Control Mechanisms in Queueing Systems with Nonlinear Waiting Costs

by

Ata Ghareh Aghaji Zare

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Fredrik Odegaard Associate Professor
Supervisor	Hossein Abouee Mehrizi Associate Professor
Internal Member	Fatih Safa Erenay Assistant Professor
Internal Member	Qi-Ming He Professor
Internal-external Member	Steve Drekic Associate Professor

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In many queueing systems, customers have been observed to exhibit strategic behavior. Each customer gains a value when receiving a product or getting served and suffers when incurring a delay. We consider a nonlinear waiting cost function to capture the sensitivity of customers toward delay. We investigate customers' behavior and system manager's strategy in two different settings: (1) customers are served in a service system, or (2) they receive a product in a supply chain.

In the first model, we study an unobservable queueing system. We consider that customers are impatient, and are faced with decision problems whether to join a service system upon arrival, and whether to remain or renege at a later time. The goal is to address two important elements of queueing analysis and control: (1) customer characteristics and behavior, and (2) queueing control. The literature on customer strategic behavior in queues predominately focuses on the effects of waiting time and largely ignores the mixed risk attitude of customer behavior. Empirical studies have found that customers' risk attitudes, their anticipated time, and their wait time affect their decision to join or abandon a queue. To explore this relationship, we analyze the mixed risk attitude together with a non-linear waiting cost function that includes the degree of risk aversion. Considering this behavior, we analyze individuals' joint balking and reneging strategy and characterize socially optimal strategy. To determine the optimal queue control policy from a revenue-maximizer perspective, which induces socially optimal behavior and eliminates customer externalities, we propose a joint entrance-fee/abandonment-threshold mechanism. We show that using a pricing policy without abandonment threshold is not sufficient to induce socially optimal behavior and in many cases results in a profit lower than the maximum social welfare the system can generate. Also, considering both customer characteristics and queue control policy, our findings suggest that customers with a moderate anticipation time provide higher expected revenue, acknowledging the importance of understanding customer behavior with respect to both wait time and risk attitude in the presence of anticipation time.

In the second model, we consider a two-echelon production inventory system with a

single manufacturer and a single distribution center (DC) where the manufacturer has a finite production capacity. There is a positive transportation time between the manufacturer and the DC. Each customer gains a value when receiving the product and suffers a waiting cost when incurring a delay. We assume that customers' waiting cost depends on their degree of impatience with respect to delay (delay sensitivity). We consider a nonlinear waiting cost function to show the degree of risk aversion (impatience intensity) of customers. We assume that customers follow the strategy p where they join the system and place an order with probability p. We analyze the inventory system with a base-stock policy in both the DC and the manufacturer. Since customers and supply chain holder are strategic, we study the Stackelberg equilibrium assuming that the DC acts as a Stackelberg leader and customers are the followers. We first obtain the total expected revenue and then derive the optimal base-stock level as well as the optimal price at the DC.

Acknowledgements

First I would like to express my special appreciation and thanks to my advisor Professor Hossein Abouee Mehrizi. It has been an honor to be his first Ph.D student. His deep insights helped me at many stages of my research and I am grateful for all his advice, contributions and funding to make my Ph.D experience appealing and productive.

I would also like to express my sincere gratitude to the members of my defense committee: Steve Drekic, Fatih Safa Erenay, Qi-Ming He, and Fredrik Odegaard for their time, interest, and helpful comments.

I would like to thank all of my friends who supported and encouraged me to strive towards my goal. My time at the University of Waterloo was made enjoyable in large part due to the many friends and colleagues that became and will continue to be a part of my life.

Most importantly, a special thanks to my family. Words cannot express how grateful I am to my mother and father for all the sacrifices they have made, and for their unconditional and wholehearted support through all of my pursuits. I would not be here if it were not for you. I would also like to thank my brother and sister for all their love and encouragement.

Table of Contents

List of Figures

Chapter 1

Introduction

In daily life, we face many queueing systems with different fashions. For instance, in service systems, customers wait in line until they are served; however, in production systems, customers place an order at a retailer and wait until receiving that product. In these cases, customers gain a value by receiving the service or product and may incur a waiting cost. In practice, customers have strategic behavior wishing to maximize their own expected utility and may not join the system if they believe that the waiting time is too long. In this study, we analyze customers' behavior and system manager's strategy considering two different queueing systems, namely service and production systems, respectively.

1.1 Service Systems

Individuals who act in their self–interest to maximize their own welfare using a shared resource, may impose costs as negative externalities on each other. For instance, in a telecommunication system, individual user demands for bandwidth create congestion to the point that the network is busy and the speed of communications decreases to the point of being unusable. In queueing systems, customers by maximizing independently their own welfare, cause future customers to spend more time in the system resulting in externalities

in the form of delays. By ignoring these externalities, joining customers impose waiting cost on others and resulting in excessive congestion.

Customer's decision in queueing systems is associated with a trade-off between a reward and cost. Customers obtain a reward once they receive service, but there is an associated cost proportional to the time spent in queue. Some customers opt not to queue when they perceive a long wait time whereas others may abandon the queue after waiting for some time. Rational customers can join the queue or balk, and also can renege the queue when they feel frustrated from waiting too long. If all customers were to act in a way that maximizes the social welfare, fewer people would wait in line and more people would balk or renege. However, as a utility–maximizing individual, a customer does not consider the effect of her action on others when she decides to join and wait in a queue, i.e. her waiting lengthens the delay for others, thus increasing others' queueing costs. Under this premise, the more people that join a queue and consequently wait longer, the worse off everyone is. The question here is how to manage these queueing systems in order to achieve organizational goals such as greater efficiency, increased revenues, lower congestion collapse rate, fewer service interruptions and failure rates and so forth. We propose a mechanism to control the queue to achieve a profit equal to the maximum social welfare while the waiting cost of customers is non-linear and customers are strategic regarding balking and reneging from the queue.

1.1.1 Existence of Customer Behavioral Factors

Designing a control mechanism, we consider behavioral factors which influence customers' behavior while they are waiting in a queue. Psychological factors, which have not been covered in classical queueing models, govern customer behavior when they are waiting. Ignoring the psychological cost of waiting leads to inappropriate conclusions in service systems and may not be consistent with empirical results (Carmon et al. 1995).

For example, Osuna (1985) illustrates how stress and anxiety shape customer behavior in a waiting queue. Stress and anxiety result in aversiveness to the waiting time and lead to increasing marginal waiting costs. However, in predictable wait settings, when customers

anticipate that waiting is nearly over, they show less aversive behavior compared to the beginning of wait (Carmon and Kahneman 1996). It seems that having knowledge about the duration and ending time has a significant effect on how customers experience waiting time. Janakiraman et al. (2011) argue that two opposite forces can shape abandonment behavior; on one hand customers suffer from waiting in a queue and on the other, they commit to remain in the queue until the end. Customers define their own waiting and completion utility to capture these two effects. This commitment to remain in the queue has been studied by considering that in the presence of a goal and anticipation time, as the goal is closer to attainment, a customer is urged to stay and shows less aversive behavior with respect to the delay (Carmon and Kahneman 1996).

However, there is a distinct lack of research concerning how both waiting cost and anticipation time simultaneously affect customer behavior. Classical queueing models assume that waiting time has a linear effect on customer waiting cost and all customers have the same perception regarding waiting cost (Hassin and Haviv 2003). We argue that, in addition to waiting time, customer's anticipation about waiting time influences decision of joining or abandoning a queue. This situation happens in many real-world settings, such as downloading from free file hosting and sharing services (e.g., www.mediafire.com); before downloading a file from a hosting service, the system provides some information regarding estimated downloading time which results in customer's anticipation about delay.

1.1.2 Impact of Strategic Behavior and Behavior Factors

Since customers are impatient and may renege from the system after sometime, time lost has a significant effect on their behavior and shapes their reactions to a wait. We assume that customers have a non-linear waiting cost function which captures (1) wait aversiveness and (2) sensitivity with respect to the delay anticipation time. In this context, risk is defined as the loss of time for a customer. According to Guo and Zipkin (2007, 2009), different sensitivities toward risk can be associated with the shape of waiting cost function. Convex waiting cost function refers to strong aversion to the waiting time and concave cost function represents risk-seeking customers. In some cases, utility loss with respect to the

waiting time is not pure convex or concave, i.e. customers show a mixed attitude toward the risk.

This study also addresses the disconnect between empirical results and existing strategic behavior in queueing theory. Take for example the 2017 empirical study which found that patients exhibit a mixed-risk attitude toward a waiting cost when faced with long-term delays in accessing health care (Liu et al. 2017). Up to a certain point in time, individuals were found to be risk-averse with respect to quality of care and patiently wait to receive service. After the specified amount of time passed, they become risk-seeking with respect to quality of care. Patients tend to sacrifice care quality for time and want to receive their desired service as soon as possible. The concave-convex waiting cost function reflects the observed risk-seeking behavior relative to time (risk-averse in relation to care quality) prior to the anticipation point, and the risk averse behavior relative to time (risk-seeking in relation to care quality) after the anticipation point.

We assume that customers have a non-linear waiting cost function which captures both aversiveness of wait and sensitivity with respect to the anticipation point before reaching the point and after passing it. Moreover, our waiting cost function fully captures the diminishing aversiveness near the anticipation point as demonstrated by Janakiraman et al. (2011). In particular, we assume that customers have a target (hereafter referred to as a anticipation point) for their waiting time. Their waiting cost function is concave up to the anticipation point, indicating the marginal cost of their wait for service is not increasing. Their waiting cost function becomes convex after the anticipation point is reached, reflecting the fact that customers increasingly lose their patience after a certain amount of time has passed and their marginal cost of waiting increases.

We seek to contribute to the literature in customer strategic behavior, specifically in the nexus between mixed behavior and non-linear waiting cost functions. Customers intuitively assess the trade-off between the perceived value of their time and their perceived value of the service, and decide (1) whether to join the queue and (2) how long to wait before reneging. However, when everyone considers only their own interests, their actions have negative externalities on others, such as increasing waiting time. To this end we propose

a control mechanism to eliminate these effects caused by externalities among customers. The revenue maximizer wants to control the effects of externalities to maximize customers' surplus and by doing so gain maximum profit. We use an abandonment threshold policy to control the waiting time and a pricing policy to control the arrival rate. In such a case, customers know the abandonment threshold and if the waiting time for a customer exceeds this time threshold, they automatically renege the system (this customer is abandoned from the system by the planner). While this type of control policy may appear undesirable from an individual's perspective, as all customers are aware no one will wait longer than the abandonment threshold, this mechanism decreases the uncertainty of waiting, which in turn declines anxiety while waiting in the queue. We show that when using a pricing policy without such an abandonment threshold, the maximal profit is not guaranteed.

1.1.3 Related Research

This study bridges research streams on customer behavior and queueing control. We categorize the related literature into three streams: impatient customers, risk attitude, and queueing control.

Strategic Behavior of Impatient Customers

A number of studies address customer strategic behavior in queues. Naor's (1969) seminal work proposed a linear reward-cost structure that accounts for the possibility that customers may balk upon arrival and then characterizes the optimal joining threshold from individual, social and revenue maximizer perspectives. Optimal joining and balking strategies for a general arrival process with both single and multiple servers have also been investigated (Yechiali 1971 and 1972). Besides balking, customers may strategize regarding their abandonment while waiting in a queue. Abandonment was first captured by Barrer (1957) who introduces a queueing model with impatient customers whose patience is modeled using a threshold.

