theta_2}^*$  for all  $\alpha$ . Also see Figure 7.

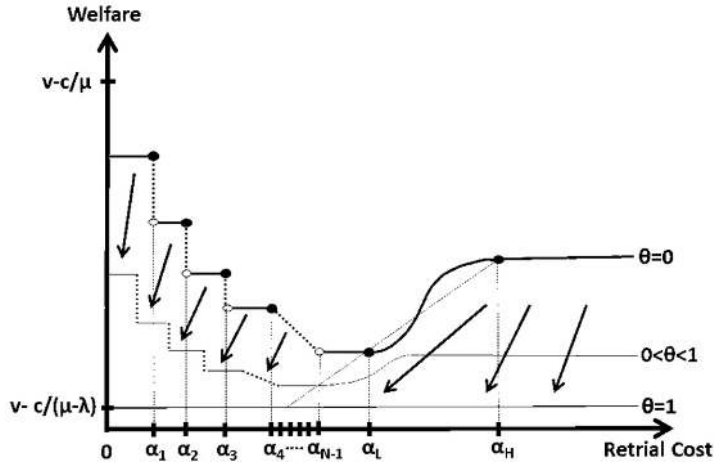


Figure 7: Illustration of the welfare *per myopic consumer* as the myopic population increases.

The welfare curve forms a step function on the region  $\alpha \leq \alpha_L$  because the underlying equilibrium strategy adopted by the strategic consumers remains the same for all  $\alpha$  that falls in one of the  $N$  intervals:  $(0, \alpha_1], (\alpha_1, \alpha_2], \dots, (\alpha_{N-2}, \alpha_{N-1}], (\alpha_{N-1}, \alpha_L]$ . The slope for each linear piece is zero because the retrial cost is irrelevant to the welfare of a myopic consumer given a fixed strategy used by the strategic population. Welfare decreases on  $\alpha \leq \alpha_L$  and increases on  $(\alpha_L, \alpha_H)$  is due to increased and decreased expected waiting cost for each myopic consumer. On the other hand, the whole welfare curve for each myopic consumer decreases in  $\theta$  due to the congestion that the myopic consumers create, as already explained before.

For a second, consider the extreme case when  $\theta \rightarrow 1$ , i.e., the whole population approaches myopic. Then the underlying system would coincide with the regular  $M/M/1$  queue with no balking or retrials. One would then conjecture that the welfare per myopic consumer (as well as per consumer in this case) would be equal to the welfare per capita in a regular  $M/M/1$  system for all retrial cost  $\alpha$ . We show in the proposition below that this is indeed the case.

**Proposition 4** *As  $\theta \rightarrow 1$ ,  $\mathcal{V}_{\alpha_n, \theta}^{JnR} \downarrow (v - \frac{c}{\mu - \lambda})$  for all  $n = 1, 2, \dots, N - 1$ . Moreover,  $\mathcal{V}_{\alpha_L, \theta}^{JNR}$  and  $\mathcal{V}_{\alpha_H, \theta}^{JNB} \downarrow (v - \frac{c}{\mu - \lambda})$ . Since the welfare curve is a step function on  $\alpha \leq \alpha_L$  with jumps at  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}$  for every  $\theta$ , and  $\alpha_H \rightarrow \alpha_L$  as  $\theta \rightarrow 1$ , we conclude that welfare per (myopic) consumer becomes  $(v - \frac{c}{\mu - \lambda})$  as  $\theta \rightarrow 1$  for any retrial cost  $\alpha \in (0, \infty)$ .*

**Consumer Welfare per Consumer in the Population.** Following Theorems 10 and 11, we draw conclusion that welfare per capita in the mixed population (including both the strategic and the myopic consumers) has the down-up-flat shape as a function of the retrial cost  $\alpha$  for any given  $\theta$ , and that the whole welfare curve decreases in  $\theta$ . When  $\theta = 0$ , it coincides with the welfare curve for one class of strategic consumers. And when  $\theta = 1$ , it becomes a constant function in  $\alpha$  with a value of  $v - \frac{c}{\mu - \lambda}$ , i.e., the  $M/M/1$  consumer welfare.

### 3.6. CONCLUSIONS AND IMPLICATIONS

Consumers often use retry decisions in practice when they are faced with a long queue. However, existing service operations literature on modeling strategic consumer decisions has primarily focused on join and balk decisions. As a result, not much is known about consumers' decision-making when they also have the option to retry. Our paper seeks to fill this blank in the literature by modeling consumers' rational retrials.

We find that, with or without the presence of myopic consumers (who join the queue unconditionally), consumers who play strategically act the following way when given the options to join, balk or retry: (i) When the hassle cost to retry is low, consumers follow the threshold retry strategies, i.e., they join the queue if it is short than some threshold and retry otherwise. All else being equal, when the retrial cost increases, the retrial threshold also increases. (ii) When the hassle cost to retry becomes significant, all consumers still join short queues, but some will start to balk from long queues. (iii) Eventually when the hassle cost is too high (e.g., when it exceeds the value from the service), consumers follow the threshold balking policy given in Naor (1969).

Surprisingly, allowing consumers to retry does not always generate higher overall consumer welfare than not allowing (i.e., the model in Naor (1969)). This is because the extra gain an individual benefits from retrials does not always compensate for the negative externalities (i.e., congestions costs) these retrial activities impose on other consumers in the system, when balk decisions are actually better off for the society as a whole.

We also identified positive externalities that retrials can generate. Since consumers retry at seeing longer queues, their retrial activities effectively reduce the steady-state probabilities on the higher states of the system and increase those on the lower states. This leads to less congestion in the system, and reduced expected waiting costs for every other consumer. To some sense, retrials serve as a great smoothing mechanism to spread the workload over time.

In the case of self-interested consumers, they completely ignore externalities to other consumers when making decisions and thus behave differently from the socially optimal point of view. In Naor (1969), consumers when left to their own over-congest the system because they ignore the negative externalities of joining the queue. We find that when the retrial cost is small, consumers on their own do not retry enough compared to the social optimum (which is by itself a counter-intuitive result), and the deviation comes from ignorance of the positive externalities of the retrials.

In an important extension to the basic model, we consider both strategic and myopic consumers in the population. The strategic consumers make state-dependent join, balk and retry decisions as in the basic model, but the myopics unconditionally join the queue. Thus our findings can be further applied to settings such as hospitals where both types of consumers coexist. We find that, when there is a higher ratio of myopic consumers in the population, welfare decreases for everybody due to the extra congestion that the myopic consumers bring to the system.

Our findings have several implications for queue management policies in many service settings. As consumer equilibrium deviates from the social optimum, a social planner could implement various policies to improve the overall consumer welfare. Specifically, we find that consumers retry too little when the retrial hassle is low; and they retry too much when it is high. As a result, the social planner should consider subsidizing retrial consumers or charging tolls for each visit to encourage or discourage retrials. In an incoming paper, we devote our focus to the control policies.

Finally, in modeling retrials in this paper, we assume that the number of periods between individual retrials is fixed. However, all of our results will continue to hold if we allow the number of periods between retrials to be random. In fact, instead of using the periodic model that we currently have, our results still hold if the time between retrials is modeled by an exponential distribution like in orbital models. (We will address this issue with details in the appendix.) Therefore, the insights found in this paper are general and robust.



## CHAPTER 4 : MANUFACTURING SOURCING IN A GLOBAL SUPPLY CHAIN: A LIFE CYCLE ANALYSIS

### 4.1. INTRODUCTION

For decades, the dominant strategy in U.S. manufacturing has been to outsource to low labor-cost countries, first dating back to the 60's and the 70's when off-shoring manufacturing in Japan occurred, and the 70's and 80's to South Korea. With the advent of the 90's, the U.S. started to see a rapid transfer of production jobs to China and other low labor-cost countries such as Vietnam, Indonesia, and Bangladesh.

According to U.S. Bureau of Labor Statistics, more than five and half million manufacturing jobs were lost between the years 2000 to 2010 (i.e., from above 17 million jobs in 2000 to below 12 million in 2010). Figure 8 below plots the number of U.S. manufacturing jobs over time. The sheer drop happened when the congress agreed to permanent normal trade relations (PNTR) status with China and President Clinton signed it into law in 2000, which paved the way for China's accession to the World Trade Organization (WTO) and its rise as the single most favorable host country for outsourcing of manufacturing.

Today, however, this trend is being challenged by a movement by some companies to move their manufacturing back to the U.S. (i.e., by "re-shoring"), or by moving it to Mexico (i.e., by "near-shoring"). Over the past four years (2010 – 2014), the number of manufacturing jobs in the U.S. has started to rise for the first time after plunging for 15 years, edging up by half a million to just above 12 million now.

Why are some firms placing a huge bet on re-shoring? How many manufacturers are likely to follow this movement? Does offshore-outsourcing still make good business sense? And, what are the consequences of changes in sourcing strategy for a firm's products, performance and operational flexibility? Motivated by an ongoing survey effort<sup>1</sup> that is designed to explore

---

<sup>1</sup>The survey is an joint research project among Wharton, MIT, Shanghai Jiao Tong University and Shanghai Institute of Foreign Trade.

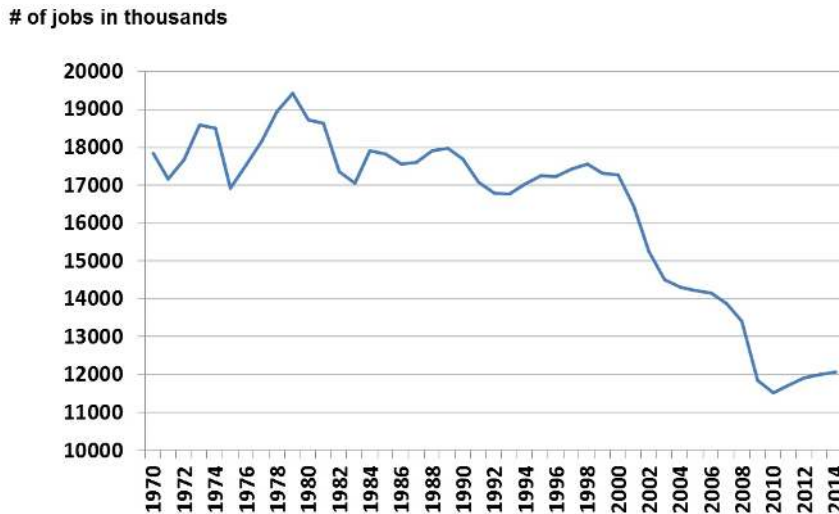


Figure 8: U.S. manufacturing employment. Data: U.S. Bureau of Labor Statistics, <http://www.bls.gov/iag/tgs/iag31-33.htm#workforce>

answers to the questions noted above, we develop a modeling framework in this paper to study the drivers of companies' global sourcing location decisions.

#### 4.1.1. Evidence and Competing Theories on Supply Chain Re-Structuring

In a review we recently conducted, we found over 50 cases where major U.S. and global companies have announced significant re-structuring decisions for their global supply chain in the past three years. Among them, 19 companies reported increased off-shoring by shifting production to an off-shore location. In contrast, 19 other companies re-shored by bringing production back to their home country. Well-known examples include decisions by Apple to invest \$100 million in producing some of its Mac computers in the U.S.; and General Electric to invest \$1 billion into domestic appliances manufacturing capabilities at Louisville, Kentucky. There is also evidence that some companies are near-shoring by bringing production to a country that is closer to major customers, and that others are investing in manufacturing technology (e.g., the adoption of robots).

Interestingly, a number of Chinese manufacturing firms have also opened factories in the U.S. to be near their U.S. market. Examples include Lenovo opening its first U.S. personal

computer production facility in Whitsett, N.C. (June, 2013); and Fuling Plastics, which makes plastic cutlery for America's top fast food chains including McDonald's, KFC, and Subway, opening a production facility in Pennsylvania's Lehigh Valley (May 2014).

A popular argument to explain the re-shoring phenomenon is called the "Tipping Point Theory" (e.g., see Sirkin et al. (2011, 2012)), which comes from the observation that wages in low labor-cost countries are rising at faster rate than those at developed countries. For example, Chinese wages are growing at a rate of 15 percent annually, as opposed to 2 percent in the U.S. As a result, the labor-cost advantage of producing a product in China is diminishing. After factoring in increased ocean freight rates, strengthened Chinese currency and lower energy cost in the U.S., the landed cost advantage of manufacturing in China versus the U.S. could be eliminated in a few years. The Tipping Point Theory thus argues that "when the landed cost advantage falls below a critical level, most manufacturers will re-shore to the U.S."

While labor costs are a significant factor that firms need consider when making a global sourcing decision, they must also consider other aspects of doing business overseas, such as the time to market, foreign exchange rates, ownership of intellectual property, local content requirements as well as other cost components in addition to labor. As a result, a number of other competing theories to explain global sourcing decisions have emerged.

For example, Markides and Berg (1988) and Pisano and Shih (2012) believe that "outsourcing of manufacturing leads to a loss of capability to develop new products and to adopt new technologies", while others have argued "technology developments such as enhanced automation reduce the impact of a foreign labor-cost advantage in making sourcing decisions"; "government policies have a major impact on sourcing decisions"; and "the movement to a service dominated economy has reduced the importance of manufacturing and will make it more difficult to off-shore manufacturing", etc.

There is a national debate among the media, the Federal Government, academia and in-

dustry as to which of these theories is right or wrong. In this paper, we introduce a comprehensive model that incorporates perspectives over the entire life cycle of a product, i.e., product design, manufacturing and after-sale service support, in order to examine conditions that support the validity of these theories. As the existing global sourcing literature has mainly focused on the impact of costs on sourcing decisions, our work, to the best of our knowledge, is the first to conduct a life cycle analysis in the context of global sourcing strategy.

The remainder of this paper is structured as follows. Section 2 reviews the related literature. Section 3 describes a basic single-market single-plant model and sheds light on different trade-offs present among the elements of a life cycle analysis, when making sourcing decisions. Section 4 presents a number of extended models and associated structural results. These include incorporation of technology investment decisions, multiple demand markets and multiple plant locations into the basic model. Section 5 provides numerical examples that illustrate managerial insights that can be derived from the optimal solution to various models based on model specifications associated with a number of industries. Section 6 concludes the paper with suggestions for continuing analytical and empirical research.

## 4.2. LITERATURE REVIEW

Early research on the global plant location problem appears in Hodder and Jucker (1982, 1985), Hodder and Dincer (1986), Breitman and Lucas (1987). Cohen et al. (1989), and Cohen and Lee (1989). These papers extended basic supply chain network models by incorporating issues such as corporate taxes, tariffs, and duties. Here, we focus on a representative publication: Cohen and Lee (1989) evaluated a series of policy options that the Apple computer company might use to establish its global manufacturing strategy. Their mathematical programming model is capable of capturing a large number of constraints on demand, sourcing, interplant transshipments, taxation and tariffs, and the objective function is to maximize total global after-tax profits.

Many subsequent papers that developed global supply chain design models examined the impact of factors such as the uncertain exchange rates, as well as transfer pricing, financial subsidies and local content rules. For example, Kogut and Kulatilaka (1994) used a dynamic programming model to study the value of real option in a multinational operating network due to uncertain currency exchange rates. Huchzermeier and Cohen (1996) also developed a modeling framework to solve the real option valuation problem. Their framework integrated a production-distribution flow model by looking at a three-tier supply chain that contains suppliers, production sites, and market regions. Munson and Rosenblatt (1997) incorporated local content rules into the plant location problem. Other representative works include Arntzen et al. (1995), Gutierrez and Kouvelis (1995), Canel and Khumawala (1996), Kouvelis and Gutierrez (1997), Dasu and de La Torre (1997), Vidal and Goetschalckx (2001), Kazaz et al. (2005), Goh et al. (2007), Robinson and Bookbinder (2007). We refer interested readers to Cohen and Mallik (1997), Cohen and Huchzermeier (1999), Meixell and Gargeya (2005) and Bookbinder and Matuk (2009) for further reviews.

Techniques used in the papers mentioned above are mainly mathematical or stochastic programming. A more recent stream of related papers use stylized models to explore structural results and draw managerial insights pertaining global supply chain network management. Our paper belongs to this category which includes: Rosenfield (1996) who examined location and capacity strategies when exchange rates are uncertain, Hadjinicola and Kumar (2002) who incorporated marketing functions on top of manufacturing into a global supply chain model, Nagurney et al. (2003) who developed a network equilibrium model for manufacturers, retailers and consumers in a global supply chain context, Lu and Van Mieghem (2009) who studied multiplant network configurations for off-shoring products with component commonality, and Ang et al. (2014b) and Simchi-Levi et al. (2014) who investigated global supply chain disruption issues in wake of the 2011 Tōhoku earthquake and tsunami.

Literature on global sourcing is growing. In the international business literature, MacCormack et al. (1994), Ferdows (1997) and Farrell (2004, 2005) have pointed out competitive

advantages in setting up foreign factories, including low cost direct labor, capital subsidies, tariff concessions, and access to overseas markets, etc. In contrast, Markides and Berg (1988) and Pisano and Shih (2012) argue that it is only a myopic tactical move because off-sourcing activities transfers technology which could eventually put the company out of business. We consider both sides of this story in our model, and are able to demonstrate the key trade-offs in the paper. On the other hand, Ghelfi (2011) discussed issues on ownership of intellectual property for the off-sourcing of processes, and Feng and Lu (2011) provided an overview of ODM practices to Asia (i.e., off-sourcing of both design and manufacturing by original design manufacturers).

Hsu et al. (2014) looked at the impact of foreign tax credit on global sourcing quantity decisions for a multinational firm. In contrast, this paper investigates global sourcing location decisions, whether it is for “re-shoring”, “near-shoring” or continuing to “off-shore”. Given the increasing labor costs in emerging markets, oil price volatility and technology advances, the discussion on “re-shoring” and the debate on whether manufacturing jobs would return to U.S. has intensified in recent years, e.g., see Sirkin et al. (2011, 2012) and Simchi-Levi et al. (2012) in the popular business press. Using a survey, Simchi-Levi (2012) reported that 33.6% of the U.S. companies in its sample are “considering” bringing manufacturing back to the U.S., while 15.3% are “definitely” planning to re-shore to the U.S.

Similar to Pyke (2007) and Kumar and Kopitzke (2008), the goal of this paper is to provide a framework for analyzing the global sourcing location decision process which would be relevant to companies who are in the midst of global supply chain re-structuring transformation. Our model framework however is based on a specific mathematical model that can be used to quantify the impact of the underlying drivers of such decision while the previous two papers offer a qualitative approach. Wu and Zhang (2014) used a game theoretical approach to study manufacturers’ sourcing location decisions under correlated demands. While their paper focuses on off-sourcing supply responsiveness, we take a life-cycle analysis approach to balance issues that impenetrate product design, manufacturing, delivering,

and after-sale service support.

Finally, empirical studies have been carried out relating to global supply chain sourcing to provide additional insight, e.g., see Brush et al. (1999), MacCarthy and Atthirawong (2003), Li et al. (2008), Massini et al. (2010) and Jain et al. (2013). More relevant to our focus, Srivastava et al. (2008) examines the fraction of activity that is off-shored; Lewin et al. (2009) and Roza et al. (2011) studies the impact of skilled labor and firm size on off-shoring strategy, respectively; Hutzschenreuter et al. (2011) looks at a firm's decision on internal versus external governance mode for off-shoring activities.

### 4.3. MODEL

We begin with a base model that considers a scenario where a firm is about to introduce a product for sale to a particular end demand market. The firm must make a sourcing location decision to manufacture this product in one country in the world, for example, in the U.S., China or Mexico. Thus, we consider a single-plant single-market model. We will extend this model to include multiple demand markets or multiple plants in Section 4.

The firm has determined that it will introduce one future (upgraded) generation of the product, but products of both the current and future generations will be manufactured in the same plant that it has selected at the time that the first-generation product is introduced. We thus consider a two-period model: one for the current generation and one for the next generation.

Similar to Guajardo et al. (2014), a product of either generation in our model can be viewed as a product/service bundle that specifies the level of product quality and service quality that is offered to the market. For example, when a consumer purchases a new car, the physical performance of the car together with the manufacturer's warranty and associated maintenance service that are included with the purchase forms a product/service bundle.

Between launches of the two generations of the product bundle, the firm will be able to

improve both the product quality and the service quality by investing in product design enhancements and in service support resources. Examples of design upgrade effort include the adoption of new materials or product technology, the hiring of additional design engineers, or surveying consumers to determine their level of satisfaction with the current product generation. Examples of service improvement effort include the purchasing of more spare parts, increasing the capacity of repair depots, the hiring of additional customer representatives, or improving service processes. In the case of Tesla Motors, for example, the electric car maker is well-known for its efforts in both building better cars (product quality) and a stronger network of supercharger stations around the world (service quality).

We note, however, that the returns on quality enhancement investment will depend on the location of the plant that is used to produce the product. Specifically, we assume that the location of the research and development center of the firm (usually at the headquarters) is fixed throughout the two periods of our model. As Pisano and Shih (2012) note, product and process innovations are intertwined and process engineering expertise depends on daily interactions with manufacturing. Therefore, if the plant location is closer to the R&D center, there will be presumably more return on product quality enhancement investment dollars spent on re-design and innovation. Pufall (2013) shows empirically the positive relationship between the proximity of engineers of the suppliers and the ramp-up performance of a product.

On the other hand, if the plant location is closer to the end demand market, one anticipates more return on service quality enhancement investment dollars spent on efforts to improve service due to quicker response times, easier management coordination, highly spare parts availability, and lower transaction costs, etc.

The firm conducts a life cycle cost-benefit analysis to choose a plant location up front to maximize its expected after-tax profit generated by selling both generations of the product on the given end market over two periods and must consider the impact of its plant location decision through the product design, manufacturing and after-sale service support phases.



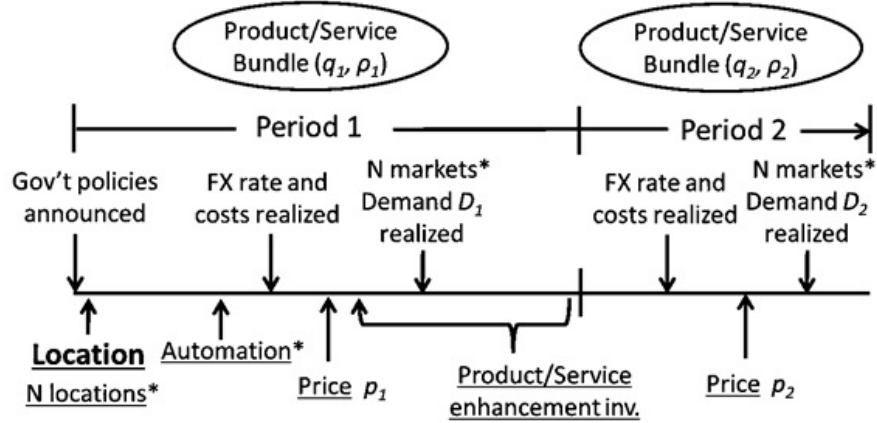


Figure 9: Sequence of events in the two-period model. Decisions are underlined. Features in the extended models in Section 4 are denoted with \*.

In our model framework landed costs, which are the fully loaded costs of producing and delivering one product unit to an end market, will be random. In particular, these costs will be affected by fluctuations in labor costs, foreign exchange rates, energy costs, and import/export tariffs, etc. In each of the two periods, the firm will make price decisions. Demand for the product will then be determined by factors such as product quality, service quality, the price of the product, additional exogenous market factors, and random noise. Finally, the local government determined by the plant location offers a one-time subsidy and sets a corporate tax rate in order to attract the firm to open the plant in their jurisdiction.

We describe the sequence of events below and in Figure 9:

- 1) The firm selects a host country to build one plant, and the plant will be used to manufacture the product for two periods.
- 2) The firm receives a one-time subsidy from the local government. A pre-determined tax rate will apply profits generated in both periods. The amount of the subsidy and the tax rate depend on the location of the plant that the firm has selected.
- 3) Period 1 starts. The landed cost for period 1 (which includes the raw material and manufacturing cost of the item, all logistics and shipping costs, customs duties and tariffs)

is realized and converted to the numeraire (U.S.) currency according to the realized foreign exchange rate.

4) The firm launches the product/service bundle that comes with some pre-determined level of product quality and service quality, and selects a price for the product.

5) The market demand for period 1 is then realized based on the product quality, the service quality as well as the price of the product. The firm generates after-tax profit in period 1 for products sold in that period.

6) Also in Period 1, the firm invests in product design and service improvements, in preparation for launching the next-generation version of the product. These enhancement efforts will determine the product quality and the service quality for the product/service bundle in the second period.

7) Period 2 starts, and the landed cost in U.S. dollars for period 2 is realized.

8) The firm launches the updated product/service bundle with a new price decision.

9) The market demand for period 2 is realized based on the product quality, the service quality and the price of the product in this period, and the firm generates after-tax profit for product sold in the second period.

Note that, although the firm can announce a price in each period in response to the realized landed cost, the enhancement of product/service quality cannot be carried out instantaneously. Therefore, we assume that effort made in Period 1 will lead to changes to product and service quality in Period 2 only.

#### *4.3.1. Notation*

For the basic (one-plant one-end-market) model, we will use superscript  $i$  for the plant location, and subscript  $j \in \{1, 2\}$  for the period. For example,  $x_2^{China}$  represents a particular variable,  $x$ , realized in period 2, if the plant was located in China. With this notation,  $x_1^A$ ,

$x_2^A$ ,  $x_1^B$  and  $x_2^B$  are all different. The firm makes the plant location decision  $i$ , for a given end market and a given location for its R&D center. Conditional on  $i$ , it also makes the following decisions in the two periods:

- 1) the price of the product in Period 1, namely  $p_1^i$ ;
- 2) the investment  $r^i$  for the improvement of the product quality;
- 3) the investment  $s^i$  for the improvement of the service quality;
- 4) the price of the product in period 2, namely  $p_2^i$ .

The exogenous parameters are:

- $f^i$ : the fixed cost of opening and operating a new plant or for using an existing plant in location  $i$ ;
- $q_1^i \equiv q_1$ : the initial product quality introduced to the end market which is given and assumed to be same for all plant location  $i$ ;
- $\rho_1^i \equiv \rho_1$ : the initial service quality introduced to the end market which is given and assumed to be same for all plant location  $i$ ;
- $D_j$ : market demand which is a function of
- $A_j$ : random intercept of the demand curve (to capture demand size and shocks)
- $a_q$ : demand sensitivity to the product quality, assumed to be same for both periods
- $a_\rho$ : demand sensitivity to the service quality, assumed to be same for both periods
- $a_p$ : demand sensitivity to the price, assumed to be same for both periods
- $C_j^i$ : random landed cost associated with producing one unit of the product;
- $g^i$ : one time subsidy from government  $i$ ;
- $v^i$ : tax rate offered by country  $i$ , assumed to be same for both periods;
- $k_r^i$ : return rate of the investment on product design enhancements;
- $k_s^i$ : return rate of the investment on product service support enhancements.

The endogenous variables are:

- $q_2^i$ : new product quality in Period 2 due to enhancement effort  $r^i$ ;
- $\rho_2^i$ : new service quality in Period 2 due to enhancement effort  $s^i$ ;

### 4.3.2. Formulation

We assume that market demand increases with higher product quality and higher service quality but decreases with higher price. The demand function is given as the following:

$$D_j^i = A_j + a_q \cdot q_j^i + a_\rho \cdot \rho_j^i - a_p \cdot p_j^i \quad (4.1)$$

where  $A_j^i + a_q \cdot q_1 + a_\rho \cdot \rho_1$  is sufficiently larger than  $a_p \cdot C_j^i$  (almost surely) to ensure positive demand and revenue for the firm.

Next, we assume that investments on product design and service support enhancements exhibit diminishing marginal returns. That is, if  $f_r^i(r^i) \triangleq q_2^i - q_1$  and  $f_s^i(s^i) \triangleq \rho_2^i - \rho_1$ , then  $\frac{\partial f_r^i}{\partial r^i} > 0$ ,  $\frac{\partial^2 f_r^i}{\partial (r^i)^2} < 0$  and  $\frac{\partial f_s^i}{\partial s^i} > 0$ ,  $\frac{\partial^2 f_s^i}{\partial (s^i)^2} < 0$ . For tractability of the solutions, in this paper, we assume that

$$\begin{aligned} q_2^i - q_1 &= k_r^i \cdot \sqrt{r^i}; \\ \rho_2^i - \rho_1 &= k_s^i \cdot \sqrt{s^i}. \end{aligned}$$

(Insights found in this paper generalize to more complex function forms but closed form solutions may no longer be available.) The location-dependent parameters,  $k_r^i$  and  $k_s^i$ , reflect the return rates of the investment on product and service quality enhancements. We assume they are not too big in that eventually the firm loses money if it invests  $r \rightarrow \infty$  or  $s \rightarrow \infty$ . Specifically, we require  $(a_q k_r^i)^2 + (a_\rho k_s^i)^2 < 4a_p$ .

As the firm's goal is to maximize its expected after-tax future profit of selling both generations of the product at the end market over two periods, its optimal location decision  $i^*$  is given by

$$i^* = \arg \max_i \left\{ \max_{p_1^i, r^i, s^i, p_2^i} \left\{ \mathbb{E}_{C_1^i, A_1, C_2^i, A_2} \sum_{j=1}^2 (1 - v^i)(p_j^i - C_j^i) D_j^i - f^i + g^i - r^i - s^i \right\} \right\} \quad (4.2)$$

We assume the random intercept of the demand function  $A_j^i$  is independent of the random landed cost  $C_j^i$  for  $j \in \{1, 2\}$ . Then, expression (4.2) in  $i^*$  becomes

$$\arg \max_i \left\{ \max_{p_1^i, r^i, s^i, p_2^i} \left\{ \frac{\mathbb{E} \mathbb{E}(1-v^i)(p_1^i - C_1^i)D_1^i}{C_1^i A_1} + \frac{\mathbb{E} \mathbb{E}(1-v^i)(p_2^i - C_2^i)D_2^i}{C_2^i A_2} - r^i - s^i \right\} - f^i + g^i \right\} \quad (4.3)$$

#### 4.3.3. Optimal Solutions and Structural Results

The optimal decisions to the two-period basic model are given in the following theorem.

**Theorem 12** (*Optimal Decisions.*) *The optimal sourcing location decision is given by*

$$i^* = \arg \max_i \left\{ \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_1^i)]^2 + \frac{(1-v^i)a_p \cdot \text{Var}(C_1^i)}{4} \right. \\ \left. + \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p^i \cdot \mathbb{E}(C_2^i)]^2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} + \frac{(1-v^i)a_p \cdot \text{Var}(C_2^i)}{4} - f^i + g^i \right\} \quad (4.4)$$

where the optimal Period-1 price decision contingent on the realized landed cost  $c_1^i$  is given by

$$p_1^{i*}(v^i, c_1^i) = \frac{1}{2a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 + a_p \cdot c_1^i], \quad (4.5)$$

the optimal investment decisions on product and service quality enhancement are given by

$$r^{i*}(v^i) = (a_q k_r^i)^2 \left\{ \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_2^i)]}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right\}^2 \quad (4.6)$$

$$s^{i*}(v^i) = (a_\rho k_s^i)^2 \left\{ \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_2^i)]}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right\}^2, \quad (4.7)$$

and the optimal Period-2 price decision contingent on the realized landed cost  $c_2^i$  is given by

$$p_2^{i*}(v^i, c_2^i) = \frac{2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} [\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \\ + \frac{1}{2a_p} \frac{4a_p - 2(1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} a_p \cdot c_2^i, \quad (4.8)$$

Proof: In order to solve for the optimal decisions to the two-period model, we proceed by backward induction.

