# OPTIMIZING QUEUES WITH SUNK COST FALLACY

LIUTAO YANG

MPhil

The Hong Kong Polytechnic University

2021

The Hong Kong Polytechnic University

Department of Logistics and Maritime Studies

# Optimizing Queues with Sunk Cost Fallacy

Liutao YANG

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Philosophy

May 2021

# CERTIFICATE OF ORIGINALITY

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

(Signed)

LIUTAO YANG (Name of student)

# Acknowledgement

I owe great thanks to my supervisor, Dr. Shining Wu, who gave me unselfish guidance during the past two years of study and the period of writing my thesis. Without his help, I wouldn't have improved my learning skills and maintained the passion for research. I also want to give thanks to my co-supervisor, Prof. Li Jiang, who cared about my study and research and gave useful suggestions in dilemma.

Besides, I sincerely want to thank Department of Logistics and Maritime Studies, the Hong Kong Polytechnic University for her financial support.

Finally, I would like to thank my dear mother and my dear friends for their understanding and support in my life.

# Contents


**1 Introduction** **13**

**2 Literature Review** **17**

**3 Model** **21**

3.1 Joining Decisions of Customers . . . . . 22

3.2 Customer Consumption Decisions and Sunk Cost Fallacy . . . . . 23

**4 Case of Unobservable Queue** **29**

4.1 Equilibrium Outcome . . . . . 31

4.2 Sensitivity Analysis . . . . . 32

**5 Case of Observable Queue** **37**

5.1 Equilibrium Outcome . . . . . 38

5.2 Sensitivity Analysis . . . . . 39

**6 Case with Information Heterogeneity** **43**

6.1 Equilibrium Outcome . . . . . 44

6.2 Sensitivity Analysis . . . . . 45

**7 Comparisons of Unobservable and Observable Cases** **47**

**8 Discussion** **51**

8.1 Determining Optimal Service Rate . . . . . 51

8.2 The Effect of Improving Waiting Experience . . . . . 52

8.3 The Optimal Product Price . . . . . 52

**9 Concluding Remarks** **53**

**References** **55**

# Abstract

We study an M/M/1 queue with customers subject to sunk cost fallacy caused by reference effect and loss aversion. The customers can purchase any quantity (even fractional) when they are in service. Also, the customers are assumed to be rational before joining the queueing system. However, once they join the queue, their purchasing quantity decisions can be affected by actual average waiting time. We fully characterize the purchasing decisions of customers with sunk cost fallacy in equilibrium, and it turns out that the decision is S-shaped in the actual waiting time and is increasing in the actual waiting time. We consider cases where customers may or may not observe the queue length information upon their arrivals. In the case of unobservable queue, we show that in the presence of sunk cost fallacy, the firm's profit is unimodal in service rate and waiting cost, which is different from that in a system without sunk cost fallacy. These facts imply that, the firm may not choose to speed up the service rate even though the it's costless to do so; moreover, the firm may not manage to improve the queueing experience in the unobservable case. We also show that a firm that ignores sunk cost fallacy while this phenomenon indeed exists may underprice or overprice the products, which may lead to a substantial loss. Similar findings are identified in the case of observable queue. Besides, we can conclude from our study that the higher degree of sunk cost fallacy, the higher profit the firm can gain. Last but not the least, disclosing the queue length information may not always benefit the firm.

**Keywords:** Queues; sunk cost fallacy; reference effect; loss aversion; bounded rationality

# Chapter 1

# Introduction

In service queues where customers' final consumption decisions are made after waiting in line (e.g., at a restaurant), empirical evidence shows that customers tend to consume more if they have spent longer time waiting in queue for the service (Ülkü et al., 2020). Such a finding obviously cannot be explained by rationality because otherwise customers' consumption quantity would be independent of how long they have waited, a sunk cost that they have already paid. This phenomenon is referred to as sunk cost fallacy in literature (Ho et al., 2018) and may be caused by mental account set up by customers to evaluate gain and loss differently (Arkes and Blumer, 1985; Thaler, 1999). Such a mental component in customers' utility function induces them to purchase more than the rational optimal level if the sunk cost is high, leading to irrational decisions. Although existing psychology and behavioral studies have looked into the phenomenon under various scenarios, like people tend to throw good money after bad (Kahneman and Tversky, 1979), few studies how sunk cost fallacy affects firm's operational decisions. In the presence of sunk cost fallacy, a firm's optimal operations strategies may be different from those without the fallacy. For example, a firm may intentionally slow down the service rate so as to increase per capita consumption and profit even when speeding up is costless under some situations. Moreover, some widely adopted operational levers that improve experience by distracting customers from waiting may not be as beneficial as expected since easing the cognitive cost of waiting may lead to reduced consumption. In this report, we study the operations of a queue when customers' consumption decisions are subject to sunk cost fallacy. Different from recent empirical OM studies that investigated the topic, we build a reference effect model for the customers' mental utility and adopt a theoretical approach in our

study.

We consider a parsimonious queueing model in which a firm sell its products (goods or services) at a constant unit price by a single server. Customers arrive according to a Poisson process and form a first-in-first-out queue for purchasing. Service times are independent and identically distributed according to an exponential distribution. Customers make two decisions in the system: whether to join the queue and how much to purchase. We assume that the gross value of purchase of a customer is increasing concave in the purchasing quantity and is bounded. We first consider a case of unobservable queue where customers are not informed about the queue length information upon arrival to the queue, and should decide to join or balk with pure or mixed strategy. Then, we consider a case of observable queue where customers are informed about the queue length information upon arrival, and should decide whether to join or balk with a threshold strategy. Finally, we consider a case of heterogeneous information where only some of the customers are informed about the queue length information upon arrival, while the others are not. In the last case, the informed customers still use the threshold strategy, while the uninformed customers use mixed strategy. To capture the impact of sunk cost fallacy, we assume that customers' utilities after having spent time waiting are affected by a reference effect which depends on the difference of the gross value of the purchase and a reference cost. Customers perceive a loss if the gross value is smaller than the reference (e.g., when waiting time is long) and a gain otherwise. According to prospect theory, we assume that customers are more sensitive to loss than to gain.

The results of our study are as follows. We first characterize the optimal joining and purchasing decisions of customers in equilibrium. Furthermore, we show that the realized purchasing quantity of a customer increases with the time he has spent waiting in line and exhibits an "S" shape pattern. Based on the equilibrium outcome, we study the impact of service rate, waiting cost, unit product price, and reference effect on the sales and profit of the firm. We show that the firm's revenue may drop due to lower per capita consumption as the service rate increases although more customers purchase in this case. This finding suggests that in some scenarios a firm may choose a slow service rate even if the cost of speeding up is costless. By comparing the optimal price of our model and that of a benchmark where customers are assumed to be rational, we show that a firm which ignores sunk cost fallacy while

customers indeed are subject to such effect may underprice/overprice the product. Furthermore, the impact of the improving waiting experience (e.g., reducing cognitive waiting cost) on profit is mixed. We discuss the conditions under which this impact is negative. Lastly, we also discuss the impact of customers' loss aversion.

The rest of this report is organized as follows. In Chapter 2, we review the literature. In Chapter 3, we describe the model. In Chapter 4 and 5, we analyze equilibria of the queue in two different scenarios respectively. In Chapter 6, we extend our analysis to a case with heterogeneous customers. In Chapter 7, we compare the cases of unobservable and observable queues to reveal the impact of information disclosure. In Chapter 8, we provide a further discussion of the key results and provide insights into managing queues in the presence of sunk cost fallacy in practice. In Chapter 9, we offer concluding comments. Proofs that are not presented in the main text are included in the Appendix.

# Chapter 2

# Literature Review

There are abundant literatures on the psychology of sunk cost fallacy. Classical ones include Kahneman and Tversky (1979), Thaler (1980), Maister et al. (1984), Thaler (1985), Arkes and Blumer (1985), Garland (1990), Heath (1995), Thaler (1999), Arkes and Ayton (1999), McAfee et al. (2010), Haita-Falah (2017), Puaschunder et al. (2019) and so on. Kahneman and Tversky (1979) finds that people would throw good money after bad, which is the basic finding of sunk cost. And the phenomenon can be explained by prospect theory. Thaler (1980) follows the theory of Kahneman and Tversky, and model the value in two separate parts: the first one is gain $v(g)$, and the second one is loss $\tilde{v}(-c)$. He further makes use of prospect theory to develop a value function $w(z, p, p^*) = v(\bar{p}, -p) + v(-p : -p^*)$ with reference price $p^*$ (Thaler, 1985). In this latter paper, Thaler also use the term "segregation" and "integration" to deal with the variation on gain or loss: An increase in gain should be segregated, an increase in a loss should be integrated, a decrease in a gain should be integrated, and a reduction in the loss should be segregated. Kahneman and Tversky (1984) shows that the preference are different in different decision process. Then Thaler restated that outcomes can be segregated or integrated, and different objects can sit in different mental account, the frequency he evaluates the mental account and "choice bracketing"(Thaler, 1999). There is also an economic model of the reference effect, which is the reason for sunk cost fallacy, separating "consumption utility" and "gain-loss utility"(Kőszegi and Rabin, 2006). We learn from Thaler (1985) and Kőszegi and Rabin (2006) to build a model based on the suffering (waiting cost) per quantity one purchases. The rationale behind is that, human beings form a prospect for the object that can be expected (Thaler, 1985). In queueing system, the object that customers expect is time.

We generalize the model from sunk price cost to sunk time cost, as is said that time "has an economic value to the consumer" and time and efforts can be referred to as "time price" (Baker et al., 2002; Giebelhausen et al., 2011). From Puaschunder et al. (2019), we know that people may form prospect for time, and Ülkü et al. (2020) uses experiments to prove that sunk time cost actually exists. Therefore, we build a model based on the prospect theory to learn how revenue is affected by different subjects (price, service rate, waiting cost, and so on).

Another stream of literature involves sunk cost fallacy in operations management. Baucells and Hwang (2017) establishs a model that captures the psychological mechanism of mental accounting to explain sunk cost effect. Kong et al. (2018) studies the impact of sunk cost effect on the profit of a monopoly firm and finds that customers with greater sunk cost bias and naiveness would lead to better profits. Hong et al. (2019) studied why the sunk cost emerges based on an intrapersonal self management game, and also empirically proves the "overcoming" reason and "resolving" reason. To the best of our knowledge, we are the first to model sunk cost fallacy in queueing system and study its impact on the firm's behaviour.

We also contribute to the literature on behavioral queue. Maister et al. (1984) uses anecdotal reasoning to put up eight behavioral phenomenon in waiting lines: occupied time feels shorter than unoccupied time; people want to get started; anxiety makes waits seem longer; uncertain waits are longer than known, finite waits; unexplained waits are longer than explained waits; unfair waits are longer than equitable waits. Dube-Rioux et al. (1989) empirically proved that delay in the preprocess and post-process phase (long "unoccupied time") would cause consumers evaluate the restaurant negatively. Typical literatures (Naor, 1969; Edelson and Hilderbrand, 1975) naturally assume that people tend to avoid waiting ("get started"), and that psychological anxiety derived from the wait would create disutility in the consumer. Kumar and Krishnamurthy (2008) empirically studied the negative impact of uncertainty on consumers. Other work involving behavioral queue includes, to name a little, Debo et al. (2012), Buell (2021), Giebelhausen et al. (2011), Huang and Chen (2015), Ren et al. (2018), Huang et al. (2013), Li et al. (2016), Kremer and Debo (2016), Lu et al. (2013). These literature all empirically or theoretically proved the individual-level behavioural claims of Maister et al. (1984). Experiments indicate that waiting time can be a signal of quality, and the satisfaction of customers can be increased by making them wait (Giebelhausen et al., 2011). Debo et al. (2012)

design a model with heterogeneously informed customers in which uninformed customers may rely on the queue length to infer quality, and develop the equilibrium strategy for uninformed customers is to join the queue either below or above a "hole" in the queue. They further state that speeding up the service may result in less profit, which is similar to our results. Kremer and Debo (2016) tests the effect of waiting time on the perceived quality when part of the customers are informed about the quality. They find that this effect occurs even with small proportion of informed customer. Surprisingly, the waiting time may also affect the purchasing frequency in a positive way. Buell (2021) experimentally learn that the last-place aversion can lead to switching or abandoning behaviours, and suggested that to hide the queue information for the customer in the last place but disclose the queue when they are not in the last place. Huang et al. (2013) model bounded rationality by using the tool of multinomial logit function, the results indicate that bounded rationality can hurt revenue and welfare in observable queue by using optimal price, but will benefit the firm when the level of bounded rationality is high enough. Li et al. (2016) studied queue in a duopoly competition model with boundedly rational customers using multinomial logit function, they found that profit of firms is unimodal in service rate, where service time is the standard for rewarding. They also studied the relationship of socially optimal price and monopoly optimal price when customers' actual utility is non-negative. Huang and Chen (2015) considers boundedly rational customers who make decisions by experience or anecdotal reasoning, and find that boundedly rational customers are less price-sensitive, and firms may intentionally lower the price when they can make service rate decisions. Ren et al. (2018) further extend Huang and Chen (2015) to a circumstance where quality is unknown and has to be inferred from the anecdotes. Results show that, profit is U-shaped in the level of bounded rationality, which is measured by the size of anecdotes they take. Interestingly, they find that the less customers are boundedly rational, the less their surplus will be. Moreover, if customers are less boundedly rational, the firm may reduce both the quality and price. They also emphasize the importance of disclosure of information in different circumstances. However, As Allon et al. (2018) said, "there are to date too few studies (empirical or theoretical) that even attempt to assess whether individual customers' behavioural tendencies affect system behaviour in a meaningful way". We attempt to make up this nearly blank field. In our report, we try to build up a novel model to study how the individual-level behaviours affect the system-level

behaviours.