Hassin and Haviv (1995) examine the behavior of customers who decide on both the

balking probability and reneging time. Homogeneous customers trade off the service reward, which is dropped to zero after a time threshold, and a waiting cost represented as a linear function with respect to delay. In contrast, Mandelbaum and Shimkin (2000) consider that customers have some knowledge regarding the waiting time but are heterogeneous in their utility function and decide to abandon the queue strategically. While a linear waiting cost function results in trivial abandonment, not joining the queue at all or joining and never reneging, a non-linear waiting cost function leads to a non-trivial abandonment strategy (Haviv and Ritov 2001, Shimkin and Mandelbaum 2004). Shimkin and Mandelbaum (2004) show that a non-linear cost function leads to multiple equilibrium points for an abandonment threshold resulting in sub optimal decisions. We refer the reader to Mandelbaum and Zeltyn (2013), and Wang et al. (2010) for a comprehensive recent review dealing with impatient customers, and Hassin (2016) for a comprehensive review of the research on customer strategic behavior.

The above literature lays the groundwork for describing customer strategic behavior. Against this backdrop, another stream of literature explores how psychological factors influence a customer's decision to queue as well as her anticipation of the service or product. Researches have noted the effect of anticipation on customer behavior while waiting in queue and the relation between this anticipation and a customer's decision whether to renege the system (Kumar et al. 2014, Zohar et al. 2002, and Janakiraman 2011). We argue that, in addition to waiting time, a customer's anticipation regarding the wait time influences her decision of whether to join or abandon a queue. There is a distinct lack of research concerning how both waiting time and anticipation time simultaneously affect customer behavior. From the customer behavior point of view, we are the first that investigate customer decisions regarding balking and reneging from queue while considering both the anticipation time and impatient sensitivity.

Risk Attitude

A second stream of research related to our study is the examination of customer risk attitudes as they relate to strategic behavior in queueing systems. Guo and Zipkin (2009)

assume risk-aversive behavior and analyze the relationship between the value of information and customer characteristics. Sun and Li (2012) consider a queueing system with a single server, risk-neutral customers, and risk-averse customers. They analyze the joining and balking behavior of customers when partial information about the service time distribution is provided. Afèche et al. (2013) investigate lead-time dependent pricing with risk-neutral and risk-averse customers. Finkelstein et al. (2014) investigate patient behavior under an open-access scheduling appointment system. They show that risk-seeking and risk-averse patients place difference premiums on speedy access to care and exhibit different degrees of willingness to wait to see their own doctors. Liu et al. (2017) investigate patient behavior during the wait time in an emergency department. They show that whereas the waiting cost function for the males is convex, females exhibit mixed behavior with regard to the waiting delay and an S-shaped waiting cost function. Although a number of effective approaches exist to capture customer strategic behavior or risk attitude independently, few address both simultaneously.

Queueing Control

Against the backdrop of customer strategic behavior and risk attitude, we consider congestion control as a means to maximize social welfare or revenue. The implications of imposing tolls on customers as a queue control policy were introduced by Naor (1969). Since then a number of studies explored a variety of queue control policies (Adiri and Yechiali 1974, Hassin 1985 and 1995, Cachon and Feldman 2011, Afèche 2013, Hassin and Koshman 2015). For a comprehensive recent review of economic analyses of queueing systems with strategic customers we refer reader to Hassin (2016). Afèche et al. (2013) and Afèche and Sarhangian (2015) present techniques which are most closely related to ours. To achieve optimal social welfare and control the congestion in the system, the authors propose a pricing policy to decrease the negative externalities caused by individual self-interested decision makers. We propose two control policies to eliminate the negative effect of externalities in such queueing systems. We are the first to propose a joint pricing and abandonment threshold mechanism to control the waiting time and congestion in a

queueing system. We demonstrate how using this mechanism, a revenue maximizer can achieve the optimal expected revenue.

1.2 Production Systems

In production systems, customers gain a value when receiving a product and suffer a waiting cost when incurring a delay. However, customers who act strategically may not join the system and place an order if they believe that the waiting time is too long. In this study, we consider a supply chain including a single DC and a manufacturer that operates from a warehouse. The DC and the warehouse manage their inventories using a base-stock policy. Customers arrive at the DC and either choose to place an order and wait until receiving the product or balk the system without placing any order. All system parameters are common knowledge and customers receive no information regarding the inventory position at the DC. Customers are assumed to be homogeneous who strategically choose to place an order with probability p based on trade-off between the value gained and the waiting cost which is a non-linear function with respect to delay. We assume that customers' waiting cost depends on their degree of *risk aversion* toward delay. This non-linear waiting cost function is convex in the waiting time and reflects the case that the waiting time is increasingly unattractive (Ata and Olsen 2009). We assume that the DC has complete information about customers' characteristics and plays as a Stackelberg leader. The DC manages its expected revenue by controlling the base-stock level and imposing a price. We investigate the effect of the shape of the waiting cost function on the supply chain policy.

1.2.1 Related Research

We categorize the related literature into two streams -inventory system, and attitude toward risk- based on the components studied in strategic supply chain management.

Multi-Echelon Production Inventory System

Multi-echelon inventory system with stochastic demand has been studied by many researchers in the last decades: see for example, Federgruen (1993), Zipkin (2000), Simchi-Levi and Zhao (2007) and Wang (2011), and references therein. Meanwhile, many researchers consider production facilities as well as inventory systems in analyzing supply chain performance. He et al. (2002) analyze a production inventory system with one warehouse and a production workshop with no lead-time. They use Markovian decision process to find the optimal replenishment policy. Abouee-Mehrizi et al. (2011) consider a manufacturer operating from a warehouse and multiple DCs with positive transportation time between the manufacturer and the DCs using the base-stock policy in both the warehouse and the DCs. They use the Flow-Unit method introduced by Axsäter (1990) to find the exact cost for a joint production inventory problem. This paper is extended by Zare et al. (2017) by considering an (R, Q) inventory policy at the DC in a two-echelon production inventory system with a single manufacturer and a DC. They assume that the manufacturer operates from a warehouse using a base-stock policy to manage its inventory. They find the optimal reordering point at the DC and develop a two-phase heuristic to approximate the optimal inventory level at the warehouse as well as the optimal batch order size at the DC. We consider a production inventory system with stochastic demand where both DC and the production facility use the base-stock policy to manage their inventories. We assume that there is a positive transportation time between the warehouse and the DC. We use the Flow-Unit method applied by Abouee-Mehrizi et al. (2011) and Zare et al. (2017) to derive the average inventory holding at the DC.

Risk Attitude

In recent decades, analysis of customer risk attitude toward delay in queueing systems have been considered in the literature of operation management. A common assumption in the literature is that customers are not risk neutral with respect to lead-time uncertainty and investigate the effect of delay cost function on firm's policies (see, e.g., Kumar and Randhawa 2010, and Afèche et al. 2013). Ata and Olsen (2009) consider convex, concave

and convex-concave delay cost functions to capture sensitivity to deadline in a system in which the manager quotes lead times (and prices) to each arriving customer. In this setting, a convex-concave curve refers to a situation where customers become increasingly impatient when an acceptable deadline passes. Akan et al. (2012) consider the similar system but for multiple class customers assuming that the delay cost function has a convex-concave curve.

In our study, we consider a convex waiting cost function to capture delay sensitivity of customers (risk aversion), and investigate how the strategic and risk aversive behavior of consumers affects the supply chain manager's decision and the system equilibrium.

Risk aversion in inventory management have been also studied in the literature. Berman and Schnabel (1986) were among the first to study the impact of risk aversion on the order quantity. They considered a mean-variance utility function in the news vendor problem. We refer readers to Tekin and Ozekici (2015), Agrawal and Seshadri (2000), Chen et al. (2007) and Afèche and Sanajian (2013) for comprehensive reviews in this field. However, we assume that instead of the system manager, customers are risk averse and make decisions considering a trade-off between the value of the product and the cost of waiting. We consider a centralized supply chain decision maker which can control the base-stock level at the DC and charge a price to make profit. We show that for any level of risk aversion, there is an optimal policy, i.e. optimal price and base-stock level, for the DC.

Chapter 2

Control Mechanisms in Queues with Joint Balking and Reneging Strategy

We consider an unobservable queueing system with a single server in which customers arrive to the system according to a Poisson process with the rate of λ and are served based on a first-come first-served (FCFS) policy. Also, service times are exponentially distributed with the rate of μ. We assume that customers are homogeneous and decide (1) whether to join the queue and (2) how long to wait before reneging. When a new customer arrives to the system, she decides to join the system with probability p or to balk the system with probability $1-p$. Each customer is rewarded a value V once served, however, she exhibits *mixed-risk behavior* with respect to the delay and changes her behavior from risk-seeking to risk-averse after spending a certain amount of time in the system. We assume that customer utility is a value (reward) for receiving the service less a waiting cost. Customers are forward-looking decision makers and decide whether to join a queue and when to renege based on their expected utility function, $E[U(x)]$, where $U(x) = V - C(x)$ is linear with respect to the service value V and non-linear with respect to the waiting cost $C(x)$,

$$C(x) = c((x-\beta)^{2\alpha-1} + \beta^{2\alpha-1}). \tag{2.1}$$

Note that x is the waiting time in the system, including the service time, α is a positive integer greater than 1, β is a non-negative number, and c is a positive constant. Both the

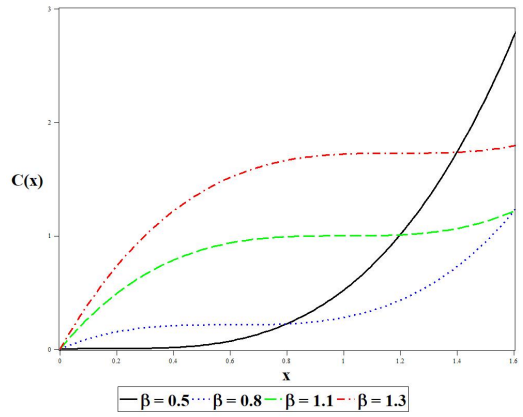

Figure 2.1: The behavior of cost function with different anticipation points (β), $c = 1$ and $\alpha = 2$.

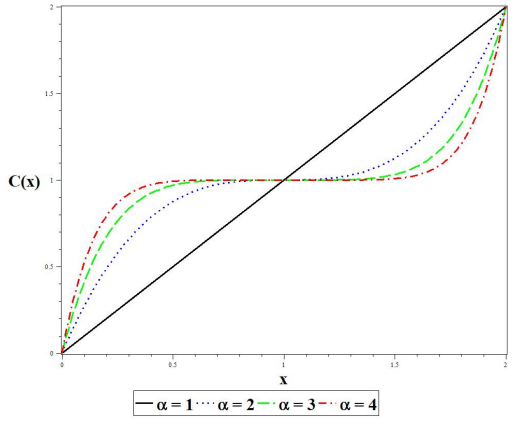

Figure 2.2: The behavior of cost function with different degrees of risk aversion (α), $c = 1$ and $\beta = 1$.

anticipation point (β) and risk aversion degree (α) shape customer behavior and determine the customer's attitude toward the delay. The utility function, $U(x)$, captures two customer behavioral characteristics. First, customers have an impatience intensity (sensitivity) with respect to any delay with a level of α. Second, customers have an anticipation point β after which they become more impatient regarding any further delay. Recall that β refers to the anticipation point, which is the point in time when a customer's waiting cost function changes from concave (willing to wait) to convex (increasingly impatient). Figure 2.1 illustrates the customer waiting cost function with respect to the delay with different values of β.

Consider the impact of the anticipation point on $U(x)$. As $\beta \to 0^+$, customers waiting in the queue become more sensitive to further delays, and the marginal waiting cost increases. That is, customers tend to be risk-averse with respect to time (placing greater value on their time than on receiving service). These customers place a high value on their time and tend to renege early, meaning they are unwilling to wait long for the service once they are in a queue. However, as β approaches infinity, customers are patient exhibiting a risk-seeking behavior. These customers are less sensitive to small increases in delay and place a high value on the service compare to the time. Finally, when $\beta > 0$ with a moderate value,

customers adopt a mixed–risk behavior. The anticipation point measures the customer's attitude toward delay by reflecting their anticipation/impatience with respect to receiving a certain service.