(i) The decision  $p_2^i$  is made by the firm given  $i$ ,  $v^i$ ,  $q_2^i$ ,  $\rho_2^i$  with the realized landed cost in Period 2,  $c_2^i$ . The optimal price for the second period is thus given by

$$\begin{aligned}
& p_2^{i*}(v^i, q_2^i, \rho_2^i, c_2^i) \\
& \triangleq \arg \max_{p_2^i} \{ \mathbb{E}_{A_2} (1 - v^i)(p_2^i - c_2^i) D_2^i \} \\
& = \arg \max_{p_2^i} \{ \mathbb{E}_{A_2} (p_2^i - c_2^i)(A_2 + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot p_2^i) \} \\
& = \arg \max_{p_2^i} \{ (p_2^i - c_2^i)[\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot p_2^i] \} \\
& = \frac{1}{2a_p} [\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i + a_p \cdot c_2^i] \tag{4.9}
\end{aligned}$$

Since  $\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i > a_p \cdot c_2^i$ , it is clear that the optimal price decision in the second period  $p_2^{i*}(v^i, q_2^i, \rho_2^i, c_2^i)$  will be bigger than the realized landed cost  $c_2^i$ , and increases in it. The optimal profit (which is contingent on the realized landed cost) is given by

$$\begin{aligned}
& \pi_2^{i*}(v^i, q_2^i, \rho_2^i, c_2^i) \\
& \triangleq \max_{p_2^i} \{ \mathbb{E}_{A_2} (1 - v^i)(p_2^i - c_2^i) D_2^i \} \\
& = \mathbb{E}_{A_2} (1 - v^i)(p_2^i - c_2^i) D_2^i \big|_{p_2^i = p_2^{i*}(v^i, q_2^i, \rho_2^i, c_2^i)} \\
& = \frac{(1 - v^i)}{4a_p} [\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot c_2^i]^2
\end{aligned}$$

(ii) Decisions on the optimal amount of investment to be made for product design and for service support enhancements are made given  $i$ , and  $v^i$  but before Period 2. Therefore, the decisions are

$$\begin{aligned}
& (r^{i*}(v^i), s^{i*}(v^i)) \\
& \triangleq \arg \max_{(r^i, s^i)} \mathbb{E} \pi_2^{i*}(v^i, q_2^i, \rho_2^i, C_2^i) - r^i - s^i \\
& = \arg \max_{(r^i, s^i)} \mathbb{E} \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot C_2^i]^2 - r^i - s^i \\
& = \arg \max_{(r^i, s^i)} \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot \mathbb{E}(C_2^i)]^2 + \frac{(1-v^i)a_p \text{Var}(C_2^i)}{4} - r^i - s^i \\
& = \arg \max_{(r^i, s^i)} \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_2) + a_q \cdot q_2^i + a_\rho \cdot \rho_2^i - a_p \cdot \mathbb{E}(C_2^i)]^2 - r^i - s^i
\end{aligned}$$

Since  $q_2^i - q_1 = k_r^i \cdot \sqrt{r^i}$  and  $\rho_2^i - \rho_1 = k_s^i \cdot \sqrt{s^i}$ , plugging in gives

$$\begin{aligned}
& (r^{i*}(v^i), s^{i*}(v^i)) \\
& = \arg \max_{(r^i, s^i)} \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_2) + a_q \cdot (q_1 + k_r^i \sqrt{r^i}) + a_\rho \cdot (\rho_1 + k_s^i \sqrt{s^i}) - a_p \cdot \mathbb{E}(C_2^i)]^2 - r^i - s^i
\end{aligned}$$

Solving first and second order conditions leads to

$$\frac{r^{i*}(v^i)}{(a_q k_r^i)^2} = \frac{s^{i*}(v^i)}{(a_\rho k_s^i)^2}. \tag{4.10}$$

In particular, we have

$$\begin{aligned}
r^{i*}(v^i) &= (a_q k_r^i)^2 \left\{ \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_2^i)]}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right\}^2 \\
s^{i*}(v^i) &= (a_\rho k_s^i)^2 \left\{ \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_2^i)]}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right\}^2
\end{aligned}$$

Plugging in  $r^{i*}(v^i)$  and  $s^{i*}(v^i)$  into (4.9), we have

$$p_2^{i*}(v^i, c_2^i) = \frac{2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} [\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1]$$

$$+ \frac{1}{2a_p} \frac{4a_p - 2(1-v)[(a_q^i k_r^i)^2 + (a_\rho^i k_s^i)^2]}{4a_p - (1-v)[(a_q^i k_r^i)^2 + (a_\rho^i k_s^i)^2]} a_p \cdot c_2^i.$$

Through some tedious algebraic operations, the optimal expected profit for the second period can be solved as

$$\begin{aligned} \pi_2^{i*}(v^i) &\triangleq \max_{(r^i, s^i)} \mathbb{E}_{C_2^i} \pi_2^{i*}(f^i, g^i, v^i, q_2^i, \rho_2^i, C_2^i) - r^i - s^i \\ &= \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p^i \cdot \mathbb{E}(C_2^i)]^2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} + \frac{(1-v^i)a_p \cdot \text{Var}(C_2^i)}{4} \end{aligned} \quad (4.11)$$

(iii) The price decision for the first period to be made by the firm is conditional on  $i$ ,  $v^i$  and the realized Period-1 landed cost  $c_1^i$ . It is given by

$$\begin{aligned} p_1^{i*}(v^i, c_1^i) &\triangleq \arg \max_{p_1^i} \{ \mathbb{E}_{A_1} (1-v^i)(p_1^i - c_1^i) D_1^i \} \\ &= \arg \max_{p_1^i} \{ \mathbb{E}_{A_1} (p_1^i - c_1^i) [A_1 + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot p_1^i] \} \\ &= \arg \max_{p_1^i} \{ (p_1^i - c_1^i) [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c_1^i] \} \\ &= \frac{1}{2a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 + a_p \cdot c_1^i] \end{aligned}$$

which is greater than and increases in the realized cost  $c_1^i$ . The optimal profit in Period 1 contingent on the realized the landed cost is

$$\begin{aligned} \pi_1^{i*}(v^i, c_1^i) &\triangleq \max_{p_1^i} \{ \mathbb{E}_{A_1} (1-v^i)(p_1^i - c_1^i) D_1^i \} \\ &= \mathbb{E}_{A_1} (1-v^i)(p_1^i - c_1^i) D_1^i \Big|_{p_1^i = p_1^{i*}(v^i, c_1^i)} \\ &= \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c_1^i]^2 \end{aligned}$$



Therefore, the expected profit to be received by the firm in Period 1, conditional on producing the product in location  $i$ , is given by

$$\begin{aligned}
\pi_1^{i*}(v^i) &\triangleq \mathbb{E}_{C_1^i} \pi_1^{i*}(v^i, C_1^i) \\
&= \mathbb{E}_{C_1^i} \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot C_1^i]^2 \\
&= \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_1^i)]^2 + \frac{(1-v^i)a_p \cdot \text{Var}(C_1^i)}{4}. \quad (4.12)
\end{aligned}$$

Using (4.11) and (4.12), we re-write (4.3), the firm's optimal plant location decision  $i^*$  as

$$\begin{aligned}
i^* &= \arg \max_i \{ \pi_1^{i*}(v^i) + \pi_2^{i*}(v^i) - f^i + g^i \} \\
&= \arg \max_i \left\{ \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_1^i)]^2 + \frac{(1-v^i)a_p \cdot \text{Var}(C_1^i)}{4} \right. \\
&\quad \left. + \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p^i \cdot \mathbb{E}(C_2^i)]^2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} + \frac{(1-v^i)a_p \cdot \text{Var}(C_2^i)}{4} - f^i + g^i \right\}
\end{aligned}$$

and therefore the theorem follows.  $\square$

A number of interesting structural results emerge from Theorem 12:

**Marginal Investment Returns.** It is clear from (4.10) that the amount to be invested in product design enhancement versus service support enhancement (i.e.,  $r^*$  versus  $s^*$ ) will depend on the overall investment return rates, which is the product of the impact of the investment dollars on product/service quality enhancement (i.e.,  $k_r^i, k_s^i$ ) times the impact of product/service quality enhancement on boosted demand (i.e.,  $a_q, a_\rho$ ). Furthermore, from (4.6) and (4.7), we see that if the investment does not generate any return (i.e.,  $k_r^i = 0$  or  $k_s^i = 0$ ), or if the enhancement of quality derived from the investment does not stimulate demand (i.e.,  $a_q = 0$  or  $a_\rho = 0$ ), then the investment should not be considered (i.e.,  $r^* = 0$  or  $s^* = 0$ ).

**Option Value.** From (4.11) and (4.12), it is evident that all else being equal, the expected profit for each period increases if the corresponding landed cost becomes more volatile. More precisely, when we compare two sets of distributions on  $C_j^i$  with the same mean for fixed  $j \in \{1, 2\}$ , a larger variance would lead to a higher expected profit in Period  $j$ . This result is due to the fact that the firm can freely adjust its price decision according to the realized cost. Similar observation has been made with other context, e.g., see Ho et al. (1998).

**Value of Quality Enhancement.** We note that having the option to invest in design and service enhancements always leads to higher expected profit. As the return rate increases ( $k_r^i \uparrow$  or  $k_s^i \uparrow$ ), the expected profit also increases. In fact, if the demand and cost distributions in both periods are identical, i.e.  $A_1 = A_2 \triangleq A$  and  $C_1^i = C_2^i \triangleq C^i$ , we can directly compare the expected profits from the two periods:

$$\begin{aligned} \pi_1^{i*}(v^i) &= \frac{(1-v^i)}{4a_p} [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C^i)]^2 + \frac{(1-v^i)a_p \cdot \text{Var}(C^i)}{4} \\ &< \frac{(1-v^i)[\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p^i \cdot \mathbb{E}(C^i)]^2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} + \frac{(1-v^i)a_p \cdot \text{Var}(C^i)}{4} = \pi_2^{i*}(v^i) \end{aligned}$$

The positive difference,  $\pi_2^{i*}(v^i) - \pi_1^{i*}(v^i)$ , captures the overall expected investment return from enhancing the product quality and service quality. The return increases in the mean demand intercept, the initial given product and service quality, and decreases in the mean landed cost.

**Tipping Point Theory.** Let us suppose that the end market being served is in the U.S., and consider plant location scenarios for  $i = U.S.$  and  $i = China$ . The mainstream media and consultants have argued that since the Chinese wages are growing at more than 15 percent annually, compared to 2 percent in the U.S., the landed cost advantage in producing a product in China as opposed to the U.S. is falling. And when this advantage falls below some critical level, companies will choose to re-shore to the U.S. (i.e., the Tipping Point Theory).

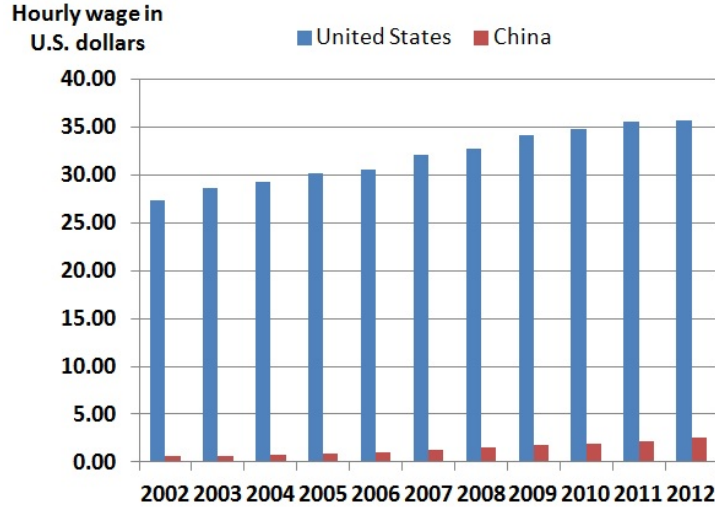


Figure 10: Comparison of hourly wage received by manufacturing workers in the U.S. and China. Data is retrieved from the website of U.S. Bureau of Labor Statistics, except that the last three data points for the Chinese wage are estimated by extrapolation (not available from the website).

Figure 10 displays the hourly wage for manufacturing workers in the U.S. and in China from years 2000 to 2012. Note that the absolute wage difference is currently widening in spite of the inflation rate differences, but it will eventually come down if the wage inflation trends for both countries continue to hold.

For the purpose of this discussion, let us ignore the variance on landed costs, by assuming that the firm can perfectly foresee the landed cost from producing the product in China and in the U.S. for Period 1 and for Period 2, i.e.,  $C_1^{China} \equiv c_1^{China}$ ,  $C_1^{U.S.} \equiv c_1^{U.S.}$ ,  $C_2^{China} \equiv c_2^{China}$ ,  $C_2^{U.S.} \equiv c_2^{U.S.}$ . From (4.4), it is evident that, for fixed U.S. landed costs  $c_1^{U.S.}$ ,  $c_2^{U.S.}$ , higher Chinese landed costs  $c_1^{China}$ ,  $c_2^{China}$  drive the decision to favor manufacturing in the U.S., and vice versa.

However, the validity of the Tipping Point Theory has to be tested when the landed costs in the U.S. and in China alter from year to year simultaneously, as captured in Figure 10. We will illustrate by numerical examples in Section 5 that the argument actually may or may not hold depending on the percentage of labor cost component in the landed cost.

**Country-of-Origin Effect.** Note that we did not specifically model the country of origin effect, i.e., we assume that a customer has the same likelihood of buying two items if they have the same product quality, service quality and price, but differ only in the country of origin. However, the existing analysis can be easily extended to model a country-of-origin effect, by making the parameters  $A_j, a_q, a_\rho, a_p$  all dependent on the plant location  $i$ .

A closer look at expression (4.4) reveals the following two corollaries:

**Corollary 2** (*Government Policies.*) *The government can have a major impact on the firm's decision by lowering the tax rate  $v^i$  and by providing more subsidy or training grant in  $g^i$ .*

Corollary 2 reveals that sufficient government support from location  $i$  can compensate for its higher landed cost or lower investment return rates from quality enhancement (i.e., inefficiency of having a distant plant at location  $i$ ), and results in optimal expected profit for the firm.

**Corollary 3** (*Cost vs. Quality.*) *When the firm incurs a lower landed cost from producing the product in country  $i = A$ , it may still be optimal to build the plant in another country  $i = B$  if  $k_r^B > k_r^A$  and/or  $k_s^B > k_s^A$ .*

Corollary 3 reveals that the firm has to balance cost and quality in making the optimal global sourcing location decision. Especially, it may be optimal for the firm to select a plant with higher landed cost if doing so brings more capability or competitive advantage for developing new products or for enhancing the quality service.

#### 4.4. EXTENDED MODELS

In this section, we present three extensions to the basic life-cycle analysis that was formulated in Section 3. First, we consider technology decisions for the firm. Second, we allow the firm to serve multiple demand markets. And lastly, we study the firm's strategy when it is allowed to build capacities in more than one plant upfront.

#### 4.4.1. Technology Decisions

In this subsection, we study a variant of the one-plant one-end-market model introduced in Section 3, by allowing the firm to invest in technology (e.g., automation and robotic arms) at the time that it builds the plant. But like the plant, the technology once built will be used to produce the products in both periods. To derive comparative results, we assume throughout this subsection that the distributions of the random demand curve intercepts and those of the landed costs stay the same through two periods, i.e.,  $A_1 = A_2 \triangleq A$ , and  $C_1^i = C_2^i \triangleq C^i$  for all  $i$ . The rest of the assumptions are the same as the basic model described in Section 3.

Once the firm decides on an amount of upfront investment  $t$  for technology (or no investment if  $t = 0$ ), the selected technology will help to reduce future costs. For example, having a higher level of automation reduces the amount of labor required to produce the product, and thus it will reduce the total labor cost; employing an RFID (Radio Frequency Identification) system can save logistic costs for the firm.

If location  $i$  has been selected to build the plant, we assume that the technology decision  $t^i$  made upfront will reduce realized landed cost from  $c_j^i$  to  $c_j^i(1 - \delta_t^i)$  where  $\delta_t^i \in [0, 1)$  is the percentage of cost reduction due to the technology and is a function of  $t$ . The cost reduction function is subject to diminishing returns, i.e.,  $\frac{\partial \delta_t^i}{\partial t^i} > 0$  and  $\frac{\partial^2 \delta_t^i}{\partial (t^i)^2} < 0$ . For simplicity, we assume in the analysis below that  $1 - \delta_t^i = \frac{1}{1 + k_t^i \cdot t^i}$  for  $t^i \geq 0$  where  $k_t^i$  is a resilience parameter for the technology decision. (A larger value of  $k_t^i$ , which leads to higher  $\delta_t^i$  for any fixed  $t^i$ , corresponds to a better return on the investment in technology.)

**Lemma 11** (*Technology Investment Decision.*) *Given the plant location  $i$ , it is optimal for the firm either not to invest in technology (i.e.,  $t^{i*} = 0$ ), or there exists a unique optimal investment level  $t^{i*} > 0$ . Furthermore,  $t^{i*}$  increases in the mean landed cost but decreases in its variance.*

Proof: For a plant at location  $i$  with technology investment  $t^i$ , the expected after-tax two-

period profit for the firm,  $\pi^i(t^i)$ , can be derived from expression (4.4), as:

$$\begin{aligned}\pi^i(t^i) &= \frac{(1-v^i)}{4a_p} [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot \mathbb{E}(C_1^i)(1-\delta_t)]^2 + \frac{(1-v^i)a_p \cdot \text{Var}(C_1^i)(1-\delta_t)^2}{4} \\ &+ \frac{(1-v^i)[\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p^i \cdot \mathbb{E}(C_2^i)(1-\delta_t)]^2}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} + \frac{(1-v^i)a_p \cdot \text{Var}(C_2^i)(1-\delta_t)^2}{4} \\ &- f^i + g^i - t^i\end{aligned}\quad (4.13)$$

It follows from (4.13) that

$$\begin{aligned}\frac{\partial \pi^i(t^i)}{\partial t_i} &= \frac{2(1-v^i)}{4a_p} \{ [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C_1^i) - (a_p)^2 \cdot [(\mathbb{E}(C_1^i))^2 + \text{Var}(C_1^i)](1-\delta_t) \} \cdot \frac{\partial \delta_t^i}{\partial t^i} \\ &+ \frac{2(1-v^i) \{ [\mathbb{E}(A_2) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C_2^i) - (a_p)^2 \cdot [(\mathbb{E}(C_2^i))^2 + \text{Var}(C_2^i)](1-\delta_t) \}}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \cdot \frac{\partial \delta_t^i}{\partial t^i} - 1\end{aligned}\quad (4.14)$$

Since  $\frac{\partial \delta_t^i}{\partial t^i} \rightarrow 0$ , we have  $\frac{\partial \pi^i(t^i)}{\partial t_i} \rightarrow -1$ . Thus, eventually, an additional dollar invested in technology barely generates a more reduced cost, and thus is purely an extra dollar in expense.

When  $A_1 = A_2 \triangleq A^i$ ,  $C_1^i = C_2^i \triangleq C^i$ , the first order condition in (4.14) can be simplified into

$$\begin{aligned}\frac{\partial \pi^i(t^i)}{\partial t_i} &= 2 \left\{ \left[ \frac{(1-v^i)}{4a_p} + \frac{2(1-v^i)}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right] \times \right. \\ &\left. \{ [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C^i) - (a_p)^2 \cdot [(\mathbb{E}(C^i))^2 + \text{Var}(C^i)](1-\delta_t) \} \cdot \frac{\partial \delta_t^i}{\partial t^i} - 1 \right.\end{aligned}\quad (4.15)$$

From (4.15), we derive the second-order derivative:

$$\begin{aligned}\frac{\partial^2 \pi^i(t^i)}{\partial (t_i)^2} &= 2 \left\{ \left[ \frac{(1-v^i)}{4a_p} + \frac{(1-v^i)}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right] \cdot \left\{ (a_p)^2 \cdot [(\mathbb{E}(C^i))^2 + \text{Var}(C^i)] \left( \frac{\partial \delta_t^i}{\partial t^i} \right)^2 \right. \right. \\ &\left. \left. + \{ [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C^i) - (a_p)^2 \cdot [(\mathbb{E}(C^i))^2 + \text{Var}(C^i)](1-\delta_t) \} \cdot \frac{\partial^2 \delta_t^i}{\partial (t_i)^2} \right\} \right.\end{aligned}\quad (4.16)$$

It follows from expression (4.16) and the assumption  $\delta_t^i = \frac{k_t^i \cdot t^i}{1+k_t^i \cdot t^i}$  that

$$\begin{aligned} \frac{\partial^2 \pi^i(t^i)}{\partial(t^i)^2} &= 2 \left\{ \frac{(1-v^i)}{4a_p} + \frac{(1-v^i)}{4a_p - (1-v^i)[(a_q k_r^i)^2 + (a_\rho k_s^i)^2]} \right\} \times \\ &\{ [\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C_1^i) - \frac{3}{2}(a_p)^2 \cdot [(\mathbb{E}(C_1^i))^2 + Var(C_1^i)](1-\delta_t) \} \cdot \frac{\partial^2 \delta_t}{\partial(t^i)^2} \end{aligned} \quad (4.17)$$

Since  $\frac{\partial^2 \delta_t}{\partial(t^i)^2} < 0$ , when  $[\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C_1^i) \geq \frac{3}{2}(a_p)^2 \cdot [(\mathbb{E}(C_1^i))^2 + Var(C_1^i)]$ ,  $\frac{\partial^2 \pi^i(t^i)}{\partial(t^i)^2}$  is always strictly negative. Therefore,  $\frac{\partial \pi^i(t^i)}{\partial t^i} \downarrow -1$ . Two scenarios can happen. (i):  $\frac{\partial \pi^i(t^i)}{\partial t^i}$  is negative for all  $t^i \geq 0$ . As a result, the firm's profit,  $\pi^i(t^i)$ , decreases in  $t^i$ , and it is optimal for the firm not to invest on any technology ( $t_i^* = 0$ ). (ii):  $\frac{\partial \pi^i(t^i)}{\partial t^i}$  is positive for small  $t^i \geq 0$  and becomes negative for large  $t^i \geq 0$ . In this case, the firm's profit function,  $\pi^i(t^i)$ , is unimodal in  $t^i$  and there exists a unique optimal technology decision  $t_i^* > 0$ .

On the other hand, when  $[\mathbb{E}(A_1) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C_1^i) < \frac{3}{2}(a_p)^2 \cdot [(\mathbb{E}(C_1^i))^2 + Var(C_1^i)]$ ,  $\frac{\partial^2 \pi^i(t^i)}{\partial(t^i)^2}$  is positive for small  $t^i \geq 0$  and negative for large  $t^i \geq 0$ . Therefore,  $\frac{\partial \pi^i(t^i)}{\partial t^i}$  increases for a small  $t^i \geq 0$ , decreases for a large  $t^i \geq 0$ , and as  $t^i \rightarrow \infty$  it approaches  $-1$ . Scenarios (i) and (ii) are still possible. In addition, there could be scenario (iii) where  $\frac{\partial \pi^i(t^i)}{\partial t^i}$ ,  $t^i \geq 0$  is first negative, then positive, and eventually negative again. Equivalently, the firm's profit function,  $\pi^i(t^i)$ , first decreases, then increases, and eventually decreases again in  $t^i$ . It follows that it is optimal for the firm either not to invest ( $t_i^* = 0$ ), or that an optimal investment level  $t_i^* > 0$  exists.

Considering all of the cases discussed, the first order condition in (4.15) can be used to solve for the (interior) optimal technology decision. Since  $\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 > a_p \mathbb{E}(C^i)$  from (4.1), we know  $[\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1] \cdot a_p \mathbb{E}(C^i) > (a_p)^2 (\mathbb{E}(C^i))^2$  in (4.15), and it follows then that when the mean of the landed cost increases ( $\mathbb{E}(C^i) \uparrow$ ) and/or its variance decreases ( $Var(C^i) \downarrow$ ), the firm should invest more in technology ( $t_i^* \uparrow$ ).  $\square$

It is intuitive that more technology investment is welcome when the mean cost goes up,

however, it is surprising that the same direction holds when variance decreases. Note that when variance of the cost decreases (while holding the mean fixed), the option value due to variable cost decreases, and thus manufacturing becomes more costly on average – a similar effect as pulling up the mean.

We demonstrate in the following theorem that when two plant locations enjoy comparable government subsidies and technology adaptability, the option to invest in technology improves the chances that the country with the higher landed cost is selected for the plant site, as opposed to the case when the firm cannot invest in technology.

**Theorem 13** (*Technology Investment.*) *If  $i = A$  and  $i = B$  are two candidate plant sites with  $g^A - f^A = g^B - f^B$  and  $\delta_t^A = \delta_t^B$  for  $t \geq 0$ . WLOG, assume that the landed cost in  $A$  is greater than the landed cost in  $B$ . Then, the likelihood that location  $A$  being selected is greater when the firm can invest in technology compared to the case when it cannot. In other words, when the firm can invest in technology, location  $B$  is less likely to be selected.*

Proof: Suppose for some landed costs  $c^A > c^B$  that the firm is indifferent between the two sourcing locations without technology decisions, i.e., from (4.4):

$$\begin{aligned} & \left\{ \frac{(1-v^A)}{4a_p} + \frac{(1-v^A)}{4a_p - (1-v^A)[(a_q k_r^A)^2 + (a_\rho k_s^A)^2]} \right\} \cdot [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^A]^2 \\ &= \left\{ \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]} \right\} \cdot [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^B]^2 \end{aligned}$$

Note that  $[\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^A] < [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^B]$  due to the difference in landed costs but  $\frac{(1-v^A)}{4a_p} + \frac{(1-v^A)}{4a_p - (1-v^A)[(a_q k_r^A)^2 + (a_\rho k_s^A)^2]} > \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]}$ . As a consequence, fixing  $c^A$ , if the landed cost in location  $B$  is lower than  $c^B$ , it is optimal for the firm to produce the products in location  $B$ , and otherwise in location  $A$ .

Now suppose the firm can equip the plant to be built with technology by upfront investment. Denote  $t^{B*}$  the optimal technology decision for the plant in location  $B$  if it were to incur a



landed cost  $c^B$ . Then, technology can save plant  $B$  a total of

$$\begin{aligned}
& \left\{ \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]} \right\} \cdot [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^B \cdot (1 - \delta_{t^{B^*}}^B)]^2 \\
& - \left\{ \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]} \right\} \cdot [\mathbb{E}(A) + a_q \cdot q_1 + a_\rho \cdot \rho_1 - a_p \cdot c^B]^2 - t^{B^*} \\
& = \left\{ \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]} \right\} \cdot a_p \cdot c^B \cdot \delta_{t^{B^*}} \times \\
& \quad [2\mathbb{E}(A) + 2a_q \cdot q_1 + 2a_\rho \cdot \rho_1 - a_p \cdot c^B \cdot (2 - \delta_{t^{B^*}}^B)] - t^{B^*} \\
& < \left\{ \frac{(1-v^B)}{4a_p} + \frac{(1-v^B)}{4a_p - (1-v^B)[(a_q k_r^B)^2 + (a_\rho k_s^B)^2]} \right\} \cdot a_p \cdot c^A \cdot \delta_{t^{B^*}} \times \\
& \quad [2\mathbb{E}(A) + 2a_q \cdot q_1 + 2a_\rho \cdot \rho_1 - a_p \cdot c^A \cdot (2 - \delta_{t^{B^*}}^B)] - t^{B^*} \\
& < \left\{ \frac{(1-v^A)}{4a_p} + \frac{(1-v^A)}{4a_p - (1-v^A)[(a_q k_r^A)^2 + (a_\rho k_s^A)^2]} \right\} \cdot a_p \cdot c^A \cdot \delta_{t^{B^*}} \times \\
& \quad [2\mathbb{E}(A) + 2a_q \cdot q_1 + 2a_\rho \cdot \rho_1 - a_p \cdot c^A \cdot (2 - \delta_{t^{B^*}}^A)] - t^{B^*} \tag{4.18}
\end{aligned}$$

Expression (4.18) is in fact the saving in cost for plant  $A$  if  $t^{B^*}$  is invested in technology with a landed cost  $c^A$ . Of course, this saving is smaller (for plant  $A$ ) compared to the case when the optimal technology level for plant  $a$ , say some  $t^{A^*}$ , is invested upfront. Therefore, with the ability to invest technology, the firm would prefer to build the plant in location  $A$  compared to  $B$  when the landed costs are  $c^A$  and  $c^B$ , respectively. (The firm is indifferent with the two locations if it cannot invest in technology.) Therefore, technology decisions increases the chances that location  $A$  is selected for the plant site.  $\square$

#### 4.4.2. Single-plant Multiple-market Model

In this subsection, we extend the basic single-plant single-market model of Section 3 to multiple markets. We will start with two markets  $A$  and  $B$ . For period  $j \in \{1, 2\}$ , the demand parameters that describe market  $l \in \{A, B\}$  are denoted as  $A_j^l$ ,  $a_q^l$ ,  $a_\rho^l$  and  $a_p^l$ . The landed cost in period  $j \in \{1, 2\}$  for producing an item in country  $i$  and then selling it in market  $l \in \{A, B\}$  is denoted by the random variable  $C_j^{i,l}$ .

Since there is only one plant, we assume that the firm produces and supplies products of the same product quality to both markets. We still use  $q_1$  to denote the initial (given) product

quality for Period 1, and as in the basic model, the firm can invest in quality enhancement to improve the product quality to some optimal level  $q_2^i$  in Period 2 which depends on the underlying choice of plant location  $i$ .

On the other hand, since there are two markets now, the firm can serve them with different service levels. We denote  $\rho_1^l$  and  $\rho_2^{i,l}$  the given service quality and the improved service quality in Period 1 and Period 2, respectively, for market  $l \in \{A, B\}$  when the firm has selected location  $i$  to build the plant.

To be consistent with the basic model, we assume that

$$\begin{aligned} q_2^i - q_1 &= k_r^i \cdot \sqrt{r^i}; \\ \rho_2^{i,A} - \rho_1^A &= k_s^{i,A} \cdot \sqrt{s^{i,A}}; \\ \rho_2^{i,B} - \rho_1^B &= k_s^{i,B} \cdot \sqrt{s^{i,B}} \end{aligned}$$

where  $r^i$  is the amount of investment made for product quality enhancement, and  $s^{i,A}$ ,  $s^{i,B}$  are the investments for service quality enhancement in markets  $A$  and  $B$ , respectively, all conditional on the plant location  $i$ . In other words, for the initial given quality levels  $(q_1, \rho_1^A, \rho_1^B)$  in Period 1, the investment return rates  $k_r^i$ ,  $k_s^{i,A}$ ,  $k_s^{i,B}$  all depend on the plant location  $i$ , thus leading to different investment strategies and different quality levels in the second period. We impose the condition that  $2(a_q k_r^i)^2 + (a_\rho k_s^{i,A})^2 < 4a_p$  and  $2(a_q k_r^i)^2 + (a_\rho k_s^{i,B})^2 < 4a_p$ .

If some location  $i$  is to be selected for the plant site, the optimal investment decisions satisfy

$$\begin{aligned} &(r^{i*}, s^{i,A*}, s^{i,B*}) \\ \triangleq &\arg \max_{(r^i, s^{i,A}, s^{i,B})} \left\{ \frac{(1-v^i)}{4a_p^A} [\mathbb{E}(A_2^A) + a_q^A \cdot (q_1 + k_r^i \cdot \sqrt{r^i}) + a_\rho^A \cdot (\rho_1 + k_s^{i,A} \cdot \sqrt{s^{i,A}}) - a_p^A \mathbb{E}(C_2^{i,A})]^2 \right. \\ &+ \frac{(1-v^i)}{4a_p^B} [\mathbb{E}(A_2^B) + a_q^B \cdot (q_1 + k_r^i \cdot \sqrt{r^i}) + a_\rho^B \cdot (\rho_1 + k_s^{i,B} \cdot \sqrt{s^{i,B}}) - a_p^B \mathbb{E}(C_2^{i,B})]^2 \\ &\left. + \frac{(1-v^i)a_p^A \cdot \text{Var}(C_2^{i,A})}{4} + \frac{(1-v^i)a_p^B \cdot \text{Var}(C_2^{i,B})}{4} - r^i - s^{i,A} - s^{i,B} \right\} \end{aligned} \quad (4.19)$$

The first order conditions with respect to  $s^{i,A}$  yields,

$$\begin{aligned} & [\mathbb{E}(A_2^A) + a_q^A \cdot q_1 + a_\rho^A \cdot \rho_1] \cdot a_\rho^A k_s^{i,A} \frac{1}{\sqrt{s^{i,A^*}}} + a_q^A k_r^i \cdot a_\rho^A k_s^{i,A} \frac{\sqrt{r^{i^*}}}{\sqrt{s^{i,A^*}}} + (a_\rho^A k_s^{i,A})^2 - \frac{4a_p^A}{(1-v^i)} = 0 \\ \text{iff } \sqrt{s^{i,A^*}} &= \frac{1-v^i}{4a_p^A - (1-v^i)(a_\rho^A k_s^{i,A})^2} \cdot \left\{ [\mathbb{E}(A_2^A) + a_q^A \cdot q_1 + a_\rho^A \cdot \rho_1] \cdot a_\rho^A k_s^{i,A} + a_q^A k_r^i \cdot a_\rho^A k_s^{i,A} \sqrt{r^{i^*}} \right\} \end{aligned} \quad (4.20)$$

Similarly, the first order conditions with respect to  $s^{i,B}$  yields,

$$\begin{aligned} & [\mathbb{E}(A_2^B) + a_q^B \cdot q_1 + a_\rho^B \cdot \rho_1] \cdot a_\rho^B k_s^{i,B} \frac{1}{\sqrt{s^{i,B^*}}} + a_q^B k_r^i \cdot a_\rho^B k_s^{i,B} \frac{\sqrt{r^{i^*}}}{\sqrt{s^{i,B^*}}} + (a_\rho^B k_s^{i,B})^2 - \frac{4a_p^B}{(1-v^i)} = 0 \\ \text{iff } \sqrt{s^{i,B^*}} &= \frac{1-v^i}{4a_p^B - (1-v^i)(a_\rho^B k_s^{i,B})^2} \cdot \left\{ [\mathbb{E}(A_2^B) + a_q^B \cdot q_1 + a_\rho^B \cdot \rho_1] \cdot a_\rho^B k_s^{i,B} + a_q^B k_r^i \cdot a_\rho^B k_s^{i,B} \sqrt{r^{i^*}} \right\} \end{aligned} \quad (4.21)$$

Plugging (4.20) and (4.21) back into (4.19) and after quite tedious algebraic simplifications, one can derive the optimal expected profit for the second period as

$$\begin{aligned} \pi_2^{i^*} &= \frac{1-v^i}{4a_p^A - (1-v^i)(a_\rho^A k_s^{i,A})^2} [\mathbb{E}(A_2^A) + a_q^A \cdot (q_1 + k_r^i \cdot \sqrt{r^{i^*}}) + a_\rho^A \cdot \rho_1 - a_p^A \mathbb{E}(C_2^{i,A})]^2 \\ &+ \frac{1-v^i}{4a_p^B - (1-v^i)(a_\rho^B k_s^{i,B})^2} [\mathbb{E}(A_2^B) + a_q^B \cdot (q_1 + k_r^i \cdot \sqrt{r^{i^*}}) + a_\rho^B \cdot \rho_1 - a_p^B \mathbb{E}(C_2^{i,B})]^2 \\ &+ \frac{(1-v^i)a_p^A \cdot \text{Var}(C_2^{i,A})}{4} + \frac{(1-v^i)a_p^B \cdot \text{Var}(C_2^{i,B})}{4} - r^{i^*} \end{aligned} \quad (4.22)$$

where  $r^{i^*}$  is the optimal amount of investment in product quality.

Optimizing expression (4.22) with respect to  $r^{i^*}$  gives

$$\sqrt{r^{i^*}} = \frac{\sum_{l=A,B} \frac{1-v^i}{4a_p^l - (1-v^i)(a_\rho^l k_s^{i,l})^2} [\mathbb{E}(A_2^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_2^{i,l})] \cdot a_q^l k_r^i}{1 - \sum_{l=A,B} \frac{(1-v^i)(a_q^l k_r^i)^2}{4a_p^l - (1-v^i)(a_\rho^l k_s^{i,l})^2}} \quad (4.23)$$

By plugging (4.23) back into (4.22), the optimal expected profit for the second period

becomes

$$\begin{aligned}
& \pi_2^{i*}(A, B) \\
&= \sum_{l=A, B} \left\{ \frac{1-v^i}{4a_p^l - (1-v^i)(a_\rho^l k_s^{i,l})^2} [\mathbb{E}(A_2^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_2^{i,l})]^2 + \frac{(1-v^i)a_p^l \cdot \text{Var}(C_2^{i,l})}{4} \right\} \\
&+ \frac{\left\{ \sum_{l=A, B} \frac{1-v^i}{4a_p^l - (1-v^i)(a_\rho^l k_s^{i,l})^2} [\mathbb{E}(A_2^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_2^{i,l})] \cdot a_q^l k_r^i \right\}^2}{1 - \sum_{l=A, B} \frac{(1-v^i)(a_q^l k_r^i)^2}{4a_p^l - (1-v^i)(a_\rho^l k_s^{i,l})^2}} \quad (4.24)
\end{aligned}$$

It is easy to check that the optimal expected profit for the second period characterized by (4.24) increases with each of the investment return rates  $k_r^i$ ,  $k_s^{i,A}$ ,  $k_s^{i,B}$ . Compare this with the optimal expected profit in the first period, i.e.,

$$\pi_1^{i*}(A, B) = \sum_{l=A, B} \left\{ \frac{1-v^i}{4a_p^l} [\mathbb{E}(A_1^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_1^{i,l})]^2 + \frac{(1-v^i)a_p^l \cdot \text{Var}(C_1^{i,l})}{4} \right\} \quad (4.25)$$

we can clearly see the value for quality enhancement, and the option value of the variance of the landed cost. Furthermore, the optimal location decision when there are two markets  $A$  and  $B$  is given by

$$i^*(A, B) = \arg \max_i \{ \pi_1^{i*}(A, B) + \pi_2^{i*}(A, B) - f^i + g^i \}.$$

The structure shares the same trade-offs among landed costs, capability of quality enhancement and government policies, just as in the basic model. But with multiple markets, we can conclude that the sourcing location decision depends on the demand sizes of individual markets and derive the following result.