We are mostly close to the literature on bounded rationality in queues. Up to now, theorists have developed five categories of bounded rationality: Logit choice model, anecdotal reasoning, cognitive hierarchy, hyperbolic discounting, and reference dependence and loss aversion (Ren and Huang, 2018). Simon (1972) developed the theory of bounded rationality. Debo et al. (2012) incorporates information availability to show that uninformed customers' joining strategy is balking only when they see a "hole". Huang et al. (2013) use the multinomial logit function to model the bounded rationality in queueing system. Different from these models, in this report, our model captures the sunk cost fallacy (reference dependence and loss aversion). Relative literature includes Yang et al. (2018), Ho et al. (2018) and Ülkü et al. (2020). Yang et al. (2018) made use of the model of Kőszegi and Rabin (2006) to develop the joining strategy and purchasing strategy. They studied the optimal pricing strategy of the firm, and the optimal price that optimizes social welfare. They found that the optimal price maximizing the profit and the optimal price maximizing social welfare do not coincide in the monoply's setting. Ho et al. (2018) develop a model of durable goods by considering mental accounting for sunk cost. Ülkü et al. (2020) empirically proves that the sunk time cost would affect the behavior of customers. We follow the fact that customer waiting in line longer tends to consumer more (Ülkü et al., 2020), which is consistent with the theory developed by Kahneman and Tversky, that people tend to throw good money after bad one (Kahneman and Tversky, 1979). We want to propose a model which includes both the rationality and irrationality parts of the decision process. Ho et al. (2018) models the irrational behavior of customers by adding a sunk cost term $M(S, Q_t)$ to the model: $M(S, Q_t) = \lambda_1 + \lambda_2 Q_t + \lambda_3 S + \lambda_4 S \cdot Q_t$, where $S$ is the sunk cost, $Q_t$ is the cumulative usage of the durable good. Ülkü et al. (2020) develops an easier model where utility function is $U(q) = B(q) - pq - cw + \lambda wq$, which indicates that longer waits will lead to more purchasing quantity. In the appendix of Ülkü et al. (2020), they stated another method to model the sunk cost fallacy: $U(q) = f(gq) + f(-pq - cw)$, where $g$ is the benefit for each product of the good, and $f(\cdot)$ is an S-shaped utility function representing loss aversion. Different from that of Ho et al. (2018) and Ülkü et al. (2020), we built up a novel model, whose irrationality part is captured by reference effect, represented by the difference between expectation and reality.

# Chapter 3

# Model

We consider a single-server service system for the sale of a product (which can be a bundle of both goods and services). Potential customers arrive according to a Poisson process with rate $\Lambda$. Upon arrival, customers decide whether to join the queue or not based on the expected pay ff of making such decisions. Customers who decide to join form a first-in-first-out (FIFO) queue for their services. Customers who balk the system leave and do not return. After waiting in queue, a customer decides on the consumption quantity (denoted by $q$) of the product when it is his turn for service. However, different from the rational behaviour when he joins, customers' consumption quantity decisions are influenced by a cognitive bias as is referred to as *sunk cost fallacy* (Ülkü et al., 2020). This fallacy captures the difference between objective utility and perceived utility. Let $B(q) \geq 0$ denote the benefit a customer obtain from purchasing $q$ unit(s) of the product, $p$ denote the unit price charged by the firm, and $c$ denote the waiting cost of a customer for per unit time waiting in the system (i.e., including the servivce time). We treassue that the purchasiny ais a continuous variable ad assume that the benefit function $B(q)$ is increasing concave in $q$. Note that in this study we consider a situation where the purchasing (consumption) quantity decisions are made after customers having spent time waiting and that the service time of a customer is independent of his purchasing quantity. We assume that the service times are identical and independently distributed according to an exponential distribution with mean $\frac{1}{\mu}$, which is common knowledge.

Briefly speaking, each customer makes one or two decisions in the system: the join-or-balk decision upon arrival and, if he decides to join, the purchasing quantity when it is his turn for service after waiting. More details about the decision process of customers are introduced as

follows.

## 3.1 Joining Decisions of Customers

We assume that customers are rational upon arrival, that is, they decide to join or not based on their expected payoff of making a joining decision, which is equal to the purchasing benefit less the purchasing cost and the expected waiting cost. Specifically, a customer's net utility is given by

$$U_0(q) := B(q) - pq - cw, \tag{3.1}$$

if he purchases $q$ unit(s) and waits for $w$ in queue. In this case, it is optimal for every rational customer to purchase $\hat{q} := \arg\max_q\{B(q) - pq\}$ since the purchasing benefit and the unit price are independent of the waiting cost. To avoid a trivial outcome, we assume that $B(\hat{q}) > p\hat{q}$ because otherwise it is optimal for all customers to make no purchase. Then, $B(\hat{q}) - p\hat{q}$ is the maximum reward a customer can receive by joining the system. An arriving customer would join the queue is $B(\hat{q}) - p\hat{q} \geq c\bar{w}$ and not otherwise, where $\bar{w}$ is the expected time he believes he has to wait if he joins. Note that this expectation depends on the system state upon his arrival and how much he knows about this state.

In classic and recent literature (e.g., Edelson and Hilderbrand (1975); Naor (1969); Hu et al. (2018)), researchers have considered difference cases in which customers are revealed with different system information (e.g., all or part of customers are or are not informed with the information) and have fully characterized the equilibrium joining decisions of customers in these cases. In this project, we adopt the terms used by Hu et al. (2018) and refer to customers who can observe the system queue length upon their arrival as *informed customers* and to those who cannot as *uninformed customers*. Specifically, we consider the following cases.

Case 1. All customers are uninformed. The system in this case has an *unobservable queue* to all customers and is studied in Chapter 4.

Case 2. All customers are informed. The system in this case has an *observable queue* to all customers and is studied in Chapter 5.

Case 3. Some customers are uninformed and the others are informed. That is, there is information

heterogeneity among customers. This case is studied in Chapter 6.

With different system information learned, informed and uninformed customers differ in the way they estimate their expected waiting times, $\bar{w}$. Since uninformed customers do not observe the queue length, they can only estimate their expected waiting time by the multiplication of the average service time per customer (i.e., $\frac{1}{\mu}$, which is common knowledge in our model) and the average system queue length over time (which can be inferred by rational expectation of the system steady state). Noting that all uninformed customers learn the same information about the system, they form the same estimation. On the other hand, informed customers, learning the actual queue length upon their arrival, can form more accurate estimation of their expected waiting time, which is equal to the average service time multiplied by the the actual queue length (including himself if he joins). Each informed customer forms customized estimation of the expected waiting time depending on his own observation. Once having formed his estimation of the expected waiting time, $\bar{w}$, a customer, whether uninformed or informed, use the same rule to decide whether to join or balk: join if the expected payoff $B(q) - pq - c\bar{w}$ is non-negative. More details of the optimal joining strategies of customers and the outcome will be introduced in Chapters 4–6 for each case, respectively.

## 3.2 Customer Consumption Decisions and Sunk Cost Fallacy

**Transaction Utility and Reference Effects** There are numerous studies suggesting that customers are not fully rational when making decisions. Thaler (1985) proposed that consumers get two kinds of utility from a purchase: *acquisition utility* and *transaction utility*. While the former measures the value of the good relative to its price and serves as the basis for rational decision making, the latter measures the perceived value of the 'deal', which depends on the difference between the amount paid and the "reference price/cost" for the good (Thaler, 1999). We adopt this framework and model a reference-dependent component in customer utility function. We assume that a customer forms a reference, denoted by $r$, for the waiting cost he expects to pay for each unit of product when making his decision to join the queue. Specifically, when an arriving customer expects to wait for a $\bar{w}$ amount of time in the system and still decides to join the queue, he forms a reference waiting cost $r = \frac{c\bar{w}}{\hat{q}}$ for per unit purchase, where $\hat{q}$ is

the ex-ante optimal purchasing quantity decision of every rational customer. After he joins the queue and experiences waiting, his decision making process (for the purchasing quantity decision) is subject to the influence of cognitive transaction utility. For example, if the customer has spent $w$ time waiting and decides to purchase $q$ units of products, his actual waiting cost per unit consumption is $\frac{cW}{q}$, implying a perceived gain if $\frac{cW}{q} < \frac{c\bar{w}}{\hat{q}}$ and a perceived loss if $\frac{cW}{q} > \frac{c\bar{w}}{\hat{q}}$. We assume that the transaction utility is a function of the perceived gain/loss and is given by $R(\frac{c\bar{w}}{\hat{q}} - \frac{cw}{q})$, where

$$R(x) := \begin{cases} \theta^+ x & \text{if } x \geq 0, \\ \theta^- x & \text{if } x < 0. \end{cases} \tag{3.2}$$

is the reference effect function and $0 \leq \theta^+ \leq \theta^-$. Noting that $\theta^+$ ($\theta^-$) captures the customers' sensitivity to gain (loss), customers are risk neutral if $\theta^+ = \theta^-$ and loss averse if $\theta^+ < \theta^-$. In general, customer loss aversion is supported by the prospect theory developed by Kahneman and Tversky (1979) and can be reflected by the concavity of the reference effect function in this case.

After incorporating the reference-dependent transaction utility, a customer's utility function becomes

$$U(q) = B(q) - pq - cw + R(\frac{c\bar{w}}{\hat{q}} - \frac{cw}{q}) \tag{3.3}$$

when he has spent time ($w$) waiting and needs to make their final purchasing quantity decision. The customer's problem at this point is to choose a quantity $q$ to maximize his net utility $U(q)$. By solving this problem and characterizing its solution, we find that the incorporation of transaction utility gives rise to sunk cost fallacy in customers' decisions on the purchasing quantity. We establish the results by the following proposition.

**Proposition 1** (Characterization of the optimal consumption quantity)**.** *Suppose that an arriving customer decides to join the queue and expects to wait for a total time $\bar{w}$ in the system, and hence forms a reference waiting cost $\frac{c\bar{w}}{\hat{q}}$ for per unit product purchased. Then, his optimal purchasing quantity that optimizes* (3.3) *after he has experienced actual waiting time equaling to $w$ is given by*

$$\mathscr{q}(w, \bar{w}) := \begin{cases} \phi(w, \theta^+), & \text{if } w \leq \Omega(\bar{w}, \theta^+), \\ \frac{\hat{q}}{\bar{w}} w, & \text{if } \Omega(\bar{w}, \theta^+) < w \leq \Omega(\bar{w}, \theta^-), \\ \phi(w, \theta^-), & \text{otherwise.} \end{cases} \tag{3.4}$$

*where* $\phi(w,\theta) := \arg\max_q\{B(q) - pq - \frac{\theta c w}{q}\}$ *and* $\Omega(\bar{w},\theta)$ *is the unique fixed point of* $\frac{\bar{w}}{\hat{q}}\phi(w,\theta)$ *with regard to* $w$*, i.e.,* $\Omega(\bar{w},\theta) = \frac{\bar{w}}{\hat{q}}\phi(\Omega(\bar{w},\theta),\theta)$.

*The optimal quantity* $q(w,\bar{w})$ *is non-decreasing in the actual waiting time* $w$*, the waiting cost* $c$*, and the parameters* $\theta^+$ *and* $\theta^-$*, and is non-increasing in* $p$ *and* $\bar{w}$*. Furthermore, if* $\theta^- > 0$ *and* $\theta^+ > 0$*, the optimal quantity strictly increases in* $w$ *and* $c$ *and strictly decreases in* $p$.

Proposition 1 shows that in the presence of sunk cost fallacy caused by the inclusion of transaction utility in decision processes, customers purchase more as they have spent more time waiting in the system. Figure 3.1 illustrates the optimal purchasing quantity as a function of the realized waiting time. As one can see, the function exhibits an "S-shape". If the realized waiting time is short (less than $\Omega(\bar{w},\theta^+)$), a customer chooses an optimal purchasing quantity such that the per-unit waiting cost is lower than his reference formed upon arrival and perceives a gain with regard to the per-unit waiting cost paid for the purchase. If the realized waiting time is medium (more than $\Omega(\bar{w},\theta^+)$ but less than $\Omega(\bar{w},\theta^-)$), a customer chooses an optimal purchasing quantity such that he pays the same per-unit waiting cost as the reference value, perceiving no gain or loss with regard to the per-unit waiting cost paid for the purchase. If the realized waiting time is very long (more than $\Omega(\bar{w},\theta^-)$), a customer, although optimally choosing a purchasing quantity more than he originally plans to amortize the waiting cost, will have to perceive a loss with regard to the per-unit waiting cost paid for the purchase because further increasing the purchasing quantity is suboptimal.