As shown in Figure 2.2, according to (2.1), around the anticipation point β, the cost function appears to be flat, indicating that customers are indifferent to incremental changes in waiting time (Carmon and Kahneman 1995). At this point, customers change their attitude from risk-seeking to risk-aversion, reflecting a growing irritation with further delays. Such behavior has been observed in patients waiting to be seen by a physician. Liu et al. (2017) observe that a patient's utility-loss function *plateaus* in the middle of their delay.

Parameter α is a positive integer and denotes the customer's degree of risk aversion (impatience) with respect to the waiting time. If $\alpha = 1$, (as shown in Figure 2.2) customers face a constant marginal waiting cost, c, exhibiting indifference to the amount of time they wait in the system. In other words, when $\alpha = 1$, customers are risk-neutral with respect to the waiting cost. For $\alpha > 1$, the marginal waiting cost depends on the time spent by a customer in the system (Figure 2.2).

We assume that each customer has a deterministic reneging time policy; i.e., after joining the system, the customer may renege the system at any deterministic time τ before her service is complete. Each customer individually determines her reneging time considering the trade-off between the waiting cost and the value obtained from service completion. All system parameters are common knowledge and customers do not receive any information regarding the queue status and their position in the queue. This reneging time is highly related to the perceived waiting time by an individual customer. We call this perceived waiting time as *offered waiting time*. An individual customer's reneging policy changes the offered waiting time of other customers and as a result changes their policies as well. Thus, the best response for an individual customer as a forward-looking decision maker will be affected by actions of others. For instance, an individual's decision may vary when all other customers are fully patient (never renege the system) or fully impatient (never join the system). In the former case, the best response for such a customer can be reneging earlier; however, in the latter case, they may do better staying in the system until service

is completed.

One can investigate the equilibrium reneging time in this system; however, if the waiting cost is not a linear function in time, the uniqueness of the equilibrium point cannot be guaranteed (e.g., see Shimkin and Mandelbaum 2004). We do not aim to characterize the equilibria of the system, rather our main focus is on finding an optimal queue control policy to eliminate customer externalities when they make decision strategically. It means that we are interested in characterizing the socially optimal behavior and proposing an optimal policy from a revenue maximizer perspective where in such circumstances customers choose that policy in the equilibrium. To find that policy, we first analyze an individual customer's reneging behavior under an offered waiting time. Therefore, we assume that all other customers follow the same T strategy (they renege at $\tau = T$), and investigate the decision of an arriving customer.

The probability density function (PDF) of the waiting time (offered waiting time), including the service time, of a customer in the system who joins the queue and does not renege when all other customers renege after T time units (assuming that $\mu > \lambda p$), are obtained as (e.g., see Hassin and Haviv 1995 and Haviv and Ritov 2001):

$$g_T^p(x) = \begin{cases} P_0 \mu e^{-(\mu - \lambda p)x}, & 0 \le x \le T, \\ P_0 \mu e^{-\mu x + \lambda p T}, & T \le x, \end{cases} \quad , \tag{2.2}$$

where,

$$P_0 = \frac{\mu - \lambda\, p}{\mu - \lambda p e^{-(\mu - \lambda p)T}}. \tag{2.3}$$

Also $G_T^p(x)$ is the corresponding cumulative distribution function. Consider an individual customer who decides to renege the system after τ time units, when all other customers renege after T time units. For such a customer, the expected utility which is associated with the strategy τ as a response against all other customers using strategy T, is obtained as:

$$U_T\,(p,\tau) = E[V\ \mathbf{1}\{X < \tau\} - C(X \wedge \tau)] = \int_0^{\tau} V\ dG_T^p(x) - \int_0^{\infty} C(\min\{x,\tau\}) dG_T^p(x) =$$

$$\int_0^{\tau} (V - \ C(x)) \ g_T^p(x)\, dx - \ C(\tau)(1 - G_T^p(\tau)). \tag{2.4}$$

The expected utility function is composed of two terms. The first term is the expected utility of joining a queue and being served before spending τ time units in the system. The second term is the expected utility of joining the queue and reneging the system without receiving or completing the service. We discuss the details on the customer response function in the following section.

2.1 Preliminary Analysis

In this section we investigate local maximum points of an individual customer's expected utility with respect to her reneging time. We show that the local extremum points of the customer's expected utility function occur where the hazard-rate of the waiting time equals to the cost-reward ratio (similar to Shimkin and Mandelbaum 2004). Let $h_T(x) = \frac{g_T^p(x)}{1-G_T^p(x)}$ and $\gamma(x) = \frac{C'(x)}{V}$ denote the hazard rate function associated with the offered waiting time distribution given by (2.2), and cost-reward ratio function respectively. Using (2.1) and (2.2), and considering the offered waiting time distribution, the hazard rate and cost-reward ratio are obtained as

$$h_T(x) = \begin{cases} \frac{\mu\,(-\lambda\, p+\mu)}{\mu-\lambda\, p\mathrm{e}^{-(\mu-\lambda\, p)(T-x)}}, & x \le T \\ \mu, & T \le x \end{cases},$$

$$\gamma(x) = \frac{c\,(x-\beta)^{2\,\alpha-2}\,(2\,\alpha-1)}{V}. \tag{2.5}$$

Proposition 1 *A new arriving customer, who joins the queue, will renege at time τ which is one of the local extremum points of $U_T(p,\tau)$ where*

$$h_T(\tau) = \gamma(\tau). \tag{2.6}$$

A candidate reneging time τ is a local maximum point of the expected utility function if $h_T(\tau) - \gamma(\tau)$ changes its sign from positive to negative at τ. When the expected utility

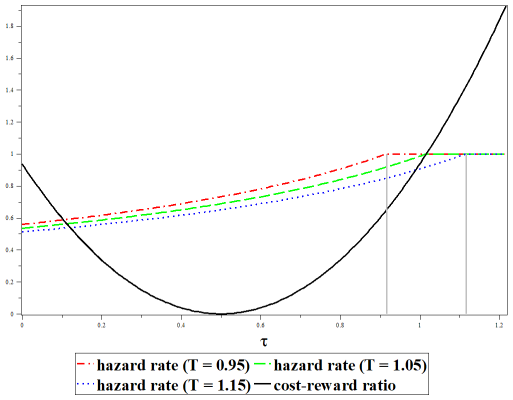

Figure 2.3: Hazard rate and cost-reward ratio functions with respect to the reneging time.

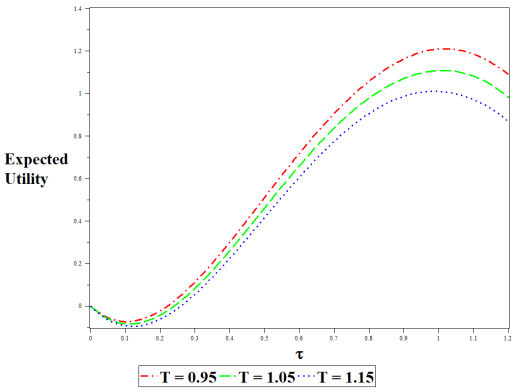

Figure 2.4: Expected utility function with respect to the reneging time.

function is unimodal, it has one unique maximum point, which we refer to as a global point. However, since the waiting cost function is non-linear with respect to delay, the expected utility function may have several local extremum points (see Figures 2.4 and 2.6). In Proposition 2 we will prove that the customer's expected utility function has at most two local maximum points; as shown in Figures 2.3 and 2.5, the hazard rate function crosses the cost-reward ratio function at more than one point (the corresponding expected utility functions are illustrated in Figures 2.4 and 2.6, respectively).

Also, an arriving customer's reneging time depends on other customers' action. Figure 2.3 illustrates the behavior of the hazard rate function for different values of T (resulting in different offered waiting times) with the same cost-reward ratio function and Figure 2.4 shows their corresponding expected utility functions. As shown in Figure 2.3, the best response for an arriving individual customer when all other customers renege at time T, is reneging before T (blue line with $T = 1.15$), after T (red line with $T = 0.95$) or equal to T (green line with $T = 1.05$). Considering this structure of the expected utility function, we discuss an individual customer's decision in the following section.

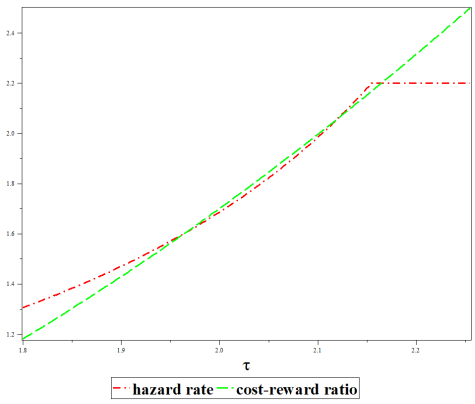

Figure 2.5: Hazard rate and cost-reward ratio functions with respect to the reneging time.

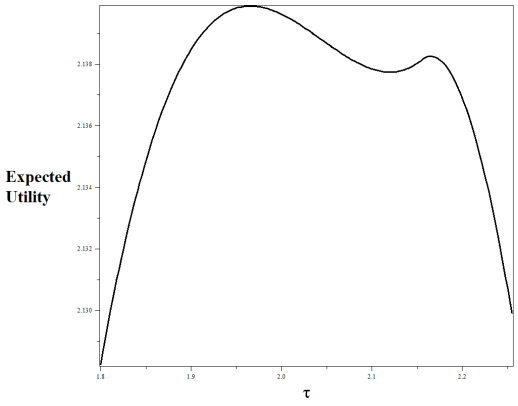

Figure 2.6: Expected utility function with respect to the reneging time.

2.1.1 Self-Maximization Strategy

We investigate the reneging action of an individual customer given an offered waiting time, i.e., we want to find the best response for an individual customer when all other customers use the strategy T. First, we discuss the worst and best cases which a joining customer may face. In the worst case scenario, all other customers never renege and stay in the system until their service ends; therefore, according to (2.5) the hazard rate function is $\lim_{T\to\infty} h_T(x) = \mu - \lambda p$ and the optimal action for an arriving individual is to renege at (and only at) τ_0, satisfying $\gamma(\tau_0) = \mu - \lambda p$. However in the latter case, when all other customers are extremely impatient and never join the system, the hazard rate function is $\lim_{T\to 0} h_T(x) = \mu$ and the best response for this customer is to renege only at τ_1 when $\gamma(\tau_1) = \mu$. We can conclude that in the general case, the best response for a customer is a time between $\tau_0 \leq \tau \leq \tau_1$:

Corollary 1 *Consider the discussed queueing system. Then, the maximum abandonment threshold for an arriving customer, τ_1, is given by:*

$$\tau_1 = \left(\frac{\mu V}{c\,(2\,\alpha - 1)}\right)^{(2\,\alpha-2)^{-1}} + \beta. \tag{2.7}$$

Note that since the cost-reward ratio is a convex function with a minimum at $x = \beta$ (see (2.5)), $\gamma(0) \geq \mu$ is not a sufficient condition to renege at $x = 0$. Next, we investigate the local maximum points of the customer's expected utility with respect to time spent in the system. We define two functions $\bar{V}_1(\beta, p, T)$ and $\bar{V}_2(\beta, T)$, to characterize the shape of the customer's expected utility with respect to her reneging time. Let

$$\bar{V}_1(\beta, p, T) = \frac{\left(\mu - \lambda p \mathrm{e}^{-(\mu-\lambda p)T}\right) c\,(2\,\alpha - 1)\,\beta^{2\,\alpha-2}}{\mu\,(\mu - \lambda\,p)}, \quad \bar{V}_2(\beta, T) = \frac{c\,(2\,\alpha - 1)\,(T-\beta)^{2\,\alpha-2}}{\mu}. \tag{2.8}$$

Proposition 2 *For a given joining probability p, if all other customers follow the T strategy (reneging at T), the customer's expected utility has the following structure with respect to the reneging time $\tau \leq \tau_1$:*

1. *For $\tau \leq \beta$: if $V < \bar{V}_1(\beta, p, T)$, the expected utility has a unique minimum point less than β; otherwise, for $V \geq \bar{V}_1(\beta, p, T)$ the expected utility function is monotone increasing in τ.*

2. *For $\tau > \beta$: if $V < \bar{V}_2(\beta, T)$ the expected utility is unimodal with a unique maximum point less than T; otherwise for $V \geq \bar{V}_2(\beta, T)$ there exists either one or two positive local maximum points where the first one is less than T and the other one is greater than T and equal to τ_1.*

According to Proposition 2, the expected utility function for an individual customer given an offered waiting time has at least one maximum point. Figures 2.7 and 2.8 illustrate the structure of the customer's expected utility function with respect to the anticipation point and other customers' action through reneging time T, respectively. The red (dashed curve) and blue (dotted curve) curves represent $\bar{V}_1(\beta, p, T)$ and $\bar{V}_2(\beta, T)$, respectively. We offer the following intuition. Considering the psychological cost of waiting, Proposition 2 explains the challenges of opposing responses in waiting behavior when a goal is anticipated. As illustrated in Figure 2.7, the combination of the service value and anticipation point

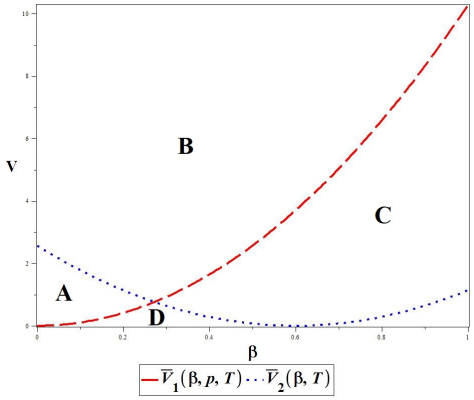

Figure 2.7: Segmentation of local maximum points of the customer's expected utility with respect to β.