**Theorem 14 (Local Market)** *If the local market in country  $i$  is big and/or growing, the firm is more likely to produce the plants in its location.*

Finally, it is a straight-forward exercise to show that when there are  $N$  end markets denoted

by some set  $S_m$ , the optimal sourcing location decision for the firm is

$$i^*(S_m) = \arg \max_i \{ \pi_1^{i^*}(S_m) + \pi_2^{i^*}(S_m) - f^i + g^i \}$$

$$\text{where } \pi_1^{i^*}(S_m) = \sum_{l \in S_m} \left\{ \frac{1-v^i}{4a_p^l} [\mathbb{E}(A_1^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_1^{i,l})]^2 + \frac{(1-v^i)a_p^l \text{Var}(C_1^{i,l})}{4} \right\}$$

$$\begin{aligned} \text{and } \pi_2^{i^*}(S_m) = & \sum_{l \in S_m} \left\{ \frac{1-v^i}{4a_p^l - (1-v^i)(a_q^l k_s^{i,l})^2} [\mathbb{E}(A_2^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_2^{i,l})]^2 \right. \\ & \left. + \frac{(1-v^i)a_p^l \cdot \text{Var}(C_2^{i,l})}{4} \right\} \\ & + \frac{\left\{ \sum_{l \in S_m} \frac{1-v^i}{4a_p^l - (1-v^i)(a_q^l k_s^{i,l})^2} [\mathbb{E}(A_2^l) + a_q^l \cdot q_1 + a_\rho^l \cdot \rho_1 - a_p^l \mathbb{E}(C_2^{i,l})] \cdot a_q^l k_r^{i,l} \right\}^2}{1 - \sum_{l \in S_m} \frac{(1-v^i)(a_q^l k_r^{i,l})^2}{4a_p^l - (1-v^i)(a_q^l k_s^{i,l})^2}} \end{aligned}$$

#### 4.4.3. Multiple-plant Single-market Model

In this subsection, we construct an extension to the basic model in Section 3 by allowing the firm to invest in  $N$  plants upfront, denoted by some location set  $S_p$ . Depending on the realized landed costs in each of the two periods, the firm can then specify a location subset of  $S_p$  to use for production, in order to optimize its overall policy. For example, a firm that owns both a Chinese plant and a Mexican plant can switch some or all of its scheduled production from China to Mexico in Period 2, if labor wages in China undergo an unusual increase due to intervention of the government. Since we do not impose production capacities in our model, an optimal policy can always be achieved by producing all of the volumes required in one available plant during each period.

Switching manufacturing plants usually does not come for free. To have the option to be flexible, often referred to as the “real option”, the firm must pay fixed costs upfront to

multiple plants to build or secure their capacities. In addition, we still assume that the firm can choose to invest in product and service quality enhancement within the first period. Nevertheless, the effectiveness of the quality enhancement would go down if production is shifted to a new plant before the second period, since process and service support improvements that are developed for the plant used in Period 1 may not and most likely will not be fully carried over to the plant in Period 2. This is an equivalent notion of switching costs in a model when a fraction of the entire production volume can be relocated.

Suppose the firm produces the product in location  $i$  during the first period, and has invested  $r$  and  $s$  dollars respectively in product and service quality enhancement, but switches to plant  $j$  in the second period. Then, given the initial quality bundle  $(q_1, \rho_1)$ , we assume that the quality of the product to be launched in the second period can be characterized by

$$(q_2^j, \rho_2^j) = (q_1 + f_r^i(r) \cdot (1 - \phi_{ij}^i), \rho_1 + f_s^i(s) \cdot (1 - \phi_{ij}^i))$$

where

$$\phi_{ij}^i \begin{cases} = 0 & \text{if } j = i \\ \in [0, 1] & \text{if } j \neq i \end{cases} \quad (4.26)$$

denotes the loss in quality enhancement as a result of switching from locations  $i$  to  $j$ .

We conduct the analysis below for when it is optimal for the firm to build multiple plants rather than one plant upfront. Let us denote  $\pi^{S_p}$  to be the optimal expected after-tax two-period profit for the firm when it can produce in the plant set  $S_p$ , and  $\pi_2^{S_p, i}(r^{S_p, i}, s^{S_p, i})$  to be the optimal expected after-tax profit *for the second period* given that plant  $i$  is being used, and  $r^i$  and  $s^i$  are invested in product and service quality enhancement in the first period.

Depending on the realizations of the second period landed costs,  $\{c_2^a : a \in S_p\}$ , a firm who has produced the product in plant  $i$  in the first period with quality investment levels  $r^i$  and  $s^i$  can strategically decide if it will continue to produce the products in plant  $i$ , or if it will

switch to another plant for the second period. Its optimal expected after-tax profit for the second period thus depends on the joint distribution of the landed costs  $\{C_2^a : a \in S_p\}$ :

$$\begin{aligned} & \pi_2^{S_p, i}(r^{S_p, i}, s^{S_p, i}) \\ = & \mathbb{E}_{\{C_2^a : a \in S_p\}} \left\{ \max_{j \in S_p} \frac{1 - v^j}{4a_p} [\mathbb{E}(A_2) + a_q \cdot [q_1 + f_r^i(r^{S_p, i}) \cdot \phi_{ij}^i] + a_\rho \cdot [\rho_1 + f_s^i(s^{S_p, i}) \cdot \phi_{ij}^i] - C_2^j]^2 \right\} \end{aligned} \quad (4.27)$$

To derive the first-period optimal quality enhancement decisions  $r^{S_p, i}$  and  $s^{S_p, i}$ , given that location  $i$  is used in the first period, the firm needs to take into consideration all possible circumstances that could arise during the second period. For example, if there is a high likelihood that the firm needs to switch location for Period 2, and/or the switching effect is severe (i.e.,  $\phi_{ij}$  is high), then it may not be optimal for the firm to invest tremendously on product and service quality enhancement during the first period. The optimal decisions are given as

$$(r^{S_p, i^*}, s^{S_p, i^*}) = \arg \max_{r^{S_p, i}, s^{S_p, i}} \pi_2^{S_p, i}(r^{S_p, i}, s^{S_p, i}) - r^{S_p, i} - s^{S_p, i} \quad (4.28)$$

Therefore, the optimal expected after-tax two-period profit for the firm when it has flexibility of the plant set  $S_p$  is:

$$\pi^{S_p} = \mathbb{E}_{\{C_1^a : a \in S_p\}} \max_{i \in S_p} \left\{ \frac{1 - v^i}{4a_p} [\mathbb{E}(A_1) + a_q q_1 + a_\rho \rho_1 - C_1^i]^2 + \pi_2^{S_p, i}(r^{S_p, i^*}, s^{S_p, i^*}) - r^{S_p, i^*} - s^{S_p, i^*} \right\} \quad (4.29)$$

**Real Option.** When the firm can only produce at location  $i$ , its optimal expected after-tax two-period profit,  $\pi^i$ , can be expressed as

$$\begin{aligned} \pi^i = & \mathbb{E}_{C_1^i} \frac{1 - v^i}{4a_p} [\mathbb{E}(A_1) + a_q q_1 + a_\rho \rho_1 - C_1^i]^2 \\ & + \mathbb{E}_{C_2^i} \frac{1 - v^i}{4a_p} [\mathbb{E}(A_1) + a_q [q_1 + f_r^i(r^{i^*})] + a_\rho [\rho_1 + f_s^i(s^{i^*})] - C_2^i]^2 - r^{i^*} - s^{i^*} \end{aligned} \quad (4.30)$$

where  $r^{i^*}$  and  $s^{i^*}$  are the optimal quality enhancement decisions in the first period.

Using (4.27) and (4.28), it is clear that  $\pi^{S_p}$  in (4.29) is greater than  $\pi^i$  in (4.30) for any  $i \in S_p$ . The positive difference,  $\pi^{S_p} - \pi^i$ , represents the real option value of having flexibility, i.e., being able to produce the products in multiple plants, as opposed to being able to use only the plant  $i$ . Given the distributions of the landed costs, numerically we can investigate the effects of the shape of the distributions, the demand parameters, and the stickiness parameters  $\phi$  on the real option value.

Ultimately, real option does not come for free. The firm has to purchase capacities upfront in order to utilize them during the following the production periods. The following statement describe the firm's optimal sourcing location decision. Suppose a fixed cost  $f^{S_p,i}$  is needed to secure capacities for plant  $i$  meanwhile the firm receives government subsidy  $g^{S_p,i}$  for investment, then it is wise for the firm to consider using the plant set  $S_p$  if and only if

$$\pi^{S_p} - \sum_{i \in S_p} f^{S_p,i} + \sum_{i \in S_p} g^{S_p,i} \geq \max_{i \in S_p} \{\pi^i - f^i + g^i\}.$$

#### 4.5. NUMERICAL ILLUSTRATIONS

In this section, we provide numerical examples to illustrate structural results found in earlier sections. In these examples, we assume that only the U.S. market is supplied by the firm, and compare costs and profits across plant locations in the U.S., China, and Mexico. These correspond to the firm's "re-shoring", "off-shoring" and "near-shoring" decisions.

For all the examples, we assume the corporate tax rates for the three countries are  $v^{U.S.} = 35\%$ ,  $v^{China} = 25\%$ ,  $v^{Mexico} = 30\%$ , and that the historical wage inflation trends will hold for the next 10 years, i.e., we assume that there will be an annual increase in wages at 15% for China and at 2% for the U.S. and Mexico. For products supplied to the U.S. market, we ignore the shipping cost if they are being produced in the U.S., and assume that it costs China four times more expensive to ship the product than it costs Mexico.



**Example (i).** In this example, we demonstrate that lower fixed costs and favorable government incentives (i.e. increasing  $g^i$  and/or decreasing  $f^i, v^i$ ) encourage sourcing (to country  $i$ ), by considering industries with high labor cost and low shipping cost relative to other costs, e.g., the apparel/footwear and the electronics industries.

We normalize the landed cost for manufacturing in the Chinese plant in year 2014 to be \$1, and assume that the labor cost accounts for 20% of the landed cost while the shipping cost to the U.S. accounts for 3%. Figure 11 displays the estimated landed costs for producing the product in the U.S., China, and Mexico, respectively, for years 2014-2024.

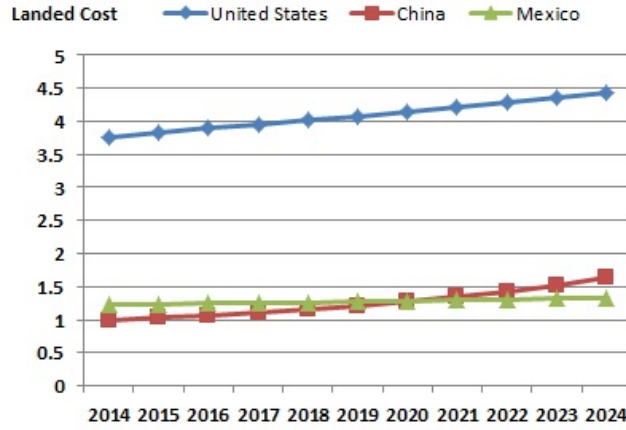


Figure 11: Estimated landed cost for a product that has high labor cost and low shipping cost.

It is clear from Figure 11 that China and Mexico have tremendous landed cost advantages over the U.S. in the industries noted above. In year 2019 and onward, Mexico is estimated to have a lower landed cost compared to China due to a lower wage inflation rate.

In Figure 12, we examine the firm’s expected profits. We normalize the base portion of the demand,  $E(A_1) + a_q q_1 + a_\rho \rho_1$ , to be 1000, and set  $(a_q k_r^{U.S.})^2 = (a_q k_s^{U.S.})^2 = 100$ ,  $(a_q k_r^{China})^2 = (a_q k_s^{China})^2 = 0$ ,  $(a_q k_r^{Mexico})^2 = (a_q k_s^{Mexico})^2 = 100$ . The difference between Figure 12/(a) and Figure 12/(b) is only that  $g^{U.S.} - f^{U.S.} = g^{China} - f^{China} = g^{Mexico} - f^{Mexico} = 0$  for Figure 12/(a), while  $g^{U.S.} - f^{U.S.} = g^{China} - f^{China} = 0$  and  $g^{Mexico} - f^{Mexico} = 100$  for Figure 12/(b).

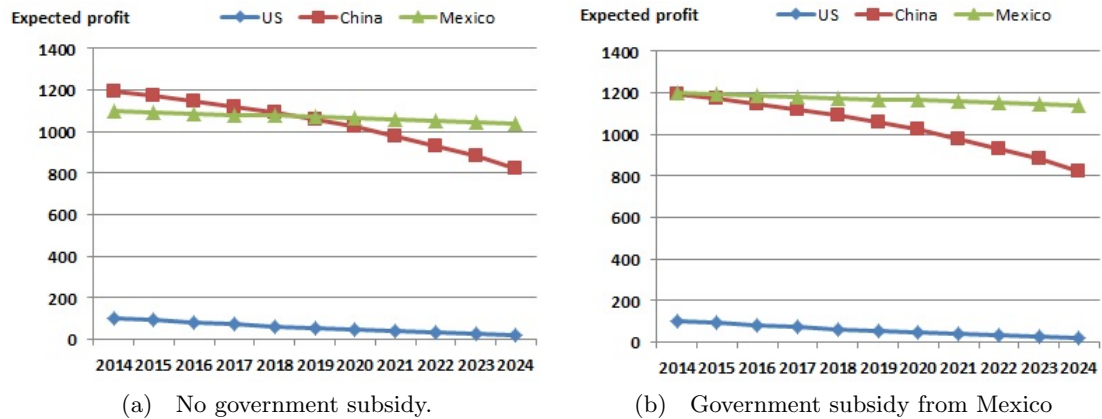


Figure 12: Expected profit for producing a product that has high labor cost and low shipping cost.

We see from Figure 12 that a relative small financial subsidy from the local government in Mexico could shift the tipping point to as early as now, enabling Mexico to become the most ideal location to produce apparel, footwear and electronics for the U.S. market. All else being equal, we predict that the apparel/footwear industry will off-shore to Mexico earlier than the electronics industry due to lower fixed cost needed upfront. On the other hand, these industries noted here are very unlikely to re-shore to the U.S., given the enormous landed cost disadvantage.

**Example (ii).** In this example, we demonstrate that enhanced technology would reduce the impact of the landed cost advantage in making sourcing decisions. We consider industries with high labor cost and high shipping cost relative to other costs, e.g., automobile and appliances companies.

We still normalize the landed cost in a plant located in China in year 2014 to be \$1, but assume now that the labor cost accounts for 15% of the landed cost while the shipping cost to the U.S. accounts for 30%. Figure 13 displays the estimated landed costs for producing such a product in the U.S., China, and Mexico for years 2014-2024.

Since the shipping cost to the U.S. market is much more significant for this product compared to that in Example (i), we see from Figure 13 that Mexico is awarded with the lowest

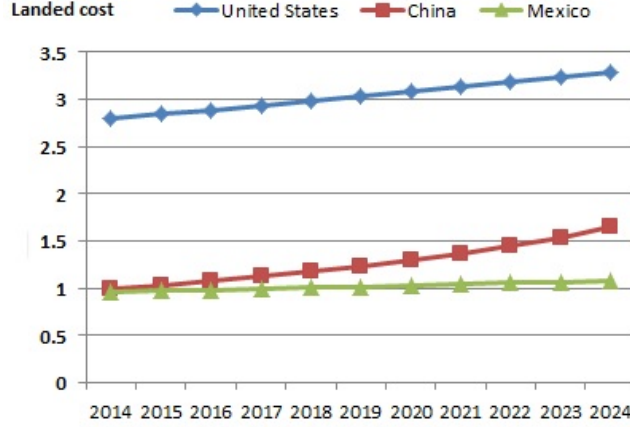


Figure 13: Estimated landed cost for a product that has high labor input and high shipping cost.

landed cost among the three country locations. In contrast, a plant in the U.S. would still incur the highest landed cost because of the relative high labor input required to produce the product.

We examine the firm’s expected profits without the option to invest in technology in Figure 14/(a) and with the option to invest in technology in Figure 14/(b). The base portion of the demand,  $E(A_1) + a_q q_1 + a_\rho \rho_1$ , is still normalized to be 1000. This time, we set  $(a_q k_r^{U.S.})^2 = (a_q k_s^{U.S.})^2 = 500$ ,  $(a_q k_r^{Mexico})^2 = (a_q k_s^{Mexico})^2 = 250$ ,  $(a_q k_r^{China})^2 = (a_q k_s^{China})^2 = 0$  as estimates for the location-dependent parameters that describe investment return rates on product and service quality enhancement in the automobile and appliances industries.

The managerial insights we can learn from Figure 14/(a) is that, without the option to invest in technology such as automation and robotic arms, near-shoring to Mexico is by far considered as the best strategy for the firm in terms of the sourcing location decision, as Mexico balances low labor cost with low shipping cost. Many companies in the automobile and appliance industries have in fact already started manufacturing products in Mexico to supply their U.S. market.

Nevertheless, if the firm has the ability to equip its plant with enhanced technology, we see from 14/(b) that it will save much more for a plant in the U.S. compared to a plant in China

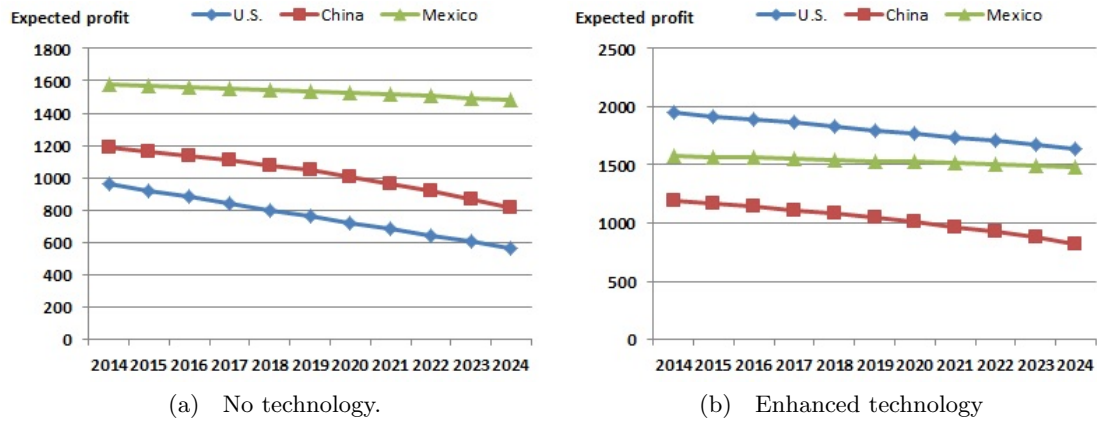


Figure 14: Expected profit for producing a product that has high labor cost and high shipping cost.

or in Mexico because of the higher labor wage in the U.S. As a result, with this particular example, that enhanced technology can totally negate the high labor cost disadvantage. We predict that with technology advances, the automobile industry and the appliances industry would be among those that consider to re-shore to the U.S.

**Example (iii).** In this example, we demonstrate that when outsourcing leads to a loss of capability for the firm to innovate product or to improve service, it may be optimal for it to manufacture the products closer to the headquarters or the market, even if that means higher landed cost.

Specifically, we consider industries with low labor cost and high shipping cost relative to the landed cost in this case, e.g., the heavy machinery industry or the aerospace & defense industry. Note that in such industries, hourly labor wages are high due to the employment of highly-skilled workers. However, considering the enormous raw material cost and logistic cost, total labor cost component accounts for a relatively small percentage in the landed cost.

Figure 15/(a) displays the estimated landed costs for producing a product in the industries noted above in the U.S., China, and Mexico for years 2014-2024. We still normalize the landed cost in a plant in China in 2014 to be \$1, but assume now that the labor cost

accounts for 3% of the landed cost while the shipping cost to the U.S. accounts for 30%.

Figure 15/(b) displays the expected profit, with the base portion of the demand,  $E(A_1) + a_q q_1 + a_\rho \rho_1$  still being 1000. Like in Example (ii), we set  $(a_q k_r^{U.S.})^2 = (a_q k_s^{U.S.})^2 = 500$ ,  $(a_q k_r^{China})^2 = (a_q k_s^{China})^2 = 0$ ,  $(a_q k_r^{Mexico})^2 = (a_q k_s^{Mexico})^2 = 250$ , to capture the impact of proximity on product innovation and service enhancement capabilities.

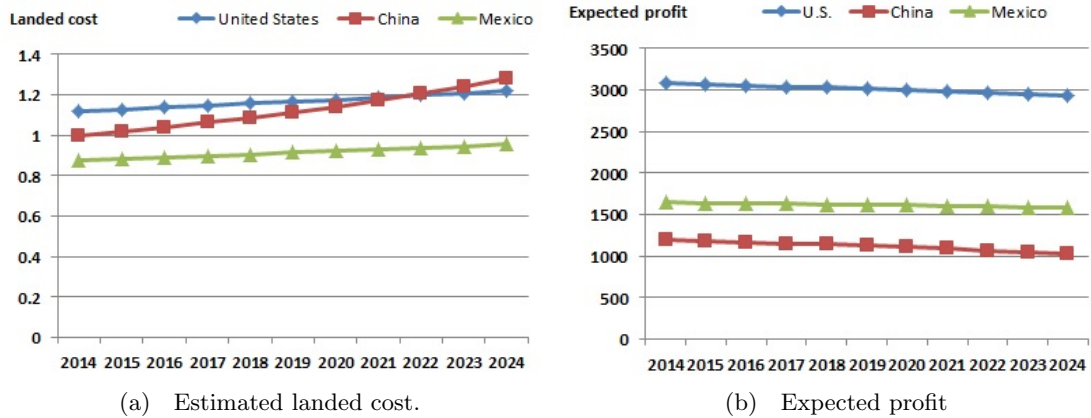


Figure 15: Estimated landed cost and expected profit for producing a product that has low labor cost and high shipping cost.

We observe from Figure 15/(a) that because of low labor input and high shipping cost, the landed cost of producing the product in the U.S. is quite close to the landed cost of producing the product in China or in Mexico. While Mexico still gives us the lowest landed cost considering the low labor wage in the country and its proximity to the U.S. market, China would in fact become the most expensive place for production in a few years, as a consequence of its rising labor wages and further off-shore location.

Nevertheless, when we examine the expected profit displayed in Figure 15/(b), a plant site in the U.S. becomes a clear winner. We predict that these industries including heavy machinery or aerospace & defense industries will stay in the United States, or other developed countries like Japan and Germany, to maintain high innovation and service levels.

## 4.6. CONCLUSIONS

Many manufacturing firms in the U.S. are re-examining the structure of their global supply chains and their associated sourcing strategy in response to the uncertainties and risks they face nowadays. For decades, a dominant strategy in manufacturing has been to outsource to low cost global suppliers. This has led to the transfer of manufacturing jobs and development activities out of the U.S., and into low labor cost countries such as China, India and Vietnam. In recent years, however, this trend is being challenged by some companies' "re-shoring" and "near-shoring" decisions.

In contrast to traditional analysis where labor cost savings are evaluated versus transportation costs, we propose a comprehensive model framework for this global sourcing location decision process that incorporates perspectives over the entire life cycle of a product, i.e., product design, manufacturing and delivering, and after-sale service support.

We use our framework to test the validity of various competing theories on global sourcing, and find indeed that (i) Favorable government policies such as lower corporate tax rates and greater financial subsidies can have a major positive impact on sourcing decisions; (ii) Technology developments (such as enhanced automation) can negate the landed cost disadvantage, especially for products that require substantial labor input; (iii) Consideration in capability to develop new products and higher service standard can override cost concerns in decision-making.

To extend the basic one-plant one-market model, we consider both the one-plant N-market scenario and the N-plant one-market scenario. With the former, it is easy to check that the sourcing location decision depends on the underlying market distribution, and when the local market in a specific country is big and/or growing, the firm is more likely to select a plant in its location. With the latter, the decision-making becomes a trade-off between the value of real option and the upfront sunk cost needed to open and operate multiple plants.

Numerical examples are provided in the paper to support the structural results identified

from the model. We predict that industries with high labor cost and low shipping cost relative to the landed cost (e.g., apparel and electronics industries) are unlikely to come back to the U.S.; Industries with relatively high labor cost and low shipping cost (e.g., automobile and appliances industries) will benefit from production in the U.S. if and only if the plant being used is equipped with enhanced automation; And industries with relatively low labor cost and high shipping cost (e.g., heavy machinery and aerospace & defense industries) will remain their manufacturing in the U.S., to maintain high innovation and high service levels.

## APPENDICES

### APPENDIX TO CHAPTER 2

***Proof of Lemma 1:***

(i) By construction, the entire probability mass at one end of the distribution is transferred to the middle of the support. As a result, the range of the random variable  $\tilde{N}_{K+1}$  is a strict subset of the range of  $\tilde{N}_K$ . Specifically,  $a_{(K+1)_1} = a_{K_1} + 1$  if  $f_K(a_{K_1}) \leq f_K(a_{K_n})$  and  $a_{(K+1)_N} = a_{K_n} - 1$  if  $f_K(a_{K_1}) \geq f_K(a_{K_n})$ . The *length* of the range of  $\tilde{N}_K$ ,  $|a_{K_n} - a_{K_1}|$ , is strictly decreasing in  $K$ . Within a finite number of steps, for some time  $K = T$ , the length will be less than 2. When  $a_{T_n} - 1 < a_{T_1} + 1$ , the process stops. Thus,  $T$  is finite.

(ii) We show that  $F_{K+1} \leq_{SMPs} F_K$ . Let  $a_{i_1}, a_{i_2}, a_{i_3}, a_{i_4}$  in Definition 2 be  $a_{K_1}, a_{K_1} + 1, a_{K_n} - 1$  and  $a_{K_n}$  respectively.  $f_{K+1} = f_K$  for all but these four points. Define  $\gamma_{i_k} = f_K(a_{i_k}) - f_{K+1}(a_{i_k})$  for  $k = 1, 2, 3, 4$ . Then,  $\gamma_{i_1} = -\gamma_{i_2} = -\gamma_{i_3} = \gamma_{i_4} = \min\{f_K(a_{K_1}), f_K(a_{K_n})\} > 0$ . Moreover,  $\sum_{k=i}^4 a_{i_k} \gamma_{i_k} = [a_{K_1} - (a_{K_1} + 1) - (a_{K_n} - 1) + a_{K_n}] \min\{f_K(a_{K_1}), f_K(a_{K_n})\} = 0 \cdot \min\{f_K(a_{K_1}), f_K(a_{K_n})\} = 0$ . □

***Proof of Lemma 2:***

Suppose that  $\tilde{N}_T$  has two elements which are not consecutive. Then, it must be that  $a_{T_1} + 1 < a_{T_n}$ . By Construction 1, then sequence is not completed, which contradicts the definition of  $T$ . Else, suppose that  $\tilde{N}_T$  has three or more elements. Again, it must be that  $a_{T_1} + 1 < a_{T_n}$ , and hence, the sequence in Construction 1 is incomplete, which contradicts the definition of  $T$ . Therefore,  $\tilde{N}_T$  can either take a single value or two consecutive values. Case (i): When if  $\mathbb{E}(\tilde{N}_0)$  is an integer, since the transformation is mean preserving, we have  $\tilde{N}_T$  is a singleton with  $\tilde{N}_T = \mathbb{E}(\tilde{N}_0) = \lfloor \mathbb{E}(\tilde{N}_0) \rfloor = \lceil \mathbb{E}(\tilde{N}_0) \rceil$ . Case (ii): When  $\mathbb{E}(\tilde{N}_0)$  is not an integer,  $\tilde{N}_T$  cannot be a singleton. Thus,  $\tilde{N}_T$  takes on two consecutive values. Since the transformation in Construction 1 is mean-preserving with  $\mathbb{E}(\tilde{N}_0)$ , we have  $\mathbb{E}(\tilde{N}_T) = \mathbb{E}(\tilde{N}_0)$ . Then we must have  $\tilde{N}_T \in \{\lfloor \mathbb{E}(\tilde{N}_0) \rfloor, \lceil \mathbb{E}(\tilde{N}_0) \rceil\}$ , with  $\Pr(\tilde{N}_T = \lfloor \mathbb{E}(\tilde{N}_0) \rfloor) \lfloor \mathbb{E}(\tilde{N}_0) \rfloor + \Pr(\tilde{N}_T = \lceil \mathbb{E}(\tilde{N}_0) \rceil) \lceil \mathbb{E}(\tilde{N}_0) \rceil = \mathbb{E}(\tilde{N}_0)$ . It is also clear that the distribution of  $\tilde{N}_T$  is independent of



the distribution of  $\tilde{N}_0$ . □

**Proof of Lemma 3:**

(i) For  $j \in \{K, K+1\}$ , recall from (2.6) that  $R_{\tilde{N}_j} = p\lambda_{eff, \tilde{N}_j} = p\mu(1 - \pi_0)$  where  $\pi_0 = 1 / \left(1 + \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_j(n)\right)$ . It is thus sufficient to show that

$$\sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n). \quad (\text{A.1})$$

To verify (A.1), our strategy is to form a partition of  $i \in \{0, 1, 2, \dots\}$  based on the sign of  $\rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) - \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$ . We specifically focus on terms that make the product  $\prod_{n=0}^{i-1} \bar{F}_K(n)$ , namely  $\bar{F}_K(n)$ . Since  $\bar{F}_K(n) = f_K(n+1) + f_K(n+2) + \dots$ , applying Construction 1, we have

$$\begin{aligned} \bar{F}_{K+1}(0) &= \bar{F}_K(0) &= 1, \\ \bar{F}_{K+1}(1) &= \bar{F}_K(1) &= 1, \\ &\vdots \\ \bar{F}_{K+1}(a_{K_1} - 1) &= \bar{F}_K(a_{K_1} - 1) &= 1, \\ \bar{F}_{K+1}(a_{K_1}) &= \bar{F}_K(a_{K_1}) + \min\{f_K(a_{K_1}), f_K(a_{K_n})\} &\in (0, 1], \\ \bar{F}_{K+1}(a_{K_1} + 1) &= \bar{F}_K(a_{K_1} + 1) &\in (0, 1), \\ \bar{F}_{K+1}(a_{K_1} + 2) &= \bar{F}_K(a_{K_1} + 2) &\in (0, 1), \\ &\vdots \\ \bar{F}_{K+1}(a_{K_n} - 2) &= \bar{F}_K(a_{K_n} - 2) &\in (0, 1), \\ \bar{F}_{K+1}(a_{K_n}) &= \bar{F}_K(a_{K_n}) &= 0, \\ \bar{F}_{K+1}(a_{K_n} + 1) &= \bar{F}_K(a_{K_n} + 1) &= 0, \\ &\vdots \end{aligned} \quad (\text{A.2})$$

Thus, using transformation of  $\tilde{N}_K$  to  $\tilde{N}_{K+1}$  in Construction 1, we see that  $\bar{F}_K(n)$  differs from  $\bar{F}_{K+1}(n)$  at only two points, specifically  $n = a_{K_1}$  and  $n = a_{K_n} - 1$ . In order to show (A.1), we verify, as an intermediate step, that  $\bar{F}_{K+1}(a_{K_1}) \cdot \bar{F}_{K+1}(a_{K_n} - 1) < \bar{F}_K(a_{K_1}) \cdot \bar{F}_K(a_{K_n} - 1)$ . We have

$$\begin{aligned}
& \bar{F}_{K+1}(a_{K_1}) \cdot \bar{F}_{K+1}(a_{K_n} - 1) \\
&= [\bar{F}_K(a_{K_1}) + \min\{f_K(a_{K_1}), f_K(a_{K_n})\}] [\bar{F}_K(a_{K_n} - 1) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\}] \\
&= \bar{F}_K(a_{K_1}) \bar{F}_K(a_{K_n} - 1) \\
&\quad + \min\{f_K(a_{K_1}), f_K(a_{K_n})\} [\bar{F}_K(a_{K_n} - 1) - \bar{F}_K(a_{K_1}) - \min\{f_K(a_{K_1}), f_K(a_{K_n})\}] \\
&< \bar{F}_K(a_{K_1}) \bar{F}_K(a_{K_n} - 1) \text{ since } \bar{F}_K(a_{K_n} - 1) \leq \bar{F}_K(a_{K_1}). \tag{A.3}
\end{aligned}$$

Now we define  $\mathcal{S}_1 \triangleq \{1, 2, \dots, a_{K_1}\}$ ;  $\mathcal{S}_2 \triangleq \{a_{K_1} + 1, a_{K_1} + 2, \dots, a_{K_n} - 1\}$ ;  $\mathcal{S}_3 \triangleq \{a_{K_n}\}$ ; and  $\mathcal{S}_4 \triangleq \{a_{K_n} + 1, a_{K_n} + 2, \dots\}$ .  $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$  and  $\mathcal{S}_4$  then form a partition of the space  $\{1, 2, 3, \dots\}$ .

Our goal (A.1) is equivalent to 
$$\sum_{i \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n).$$

From (A.2) and (A.3) we have  $\forall i \in \mathcal{S}_1 : \prod_{n=0}^{i-1} \bar{F}_K(n) = \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) = 1$ ;  $\forall i \in \mathcal{S}_2 :$   
 $\prod_{n=0}^{i-1} \bar{F}_K(n) < \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$ ;  $\forall i \in \mathcal{S}_3 : \prod_{n=0}^{i-1} \bar{F}_K(n) > \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$ ; and  $\forall i \in \mathcal{S}_4 : \prod_{n=0}^{i-1} \bar{F}_K(n) =$   
 $\prod_{n=0}^{i-1} \bar{F}_{K+1}(n) = 0$ .

It is clear that  $\mathcal{S}_1$  and  $\mathcal{S}_4$  are collection of the indices  $i$  where  $\prod_{n=0}^{i-1} \bar{F}_K(n) = \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$ .