Customers subject to sunk cost fallacy tend to consume more when the actual waiting time becomes longer, in which way they "make up" the loss of time they've spent in the queue. The parameter of waiting cost $c$ is plugged in the reference part of the customer's utility function, thereafter influencing the optimal consumption quantity. Also, the more sensitive are the customers to gain/loss, the more could their optimal consumption decisions be affected. Explicitly from the Proposition 1, the optimal purchasing quantity varies with actual waiting time $w$.

In order to evaluate the impact of sunk cost fallacy, it is useful to consider the benchmark case of a system where customers are not subject to sunk cost fallacy and are fully rational. That is, customers make their purchasing decisions only based on their acquisition utility and

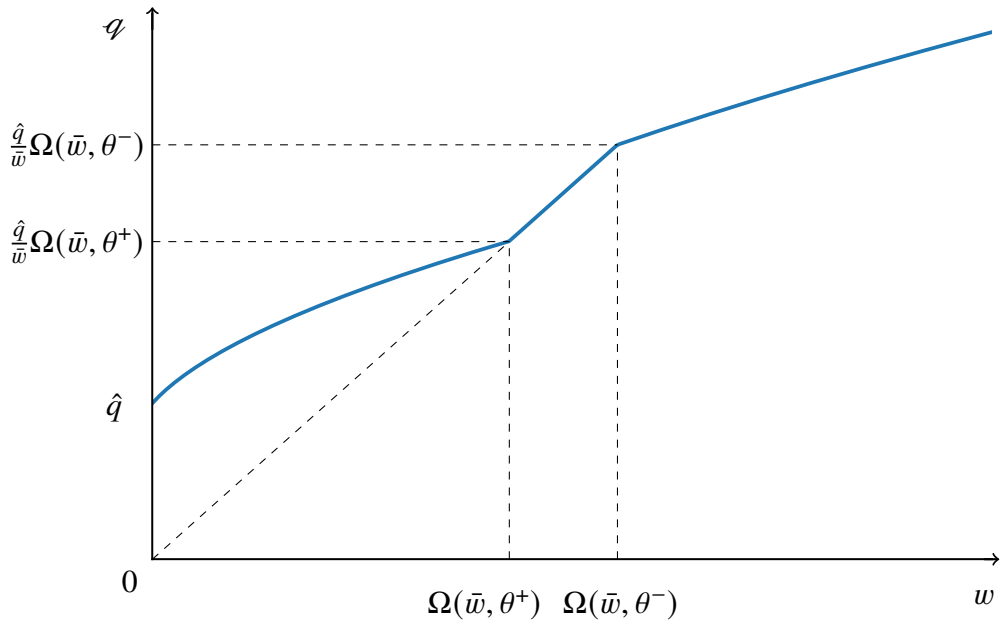



Figure 3.1: The optimal consumption quantity as a function of the actual waiting time

ignore the "perceived" transaction utility. Precisely, consider the following benchmark case.

Benchmark Case (customers are rational): Every customer is fully rational and makes all decisions by maximizing his acquisition utility (given by (3.1)). In this case, sunk cost fallacy does not exist and each customer who joins the queue purchases $\hat{q}$ units of the product.

**Corollary 1.** $q(w,\bar{w}) \geq \hat{q}$.

This corollary verifies that customers who are subject to sunk cost fallacy in their decision processes always purchase no less than rational customers. From this, we can anticipate that a firm which ignores customers' sunk cost fallacy may unexpectedly make suboptimal decisions and incur a substantial loss. In the rest of the report, we will make this point concrete.

**Corollary 2.** *If* $\theta^+ = 0$, *then* $q(w,\bar{w}) = \hat{q}$ *for* $w \leq \Omega(\bar{w},0) = \bar{w}$.

A zero $\theta^+$ corresponds to the case where customers do not perceive a gain if their realized waiting time is less than they expected. This corollary suggests that in this case, a customer eventually purchases the amount he initially intended to purchase if his realized waiting time is no more than his reference (i.e., his ex-ante expected waiting time).

Lastly, we further show how customers' optimal purchasing quantity changes with the product price, the waiting cost, the waiting time, and the coefficient of reference effect.

**Proposition 2.** *If $\theta^+ = \theta^- = \theta$ and $B'''(\cdot) > 0$, then the optimal consumption quantity $q$ is decreasing convex in price $p$.*

**Proposition 3.** *In the case of $\theta^+ = \theta^- = \theta$, the optimal consumption quantity $q$ is increasing concave in the waiting cost $c$, the waiting time $w$, and the coefficient of reference effect $\theta$ if $B'''(q) \leq -\frac{4B''(q)}{q}$.*

Given that $B(q)$ is increasing concave and $\lim_{q\to\infty} B'(q) = 0$, we must have $\lim_{q\to\infty} B''(q) = 0$. That is, $B''(q)$ approaches 0 from below as $q$ increases. The above two propositions just require conditions that are just a little bit stronger: $B''(q)$ monotonically approaches 0 at a rate that is bounded by $-\frac{4B''(q)}{q}$. Note that a similar assumption is also made in Ülkü et al. (2020). Under these conditions, Propositions 2 and 3 show that the optimal consumption quantity has diminishing sensitivity to the price, the waiting cost, the waiting time, and the coefficient of reference effect.

Note that the results in this section are valid for both the observable and unobservable cases. However, customers in these two cases differ in the way they form the reference waiting time $\bar{w}$ because the information available to them are different. In the following two chapters, we would explicitly give the expressions of $\bar{w}$.

# Chapter 4

# Case of Unobservable Queue

In this chapter, we consider the case where all customers are uninformed and hence cannot observe the queue length upon their arrivals. Edelson and Hilderbrand (1975) first study the unobservable queue. They derive the expression of expected waiting time when "the service discipline is strong and work-conserving". Hassin and Haviv (2003) comprehensively summarize the results of unobservable queue and study the social welfare optimization and profit maximization problems. In our problem, the joining decisions are made in the same way as in their models. We further consider the consumption quantity decisions made after customers having spent time waiting in queue. Customers' optimal strategies are characterized as follows.

**Customer Joining Strategy.** Uninformed customers cannot observe the queue length when they arrive and thus make joining decisions based on the average waiting time, which is equal to the average service time per customer multiplied by the expected queue length. Given every uninformed customer is provided with the same information, we discuss the optimal symmetric joining strategy.

By rationality, customers will definitely join the queue if they can receive a positive expected utility by doing so. As each customer maximizes his own payoff in making their decisions, an equilibrium is reached if either (1) all customers join and each can still obtained a non-negative expected utility, or (2) part of the arriving customers join and customers are indifferent between joining and balking (i.e., the expected utility of joining equaling that of balking, and both being zero), or (3) nobody joins since the expected utility is negative even when customers don't need to wait. In equilibrium, customers may adopt either a pure strategy (i.e., join for

sure or balk for sure) or a mixed strategy (i.e., join with a certain probability) for making the joining decisions. The optimal strategy can be characterized by a joining probability, denoted by $\alpha_e$, of each customer, where $\alpha_e = 0$ represents an always-balk strategy, $\alpha_e = 1$ represents an always-join strategy, and $\alpha_e \in (0.1)$ represents a mixed strategy. As summarized by Hassin and Haviv (2003), the effective arrival rate $\lambda_e$ that leads to an equilibrium and the corresponding joining probability of customers (denoted by $\alpha_e := \frac{\lambda_e}{\Lambda}$ can be characterized by the following proposition.

**Proposition 4** (Unobservable Queue: Pure or Mixed Joining Strategy)**.** *In the case of unobservable queue, an arriving customer joins the queue with probability $\alpha_e = \frac{\lambda_e}{\Lambda}$, in which $\lambda_e$ is the effective arrival rate and is given by*

$$\lambda_e = \begin{cases} \Lambda & \text{if } \Lambda \le \mu - \frac{c}{B(\hat{q}) - p\hat{q}}, \\ \mu - \frac{c}{B(\hat{q}) - p\hat{q}} & \text{if } 0 < \mu - \frac{c}{B(\hat{q}) - p\hat{q}} < \Lambda, \\ 0 & \text{otherwise}. \end{cases} \tag{4.1}$$

This proposition recasts the results of §3.1.1 of Hassin and Haviv (2003) (page 46) using our notation. We omit the proof. By (4.1), the effective arrival rate deceases in waiting cost and the unit price, and increases in the service rate, which is intuitive.

**Reference Formation.** Noting that the queue length is unobservable and all uninformed customers have the same prior information before they join, they form the same estimation of the expected waiting time in the system, which is $\bar{w} = \frac{1}{\mu - \lambda_e}$ by classic $M/M/1$ result. Recall that each customer who decides to join the queue forms a reference of the waiting cost he expects to pay for per unit product. Thus, an uninformed customer forms a reference $r = \frac{c\bar{w}}{\hat{q}}$, which is the ratio of his expected waiting cost to the quantity he intends to purchase when he makes the joining decision. Note that all uninformed customers form the same reference.

**Actual Consumption.** Since customers form the same reference ex-ante, they adopt the same optimal consumption strategies, with the optimal consumption quantity characterized by $q(w, \bar{w})$ in Proposition 1 where $\bar{w} = \frac{1}{\mu - \lambda_e}$. However, customers' actual consumption quantities still vary since their realized waiting times $w$'s are contingent, depending on the actual number of customers ahead and the realization of the service times of those ahead. If there are already $n - 1$ customers in queue when a customer joins, then this customer's actual waiting time in the

system follows an Erlang distribution with shape parameter $n$ and rate parameter $\mu$, which can range in the support $[0, +\infty)$ and has a mean $\frac{n}{\mu}$ and variance $\frac{n}{\mu^2}$.

Noting that in the unobservable queue, the number of customers in queue can be very large (although with very small probability), it is possible for uninformed customers having to wait for an extremely long time in system, which is different from an observable queue (to be introduced in next Chapter) where the maximum queue length is bounded. Therefore, as you can expect, the purchasing quantities in the unobservable queue exhibit larger variabilities compared with that in the observable queue.

## 4.1 Equilibrium Outcome

In this section, we characterize the equilibrium outcome. Let $\rho_e := \frac{\lambda_e}{\mu}$ denote the effective system load and $X$ denote the number of customers already in queue when a new customer arrives. Then, $\mathbb{P}(X = i) = P_i^U$, where $P_i^U = \rho_e^i(1 - \rho_e), i = 0, 1, 2, \cdots$ is the stationary distribution of the underlying Markov chain of the unobservable $M/M/1$ queue with effective arrival rate $\lambda_e$ and service rate $\mu$. Let $\Pi^U$ denote the firm's profit per unit of time.

The firm's profit rate $\Pi^U$ is given by

$$\Pi^U = \lambda_e \cdot \sum_{i=0}^{+\infty} P_i^U \mathbb{E}_w[q(w, \frac{1}{\mu - \lambda_e})|X = i]p,$$

in which the waiting time $w$ follows an Erlang distribution with shape parameter $i + 1$ and rate parameter $\mu$ when $X = i$.

We can rewrite the profit function in the following form,

$$\Pi^U = \Lambda\alpha_e \cdot \sum_{i=0}^{+\infty} \rho_e^i(1 - \rho_e) \cdot \int_0^{+\infty} q(w, \frac{1}{\mu - \lambda_e}) f(w; i + 1, \mu)dw \cdot p \tag{4.2}$$

where $f(x; k, \lambda)$ is the density function of Erlang distribution. We also denote $\sum_{i=0}^{+\infty} \rho_e^i(1 - \rho_e) \cdot \int_0^{+\infty} q(w, \frac{1}{\mu-\lambda_e}) f(w; i + 1, \mu)dw$, i.e., the average purchasing quantity of those who join, by notation $\bar{q}$. Then, the firm's profit rate in the case of unobservable queue can be simply expressed as

$$\Pi^U = \Lambda\alpha_e \cdot \bar{q} \cdot p, \tag{4.3}$$

where $\alpha_e$, given by Proposition 4 is non-increasing in $p$, $c$, and $\frac{1}{\mu}$, and $\bar{q}$ is non-increasing in $p$, $c$, $\frac{1}{\theta^+}$, and $\frac{1}{\theta^-}$.

## 4.2 Sensitivity Analysis

In this section, we study how the firm's profit changes with different parameters of the model.

**Proposition 5.** *The firm's profit is increasing in $\theta^+$ and $\theta^-$.*

This proposition suggests that ceteris paribus, the firm's profit increases in the degree of sunk cost fallacy. That is to say, customers' sunk cost fallacy, if appropriately manipulated by the firm, can help improve firm's profitability.

To further discuss the impact of other parameters, we first derive properties of the distribution of customer waiting times and the expected purchasing quantity.

**Lemma 1.** *Let $N_p$ be geometric random variables on $\{1, 2, 3, \cdots\}$ with parameter $p$ and $T_1$, $T_2$, ... be independent and identically distributed exponential random variables. Then, $p\sum_{i=1}^{N_p} T_i$ and $T_1$ are identically distributed, i.e., $p\sum_{i=1}^{N_p} T_i$ follows the same exponential distribution as $T_i$'s.*

Noting that the steady-state distribution of an $M/M/1$ queue is geometric, we know that the sojourn times of all customers who join the system follow an exponential distribution. By this result, we have the following proposition with regard to the the expected optimal purchasing quantity of customers who join the queue.