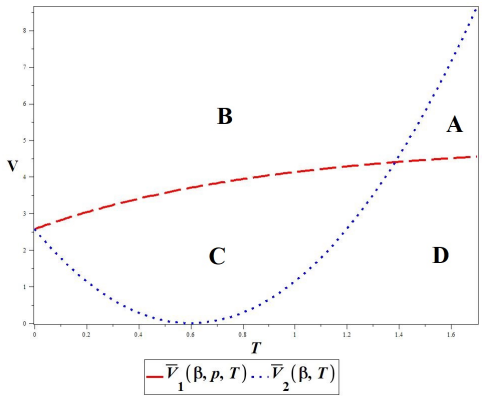

Figure 2.8: Segmentation of local maximum points of the customer's expected utility with respect to T.

shapes an individual customer abandonment decision. When this customer can only decide on their reneging time (no joining decision) two opposing psychic forces, namely displeasure of waiting and commitment to wait, shape customer behavior in waiting (Janakiraman et al. 2011). While the anticipation point and the service value is low, this customer chooses the only existing maximum point as the reneging time where this reneging time is always less than T and τ_1 (Regions A and D). However, by increasing the service value as a motivational factor, two local maximum points emerge to distinguish reneging policy. In this case, this customer may choose a reneging time greater than all other customers' choices and renege at the maximum reneging τ_1 (Region B). Observe that with higher values of V, when the anticipation point (β) increases and tends to be high, although the expected utility function is decreasing and negative at the beginning of waiting, this customer commits to wait and chooses a reneging time greater than T (Region C).

Other customers' actions also affect the individual customer's decision regarding the reneging time. In Figure 2.8, Regions A, B, C and D with the discussed properties characterize the individual customer's expected utility with respect to others' actions (reneging after spending T units of time). Observe that as the others' abandonment threshold T increases, an individual customer tends to choose a reneging time less than T (Regions A

and D in Figure 2.8).

Corollary 2 *For a given joining probability p, if all other customers choose T as their reneging time, an arriving customer's reneging time τ has the following structure:*

1. *If $T > \tau_1$, there exists a unique positive maximum point which is less than τ_1.*
2. *If $T \leq \tau_1$, there exists one ($\tau = \tau_1$) or two positive local maximum points, one is less than T and the other one is $\tau = \tau_1$.*

Note that according to Proposition 2, $\tau = 0$ is also a candidate for the reneging time when $V < \bar{V}_1(\beta, p, T)$. Therefore, if for all cases provided in Corollary 2, the customer's expected utility function is negative, then the best response for this customer is reneging at time $\tau = 0$; otherwise when $V \geq \bar{V}_1(\beta, p, T)$, there always exists a positive optimal reneging time for this customer.

Since customers are rational and self-interested, they may not follow a strategy which is beneficial for all. For example, when the waiting cost function is linear, the equilibrium reneging time is a mix between not joining or joining and staying without reneging (see Hassin and Haviv 1995, and Mandelbaum and Shimkin 2000). In this case, self-maximization leads to a more congested system and results in imposing waiting costs, as negative externalities, on the rest of the society. However, if all customers commit to join less often and also renege in a finite time, such a situation will be more beneficial for all. Next, we characterize socially optimal behavior and then propose a mechanism to induce socially optimal behavior.

2.2 Socially Optimal Behavior

In this section, we characterize the structure of the joint optimal balking and abandonment strategy that maximizes the social welfare, total expected net benefit of the people in the society, denoted by $W(p, T)$ and given by $W(p, T) = \lambda p U_T(p, T)$. We first prove the

existence and uniqueness of such an optimal reneging time for any given joining probability. We then extend the results to the joint optimal balking and abandonment strategy and find the maximum achievable social welfare.

We are interested in characterizing an optimal reneging time T^* which maximizes the aggregate expected utility functions assuming that all customers will follow the same reneging time T. Since the social welfare function may have two local maximum points with respect to the abandonment threshold T (considering $T = 0$ as a candidate), we refer to type-1 (denoted by $\bar{T}_1$) and type-2 (denoted by $\bar{T}_2$) as the first (the smallest time T at which the utility function is not strictly increasing, i.e., $\frac{\partial U_T(p,T)}{\partial T} \leq 0$), and second local maximum points, respectively. In the following proposition, we characterize the structure of the optimal abandonment strategy for any given p using three functions $V_1(\beta, p)$, $V_2(\beta)$ and $V_3(\beta, p)$ which are defined as follows:

$$V_1(\beta,p) = \frac{\lambda\, p c \beta^{2\,\alpha-1}\left(1 - M\left(1,\, 2\,\alpha,\, -s\right)\right)}{\mu - \lambda p}\,, \quad V_2(\beta) = \frac{c\left(2\,\alpha - 1\right)\beta^{2\,\alpha-2}}{\mu},$$

$$V_3(\beta,p) =$$

$$c\left(\frac{\beta}{s}\right)^{2\,\alpha-1}\left(\frac{k^{2\,\alpha-2}\left(\lambda\, p\left(k-1\right)\mathrm{e}^{k-s} + \mu\right)}{\mu} - \frac{\lambda\, p \beta\, \mathrm{e}^{-s}\left(\Gamma\left(2\,\alpha, -s\right) - \Gamma\left(2\,\alpha, -k\right)\right)}{s}\right), \quad (2.9)$$

where

$$k = \frac{2\left(\alpha - 1\right)\left(\mu - \lambda\, p\right)}{\lambda\, p}, \quad s = \beta\,\left(\mu - \lambda\, p\right),$$

and $M(a, b, z)$ and $\Gamma\left(a, z\right)$ are the confluent hypergeometric function of the first kind (given by (A.18)) and incomplete Gamma function (given by (A.21)) respectively (see e.g., Abramowitz and Stegun 1964).

Proposition 3 *Consider the expected utility function given in (2.4) and the social welfare function.*

1. *For a given p,*

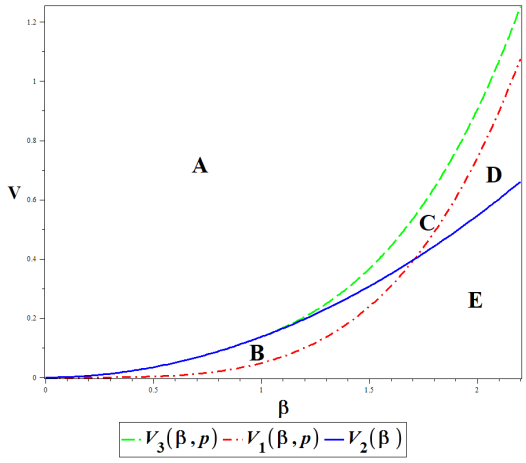

Figure 2.9: Segmentation of the local and global maximum points of the social welfare for given p.

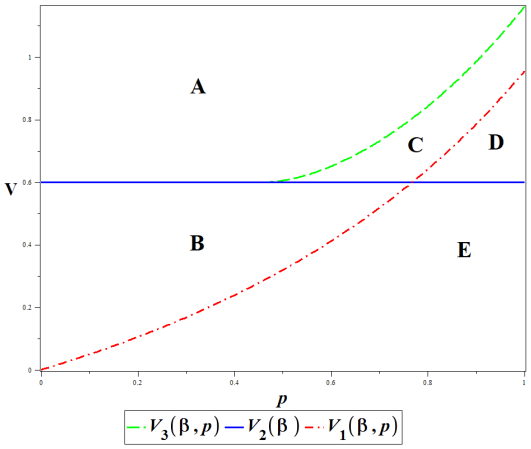

Figure 2.10: Segmentation of the local and global maximum points of the social welfare for given β.

(a) *If* $\beta \leq 2(\alpha-1)/\lambda p$, *for* $V \geq V_2(\beta)$, *the social welfare function is unimodal; for* $V_1(\beta,p) < V < V_2(\beta)$, *it is bimodal* ($\bar{T}_1$ *is zero) and for* $V_1(\beta,p) \geq V$, *the social welfare is always equal to zero (the expected utility function is always negative).*

(b) *If* $\beta \geq 2(\alpha-1)/\lambda p$, *for* $V \geq V_3(\beta,p)$, *or when* $V_2(\beta) < V \leq V_1(\beta,p)$, *the social welfare function is unimodal; for* $V_2(\beta,p) < V < V_3(\beta,p)$ *and* $V > V_1(\beta,p)$, *it is bimodal and two positive local maximum points* $\bar{T}_1$ *and* $\bar{T}_2$ *exist, and for* $V_1(\beta,p) < V \leq V_2(\beta)$, *we have* $\bar{T}_1 = 0$. *If* $V_1(\beta,p) \geq V$ *and* $V_2(\beta) > V$, *the social welfare is always equal to zero (the expected utility function is always negative).*

2. $V_1(\beta,p)$ *and* $V_3(\beta,p)$ *are increasing in* p *and* β. *Also* $V_2(\beta)$ *is monotone increasing in* β.

Figure 2.9 illustrates the threshold structure of the optimal reneging time for a given joining probability. The vertical and horizontal axes indicate the value V and anticipation point β, respectively. The red, blue, and green curves in Figure 2.9 represent $V_1(\beta,p)$, $V_2(\beta)$, and $V_3(\beta,p)$, respectively. Considering that all customers follow a (p,T) strategy,

Proposition 3 demonstrates that in region A, the social welfare function is unimodal and there exists a unique global optimal abandonment time. In contrast, when the expected utility function is negative for all values of T, for a given balking probability p, no maximum point exists for Region E in Figure 2.9. Regions B and C correspond to the situations in which the social welfare function is bimodal. In Region B, there is one minimum and two local maximum points for T ($\bar{T}_1$ is zero). Conversely, in Region C, both type-1 and type-2 points are positive numbers and the social welfare function is bimodal. Finally, in Region D, the type-1 and type-2 points coincide, which again leads to a unimodal social welfare function.

Next, we analyze the social welfare function when the joining probability is endogenous. First, we need to elaborate on the structure of Proposition 3 by considering the optimal reneging time structure for different values of p, the joining probability.