Hence, to prove (A.1), it suffices to show that 
$$\sum_{i \in \mathcal{S}_2 \cup \mathcal{S}_3} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i \in \mathcal{S}_2 \cup \mathcal{S}_3} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n).$$

As  $\rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$  for all  $i \in \mathcal{S}_2$ , the inequality will hold, if there exists some  $\mathcal{S}_{2'} \subseteq \mathcal{S}_2$  such that

$$\sum_{i \in \mathcal{S}_{2'} \cup \mathcal{S}_3} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i \in \mathcal{S}_{2'} \cup \mathcal{S}_3} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n). \tag{A.4}$$

On the other hand, the existence of  $\tilde{N}_{K+1}$  guarantees that  $a_{K_n} - 1 \geq a_{K_1} + 1$  and  $a_{K_1} \geq 1$  so there exists at least one index in  $\mathcal{S}_2$  (i.e.,  $i = a_{K_1} + 1$ ). There is only one element in  $\mathcal{S}_3$  (i.e.,  $i = a_{K_n}$ ). Let us define  $\mathcal{S}_2' \triangleq \{a_{K_1} + 1\}$ . So  $\mathcal{S}_2' \cup \mathcal{S}_3 = \{a_{K_1} + 1, a_{K_n}\}$ . Inequality (A.4) is therefore equivalent to

$$\begin{aligned} & \sum_{i \in \{a_{K_1}+1, a_{K_n}\}} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i \in \{a_{K_1}+1, a_{K_n}\}} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n), \\ \Leftrightarrow & \rho^{a_{K_1}+1} \prod_{n=0}^{a_{K_1}} \bar{F}_K(n) + \rho^{a_{K_n}} \prod_{n=0}^{a_{K_n}-1} \bar{F}_K(n) < \rho^{a_{K_1}+1} \prod_{n=0}^{a_{K_1}} \bar{F}_{K+1}(n) + \rho^{a_{K_n}} \prod_{n=0}^{a_{K_n}-1} \bar{F}_{K+1}(n). \end{aligned}$$

And the last condition is true because

$$\begin{aligned} & \rho^{a_{K_n}} \prod_{n=0}^{a_{K_n}-1} \bar{F}_K(n) - \rho^{a_{K_n}} \prod_{n=0}^{a_{K_n}-1} \bar{F}_{K+1}(n) < \rho^{a_{K_n}} \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \prod_{n=0}^{a_{K_n}-2} \bar{F}_K(n) \\ & \leq \rho^{a_{K_n}} \min\{f_K(a_{K_1}), f_K(a_{K_n})\} < \rho^{a_{K_1}+1} \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \\ & = \rho^{a_{K_1}+1} \min\{f_K(a_{K_1}), f_K(a_{K_n})\} \prod_{n=0}^{a_{K_1}-1} \bar{F}_K(n) = \rho^{a_{K_1}+1} \prod_{n=0}^{a_{K_1}} \bar{F}_{K+1}(n) - \rho^{a_{K_1}+1} \prod_{n=0}^{a_{K_1}} \bar{F}_K(n). \end{aligned}$$

Therefore, inequality (A.4)-(A.1) all hold by backward induction, and

$$\lambda_{e, \tilde{N}_K} < \lambda_{e, \tilde{N}_{K+1}} \quad (R_{\tilde{N}_K} < R_{\tilde{N}_{K+1}}).$$

(ii) Using definition of  $L$  from equation (2.3), we first show below that  $L_K < L_{K+1} \Leftrightarrow$

$$\sum_{i, j \geq 0: i > j} \sum (i-j) \rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \left( \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) - \prod_{n=j}^{i-1} \bar{F}_K(n) \right) > 0 \quad (\text{A.5})$$

(which also provides an alternative approach to prove Theorem 1/(ii)).  $L_{K+1} > L_K \Leftrightarrow$

$$\begin{aligned} & \frac{\sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)}{\sum_{i=0}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)} > \frac{\sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n)}{\sum_{i=0}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n)} \Leftrightarrow \frac{\sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)}{\sum_{j=0}^{\infty} \rho^j \prod_{n=0}^{j-1} \bar{F}_{K+1}(n)} > \frac{\sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n)}{\sum_{j=0}^{\infty} \rho^j \prod_{n=0}^{j-1} \bar{F}_K(n)} \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \left[ \sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \right] \left[ \sum_{j=0}^{\infty} \rho^j \prod_{n=0}^{j-1} \bar{F}_K(n) \right] > \left[ \sum_{i=0}^{\infty} i \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) \right] \left[ \sum_{j=0}^{\infty} \rho^j \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right], \\
&\Leftrightarrow \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i \rho^{i+j} \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) > \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i \rho^{i+j} \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n), \\
&\Leftrightarrow \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right) > 0, \\
&\Leftrightarrow \sum_{i,j>0: i \neq j} i \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right) > 0, \\
&\Leftrightarrow \sum_{i,j>0: i > j} i \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right) \\
&\quad + \sum_{i,j>0: i < j} i \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right) > 0, \\
&\Leftrightarrow \sum_{i,j>0: i > j} i \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{i-1} \bar{F}_K(n) \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \right) \\
&\quad + \sum_{i,j>0: i > j} j \rho^{i+j} \left( \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{i-1} \bar{F}_K(n) - \prod_{n=0}^{j-1} \bar{F}_K(n) \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \right) > 0.
\end{aligned}$$

Regrouping again, gives

$$\begin{aligned}
&\sum_{i,j>0: i > j} \left( (i-j) \rho^{i+j} \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) + (j-i) \rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{i-1} \bar{F}_K(n) \right) > 0, \\
&\Leftrightarrow \sum_{i,j>0: i > j} (i-j) \rho^{i+j} \left( \prod_{n=0}^{i-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) - \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{i-1} \bar{F}_K(n) \right) > 0, \\
&\Leftrightarrow \sum_{i,j>0: i > j} (i-j) \rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \left( \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) - \prod_{n=j}^{i-1} \bar{F}_K(n) \right) > 0.
\end{aligned}$$

Since,  $a_{K_n}$  is the largest value on the support of  $\bar{F}_K$ , we have  $\bar{F}_{K+1}(i-1) = \bar{F}_K(i-1) = 0$  for  $i \in \{a_{K_n} + 1, a_{K_n} + 2, \dots\}$ . Hence, those indices can be dropped, which proves (A.5):

$$L_K < L_{K+1} \Leftrightarrow \sum_{a_{K_n} \geq i} \sum_{j \geq 0} (i-j) \rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \left( \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) - \prod_{n=j}^{i-1} \bar{F}_K(n) \right) > 0.$$

$$\begin{aligned}
\text{Let us define } A_{K+1}(i, j) &\triangleq (i-j)\rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \prod_{n=j}^{i-1} \bar{F}_{K+1}(n); \\
A_K(i, j) &\triangleq (i-j)\rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \prod_{n=j}^{i-1} \bar{F}_K(n). \tag{A.6}
\end{aligned}$$

Then (A.5) reduces to  $\sum_{a_{K_n} \geq i > j \geq 0} [A_{K+1}(i, j) - A_K(i, j)] > 0$ .

Similar to the approach used in the proof of part (i), our strategy is to form a partition of  $(i, j)$  based on the sign of  $A_{K+1}(i, j) - A_K(i, j)$ . The underlying space is the 2-dimensional set  $\{(i, j) : a_{K_n} \geq i > j \geq 0\}$ . Since  $A_{K+1}(i, j) > A_K(i, j)$  if and only if  $\prod_{n=j}^{i-1} \bar{F}_{K+1}(n) > \prod_{n=j}^{i-1} \bar{F}_K(n)$ , we shall seek a partition over  $\{(i, j) : a_{K_n} \geq i > j \geq 0\}$  based on the sign of  $\prod_{n=j}^{i-1} \bar{F}_{K+1}(n) - \prod_{n=j}^{i-1} \bar{F}_K(n)$  instead.

Define  $\mathcal{G}_1 \triangleq \{a_{K_n}\} \times \{0, 1, 2, \dots, a_{K_n} - 1\}$ ;  $\mathcal{G}_2 \triangleq \{a_{K_1} + 1, a_{K_1} + 2, \dots, a_{K_n} - 2, a_{K_n} - 1\} \times \{0, 1, 2, \dots, a_{K_1}\}$ ; and  $\mathcal{G}_3 \triangleq \{(i, j) : a_{K_n} \geq i > j \geq 0\} - \{\mathcal{G}_1 \cup \mathcal{G}_2\}$ , i.e.,  $\mathcal{G}_3$  contains all the elements that are not in  $\mathcal{G}_1$  or  $\mathcal{G}_2$ .

From (A.2) and (A.3), we can verify that

$$\begin{aligned}
\forall (i, j) \in \mathcal{G}_1 : \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) &< \prod_{n=j}^{i-1} \bar{F}_K(n) \Rightarrow A_{K+1}(i, j) - A_K(i, j) < 0; \\
\forall (i, j) \in \mathcal{G}_2 : \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) &> \prod_{n=j}^{i-1} \bar{F}_K(n) \Rightarrow A_{K+1}(i, j) - A_K(i, j) > 0; \\
\forall (i, j) \in \mathcal{G}_3 : \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) &= \prod_{n=j}^{i-1} \bar{F}_K(n) \Rightarrow A_{K+1}(i, j) - A_K(i, j) = 0.
\end{aligned}$$

Since  $\mathcal{G}_3$  contains all  $(i, j)$  where  $A_{K+1}(i, j) - A_K(i, j) = 0$ , it suffices to just show that

$$\sum_{(i, j) \in \mathcal{G}_1 \cup \mathcal{G}_2} [A_{K+1}(i, j) - A_K(i, j)] > 0 \text{ as a goal. Also since } A_{K+1}(i, j) - A_K(i, j) > 0,$$

$\forall (i, j) \in \mathcal{G}_2$ , the inequality will hold if there exists a subset  $\mathcal{G}_{2'} \subseteq \mathcal{G}_2$  such that

$$\sum_{(i,j) \in \mathcal{G}_1 \cup \mathcal{G}_{2'}} [A_{K+1}(i, j) - A_K(i, j)] > 0. \quad (\text{A.7})$$

We shall prove that the sufficient condition on  $\rho$  stated in the lemma guarantees for inequality (A.7) to hold. To do that, we consider the elements of  $\mathcal{G}_1$  and  $\mathcal{G}_2$  in greater detail.

From the construction of the partition above, we have  $|\mathcal{G}_1| = a_{K_n}$ , i.e., there are  $a_{K_n}$  pairs of  $(i, j)$  in  $\mathcal{G}_1$ , represented by  $\{(a_{K_n}, a_{K_n} - 1), (a_{K_n}, a_{K_n} - 2), (a_{K_n}, a_{K_n} - 3), \dots, (a_{K_n}, 1), (a_{K_n}, 0)\}$ . On the other hand,  $|\mathcal{G}_2| = (a_{K_n} - a_{K_1} - 1)(a_{K_1} + 1)$ . Treating  $a_{K_n} - a_{K_1} - 1$  and  $a_{K_1} + 1$  as base and height of a rectangular and using the fact that a rectangular shape of fixed perimeter ( $a_{K_n}$ ) contains less area ( $|\mathcal{G}_2|$ ) when the shape is more asymmetric, we can show that  $|\mathcal{G}_2| \geq a_{K_n} - 1$ , with equality holds only when  $a_{K_1} = a_{K_n} - 2$  (or  $a_{K_1} = 0$  which is not possible).

The  $a_{K_n} - 1$  pairs of  $(i, j)$  that are guaranteed to reside in  $\mathcal{G}_2$  can be parametrized as  $\{(a_{K_1} + 1, a_{K_1}), (a_{K_1} + 1, a_{K_1} - 1), (a_{K_1} + 1, a_{K_1} - 2), \dots, (a_{K_1} + 1, 1), (a_{K_1} + 1, 0), (a_{K_1} + 2, 0), (a_{K_1} + 3, 0), \dots, (a_{K_n} - 2, 0), (a_{K_n} - 1, 0)\}$ . We therefore define this set to be  $\mathcal{G}_{2'}$ . Note that  $|\mathcal{G}_1| = a_{K_n}$  and  $|\mathcal{G}_{2'}| = a_{K_n} - 1$ . We now order elements of  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  in a specific way displayed in Table 2 (each element in either group is itself an  $(i, j)$  pair).

We denote  $\mathcal{G}_1^l : l \in \{1, 2, \dots, a_{K_n}\}$  and  $\mathcal{G}_{2'}^l : l \in \{1, 2, \dots, a_{K_n} - 1\}$  the  $l$ -th element in  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$ , respectively, according to the order specified in Table 2. Furthermore, for each  $\mathcal{G}_1^l$  and each  $\mathcal{G}_{2'}^l$ , we specify its Cartesian coordinates by subscript  $i$  and  $j$ , i.e.,  $\mathcal{G}_1^l = (\{\mathcal{G}_1^l\}_i, \{\mathcal{G}_1^l\}_j)$  and  $\mathcal{G}_{2'}^l = (\{\mathcal{G}_{2'}^l\}_i, \{\mathcal{G}_{2'}^l\}_j)$ . For example,  $\{\mathcal{G}_1^{a_{K_n}}\}_i = a_{K_n}$  and  $\{\mathcal{G}_{2'}^{a_{K_1}}\}_j = 1$ . We note that  $\forall l \in \{1, 2, \dots, a_{K_n} - 1\}$ ,

$$\{\mathcal{G}_1^l\}_i = a_{K_n} = (a_{K_n} - 1) + 1 \geq \{\mathcal{G}_{2'}^l\}_i + 1 > \{\mathcal{G}_{2'}^l\}_i \quad (\text{A.8})$$

$$\{\mathcal{G}_1^l\}_i - \{\mathcal{G}_1^l\}_j = l = \{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j \quad (\text{A.9})$$

Order $l$	$\mathcal{G}_1$ contains:	$\mathcal{G}_{2'}$ contains:
$l = 1$	$(a_{K_n}, a_{K_n} - 1)$	$(a_{K_1} + 1, a_{K_1})$
$l = 2$	$(a_{K_n}, a_{K_n} - 2)$	$(a_{K_1} + 1, a_{K_1} - 1)$
$l = 3$	$(a_{K_n}, a_{K_n} - 3)$	$(a_{K_1} + 1, a_{K_1} - 2)$
$\vdots$	$\vdots$	$\vdots$
$l = a_{K_1} - 1$	$(a_{K_n}, a_{K_n} - a_{K_1} + 1)$	$(a_{K_1} + 1, 2)$
$l = a_{K_1}$	$(a_{K_n}, a_{K_n} - a_{K_1})$	$(a_{K_1} + 1, 1)$
$l = a_{K_1} + 1$	$(a_{K_n}, a_{K_n} - a_{K_1} - 1)$	$(a_{K_1} + 1, 0)$
$l = a_{K_1} + 2$	$(a_{K_n}, a_{K_n} - a_{K_1} - 2)$	$(a_{K_1} + 2, 0)$
$l = a_{K_1} + 3$	$(a_{K_n}, a_{K_n} - a_{K_1} - 3)$	$(a_{K_1} + 3, 0)$
$\vdots$	$\vdots$	$\vdots$
$l = a_{K_n} - 2$	$(a_{K_n}, 2)$	$(a_{K_n} - 2, 0)$
$l = a_{K_n} - 1$	$(a_{K_n}, 1)$	$(a_{K_n} - 1, 0)$
$l = a_{K_n}$	$(a_{K_n}, 0)$	

Table 2:  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  used in the proof of Lemma 3/(ii)

$$\{\mathcal{G}_1^l\}_i + \{\mathcal{G}_1^l\}_j = a_{K_n} + \{\mathcal{G}_1^l\}_j = 2a_{K_n} - l > 2(a_{K_n} - 1) - l \geq 2\{\mathcal{G}_{2'}^l\}_i - l \geq \{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j \quad (\text{A.10})$$

Recall from (A.7), our goal is to find a sufficient condition such that the summation of  $A_{K+1}(i, j) - A_K(i, j)$  over all  $(i, j)$  in  $\mathcal{G}_1 \cup \mathcal{G}_{2'}$  is positive. We describe all the elements of  $\mathcal{G}_1 \cup \mathcal{G}_{2'}$  by considering the first  $(a_{K_n} - 2)$  rows of  $\mathcal{G}_1^l$  and  $\mathcal{G}_{2'}^l$  in Table 2 plus the last three elements at the  $(a_{K_n} - 1)$ -th and the  $a_{K_n}$ -th rows of the table (namely  $\mathcal{G}_1^{a_{K_n}-1}$ ,  $\mathcal{G}_{2'}^{a_{K_n}-1}$  and  $\mathcal{G}_1^{a_{K_n}}$ ) from Table 2. Therefore, one set of sufficient conditions for (A.7) to hold is (a)  $\forall l \in \{1, 2, \dots, a_{K_n} - 2\}$ ,  $\sum_{(i,j) \in \{\mathcal{G}_1^l, \mathcal{G}_{2'}^l\}} [A_{K+1}(i, j) - A_K(i, j)] > 0$  and (b)

$$\sum_{(i,j) \in \{\mathcal{G}_1^{a_{K_n}-1}, \mathcal{G}_{2'}^{a_{K_n}-1}, \mathcal{G}_1^{a_{K_n}}\}} [A_{K+1}(i, j) - A_K(i, j)] > 0.$$

We first show that (a) is true for all  $\rho$ .  $\forall l \in \{1, 2, \dots, a_{K_n} - 2\}$ . Recall that  $[A_{K+1}(i, j) - A_K(i, j)]$  evaluated at  $(i, j) = \mathcal{G}_1^l$  is negative, and  $[A_{K+1}(i, j) - A_K(i, j)]$  evaluated at

$(i, j) = \mathcal{G}_{2'}^l$  is positive. It is thus equivalent to show that

$$[A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i, j) = \mathcal{G}_1^l} < [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i, j) = \mathcal{G}_{2'}^l}.$$

Denote  $d = \min\{f_K(a_{K_1}), f_K(a_{K_n})\} > 0$ , we have from (A.6) that

$$\begin{aligned} & [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i, j) = \mathcal{G}_1^l} \\ &= (\{\mathcal{G}_1^l\}_i - \{\mathcal{G}_1^l\}_j) \rho^{\{\mathcal{G}_1^l\}_i + \{\mathcal{G}_1^l\}_j} \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_K(n) \left( \prod_{n=\{\mathcal{G}_1^l\}_j}^{\{\mathcal{G}_1^l\}_i - 1} \bar{F}_K(n) - \prod_{n=\{\mathcal{G}_1^l\}_j}^{\{\mathcal{G}_1^l\}_i - 1} \bar{F}_{K+1}(n) \right) \\ &= (a_{K_n} - \{\mathcal{G}_1^l\}_j) \rho^{a_{K_n} + \{\mathcal{G}_1^l\}_j} \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_K(n) \left( \prod_{n=\{\mathcal{G}_1^l\}_j}^{a_{K_n} - 1} \bar{F}_K(n) - \prod_{n=\{\mathcal{G}_1^l\}_j}^{a_{K_n} - 1} \bar{F}_{K+1}(n) \right) \\ &= l \cdot \rho^{a_{K_n} + \{\mathcal{G}_1^l\}_j} \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_K(n) \left( \prod_{n=\{\mathcal{G}_1^l\}_j}^{a_{K_n} - 1} \bar{F}_K(n) - \prod_{n=\{\mathcal{G}_1^l\}_j}^{a_{K_n} - 1} \bar{F}_{K+1}(n) \right) \\ &< l \cdot \rho^{a_{K_n} + \{\mathcal{G}_1^l\}_j} \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_K(n) \left( d \cdot \prod_{n=\{\mathcal{G}_1^l\}_j}^{a_{K_n} - 2} \bar{F}_K(n) \right) \\ &= l \cdot d \cdot \rho^{a_{K_n} + \{\mathcal{G}_1^l\}_j} \prod_{n=0}^{\{\mathcal{G}_1^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{a_{K_n} - 2} \bar{F}_K(n) \\ &< (\{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j) \cdot d \cdot \rho^{\{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j} \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{a_{K_n} - 2} \bar{F}_K(n) \\ &\quad \text{because } l = \{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j \text{ see (A.9), } \rho < 1 \text{ and } a_{K_n} + \{\mathcal{G}_1^l\}_j > \{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j \text{ see (A.10)} \\ &\leq (\{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j) \cdot d \cdot \rho^{\{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j} \prod_{n=0}^{a_{K_n} - 2} \bar{F}_K(n) \\ &\leq (\{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j) \cdot d \cdot \rho^{\{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j} \prod_{n=a_{K_1} + 1}^{\{\mathcal{G}_{2'}^l\}_i - 1} \bar{F}_K(n) \text{ because } a_{K_n} - 2 \geq \{\mathcal{G}_{2'}^l\}_i - 1 \text{ see (A.8)} \\ &= (\{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j) \rho^{\{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j} \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_K(n) \left( d \cdot \prod_{n=a_{K_1} + 1}^{\{\mathcal{G}_{2'}^l\}_i - 1} \bar{F}_K(n) \right) \\ &\quad \text{because } \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_K(n) = 1 \\ &= (\{\mathcal{G}_{2'}^l\}_i - \{\mathcal{G}_{2'}^l\}_j) \rho^{\{\mathcal{G}_{2'}^l\}_i + \{\mathcal{G}_{2'}^l\}_j} \times \\ &\quad \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_{K+1}(n) \prod_{n=0}^{\{\mathcal{G}_{2'}^l\}_j - 1} \bar{F}_K(n) \left( \prod_{n=\{\mathcal{G}_{2'}^l\}_j}^{\{\mathcal{G}_{2'}^l\}_i - 1} \bar{F}_{K+1}(n) - \prod_{n=\{\mathcal{G}_{2'}^l\}_j}^{\{\mathcal{G}_{2'}^l\}_i - 1} \bar{F}_K(n) \right) \end{aligned}$$



because  $\prod_{n=\{\mathcal{G}_{2'}^l\}_j}^{a_{K_1}-1} \bar{F}_{K+1}(n) = \prod_{n=\{\mathcal{G}_{2'}^l\}_j}^{a_{K_1}-1} \bar{F}_K(n) = 1$ ,  $\bar{F}_{K+1}(a_{K_1}) - \bar{F}_K(a_{K_1}) = d$   
and  $\bar{F}_{K+1}(x) = \bar{F}_K(x)$ ,  $\forall x \in \{a_{K_1} + 1, \dots, \{\mathcal{G}_{2'}^l\}_i - 1\}$   
 $= [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i,j)=\mathcal{G}_{2'}^l}$ , as required.

Next, for (b) to hold, i.e.,  $\sum_{(i,j) \in \{\mathcal{G}_1^{a_{K_n}-1}, \mathcal{G}_{2'}^{a_{K_n}}, \mathcal{G}_1^{a_{K_n}}\}} [A_{K+1}(i, j) - A_K(i, j)] > 0$ , we have equivalently:

$$\begin{aligned}
& [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=\mathcal{G}_1^{a_{K_n}-1}} + [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=\mathcal{G}_1^{a_{K_n}}} \\
& < [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i,j)=\mathcal{G}_{2'}^{a_{K_n}-1}} \\
& \Leftrightarrow [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n},1)} + [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n},0)} \\
& < [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i,j)=(a_{K_n}-1,0)} \tag{A.11}
\end{aligned}$$

Note that (with any empty product being equal to = 1)

$$\begin{aligned}
& [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n},1)} = (a_{K_n} - 1)\rho^{a_{K_n}+1} \cdot \left( \prod_{n=1}^{a_{K_n}-1} \bar{F}_{K+1}(n) - \prod_{n=1}^{a_{K_n}-1} \bar{F}_K(n) \right) \\
& < (a_{K_n} - 1)\rho^{a_{K_n}+1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n), \\
& [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n},0)} = a_{K_n}\rho^{a_{K_n}} \cdot \left( \prod_{n=0}^{a_{K_n}-1} \bar{F}_{K+1}(n) - \prod_{n=0}^{a_{K_n}-1} \bar{F}_K(n) \right) \\
& < a_{K_n}\rho^{a_{K_n}} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n), \\
& [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i,j)=(a_{K_n}-1,0)} = (a_{K_n} - 1)\rho^{a_{K_n}-1} \cdot \left( \prod_{n=0}^{a_{K_n}-2} \bar{F}_{K+1}(n) - \prod_{n=0}^{a_{K_n}-2} \bar{F}_K(n) \right) \\
& = (a_{K_n} - 1)\rho^{a_{K_n}-1} \cdot \left( \prod_{n=a_{K_1}}^{a_{K_n}-2} \bar{F}_{K+1}(n) - \prod_{n=a_{K_1}}^{a_{K_n}-2} \bar{F}_K(n) \right) = (a_{K_n} - 1)\rho^{a_{K_n}-1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n).
\end{aligned}$$

Therefore, it is sufficient for (A.11) to hold if

$$\begin{aligned}
& (a_{K_n} - 1)\rho^{a_{K_n}+1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) + a_{K_n}\rho^{a_{K_n}} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) \\
& \leq (a_{K_n} - 1)\rho^{a_{K_n}-1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) \\
& \Leftrightarrow (a_{K_n} - 1)\rho^{a_{K_n}+1} + a_{K_n}\rho^{a_{K_n}} \leq (a_{K_n} - 1)\rho^{a_{K_n}-1} \Leftrightarrow \rho^2 + \frac{a_{K_n}}{a_{K_n}-1}\rho \leq 1. \quad (\text{A.12})
\end{aligned}$$

Solving quadratic equation (A.12) gives the condition

$$\frac{1}{2} \left( -\sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right) \leq \rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right).$$

where it is clear that  $\frac{1}{2} \left( -\sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right) < 0$  and  $0 < \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right) < 1$ .

Since  $\rho \in (0, 1)$ , we conclude that, when  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}}{a_{K_n}-1} \right)$ , (A.12) (A.11), (A.7) and (A.5) all hold and thus it is a sufficient condition for  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$ . This completes the proof of part (ii).

(iii) Using the definition of  $W$  from equation (2.8), and comparing the structure of equation (2.3) to that of (2.8), it can be shown via a similar approach used in the proof of inequality (A.5) that

$$W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}} \Leftrightarrow \sum_{i,j \geq 1} \sum_{i > j} (i-j)\rho^{i+j} \prod_{n=0}^{j-1} \bar{F}_{K+1}(n) \prod_{n=0}^{j-1} \bar{F}_K(n) \left( \prod_{n=j}^{i-1} \bar{F}_{K+1}(n) - \prod_{n=j}^{i-1} \bar{F}_K(n) \right) > 0, \quad (\text{A.13})$$

which also provides an alternative approach to prove Theorem 1/(iii). Using the same definition of  $A_K(i, j)$  and  $A_{K+1}(i, j)$  from part (ii), we have

$$W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}} \Leftrightarrow \sum_{a_{K_n} \geq i > j \geq 1} [A_{K+1}(i, j) - A_K(i, j)] > 0.$$

The rest of the proof is then almost identical to the proof of part (ii) except now  $i, j$  cannot take on 0. Define  $\mathcal{G}_1 \triangleq \{a_{K_n}\} \times \{1, 2, \dots, a_{K_n} - 1\}$ ,  $\mathcal{G}_2 \triangleq \{a_{K_1} + 1, a_{K_1} + 2, \dots, a_{K_n} - 2, a_{K_n} -$

$1\} \times \{1, 2, \dots, a_{K_1}\}$ , and  $\mathcal{G}_3 \triangleq \{(i, j) : a_{K_n} \geq i > j \geq 1\} - \{\mathcal{G}_1 \cup \mathcal{G}_2\}$ .

There are now at least  $a_{K_n} - 2$  elements in the set  $\mathcal{G}_2$  which defines the subset  $\mathcal{G}_{2'}$ . The elements in  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  are ordered in a similar fashion as before and displayed in Table 3.

Order $l$	$\mathcal{G}_1$ contains:	$\mathcal{G}_{2'}$ contains:
$l = 1$	$(a_{K_n}, a_{K_n} - 1)$	$(a_{K_1} + 1, a_{K_1})$
$l = 2$	$(a_{K_n}, a_{K_n} - 2)$	$(a_{K_1} + 1, a_{K_1} - 1)$
$l = 3$	$(a_{K_n}, a_{K_n} - 3)$	$(a_{K_1} + 1, a_{K_1} - 2)$
$\vdots$	$\vdots$	$\vdots$
$l = a_{K_1} - 1$	$(a_{K_n}, a_{K_n} - a_{K_1} + 1)$	$(a_{K_1} + 1, 2)$
$l = a_{K_1}$	$(a_{K_n}, a_{K_n} - a_{K_1})$	$(a_{K_1} + 1, 1)$
$l = a_{K_1} + 1$	$(a_{K_n}, a_{K_n} - a_{K_1} - 1)$	$(a_{K_1} + 2, 1)$
$l = a_{K_1} + 2$	$(a_{K_n}, a_{K_n} - a_{K_1} - 2)$	$(a_{K_1} + 3, 1)$
$l = a_{K_1} + 3$	$(a_{K_n}, a_{K_n} - a_{K_1} - 3)$	$(a_{K_1} + 4, 1)$
$\vdots$	$\vdots$	$\vdots$
$l = a_{K_n} - 2$	$(a_{K_n}, 2)$	$(a_{K_n} - 1, 1)$
$l = a_{K_n} - 1$	$(a_{K_n}, 1)$	

Table 3:  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  used in the proof of Lemma 3/(iii)

A sufficient condition for  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  from (A.13) is that  $\sum_{(i,j) \in \mathcal{G}_1 \cup \mathcal{G}_{2'}} [A_{K+1}(i, j) - A_K(i, j)] > 0$ . It can be shown that conditions (A.8)-(A.10) still hold, and thus  $\forall l \in \{1, 2, \dots, a_{K_n} - 3\}$  and for all  $\rho$ ,  $\sum_{(i,j) \in \{\mathcal{G}_1^l, \mathcal{G}_{2'}^l\}} [A_{K+1}(i, j) - A_K(i, j)] > 0$ . Therefore, a sufficient condition for  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  is that

$$\sum_{(i,j) \in \{\mathcal{G}_1^{a_{K_n}-2}, \mathcal{G}_{2'}^{a_{K_n}-2}, \mathcal{G}_1^{a_{K_n}-1}\}} [A_{K+1}(i, j) - A_K(i, j)] > 0. \quad (\text{A.14})$$

$$\begin{aligned} \text{Since } [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n}, 2)} &< (a_{K_n} - 2)\rho^{a_{K_n}+2} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n), \\ [A_K(i, j) - A_{K+1}(i, j)] \Big|_{(i,j)=(a_{K_n}, 1)} &< (a_{K_n} - 1)\rho^{a_{K_n}+1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n), \end{aligned}$$

$$\text{and } [A_{K+1}(i, j) - A_K(i, j)] \Big|_{(i,j)=(a_{K_n-1}, 1)} = (a_{K_n} - 2)\rho^{a_{K_n}} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n),$$

condition (A.14) will hold if

$$\begin{aligned} & (a_{K_n} - 2)\rho^{a_{K_n}+2} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) + (a_{K_n} - 1)\rho^{a_{K_n}+1} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) \\ & \leq (a_{K_n} - 2)\rho^{a_{K_n}} \cdot d \cdot \prod_{n=a_{K_1}+1}^{a_{K_n}-2} \bar{F}_K(n) \\ & \Leftrightarrow (a_{K_n} - 2)\rho^{a_{K_n}+2} + (a_{K_n} - 1)\rho^{a_{K_n}+1} \leq (a_{K_n} - 2)\rho^{a_{K_n}} \Leftrightarrow \rho^2 + \frac{a_{K_n}-1}{a_{K_n}-2}\rho \leq 1. \end{aligned}$$

The solution of the quadratic inequality on the set  $\rho \in (0, 1)$  is

$$\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-2}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-2} \right). \quad \square$$

**Proof of Theorem 2:**

(i) Result follows immediately from Lemma 3/(i) since  $\lambda_{eff, \tilde{N}_K} < \lambda_{eff, \tilde{N}_{K+1}}$  ( $R_{\tilde{N}_K} < R_{\tilde{N}_{K+1}}$ ) for all  $K$ .

(ii) Recall from Lemma 3/(ii) that  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  if  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-1} \right)$ . It can be easily verified that  $\frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-1}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-1} \right)$  increases in  $a_{K_n}$ . Plugging in the smallest possible value of  $a_{K_n}$  which is 3, we get  $\rho = 0.5$ . Therefore, when  $\rho \leq 0.5$ ,  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  for all  $K$  (regardless of the distributions of  $\{\tilde{N}_K\}_{K=0,1,2,\dots,T}$ ). Result thus follows. Note that it is possible to derive stronger distribution-specific conditions.

(iii) Recall from Lemma 3/(iii) that  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  if  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-2}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-2} \right)$ . It can be verified that  $\frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-2}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-2} \right)$  increases in  $a_{K_n}$ . Plugging in the smallest possible value of  $a_{K_n}$  which is 3, we get  $\rho = 0.414$ . Therefore, when  $\rho \leq 0.414$ ,  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  for all  $K$ . Result thus follows. Again, it is possible to derive stronger distribution-specific conditions.  $\square$

**Proof of Theorem 1':**

(i) From (2.12), we then have  $R_{\tilde{N}} = p \cdot \mu[s - (s\pi_0 + (s-1)\pi_1 + \dots + 2\pi_{s-2} + 1\pi_{s-1})]$  so

$R_{\tilde{N}}$  is decreasing in  $\pi_0$  (given that  $p, \mu, \lambda, \rho, s$  are all fixed). Since  $\pi_0$  itself is decreasing in  $\sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n)$  from (2.11), we have  $R_{\tilde{N}}$  increases in  $\sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n)$ . The rest of the proof follows the proof of Theorem 1/(i).

(ii) Proof is similar to that of Theorem 1/(ii). Result follows because  $L_{\tilde{N}} \leq L_{\tilde{N}'}$  iff

$$\sum_{i,j \geq 0: i > j} \sum (i-j) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{\tilde{N}'}(n) \prod_{n=0}^{j-s} \bar{F}_{\tilde{N}}(n) \left( \prod_{n=j-s+1}^{i-s} \bar{F}_{\tilde{N}'}(n) - \prod_{n=j-s+1}^{i-s} \bar{F}_{\tilde{N}}(n) \right) \geq 0.$$

(iii) Proof is similar to that of Theorem 1/(iii). We have  $W_{\tilde{N}} \leq W_{\tilde{N}'}$  iff

$$\sum_{i,j \geq 1: i > j} \sum (i(j \wedge s) - j(i \wedge s)) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{\tilde{N}'}(n) \prod_{n=0}^{j-s} \bar{F}_{\tilde{N}}(n) \left( \prod_{n=j-s+1}^{i-s} \bar{F}_{\tilde{N}'}(n) - \prod_{n=j-s+1}^{i-s} \bar{F}_{\tilde{N}}(n) \right) \geq 0$$

and result follows because  $(i(j \wedge s) - j(i \wedge s)) \geq 0$  for all  $\{i, j \geq 1 : i > j\}$ .  $\square$

### **Proof of Theorem 2':**

We will prove the following lemma (a general version of Lemma 3 but with the  $M/M/s$  queue setting) then the results of Theorem 2' immediately follow.