**Proposition 6.** *Under the condition that $0 < \lambda_e = \mu - \frac{c}{B(\hat{q})-p\hat{q}} < \Lambda$, the waiting cost ($cw$) of an arriving customer follows an exponential distribution with mean $B(\hat{q}) - p\hat{q}$, which is independent of $c$ and $\mu$. This, by the Poisson-Arrivals-See-Time-Averages (PASTA) property, implies that the distribution of the purchasing quantities, $q(w, \bar{w})$, of the customers who decide to join and its average, $\mathbb{E}_w[q(w, \bar{w})]$, is independent of $c$ and $\mu$ if $0 < \lambda_e = \mu - \frac{c}{B(\hat{q})-p\hat{q}} < \Lambda$, i.e., $\alpha_e \in (0, 1)$.*

The intuition behind this proposition is as follows. As the unit waiting cost (the service rate) increases, fewer (more) customers join. When some but not all customers join (i.e., $0 < \alpha_e < 1$) in equilibrium, the effective arrival rate decreases (increases) with $c$ ($\mu$) in such a way that the average waiting cost, $c\bar{w}$, is equal to the reward of joining, $B(\hat{q}) - p\hat{q}$, which is independent of $c$ and $\mu$. By Lemma 1, not only the mean but also the whole distribution of the waiting cost is unchanged as $c$ or/and $\mu$ increase. Noting by (3.3) and (3.4) that the optimal ex-post purchasing quantity depends on $cw$ as a whole, we can see that the distribution of the purchasing quantities of those who join and thus its mean are unaffected by the waiting cost per unit time and the service rate if $0 < \alpha_e < 1$.

Based on the above result, we first study the impact of the service rate $\mu$ on the firm's profit. Increasing $\mu$ corresponds to operational levers that increase service capacity in practice, such as hiring more staff, training to improve productivity of each server, increasing service speed by automation, etc. The following proposition shows that the effect of these levers can be mixed even if applying them is costless.

**Proposition 7.** *The firm's profit is unimodal in the service rate $\mu$. Specifically, the profit is increasing in $\mu$ for $\mu < \Lambda + \frac{c}{B(\hat{q})-p\hat{q}}$ (i.e., when $\alpha_e < 1$) and decreasing in $\mu$ for $\mu > \Lambda + \frac{c}{B(\hat{q})-p\hat{q}}$ (i.e., when $\alpha_e = 1$).*

The effect of increasing the service rate is double-edged since it has different influences on the components that determine the firm's profit (4.3). On the one hand, customers expect a shorter queue and hence more are willing to join (i.e., a larger effective arrival rate) if the service speed increases. On the other hand, customers are less subject to sunk cost fallacy if they experience shorter waiting times, resulting in a lower average purchasing quantity per customer. We find that the former (positive) effect, if exists, dominates the latter (negative one). That is, the benefit of enlarging the customer numbers offsets the loss caused by a lower per capita consumption. Thus, when $\alpha_e < 1$ and not all potential customers join, improving service speed helps the firm increase profit since by doing so it can attract more customers to join. However, when $\alpha_e = 1$ and all potential customers have already decided to join, further improving service speed does not help attracting more customers but only makes customers purchase less on average, which leads to a lower profit rate. Therefore, the firm's profit first

increases and then decreases in the service rate $\mu$. This finding suggests that in the presence of customer sunk cost fallacy, it may not be optimal for the firm to improve service speed even when it is costless to do so. With a service rate dependent capacity cost in practice, some firms may even choose a lower service rate not to serve all customers but to induce a higher per capita consumption.

Next, we study the impact of the unit-time waiting cost, $c$, on the firm's profit. Decreasing $c$ corresponds to managerial levers that make customer waiting experience less suffering in practice, such as engaging customers in pleasurable activities (e.g., Haidilao, a Chinese hotpot restaurant, provides free hand massages to waiting customers), distracting customers' attention away from waiting (e.g., broadcasting TV shows, providing free WiFi), improving waiting environment (e.g., providing seats), etc. The following proposition shows that the effect of these levers can be mixed even if applying them is costless.

**Proposition 8.** *The profit is unimodal in the waiting cost c. Specifically, the profit is increasing in c for* $0 \leq c \leq (B(\hat{q}) - p\hat{q})(\mu - \Lambda)^+$ *(i.e., when* $\alpha_e = 1$*) and decreasing in c for* $c > (B(\hat{q}) - p\hat{q})(\mu - \Lambda)^+$ *(i.e., when* $\alpha_e < 1$*).*

The intuition behind this proposition is similar to that of Proposition 7. There are both positive and negative effects related to improving service experience (i.e., decreasing $c$). On the one hand, more customers are willing to join (i.e., a larger effective arrival rate) if the waiting is less suffering (i.e., $c$ decreases). On the other hand, customers are less subject to sunk cost fallacy and purchase less on average if they perceive the waiting cost less. We find that the benefit of enlarging the customer numbers, if exists, offsets the loss caused by a lower per capita consumption. Thus, when $\alpha_e < 1$ and not all potential customers join, improving waiting experience helps the firm increase profit since by doing so it can attract more customers to join. However, when $\alpha_e = 1$ and all potential customers have already decided to join, further improving waiting experience does not help attracting more customers but only makes customers purchase less on average, which leads to a lower profit rate. Therefore, the firm's profit first increases and then decreases in the unit-time waiting cost $c$. This finding suggests that in the presence of customer sunk cost fallacy, it may not be optimal for the firm to improve waiting experience even when it is effortless to do so. It would be even less profitable to do so

if improving waiting experience costs a lot of resource.

When the total arrival rate is larger than the service rate, i.e., $\mu - \Lambda < 0$, the firm cannot serve all customers no matter how low the waiting cost $c$ is. Therefore, reducing $c$ can always help attract more customers to join and the firm's profit monotonically decreases in $c$. When the total arrival rate is smaller than the service rate, i.e., $\mu - \Lambda > 0$, there exist a threshold $(B(\hat{q}) - p\hat{q})(\mu - \Lambda)$ such that all customers join if the waiting cost $c$ is smaller than this threshold. Therefore, the firm's profit first increases and then decreases in $c$, attaining its maximum at $c = (B(\hat{q}) - p\hat{q})(\mu - \Lambda)$.

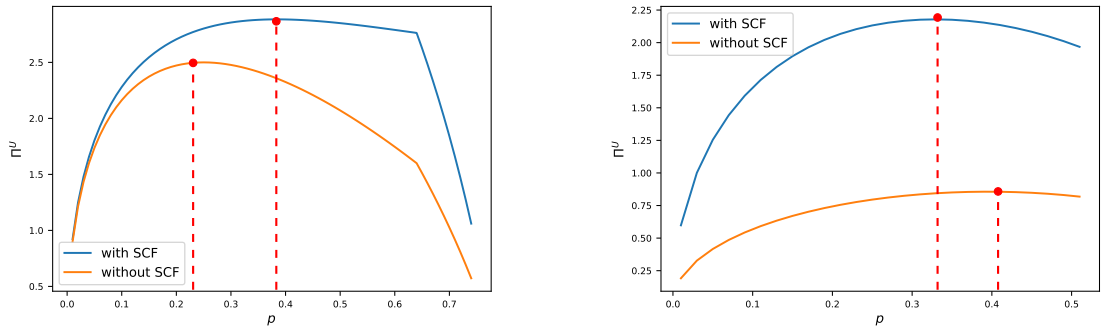
**Observation 1.** *The optimal price for the case with sunk cost fallacy may be higher/lower than the optimal price for the case without sunk cost fallacy.*

Firm's profit is unimodal in the unit price and there exists an optimal price that maximizes the profit. We compare the optimal prices when sunk cost fallacy is present or absent. One obvious thing is that, for each given price, the profit of the case with sunk cost fallacy is greater than the case without sunk cost fallacy. This is intuitive, since the average demand becomes higher when we incorporate sunk cost effect, and the effective arrival rate are the same for both cases. In the following figures, "SCF" stands for "sunk cost fallacy", we show that if $B(q) = \frac{q}{1+q}$, then when $\theta^+ = 0.88, \theta^- = 1, c = 0.2, \mu = 15, \Lambda = 10$, the firm may neglect sunk cost fallacy and mistakenly set a low price, which may lead to less profit; if $B(q) = \arctan q$, then when $\theta^+ = 8.8, \theta^- = 10, c = 0.1, \mu = 2, \Lambda = 5$, the firm may mistakenly set a high price, which may lead to suboptimal profit.

For some forms of the benefit function $B(q)$ and under some conditions, the optimal price of the case with sunk cost fallacy is smaller than the optimal price of the case without sunk cost fallacy. In this case, although the marginal profit per unit of product decreases as the firm charges a lower $p$, the effective arrival rate and average consumption quantities become higher, and the overall profit becomes higher than charging a price that maximizes profit assuming no sunk cost fallacy. This implies that the firm may intentionally lower the price to "trick" customers to join, and once they join the queue, their consumption quantity is affected by the sunk cost fallacy. This observation implies that, the firm may intentionally lower the price to attract more customers compared with the optimal price for the case without sunk cost fallacy.

And the effect of sunk cost fallacy dominates the loss of gross profit.

On the other hand, tt is also possible that the optimal price of the case with sunk cost fallacy is higher than the optimal price of the case without sunk cost fallacy. From these examples, one can see that a firm that ignores the impact of sunk cost fallacy in customer consumption decisions while this phenomenon indeed exists may suffer a substantial loss due to overpricing or underpricing.




(a) $B(q) = \frac{q}{1+q}$, $\theta^+ = 0.88$, $\theta^- = 1$, $c = 0.2$, $\mu = 15$, $\Lambda = 10$

(b) $B(q) = \arctan(q)$, $\theta^+ = 8.8$, $\theta^- = 10$, $c = 0.1$, $\mu = 2$, $\Lambda = 5$

Figure 4.1: Firm's profit as a function of price when sunk cost fallacy is or is not present

# Chapter 5

# Case of Observable Queue

In this chapter, we consider the case where all customers are informed and hence can observe the queue length upon their arrivals. Naor (1969) first study the observable queue and characterize the threshold type joining strategy of customers in equilibrium. Hassin and Haviv (2003) comprehensively summarize the results and study the social welfare optimization and profit maximization problems. In our problem, the joining decisions are made in the same way as in their models. We further consider the consumption quantity decisions made after customers having spent time waiting in queue. We first characterized customers' optimal strategies as follows.

**Customer Joining Strategy.** Informed customers observe the queue length when they arrive and hence can make joining decisions based on queue-length dependent expected waiting times, which is equal to the average service time per customer multiplied by the actual queue length upon their arrival. Learning more accurate information about the actual queue length upon arrival, informed customers follow a *threshold strategy* to join or balk. Specifically, suppose that an informed customer observes a queue length of $n-1$ upon arrival. His expected payoff of joining the queue is then $\mathbb{E}[U_0(\hat{q})] = B(\hat{q}) - c\hat{q} - c\frac{n}{\mu}$. It is optimal for him to do so if and only if $\mathbb{E}[U_0(\hat{q})] \geq 0$, i.e., $n \leq \frac{B(\hat{q})-c\hat{q}}{c/\mu}$. Let $\lfloor x \rfloor$ denote the largest integer less than or equal to $x$ and let $n_e := \lfloor \frac{\mu(B(\hat{q})-p\hat{q})}{c} \rfloor$. Then, in the case of observable queue, an arriving customer joins if and only if he finds the current queue length (not including him) less than $n_e$ upon arrival. In this case, the queue length is never greater than $n_e$ and the system becomes an $M/M/1/n_e$ queue.

By classic results with regard to an $M/M/1/n_e$ queue, the effective arrival rate to the

system is given by

$$\lambda_e = \frac{1-\rho^{n_e}}{1-\rho^{n_e+1}}\Lambda, \tag{5.1}$$

where $\rho := \frac{\Lambda}{\mu}$ is the offered load of the system. Only a proportion $\frac{1-\rho^{n_e}}{1-\rho^{n_e+1}}$ of the arrival customers would join the queue.

**Reference Formation.** Having observed the queue length, informed customers update their expectation of the delay they would experience if they choose to join the queue. An arriving informed customer who sees $i-1$ customers ahead upon arrival would then estimate his expected waiting time in the system by $\bar{w} = \frac{i}{\mu}$ and hence forms a reference $r = \frac{ci}{\mu\hat{q}}$ for the waiting cost per unit purchase. Noting that the queue length varies over time, the reference varies from individual to individual.

**Actual Consumption.** Since the observed queue length and hence the reference are variable, informed customers in queue may have different optimal consumption functions. A customer who sees $i-1$ customers ahead upon arrival has an optimal consumption function $\mathcal{q}(w, \frac{i}{\mu})$ where $w$ follows an Erlang distribution with shape parameter $i$ and rate parameter $\mu$. By Proposition 1, $\mathcal{q}(w, \frac{i}{\mu})$ is non-decreasing in $w$ and non-increasing in $i$. A customer who saw a longer queue (larger $i$) upon arrival purchases no more than another one who saw a shorter queue upon arrival if they experience the same waiting time. This perhaps provides a reference-effects related explanation for the observation that a customer who sees a long queue but still decides to join is likely a patient one and exhibits a lower degree of sunk cost fallacy in decision.