Corollary 3 *As the joining probability varies:*

1. *For $p \geq \frac{2(\alpha-1)}{\lambda\beta}$, we have $V_3(\beta,p) \geq V_2(\beta)$ and $V_3(\beta,p) > V_1(\beta,p)$. Also, thresholds $V_2(\beta)$ and $V_1(\beta,p)$ intersect each other at probability p_e where*

$$p_e = \frac{\mu(2\alpha-1)}{\lambda\mu\beta(1-M\left(1,\,2\,\alpha,\,-\left(\mu-\lambda\,p\right)\beta\right))+\lambda(2\alpha-1)}.$$

2. *If $V \leq V_2(\beta)$, as p increases, the optimal reneging time goes from a positive type-2 point (type-1 point is equal to 0) ($V > V_1(\beta,p)$) to no feasible point ($V \leq V_1(\beta,p)$). If $V > V_2(\beta)$, by increasing the joining probability,the social welfare function first moves from having one unique global point ($V \geq V_3(\beta,p)$ to having both positive type-1 and type-2 points ($\max(V_2(\beta),V_1(\beta,p)) < V < V_3(\beta,p)$); then, it returns to the unique global region ($V_2(\beta) < V \leq V_1(\beta,p)$).*

Figure 2.10 illustrates the structure of the optimal reneging time with different values of the joining probability. In Figure 2.10, the red, blue, and green lines represent $V_1(\beta,p)$, $V_2(\beta)$ and $V_3(\beta,p)$, respectively. From Corollary 3, we can also infer that when $V \geq$

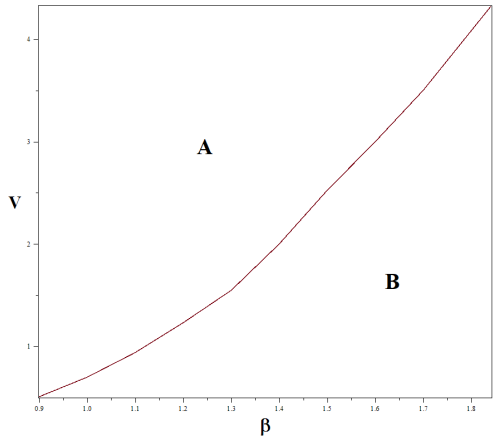

Figure 2.11: Structure of the socially optimal strategy.

$\max(V_2(\beta), V_3(\beta, p))$ (Region A in Figure 2.10) type-1 and type-2 points coincide, resulting in one unique global optima reneging time. In Region C of Figure 2.10, when $V_3(\beta, p) > V > V_1(\beta, p)$, there exist both positive type-1 and type-2 points; in this case, as p increases, the type-2 point tends to coincide with the type-1 point. When $V_1(\beta, p) \geq V > V_2(\beta)$, there is one unique global optimal reneging time and the type-2 point coincides with the type-1 point (Region D).

We next show that the optimal reneging time can be either greater or less than the anticipation point β depending on the value of the service.

Corollary 4 *Let T^* be the optimal reneging time.*

1. *For $V \geq V_2(\beta)$ and $V \geq V_3(\beta, p)$ the optimal reneging time is greater than the anticipation point, $T^* > \beta$.*

2. *For $V \leq V_1(\beta, p)$ and $V > V_2(\beta)$, the optimal reneging time is bounded by $T^* \leq \beta - 2(\alpha - 1)/\lambda p$.*

Corollary 4 provides bounds for the optimal reneging time for different values of V. One interesting observation is that depending on the value of the service, the optimal reneging time can be either less or greater than the anticipation point, β.

Next we demonstrate that there exists an optimal behavior in which the maximum value of the social welfare is secured.

Theorem 1 *There exists an optimal* (p^*, T^*) *strategy that maximizes the social welfare. Also, if* $V < \min(V_1(\beta, 1), V_2(\beta))$*, we can guarantee that the partial join strategy, i.e.,* $p^* < 1$*, is optimal.*

Figure 2.11 illustrates the implications of Theorem 1. There always exits an optimal abandonment strategy while in Region A, the fully join , i.e., $p^* = 1$, is optimal, and in Region B the partially joining strategy is the optimal strategy.

Theorem 1 defines and characterizes the socially optimal strategy. In the next section, we investigate how a revenue maximizer service provider should design a mechanism to guarantee a profit equal to the maximum achievable social welfare (the upper bound of the total profit).

2.3 Mechanism Design Problem

According to the Tragedy of the Commons, when customers use a common resource and maximize their own utilities, the equilibrium behavior leads to excessive use of the resource. In the context of a queue, an arriving customer who maximizes her own benefit imposes negative externalities on others. Such self-optimization induces excessive congestion which needs to be regulated (see e.g., Hassin and Haviv 2003). A server can achieve social optimality as the upper bound on profit if two conditions are satisfied: (i) the socially optimal behavior is not changed and (ii) total surplus (sum of all customers' surplus) is fully extracted (see e.g., Erlichman and Hassin 2015). The main question is whether the socially optimal behavior is achievable in the equilibrium by using an appropriate policy on the system.

We assume that the planner is interested in designing an appropriate mechanism inducing socially optimal behavior under which no one will deviate from the induced equilibrium.

In such a case, the planner knows that customers are self-interested and under any policy, they choose their own strategy (p, τ) ignoring the negative externalities on others. The planner must design a mechanism to control these externalities inducing socially optimal behavior.

2.3.1 Entrance-Fee/Service-Fee Mechanism

One known way to control congestion is by imposing an appropriate fee for service (Naor 1969). Because customers can renege from the system after a threshold, we consider two types of fees: an *entrance fee*, denoted by θ_e and a *service fee*, denoted by θ_s. The entrance fee is charged when customers join the queue, whereas the service fee is charged after they use the service. The expected revenue functions of the two pricing scenarios, entrance and service fees, can be defined as $\Phi_e(p, T, \theta_e) = \lambda p \theta_e$ and $\Phi_s(p, T, \theta_s) = \lambda p G_T^p(T)\theta_s$, respectively. Therefore, the total provider's expected revenue under this mechanism is $\lambda p(\theta_e + G_T^p(T)\theta_s)$.

Proposition 4 *Using a joint entrance-fee/service-fee mechanism,*

1. *There is no guarantee that the revenue maximizer can induce socially optimal behavior.*

2. *If the following conditions hold, a joint entrance-fee/service-fee mechanism can induce the socially optimal behavior and the maximum profit is achievable.*

   - *Condition 1:* $\gamma(T^*) < \mu - \frac{\lambda p}{VP_0} C(T^*)$.
   - *Condition 2:* $\frac{C'(T^*)}{V\mu} h_{T^*}(t) = \gamma(t)$ *must have no root for* $\beta \leq t < T^*$.

3. *Under conditions 1 and 2 above, the optimal entrance and service fees are obtained as:*

   $\theta_e = \frac{VP_0}{\rho}(1 - \frac{\gamma(T^*)}{\mu}) - C(T^*), \quad \theta_s = V(1 - \frac{\gamma(T^*)}{\mu})$.

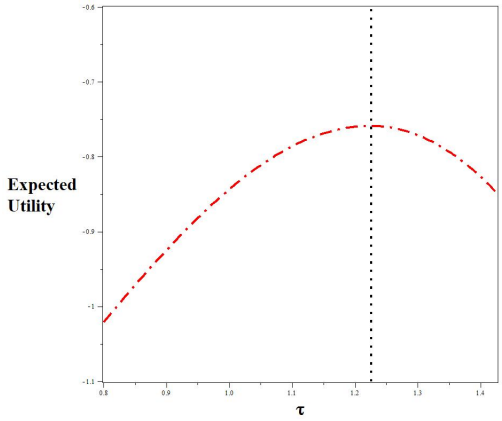

Figure 2.12: Individual customer's expected utility function when the first condition does not hold.

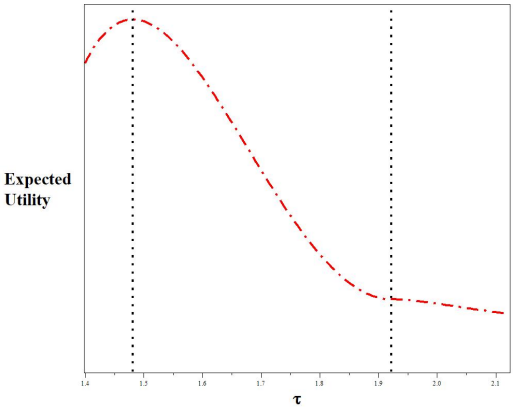

Figure 2.13: Individual customer's expected utility function when the second condition does not hold.

Proposition 4 demonstrates that designing a mechanism that is only based on the entrance and service fees may not be able to induce the socially optimal behavior in the system. The reason is that when the waiting cost function is non-linear, imposing a service fee cannot completely control the customer reneging strategy. To induce socially optimal behavior, the pricing mechanism should ensure that the optimal reneging time of an individual customer coincides with the abandonment threshold which maximizes the social welfare (Condition 2), and the individual customer's expected utility function is not negative at this threshold (Condition 1). However, in many cases this policy leads to a sub-optimal throughput. For instance, if the first condition does not hold, as shown in Figure 2.12, the pricing policy always causes a negative expected utility function for customers and as a result it discourages customers to join the system. In such a case, the proposed mechanism can not induce the socially optimal behavior. When the second condition does not hold, the proposed pricing policy can not induce the socially optimal reneging strategy; as shown in Figure 2.13. The second local maximum point in Figure 2.13 ($T^* = 1.92$), indicates the socially optimal reneging time; however, an arriving customer will choose a reneging time (the first local maximum point shown in Figure 2.13) less than the optimal one resulting the failure of proposed mechanism. Thus, we conclude that this pricing policy

is insufficient to induce optimal joining/reneging behavior of self-optimizing customers.

Next, we propose a joint entrance-fee/abandonment-threshold mechanism which induces the socially optimal behavior and guarantees a profit equal to the maximum social welfare.

2.3.2 Entrance-Fee/Abandonment-Threshold Mechanism

We first analyze how the planner's chosen abandonment threshold T^* affects customers' reaction and how customers choose their own strategy (p, τ) considering this policy. We show that this abandonment threshold policy can efficiently control customers' reneging time and induce the optimal reneging behavior. However, since there is no control on the joining probability, it can not guarantee the optimal joint decision, (p^*, T^*). Therefore, we extend the mechanism to charge an entrance fee so that the firm can control the effective arrival rate and also fully extract total surplus (sum of all customers' surplus) and achieve the profit equal to the maximum social welfare the system can generate.

Considering the effect of the joining probability on the abandonment threshold, we introduce an abandonment threshold policy for the planner to control customers' reneging behavior. First, we show that for a given joining probability, the planner choice will be the unique Nash equilibrium for all customers. Then, we extend our results to when the joining probability is endogenous. We prove that all customers will follow this policy and the optimal abandonment threshold coincides with the equilibrium customer's reneging time.

Proposition 5 *For a given p, using the abandonment threshold policy, reneging at T^* is the unique Nash equilibrium for all customers.*

Proposition 5 demonstrates that, if the planner chooses the optimal abandonment threshold equal to T^*, all customers will end up reneging at T^* and this point is the unique Nash equilibrium for all customers under this policy. This means that no one will deviate from this abandonment time by choosing a reneging time less than T^*.

Next, we consider the endogenous joining probability case which means that customers choose their joining strategy according to this abandonment threshold. We want to answer the question of when customers choose their own joining probability, whether the planner can find an appropriate abandonment threshold leading to a Nash equilibrium as well.

We first analyze the behavior of an individual customer's expected utility function with respect to the joining probability p when the abandonment threshold is exogenous. When the effective arrival rate to the queueing system increases, for a given abandonment threshold, the waiting time increases and the chance of being served within a certain time decreases. We show that there exists a unique equilibrium for the joining probability, denoted by $\overline{p}$, in the following proposition.