Let  $\{\tilde{N}_K\}$  be any sequence from Construction 1 in an  $M/M/s$  queue. We can show (i)  $R_{\tilde{N}_K} < R_{\tilde{N}_{K+1}}$  for all  $\rho$ ; (ii)  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  if  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}+s-1}{a_{K_n}+s-2}\right)^2 + 4} - \frac{a_{K_n}+s-1}{a_{K_n}+s-2} \right)$ ; And (iii) when  $s = 1$ ,  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  if  $\rho \leq \frac{1}{2} \left( \sqrt{\left(\frac{a_{K_n}-1}{a_{K_n}-2}\right)^2 + 4} - \frac{a_{K_n}-1}{a_{K_n}-2} \right)$ ; when  $s \geq 2$ ,  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  if  $\rho \leq \frac{1}{2} \left( \sqrt{1 + 4\left(\frac{a_{K_n}-2}{a_{K_n}-1}\right)} - 1 \right)$ .

(i) Recall from the proof of Theorem 1'/(i),  $R_{\tilde{N}}$  increases in  $\sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}(n)$ . Therefore it is sufficient to show that  $\sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_K(n) < \sum_{i=1}^{\infty} \rho^i \prod_{n=0}^{i-1} \bar{F}_{K+1}(n)$ . Result then follows the proof of Lemma 3/(i).

(ii) We can apply the same approach used in the proof of Lemma 3/(ii) here. Define

$$A_{K+1}(i, j) \triangleq (i-j) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{K+1}(n) \prod_{n=0}^{j-s} \bar{F}_K(n) \prod_{n=j-s+1}^{i-s} \bar{F}_{K+1}(n)$$

$$A_K(i, j) \triangleq (i - j) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{K+1}(n) \prod_{n=0}^{j-s} \bar{F}_K(n) \prod_{n=j-s+1}^{i-s} \bar{F}_K(n)$$

It can be shown that  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$  if  $\sum_{(i,j) \in \mathcal{G}_1 \cup \mathcal{G}_{2'}} [A_{K+1}(i, j) - A_K(i, j)] > 0$  where the elements of  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  are listed in Table 4.

Order $l =$	$\mathcal{G}_1$ contains:	$\mathcal{G}_{2'}$ contains:
1	$(a_{K_n} + s - 1, a_{K_n} + s - 2)$	$(a_{K_1} + s, a_{K_1} + s - 1)$
2	$(a_{K_n} + s - 1, a_{K_n} + s - 3)$	$(a_{K_1} + s, a_{K_1} + s - 2)$
3	$(a_{K_n} + s - 1, a_{K_n} + s - 4)$	$(a_{K_1} + s, a_{K_1} + s - 3)$
$\vdots$	$\vdots$	$\vdots$
$a_{K_1} - 1$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1})$	$(a_{K_1} + s, 2)$
$a_{K_1}$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 1)$	$(a_{K_1} + s, 1)$
$a_{K_1} + 1$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 2)$	$(a_{K_1} + s, 0)$
$a_{K_1} + 2$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 3)$	$(a_{K_1} + s + 1, 0)$
$a_{K_1} + 3$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 4)$	$(a_{K_1} + s + 2, 0)$
$\vdots$	$\vdots$	$\vdots$
$a_{K_n} + s - 3$	$(a_{K_n} + s - 1, 2)$	$(a_{K_n} + s - 3, 0)$
$a_{K_n} + s - 2$	$(a_{K_n} + s - 1, 1)$	$(a_{K_n} + s - 2, 0)$
$a_{K_n} + s - 1$	$(a_{K_n} + s - 1, 0)$	

Table 4:  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  used in the proof of Theorem 2'/(ii)

It then can be verified that conditions (A.8)-(A.10) still hold, and that  $\forall l \in \{1, 2, \dots, a_{K_n} + s - 3\}$  and for all  $\rho$ ,  $\sum_{(i,j) \in \{\mathcal{G}_1^l, \mathcal{G}_{2'}^l\}} [A_{K+1}(i, j) - A_K(i, j)] > 0$ . One sufficient condition for

$$\sum_{(i,j) \in \{\mathcal{G}_1^{a_{K_n}+s-2}, \mathcal{G}_{2'}^{a_{K_n}+s-2}, \mathcal{G}_1^{a_{K_n}+s-1}\}} [A_{K+1}(i, j) - A_K(i, j)] > 0,$$

which also makes  $L_{\tilde{N}_K} < L_{\tilde{N}_{K+1}}$ , is that

$$(a_{K_n} + s - 2)\rho^{a_{K_n}+s} + (a_{K_n} + s - 1)\rho^{a_{K_n}+s-1} \leq (a_{K_n} + s - 2)\rho^{a_{K_n}+s-2}.$$

It follows by solving the quadratic equation that  $\rho \leq \frac{1}{2} \left( \sqrt{\frac{(a_{K_n}+s-1)^2}{(a_{K_n}+s-2)^2} + 4} - \frac{a_{K_n}+s-1}{a_{K_n}+s-2} \right)$ .

(iii) We can apply the same approach used in the proof of Lemma 3/(iii) here. Define

$$A_{K+1}(i, j) \triangleq (i(j \wedge s) - j(i \wedge s)) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{K+1}(n) \prod_{n=0}^{j-s} \bar{F}_K(n) \prod_{n=j-s+1}^{i-s} \bar{F}_{K+1}(n)$$

$$A_K(i, j) \triangleq (i(j \wedge s) - j(i \wedge s)) \frac{\rho^{i+j}}{(i \wedge s)!(j \wedge s)!} \prod_{n=0}^{j-s} \bar{F}_{K+1}(n) \prod_{n=0}^{j-s} \bar{F}_K(n) \prod_{n=j-s+1}^{i-s} \bar{F}_K(n)$$

It can be shown that  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$  if  $\sum_{(i,j) \in \mathcal{G}_1 \cup \mathcal{G}_{2'}} [A_{K+1}(i, j) - A_K(i, j)] > 0$  where the elements of  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  are listed in Table 5.

Order $l =$	$\mathcal{G}_1$ contains:	$\mathcal{G}_{2'}$ contains:
1	$(a_{K_n} + s - 1, a_{K_n} + s - 2)$	$(a_{K_1} + s, a_{K_1} + s - 1)$
2	$(a_{K_n} + s - 1, a_{K_n} + s - 3)$	$(a_{K_1} + s, a_{K_1} + s - 2)$
3	$(a_{K_n} + s - 1, a_{K_n} + s - 4)$	$(a_{K_1} + s, a_{K_1} + s - 3)$
$\vdots$	$\vdots$	$\vdots$
$a_{K_1} - 1$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1})$	$(a_{K_1} + s, 2)$
$a_{K_1}$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 1)$	$(a_{K_1} + s, 1)$
$a_{K_1} + 1$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 2)$	$(a_{K_1} + s + 1, 1)$
$a_{K_1} + 2$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 3)$	$(a_{K_1} + s + 2, 1)$
$a_{K_1} + 3$	$(a_{K_n} + s - 1, a_{K_n} + s - a_{K_1} - 4)$	$(a_{K_1} + s + 3, 1)$
$\vdots$	$\vdots$	$\vdots$
$a_{K_n} + s - 3$	$(a_{K_n} + s - 1, 2)$	$(a_{K_n} + s - 2, 1)$
$a_{K_n} + s - 2$	$(a_{K_n} + s - 1, 1)$	

Table 5:  $\mathcal{G}_1$  and  $\mathcal{G}_{2'}$  used in the proof of Theorem 2'/(iii)

It then can be verified that  $\forall l \in \{1, 2, \dots, a_{K_n} + s - 4\}$  and for all  $\rho$ ,

$$\sum_{(i,j) \in \{\mathcal{G}_1^l, \mathcal{G}_{2'}^l\}} [A_{K+1}(i, j) - A_K(i, j)] > 0.$$

Further, when  $s \geq 2$  (the case when  $s = 1$  is proved in Lemma 3/(iii)), one sufficient condition for

$\sum_{(i,j) \in \{\mathcal{G}_1^{a_{K_n}+s-3}, \mathcal{G}_{2'}^{a_{K_n}+s-3}, \mathcal{G}_1^{a_{K_n}+s-2}\}} [A_{K+1}(i, j) - A_K(i, j)] > 0$ , which leads to  $W_{\tilde{N}_K} < W_{\tilde{N}_{K+1}}$ , is that

$$(a_{K_n} - 1)\rho^{a_{K_n}+s+1} + (a_{K_n} - 1)\rho^{a_{K_n}+s} \leq (a_{K_n} - 2)\rho^{a_{K_n}+s-1}.$$

It then follows by solving the quadratic equation that  $\rho \leq \frac{1}{2} \left( \sqrt{1 + 4 \left( \frac{a_{K_n} - 2}{a_{K_n} - 1} \right)} - 1 \right)$ .  $\square$

**Proof of Proposition 1:**

Consider the random variable  $\tilde{N}_T \in \{\lfloor \mathbb{E}(\tilde{N}) \rfloor, \lceil \mathbb{E}(\tilde{N}) \rceil\}$  such that  $E(\tilde{N}_T) = E(\tilde{N})$ . By Theorem 2, we have  $R_{\tilde{N}} \leq R_{\tilde{N}_T}$  for all  $\rho$ , and  $L_{\tilde{N}} \leq L_{\tilde{N}_T}$ ,  $W_{\tilde{N}} \leq W_{\tilde{N}_T}$  for small  $\rho$ . On the other hand, since  $\mathbb{E}(\tilde{N}) \leq N$  and  $N$  is an integer, we must have  $\lceil \mathbb{E}(\tilde{N}) \rceil \leq N$ . It follows that  $\tilde{N}_T \leq_{st} N$  so by Theorem 1, we have  $R_{\tilde{N}_T} \leq R_N$  for all  $\rho$ , and  $L_{\tilde{N}_T} \leq L_N$ ,  $W_{\tilde{N}_T} \leq W_N$  for small  $\rho$ . Result thus follows.  $\square$

**Proof of Theorem 1<sup>U</sup>:**

(i) For  $t \in \{\tilde{\mu}, \tilde{\mu}'\}$ , we have  $\lambda_{eff,t}^U = \int_{-\infty}^{\infty} \lambda \cdot q(t) dG_t = \left[ \int_{\frac{c}{v-p}}^{\lambda + \frac{c}{v-p}} (t - \frac{c}{v-p}) + \int_{\lambda + \frac{c}{v-p}}^{\infty} \lambda \right] dG_v = \mathbb{E}_t \min\{t - \frac{c}{v-p}, \lambda\}$ . (Throughout the paper, we assume  $\tilde{\mu} \geq \frac{c}{v-p}$ .) Therefore,  $\tilde{\mu} \leq_{st} \tilde{\mu}' \Rightarrow \tilde{\mu} - \frac{c}{v-p} \leq_{st} \tilde{\mu}' - \frac{c}{v-p} \Rightarrow \min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq_{st} \min\{\tilde{\mu}' - \frac{c}{v-p}, \lambda\} \Rightarrow \mathbb{E}_{\tilde{\mu}} \min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq \mathbb{E}_{\tilde{\mu}'} \min\{\tilde{\mu}' - \frac{c}{v-p}, \lambda\} \Rightarrow \lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\tilde{\mu}'}^U$ . Also since  $R_t^U = p\lambda_{eff,t}^U$ , it follows that  $R_{\tilde{\mu}}^U \leq R_{\tilde{\mu}'}^U$ .

(ii) For  $t \in \{\tilde{\mu}, \tilde{\mu}'\}$ , we have  $W_t^U = \frac{1}{\mu - \lambda_{eff,t}^U}$ . Since  $\lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\tilde{\mu}'}^U < \lambda$ , then  $W_{\tilde{\mu}}^U \leq W_{\tilde{\mu}'}^U$ .  $\square$

**Proof of Theorem 2<sup>U</sup>:**

(i) Under belief  $\tilde{\mu}$ , we have  $\lambda_{eff,\tilde{\mu}}^U = \mathbb{E}_{\tilde{\mu}} \min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\}$ . Under belief  $\mathbb{E}(\tilde{\mu})$ , we have

$$\lambda_{eff,\mathbb{E}(\tilde{\mu})}^U = \begin{cases} \mathbb{E}(\tilde{\mu}) - \frac{c}{v-p} & \text{if } (\frac{c}{v-p} \leq) \mathbb{E}(\tilde{\mu}) < \lambda + \frac{c}{v-p} \\ \lambda & \text{if } \mathbb{E}(\tilde{\mu}) \geq \lambda + \frac{c}{v-p} \end{cases}$$

If  $(\frac{c}{v-p} \leq) \mathbb{E}(\tilde{\mu}) < \lambda + \frac{c}{v-p}$ , we have  $\min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq_{st} \tilde{\mu} - \frac{c}{v-p} \Rightarrow$

$$\mathbb{E}_{\tilde{\mu}} \min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq_{st} \mathbb{E}_{\tilde{\mu}} (\tilde{\mu} - \frac{c}{v-p}) = \mathbb{E}(\tilde{\mu}) - \frac{c}{v-p} \Rightarrow \lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\mathbb{E}(\tilde{\mu})}^U;$$

If  $\mathbb{E}(\tilde{\mu}) \geq \lambda + \frac{c}{v-p}$ , we have  $\min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq_{st} \lambda \Rightarrow$

$$\mathbb{E}_{\tilde{\mu}} \min\{\tilde{\mu} - \frac{c}{v-p}, \lambda\} \leq_{st} \mathbb{E}(\lambda) = \lambda \Rightarrow \lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\mathbb{E}(\tilde{\mu})}^U;$$



(ii) Result follows from (i). Proof of the result is analogous to that of Theorem 1<sup>U</sup>.  $\square$

**Proof of Proposition 1<sup>U</sup>:**

$\lambda_{eff,\tilde{\mu}}^U \leq \lambda_{eff,\mathbb{E}(\tilde{\mu})}^U$ ,  $R_{\tilde{\mu}}^U \leq R_{\mathbb{E}(\tilde{\mu})}^U$  and  $W_{\tilde{\mu}}^U \leq W_{\mathbb{E}(\tilde{\mu})}^U$  by Theorem 2<sup>U</sup>. Then because  $\mathbb{E}(\tilde{\mu}) \leq \mu$ , we have  $\mathbb{E}(\tilde{\mu}) \leq_{st} \mu$ . It thus follows by Theorem 1<sup>U</sup> that  $\lambda_{eff,\mathbb{E}(\tilde{\mu})}^U \leq \lambda_{eff,\mu}^U$ ,  $R_{\mathbb{E}(\tilde{\mu})}^U \leq R_{\mu}^U$  and  $W_{\mathbb{E}(\tilde{\mu})}^U \leq W_{\mu}^U$ .  $\square$

## APPENDIX TO CHAPTER 3

**Proof of Proposition 2:**

Omitted. Results follow from Definition 3.  $\square$

**Proof of Theorem 3:**

Suppose the population adopts  $JnR$  or  $J_J^R nR$  and  $n > N$ , then some consumers, if not all, will join the queue at state  $N$ . Such strategies cannot be equilibrium strategies because these consumers can do better if they balk at state  $N$ , rather than joining, as  $0 > v - \frac{c}{\mu}(N+1)$ . On the other hand, we assume that every consumer will join (i.e., will not retry) on seeing an idle server, so we must have  $n \geq 1$ . Similarly, if  $JnB$  or  $Jn_B^R$  is adopted by the population and  $n \neq N$ , then either some consumers are specified by the strategy to balk when they are better off not to, or specified to join the queue when better off not to join.  $\square$

**Proof of Proposition 3:**

Let us suppose that everybody else is adopting  $JNB$  on every service occasion, and only one consumer is given the opportunity to change his strategy unilaterally. Then,  $JNB$  will be an equilibrium (i.e., this consumer has no incentive to retry at any state  $\{1, 2, \dots, N-1, N\}$ ) if and only if his expected payoff from a retry decision is less than or equal to 0. Since this consumer's retrial payoff is decreasing in the retrial cost  $\alpha$ , there exists one unique value of  $\alpha$  above which  $JNB$  is an equilibrium and below which it is not. We will show this value is equal to  $\alpha_H$ , given by  $(1 - \pi_N^{JNB})(v - cW^{JNB})$ .

At  $\alpha_H$ , this consumer's retrial payoff is exactly 0, i.e.,

$$\begin{aligned} \pi_0^{JNB}(v - \frac{c}{\mu}) + \pi_1^{JNB}(v - \frac{2c}{\mu}) + \dots + \pi_{N-1}^{JNB}(v - \frac{cN}{\mu}) + \pi_N^{JNB} \cdot 0 - \alpha_H &= 0 \\ (1 - \pi_N^{JNB})v - c(\pi_0^{JNB}\frac{1}{\mu} + \pi_1^{JNB}\frac{2}{\mu} + \dots + \pi_{N-1}^{JNB}\frac{N}{\mu}) - \alpha_H &= 0 \\ (1 - \pi_N^{JNB})v - (1 - \pi_N^{JNB})cW^{JNB} - \alpha_H &= 0 \\ (1 - \pi_N^{JNB})(v - cW^{JNB}) - \alpha_H &= 0 \end{aligned}$$

and therefore result follows.  $\square$

***Proof of Lemma 4:***

If  $\alpha = \alpha^*$  solves the equality  $\frac{c}{\mu}(n^* + 1) = cW^{Jn^*R} + \frac{\alpha}{1 - \pi_n^{Jn^*R}}$  for some  $n^*$ , and assuming the stability condition in (3.12) holds, we have from (3.15) that  $Jn^*R$  generates an equilibrium retry strategy for  $\alpha \in [\alpha^* - \frac{c}{\mu}(1 - \pi_n^{Jn^*R}), \alpha^*]$ . On the other hand, if  $\alpha = \alpha^{**}$  solves the equality  $\frac{c}{\mu}n^{**} = cW^{Jn^{**}R} + \frac{\alpha}{1 - \pi_n^{Jn^{**}R}}$  for some  $n^{**}$ , and if the stability condition holds, again from (3.15) we know that  $Jn^{**}R$  is an equilibrium for  $\alpha \in [\alpha^{**}, \alpha^{**} + \frac{c}{\mu}(1 - \pi_n^{Jn^{**}R})]$ . For a fixed  $n = n^* = n^{**}$ , the set of  $\alpha$  that satisfies the inequality  $\frac{c}{\mu}(n+1) \geq cW^{JnR} + \frac{\alpha}{1 - \pi_n^{JnR}} \geq \frac{c}{\mu}n$  in (3.15) is  $\alpha \in [\alpha^* - \frac{c}{\mu}(1 - \pi_n^{Jn^*R}), \alpha^*]$  which is exactly that same as  $\alpha \in [\alpha^{**}, \alpha^{**} + \frac{c}{\mu}(1 - \pi_n^{Jn^{**}R})]$  by setting  $\alpha^* - \frac{c}{\mu}(1 - \pi_n^{Jn^*R}) = \alpha^{**}$ . In other words, to find all  $(\alpha, n)$  pairs that satisfy the seemingly two inequalities in (3.15), we only need to find solutions  $(\alpha^*, n^*)$  to one equality, say,

$$\frac{c}{\mu}(n+1) = cW^{JnR} + \frac{\alpha}{1 - \pi_n^{JnR}} \quad (\text{with } n \leq N), \quad (\text{A.15})$$

and then all the pairs  $\{(\alpha, n^*) : \alpha \in [\alpha^* - \frac{c}{\mu}\pi_n^{Jn^*R}, \alpha^*]\}$  are solutions to the indifference condition in (3.15).

Since  $W^{JnR} = \frac{1}{\mu} \sum_{k=0}^{n-1} \frac{\pi_k^{JnR}}{1 - \pi_n^{JnR}}(k+1)$ , equation (A.15) becomes

$$\frac{c}{\mu}(n+1) = \frac{c}{\mu} \sum_{k=0}^{n-1} \frac{\pi_k^{JnR}}{1 - \pi_n^{JnR}}(k+1) + \frac{\alpha}{1 - \pi_n^{JnR}}$$

$$\begin{aligned}
&\Leftrightarrow (1 - \pi_n^{JnR}) \frac{c}{\mu} (n+1) = \frac{c}{\mu} \sum_{k=0}^{n-1} \pi_k^{JnR} (k+1) + \alpha \\
&\Leftrightarrow (1 - \pi_n^{JnR}) \frac{c}{\mu} n = \frac{c}{\mu} \sum_{k=0}^{n-1} \pi_k^{JnR} k + \alpha \\
&\Leftrightarrow \frac{c}{\mu} n = \frac{c}{\mu} \sum_{k=0}^n \pi_k^{JnR} k + \alpha \\
&\Leftrightarrow \frac{c}{\mu} (n - \sum_{k=0}^n \pi_k^{JnR} k) = \alpha
\end{aligned} \tag{A.16}$$

By letting  $L^{JnR} \triangleq \sum_{k=0}^n \pi_k^{JnR} k$  denote the long-run average number of consumers in the  $M/M/1/n$  system under the retrial strategy  $JnR$  and  $r \triangleq \alpha/\frac{c}{\mu}$  the cost ratio of the retrial cost over the expected waiting cost for a service cycle, equation (A.16) becomes

$$n - L^{JnR} = r \tag{A.17}$$

Ignore the feasibility condition  $n \leq N$  and the integer condition on  $n$  for now. Then equation (A.17) (the indifference condition) along with equation (3.12) (the stability condition) give us two equations for three unknowns ( $n$ ,  $\rho$  and  $r$ ) where the other parameters  $\lambda, \mu, v, c$  are fixed system inputs. In what follows, we should find solutions  $(\rho^{Jn^*R}, n^*)$  to equations (3.12) and (A.17) in terms of  $r$ . Algebraic operations force us to separate the case when  $\rho^{Jn^*R} = 1$  (so-called *the trivial solution*) with the case when  $\rho^{Jn^*R} \neq 1$  (so-called *the non-trivial solution*).

**Trivial Solution:** We first consider the “trivial” case, i.e., when  $\rho^{Jn^*R} = 1$  is part of the solution to equations (3.12) and (A.17). When  $\rho^{JnR} = 1$ , we have  $\pi_0^{JnR} = \pi_1^{JnR} = \pi_2^{JnR} = \dots = \pi_n^{JnR} = \frac{1}{n+1}$  and  $L^{JnR} = n/2$ . From equation (A.17), we have  $n = 2r$ . Plugging  $n = 2r$  into (3.12), we then have  $\rho^{JnR} = \frac{2r+1}{2r}l$ . Since  $\rho^{JnR} = 1$ , we must have  $\frac{2r+1}{2r}l = 1$ , or

$$r = \frac{1}{2} \frac{l}{1-l} \tag{A.18}$$

Therefore, when  $r = \frac{1}{2} \frac{l}{1-l}$ , we have a trivial (and unique) solution to equations (3.12) and (A.17) where

$$(\rho^{Jn^*R}, n^*) = (1, 2r) = (1, \frac{l}{1-l}).$$

**Non-trivial Solution:** Next, we search for any possible solution,  $(\rho^{Jn^*R}, n^*)$ , to equations (3.12) and (A.17) when  $\rho^{Jn^*R} \neq 1$ . Although  $l < 1$ ,  $\rho^{JnR}$  can be both smaller than or greater than 1 at the equilibrium (we only know  $\rho^{JnR} > l$ ). When  $\rho^{JnR} \neq 1$ , we have

$$\pi_0^{JnR} = \frac{1}{1 + \rho^{JnR} + (\rho^{JnR})^2 + \dots + (\rho^{JnR})^n} = \frac{1 - \rho^{JnR}}{1 - (\rho^{JnR})^{n+1}} \quad (\text{A.19})$$

$$\pi_n^{JnR} = (\rho^{JnR})^n \pi_0^{JnR} = \frac{(\rho^{JnR})^n - (\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} \quad (\text{A.20})$$

$$1 - \pi_n^{JnR} = 1 - \frac{(\rho^{JnR})^n - (\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} = \frac{1 - (\rho^{JnR})^n}{1 - (\rho^{JnR})^{n+1}} \quad (\text{A.21})$$

Therefore, equation (3.12), the stability condition becomes

$$\begin{aligned} \rho^{JnR}(1 - \pi_n^{JnR}) &= l \\ \Leftrightarrow \rho^{JnR} \left[ \frac{1 - (\rho^{JnR})^n}{1 - (\rho^{JnR})^{n+1}} \right] &= l \text{ from (A.21)} \\ \Leftrightarrow \frac{\rho^{JnR} - (\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} &= l \\ \Leftrightarrow \rho^{JnR} - (\rho^{JnR})^{n+1} &= l - l(\rho^{JnR})^{n+1} \\ \Leftrightarrow \rho^{JnR} - l &= (1 - l)(\rho^{JnR})^{n+1} \\ \Leftrightarrow (\rho^{JnR})^{n+1} &= \frac{\rho^{JnR} - l}{1 - l} \end{aligned} \quad (\text{A.22})$$

$$\Leftrightarrow n + 1 = \log_{\rho^{JnR}} \left( \frac{\rho^{JnR} - l}{1 - l} \right) = \frac{\ln \left( \frac{\rho^{JnR} - l}{1 - l} \right)}{\ln \rho^{JnR}} \quad (\text{A.23})$$

On the other hand, when  $\rho^{JnR} \neq 1$ , the average number of customers in the queue is given by

$$L^{JnR} = \sum_{k=0}^n \pi_k^{JnR} k = \frac{\rho^{JnR}}{1 - \rho^{JnR}} - \frac{(n+1)(\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}}. \quad (\text{A.24})$$

Equation (A.17), the indifference condition,  $n - L^{JnR} = r$  is thus equivalent to

$$\begin{aligned}
& n - \frac{\rho^{JnR}}{1 - \rho^{JnR}} + \frac{(n+1)(\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} = r \\
\Leftrightarrow & n + \frac{(n+1)(\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} = r + \frac{\rho^{JnR}}{1 - \rho^{JnR}} \\
\Leftrightarrow & n + 1 + \frac{(n+1)(\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}} = r + 1 + \frac{\rho^{JnR}}{1 - \rho^{JnR}} \\
\Leftrightarrow & (n+1)\left(1 + \frac{(\rho^{JnR})^{n+1}}{1 - (\rho^{JnR})^{n+1}}\right) = r + 1 + \frac{\rho^{JnR}}{1 - \rho^{JnR}} \\
\Leftrightarrow & (n+1)\frac{1}{1 - (\rho^{JnR})^{n+1}} = r + 1 + \frac{\rho^{JnR}}{1 - \rho^{JnR}} \\
\Leftrightarrow & (n+1)\frac{1-l}{1 - \rho^{JnR}} = r + 1 + \frac{\rho^{JnR}}{1 - \rho^{JnR}} \text{ from (A.22)} \\
\Leftrightarrow & (n+1)(1-l) = r(1 - \rho^{JnR}) + 1 \\
\Leftrightarrow & n + 1 = \frac{r(1 - \rho^{JnR}) + 1}{1-l} \tag{A.25}
\end{aligned}$$

We have just transferred the two equilibrium conditions in (3.12) and (A.17), into an equivalent set of equations

$$(A.23') : n = \frac{\ln(\frac{\rho^{JnR}-l}{1-l})}{\ln \rho^{JnR}} - 1; \text{ And } (A.25') : n = \frac{r(1 - \rho^{JnR}) + 1}{1-l} - 1.$$

in the sense that a pair of solutions  $(n^*, \rho^*)$  that solves equations (3.12) and (A.17) also solves equations (A.23') and (A.25'), and vice versa.

We observe from equations (A.23') and (A.25') that a necessary condition for  $n \geq 0$  is  $l < \rho \leq 1 + \frac{l}{r}$ . Define

$$\begin{aligned}
f_1(\rho) & \triangleq \frac{\ln(\frac{\rho-l}{1-l})}{\ln \rho} - 1 \text{ from (A.23')}; \\
f_2(\rho) & \triangleq \frac{r(1 - \rho) + 1}{1-l} - 1 \text{ from (A.25')}.
\end{aligned}$$

Then, any intersection of the graphs of  $f_1(\rho)$  and  $f_2(\rho)$  (other than at  $\rho = 1$ ) on the feasible region  $l < \rho \leq 1 + \frac{l}{r}$  will lead to a solution  $(\rho^* = \rho^{Jn^*R}, n^*) = (\rho^*, f_1(\rho^*) = f_2(\rho^*))$ .

It can be shown that  $f_1(\rho)$  is continuous, decreasing and convex in  $\rho$  on the region  $\rho \in (l, +\infty)$  with

$$\lim_{\rho \rightarrow l} f_1(\rho) = +\infty; \lim_{\rho \rightarrow 1} f_1(\rho) = \frac{l}{1-l}; \lim_{\rho \rightarrow +\infty} f_1(\rho) = 0. \quad (\text{A.26})$$

On the other hand,  $f_2(\rho)$  is simply a straight line in  $\rho$  with slope  $-\frac{r}{1-l}$  on the region  $\rho \in [0, 1 + \frac{l}{r}]$  with

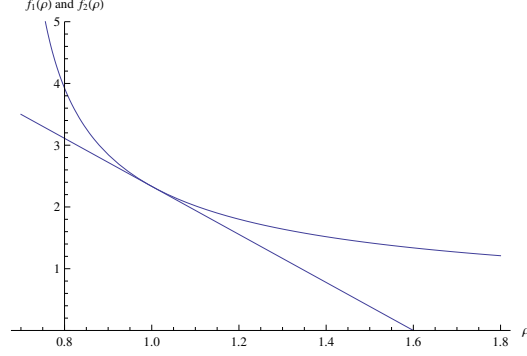
$$f_2(0) = \frac{r+l}{1-l}; f_2(1) = \frac{l}{1-l}; f_2(1 + \frac{l}{r}) = 0. \quad (\text{A.27})$$

Note from (A.26) and (A.27) that  $f_1(\rho)$  and  $f_2(\rho)$  intersect at  $\rho = 1$  with  $f_1(1) = f_2(1) = \frac{l}{1-l}$ . If  $f_2(\rho)$  is a tangent line to  $f_1(\rho)$  at  $\rho = 1$ , as shown in Figure 16/(a), then we will not have a solution  $(\rho^*, n^*)$  with  $\rho^* \neq 1$  (because in this case  $f_1(\rho)$  and  $f_2(\rho)$  will only intersect at  $\rho = 1$  on the feasible region  $l < \rho \leq 1 + \frac{l}{r}$ ). It can be verified that the slope of the curve  $f_1(\rho)$  at  $\rho = 1$  is  $-\frac{1}{2} \frac{l}{(1-l)^2}$  so  $f_2(\rho)$  coincides with the tangent line of  $f_1(\rho)$  at  $\rho = 1$  if and only if

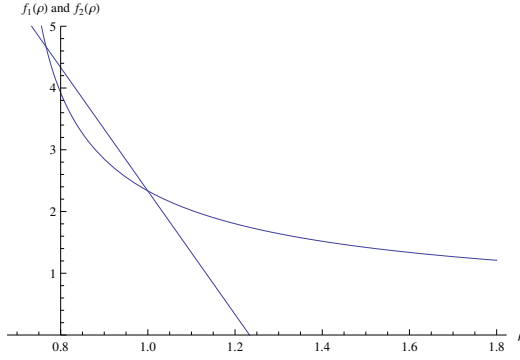
$$-\frac{r}{1-l} = -\frac{1}{2} \frac{l}{(1-l)^2} \Leftrightarrow r = \frac{1}{2} \frac{l}{(1-l)} \quad (\text{A.28})$$

note that condition (A.28) is exactly that same as condition (A.18), so in this case, although we do not have any solution  $(\rho^*, n^*)$  such that  $\rho^* \neq 1$ , we do have an unique solution found earlier during the discussion of the trivial solution, i.e.,  $(\rho^*, n^*) = (1, \frac{l}{1-l})$ .

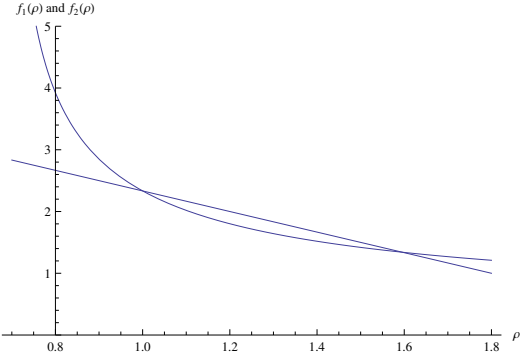
When  $r \neq \frac{1}{2} \frac{l}{(1-l)}$ , then  $f_2(\rho)$  is still a straight line but no longer a tangent one to the curve  $f_1(\rho)$  at  $\rho = 1$ . Since  $f_1(\rho)$  is a smooth decreasing convex curve,  $f_2(\rho)$  is a straight line, and they both have a common intersection at  $(1, \frac{l}{1-l})$ , they will intersect at some point  $\rho^*$  other than  $\rho^* = 1$  on the region  $l < \rho \leq 1 + \frac{l}{r}$  as shown in Figure 16/(b) and (c). The intersection point depends on the slope of  $f_2(\rho)$ , or the value of  $r$ , or simply the retrial cost  $\alpha$  (assuming parameters  $\lambda, \mu, v, c$  are system inputs and fixed). Thus, solution also exists



(a)  $r = \frac{7}{6}$ :  $f_2(\rho)$  is tangent to  $f_1(\rho)$ .



(b)  $r = 3$ : not tangent and cross above.



(c)  $r = \frac{1}{2}$ : not tangent and cross below.

Figure 16: In these three subfigures, we let  $l = 0.7$  and plot  $f_1(\rho)$  and  $f_2(\rho)$  w.r.t. different values of  $r$ . In every case,  $f_1(\rho)$  and  $f_2(\rho)$  cross the point  $(\rho = 1, f(\rho) = \frac{l}{1-l}) = (1, \frac{7}{3})$ . (a) When  $r = \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ ,  $f_2(\rho)$  is a tangent line to the curve  $f_1(\rho)$ . (b) When  $r = 3 > \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ , they cross above. (c) When  $r = \frac{1}{2} < \frac{7}{6} = \frac{1}{2} \frac{l}{1-l}$ , they cross below.

when  $r \neq \frac{1}{2} \frac{l}{1-l}$  and it is unique due to the single-crossing property of  $f_1(\rho)$  and  $f_2(\rho)$ . Note again that the intersection where  $\rho = 1$  cannot be counted as a second non-trivial solution.