## 5.1 Equilibrium Outcome

The equilibrium maximum queue length $n_e$ is given by $\lfloor \frac{(B(\hat{q})-p\hat{q})\mu}{c} \rfloor$, that is, the maximum queue length leading to non-negative reward. Let $X$ denote the number of customers already in queue when a new customer arrives. Then $\mathbb{P}(X = i) = P_i^O$, where $P_i^O = \rho^i \frac{1-\rho}{1-\rho^{n_e+1}}$, $i = 0, 1, 2, \cdots, n_e$ is the stationary distribution of the underlying Markov chain of the observable $M/M/1/n_e$ queue with effective arrival rate $\lambda_e = \frac{1-\rho^{n_e}}{1-\rho^{n_e+1}}\Lambda$ and service rate $\mu$. Let $\Pi^O$ denote the firm's profit per unit of time.

The firm's profit rate $\Pi^O$ is given by

$$\begin{aligned}\Pi^O &= \Lambda \cdot \sum_{i=0}^{n_e-1} P_i^O \mathbb{E}_w[q(w, \frac{i+1}{\mu})|X=i]p \\ &= \lambda_e \cdot \sum_{i=0}^{n_e-1} \frac{P_i^O}{1-P_{n_e}^O} \mathbb{E}_w[q(w, \frac{i+1}{\mu})|X=i]p \\ &= \lambda_e \cdot \sum_{i=0}^{n_e-1} \rho^i \frac{1-\rho}{1-\rho^{n_e}} \cdot \int_0^{+\infty} q(w, \frac{i+1}{\mu}) f(w; i+1, \mu) dw \cdot p \end{aligned} \tag{5.2}$$

in which the waiting time $w$ follows an Erlang distribution with shape parameter $i+1$ and rate parameter $\mu$ when $X=i$, and $f(w; i+1, \mu)$ is the corresponding probability density function. Let $\sum_{i=0}^{n_e-1} \rho^i \frac{1-\rho}{1-\rho^{n_e}} \cdot \int_0^{+\infty} q(w, \frac{i+1}{\mu}) f(w; i+1, \mu) dw$, i.e., the avearge purchasing quantity of those who join, be denoted by $\bar{q}$. We can write the profit function in the following form,

$$\Pi^O = \lambda_e \cdot \bar{q} \cdot p \tag{5.3}$$

Note that the expressions of $\bar{q}$ are different in the unobservable (e.g., in (4.3)) and observable (e.g., in (5.3)) cases. For ease of exposition, we abuse notation without differentiating them by using superscript.

## 5.2 Sensitivity Analysis

**Proposition 9.** *The firm's profit is increasing in $\theta^+$ and $\theta^-$.*

This result is the same as that in the case of unobservable queue. The firm's profit increases in the magnitude of the reference effects. As customers are subject to a higher degree of sunk cost fallacy, they purchase more and thus the firm has higher sales and profit.

**Proposition 10.** *The profit is piecewise continuous in the service rate $\mu$ with discontinuity points at which the profit jumps upward as $\mu$ increases.*

The firm's profit is discontinuous since the equilibrium threshold of customers' joining strategy, $n_e$, is discrete. Note that the expected waiting time conditional on seeing a given queue length decreases if the service rate increases. As $\mu$ increases beyond certain cut-off points at which the equilibrium joining threshold $n_e$ increases by 1, the number of customers who join

the system and the profit jump upward. Because of the challenges caused by discontinuity, we then resort to numerical studies to obtain more results. We illustrate our finding by some specific cases as follows. In the following numerical illustrations, we assume that the benefit function is given by $B(q) = \frac{q}{1+q}$ if not specifically mentioned. In figure 5.1, we study profit with parameters $\theta^+ = 0.03$, $\theta^- = 0.06$, $c = 0.2$, $p = 0.1$, $\Lambda = 1$. From the figure, we can see that the profit is piecewise continuous in the service rate, where each piece represents a different equilibrium queue length. At the "jumping" points, the right limits are always greater than the left limits. Furthermore, the gap of the profits of the case with sunk cost fallacy and that of the benchmark case without sunk cost fallacy will asymptotically diminish. The last two results apply to all groups of parameters.

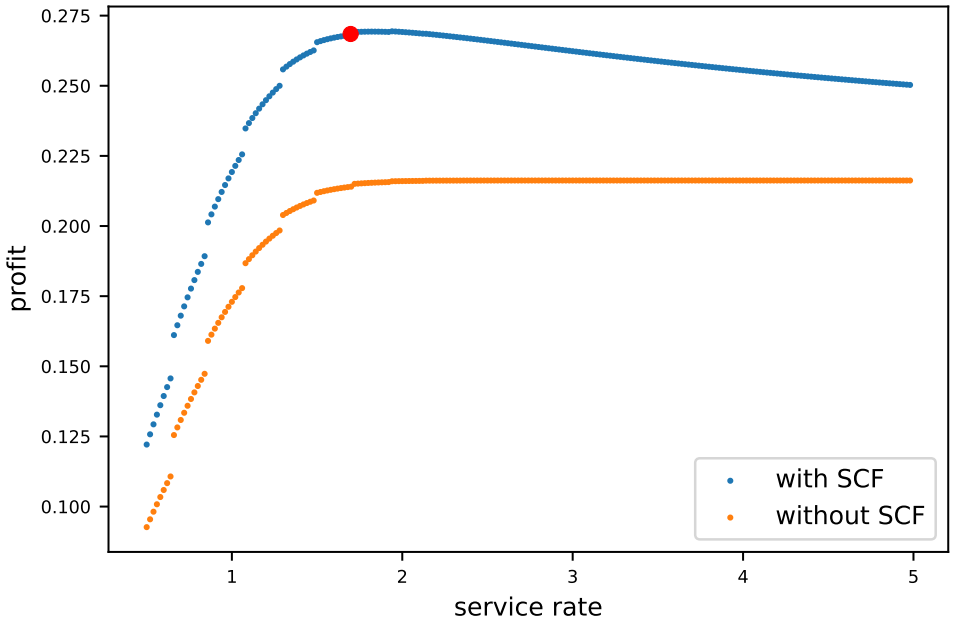



Figure 5.1: Firm's profit as service rate varies ($\theta^+ = 0.88$, $\theta^- = 1$, $c = 0.2$, $p = 0.1$, $\Lambda = 1$)

**Observation 2.** *The profit is non-monotone in service rate. There exists a threshold such that the profit is increasing for $\mu$ below the threshold and is piecewise decreasing for $\mu$ above the threshold.*

This observation suggests that, even if speeding up the service rate is costless, it may not be optimal for the firm manager to speed up the service rate. The intuition behind this result is that, although a faster service and a shorter waiting time can help attract more customers,

each customer may purchase less on average than the situation with a slower service rate. The increase in the number of customers, especially when the equilibrium $n_e$ remains unchanged after $\mu$ increases, may not cover the loss due to lower purchasing quantities.

As one can see from the comparison between Proposition 7 and Observation 2, the insights into managing a system with an unobservable queue or an observable one are similar: the firm's profit may increase or decrease as $\mu$ increases. Because of the discrete nature of customers' joining strategy and the discontinuity of the profit in the case of observable queue, we obtain one additional finding that is not present in the case of unobservable queue. The profit may decrease while the firm enforces a minor improvement in service speed (e.g., when $n_e$ remains unchanged), but may gets higher if the firm furthers improve to a certain level (e.g., when $n_e$ increases by 1). This result implies that while an incremental improvement of service may hurt the profitability, a substantial improvement could be beneficial.

**Proposition 11.** *The profit is piecewise increasing in the waiting cost c with discontinuity points at which the profit jumps downward as c increases.*

This proposition suggests that it is not always optimal for the firm to improve customer waiting experience even if doing so is effortless and costless. The firm's profit is discontinuous since the equilibrium threshold of customers' joining strategy, $n_e$, is discrete. Note that the optimal consumption quantity $q(w, \bar{w})$ increases in $c$ for given $(w, \bar{w})$ since customers have a bigger incentive to amortize the realized waiting cost $cw$ as it becomes larger. Therefore, the firm's sales and profit are increasing in $c$ if the number of joining customers (i.e., the joining threshold $n_e$) remains unchanged. But as $c$ increases beyond certain cut-off points at which the equilibrium joining threshold $n_e$ decreases by 1, the number of customers who join the system and the profit jumps downward. Note that this non-monotone result is not present in system without customer sunk cost fallacy. The result is illustrated by Figure 5.2.

**Observation 3.** *The optimal price of the case with sunk cost fallacy may be greater than that of the case without sunk cost fallacy.*

The observation is illustrated by Figure 5.3. This finding implies that a firm that ignores sunk cost fallacy in customers' purchasing decisions while the phenomenon indeed exists may

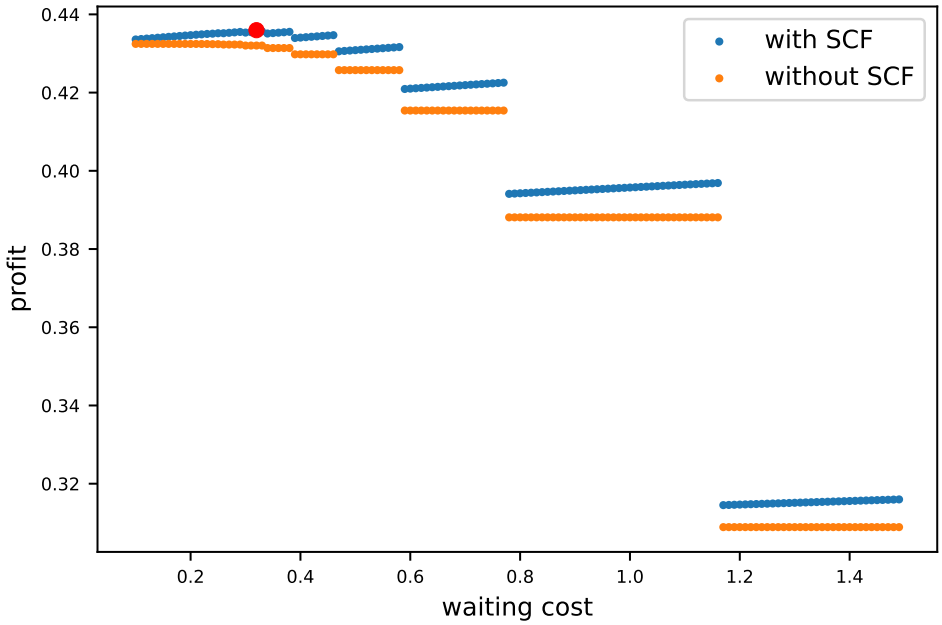



Figure 5.2: Firm's profit as waiting cost varies ($\theta^+ = 0.03$, $\theta^- = 0.06$, $p = 0.1$, $\mu = 5$, $\Lambda = 2$)

lower price the products and suffer a substantial loss.

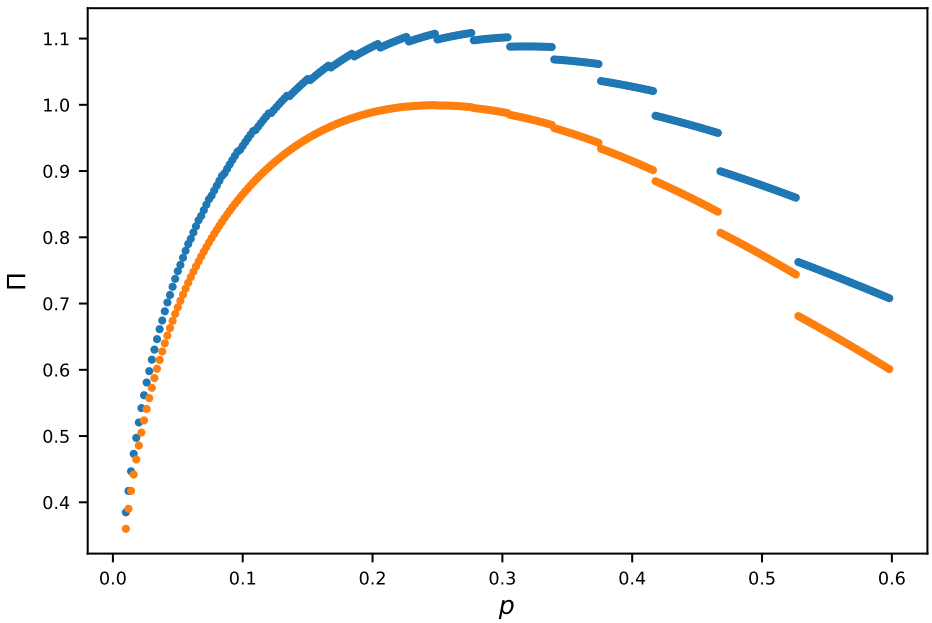



Figure 5.3: Firm's profit as price varies ($\theta^+ = 0.1$, $\theta^- = 0.2$, $c = 0.1$, $\mu = 4$, $\Lambda = 8$)

# Chapter 6

# Case with Information Heterogeneity

In this chapter, we consider the case where a $\gamma$ proportion of customers are informed and the remaining $1-\gamma$ proportion of customers are uninformed. Informed customers can observe the queue length upon arrival and uninformed ones cannot. Hu et al. (2018) study customers' optimal joining strategies and the equilibrium outcome in this case. They discuss the effects of growing information prevalence on system performance and find that throughput and social welfare can be unimodal in the fraction of informed customers. In this project, we further consider consumption quantity decisions of customers.