Proposition 6 *For a given abandonment threshold T, if all customers follow the T policy, the expected utility function decreases monotonically with respect to the joining probability and there is a unique equilibrium $\bar{p}$.*

Since customers selfishly choose their joining strategy, then if $p < \overline{p}$, there is an incentive for customers to join the queue and receive positive expected utility. Otherwise, if $p > \overline{p}$, customers who join the queue incur losses and the equilibrium is pushed to $\overline{p}$. Specifically, the best response function of each customer is not increasing in p, and customers use the *avoid-the-crowd (ATC)* strategy (e.g., see Hassin and Haviv 2003). While all customers follow (p, T), it is clear that if customers choose $p < 1$, then the expected utility function should be equal to zero; otherwise, for $p = 1$, it is necessary to have $U_T(p, T) \geq 0$ (see also Hassin and Haviv 1995).

In the following proposition, we show that the equilibrium joining probability has a unique maximum point with respect to the anticipation point.

Corollary 5 *If the abandonment threshold is exogenous, the equilibrium $\bar{p}$ is unimodal in β.*

Corollary 5 states that as the anticipation point increases, the joining probability increases until reaching a maximum value. After that point, the joining probability begins to decline as the anticipation point increases. Therefore, if the abandonment threshold T is

exogenous, the queueing system faces a maximum joining rate when customers are not highly patient or highly impatient – essentially a moderate anticipation point. Corollary 5 demonstrates how the anticipation point captures the effect of opposing responses in encouraging customers to join or balk the system. While customers anticipate a short waiting time but offered a longer wait, the psychological cost of waiting preponderates over their commitment to stay in the queue longer. In this case, customers are discouraged from joining the queue. On the other hand, if customers anticipate a longer wait and commit to stay in queue, the queueing system will become less preferable and there is little incentive for customers to join the queue if they know they will be forced to renege early in the wait. As such, the maximum joining probability is brought about when there is an appropriate balance between these two opposing tensions. This balance is achieved with a moderate anticipation point. In this light, our model fully captures this psychological behavior.

Proposition 7 *Suppose customers choose balking strategically according to an abandonment threshold. Then, there exists a unique optimal abandonment threshold T^* and corresponding Nash equilibria for customers' balking and reneging strategy, that has the following threshold structure:*

1. *There exists a unique equilibrium $(\bar{p}, \bar{T})$ in which $\bar{T}$ coincides with T^*.*
2. *In the case $V_3(\beta, 1) \leq V$ or $V_2(\beta) < V \leq V_1(\beta, 1)$, customers apply the fully join strategy, i.e., $\bar{p} = 1$.*
3. *In the case $V < \min(V_1(\beta, 1), V_2(\beta))$, customers always apply a partial join strategy.*

Figure 2.14 illustrates the structure of the customer equilibrium strategy. The red, blue, and green curves represent $V_1(\beta, p)$, $V_2(\beta)$ and $V_3(\beta, p)$, respectively when $p = 1$. The black curve indicates partial and fully join strategies. In Proposition 7, we are considering the joint balking probability p and reneging time simultaneously under an abandonment threshold policy. As depicted in Figure 2.14, there are 5 regions. In Regions A and D the social welfare function is unimodal and there exists a unique optimal point. Likewise, in

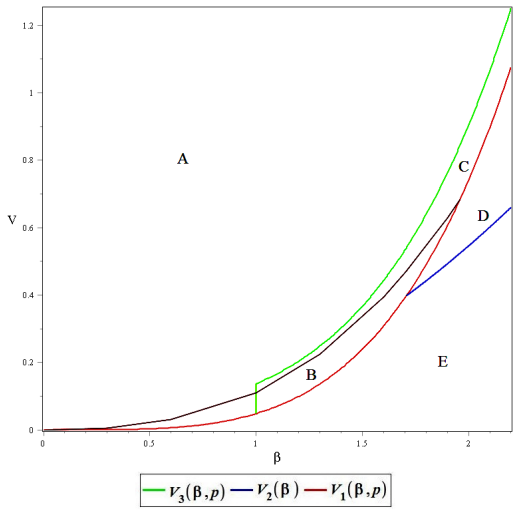

Figure 2.14: Structure of the equilibrium customer strategy under abandonment threshold policy.

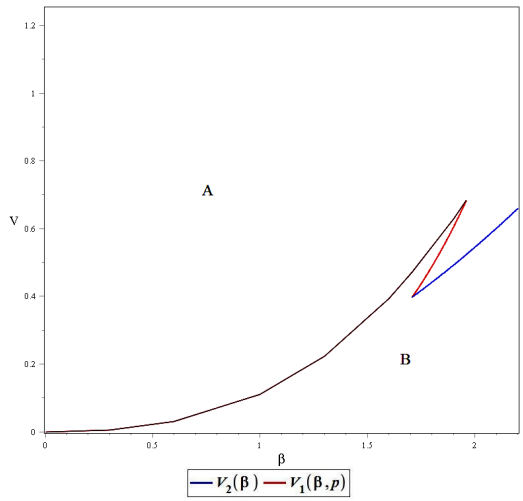

Figure 2.15: Summary of the equilibria under abandonment threshold policy.

Region C the social welfare function is bimodal and there exists both positive type-1 and type-2 points.

Note that in all these regions (A, C, D), the equilibrium joining probability is 1; that is, all customers join the system and there is a unique pure Nash equilibrium reneging time for all customers which is equal to T^*. However, in Regions B and E there exists a unique mixed Nash equilibrium and customers always apply a partial join strategy. In this case, the customer's expected utility and welfare functions are equal to zero and one is indifferent between not joining or joining and reneging at T^* resulting in a strict mixing between two pure strategies. Since all customers will choose the optimal abandonment threshold as an equilibrium one, there are two types of equilibria: a unique pure Nash equilibrium which is equal to T^* under which all customers join the queue (Region A in Figure 2.15), or a unique mixed Nash equilibrium under which customers apply the partial join strategy with a positive reneging time (Region B in Figure 2.15). Figure 2.15 shows that the anticipation point and service value significantly affect a customer's joining probability decision. As depicted in Figure 2.15, if customers anticipate a long wait time,

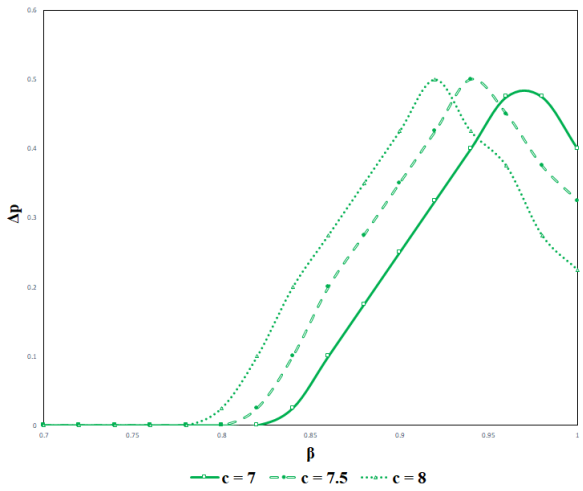

Figure 2.16: Effect of β on the gap between the customer's equilibrium and the optimal joining probabilities.

they join the system but only partially. However, these customers may not keep to the partial joining strategy monotonically with respect to anticipation point increases. As shown in Figure 2.14, depending on the value of the service reward, customers may choose a fully join strategy for higher values of the anticipation point (Region D in Figure 2.14). In this case the optimal abandonment threshold is always less than the anticipation point (Corollary 4).

The partial join region defined in Proposition 1 (depicted in Figure 2.11) shrinks if considered under the equilibrium case characterized in Proposition 7 (see Figure 2.15). Since customers are self-interested, as long as the expected utility function is positive, there is an incentive for customers to join the system. This means that at the equilibrium, the expected utility function tends to be zero and as a result the total surplus goes to zero. Figure 2.16 illustrates the difference between the customer's equilibrium and the optimal joining probability, $\Delta p = \bar{p} - p^*$. This gap reveals the externalities that an arrival imposes on other customers and, consequently, the amount of effort that should be applied to regulate it. Figure 2.16 shows that this gap is unimodal in the anticipation point.

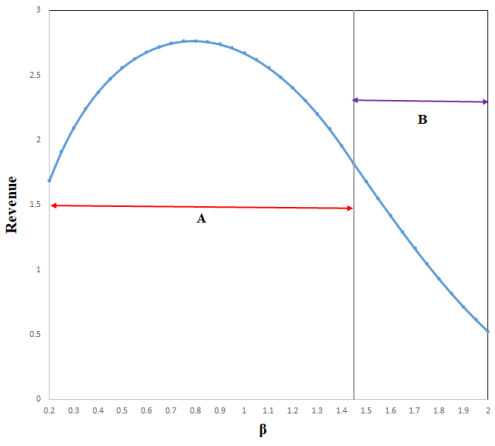

Figure 2.17: Segmentation of the expected revenue using the entrance-fee/abandonment-threshold mechanism.

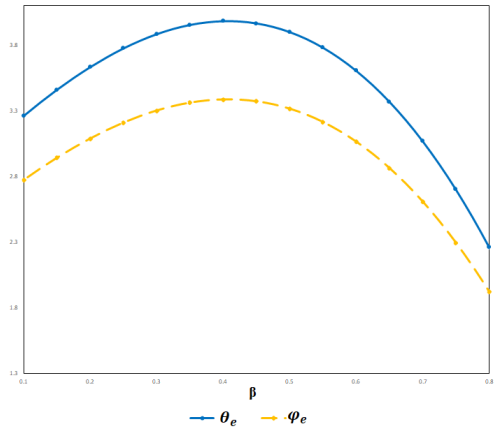

Figure 2.18: Effect of β on the optimal entrance fee and the expected revenue.

As the firm charges a fee to collect the total surplus (the sum of all customers' surplus), a lower surplus results in a lower profit for the firm. To avoid this situation, the planner must regulate the joining probability to increase total surplus. Because the expected utility function is decreasing in the joining probability, the planner sets p^* smaller than the equilibrium point $\overline{p}$, to achieve a positive expected utility and in doing so maximizes the profit by imposing a fee. Therefore, the abandonment threshold policy by itself is not an optimal policy to control the congestion and to eliminate customer externalities. Next, we show that if a planner (revenue maximizer) can impose a fee to control the arrival rate, there exists an optimal joint pricing and abandonment threshold policy when customers follow a (p, τ) strategy. We show that this joint policy induces socially optimal behavior, and the revenue maximizer can obtain all the social welfare by charging a unique maximal price and placing a restriction on customer's reneging time.

Let $\hat{p}_e$ and $\hat{T}_e$ denote the equilibrium joining probability and reneging time of customers respectively, when the joint entrance-fee/abandonment-threshold mechanism is employed by the firm.

Theorem 2 *Using the entrance-fee/abandonment-threshold mechanism,*

1. *There exists an optimal entrance fee, where* $\theta_e = U(p^*, T^*)$.
2. *The expected revenue is equal to the optimal social welfare and the equilibrium customers' joining probability* $\hat{p}_e$ *and reneging time* $\hat{T}_e$ *coincides with the optimal joining probability* p^* *and abandonment threshold* T^**, respectively.*

According to Theorem 2, when an entrance fee is imposed, the customer equilibrium strategy coincides with the optimal strategy stated in Theorem 1 (the same as shown in Figure 2.11). Under the entrance-fee/abandonment-threshold mechanism, the firm can induce socially optimal behavior. In this case, the proposed mechanism can be used to collect the total surplus and guarantee a profit equal to the optimal social welfare $W(p^*, T^*)$.

Corollary 6 *Using the entrance-fee/abandonment-threshold mechanism, when* $V < \min(V_1(\beta, 1), V_2(\beta))$*, we can guarantee that the optimal price always leads to the adoption of a partial join strategy.*

Figure 2.17 illustrates the relationship between the anticipation point β and the expected revenue under the entrance-fee/abandonment-threshold mechanism. As illustrated in Figure 2.17, the expected revenue function has a unique maximum point in β and also customers will use the fully join strategy in Region A and the partial join strategy in Region B. Moreover, numerical results show that when an entrance fee is charged, the optimal entrance fee θ_e and expected revenue ϕ_e are unimodal in the anticipation point β (Figure 2.18) indicating that a moderate anticipation point provides higher expected revenue. These findings highlight the importance of considering psychological factors in queue control which have largely been ignored in the literature.