So far we have shown that given the inputs of the system (i.e.,  $\lambda, \mu, v, c$ ), a particular value of the retrial fee  $\alpha$ , which is transformed into a particular value of the cost ratio  $r = \alpha / \frac{c}{\mu}$ , induces a unique pair of solutions, in the threshold  $n^*$  and its corresponding traffic  $\rho^* = \rho^{Jn^*R}$  among consumers, to equations (3.12) and (A.17). Furthermore, it is clear from the graph that when  $\alpha \uparrow$  (i.e.,  $f_2(\rho)$  becomes steeper),  $n^* \uparrow$  and  $\rho^* \downarrow$ . In fact, when  $\alpha \rightarrow 0$ ,  $\lim_{\alpha \rightarrow 0} n^* = 0$  and  $\lim_{\alpha \rightarrow 0} \rho^* = \infty$ . When  $\alpha \rightarrow \infty$ ,  $\lim_{\alpha \rightarrow \infty} n^* = \infty$  and  $\lim_{\alpha \rightarrow \infty} \rho^* = l$ .

At this moment we are ready to pick up the feasibility condition  $n^* \leq N$  and the integer condition on  $n$ . (Recall that  $N = \lfloor v/\frac{c}{\mu} \rfloor$  indicates the balking threshold in Naor's model, see (3.1).) Let us denote  $0 < \alpha_1 < \alpha_2 < \dots < \alpha_{N-1} < \alpha_N$  the retrial costs that induce the solutions to equations (3.12) and (A.17) with  $n = 1, 2, \dots, N-1, N$ , respectively. We next show existence and uniqueness of these  $\alpha$ 's.

Recall that  $\lim_{\alpha \rightarrow 0} n^* = 0$ ,  $\lim_{\alpha \rightarrow \infty} n^* = \infty$ , and  $n^*$  is continuous and increases in  $\alpha$ . For any fixed integer  $n \in (0, N]$ , compute the unique  $\rho_n$  such that  $f_1(\rho_n) = n$ , i.e.,  $(\rho_n, n)$  solves equation (A.23'). Next, let  $\alpha_n = \frac{(n+1)(1-l)-1}{1-\rho_n} \frac{c}{\mu}$ , i.e.,  $\alpha_n$  is such that  $(\rho_n, n)$  solves equation (A.25'). Since  $(\rho_n, n)$  solves both equations (A.23') and (A.25'),  $\alpha_n$  is the amount of the retrial cost that would induce  $n^* = n$  as a solution to equations (3.12) and (A.17). Uniqueness and monotonicity of  $\alpha_n$  were demonstrated in earlier discussion with the graph.  $\square$

**Proof for the statement that “ $\alpha_L < \alpha_H$ ”:**

Recall by definition that

$$\alpha_L = (1 - \pi_N^{JNR})(v - cW^{JNR}), \quad (\text{A.29})$$

$$\alpha_H = (1 - \pi_N^{JNB})(v - cW^{JNB}). \quad (\text{A.30})$$

where  $W^{JNR}$  and  $W^{JNB}$  are the expected waiting time conditional on joining, given by

$$W^{JNR} = \frac{\pi_0^{JNR}}{1 - \pi_N^{JNR}} \frac{1}{\mu} + \frac{\pi_1^{JNR}}{1 - \pi_N^{JNR}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1}^{JNR}}{1 - \pi_N^{JNR}} \frac{N}{\mu} \quad (\text{A.31})$$

$$W^{JNB} = \frac{\pi_0^{JNB}}{1 - \pi_N^{JNB}} \frac{1}{\mu} + \frac{\pi_1^{JNB}}{1 - \pi_N^{JNB}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1}^{JNB}}{1 - \pi_N^{JNB}} \frac{N}{\mu} \quad (\text{A.32})$$

The underlying queueing system under both  $JNR$  and  $JNB$  is  $M/M/1/N$ . The birth rate is bigger in the system under  $JNR$  than that under  $JNB$ , i.e.,  $\lambda^{JNR} = \frac{\lambda}{1 - \pi_N^{JNR}} > \lambda = \lambda^{JNB}$ . The death rates are the same in both systems, namely  $\mu$ . Therefore,  $\rho^{JNR} > \rho^{JNB}$ . As a result,

$$\pi_0^{JNR} = \frac{1}{1 + \rho^{JNR} + (\rho^{JNR})^2 + \dots + (\rho^{JNR})^N} < \frac{1}{1 + \rho^{JNB} + (\rho^{JNB})^2 + \dots + (\rho^{JNB})^N} = \pi_0^{JNB}.$$



That is, the server's long-run idle rate is higher with a system that has balking consumers compared to one that does not. Now suppose  $\pi_n^{JNR} > \pi_n^{JNB}$  for some  $n \in \{1, 2, \dots, N\}$ , then

$$\pi_x^{JNR} = \pi_n^{JNR}(\rho^{JNR})^{x-n} > \pi_n^{JNB}(\rho^{JNB})^{x-n} = \pi_x^{JNB}$$

for all  $x \in \{n, n+1, \dots, N\}$ . Since  $\sum_{x=1}^N \pi_x^{JNR} = \sum_{x=1}^N \pi_x^{JNB} = 1$ , we must have  $\pi_N^{JNR} > \pi_N^{JNB}$ .

Next, we observe that the conditional probabilities in equations (A.31) and (A.32) sum up to one and form geometric progressions with ratios  $\rho^{JNR}$  and  $\rho^{JNB}$ , respectively, i.e., for  $s \in \{JNR, JNB\}$ ,

$$\begin{aligned} \frac{\pi_0^s}{1 - \pi_N^s} + \frac{\pi_1^s}{1 - \pi_N^s} + \dots + \frac{\pi_{N-1}^s}{1 - \pi_N^s} &= 1 \\ \frac{\pi_{N-1}^s}{1 - \pi_N^s} &= \rho^s \frac{\pi_{N-2}^s}{1 - \pi_N^s} = (\rho^s)^2 \frac{\pi_{N-2}^s}{1 - \pi_N^s} = \dots = (\rho^s)^{(N-1)} \frac{\pi_0^s}{1 - \pi_N^s} \end{aligned}$$

It then follows from  $\rho^{JNR} > \rho^{JNB}$  and the structure of  $W$  that  $W^{JNR} > W^{JNB}$ . Since  $\pi_N^{JNR} > \pi_N^{JNB}$  and  $W^{JNR} > W^{JNB}$ , we have from (A.29) and (A.30) that  $\alpha_H > \alpha_L$ .  $\square$

**Proof of Lemma 5:**

Fix  $n \in \{1, 2, \dots, N-1\}$ . When the retrial cost  $\alpha = \alpha_n$ , since both stability and indifference conditions are satisfied for the solution  $(n^*, \rho^*) = (n, \rho^{JnR})$ , we have an equilibrium under  $JnR$ . Specifically, when  $\alpha = \alpha_1$ ,  $J1R$  is an equilibrium strategy which specifies that an arrival only joins the server if it is idle. By the construction of the game, every consumer would join an idle server no matter what the retrial cost  $\alpha$  is (because the retrial and balking payoffs are always less than the joining payoff at seeing an idle server). Therefore, for  $\alpha \leq \alpha_1$ ,  $J1R$  remains an equilibrium strategy. On the other hand, for  $n \in \{1, 2, \dots, N-1\}$ , when  $\alpha \in [\alpha_n - \frac{c}{\mu}(1 - \pi_n^{JnR}), \alpha_n]$ , the indifference condition of (A.15) still holds at  $n$ . Thus,  $JnR$  is also an equilibrium for such  $\alpha$ . Note that, from equation (3.12) that  $1 - \pi_n^{JnR} = \frac{l}{\rho^{JnR}}$ . Thus, when the retrial cost  $\alpha \in [\alpha_n - \frac{c}{\mu} \frac{l}{\rho^{JnR}}, \alpha_n]$ , we have an equilibrium under  $JnR$ . Finally, when  $\alpha = \alpha_L$ , we have a retrial payoff of zero under  $JNR$  according to the definition of  $\alpha_L$ , so such a strategy is an equilibrium. Furthermore, when  $\alpha \in [\alpha_L - (v - \frac{c}{\mu}N) \frac{l}{\rho^{JNR}}, \alpha_L]$ ,

the retrial payoff under  $JNR$  is greater than zero and less than  $v - \frac{c}{\mu}N$  thus the strategy remains an equilibrium.  $\square$

**Proof of Lemma 6:**

Fix  $1 \leq n \leq N$ . Choose any  $\beta \in (0, 1)$ . For  $J_{R(1-\beta)}^{J(\beta)}nR$  to be an equilibrium, upon arriving at state  $n - 1$ , a consumer is indifferent between join and retry decisions. Therefore, the joining and the retrial payoffs at state  $n - 1$  (under  $J_{R(1-\beta)}^{J(\beta)}nR$ ) are identical and both should be greater than 0 (the balking payoff), i.e.,

$$v - \frac{c}{\mu}(n) = v - cW^{J_{R(1-\beta)}^{J(\beta)}nR} - \frac{\alpha}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}} \quad (\text{A.33})$$

If and only if the retrial cost  $\alpha$  satisfies equation (A.33), the policy  $J_{R(1-\beta)}^{J(\beta)}nR$  becomes an equilibrium. Since  $W^{J_{R(1-\beta)}^{J(\beta)}nR}$  and  $\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}$  are just some fixed quantities given that the population adopts the join/retry strategy  $J_{R(1-\beta)}^{J(\beta)}nR$ , it is clear from equation (A.33) that  $\alpha$  exists and is unique, where

$$\alpha = (1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}) \left[ \frac{c}{\mu}n - cW^{J_{R(1-\beta)}^{J(\beta)}nR} \right] \quad (\text{A.34})$$

is a function of  $\beta$ . Finally, when  $\beta$  increases from 0 to 1, the underlying queueing system, where everyone adopts the strategy  $J_{R(1-\beta)}^{J(\beta)}nR$ , evolves continuously, and thus the conditional waiting time  $W^{J_{R(1-\beta)}^{J(\beta)}nR}$  and the steady-state retrial probability  $\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}$  are both continuous quantities in  $\beta$ . It follows that  $\alpha(\beta)$ , given in (A.34), is also continuous in  $\beta$ .  $\square$

**To prove Lemma 7, we first prove two supporting lemmas (Lemmas 12 and 13).**

**Lemma 12** For fixed  $n : 1 \leq n \leq N$ , vary  $\alpha$  such that  $J_{R(1-\beta)}^{J(\beta)}nR$  is an equilibrium policy for  $\beta \in (0, 1)$ . Then, the partial derivative of consumer welfare under the equilibrium join/retry strategy  $J_{R(1-\beta)}^{J(\beta)}nR$  with respect to  $\beta$  is negative. Mathematically, if we denote  $U$  as consumer welfare, then  $\frac{\partial U^{J_{R(1-\beta)}^{J(\beta)}nR}}{\partial \beta} < 0$ .

**Proof of Lemma 12:**

Suppose  $J_{R(1-\beta)}^{J(\beta)} nR$  is an equilibrium strategy. We still use

$$\pi_0^{J_{R(1-\beta)}^{J(\beta)} nR}, \pi_1^{J_{R(1-\beta)}^{J(\beta)} nR}, \dots, \pi_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR}, \pi_n^{J_{R(1-\beta)}^{J(\beta)} nR}$$

as the steady-state probabilities for states  $\{0, 1, \dots, n-1, n\}$ , respectively. Let

$$U_0^{J_{R(1-\beta)}^{J(\beta)} nR}, U_1^{J_{R(1-\beta)}^{J(\beta)} nR}, \dots, U_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR}, U_n^{J_{R(1-\beta)}^{J(\beta)} nR}$$

denote a consumer's (expected) payoff arriving to state  $x \in \{0, 1, \dots, n-1, n\}$  (when the population adopts  $J_{R(1-\beta)}^{J(\beta)} nR$ ). Then,

$$\begin{aligned} U_0^{J_{R(1-\beta)}^{J(\beta)} nR} &= v - \frac{c}{\mu}(1) \text{ for joining;} \\ U_1^{J_{R(1-\beta)}^{J(\beta)} nR} &= v - \frac{c}{\mu}(2) \text{ for joining;} \\ &\vdots = \quad \vdots \\ U_{n-2}^{J_{R(1-\beta)}^{J(\beta)} nR} &= v - \frac{c}{\mu}(n-1) \text{ for joining;} \\ U_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR} &= v - \frac{c}{\mu}(n) \text{ no matter choosing to join or retry;} \\ U_n^{J_{R(1-\beta)}^{J(\beta)} nR} &= v - \frac{c}{\mu}(n) \text{ for retrying.} \end{aligned}$$

On the other hand, the consumer welfare (rate) is given by

$$U^{J_{R(1-\beta)}^{J(\beta)} nR} = \lambda \sum_{x=0}^n \pi_x^{J_{R(1-\beta)}^{J(\beta)} nR} U_x^{J_{R(1-\beta)}^{J(\beta)} nR}. \quad (\text{A.35})$$

The retrial probability under  $J_{R(1-\beta)}^{J(\beta)} nR$  is

$$\pi_R^{J_{R(1-\beta)}^{J(\beta)} nR} = (1 - \beta)\pi_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR} + \pi_n^{J_{R(1-\beta)}^{J(\beta)} nR}.$$

According to (3.12), we have  $\rho^{J_{R(1-\beta)}^{J(\beta)} nR} = \frac{l}{1 - \pi_R^{J(\beta)}}$  and

$$\begin{aligned} \pi_1^{J_{R(1-\beta)}^{J(\beta)} nR} &= \rho^{J_{R(1-\beta)}^{J(\beta)} nR} \pi_0^{J_{R(1-\beta)}^{J(\beta)} nR}; \\ \pi_2^{J_{R(1-\beta)}^{J(\beta)} nR} &= \rho^{J_{R(1-\beta)}^{J(\beta)} nR} \pi_1^{J_{R(1-\beta)}^{J(\beta)} nR}; \\ &\vdots \\ \pi_{n-2}^{J_{R(1-\beta)}^{J(\beta)} nR} &= \rho^{J_{R(1-\beta)}^{J(\beta)} nR} \pi_{n-3}^{J_{R(1-\beta)}^{J(\beta)} nR}; \\ \pi_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR} &= \rho^{J_{R(1-\beta)}^{J(\beta)} nR} \pi_{n-2}^{J_{R(1-\beta)}^{J(\beta)} nR}; \\ \pi_n^{J_{R(1-\beta)}^{J(\beta)} nR} &= \beta \rho^{J_{R(1-\beta)}^{J(\beta)} nR} \pi_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR}. \end{aligned}$$

It follows that for fixed  $n : 1 \leq n \leq N$ ,  $\rho^{J_{R(1-\beta)}^{J(\beta)} nR}$  decreases in  $\beta : 0 \rightarrow 1$  due to (i)

$$\begin{aligned} &\pi_0^{J_{R(1-\beta)}^{J(\beta)} nR} + \pi_1^{J_{R(1-\beta)}^{J(\beta)} nR} + \dots + \pi_n^{J_{R(1-\beta)}^{J(\beta)} nR} \\ &= \pi_0^{J_{R(1-\beta)}^{J(\beta)} nR} [\rho^{J_{R(1-\beta)}^{J(\beta)} nR} + (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^2 + \dots + (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^{n-1} + \beta (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^n] = 1, \end{aligned}$$

and (ii)  $\pi_0^{J_{R(1-\beta)}^{J(\beta)} nR} \equiv 1 - l$  no matter what  $\beta$  is (and in fact what  $n$  is), because in a non-balking system (i.e.,  $\pi_B^{J_{R(1-\beta)}^{J(\beta)} nR} = 0$ ), the server's long-run idle probability is always equal to one minus the utility rate. As a result, the consumer welfare in (A.35) decreases in  $\beta$  because the payoffs  $U_x^{J_{R(1-\beta)}^{J(\beta)} nR}$ 's are decreasing in state  $x$ , i.e.,

$$U_0^{J_{R(1-\beta)}^{J(\beta)} nR} > U_1^{J_{R(1-\beta)}^{J(\beta)} nR} > \dots > U_{n-2}^{J_{R(1-\beta)}^{J(\beta)} nR} > U_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR} = U_n^{J_{R(1-\beta)}^{J(\beta)} nR}.$$

As a side result, we also notice that the steady-state retrial probability

$$\pi_R^{J_{R(1-\beta)}^{J(\beta)} nR} = (1 - \beta) \pi_{n-1}^{J_{R(1-\beta)}^{J(\beta)} nR} + \pi_n^{J_{R(1-\beta)}^{J(\beta)} nR} = [(1 - \beta) (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^{n-1} + \beta (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^n] \pi_0^{J_{R(1-\beta)}^{J(\beta)} nR}$$

decreases in  $\beta$ . This is because for fixed  $\rho^{J_{R(1-\beta)}^{J(\beta)} nR}$ ,  $(1 - \beta) (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^{n-1} + \beta (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^n$  decreases in  $\beta$ . And now since  $\rho^{J_{R(1-\beta)}^{J(\beta)} nR}$  decreases in  $\beta$ , so  $(1 - \beta) (\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^{n-1} +$

$\beta(\rho^{J_{R(1-\beta)}^{J(\beta)} nR})^n$  further decreases as  $\beta$  increases. This can be formally proved by taking derivatives to show  $\frac{\partial \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}}{\partial \beta} < 0$ .  $\square$

**Lemma 13** Fix  $n : 1 \leq n \leq N$ . The consumer welfare under the equilibrium join/retry strategy  $J_{R(1-\beta)}^{J(\beta)} nR$ , i.e.,  $U^{J_{R(1-\beta)}^{J(\beta)} nR}$  in Lemma 12, is also equal to

$$\lambda v - \lambda cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \lambda \left( \frac{1}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}} - 1 \right) \alpha \quad (\text{A.36})$$

where  $\alpha = \alpha(\beta)$  is the quantity that induces the equilibrium join/retry strategy  $J_{R(1-\beta)}^{J(\beta)} nR$ .

**Proof of Lemma 13:**

We show below that (A.36) equals  $U^{J_{R(1-\beta)}^{J(\beta)} nR}$  via equation (A.35). Since  $\alpha$  induces the equilibrium policy  $J_{R(1-\beta)}^{J(\beta)} nR$ , we have from (A.33) that  $v - \frac{c}{\mu}(n) = v - cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \frac{\alpha}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}}$ . Using  $\sigma = J_{R(1-\beta)}^{J(\beta)} nR$  in the rest of the proof, it then follows that

$$\begin{aligned} & U^\sigma \\ &= \lambda \sum_{x=0}^n \pi_x^\sigma U_x^\sigma \\ &= \lambda \left\{ \pi_0^\sigma \left[ v - \frac{c}{\mu}(1) \right] + \dots + \pi_{n-2}^\sigma \left[ v - \frac{c}{\mu}(n-1) \right] + \beta \pi_{n-1}^\sigma \left[ v - \frac{c}{\mu}(n) \right] + \pi_R^\sigma \left[ v - \frac{c}{\mu}(n) \right] \right\} \\ &= \lambda \left\{ \pi_0^\sigma \left[ v - \frac{c}{\mu}(1) \right] + \dots + \pi_{n-2}^\sigma \left[ v - \frac{c}{\mu}(n-1) \right] + \beta \pi_{n-1}^\sigma \left[ v - \frac{c}{\mu}(n) \right] + \pi_R^\sigma \left[ v - cW - \frac{\alpha}{1 - \pi_R^\sigma} \right] \right\} \\ &= \lambda \left\{ (\pi_0^\sigma + \dots + \pi_{n-1}^\sigma + \beta \pi_{n-1}^\sigma + \pi_R^\sigma) v \right. \\ &\quad \left. - \left[ \pi_0^\sigma \frac{c}{\mu}(1) + \dots + \pi_{n-2}^\sigma \frac{c}{\mu}(n-1) + \beta \pi_{n-1}^\sigma \frac{c}{\mu}(n) + \pi_R^\sigma cW^\sigma \right] - \frac{\pi_R^\sigma}{1 - \pi_R^\sigma} \alpha \right\} \\ &= \lambda v - \lambda \left[ (1 - \pi_R^\sigma) cW^\sigma + \pi_R cW^\sigma \right] - \lambda \frac{\pi_R}{1 - \pi_R} \alpha \\ &\quad \left( \text{because } W^\sigma = \frac{\pi_0^\sigma}{1 - \pi_R^\sigma} \frac{1}{\mu} + \frac{\pi_1^\sigma}{1 - \pi_R^\sigma} \frac{2}{\mu} + \dots + \frac{\pi_{n-2}^\sigma}{1 - \pi_R^\sigma} \frac{n-1}{\mu} + \frac{\beta \pi_{n-1}^\sigma}{1 - \pi_R^\sigma} \frac{n}{\mu} \right) \\ &= \lambda v - \lambda cW^\sigma - \lambda \frac{\pi_R}{1 - \pi_R} \alpha \end{aligned}$$

which is equal to the expression in (A.36).  $\square$

**Proof of Lemma 7:**

As a corollary to Lemmas 12 and 13 proved above, at an equilibrium strategy  $J_{R(1-\beta)}^{J(\beta)} nR$ , we must have

$$\frac{\partial[v - cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \frac{\alpha}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}}]}{\partial\beta} + \frac{\partial\alpha}{\partial\beta} < 0. \quad (\text{A.37})$$

This is because the partial derivative of  $\lambda v - \lambda cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \lambda \frac{1}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}} \alpha + \lambda \alpha$  in (A.36) with respect to  $\beta$  is negative and  $\lambda$  is only a constant.

Fix  $n \in \{2, \dots, N-1, N\}$ . Let  $\alpha(\beta)$  be the retrial cost that induces the equilibrium strategy  $J_{R(1-\beta)}^{J(\beta)} nR$ . Under any equilibrium strategy  $J_{R(1-\beta)}^{J(\beta)} nR$ , the retrial payoff must be the same as the joining payoff at state  $n-1$ , i.e.,

$$v - \frac{c}{\mu}(n) = v - cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \frac{\alpha}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}}. \quad (\text{A.38})$$

It follows from (A.38) that

$$\frac{\partial[v - cW^{J_{R(1-\beta)}^{J(\beta)} nR} - \frac{\alpha}{1 - \pi_R^{J_{R(1-\beta)}^{J(\beta)} nR}}]}{\partial\beta} = 0. \quad (\text{A.39})$$

Comparing (A.39) to (A.37) tells that  $\frac{\partial\alpha}{\partial\beta} < 0$ , i.e., when  $\beta$  increases from 0 to 1, the unique retrial cost that induces the equilibrium policy  $J_{R(1-\beta)}^{J(\beta)} nR$  decreases continuously in  $\beta$ .

On the other hand, we have showed in Lemma 5 that  $\alpha_{n-1}$  induces  $J(n-1)R$  or simply  $J_{R(1)}^{J(0)} nR$ , and  $\alpha_n - \frac{cl}{\mu\rho^{J_n R}}$  induces  $JnR$  or simply  $J_{R(0)}^{J(1)} nR$  with binding indifference conditions, i.e.,

$$v - \frac{c}{\mu}(n) = v - cW^{J_{R(1)}^{J(0)} nR} - \frac{\alpha_{n-1}}{1 - \pi_{n-1}^{J_{R(1)}^{J(0)} nR}}$$

$$v - \frac{c}{\mu}(n) = v - cW^{J_{R(0)}^{J(1)} nR} - \frac{\alpha_n - \frac{cl}{\mu\rho^{J_n R}}}{1 - \pi_n^{J_{R(0)}^{J(1)} nR}}$$

Therefore, when  $\beta$  increase from 0 to 1, we have an equilibrium strategy  $J_{R(1-\beta)}^{J(\beta)}nR$  for some unique retrial cost  $\alpha$  that is decreasing from  $\alpha_{n-1}$  to  $\alpha_n - \frac{cl}{\mu\rho^{JnR}}$ . It follows that  $\alpha_n - \frac{cl}{\mu\rho^{JnR}} < \alpha_{n-1}$ .  $\square$

***Proof of Theorem 4:***

From Lemmas 5 and 7, we see that for any  $\alpha \leq \alpha_L$ , there exists at least one equilibrium strategy in the forms of a retry or a join/retry strategy. The lemmas tell us that on the region of  $I_n$ , the equilibrium strategies certainly include retry strategy  $JnR$  and join/retry strategies  $J_{R(1-\beta)}^{J(\beta)}nR$ . If there are any other equilibrium strategies of a retry or a join/retry strategy on this region, say  $JmR$  or  $J_{R(1-\beta)}^{J(\beta)}mR$ , then it must be the case that  $m > n$ .

Therefore, to show the retry strategy  $JnR$  is the Pareto-dominant equilibrium for all  $\alpha \in I_n$ , it is equivalent to show that  $JnR$  generates the highest welfare among the family of strategies  $\{J_{R(1-\beta)}^{J(\beta)}mR : m > n, 0 \leq \beta \leq 1\}$ . With Lemma 12 in mind, it suffices to show that  $U_{R(1)}^{J_{R(1)}^{J(0)}(n+1)R} > U_{R(1)}^{J_{R(1)}^{J(0)}(n+2)R} > U_{R(1)}^{J_{R(1)}^{J(0)}(n+3)R}, \dots$ , or equivalently  $U^{JnR} > U^{J(n+1)R} > U^{J(n+2)R}, \dots$ . We will show next that for all  $k \in \{n, n+1, \dots\}$ ,  $U^{JkR} > U^{J(k+1)R}$ . Recall

$$U^{JkR} = \lambda \sum_{x=0}^k \pi_x^{JkR} U_x^{JkR}$$

$$U^{J(k+1)R} = \lambda \sum_{x=0}^{k+1} \pi_x^{J(k+1)R} U_x^{J(k+1)R}$$

$$\begin{array}{ccccccc} U_0^{JkR} & > & U_1^{JkR} & > & U_2^{JkR} & > & \dots > U_{k-1}^{JkR} & \geq & U_k^{JkR} \\ \parallel & & \parallel & & \parallel & & & \parallel & \vee \\ U_0^{J(k+1)R} & > & U_1^{J(k+1)R} & > & U_2^{J(k+1)R} & > & \dots > U_{k-1}^{J(k+1)R} & > & U_k^{J(k+1)R} \geq U_{k+1}^{J(k+1)R}. \end{array}$$

Since  $\pi_0^{JkR} = \pi_0^{J(k+1)R} = 1 - l$  and  $\rho^{JkR} > \rho^{J(k+1)R}$ , it follows that  $U^{JkR} > U^{J(k+1)R}$ .  $\square$

***Proof of Corollary 1:***

Results follow from the properties of functions (A.26) and (A.27).  $\square$

**Proof of Lemma 8:**

We note that the underlying system is  $M/M/1/N$  when the retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  is adopted by the population. With retrial cost  $\alpha$ , the expected retrial payoff is given by

$$\begin{aligned}
& \pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}} - \alpha) - \pi_B \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} \alpha + \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} \pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}} - 2\alpha) \\
& - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} \pi_B \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} 2\alpha + (\pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}})^2 \pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}} - 3\alpha) \\
& - (\pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}})^2 \pi_B \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} 3\alpha + \dots \\
& = \frac{\pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} - \frac{\pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} \alpha}{(1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}})^2} - \frac{\pi_B \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} \alpha}{(1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}})^2} \\
& = \frac{\pi_J \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} - \frac{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{(1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}})^2} \alpha \\
& = \frac{1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\alpha}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} \tag{A.40}
\end{aligned}$$

where

$$W^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\pi_0 \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{1}{\mu} + \frac{\pi_1 \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{2}{\mu} + \dots + \frac{\pi_{N-1} \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} \frac{N}{\mu}.$$

The strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  will be an equilibrium of the system if and only the retrial payoff under  $JN_{B(\gamma)}^{R(1-\gamma)}$  is exactly zero, i.e., expression (A.40) is zero. Therefore, the unique retrial cost that induces  $JN_{B(\gamma)}^{R(1-\gamma)}$  as an equilibrium is a solution to

$$\frac{1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\alpha}{1 - \pi_R \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}} = 0,$$

or simply  $\alpha = (1 - \pi_N \frac{JN_{B(\gamma)}^{R(1-\gamma)}}{JN_{B(\gamma)}^{R(1-\gamma)}}) (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})$  which is given by (3.17).  $\square$



**Proof of Theorem 5:**

Suppose the consumer population adopts the retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  and  $\gamma$  increases from 0 to 1. The underlying system remains  $M/M/1/N$ . By a similar proof to “ $\alpha_L < \alpha_H$ ”, it is easy to see that  $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$  and  $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$  all continuously decrease in  $\gamma$ . Therefore, by (3.17), the retrial cost  $\alpha(\gamma)$  that induces the equilibrium strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  increases continuously in  $\gamma$ . Furthermore, when  $\gamma = 0$ , the retry/balk strategy  $JN_{B(0)}^{R(1)}$  is equivalent to the retrial strategy  $JNR$  and equation (3.17) is simply equation (A.29), so  $\alpha(0) = \alpha_L$ . Similarly, when  $\gamma = 1$ , the retry/balk strategy  $JN_{B(1)}^{R(0)}$  is equivalent to the balk strategy  $JNB$  and equation (3.17) becomes equation (A.30), and so  $\alpha(1) = \alpha_H$ .

We have argued that for each  $\gamma \in (0, 1)$ , there exists a unique retrial cost  $\alpha \in (\alpha_L, \alpha_H)$  such that  $JN_{B(\gamma)}^{R(1-\gamma)}$  is an equilibrium strategy. Due to the continuity and strict monotonicity of  $\alpha(\gamma)$  in  $\gamma$ , we can conclude that for each retrial cost  $\alpha \in (\alpha_L, \alpha_H)$ , there exists a unique equilibrium retry/balk strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  where  $\gamma$  increases in  $\alpha$  (from 0 to 1). Note that there does not exist other types of equilibrium strategy when  $\alpha \in (\alpha_L, \alpha_H)$  because equilibrium retry and join/retry strategies appear for  $\alpha \leq \alpha_L$  and equilibrium balk strategy appears for  $\alpha \geq \alpha_H$ . Therefore, the equilibrium retry/balk strategy is unique.  $\square$

**Proof of (3.19):**

$$\begin{aligned}
\lambda W^{JnR} &= \lambda \left[ \frac{\pi_0^{JnR}}{1 - \pi_n^{JnR}} \frac{1}{\mu} + \frac{\pi_1^{JnR}}{1 - \pi_n^{JnR}} \frac{2}{\mu} + \dots + \frac{\pi_{n-1}^{JnR}}{1 - \pi_n^{JnR}} \frac{n}{\mu} \right] \\
&= \lambda^{JnR} (1 - \pi_n^{JnR}) \left[ \frac{\pi_0^{JnR}}{1 - \pi_n^{JnR}} \frac{1}{\mu} + \frac{\pi_1^{JnR}}{1 - \pi_n^{JnR}} \frac{2}{\mu} + \dots + \frac{\pi_{n-1}^{JnR}}{1 - \pi_n^{JnR}} \frac{n}{\mu} \right] \\
&= \lambda^{JnR} \left[ \pi_0^{JnR} \frac{1}{\mu} + \pi_1^{JnR} \frac{2}{\mu} + \dots + \pi_{n-1}^{JnR} \frac{n}{\mu} \right] \\
&= \pi_0^{JnR} \rho^{JnR} \cdot 1 + \pi_1^{JnR} \rho^{JnR} \cdot 2 + \dots + \pi_{n-1}^{JnR} \rho^{JnR} \cdot n \\
&= \pi_1^{JnR} \cdot 1 + \pi_2^{JnR} \cdot 2 + \dots + \pi_n^{JnR} \cdot n \\
&= \pi_0^{JnR} \cdot 0 + \pi_1^{JnR} \cdot 1 + \pi_2^{JnR} \cdot 2 + \dots + \pi_n^{JnR} \cdot n \\
&= L^{JnR}
\end{aligned}$$

**Proof of Theorem 6:**

By the proof of Theorem 4, we know that when  $\alpha \leq \alpha_L$ , the consumer welfare (of the Pareto-dominant retry strategy  $JnR$ ) decreases in the retrial cost  $\alpha$ . The welfare is linearly decreasing on each region of  $(0, \alpha_1], (\alpha_1, \alpha_2], \dots, (\alpha_{N-1}, \alpha_L]$  with flatter and flatter slope because  $\frac{1}{1-\pi_n^{JnR}}$  decreases in  $n$ . Therefore, the highest welfare is achieved when  $\alpha \rightarrow 0$  and the lowest when  $\alpha \rightarrow \alpha_L$ . Recall from (A.24) that the average number of consumers in the system is

$$L = \frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}} = \frac{\rho}{1-\rho} - [\log_\rho(\frac{\rho-l}{1-l})] \frac{\rho-l}{1-\rho}. \quad (\text{A.41})$$

It can be verified that (A.41) decreases in  $\rho$  on  $\rho > l$ . In fact,  $\lim_{\rho \rightarrow l} L = \frac{l}{1-l}$  and  $\lim_{\rho \rightarrow \frac{l}{1-l}} L = l$ .

When  $\alpha$  goes from 0 to  $\alpha_L$ , the threshold of the equilibrium retry strategy  $JnR$  increases from 1 to  $N$ , and  $\rho^{JnR}$  decreases from  $\rho^{J1R} = \frac{l}{1-l}$  to some  $\rho^{JNR} > l$ . (When  $J1R$  is an equilibrium strategy,  $\rho^{J1R} = \frac{l}{1-\pi_1^{J1R}} = \frac{l}{1-l}$ .) Therefore, when  $\alpha \rightarrow 0$ , the equilibrium strategy is  $J1R$  with  $L^{J1R} = \pi_1^{J1R} = l$  and  $\pi_0^{J1R} = 1-l$ . It then follows from (3.20) that the consumer welfare equals  $\lambda v - cl$ . On the other hand, when  $\alpha = \alpha_L$ , the consumer welfare is  $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L$ .

Moreover, we can bound  $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L$  in that

$$\lambda v - cl(L^{JNR} + 1) < \lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L < \lambda v - cL^{JNR}. \quad (\text{A.42})$$

It is clear that  $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L < \lambda v - cL^{JNR}$ . To show the other half of the inequality in (A.42), we note that since  $\alpha_N > \alpha_L$ , we have

$$\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_N < \lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L, \quad (\text{A.43})$$

and we shall show below that the LHS of expression (A.43) is equivalent to  $\lambda v - cl(L^{JNR} + 1)$ .