**Customer Joining Strategy.** Note that the exogenous arrival rates of informed and uninformed customers are $\gamma\Lambda$ and $(1-\gamma)\Lambda$, respectively. Informed customers can observe the queue length upon arrival and hence would simply adopt a threshold joining strategy as described in Chapter 5. Uninformed customers cannot observe the queue length upon arrival and use a pure or mixed joining strategy. Because of the existence of informed customers, uninformed customers need to take the impact of informed customers traffic into consideration when they decide their joining probability.

Let $\alpha$ denote the joining probability of uninformed customers. For a fixed $\gamma$, like the derivation of Hu et al. (2018), we can get the expected waiting time $W(\alpha)$ in steady state in the following form

$$W(\alpha) = \frac{P_0^M}{\mu}\left[\frac{1-(\rho_C)^n}{1-\rho_C}+\frac{\rho_C}{(1-\rho_C)^2}+(\rho_C)^n\left(\frac{1-n}{1-\rho_C}-\frac{1}{(1-\rho_C)^2}+\frac{1}{(1-\rho_U)^2}+\frac{n}{1-\rho_U}\right)\right] \quad (6.1)$$

where $P_0^M = \left(\frac{1-(\rho_C)^n}{1-\rho_C}+\frac{(\rho_C)^n}{1-\rho_U}\right)^{-1}$, $\rho_C = \frac{(\gamma+\alpha(1-\gamma))\Lambda}{\mu}$, $\rho_U = \frac{\alpha(1-\gamma)\Lambda}{\mu}$. Also, the equilibrium joining

strategy $\alpha_e$ for uninformed customers following from Hu et al. (2018) is

$$\alpha_e = \begin{cases} 0 & \text{if } cW(0) \geq B(\hat{q}) - p\hat{q} \\ 1 & \text{if } cW(1) \leq B(\hat{q}) - p\hat{q} \\ W^{-1}(\frac{B(\hat{q})-p\hat{q}}{c}) & \text{if } cW(0) < B(\hat{q}) - p\hat{q} < cW(1) \end{cases} \tag{6.2}$$

where $W(\alpha)$ is strictly increasing with respect to $\alpha$.

**Proposition 12** (Reproduction of Theorem 1 of Hu et al. (2018).)**.** *For given $\rho \equiv \frac{\Lambda}{\mu}$ and $v \equiv \frac{(B(\hat{q})-p\hat{q})\mu}{c}$, define $\underline{\rho} \equiv 1 - \frac{1}{v}$ and $\overline{\rho} \equiv y^*(v)$, where $y^*(v) \geq 0$ is the unique solution to $\lfloor v \rfloor + 1 + \frac{1}{1-y} - \frac{\lfloor v \rfloor + 1}{1-y^{\lfloor v \rfloor + 1}} = v$. Then the equilibrium joining probability $\alpha_e$ of uninformed customers depends on $\rho$ and $\gamma$ in the following way:*

*(i) (Always Full Participation) If $0 \leq \rho < \underline{\rho}$, $\alpha_e = 1$ for all $0 \leq \gamma \leq 1$.*

*(ii) (Partial to Full Participation) If $\underline{\rho} \leq \rho \leq \overline{\rho}$, $\alpha_e \neq 0$ for all $0 \leq \gamma \leq 1$. In particular, $0 < \alpha_e < 1$ for $0 \leq \gamma < \gamma_1^*(\rho, v)$ and $\alpha_e = 1$ for $\gamma_1^*(\rho, v) \leq \gamma \leq 1$, where $\gamma_1^*(\rho, v) \equiv 1 - \frac{1}{\rho} + \frac{2}{\rho}\left((v - \lfloor v \rfloor) + \sqrt{(v - \lfloor v \rfloor)^2 + 4L(\rho, v)}\right)^{-1} \in [0, 1]$, and $L(\rho, v) \equiv \frac{(v-\lfloor v \rfloor)(\rho-1)\rho^{\lfloor v \rfloor} + v - v\rho + \rho^{\lfloor v \rfloor} - 1}{(1-\rho)^2 \rho^{\lfloor v \rfloor}}$.*

*(iii) (Partial to No Participation) If $\rho > \overline{\rho}$, $\alpha_e \neq 1$ for all $0 \leq \gamma \leq 1$. In particular, $0 < \alpha_e < 1$ for $0 \leq \gamma < \gamma_0^*(\rho, v)$ and $\alpha_e = 0$ for $\gamma_0^*(\rho, v) \leq \gamma \leq 1$, where $\gamma_0^*(\rho, v) \equiv \frac{y^*(v)}{\rho} \in [0, 1]$.*

**References and Consumption.** Customers' references are formed and the consumption quantities are decided according to the same ways described in Chapters 4 and 5 for uninformed and informed customers, respectively. The only thing to note in this case is that uninformed customers need to take into consideration the presence of informed customers when they form their estimation of the expected waiting time.

## 6.1 Equilibrium Outcome

The stationary distribution derived from the mixed queue is given by

$$P_i^M(\alpha, \gamma) = \begin{cases} (\rho_C)^i P_0^M & \text{if } 0 \leq i < n, \\ (\rho_C)^n (\rho_U)^{i-n} P_0^M & \text{if } i \geq n. \end{cases} \tag{6.3}$$

Therefore, the expected profit per unit time $\Pi^M$ can be written as

$$\Pi^M(\alpha,\gamma) = p\gamma\lambda_e \cdot \sum_{i=0}^{n-1} P_i^M \mathbb{E}_{w^O}[q(w, \frac{i+1}{\mu})|X=i] + p(1-\gamma)\Lambda\alpha_e \cdot \sum_{i=0}^{+\infty} P_i^M \mathbb{E}_{w^U}[q(w, \frac{1}{\mu-(1-\gamma)\Lambda\alpha_e})].$$

**Lemma 2** (Reproduction of Lemma 1 of Hu et al. (2018))**.** *For any given $\gamma \in [0,1)$, the queue length $Q(\alpha)$ in the steady state is stochastically increasing in $\alpha$. Therefore, the expected waiting time $W(\alpha)$ is strictly increasing in $\alpha$.*

## 6.2 Sensitivity Analysis

Given that the sensitivity analysis results in the cases of unobservable and observable queues are qualitatively similar except that the firm's profit is discontinuous in the observable case, we omit the sensitivity analysis with regard to the parameters $\mu$, $c$, $p$, $\theta^+$ and $\theta^-$ in this case. The results are similar to that in Chapter 5. We only discuss the impact of the parameter $\gamma$ in this section.

**Observation 4.** *The profit is unimodal in $\gamma$.*

The result is similar to that in Hu et al. (2018) where the phenomenon of sunk cost fallacy is not present. The profit may monotonically increases in $\gamma$ or increase and then decrease in $\gamma$. This finding suggests that in the presence of sunk cost fallacy, a higher level of information prevalence in the market may not be beneficial to the firm.

# Chapter 7

# Comparisons of Unobservable and Observable Cases

Apart from the comparisons of cases with sunk cost fallacy and without sunk cost fallacy in the above three kinds of information settings, we also want to know how the information disclosure can affect the profit. Therefore, in this chapter, we compare the optimal service rate, optimal waiting cost, and optimal price and the corresponding profit in unobservable case and observable case in a numerical way.

We first learn the optimal service rate and the corresponding profit in the unobservable and observable cases. From Figure 7.1, we can see that the optimal service rate exists for the unobservable case, which is just the same as what we found in Chapter 4. And for the observable case, the profit is upper bounded by the profit of unobservable case when the service rate is large enough. This shows that the highest profit for the unobservabel case is greater than the profit for the observable case. Moreover, we have the following proposition:

**Proposition 13.** $\lim_{\mu\to+\infty} \Pi^O = \lim_{\mu\to+\infty} \Pi^U$.

Next, we study the optimal waiting cost and the corresponding profit in the unobservable and observable cases. From Figure 7.2, we can see that the profit of the observable case is upper bounded by the profit of the unobservable case when waiting cost is small. And obviously, the highest profit that a firm can obtain in the unobservable case is far more greater than the highest profit in the observable case. This implies that, the disclosure of information may do harm to the firm's profit. Moreover, we have the following proposition:

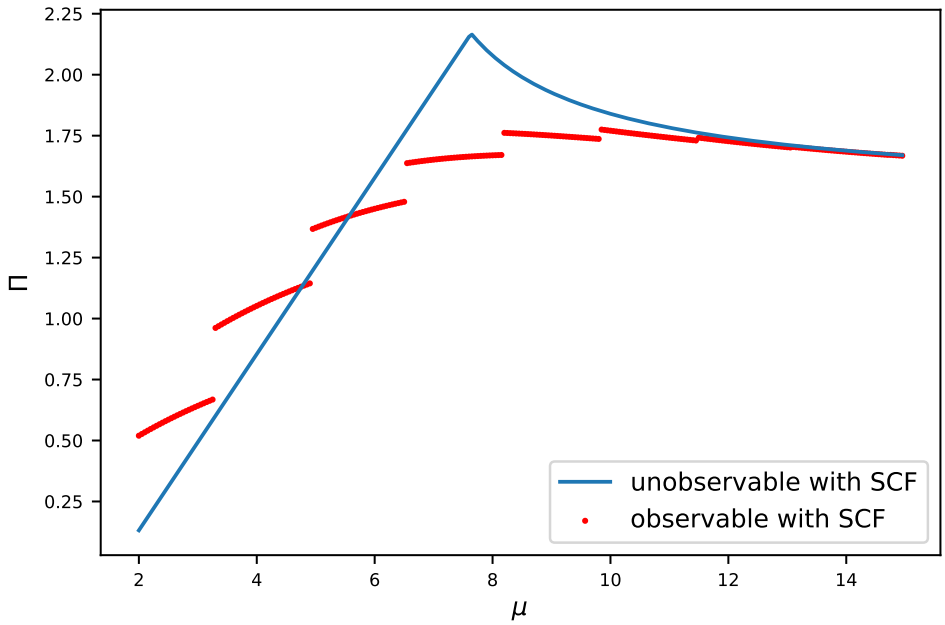



Figure 7.1: Firm's profit as the service rate varies ($\theta^+ = 0.88$, $\theta^- = 1$, $c = 0.5$, $p = 0.2$, $\Lambda = 6$)

**Proposition 14.** $\lim_{c \to 0^+} \Pi^O = \lim_{c \to 0^+} \Pi^U$.

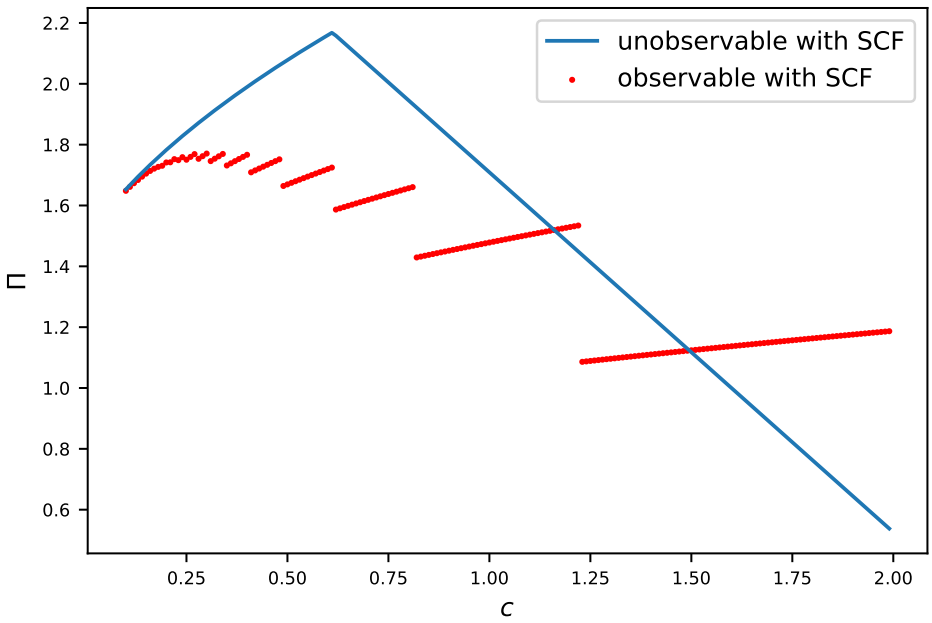



Figure 7.2: Firm's profit as the waiting cost varies ($\theta^+ = 0.88$, $\theta^- = 1$, $p = 0.2$, $\mu = 8$, $\Lambda = 6$)

Lastly, we study the optimal price and the corresponding profit in both the unobservable case and observable case. From Figure 7.3, we can see that the optiaml price of the unobservable case is less than the optimal price of the observable case; and that the highest profit a firm can gain in the unobservable case is less than that in the observable case. This implies that, sometimes, disclosing queue length information can help to improve the firm's profit if the price is appropriately set.