Chapter 3

Analysis of the Multi-Echelon Production Inventory System with Strategic Customers

In this chapter, we consider a two-echelon production inventory system where customers strategically either choose to place an order and wait until receiving the product or balk the system without placing any order. We first analyze the system assuming that the arrival rate is exogenous. We obtain a closed-form expression to find the optimal base-stock level at the DC, when the production and transportation times are generally distributed. Then, we assume that arrival rate is endogenous (customers behave strategically) and use the formula obtained in the first model to develop the DC cost and also customer expected utility function. However, in the endogenous arrival rate scenario, to make the problem tractable, we assume that the production times are exponentially distributed and the transportation time from the manufacturer to the DC is deterministic.

3.1 Two-Echelon Inventory System with Exogenous Arrival Rate

3.1.1 Two-Echelon Inventory System with a Manufacturer

In this section, we consider a two-echelon inventory system with a single manufacturer having no warehouse and a single DC that manages its inventory using a base-stock policy (see Figure 3.1). We assume that the DC manages its inventory using a base-stock policy with the base-stock level S. We assume that the transportation time between the manufacturer and the DC is uncertain and follows a general distribution with mean η and Laplace transform (LT) $\delta^*(s)$, and the production times are generally distributed with the mean of $1/\mu$. Let $f(\cdot)$ denote the probability density function (PDF) of the lead-time of an order placed by the DC. This time includes the waiting time in the production system, with the LT of $w^*(s)$, and the transportation time from the manufacturer to the DC which are independently distributed; therefore, the Laplace transform of the lead-time of an order placed by the DC, denoted by $f^*(s)$, is obtained as $f^*(s)=w^*(s)\delta^*(s)$. Note that, when the transportation times are stochastic, orders may cross over time. To make the problem tractable, we assume that no order crossing is allowed over time and products are received by the DC sequentially (for more details concerning this assumption, see e.g., Svoronos and Zipkin(1991)).

Assuming that the inventory holding and shortage costs per unit and time unit at the DC are h and b, respectively, our objective is to minimize the average total inventory holding and shortage costs at the DC.

We use the Flow-Unit method presented by Axsäter (1990) to determine the average cost of a unit demanded at the DC. We first need to obtain the lead-time distribution of an order placed by the DC. When the inventory is managed using the one-for-one replenishment policy with the base-stock level S, an item ordered by the DC is used to fill the S^{th} demand arrival to the DC after this order is placed. Therefore, a new order, sent to the manufacturer upon the arrival of j^{th}customer at the DC, is used to satisfy the $(S+j)^{th}$

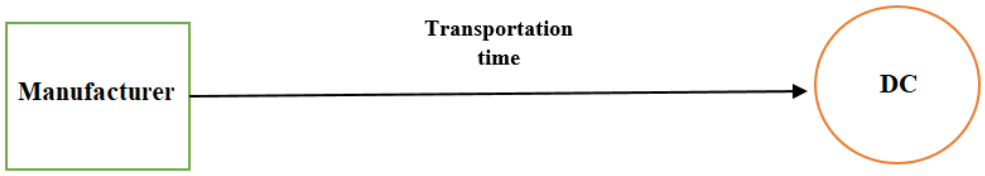

Figure 3.1: Production inventory system with no warehouse.

customer at the DC. If the $(S+j)^{th}$ customer arrives before her corresponding item is received by the DC, she should wait until the product is received. The tail distribution (complementary cumulative distribution function) of this waiting time is denoted by $G^Q(t)$. If the product is received by the DC before the $(S+j)^{th}$ customer arrives to the DC, the product will be held until its corresponding customer arrives. Let $G^P(t)$ denote the tail distribution of the time elapsed between the instant that a product is received by the DC and its corresponding customer arrives. Using $G^Q(t)$ and $G^P(t)$, we obtain the expected shortage and holding times of a unit demanded at the DC.

Recall that customers arrive at the DC according to an Poisson process with the rate of λ. Therefore, the time until S customers arrive has an $Erlang(S,\lambda)$ distribution. Let the random variable X denote the lead-time of an order placed by the DC, i.e., the time between the moment an order is placed by the DC and its corresponding product is received. The distribution of X is the convolution of the waiting time of the order in the manufacturer and the transportation time. Moreover, let the random variable Y denote the time between the placement of an order to the manufacturer and the arrival of the customer at the DC who receives the corresponding product of the order. The distribution of Y is $Erlang(S,\lambda)$ since the $(S+j)^{th}$ customer is served using the product of the order placed at the arrival of j^{th} customer. Consequently, $(X-Y)^+$ represents the amount of the time a customer waits in DC until she receives her product with the tail distribution of $G^Q(t)$. Similarly,

$(Y-X)^+$ is the amount of the time a unit of the product is kept in the DC until its corresponding customer arrives with the tail distribution of $G^P(t)$.

We use the normalized incomplete Gamma function (NIG), $Q(S,\lambda t)$, to find the cumulative distribution function (CDF) of the Erlang distribution. Note that, $1-Q(S,\lambda t)$ is the CDF of an Erlang distribution with parameters λ and S and it has the following properties (see e.g., Abouee-Mehrizi et al. 2011):

$$Q(S,\lambda t)=\frac{\int_{\lambda t}^{\infty} e^{-y}\, y^{S-1}\, dy}{(S-1)!},$$

and

$$Q(S,\lambda t)-Q(S-1,\lambda t)=-\frac{1}{\lambda}\frac{dQ(S,\lambda t)}{dt}=\frac{(\lambda t)^{S-1}}{(S-1)!}e^{-\lambda t}. \tag{3.1}$$

Since Y has an Erlang distribution with parameters λ and S, $Q(S,\lambda t)$ can be used to obtain the tail distribution of the time that a unit of the product is kept in the DC, $(Y-X)^+$, as,

$$G^P(t)=P(Y-X>t)=\int_0^{\infty} f(x)P(Y>X+t\mid X=x)\,dx=\int_0^{\infty} f(x)Q(S,\lambda(x+t))\,dx. \tag{3.2}$$

Similarly, the tail distribution of the time that a customer waits until she is served, $(X-Y)^+$, is

$$G^Q(t)=P(X-Y>t)=\int_t^{\infty} f(x)P(Y<X-t\mid X=x)\,dx$$

which results in

$$G^Q(t)=\int_t^{\infty} f(x)(1-Q(S,\lambda(x-t))\,dx. \tag{3.3}$$

Given (3.2) and (3.3), we can obtain the fraction of customers who find the DC empty and non-empty as $G^Q(0)$ and $G^P(0)$, respectively:

$$G^P(0)=P(Y>X),\ G^Q(0)=P(X>Y). \tag{3.4}$$

From (3.2) and (3.3) we can see that the function $Q(S,\lambda x)$ plays the main role in characterizing the tail distributions of the time a customer waits and a unit of the product is kept in the DC. Let $Za(S,\lambda,f(\cdot))=\int_0^{\infty} f(x)\,Q(S,\lambda x)\,dx$. In the following lemma, we provide

a relation between $Za\left(S,\lambda,f(\cdot)\right)$ and the LT of $f(\cdot)$ which helps us to obtain the total cost of the system. Let $f^*(s)$ denote the LT of $f(\cdot)$.

Lemma 1 *$Za\left(S,\lambda,f(\cdot)\right)$ can be written as a function of $f^*(s)$ as follows:*

$$Za\left(S,\lambda,f(\cdot)\right)=\int_0^\infty f\left(x\right)Q\left(S,\lambda x\right)dx=\sum_{i=1}^{S}\left(-1\right)^{i-1}\frac{\lambda^{i-1}}{\left(i-1\right)!}\ f^{*(i-1)}(k)|_{k=\lambda}, \tag{3.5}$$

where $f^{*(n)}\left(\cdot\right)=\frac{d^n}{ds^n}f^*(s)$.

As shown in Lemma 1, the function $Za\left(S,\lambda,f(\cdot)\right)$ is a weighted sum of derivatives of $f^*(s)$. In the following lemma, we find an explicit expression for $Za\left(S,\lambda,f(\cdot)\right)$.

Lemma 2 *If $f\left(\cdot\right)$ and $F(\cdot)$ are the PDF and CDF of the random variable X, then,*

$$Za\left(S,\lambda,f(\cdot)\right)=\left(-1\right)^{S-1}\frac{\lambda^S}{\left(S-1\right)!}\ \ \frac{d^{S-1}}{dk^{S-1}}\left(F^*(k)\right)|_{k=\lambda},$$

and

$$Za(S,\lambda,F(\cdot))=\frac{1}{\lambda}\ \sum_{i=1}^{S}\ Za\left(i,\lambda,f(\cdot)\right). \tag{3.6}$$

Using Lemmas 1 and 2, we next obtain the expected time that a unit of product is kept in the DC and the expected time that a customer waits, which are denoted by $W^P\left(S,\lambda,f(\cdot)\right)$ and $W^Q\left(S,\lambda,f(\cdot)\right)$, respectively.

Theorem 3 *The expected holding time and shortage time of a unit demanded by the DC are:*

$$W^P\left(S,\lambda,f(\cdot)\right)=Za\left(S,\lambda,F(\cdot)\right),$$

and

$$W^Q\left(S,\lambda,f(\cdot)\right)=Za\left(S,\lambda,F(\cdot)\right)-f^{*(1)}(k)|_{k=0}-\frac{S}{\lambda}. \tag{3.7}$$

Now, we can obtain the average inventory and shortage costs of a unit demanded at the DC, denoted by $Cost(S, f(\cdot))$, as:

$$Cost(S, f(\cdot)) = hW^p(S, \lambda, f(\cdot)) + bW^Q(S, \lambda, f(\cdot)). \tag{3.8}$$

Let $\Pi(S, f(\cdot))$ denote the average total holding and shortage costs at the DC. Since the demand rate at the DC is λ, we have

$$\Pi(S, f(\cdot)) = \lambda Cost(S, f(\cdot)). \tag{3.9}$$

To obtain the optimal base-stock level which minimizes the total inventory cost, we prove in the following proposition that the average total cost function is convex in the base-stock level.

Proposition 8 *The average total inventory cost function, $\Pi(S, f(\cdot))$, is convex with respect to the base-stock level S. Moreover, the optimal base-stock S^* can be obtained using*

$$S^* = \min\{k : \gamma(k, f(\cdot)) \geq 0\}, \tag{3.10}$$

where $\gamma(S, f(\cdot))$ is

$$\gamma(S, f(\cdot)) = Za(S,\lambda,f(\cdot)) - \frac{b}{(h+b)}. \tag{3.11}$$

3.1.2 Two-Echelon Production Inventory System

In this section, we consider a two-echelon inventory system with a manufacturer keeping its stock in a warehouse to satisfy the orders placed by the DC (see Figure 3.2). We assume that the inventory in the warehouse is managed using a base-stock policy with the base-stock level S_0. Assuming that the inventory holding cost per unit and time unit in the warehouse is h_0, we are interested in minimizing the average total inventory holding and shortage costs at the DC and the warehouse.