Recall from (A.25) that at the equilibrium,  $n + 1 = \frac{r(1-\rho^{JnR})+1}{1-l}$ , thus we have

$$\begin{aligned}
N + 1 &= \frac{(\alpha_N/(c/\mu))(1 - \rho^{JNR}) + 1}{1 - l} \\
\Leftrightarrow c(N + 1) &= \frac{\alpha_N\mu(1 - \rho^{JNR}) + c}{1 - l} \\
\Leftrightarrow c(N + 1)(1 - l) &= \alpha_N\mu(1 - \rho^{JNR}) + c \\
\Leftrightarrow cN + c - cNl - cl &= \alpha_N\mu - \alpha_N\mu\rho^{JNR} + c \\
\Leftrightarrow cN - cNl - cl &= \alpha_N\mu - \alpha_N\mu\rho^{JNR} \\
\Leftrightarrow cN + \alpha_N\mu\rho^{JNR} - \alpha_N\mu &= cNl + cl \\
\Leftrightarrow cN + \alpha_N\mu\rho^{JNR} - \alpha_N\mu &= cl(N + 1)
\end{aligned} \tag{A.44}$$

The LHS of expression (A.43) can be rewritten as

$$\begin{aligned}
&\lambda v - \lambda\left(\frac{1}{1 - \pi_N^{JNR}} - 1\right)\alpha_N - cL^{JNR} \\
&= \lambda v - [\lambda\left(\frac{1}{1 - \pi_N^{JNR}} - 1\right)\alpha_N + cL^{JNR}] \\
&= \lambda v - [\lambda\left(\frac{1}{1 - \pi_N^{JNR}} - 1\right)\alpha_N + c(N - \alpha_N/(c/\mu))] \text{ from (A.17)} \\
&= \lambda v - [\mu(\rho - l)\alpha_N + c(N - \alpha_N/(c/\mu))] \\
&= \lambda v - [\mu\rho\alpha_N - \mu l\alpha_N + cN - \alpha_N\mu] \\
&= \lambda v - [cl(N + 1) - \mu l\alpha_N] \text{ from (A.44)} \\
&= \lambda v - [cl(N + 1 - r)] \\
&= \lambda v - cl(L^{JNR} + 1) \text{ from (A.17)}
\end{aligned} \tag{A.45}$$

From (A.43) and (A.45), we see that  $\lambda v - cl(L^{JNR} + 1) < \lambda v - cL^{JNR} - \lambda\left(\frac{1}{1 - \pi_N^{JNR}} - 1\right)\alpha_L < \lambda v - cL^{JNR}$ . It is a quick check that  $\lambda v - cl(L^{JNR} + 1) < \lambda v - cL^{JNR}$  because we know  $L^{JNR} < \frac{l}{1-l}$ .  $\square$

**Proof of Theorem 7:**

By (3.24), the consumer welfare equals

$$\begin{aligned} & \frac{\lambda(1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \alpha \\ &= \frac{\lambda}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} (1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}) (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \alpha \end{aligned}$$

which by (3.17) is equal to

$$\frac{\lambda}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \alpha - \frac{\lambda(1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - (1 - \gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \alpha = \lambda\alpha. \quad (\text{A.46})$$

As (A.46) clearly increases in  $\alpha$  and by considering the two extreme scenarios, our claim is proved.

Recall from (A.29) and (A.30) that  $\alpha_L = (v - cW^{JNR})(1 - \pi_N^{JNR})$  and  $\alpha_H = (v - cW^{JNB})(1 - \pi_N^{JNB})$ , it is then a trivial verification that

$$\begin{aligned} \lambda \cdot \alpha_L &= \lambda v - \lambda cW^{JNR} - \lambda \left( \frac{1}{1 - \pi_N^{JNR}} - 1 \right) \alpha_L = \lambda v - cL^{JNR} - \lambda \left( \frac{1}{1 - \pi_N^{JNR}} - 1 \right) \alpha_L, \\ \lambda \cdot \alpha_H &= \lambda v (1 - \pi_N^{JNB}) - \lambda (1 - \pi_N^{JNB}) cW^{JNB} = \lambda v (1 - \pi_N^{JNB}) - cL^{JNB}. \end{aligned}$$

Here's an alternative proof. Since  $\alpha(\beta)$  always adjust to make the retrial payoff of the strategy  $JN_{B(\gamma)}^{R(1-\gamma)}$  zero, the welfare can also be given as

$$\begin{aligned} & U^{JN_{B(\gamma)}^{R(1-\gamma)}} \\ &= \lambda \left[ \sum_{x=0}^N \pi_x^{JN_{B(\gamma)}^{R(1-\gamma)}} U_x^{JN_{B(\gamma)}^{R(1-\gamma)}} \right] \\ &= \lambda \left[ \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} \left( v - \frac{c}{\mu} \right) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left( v - \frac{Nc}{\mu} \right) \right] \end{aligned}$$

$$\begin{aligned}
& + \gamma \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \text{balking payoff} + (1-\gamma) \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \text{retrial payoff}] \\
& = \lambda [\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{2c}{\mu}) + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{Nc}{\mu}) \\
& \quad + \gamma \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot 0 + (1-\gamma) \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot 0] \\
& = \lambda [\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{2c}{\mu}) + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} (v - \frac{Nc}{\mu})] \\
& = \lambda [(1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})v - (1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})cW^{JN_{B(\gamma)}^{R(1-\gamma)}}] \\
& = \lambda \alpha
\end{aligned}$$

**Proof of Proposition 8:**

There was an argument in the paper. Here, we provide a more formal proof to show that  $\lambda v - cl$  is greater than Naor's consumer welfare given in (3.25), i.e.,  $\lambda v - cl > \lambda v(1 - \pi_N^{JNB}) - cL^{JNB}$ , or equivalently,  $\pi_N^{JNB} \lambda v > cl - cL^{JNB}$ . Recall that  $v > N \frac{c}{\mu}$ , therefore it is sufficient to show that

$$\begin{aligned}
& \pi_N^{JNB} \lambda N \frac{c}{\mu} > cl - cL^{JNB} \\
& \Leftrightarrow \pi_N^{JNB} lN > l - L^{JNB} \\
& \Leftrightarrow \frac{(1-l)l^N}{1-l^{N+1}} lN > l - \frac{l}{1-l} + (N+1) \frac{l^{N+1}}{1-l^{N+1}} \\
& \Leftrightarrow \frac{l}{1-l} - l > [(N+1) - (1-l)N] \frac{l^{N+1}}{1-l^{N+1}} \\
& \Leftrightarrow \frac{l^2}{1-l} > (1+lN) \frac{l^{N+1}}{1-l^{N+1}} \tag{A.47}
\end{aligned}$$

Since the RHS of expression (A.47) is decreasing in  $N$  and equals to  $\frac{l^2}{1-l}$  when  $N = 1$  (the smallest possible value), we know (A.47) holds and therefore the claim is proved.  $\square$

**Proof of Theorem 9:**

We take the graph on the results of the equilibrium welfare in Figure 4, and extend all the  $N$  line segments on  $\alpha \in (0, \alpha_L]$  into  $N$  lines, and denote them as  $L_1, L_2, \dots, L_N$ , respectively, see Figure 17. The original  $N$  line segments correspond to the welfare under the retry

strategy  $J1R$  for  $\alpha \in (0, \alpha_1]$ , welfare under the retry strategy  $J2R$  for  $\alpha \in (\alpha_1, \alpha_2]$ , ..., and welfare under the retry strategy  $JNR$  for  $\alpha \in (\alpha_{N-1}, \alpha_L]$ , etc. With the extension, the full line  $L_n$  for  $n = 1, 2, \dots, N$  would represent the welfare under retry strategies  $JnR$  for all  $\alpha \in (0, \infty)$ .

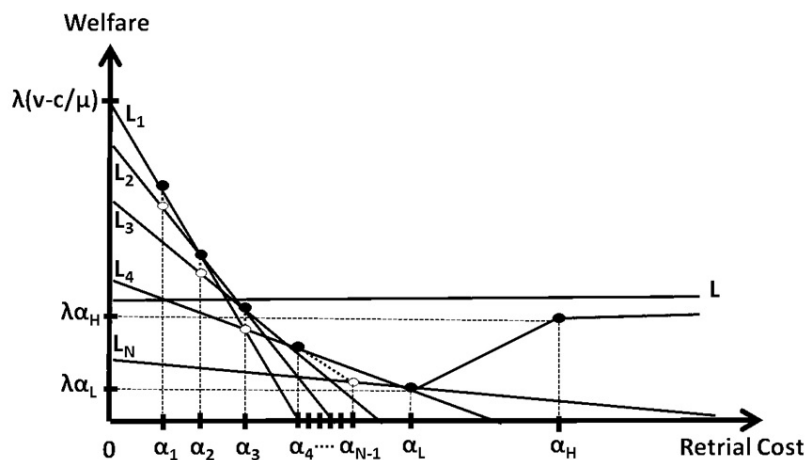


Figure 17: Illustration of the welfare under socially optimal policies (the upper envelope of  $L_1, L_2, \dots, L_N$  and  $L$ ).

On the other hand, if we denote  $L$  the welfare under  $JN'B$ , i.e., the socially optimal policy in Naor (1969), we know it is a straight line describing a constant function of the retrial cost  $\alpha$ . Naor (1969) showed that  $N' < N$  and the consumer welfare under  $JN'B$  exceeds that under  $JNB$  which is equal to  $\lambda \alpha_H$ . In fact, welfare curve under any balk strategy (i.e.,  $JnB$ -type) will be a horizontal line in the graph, because there are no retrials at the equilibrium. Since  $JN'B$  generates the highest welfare among all balk strategies,  $\{JnB : n \geq 1\}$ , the upper envelope of  $L_1, L_2, \dots, L_N$  and  $L$  will give the socially optimal welfare as a function of the retrial cost  $\alpha$  among the class of pure threshold policies that we study, namely  $\{s : s = JnR \text{ or } s = JnB \text{ for some } n \geq 1\}$ . Result then follows.  $\square$

***Proof of Lemma 9:***

The queueing system under a retry/balk strategy  $\sigma$  have the following birth and death rates:

$$\begin{aligned}
\pi_1^\sigma &= \rho^\sigma \pi_0^\sigma; \\
\pi_2^\sigma &= \rho^\sigma \pi_1^\sigma; \\
\pi_3^\sigma &= \rho^\sigma \pi_2^\sigma; \\
&\vdots \\
\pi_N^\sigma &= \rho^\sigma \pi_{N-1}^\sigma; \\
\pi_{N+1}^\sigma &= \theta l \pi_N^\sigma; \\
\pi_{N+2}^\sigma &= \theta l \pi_{N+1}^\sigma; \\
&\vdots
\end{aligned}$$

where

$$\rho^\sigma = \theta l + (1 - \theta) \frac{l}{1 - \pi_R^\sigma} \quad (\text{A.48})$$

$$\pi_B^\sigma + \pi_R^\sigma = \pi_N^\sigma + \pi_{N+1}^\sigma + \pi_{N+2}^\sigma + \dots = \frac{\pi_N^\sigma}{1 - \theta l} \quad (\text{A.49})$$

If the underlying queueing systems when the population adopts retry/balk strategy  $\sigma_1$  or  $\sigma_2$ , share the same steady-state balking and retrial probabilities, i.e., if  $\pi_B^{\sigma_1} = \pi_B^{\sigma_2}$  and  $\pi_R^{\sigma_1} = \pi_R^{\sigma_2}$ , then by (A.48) and (A.49),  $\rho_R^{\sigma_1} = \rho_R^{\sigma_2}$  and  $\pi_N^{\sigma_1} = \pi_N^{\sigma_2}$ . It will be true then  $\pi_x^{\sigma_1} = \pi_x^{\sigma_2}$  for all  $x \in \mathbb{N}_0$ .

Therefore, given two distinct strategies  $\sigma_1, \sigma_2 \in JN_{B(\gamma)}^{R(1-\gamma)}$  where  $\frac{\pi_B^{\sigma_1}}{\pi_B^{\sigma_1} + \pi_R^{\sigma_1}} = \frac{\pi_B^{\sigma_2}}{\pi_B^{\sigma_2} + \pi_R^{\sigma_2}} = \gamma \Rightarrow \pi_B^{\sigma_1} \pi_R^{\sigma_2} = \pi_B^{\sigma_2} \pi_R^{\sigma_1}$ , we will only need to show that  $\pi_B^{\sigma_1} = \pi_B^{\sigma_2}$  and  $\pi_R^{\sigma_1} = \pi_R^{\sigma_2}$ .

WLOG, assume that  $\pi_B^{\sigma_1} < \pi_B^{\sigma_2}$  and  $\pi_R^{\sigma_1} < \pi_R^{\sigma_2}$ . (Signs need to be the same for  $\pi_B^{\sigma_1} \pi_R^{\sigma_2} =$

$\pi_B^{\sigma_2} \pi_R^{\sigma_1}$  to hold.) According to (A.48) and (A.49),  $\rho_R^{\sigma_1} < \rho_R^{\sigma_2}$  and  $\pi_N^{\sigma_1} < \pi_N^{\sigma_2}$ . Also, since

$$\pi_0^\sigma = 1 - \frac{1 - \pi_R^\sigma - \pi_B^\sigma}{1 - \pi_R^\sigma} l = 1 - \left(1 - \frac{\pi_B^\sigma}{1 - \pi_R^\sigma}\right) l,$$

we must have  $\pi_0^{\sigma_1} < \pi_0^{\sigma_2}$ . Then,  $\pi_x^{\sigma_1} < \pi_x^{\sigma_2}$  for all  $x \in \mathbb{N}_0$ . And this is a contradiction to the fact that  $\sum_{x \in \mathbb{N}_0} \pi_x^{\sigma_1} = \sum_{x \in \mathbb{N}_0} \pi_x^{\sigma_2} = 1$ .  $\square$

**Proof of Lemma 5':**

Proof is similar to that for the one-class case. The new stability condition is

$$\begin{aligned} \rho^{JnR} &= \theta l + (1 - \theta) \frac{l}{1 - \pi_R^{JnR}} \\ \Leftrightarrow \rho^{JnR} - \theta l &= \frac{(1 - \theta l)(l - \theta l)}{1 - \theta l - (1 - l)(\rho^{JnR})^n} \end{aligned} \quad (\text{A.50})$$

$$\begin{aligned} \text{since } \pi_R^{JnR} &= \frac{\pi_0^{JnR} (\rho^{JnR})^n}{1 - \theta l} = \frac{(1 - l)(\rho^{JnR})^n}{1 - \theta l} \\ \Leftrightarrow (1 - l)(\rho^{JnR})^n &= \frac{(1 - \theta l)(\rho^{JnR} - l)}{\rho^{JnR} - \theta l} \\ \Leftrightarrow (\rho^{JnR})^n &= \frac{(1 - \theta l)(\rho^{JnR} - l)}{(\rho^{JnR} - \theta l)(1 - l)} \\ \Leftrightarrow n &= \frac{\ln \frac{(1 - \theta l)(\rho^{JnR} - l)}{(\rho^{JnR} - \theta l)(1 - l)}}{\ln \rho^{JnR}} \end{aligned} \quad (\text{A.51})$$

The new indifference condition is

$$\begin{aligned} \frac{c}{\mu}(n + 1) &= \frac{c}{\mu} \sum_{k=0}^{n-1} \frac{\pi_k^{JnR}}{1 - \pi_R^{JnR}} (k + 1) + \frac{\alpha}{1 - \pi_R^{JnR}} \\ \Leftrightarrow (1 - \pi_R^{JnR}) \frac{c}{\mu}(n + 1) &= \frac{c}{\mu} \sum_{k=0}^{n-1} \pi_k^{JnR} (k + 1) + \alpha \\ \Leftrightarrow (1 - \pi_R^{JnR}) \frac{c}{\mu} n &= \frac{c}{\mu} \sum_{k=0}^{n-1} \pi_k^{JnR} k + \alpha \\ \Leftrightarrow \frac{c}{\mu} n &= \frac{c}{\mu} \sum_{k=0}^n \pi_k^{JnR} k + \frac{c}{\mu} \pi_R^{JnR} n + \alpha \\ \Leftrightarrow \sum_{k=0}^n \pi_k^{JnR} (n - k) &= r \end{aligned}$$



$$\begin{aligned}
&\Leftrightarrow (1 - \rho^{JnR})r = \pi_0^{JnR}n - \pi_0^{JnR}\rho(1 + \rho^{JnR} + \rho^{JnR^2} + \dots + (\rho^{JnR})^n) \\
&\Leftrightarrow (1 - \rho^{JnR})r = \pi_0^{JnR}n - \pi_0^{JnR}\rho \frac{1 - (\rho^{JnR})^n}{1 - \rho^{JnR}} \\
&\Leftrightarrow (1 - \rho^{JnR})r = (1 - l)n - \frac{(1 - \theta)\rho^{JnR}l}{\rho^{JnR} - \theta l} \tag{A.52}
\end{aligned}$$

$$\begin{aligned}
&\text{since (A.51)} \Rightarrow \frac{1 - (\rho^{JnR})^n}{1 - \rho^{JnR}} = \frac{(1 - \theta)l}{(\rho^{JnR} - \theta l)(1 - l)} \\
&\Leftrightarrow n = \frac{(1 - \rho^{JnR})(\rho^{JnR} - \theta l)r + (1 - \theta)\rho^{JnR}l}{(1 - l)(\rho^{JnR} - \theta l)} \tag{A.53}
\end{aligned}$$

Note that (A.51) and (A.53) reduce to (A.23) and (A.25) when  $\theta = 0$ . Define

$$\begin{aligned}
f_3(\rho) &\triangleq \frac{\ln \frac{(1-\theta l)(\rho-l)}{(\rho-\theta l)(1-l)}}{\ln \rho} \text{ from (A.51);} \\
f_4(\rho) &\triangleq \frac{(1-\rho)(\rho-\theta l)r + (1-\theta)\rho l}{(1-l)(\rho-\theta l)} \text{ from (A.53).}
\end{aligned}$$

It can be shown that  $f_3(\rho)$  and  $f_4(\rho)$  always intercept at  $\rho = 1$  with a function value of  $\frac{l}{1-l} \cdot \frac{1-\theta}{1-\theta l}$  regardless of  $r$ , but intercept at only one point, say  $n$ , other than  $\rho \neq 1$ . As  $r \uparrow$  (i.e.,  $\alpha \uparrow$ ),  $n \uparrow$  and  $\rho \downarrow$ . Similar as before, we can choose  $\alpha_1, \alpha_2, \dots, \alpha_N$  such that for  $n = 1, 2, \dots, N$  and for  $\alpha \in [\alpha_n - \frac{c}{\mu}(1 - \pi_R^{JnR}), \alpha_n]$ , the pair  $(n, \rho) = (n, \rho^{JnR})$  satisfies both conditions (A.51) and (A.53), thus  $JnR$  is an equilibrium strategy. Considering the boundary conditions, we then have the result.  $\square$

***Proof of Lemma 7', Theorem 4', Proposition 3' and Theorem 5':***

Omitted. See the proof of Lemma 7, Theorem 4, Proposition 3 and Theorem 5, respectively.

$\square$

***Proof of Theorem 6', Theorem 7' and Proposition 8':***

Using an argument similar to Proposition 4, it can be shown that the total consumer welfare for the strategic population decreases in *alpha* when  $\alpha \leq \alpha_L$ , from  $(1 - \theta)(\lambda v - cl)$  to  $(1 - \theta)\lambda(v - cW^{JNR} - (\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L)$ . Dividing the quantities by  $(1 - \theta)\lambda$  gives the welfare per strategic consumer, going from  $v - \frac{c}{\mu}$  to  $v - cW^{JNR} - (\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L = \alpha_L$ .

The consumer welfare per strategic consumer when the retrial cost  $\alpha \in [\alpha_L, \alpha_H]$  is given by

$$\begin{aligned}
& \frac{1}{(1-\theta)\lambda} \left[ \frac{(1-\theta)\lambda \left(1 - \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}\right)}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\theta)\lambda(1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \alpha \right] \\
&= \frac{\left(1 - \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}\right)}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \alpha \\
&= \frac{1}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \left(1 - \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}\right) (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \alpha \\
&= \frac{1}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \alpha - \frac{(1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}}{1 - (1-\gamma) \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}} \alpha \\
&\quad \text{since } \alpha = \left(1 - \sum_{n=N}^{\infty} \pi_n JN_{B(\gamma)}^{R(1-\gamma)}\right) (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \\
&= \alpha
\end{aligned}$$

Therefore, the welfare per strategic consumer increases linearly in  $\alpha$  from  $\alpha_L$  to  $\alpha_H$ .

When  $\alpha \geq \alpha_H$ , the consumer welfare will stay constant as the retrial cost varies (because no consumer retries and pays for the retrial cost). The total consumer welfare for the strategic population is  $(1-\theta)\lambda \left(1 - \frac{\pi_N^{JNB}}{1-\theta l}\right) (v - cW^{JNB})$ , and therefore welfare per strategic consumer equals  $\left(1 - \frac{\pi_N^{JNB}}{1-\theta l}\right) (v - cW^{JNB}) = \alpha_H$ .

Finally, it is easy to see that welfare per strategic consumer on  $\alpha \geq \alpha_H$  is less than the welfare when  $\alpha$  approaches 0, because the total welfare reaches the maximum possible at the value of  $(1-\theta)\lambda \left(v - \frac{c}{\mu}\right)$  when  $\alpha$  approaches 0. If the objective function were changed to maximize welfare per strategic consumer with respect to the retrial cost, the maximizer remains at  $\alpha = 0$  regardless of the value of  $\theta$ .  $\square$

**Proof of Lemma 10:**

We only prove part (ii) here. Proofs to part (i) and (iii) are given in the proofs to Theorem 10 and Proposition 4. For part (ii), fix  $\gamma \in [0, 1]$ . Since

$$\begin{aligned}\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} &= 1 - \left[ \theta + (1 - \theta) \frac{1 - \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} \right] l \\ \rho^{JN_{B(\gamma)}^{R(1-\gamma)}} &= \left( \theta + \frac{1 - \theta}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} \right) l,\end{aligned}$$

and  $\frac{1 - \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} < 1$ ,  $\frac{1}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} > 1$ , we have by taking derivatives that

$\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}$  and  $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$  both decrease in  $\theta$ . As a result,  $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\dots$ ,  $\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}$

all decrease in  $\theta$ . Recall that

$$\begin{aligned}\alpha(\gamma) &= \left( 1 - \sum_{i=N}^{\infty} \pi_i^{JN_{B(\gamma)}^{R(1-\gamma)}} \right) (v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \\ &= \left( \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \right) v \\ &\quad - \left( \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} \frac{c}{\mu} + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} \frac{2c}{\mu} + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \frac{Nc}{\mu} \right),\end{aligned}$$

we can conclude  $\alpha(\gamma)$  decreases in  $\theta$  because  $v > \frac{Nc}{\mu} > \frac{(N-1)c}{\mu} > \dots > \frac{c}{\mu}$  and

$$\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}, \dots, \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}$$

all decrease in  $\theta$ . □

**Proof of Theorem 10:**

Recall that  $\mathcal{U}_{\alpha, \theta}^{\sigma}$  denotes the welfare *per strategic consumer* when the environment is  $\theta$ , the retrieval cost is  $\alpha$  and the strategy being adopted by the population is  $\sigma$ . The idea is to show that the down-up-flat welfare-per-strategic-consumer curve (given by the Pareto-dominant equilibrium strategies) with environment  $\theta_2$  dominates that with environment  $\theta_1$ . Note that

when the retrial cost  $\alpha \rightarrow 0$ , the Pareto-dominant strategy (either with environment  $\theta_1$  or  $\theta_2$ ) is for every strategic consumer to join an idle server and retry whenever the server is busy (i.e., the retrial strategy  $J1R$ ). Therefore,

$$\lim_{\alpha \rightarrow 0} \mathcal{U}_{\alpha, \theta_1}^* = \lim_{\alpha \rightarrow 0} \mathcal{U}_{\alpha, \theta_2}^* = v - \frac{c}{\mu}$$

To show  $\mathcal{U}_{\alpha, \theta_1}^* \leq \mathcal{U}_{\alpha, \theta_2}^*$  for all  $\alpha$ , we will use the fact that the welfare curve between  $\alpha_L$  and  $\alpha_H$  is a 45-degree straight line. It suffices to prove that

- (i) the slope of each piecewise welfare curve on  $\alpha \leq \alpha_L$  becomes steeper as  $\theta$  increases, i.e.,  $\frac{1}{1-\pi_R^{JnR}}$  increases in  $\theta$  for  $n = 1, 2, \dots, N$ ;
- (ii) The values of  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}, \alpha_L$  decrease in  $\theta$ .
- (iii) Value on the left end of each of the piecewise welfare curves on  $\alpha \leq \alpha_L$  decreases in  $\theta$ , i.e.,  $\mathcal{U}_{\alpha_{n-1}, \theta}^{JnR}$  decreases in  $\theta$  for  $n = 2, \dots, N$ ;
- (iv) Value on the right end of each of the piecewise welfare curves on  $\alpha \leq \alpha_L$  decreases in  $\theta$ , i.e.,  $\mathcal{U}_{\alpha_n, \theta}^{JnR}$  decreases in  $\theta$  for  $n = 1, 2, \dots, N-1$  and  $\mathcal{U}_{\alpha_L, \theta}^{JNR}$  decreases in  $\theta$ ;
- (v) Finally,  $\mathcal{U}_{\alpha_H, \theta}^{JNB}$  decreases in  $\theta$ .

For (i), fix  $n \in \{1, 2, \dots, N\}$ . The steady-state probabilities of the underlying system under  $JnR$  satisfy

$$\begin{aligned} \pi_1^{JnR} &= \rho^{JnR} \pi_0^{JnR}; \\ \pi_2^{JnR} &= \rho^{JnR} \pi_1^{JnR}; \\ \pi_3^{JnR} &= \rho^{JnR} \pi_2^{JnR}; \\ &\vdots \\ \pi_n^{JnR} &= \rho^{JnR} \pi_{n-1}^{JnR}; \\ \pi_{n+1}^{JnR} &= \theta l \pi_n^{JnR}; \end{aligned}$$

$$\begin{aligned}\pi_{n+2}^{JnR} &= \theta l \pi_{n+1}^{JnR}; \\ &\vdots = \vdots\end{aligned}$$

Since  $\pi_0^{JnR} \equiv 1 - l$  (none of the strategic or myopic consumer balks),  $\rho^{JnR}$  must decrease when  $\theta$  increases. Therefore,  $\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR}$  all decrease in  $\theta$ , which implies that  $\pi_R^{JnR} = 1 - (\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR})$  increases. And thus,  $\frac{1}{1 - \pi_R^{JnR}}$  increases in  $\theta$ .

For (ii), recall from (A.29) that

$$\begin{aligned}\alpha_L &= (1 - \pi_R^{JNR})(v - cW^{JNR}) \\ &= (1 - \pi_R^{JNR})v - (1 - \pi_R^{JNR})(cW^{JNR}) \\ &= (1 - \pi_R^{JNR})v - \left(\pi_0^{JNR} \frac{c}{\mu} + \pi_1^{JNR} \frac{2c}{\mu} + \dots + \pi_{N-1}^{JNR} \frac{Nc}{\mu}\right)\end{aligned}\tag{A.54}$$

From (i), we know that  $\pi_0^{JNR}, \pi_1^{JNR}, \dots, \pi_{N-1}^{JNR}$  all decrease in  $\theta$  while  $\pi_R^{JNR}$  increases in  $\theta$ . And the total reduction by  $\pi_0^{JNR}, \pi_1^{JNR}, \dots, \pi_{N-1}^{JNR}$  must be equal to the reduction of  $1 - \pi_R^{JNR}$ , as they sum to 1. Since  $v > \frac{Nc}{\mu} > \frac{(N-1)c}{\mu} > \dots > \frac{2c}{\mu} > \frac{c}{\mu}$ , it is clear from (A.54) that  $\alpha_L$  decreases in  $\theta$ .

Now fix  $n \in \{1, 2, \dots, N-1\}$ . The retrial payoff when the population adopts  $JnR$  with the retrial cost  $\alpha_n$  equals the joining payoff at state  $n$ , i.e.,

$$v - cW^{JnR} - \frac{\alpha_n}{1 - \pi_R^{JnR}} = v - \frac{(n+1)c}{\mu}\tag{A.55}$$

$$\begin{aligned}\Leftrightarrow \alpha_n &= (1 - \pi_R^{JnR})\left(\frac{(n+1)c}{\mu} - cW^{JnR}\right) \\ &= (1 - \pi_R^{JnR})\frac{(n+1)c}{\mu} - (1 - \pi_R^{JnR})(cW^{JnR}) \\ &= (1 - \pi_R^{JnR})\frac{(n+1)c}{\mu} - \left(\pi_0^{JnR} \frac{c}{\mu} + \pi_1^{JnR} \frac{2c}{\mu} + \dots + \pi_{n-1}^{JnR} \frac{nc}{\mu}\right)\end{aligned}\tag{A.56}$$

From (i), we know that  $\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR}$  all decrease in  $\theta$  while  $\pi_R^{JnR}$  increases in  $\theta$ .

And the total reduction by  $\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR}$  must be equal to the reduction of  $1 - \pi_R^{JnR}$ . Since  $\frac{(n+1)c}{\mu} > \frac{nc}{\mu} > \dots > \frac{2c}{\mu} > \frac{c}{\mu}$ , it is clear from (A.56) that  $\alpha_n$  decreases in  $\theta$ .

For (iii), fix  $n \in \{2, \dots, N\}$ . The retrial payoff when the population adopts  $JnR$  with the retrial cost  $\alpha_{n-1}$  equals the joining payoff at state  $n - 1$ , i.e., the indifference condition is binding at the smaller side:

$$v - cW^{JnR} - \frac{\alpha_{n-1}}{1 - \pi_R^{JnR}} = v - \frac{nc}{\mu} \quad (\text{A.57})$$

Therefore,

$$\begin{aligned} & \mathcal{U}_{\alpha_{n-1}, \theta}^{JnR} \\ &= v - cW^{JnR} - \left( \frac{1}{1 - \pi_R^{JnR}} - 1 \right) \alpha_{n-1} \\ &= v - [(1 - \pi_R^{JnR})cW^{JnR} + \pi_R^{JnR}cW^{JnR}] - \frac{\pi_R^{JnR}}{1 - \pi_R^{JnR}} \alpha_{n-1} \\ &= v - \left[ \pi_0^{JnR} \frac{c}{\mu} + \pi_1^{JnR} \frac{2c}{\mu} + \dots + \pi_{n-1}^{JnR} \frac{nc}{\mu} + \pi_R^{JnR} cW^{JnR} \right] - \frac{\pi_R^{JnR}}{1 - \pi_R^{JnR}} \alpha_{n-1} \\ &= \pi_0^{JnR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JnR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{n-1}^{JnR} \left( v - \frac{nc}{\mu} \right) + \pi_R^{JnR} \left( v - cW^{JnR} - \frac{\alpha_{n-1}}{1 - \pi_R^{JnR}} \right) \\ &= \pi_0^{JnR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JnR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{n-1}^{JnR} \left( v - \frac{nc}{\mu} \right) + \pi_R^{JnR} \left( v - \frac{nc}{\mu} \right) \\ & \quad \text{due to (A.57)} \end{aligned} \quad (\text{A.58})$$

As  $\theta$  increases,  $\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR}$  all decrease and  $\pi_R^{JnR}$  increases. It is clear from (A.58) that  $\mathcal{U}_{\alpha_{n-1}, \theta}^{JnR}$  decreases in  $\theta$ .

For (iv), fix  $n \in \{1, 2, \dots, N - 1\}$ . Then,

$$\begin{aligned} & \mathcal{U}_{\alpha_n, \theta}^{JnR} \\ &= v - cW^{JnR} - \left( \frac{1}{1 - \pi_R^{JnR}} - 1 \right) \alpha_n \\ &= v - [(1 - \pi_R^{JnR})cW^{JnR} + \pi_R^{JnR}cW^{JnR}] - \frac{\pi_R^{JnR}}{1 - \pi_R^{JnR}} \alpha_n \end{aligned}$$

$$\begin{aligned}
&= v - \left[ \pi_0^{JnR} \frac{c}{\mu} + \pi_1^{JnR} \frac{2c}{\mu} + \dots + \pi_{n-1}^{JnR} \frac{nc}{\mu} + \pi_R^{JnR} cW^{JnR} \right] - \frac{\pi_R^{JnR}}{1 - \pi_R^{JnR}} \alpha_n \\
&= \pi_0^{JnR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JnR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{n-1}^{JnR} \left( v - \frac{nc}{\mu} \right) + \pi_R^{JnR} \left( v - cW^{JnR} - \frac{\alpha_n}{1 - \pi_R^{JnR}} \right) \\
&= \pi_0^{JnR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JnR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{n-1}^{JnR} \left( v - \frac{nc}{\mu} \right) + \pi_R^{JnR} \left( v - \frac{(n+1)c}{\mu} \right) \text{ due to (A.55)}
\end{aligned} \tag{A.59}$$

As  $\theta$  increases,  $\pi_0^{JnR}, \pi_1^{JnR}, \dots, \pi_{n-1}^{JnR}$  all decrease and  $\pi_R^{JnR}$  increases. And  $\mathcal{U}_{\alpha_n, \theta}^{JnR}$  decreases in  $\theta$  due to (A.59).

When  $n = N$ . The retrial payoff when the population adopts  $JNR$  with retrial cost  $\alpha_L$  equals 0 because we recall from (A.29) that  $v - cW^{JNR} - \frac{\alpha_L}{1 - \pi_R^{JNR}} = 0$ . Then,

$$\begin{aligned}
&\mathcal{U}_{\alpha_L, \theta}^{JNR} \\
&= v - cW^{JNR} - \left( \frac{1}{1 - \pi_R^{JNR}} - 1 \right) \alpha_L \\
&= v - \left[ (1 - \pi_R^{JNR}) cW^{JNR} + \pi_R^{JNR} cW^{JNR} \right] - \frac{\pi_R^{JNR}}{1 - \pi_R^{JNR}} \alpha_L \\
&= v - \left[ \pi_0^{JNR} \frac{c}{\mu} + \pi_1^{JNR} \frac{2c}{\mu} + \dots + \pi_{N-1}^{JNR} \frac{Nc}{\mu} + \pi_R^{JNR} cW^{JNR} \right] - \frac{\pi_R^{JNR}}{1 - \pi_R^{JNR}} \alpha_L \\
&= \pi_0^{JNR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JNR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{N-1}^{JNR} \left( v - \frac{Nc}{\mu} \right) + \pi_R^{JNR} \left( v - cW^{JNR} - \frac{\alpha_L}{1 - \pi_R^{JNR}} \right) \\
&= \pi_0^{JNR} \left( v - \frac{c}{\mu} \right) + \pi_1^{JNR} \left( v - \frac{2c}{\mu} \right) + \dots + \pi_{N-1}^{JNR} \left( v - \frac{Nc}{\mu} \right)
\end{aligned} \tag{A.60}$$

As  $\theta$  increases,  $\pi_0^{JNR}, \pi_1^{JNR}, \dots, \pi_{N-1}^{JNR}$  all decrease. It is clear from (A.60) that  $\mathcal{U}_{\alpha_L, \theta}^{JNR}$  decreases in  $\theta$ .