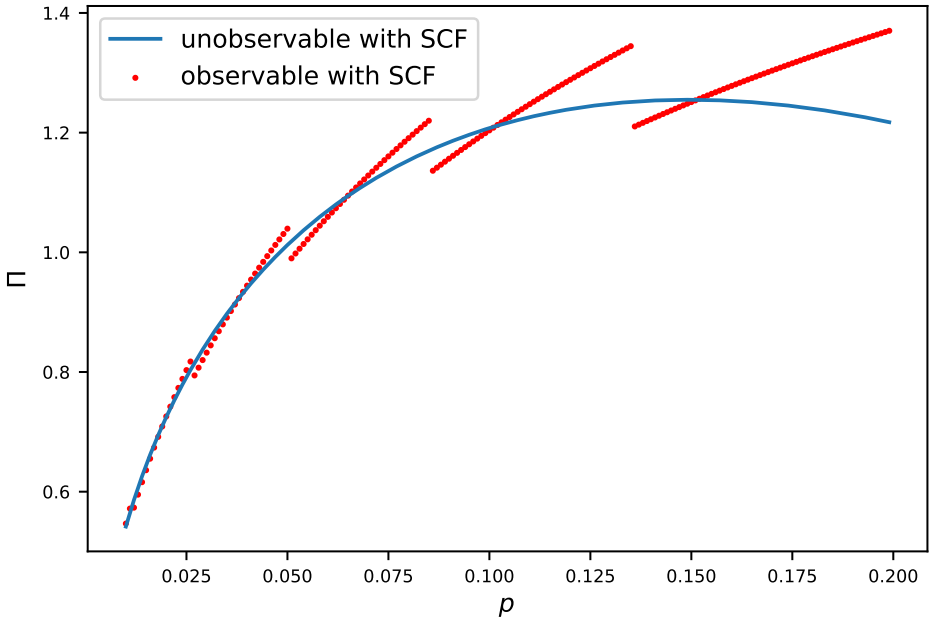



Figure 7.3: Firm's profit as the price varies ($\theta^+ = 0.88$, $\theta^- = 1$, $c = 0.5$, $\mu = 5$, $\Lambda = 6$)

From the above analysis and illustration, we can conclude that the impact of the disclosure of queue length information may be mixed. Information disclosure may benefit or do harm to the firm.

# Chapter 8

# Discussion

In this report, we capture the sunk cost fallacy of consumer consumption in a service queue by using reference-effect transaction utility model. Different from the model of Ho et al. (2018), we explicitly model the rationale of customers that leads to the sunk cost really. Similar to Ülkü et al. (2020), we incorporate a sunk cost term into the utility function. However, Ülkü et al. (2020) do not explain how the interaction of quantity and waiting time gives rise to sunk cost fallacy of customers. In our model, we adopt the prospect theory of Kahneman and Tversky (1979), and also follow the expectation concern of Thaler (1985). In the following sections, we discuss the managerial implications of our work.

## 8.1 Determining Optimal Service Rate

In the unobservable case, when customers adopt a mixed strategy (i.e., $\alpha_e < 1$) to join the queue, the firm's profit will become higher if the firm speeds up the service. However, when all customers choose to join the queue, furthere increasing the service rate cannot bring more customers but will lead to a lower profit. This is because higher service rate will not help to attract more customers when all the customers already chose to join, but it will result in lower purchasing quantities of existing customers because of lower influence of sunk cost fallacy. Similar findings are identified in the observable case of the case with heterogeneous information disclosure.

Overall, if the firm ignores the presence of sunk cost fallacy, they may get a suboptimal profit, and set a service rate higher than the optimal one in the unobservable case.

## 8.2 The Effect of Improving Waiting Experience

In the unobservable case, when all the arriving customers already chose to join the queue, a higher waiting cost will help to intensify the sunk cost fallacy in consumption, and hence benefit the firm. However, when the potential arrival rate is sufficiently large (i.e., they use a mixed strategy to join the queue), an increment in the waiting cost may hurt the firm's profit, since a higher waiting cost drives customers away, and the increase in the profit brought by per customer cannot offset the loss of potential customers in the market. In the observable case, the situation is more complicated, in some cases, a small increment in the waiting cost may help to improve the firm's profit; but in the other, it may do damage to the firm's profit. This is because, at proper point, a small increase in waiting cost may not affect the queue length, but will lead customers to purchase more; in the other cases, a small increase in waiting cost may result in a loss of customers, which may not be covered by the increased profit gained by per unit customer. In both unobservable and observable cases, if the firm ignores the sunk cost fallacy, it may mistakenly get a suboptimal profit. At some point, it is not necessary to further invest in queue management.

## 8.3 The Optimal Product Price

In the unobservable case, the firm considering sunk cost fallacy may set a lower price than the firm ignoring sunk cost fallacy. Even though this will lead to a lower margin, it will help to attract more customers, and each customer will purchase more. The managerial implication is that, if the firm ignores sunk cost fallacy in business, they may incur a substantial loss accidentally. This also implies that the firm may use the price tool to "trick" customers to join (because they set a lower price), but once they choose to join, their consumption decision is affected by the sunk cost fallacy, and hence benefit the firm. In the observable case, from our numerical results, this phenomenon is not very obvious.

On the other hand, it is also possible that a firm knowing the presence of sunk cost fallacy should set a higher price than the firm ignoring sunk cost fallacy. This act will drive away some customers, but it would help to get a higher margin per purchases.

# Chapter 9

# Concluding Remarks

Traditional economic literature believed that in the queueing system, the higher the service rate is, the better for the profit. However, there are some recent literature study how the behavioral factors influence the profit and find out that the firm may intentionally slow down the service to get better profit. Debo et al. (2012) assumes information heterogeneity and concludes that customers' joining strategy is a "hole" strategy, and that the firm may slow down the service rate such that customers could infer higher quality from the queue length. Apart from this model work, Ülkü et al. (2020) empirically finds that longer waiting time may lead to more purchasing quantity due to sunk cost fallacy. We further model the sunk cost fallacy, and also reach the conclusion that rise in the service rate alone may do harm to the firm's profit.

We also extend the results in Naor (1969) and Edelson and Hilderbrand (1975) to multi-item cases and consider it as the baseline of our model. From comparisons between benchmark and sunk cost cases, we learned how individual-level behaivors influence the system-level behaviors in the discussion part. We now conclude the results here. Sunk cost fallacy forces the firm to speed up the service rate. When the rise of service rate is costless, the sunk cost case would always create more profit than the benchmark case. However, if the rise of service rate is costly, the firm should balance the improvement of average gain and the rise of operating cost. Also, sunk cost fallacy make the firm to build up worse queueing environment so as to create more sunk cost fallacy. This is counter intuitive, because common belief is that less waiting cost will attract more customers to join the queue. In our model, the impact of sunk cost fallacy on the purchasing quantity dominates the loss of customers.

To the best of our knowledge, we are the first to model sunk cost fallacy based on the

reference effect with respect to the expectation of outcome, which coordinates to the theory of Thaler (1985). Ho et al. (2018) uses mental accounting to propose an disutility term of the value function. We also incorporate a sunk cost fallacy term, but uses the average expected outcome as reference point, which goes more close to the classical literature.

There are several major limitations to our work. First, we assume the waiting cost is proportional to the waiting time. Some research in existing literature (Leclerc et al., 1995; Weber and Milliman, 1997; Soman, 2001; Krishnamurthy and Kumar, 2002) finds that time is different from money in mental accounting. Therefore, there is a need to develop models that treat waiting cost properly or differently from monetary costs. Second, the sunk cost fallacy is modelled based on the existing literature, empirical testing is needed for the sunk cost term. Third, more general prospect function can be considered in our model in extension. We leave these problems for the future research.

# References


Allon, G., M. Kremer, K. Donohue, E. Katok, and S. Leider (2018). Behavioral foundations of queueing systems. *The Handbook of Behavioral Operations 9325*.

Arkes, H. R. and P. Ayton (1999). The sunk cost and concorde effects: Are humans less rational than lower animals? *Psychological bulletin 125*(5), 591.

Arkes, H. R. and C. Blumer (1985). The psychology of sunk cost. *Organizational behavior and human decision processes 35*(1), 124–140.

Baker, J., A. Parasuraman, D. Grewal, and G. B. Voss (2002). The influence of multiple store environment cues on perceived merchandise value and patronage intentions. *Journal of marketing 66*(2), 120–141.

Baucells, M. and W. Hwang (2017). A model of mental accounting and reference price adaptation. *Management Science 63*(12), 4201–4218.

Buell, R. W. (2021). Last-place aversion in queues. *Management Science 67*(3), 1430–1452.

Debo, L. G., C. Parlour, and U. Rajan (2012). Signaling quality via queues. *Management Science 58*(5), 876–891.

Dube-Rioux, L., B. H. Schmitt, and F. Leclerc (1989). Consumers' reactions to waiting: when delays affect the perception of service quality. *ACR North American Advances*.

Edelson, N. M. and D. K. Hilderbrand (1975). Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society*, 81–92.

Garland, H. (1990). Throwing good money after bad: The effect of sunk costs on the decision to escalate commitment to an ongoing project. *Journal of Applied Psychology 75*(6), 728.

Giebelhausen, M. D., S. G. Robinson, and J. J. Cronin (2011). Worth waiting for: increasing satisfaction by making consumers wait. *Journal of the Academy of Marketing Science 39*(6),

889–905.

Haita-Falah, C. (2017). Sunk-cost fallacy and cognitive ability in individual decision-making. *Journal of Economic Psychology 58*, 44–59.

Hassin, R. and M. Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, Volume 59. Springer Science & Business Media.

Heath, C. (1995). Escalation and de-escalation of commitment in response to sunk costs: The role of budgeting in mental accounting. *Organizational Behavior and Human Decision Processes 62*(1), 38–54.

Ho, T.-H., I. P. Png, and S. Reza (2018). Sunk cost fallacy in driving the world's costliest cars. *Management Science 64*(4), 1761–1778.

Hong, F., W. Huang, and X. Zhao (2019). Sunk cost as a self-management device. *Management Science 65*(5), 2216–2230.

Hu, M., Y. Li, and J. Wang (2018). Efficient ignorance: Information heterogeneity in a queue. *Management Science 64*(6), 2650–2671.

Huang, T., G. Allon, and A. Bassamboo (2013). Bounded rationality in service systems. *Manufacturing & Service Operations Management 15*(2), 263–279.

Huang, T. and Y.-J. Chen (2015). Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management 24*(5), 778–790.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292.

Kahneman, D. and A. Tversky (1984). Choices, values, and frames. *American Psychologist 39*(4), 341–350.

Kong, G., S. Rajagopalan, and C. Tong (2018). Pricing diagnosis-based services when customers exhibit sunk cost bias. *Production and Operations Management 27*(7), 1303–1319.

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

Kremer, M. and L. Debo (2016). Inferring quality from wait time. *Management Science 62*(10), 3023–3038.

Krishnamurthy, P. and P. Kumar (2002). Self-other discrepancies in waiting time decisions. *Organizational Behavior and Human Decision Processes 87*(2), 207–226.

Kumar, P. and P. Krishnamurthy (2008). The impact of service-time uncertainty and anticipated congestion on customers' waiting-time decisions. *Journal of Service Research 10*(3), 282–292.

Leclerc, F., B. H. Schmitt, and L. Dube (1995). Waiting time and decision making: Is time like money? *Journal of Consumer Research 22*(1), 110–119.

Li, X., P. Guo, and Z. Lian (2016). Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management 25*(11), 1885–1901.

Lu, Y., A. Musalem, M. Olivares, and A. Schilkrut (2013). Measuring the effect of queues on customer purchases. *Management Science 59*(8), 1743–1763.

Maister, D. H. et al. (1984). *The psychology of waiting lines*. Citeseer.

McAfee, R. P., H. M. Mialon, and S. H. Mialon (2010). Do sunk costs matter? *Economic Inquiry 48*(2), 323–336.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, 15–24.

Puaschunder, J. M. et al. (2019). Mental temporal accounting. In *Proceedings of the 13th International RAIS Conference on Social Sciences and Humanities*, pp. 114–124. Scientia Moralitas Research Institute.

Ren, H. and T. Huang (2018). Modeling customer bounded rationality in operations management: A review and research opportunities. *Computers & Operations Research 91*, 48–58.

Ren, H., T. Huang, and K. Arifoglu (2018). Managing service systems with unknown quality and customer anecdotal reasoning. *Production and Operations Management 27*(6), 1038–1051.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization 1*(1), 161–176.

Soman, D. (2001). The mental accounting of sunk time costs: Why time is not like money. *Journal of Behavioral Decision Making 14*(3), 169–185.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of economic behavior*

*& organization 1*(1), 39–60.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science 4*(3), 199–214.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making 12*(3), 183–206.

Ülkü, S., C. Hydock, and S. Cui (2020). Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science 66*(3), 1149–1171.

Weber, E. U. and R. A. Milliman (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management science 43*(2), 123–144.

Yang, L., P. Guo, and Y. Wang (2018). Service pricing with loss-averse customers. *Operations research 66*(3), 761–777.

# Appendix

In this appendix, we present the proofs of our main results. Proofs of results that are reproduced from literature or that are obvious are omitted.

## Customers' joining strategy

When a customer arrives the system, he makes joining decision on the basis of expected utility. Suppose he saw $i$ customers ahead of him, then the expected waiting time is $\frac{i}{\mu}$. He joins if and only if the expected utility is nonnegative, that is $\mathbb{E}[U_0(\hat{q})] = \mathbb{E}[B(\hat{q}) - p\hat{q} - cw] = B(\hat{q}) - p\hat{q} - c\frac{i}{\mu} \geq 0$ (We ignore the time cost during the service). It's equivalent to the condition that $i \leq \frac{\mu}{c}[B(\hat{q}) - p\hat{q}]$. Therefore, the customers' joining strategy is a threshold strategy.

## Proof of Proposition 1

For $w$ such that $\frac{c\bar{w}}{\hat{q}} > \frac{cw}{q}$, i.e., $w \leq \frac{\bar{w}}{\hat{q}}q$, the utility function $U(q)$ can be written as $U(q) = B(q) - pq - cw + \theta^+\frac{c\bar{w}}{\hat{q}} - \theta^+\frac{cw}{q}$. By *first order condition*, $\arg\max_q\{B(q) - pq - cw + \theta^+\frac{c\bar{w}}{\hat{q}} - \theta^+\frac{cw}{q}\} = \arg\max_q\{B(q) - pq - \theta^+\frac{cw}{q}\}$, which is denoted by $\phi(w, \theta^+)$. Furthermore, the optimal value can only be reached when the optimal consumption quantity meets the condition that $w \leq \frac{\bar{w}}{\hat{q}}\phi(w, \theta^+)$. Let $\Omega(\bar{w}, \theta^+)$ denote the fixed point of $\frac{\bar{w}}{\hat{q}}\phi(w, \theta^+)$ with regard to $w$, then $w \leq \Omega(\bar{w}, \theta^+)$.

Similarly, when $w > \frac{\bar{w}}{\hat{q}}q$, the optimal consumption quantity $\phi(w, \theta^-) = \arg\max_q\{B(q) - pq - \theta^-\frac{cw}{q}\}$ must meet the condition that $w > \frac{\bar{w}}{\hat{q}}\phi(w, \theta^-)$. Let $\Omega(\bar{w}, \theta^-)$ denote the fixed point of $\frac{\bar{w}}{\hat{q}}\phi(w, \theta^-)$ with regard to $w$, then $w > \Omega(\bar{w}, \theta^-)$.

Otherwise, when $w$ lies in $(\Omega(\bar{w}, \theta^+), \Omega(\bar{w}, \theta^-)]$, the above optimal consumption quantities

cannot be achieved. Therefore, by concavity of the $B(q) - pq - \frac{\theta^+ cw}{q}$ and $B(q) - pq - \frac{\theta^- cw}{q}$, the optimal value can only be achieved at the intersection point of the two curves representing $B(q) - pq + \frac{\theta^+ c\bar{w}}{\hat{q}} - \frac{\theta^+ cw}{q}$ and $B(q) - pq + \frac{\theta^- c\bar{w}}{\hat{q}} - \frac{\theta^- cw}{q}$. That is, the optimal consumption quantity is $\frac{\hat{q}}{\bar{w}}w$.

Next, we prove the existence and uniqueness of the fixed point of $\frac{\bar{w}}{\hat{q}}\phi(w, \theta)$ where $\theta = \theta^+$ or $\theta = \theta^-$. By *first order condition*, $\phi(w, \theta)$ is the solution of equation $B'(q) - p + \frac{\theta cw}{q^2} = 0$ with regard to $q$. Replace $\phi(w, \theta)$ with $\frac{\hat{q}}{\bar{w}}w$ in the above equation, we get $B'(\frac{\hat{q}}{\bar{w}}w) - p + \frac{\theta c\hat{w}^2}{\hat{q}^2 w} = 0$. Since $B'(q) - p + \frac{\theta cw}{q^2}$ is strictly decreasing in $q$ by taking the derivative of the function, we know that, if the fixed point exists, then it is unique. Hence, we now only need to prove the existence of the solution. We assume that $\lim_{q \longrightarrow +\infty} B'(q) = 0$ (diminishing marginal benefit), since $B'(0) - p > 0$, therefore, as $q \longrightarrow +\infty$, $B'(q) + \frac{\theta cw}{q^2} \longrightarrow 0 + 0$. By continuity, there must exists a solution $\phi'$ satisfying $B'(\phi') - p + \frac{\theta cw}{\phi'^2} = 0$. Especially, when $B(q)$ is bounded, the assumption $\lim_{q \longrightarrow +\infty} B'(q) = 0$ is automatically satisfied.

By *first order condition*, $\phi(w, \theta)$ should satisfy the following condition

$$B'(\phi) - p + \frac{\theta cw}{\phi^2} = 0.$$

By *implicit function theorem*, taking derivative with respect to $w$ (or $c$, $\theta$), we obtain

$$B''(\phi)\frac{\partial \phi}{\partial w} + \frac{\theta c}{\phi^2} - 2\frac{\theta cw}{\phi^3}\frac{\partial \phi}{\partial w} = 0,$$

so, $\frac{\partial \phi}{\partial w} = \frac{\theta c\phi}{2\theta cw - B''(\phi)\phi^3}$. Since $B''(\cdot) < 0$, then $\frac{\partial \phi}{\partial w} \geq 0$. That is, $\phi$ is non-decreasing in $w$. Similarly, $\phi$ is non-decreasing in $c$, $\theta^+$, and $\theta^-$. Notice that when $w$ is moderate (lies between $\Omega(\bar{w}, \theta^+)$ and $\Omega(\bar{w}, \theta^-)$, the 'non-decreasing' claims are also correct. Thereafter, the optimal consumption quantity is non-decreasing in $w$, $c$, $\theta^+$, and $\theta^-$. Using *implicit function theorem* again, the optimal consumption quantity is non-increasing in $p$ (this is obvious). Since $\bar{w}$ only affects the thresholds $\Omega(\bar{w}, \theta^+)$ and $\Omega(\bar{w}, \theta^-)$, and $\Omega(\bar{w}, \theta)$ is non-decreasing in $\bar{w}$, so the optimal quantity is non-increasing in $\bar{w}$.

## Proof of Proposition 2

When $\theta^+ = \theta^- = \theta$, the optimal consumption quantity $q$ can be expressed as $\phi$. By *first order condition*,

$$B'(\phi) - p + \frac{\theta c w}{\phi^2} = 0. \tag{9.1}$$

The optimal consumption quantity can be implicitly expressed by the above function. Then take derivative of variable $p$ with respect to $\phi$, we have

$$B''(\phi)\frac{\partial \phi}{\partial p} - 1 - \frac{2\theta c w}{\phi^3}\frac{\partial \phi}{\partial p} = 0, \tag{9.2}$$

that is,

$$\frac{\partial \phi}{\partial p} = -\frac{\phi^3}{2\theta c w - B''(\phi)\phi^3} < 0. \tag{9.3}$$

Again, take derivative of variable $p$ with respect to $\phi$, we have

$$\begin{aligned} \frac{\partial^2 \phi}{\partial p^2} &= -\frac{3\phi^2\frac{\partial \phi}{\partial p}(2\theta c w - B''(\phi)\phi^3) - \phi^3(-B'''(\phi)\phi^3\frac{\partial \phi}{\partial p} - 3B''(\phi)\phi^2\frac{\partial \phi}{\partial p})}{(2\theta c w - B''(\phi)\phi^3)^2} \\ &= -\phi^2\frac{\partial \phi}{\partial p}\frac{6\theta c w + B'''(\phi)\phi^4}{(2\theta c w - B''(\phi)\phi^3)^2} > 0. \end{aligned} \tag{9.4}$$

We can easily generate the result to the case $\theta^- > \theta^+ \geq 0$. Therefore, $\phi$, or $q$, is decreasing convex in price $p$.

## Proof of Proposition 3

By *implicit function theorem*, we can obtain

$$\frac{\partial^2 \phi}{\partial w^2} = \frac{\theta^2 c^2 \phi(-2\theta^2 c^2 w + \phi^4(B'''(\phi) + \frac{4B''(\phi)}{\phi}))}{(2\theta c w - B''(\phi)\phi^3)^3}$$

Therefore, if $B'''(q) \leq -\frac{4B''(q)}{q}$, then $\frac{\partial^2 \phi}{\partial w^2}$ is negative. Equivalently, the optimal consumption quantity is increasing concave in the waiting time $w$ (it is increasing by Proposition 1). Similar results apply to waiting cost $c$ and coefficient of reference effect $\theta$.

## Proof of Proposition 5

Since $\frac{\partial \phi(w,\theta^+)}{\partial \theta^+} = \frac{c w \phi(w,\theta^+)}{2\theta^+ c w - B''(\phi(w,\theta^+))\phi(w,\theta^+)^3} > 0$, $\bar{q}$ is then increasing in $\theta^+$. But $\alpha_e$, $\Lambda$, and $p$ are not affected by $\theta^+$. So, the profit is increasing in $\theta^+$. Similarly, the profit is increasing in $\theta^-$.

## Proof of Lemma 1

This can be proved by showing that the moment generating function of $\sum_{i=1}^{N_p} T_i$ is the same as that of an exponential distribution with mean $\frac{1}{p\mu}$, where $\mu$ is the mean of $T_i$'s. We omit the details since it is straightforward.

## Proof of Proposition 7

When $\lambda_e = \Lambda$, since $c\bar{w} = \frac{c}{\mu-\Lambda}$ is decreasing in $\mu$, and $R(\frac{c\bar{w}}{\hat{q}} - \frac{cw}{q})$ can be rewritten as $\frac{c\bar{w}}{\hat{q}} R(1 - \frac{\hat{q}}{q} \cdot \frac{w}{\bar{w}})$, then $\bar{q}$ is decreasing in $\mu$. Therefore, $\Pi^U = \Lambda \bar{q} p$ is decreasing in $\mu$.

When $\lambda_e = \mu - \frac{c}{B(\hat{q})-p\hat{q}}$, by proposition 6, the average consumption quantity $\bar{q} = \mathbb{E}_w[q(w, \bar{w})]$ is not affected by $\mu$, effective arriving rate $\lambda_e$ increases in $\mu$, and $p$ is independent of $\mu$. By $\Pi^U = \lambda_e \bar{q} p$, the profit is increasing in $\mu$ when $\lambda_e = \mu - \frac{c}{B(\hat{q})-p\hat{q}}$.

## Proof of Proposition 8

When $\lambda_e = \Lambda$, since $c\bar{w} = \frac{c}{\mu-\Lambda}$ is increasing in $c$, and $R(\frac{c\bar{w}}{\hat{q}} - \frac{cw}{q})$ can be rewritten as $\frac{c\bar{w}}{\hat{q}} R(1 - \frac{\hat{q}}{q} \cdot \frac{w}{\bar{w}})$, then $\bar{q}$ is increasing in $c$. Therefore, $\Pi^U = \Lambda \bar{q} p$ is increasing in $c$.

If $\lambda_e = \mu - \frac{c}{B(\hat{q})-p\hat{q}} < \Lambda$, by Proposition 6, average consumption quantity is independent of $c$. Since the effective arrival rate is decreasing in $c$, when price $p$ does not change, from the following formula

$$\Pi^U = p\lambda_e \bar{q} \tag{9.5}$$

the profit is decreasing in $c$. And note that this proof does not require the condition $\theta^- = \theta^+$ to be met.

## Proof of Proposition 9

The proof is similar to that of Proposition 5.

## Proof of Proposition 10

The continuity of each piece is obvious. We now focus on the proof of the part "jump upward". Suppose we are interested in the point $\mu_0$, with $\lim_{\mu \longrightarrow \mu_0+} \lfloor \frac{\mu(B(\hat{q})-\hat{q})}{c} \rfloor - \lim_{\mu \longrightarrow \mu_0-} \lfloor \frac{\mu(B(\hat{q})-\hat{q})}{c} \rfloor = (n_e+1)-n_e = 1$. Since $\lim_{\mu \longrightarrow \mu_0+} \lambda_e \geq \lim_{\mu \longrightarrow \mu_0-} \lambda_e$ by *Cauchy-Schwarz inequality*, and $\lim_{\mu \longrightarrow \mu_0+} \bar{q} \geq_{st} \lim_{\mu \longrightarrow \mu_0-} \bar{q}$ by *coupling* method, we can obtain $\lim_{\mu \longrightarrow \mu_0+} \Pi^O \geq \lim_{\mu \longrightarrow \mu_0-} \Pi^O$.

## Proof of Proposition 11

The continuity of each piece is obvious. We now focus on the proof of the part "jump downward". Suppose the point we are interested in is $c = c_0$, where $\lim_{c \longrightarrow c_0-} \lfloor \frac{\mu(B(\hat{q})-\hat{q})}{c} \rfloor - \lim_{c \longrightarrow c_0+} \lfloor \frac{\mu(B(\hat{q})-\hat{q})}{c} \rfloor = (n_e+1) - n_e = 1$. Like the argument in the proof of Proposition 10, we obtain $\lim_{c \longrightarrow c_0+} \Pi^O \leq \lim_{c \longrightarrow c_0-} \Pi^O$.

## Proof of Proposition 13

Since each term in the function of $\Pi^O$ and $\Pi^U$ is positive, the limitation and summation are exchangeable. The result follows.

## Proof of Proposition 14

The proof is similar to that of Proposition 13.