For (v). The steady-state probabilities of the underlying system under  $JNB$  satisfy

$$\begin{aligned}
\pi_1^{JNB} &= l\pi_0^{JNB}; \\
\pi_2^{JNB} &= l\pi_1^{JNB}; \\
\pi_3^{JNB} &= l\pi_2^{JNB};
\end{aligned}$$

$$\begin{aligned}
& \vdots = \vdots \\
& \pi_N^{JNB} = l\pi_{N-1}^{JNB}; \\
& \pi_{N+1}^{JNB} = \theta l\pi_N^{JNB}; \\
& \pi_{N+2}^{JNB} = \theta l\pi_{N+1}^{JNB}; \\
& \vdots = \vdots
\end{aligned}$$

When  $\theta$  increases,  $\pi_0^{JNB}$  must decrease. Therefore,  $\pi_0^{JNB}, \pi_1^{JNB}, \dots, \pi_{N-1}^{JNB}$  all decrease in  $\theta$ . Since

$$\mathcal{U}_{\alpha_H, \theta}^{JNB} = \pi_0^{JNB} \left(v - \frac{c}{\mu}\right) + \pi_1^{JNB} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{N-1}^{JNB} \left(v - \frac{Nc}{\mu}\right),$$

$\mathcal{U}_{\alpha_H, \theta}^{JNB}$  clearly decreases in  $\theta$ . □

***Proof of Theorem 11:***

Welfare per myopic consumer under an equilibrium strategy  $\sigma$  (played by the strategic consumers) is given by

$$\pi_0^\sigma \left(v - \frac{c}{\mu}\right) + \pi_1^\sigma \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^\sigma \left(v - \frac{nc}{\mu}\right) + \pi_n^\sigma \left(v - \frac{(n+1)c}{\mu}\right) \dots \quad (\text{A.61})$$

It forms a step function over  $\alpha \leq \alpha_L$  because equilibrium strategy remains the same for strategic consumers for all  $\alpha$  that falls in one of the  $N$  intervals:

$$(0, \alpha_1], (\alpha_1, \alpha_2], \dots, (\alpha_{N-2}, \alpha_{N-1}], (\alpha_{N-1}, \alpha_L].$$

When the retrial cost increases from some value in one interval to the next interval, i.e., when the equilibrium strategy jumps from  $JnR$  to  $J(n+1)R$ , we have  $\rho^{JnR} > \rho^{J(n+1)R}$ . As a result,

$$\begin{aligned}
\pi_0^{JnR} &= \pi_0^{J(n+1)R} = 1 - l \\
\pi_x^{JnR} &> \pi_x^{J(n+1)R} \text{ for } x = 1, 2, \dots, n
\end{aligned}$$



$$\pi_x^{JnR} < \pi_x^{J(n+1)R} \text{ for } x = n+1, n+2, \dots$$

As the weights of these steady-state probabilities shift to the right, it is easy to see from (A.61) that welfare per myopic consumer decreases on  $\alpha \leq \alpha_L$ .

When  $\alpha \in [\alpha_L, \alpha_H]$ , the equilibrium strategy is  $JN_{B(\gamma)}^{R(1-\gamma)}$  for some  $\gamma$ . As  $\alpha$  increases within this region,  $\gamma$  also increases. As a result,  $\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}$  decreases and  $\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}$  increases in  $\alpha$ . However,  $\frac{\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\theta l} = \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} + \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}$  must decrease in  $\alpha$ . Because otherwise both  $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$  and  $\gamma$  would increase in  $\alpha$ , then  $\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} = \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \gamma$  increases and cause a contradiction.

According to (A.61), welfare per myopic consumer is given by

$$\begin{aligned} & \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{c}{\mu}\right) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n)c}{\mu}\right) + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+1)c}{\mu}\right) \dots \\ &= [\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{c}{\mu}\right) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n)c}{\mu}\right)] + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+1)c}{\mu}\right) \dots \\ &= [(1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})] + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+1)c}{\mu}\right) + \pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+2)c}{\mu}\right) \dots \\ &= \alpha + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+1)c}{\mu}\right) + \pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(n+2)c}{\mu}\right) \dots \end{aligned} \quad (\text{A.62})$$

Since  $\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\pi_{n+2}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ ,  $\dots$  all decreases in  $\alpha$ , and  $0 > v - \frac{(n+1)c}{\mu} > v - \frac{(n+2)c}{\mu} > \dots$ , welfare per myopic consumer given by (A.62) increases in  $\alpha$ .

Finally, welfare per myopic consumer remains the same when  $\alpha \geq \alpha_H$  because strategic consumers adopt the balk strategy over this region.

Next, recall that  $\mathcal{V}_{\alpha, \theta}^\sigma$  denotes the welfare *per myopic consumer* when the environment is  $\theta$ , the retrial cost is  $\alpha$  and the strategy being adopted by the strategic population is  $\sigma$ . We have already proved in Lemma 10 and Theorem 10 that the values of  $\alpha_1, \alpha_2, \dots, \alpha_{N-1}, \alpha_L, \alpha(\gamma)$  for each  $\gamma \in [0, 1]$  (including  $\alpha_L$  and  $\alpha_H$ ) all decrease in  $\theta$ , meanwhile the welfare forms a step function on  $\alpha \leq \alpha_L$  and is increasing between  $\alpha_L$  and  $\alpha_H$  for any fixed  $\theta$ . To prove  $\mathcal{V}_{\alpha, \theta_1}^* \leq \mathcal{V}_{\alpha, \theta_2}^*$  for all  $\alpha$ , it then suffices to show

(i) Value on the right end of each of the piecewise welfare curves on  $\alpha \leq \alpha_L$  decreases in  $\theta$ , i.e.,  $\mathcal{V}_{\alpha_n, \theta}^{JnR}$  decreases in  $\theta$  for  $n = 1, 2, \dots, N - 1$  and  $\mathcal{V}_{\alpha_L, \theta}^{JNR}$  decreases in  $\theta$ ;

(ii) Fix  $\gamma \in [0, 1]$ .  $\mathcal{V}_{\alpha(\gamma), \theta}^{JN_{B(\gamma)}^{R(1-\gamma)}}$  decreases in  $\theta$ .

For (i), fix  $n \in \{1, 2, \dots, N - 1\}$ . The steady-state probabilities of the underlying system under  $JnR$  satisfy

$$\begin{aligned}\pi_1^{JnR} &= \rho^{JnR} \pi_0^{JnR}, \\ \pi_2^{JnR} &= \rho^{JnR} \pi_1^{JnR}, \\ \pi_3^{JnR} &= \rho^{JnR} \pi_2^{JnR}, \\ &\vdots \\ \pi_n^{JnR} &= \rho^{JnR} \pi_{n-1}^{JnR}, \\ \pi_{n+1}^{JnR} &= \theta l \pi_n^{JnR}, \\ \pi_{n+2}^{JnR} &= \theta l \pi_{n+1}^{JnR}, \\ &\vdots\end{aligned}$$

Since  $\pi_0^{JnR} \equiv 1 - l$  (none of the strategic or myopic consumers balks),  $\rho^{JnR}$  must decrease when  $\theta$  increases. Therefore, the weights of the steady-state probabilities shift to the left, and as a result, welfare per myopic consumer,

$$\mathcal{V}_{\alpha_n, \theta}^{JnR} = \pi_0^{JnR} \left(v - \frac{c}{\mu}\right) + \pi_1^{JnR} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^{JnR} \left(v - \frac{nc}{\mu}\right) + \pi_n^{JnR} \left(v - \frac{(n+1)c}{\mu}\right) \dots,$$

decreases in  $\theta$ . Now at  $\alpha_L$ ,

$$\mathcal{V}_{\alpha_L, \theta}^{JNR} = \pi_0^{JNR} \left(v - \frac{c}{\mu}\right) + \pi_1^{JNR} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{N-1}^{JNR} \left(v - \frac{Nc}{\mu}\right) + \pi_N^{JNR} \left(v - \frac{(N+1)c}{\mu}\right) \dots,$$

decreases in  $\theta$  for the same reason.

For (ii), fix  $\gamma \in [0, 1]$ ,

$$\mathcal{V}_{\alpha(\gamma),\theta}^{JN_{B(\gamma)}^{R(1-\gamma)}} = \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{c}{\mu}\right) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{ic}{\mu}\right) + \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \left(v - \frac{(N+1)c}{\mu}\right) \dots$$

Since both  $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}$  and  $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$  decrease in  $\theta$ , the steady-state probabilities shift to the left, and as a result,  $\mathcal{V}_{\alpha(\gamma),\theta}^{JN_{B(\gamma)}^{R(1-\gamma)}}$  decreases in  $\theta$ .  $\square$

**Proof of Proposition 4:**

Let  $\pi_0^M, \pi_1^M, \pi_2^M, \dots$ , denote the steady-state probabilities of a regular  $M/M/1$  system (no balking and no retrials). Recall from (A.61) that the consumer welfare per myopic consumer when the strategic class follows the strategy  $\sigma$  equals

$$\pi_0^\sigma \left(v - \frac{c}{\mu}\right) + \pi_1^\sigma \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^\sigma \left(v - \frac{nc}{\mu}\right) + \pi_n^\sigma \left(v - \frac{(n+1)c}{\mu}\right) \dots \quad (\text{A.63})$$

As  $\theta \rightarrow 1$ , e.g., imagine there is only 1 strategic consumer among the new arrivals in each period, what strategy this strategic consumer or the strategic population adopts would have no impact on the steady-state probabilities of the underlying queueing system. That is,  $\pi_x^\sigma \rightarrow \pi_x^M$  as  $\theta \rightarrow 1$  for all  $x \in \mathbb{N}_0$ .

Therefore, for  $n = 1, 2, \dots, N-1$ , we have

$$\mathcal{V}_{\alpha_n,\theta}^{JnR} \rightarrow \pi_0^M \left(v - \frac{c}{\mu}\right) + \pi_1^M \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{n-1}^M \left(v - \frac{nc}{\mu}\right) + \pi_n^M \left(v - \frac{(n+1)c}{\mu}\right) \dots = v - \frac{c}{\mu - \lambda}.$$

Similarly, we have

$$\begin{aligned} \mathcal{V}_{\alpha_L,\theta}^{JNR} &= \mathcal{V}_{\alpha_H,\theta}^{JNB} \\ &\rightarrow \pi_0^M \left(v - \frac{c}{\mu}\right) + \pi_1^M \left(v - \frac{2c}{\mu}\right) + \dots + \pi_{N-1}^M \left(v - \frac{Nc}{\mu}\right) + \pi_N^M \left(v - \frac{(N+1)c}{\mu}\right) \dots = v - \frac{c}{\mu - \lambda}. \end{aligned}$$

Also as  $\theta \rightarrow 1$ , we have

$$\alpha_L = \left(1 - \frac{\pi_N^{JNR}}{1 - \theta l}\right) (v - cW^{JNR}) \rightarrow \left(1 - \frac{\pi_N^M}{1 - l}\right) (v - cW^M) = \sum_{x=0}^{N-1} \pi_x^M \left[v - \frac{(x+1)c}{\mu}\right];$$

$$\alpha_H = (1 - \frac{\pi_N^{JNB}}{1 - \theta l})(v - cW^{JNB}) \rightarrow (1 - \frac{\pi_N^M}{1 - l})(v - cW^M) = \sum_{x=0}^{N-1} \pi_x^M [v - \frac{(x+1)c}{\mu}]$$

where  $W^M$  is the limiting conditional waiting time defined by

$$W^M \triangleq \frac{\pi_0^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{c}{\mu} + \frac{\pi_1^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{2c}{\mu} + \frac{\pi_2^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{3c}{\mu} + \dots + \frac{\pi_{N-1}^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{Nc}{\mu}.$$

It becomes clear that  $\alpha_H \rightarrow \alpha_L$  as  $\theta \rightarrow 1$ , although both  $\alpha_H$  and  $\alpha_L$  decrease in  $\theta$ .

On the other hand, it should be noted that not any two points in the set  $\{\alpha_1, \alpha_2, \dots, \alpha_{N-1}\}$  will converge to one point as  $\theta \rightarrow 1$ . Since  $v - cW^{JnR} - \frac{\alpha_i}{1 - \pi_R^{JnR}} = v - \frac{(n+1)c}{\mu}$  for any  $\theta \in [0, 1)$  (e.g., see (A.55)), as  $\theta \rightarrow 1$ , we have for any  $n \in \{1, 2, \dots, N-1\}$ ,

$$\begin{aligned} \alpha_n &\rightarrow \left( \sum_{x=0}^{n-1} \pi_x^M \right) \cdot \left[ \left( \frac{(n+1)c}{\mu} \right) - c \left( \frac{\pi_0^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{c}{\mu} + \frac{\pi_1^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{2c}{\mu} + \frac{\pi_2^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{3c}{\mu} + \dots + \frac{\pi_{n-1}^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{nc}{\mu} \right) \right] \\ &= \sum_{x=0}^{n-1} \pi_x^M \left[ \frac{(n+1)c}{\mu} - \frac{(x+1)c}{\mu} \right] = \sum_{x=0}^{n-1} \pi_x^M \left[ \frac{(n-x)c}{\mu} \right]. \end{aligned}$$

Therefore, as  $\theta \rightarrow 1$ , we still have  $\alpha_1 < \alpha_2 < \dots < \alpha_{N-1}$ . □

### ***Alternative modeling on the time between retrials:***

In the model presented in the paper, a retrial consumer always returns in the following period. We show here that all the results of the paper will still hold if the time to return after a consumer has made a retry decision is either  $X$  number of periods where  $X$  is a positive finite integer random variable, or  $T$  amount of time where  $T$  is exponentially distributed with some rate  $t$ , like in orbital models. (Note that when  $X$  assumes infinity, it represents a balk decision instead, so we do not consider it.)

The key is to show under both cases, the total arrival rate is still  $\lambda_{total}^\sigma = \frac{\lambda}{1 - \pi_R^\sigma}$ , and the long-run idle probability of the server is still given by  $\pi_0^\sigma = 1 - \frac{\pi_J^\sigma}{1 - \pi_R^\sigma} l$ , when the population adopts some strategy  $\sigma$  at equilibrium. Then, all the steady-state probabilities of the underlying

queue will be calculated the same way as before.

First, we assume that each retrial consumer returns in  $X$  number of periods, where  $X$  takes on values in  $\{x_1, x_2, \dots, x_k\}$  with probabilities  $p_1, p_2, \dots, p_k$ , respectively. Then the total arrival rate is given by

$$\lambda_{total}^\sigma = \lambda + \lambda \pi_R^\sigma \sum_{i=1}^k p_k + \lambda (\pi_R^\sigma)^2 \left( \sum_{i=1}^k p_k \right)^2 + \lambda (\pi_R^\sigma)^3 \left( \sum_{i=1}^k p_k \right)^3 + \dots = \frac{\lambda}{1 - \pi_R^\sigma},$$

and long-run idle probability of the server is  $\pi_0^\sigma = 1 - \frac{\pi_J^\sigma \cdot \lambda_{total}^\sigma}{\mu} = 1 - \frac{\pi_J^\sigma}{1 - \pi_R^\sigma} l$ .

In the case with exponential returning time, let  $L_t$  denote the average number of consumers in the orbit, i.e., the average number of consumers waiting to retry. (Note that  $L_t$  depends on  $t$ .) At equilibrium, equating the rates customers enter and leave the system, we have

$$\begin{aligned} \lambda \pi_R^\sigma + t L_t \pi_R^\sigma &= t L_t \\ \lambda \pi_R^\sigma &= t L_t (1 - \pi_R^\sigma) \\ t L_t &= \lambda \frac{\pi_R^\sigma}{1 - \pi_R^\sigma} \end{aligned}$$

Therefore the total arrival rate equals  $\lambda + t L_t = \lambda \left( 1 + \frac{\pi_R^\sigma}{1 - \pi_R^\sigma} \right) = \frac{\lambda}{1 - \pi_R^\sigma}$ .

Now, equating the rates customers enter and leave the orbit, we have

$$\begin{aligned} \lambda \pi_J^\sigma + t L_t \pi_J^\sigma &= \mu (1 - \pi_0^\sigma) \\ \lambda \frac{\pi_J^\sigma}{1 - \pi_R^\sigma} &= \mu (1 - \pi_0^\sigma) \\ \pi_0^\sigma &= 1 - \frac{\pi_J^\sigma}{1 - \pi_R^\sigma} l \end{aligned}$$

Claims are thus proved. □

## BIBLIOGRAPHY

- J. Abate and W. Whitt. The correlation functions of RBM and M/M/1. *Stochastic Models*, 4(2):315–359, 1988.
- P. Afèche and H. Mendelson. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science*, 50(7):869–882, 2004.
- S. Aguir, F. Karaesmen, Z. Akşin, and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376, 2004.
- S. Aguir, Z. Akşin, F. Karaesmen, and Y. Dallery. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, 191(2):398–408, 2008.
- A. Aissani. A retrial queue with redundancy and unreliable server. *Queueing Systems*, 17(3-4):431–449, 1994.
- Z. Akşin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.
- Z. Akşin, B. Ata, S. Emadi, and C.-L. Su. Structural estimation of callers’ delay sensitivity in call centers. *Management Science*, 59(12):2727–2746, 2013.
- S. Albin. On poisson approximations for superposition arrival processes in queues. *Management Science*, 28(2):126–137, 1982.
- G. Allon, A. Bassamboo, and Q. Yu. Do delay announcements influence customer behavior? an empirical study. *Kellogg Working Paper*, 2013.
- E. Ang, M. Bayati, S. Kwasnick, and E. Plambeck. Improving the prediction of emergency department waiting times. *Working Paper*, 2014a.
- E. Ang, D. Iancu, and R. Swinney. Disruption risk and optimal sourcing in multi-tier supply networks. *Working Paper*, 2014b.
- M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. *Operations Research*, 52(4):527–545, 2004a.
- M. Armony and C. Maglaras. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004b.
- B. C. Arntzen, G. G. Brown, T. P. Harrison, and L. L. Trafton. Global supply chain management at digital equipment corporation. *Interfaces*, 25(1):69–93, 1995.
- J. Artalejo. A queueing system with returning customers and waiting line. *Operations Research Letters*, 17(4):191–199, 1995.

- J. Artalejo. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers & Operations Research*, 24(6):493–504, 1997.
- J. Artalejo. Accessible bibliography on retrial queues. *Mathematical and Computer Modelling*, 30(3):1–6, 1999.
- J. Artalejo. Accessible bibliography on retrial queues: progress in 2000–2009. *Mathematical and Computer Modelling*, 51(9):1071–1081, 2010.
- J. Artalejo and M. Lopez-Herrero. On the single server retrial queue with balking. *INFOR. Information Systems and Operational Research*, 38(1):33–50, 2000.
- A. Berger and W. Whitt. Comparisons of multi-server queues with finite waiting rooms. *Communications in Statistics – Stochastic Models*, 8(4):719–732, 1992.
- O. Besbes and C. Maglaras. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research*, 57(6):1438–1450, 2009.
- O. Besbes, B. Dooley, and N. Gans. Dynamic service control of a queue with congestion-sensitive customers. *Working Paper*, 2011.
- B. Bhaskaran. Almost sure comparison of birth and death processes with application to  $m/m/s$  queueing systems. *Queueing Systems*, 1(1):103–127, 1986.
- J. H. Bookbinder and T. A. Matuk. Logistics and transportation in global supply chains: review, critique and prospects. *Tutorials in operations research. Hanover, MD: Institute for Operations Research and the Management Sciences*, pages 182–211, 2009.
- R. L. Breitman and J. M. Lucas. Planets: A modeling system for business planning. *Interfaces*, 17(1):94–106, 1987.
- T. H. Brush, C. A. Marutan, and A. Karnani. The plant location decision in multinational manufacturing firms: An empirical analysis of international business and manufacturing strategy perspectives. *Production and Operations Management*, 8(2):109–132, 1999.
- G. Cachon and P. Feldman. Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management*, 13(2):244–260, 2011.
- C. Canel and B. M. Khumawala. A mixed-integer programming approach for the international facilities location problem. *International Journal of Operations & Production Management*, 16(4):49–68, 1996.
- Y. Chen and T. Huang. Service systems with experience-based anecdotal reasoning customers. *Working Paper*, 2013.
- M. A. Cohen and A. Huchzermeier. Global supply chain management: A survey of research

- and applications. In *Quantitative models for supply chain management*, pages 669–702. Springer, 1999.
- M. A. Cohen and H. L. Lee. Resource deployment analysis of global manufacturing and distribution networks. *Journal of manufacturing and operations management*, 2(2):81–104, 1989.
- M. A. Cohen and S. Mallik. Global supply chains: research and applications. *Production and Operations Management*, 6(3):193–210, 1997.
- M. A. Cohen, M. Fisher, and R. Jaikumar. International manufacturing and distribution networks: A normative model framework. *Managing international manufacturing*, 13: 67–93, 1989.
- S. Dasu and J. de La Torre. Optimizing an international network of partially owned plants under conditions of trade liberalization. *Management Science*, 43(3):313–333, 1997.
- F. de Véricourt and Y.-P. Zhou. Managing response time in a call-routing problem with service failure. *Operations Research*, 53(6):968–981, 2005.
- L. G. Debo and S. Veeraraghavan. Equilibrium in queues under unknown service rates and service value. *Operations Research*, 62(2):38–57, 2014.
- A. Economou and S. Kanta. Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems*, 59(3):237–269, 2008.
- A. Elcan. Optimal customer return rate for an M/M/1 queueing system with retrials. *Probability in the Engineering and Informational Sciences*, 8(4):521–539, 1994.
- G. Falin. A survey of retrial queues. *Queueing Systems*, 7(2):127–167, 1990.
- G. Falin and J. Templeton. *Retrial Queues*, volume 75. CRC Press, 1997.
- D. Farrell. Beyond offshoring: Assess your company’s global potential. *Harvard Business Review*, 82(12):82–90, 2004.
- D. Farrell. Offshoring: Value creation through economic change. *Journal of Management Studies*, 42(3):675–683, 2005.
- W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1971.
- Q. Feng and L. X. Lu. Outsourcing design to Asia: ODM practices. *Managing Supply Chains on the Silk Road: Strategy, Performance, and Risk*, page 169, 2011.
- K. Ferdows. Making the most of foreign factories. *Harvard Business Review*, 75:73–91, 1997.



- N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.
- V. Gaur and Y. Park. Asymmetric consumer learning and inventory competition. *Management Science*, 53(2):227–240, 2007.
- D. Ghelfi. The ‘outsourcing offshore’ conundrum: An intellectual property perspective. *WIPO report (World Intellectual Property Organization)*, 2011.
- M. Goh, J. Lim, and F. Meng. A stochastic model for risk management in global supply chain networks. *European Journal of Operational Research*, 182(1):164–173, 2007.
- J. A. Guajardo, M. A. Cohen, and S. Netessine. Service competition and product quality in the us automobile industry. 2014.
- P. Guo and P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970, 2007.
- P. Guo and P. Zipkin. The effects of the availability of waiting-time information on a balking queue. *European Journal of Operational Research*, 198(1):199–209, 2009.
- P. Guo, W. Sun, and Y. Wang. Equilibrium and optimal strategies to join a queue with partial information on service times. *European Journal of Operational Research*, 214(2):284–297, 2011.
- G. J. Gutierrez and P. Kouvelis. A robustness approach to international sourcing. *Annals of Operations Research*, 59(1):165–193, 1995.
- G. C. Hadjinicola and K. R. Kumar. Modeling manufacturing and marketing options in international operations. *International Journal of Production Economics*, 75(3):287–304, 2002.
- R. Hassin. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica*, 54(5):1185–1195, 1986.
- R. Hassin. Information and uncertainty in a queuing system. *Probability in the Engineering and Informational Sciences*, 21(3):361, 2007.
- R. Hassin and M. Haviv. On optimal and equilibrium retrial rates in a queueing system. *Probability in the Engineering and Informational Sciences*, 10(02):223–227, 1996.
- R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behaviour in Queueing Systems*, volume 59. Kluwer Academic Pub, 2003.
- R. Hassin and R. Roet-Green. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly. *Working paper, Tel Aviv University, Israel*, 2011.

- M. Haviv and R. Randhawa. Pricing in queues without demand information. *Working Paper*, 2012.
- T.-H. Ho, C. S. Tang, and D. R. Bell. Rational shopping behavior and the option value of variable pricing. *Management Science*, 44(12-part-2):S145–S160, 1998.
- J. E. Hodder and M. C. Dincer. A multifactor model for international plant location and financing under uncertainty. *Computers & Operations Research*, 13(5):601–609, 1986.
- J. E. Hodder and J. V. Jucker. Plant location modeling for the multinational firm. In *Proceedings of the Academy of International Business Conference on the Asia-Pacific Dimension of International Business*, pages 248–258. AIB Honolulu, HI, 1982.
- J. E. Hodder and J. V. Jucker. International plant location under price and exchange rate uncertainty. *Engineering Costs and Production Economics*, 9(1):225–229, 1985.
- K. Hoffman and C. Harris. Estimation of a caller retrieval rate for a telephone information system. *European Journal of Operational Research*, 27(2):207–214, 1986.
- V. N. Hsu, J. Hu, and W. Xiao. Global sourcing decisions for a multinational firm with foreign tax credit planning. *Working Paper*, 2014.
- T. Huang, G. Allon, and A. Bassamboo. Bounded rationality in service systems. *Manufacturing & Service Operations Management*, 15(2):263–279, 2013.
- A. Huchzermeier and M. A. Cohen. Valuing operational flexibility under exchange rate risk. *Operations research*, 44(1):100–113, 1996.
- T. Hutzschenreuter, A. Y. Lewin, and S. Dresel. Governance modes for offshoring activities: A comparison of U.S. and German firms. *International Business Review*, 20(3):291–313, 2011.
- N. Jain, K. Girotra, and S. Netessine. Managing global sourcing: Inventory performance. *Management Science*, 60(5):1202–1222, 2013.
- O. Jouini, Z. Akşin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011.
- D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.
- B. Kazaz, M. Dada, and H. Moskowitz. Global production planning under exchange-rate uncertainty. *Management Science*, 51(7):1101–1119, 2005.
- B. Kogut and N. Kulatilaka. Operating flexibility, global manufacturing, and the option value of a multinational network. *Management Science*, 40(1):123–139, 1994.

- V. Kostami and A. Ward. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4):644–656, 2009.
- P. Kouvelis and G. J. Gutierrez. The newsvendor problem in a global market: Optimal centralized and decentralized control policies for a two-market stochastic inventory system. *Management Science*, 43(5):571–585, 1997.
- V. Kulkarni. A game theoretic model for two types of customers competing for service. *Operations Research Letters*, 2(3):119–122, 1983a.
- V. Kulkarni. On queueing systems with retrials. *Journal of Applied Probability*, 20:380–389, 1983b.
- V. Kulkarni and B. D. Choi. Retrial queues with server subject to breakdowns and repairs. *Queueing Systems*, 7(2):191–208, 1990.
- S. Kumar and K. K. Kopitzke. A practitioner’s decision model for the total cost of outsourcing and application to China, Mexico, and the United States. *Journal of Business Logistics*, 29(2):107–139, 2008.
- M. A. Lariviere and J. A. Van Mieghem. Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing & Service Operations Management*, 6(1):23–40, 2004.
- C. Larsen. Investigating sensitivity and the impact of information on pricing decisions in an M/M/1 queueing model. *International Journal of Production Economics*, 56:365–377, 1998.
- A. Y. Lewin, S. Massini, and C. Peeters. Why are companies offshoring innovation? The emerging global race for talent. *Journal of International Business Studies*, 40(6):901–925, 2009.
- Y. Li, Y. Liu, M. Li, and H. Wu. Transformational offshore outsourcing: Empirical evidence from alliances in China. *Journal of Operations Management*, 26(2):257–274, 2008.
- L. X. Lu and J. A. Van Mieghem. Multimarket facility network design with offshoring applications. *Manufacturing & Service Operations Management*, 11(1):90–108, 2009.
- Y. Lu, A. Musalem, M. Olivares, and A. Schilkrut. Measuring the effect of queues on customer purchases. *Management Science*, 59(8):1743–1763, 2013.
- R. Luce. *Individual Choice Behavior*. Wiley, New York, 1959.
- B. L. MacCarthy and W. Atthirawong. Factors affecting location decisions in international operations – a Delphi study. *International Journal of Operations & Production Management*, 23(7):794–818, 2003.

- A. D. MacCormack, L. J. Newman, and D. B. Rosenfield. The new dynamics of global manufacturing site location. *Sloan management review*, 35(4):69–80, 1994.
- A. Mandelbaum and U. Yechiali. Optimal entering rules for a customer with wait option at an M/G/1 queue. *Management Science*, 29(2):174–187, 1983.
- A. Mandelbaum, W. Massey, M. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2-4):149–171, 2002.
- C. C. Markides and N. Berg. Manufacturing offshore is bad business. *Harvard Business Review*, 66(5):113–120, 1988.
- E. Maskin and J. Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.
- S. Massini, N. Perm-Ajchariyawong, and A. Y. Lewin. Role of corporate-wide offshoring strategy on offshoring drivers, risks and performance. *Industry and Innovation*, 17(4):337–371, 2010.
- M. J. Meixell and V. B. Gargeya. Global supply chain design: A literature review and critique. *Transportation Research Part E: Logistics and Transportation Review*, 41(6):531–550, 2005.
- H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38(5):870–883, 1990.
- B. Miller and A. Buckman. Cost allocation and opportunity costs. *Management Science*, 33(5):626–639, 1987.
- P. M. Morse. Stochastic properties of waiting lines. *Operations Research*, 3(3):255–261, 1955.
- A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*, volume 389. Wiley Series in Probability and Statistics, 2002.
- C. L. Munson and M. J. Rosenblatt. The impact of local content rules on global sourcing decisions. *Production and Operations Management*, 6(3):277–290, 1997.
- A. Nagurney, J. Cruz, and D. Matsypura. Dynamics of global supply chain supernetworks. *Mathematical and Computer Modelling*, 37(9):963–983, 2003.
- P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.
- P. Nelson. Advertising as information. *Journal of Political Economy*, 82(4):729–754, 1974.
- A. R. Odoni and E. Roth. An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research*, 31(3):432–455, 1983.

- A. Parlaktürk and S. Kumar. Self-interested routing in queueing networks. *Management Science*, 50(7):949–966, 2004.
- G. Pisano and W. Shih. *Producing prosperity: Why America needs a manufacturing renaissance*. Harvard Business Review Press, 2012.
- E. Plambeck and Q. Wang. Hyperbolic discounting and queue-length information management for unpleasant services that generate future benefits. *Working Paper*, 2012.
- E. Plambeck and Q. Wang. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Science*, 59(8):1927–1946, 2013.
- A. A. Pufall. Ramp-up performance in consumer electronics. *Doctoral Dissertation, Technische Universiteit Eindhoven*, 2013.
- D. F. Pyke. Shanghai or Charlotte? the decision to outsource to china and other low cost countries. *International Series in Operations Research and Management Science*, 98:67, 2007.
- J. P. Quirk and R. Saposnik. Admissibility and measurable utility functions. *The Review of Economic Studies*, pages 140–146, 1962.
- J. Reed and U. Yechiali. Queues in tandem with customer deadlines and retrials. *Queueing Systems*, 73(1):1–34, 2013.
- A. G. Robinson and J. H. Bookbinder. NAFTA supply chains: facilities location and logistics. *International Transactions in Operational Research*, 14(2):179–199, 2007.
- D. B. Rosenfield. Global and variable cost manufacturing systems. *European Journal of Operational Research*, 95(2):325–343, 1996.
- M. Rothschild and J. E. Stiglitz. Increasing risk: I. a definition. *Journal of Economic Theory*, 2(3):225–243, 1970.
- M. Roza, F. A. Van den Bosch, and H. W. Volberda. Offshoring strategy: Motives, functions, locations, and governance modes of small, medium-sized and large firms. *International Business Review*, 20(3):314–323, 2011.
- Y. W. Shin and T. S. Choo. M/M/s queue with impatient customers and retrials. *Applied Mathematical Modelling*, 33(6):2596–2606, 2009.
- D. Simchi-Levi. U.S. re-shoring: A turning point. In *Annual Re-shoring Report, MIT Forum for Supply Chain Innovation*, 2012.
- D. Simchi-Levi, J. P. Peruvankal, N. Mulani, B. Read, and J. Ferreira. Is it time to rethink your manufacturing strategy. *MIT Sloan Management Review*, 23(2):20–22, 2012.

- D. Simchi-Levi, H. Wang, and Y. Wei. Increasing supply chain robustness through process flexibility and inventory. *Working Paper*, 2014.
- H. L. Sirkin, M. Zinser, and D. Hohner. Made in America, again. Why manufacturing will return to the U.S. *The Boston Consulting Group*, 2011.
- H. L. Sirkin, M. Zinser, D. Hohner, and J. Rose. U.S. manufacturing nears the tipping point. *BCG Perspectives by the Boston Consulting Group*, 2012.
- S. C. Srivastava, T. S. Teo, and P. S. Mohapatra. Business-related determinants of offshoring intensity. *Information Resources Management Journal (IRMJ)*, 21(1):44–58, 2008.
- S. Stidham. Optimal control of admission to a queueing system. *Automatic Control, IEEE Transactions on*, 30(8):705–713, 1985.
- X. Su. Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management*, 10(4):566–589, 2008.
- A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(1):207–233, 1973.
- C. J. Vidal and M. Goetschalckx. A global supply chain model with transfer pricing and transportation cost allocation. *European Journal of Operational Research*, 129(1):134–158, 2001.
- W. Whitt. Improving service by informing customers about anticipated delays. *Management science*, 45(2):192–207, 1999.
- W. Whitt. Stochastic models for the design and management of customer contact centers: Some research directions. *Department of IEOR, Columbia University*, 2002.
- R. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.
- X. Wu and F. Zhang. Home or overseas? An analysis of sourcing strategies under competition. *Management Science*, 60(5):1223–1240, 2014.
- S. Xu, L. Gao, and J. Ou. Service performance analysis and improvement for a ticket queue with balking customers. *Management Science*, 53(6):971–990, 2007.
- T. Yang and J. Templeton. A survey on retrial queues. *Queueing Systems*, 2(3):201–233, 1987.
- U. Yechiali. On optimal balking rules and toll charges in the GI/M/1 queueing process. *Operations Research*, 19(2):349–370, 1971.
- U. Yechiali. Customers’ optimal joining rules for the GI/M/s queue. *Management Science*, 18(7):434–443, 1972.