To simplify the analysis, the state of the system can be broken down into two cases. In the first case, upon an order arrival to the warehouse from the DC, the warehouse

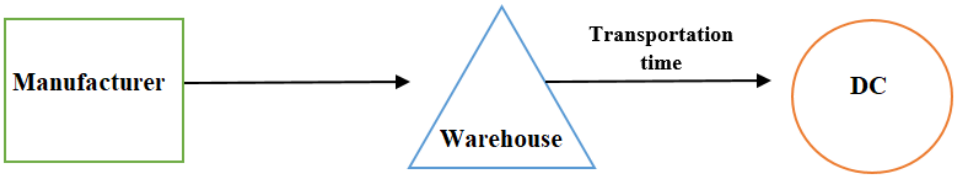

Figure 3.2: Production inventory system with a warehouse.

has some stock and in the second one, it is out of stock and this arriving order joins the production queue. Since customers arrive to the DC according to a Poisson process and the DC uses a one-for-one replenishment policy, arriving orders to the warehouse also follow a Poisson process with the rate of λ. Let $G_0^P(t)$ denote the tail distribution of the time that a unit of the product is kept in the warehouse. Recall that $w(t)$ denotes the waiting time distribution of an order in the production system. Therefore, the probability that an arriving order from the DC finds the warehouse non-empty, given by $G_0^P(0)$, is (similar to (3.4)):

$$G_0^P(0) = \int_0^\infty w(x) Q(S_0, \lambda x)\, dx = Za(S_0, \lambda, w(\cdot)). \tag{3.12}$$

If the warehouse is not empty which happens with probability $Za(S_0, \lambda, w(\cdot))$, the arriving order to the warehouse is immediately satisfied and is sent to the DC which takes t time units to be received by the DC with the PDF of $\delta(t)$. Otherwise, when the warehouse is out of stock upon an order arrival with the probability $1 - Za(S_0, \lambda, w(\cdot))$, it takes t time units with the PDF of $f(t)$, to be received by the DC where $f(t)$ is the convolution of the waiting time distribution in the manufacturer and the transportation time between the warehouse and the DC. Therefore, we get the distribution of the time between the instant that an order is placed by the DC until its corresponding product is delivered to the DC,

denoted by $u(S_0,\lambda, f(\cdot), \delta(\cdot), t)$, as the following:

$$u(S_0,\lambda, f(\cdot), \delta(\cdot), t) = Za\,(S_0,\lambda,w(\cdot))\,\delta(t)+ (1 - Za(S_0,\lambda,w(\cdot)))f(t). \tag{3.13}$$

For mathematical convenience we use $u(\cdot)$ to refer to $u(S_0,\lambda, f(\cdot), \delta(\cdot), t)$. Let $U(\cdot)$ denote the CDF of $u(\cdot)$ given by (3.13). Similar to Theorem 3 and using (3.13), we find the expected holding and shortage times of a unit demanded by the DC in the following proposition.

Proposition 9 *The expected holding and shortage times of a unit demanded by the DC are:*

$$\begin{aligned} W^P\,(S,\lambda, u(\cdot)) =&\ \ Za\,(S,\lambda,U(\cdot))\,, \\ W^Q\,(S,\lambda, u(\cdot)) =& \ Za\,(S,\lambda,U(\cdot)) - u^{*(1)}(k)|_{k=0} - \frac{S}{\lambda}. \end{aligned} \tag{3.14}$$

We next derive the warehouse inventory cost. Let $W_0^P(S_0, \lambda, w(\cdot))$ denote the expected inventory carrying time for a unit in the warehouse. Since both the DC and the warehouse use base-stock policy and receive demand according to a Poisson process, similar to the derivation of (3.2) and Theorem 3, we obtain the expected inventory holding time for a unit in the warehouse, denoted by $W_0^P(S_0, \lambda, w(\cdot))$, as:

$$W_0^P(S_0, \lambda, w(\cdot)) = \int_0^\infty G_0^p(t)dt = Za(S_0, \lambda, W(\cdot)), \tag{3.15}$$

where $W(\cdot)$ is CDF of the waiting time distribution in the manufacturer with the PDF of $w(\cdot)$.

Since the demand rate at the warehouse is λ, using Proposition 9 and (3.15), the average total inventory cost in the system is determined by:

$$\Pi\,(S,S_0, u(\cdot)) = \lambda\,\big(\ Cost\,(S,u(\cdot)) + h_0 W_0^P(S_0, \lambda, w(\cdot))\big)\,,$$

where

$$Cost\,(S,u(\cdot)) = hW^P\,(S,\lambda, u(\cdot)) + b\ W^Q\,(S,\lambda, u(\cdot))\,. \tag{3.16}$$

Note that $Cost\,(S,u(\cdot))$ denotes the average cost per unit at the DC when the lead-time distribution is $u(\cdot)$.

We next show that the average total cost function given in (3.16) is convex in the base-stock level S and derive a closed-form expression for the optimal base-stock level at the DC.

Proposition 10 *The average total inventory cost function, $\Pi(S, S_0, u(\cdot))$, is convex in the base-stock level S. Therefore, the optimal base-stock S^* can be obtained using*

$$S^* = \min\{k : \gamma\,(k, S_0, u(\cdot)) \geq 0\}, \tag{3.17}$$

where $\gamma\,(S, S_0, u(\cdot))$ is

$$\gamma\,(S, S_0, u(\cdot)) = Za\,(S,\lambda,u(\cdot)) - \frac{b}{(h+b)}. \tag{3.18}$$

Note that $Za\,(S,\lambda,u(\cdot))$ is increasing in S for a given S_0. Therefore, using a simple iterative algorithm, we can find the optimal base-stock level at the DC, S^*, that satisfies (3.17). Recall that in (3.17), $u\,(\cdot)$ is the PDF of the lead-time with the LT of $u^*(s)$. This lead-time is a convolution of the waiting time in the manufacturer with the base-stock level S_0, and the transportation time from the warehouse to the DC. Therefore, the optimal reorder point depends on the base-stock level in the warehouse as well. However, it is not straightforward to derive a closed-form expression for the optimal S^* and S_0^*.

3.2 Two-Echelon Inventory System with Endogenous Arrival Rate

We assume that customers arrive at the DC according to a Poisson process with a rate λ and are served based on a First-Come First-Served (FCFS) policy. Each customer receives the value V which is the difference between a product reward R and charged price Pr, i.e. $V = R - Pr$. Once customers arrive at the DC, they decide to join the system and

place an order with probability p or balk the system with probability $1-p$. Customers make a decision based on their expected utility function which is linear with respect to the product value V, and non-linear in the waiting time with the risk aversion degree θ. The expected utility function for a customer is defined as:

$$U\left(p, Pr\right) = R - Pr - c\ \bar{x}^{\theta}, \tag{3.19}$$

where $\theta \geq 1$, $\bar{x}$ is the mean delay before receiving the product, and c is a positive constant. In this expected utility function, the marginal cost of waiting depends on the risk aversion degree. If $\theta = 1$, customers face a constant marginal cost of waiting, independent of how much they have spent so far in the system, and suffer with a constant rate c while waiting. In other words, customers are *risk neutral* with respect to the loss suffered from waiting. In the case of $\theta > 1$, the marginal cost of waiting in the system increases in the waiting time. Therefore, customers are *risk averse* with respect to the loss.

Recall that the DC uses the base-stock policy to manage its inventory with the base-stock level S. The orders received by the DC which find the DC out of stock are backlogged and served based on a FCFS policy when the DC receives new units of the product. We assume that the manufacturer operates from a warehouse that manages its inventory using a base-stock policy with the base-stock level S_0. Arriving orders that find the warehouse out of stock are backlogged. When a production ends, the product is used to satisfy backlogged orders, if there are any, based on the FCFS policy, otherwise, it is kept in the warehouse. The production times at the manufacturer are exponentially distributed with mean $1/\mu$. We assume that the transportation time between the manufacturer and the DC is deterministic and is equal to T. Recall that the lead-time of an order placed by the DC is the sum of the waiting time of the order at the manufacturer and the transportation time from the manufacturer to the DC with the PDF $f(\cdot)$, CDF $F(\cdot)$, and Laplace Transform (LT) $f^*(s)$. Assuming that the waiting time in the manufacturer and transportation time are independently distributed, we get

$$f^*(s) = \left(1 - \left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}\right) \mathrm{e}^{-sT} + \frac{\mathrm{e}^{-sT}\left(\mu - \bar{\lambda}\right)}{\mu - \bar{\lambda} + s} \left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}, \tag{3.20}$$

where $\bar{\lambda} = \lambda p$ is the effective arrival rate. We assume that the only cost for the DC is inventory holding cost and revenue is collected by charging the price Pr. Let h denote the inventory holding cost per unit per time. Note that $Za(S,\bar{\lambda}, f(\cdot))$ is the probability that an arriving order finds the DC non-empty (see (3.4)). Now, considering the expected inventory holding time given by Theorem 3, the expected revenue of the DC is obtained as:

$$\Delta(S, \bar{\lambda}, f) = \lambda p(Pr - hW^P\left(S,\bar{\lambda}, f(\cdot)\right)). \tag{3.21}$$

Upon receiving a new order, the DC is not empty with probability $Za(S, \bar{\lambda}, f(\cdot))$ and is thus immediately satisfied. Otherwise, the DC is out of stock and the expected lead time to satisfy this order is $(1 - Za(S,\bar{\lambda},f(\cdot)))(W^Q\left(S,\bar{\lambda}, f(\cdot)\right))$. Therefore, using (3.19) we can rewrite the expected utility function of customers who join the system as:

$$U\left(p, Pr\right) = R - Pr - c\ (\bar{W}(p, S))^{\theta}, \tag{3.22}$$

where $\bar{W}(p, S)$ is the mean delay until receiving the product given by:

$$\bar{W}(p, S) = (1 - Za(S,\bar{\lambda},f(\cdot)))(W^Q\left(S,\bar{\lambda}, f(\cdot)\right)). \tag{3.23}$$

3.2.1 Joining Probability Equilibrium

Customers do not consider their negative externalities on others, and while their expected utility function is positive, they join the system to gain a positive value. In this section, we analyze the customer equilibrium joining probability and prove that the equilibrium joining probability denoted by $\overline{p}$ is unique. First, we investigate the impact of the effective arrival rate $\bar{\lambda}$ on the mean delay function $\bar{W}(p, S)$.

Lemma 3 *$Za\left(S,\bar{\lambda}, f(\cdot)\right)$ and $W^P\left(S,\bar{\lambda}, f(\cdot)\right)$ are monotone decreasing and $W^Q\left(S,\bar{\lambda}, f(\cdot)\right)$ is monotone increasing in $\bar{\lambda}$.*

Lemma 3 indicates that the probability that an arriving customer finds the warehouse empty, $1 - Za(S,\bar{\lambda},u(\cdot))$, is monotone increasing in the effective arrival rate ($\bar{\lambda}$). Also, the

expected waiting time of a unit at the DC, $W^Q\left(S,\bar{\lambda},u(\cdot)\right)$, increases when the effective arrival rate rises. Consequently, according to (3.22), the customer's expected utility function falls monotonically when the effective arrival rate increases. Using this property, we characterize the Nash equilibrium for the joining probability p which is denoted by $\bar{p}$.

Theorem 4 *For a given S, the expected utility function is monotone decreasing in p and there is a unique Nash equilibrium $\overline{p}$ where $U\left(\bar{p},Pr\right)=0$.*

Theorem 4 states that if $p<\overline{p}$, there is an incentive for customers to join the queue and earn positive profit by placing an order. Otherwise, when $p>\overline{p}$, customers who join the queue incur losses and the equilibrium is pushed to $\overline{p}$. Using Theorem 4 and (3.22), we can obtain $\overline{p}$ as the solution of:

$$\bar{W}(p,S)=m_\theta, \tag{3.24}$$

where $m_\theta=((R-Pr)/c)^{\frac{1}{\theta}}$ which we call *weighed-gain-loss ratio*.

Next, we analyze the behavior of the system in two special cases where either all customers join the system and place an order or no customer joins. First, we need to investigate the effect of the base-stock level on the customer's expected utility.

Lemma 4 *For a given p, $U\left(p,Pr\right)$ is monotone increasing with respect to the base-stock level at the DC.*

According to Lemma 4 and Theorem 4, it is clear that when the base-stock level increases, the lead-time of the order declines and consequently the customer's expected utility rises. Since customers place an order with probability p, we are interested in finding the conditions that, independently of the base-stock level, all customers will join the system or no customer will join.

Proposition 11 *Under the following conditions there is no partial join strategy.*

i. The strategy $\overline{p}=1$ is a dominating strategy if

$$T+\frac{1}{\mu-\lambda}\left(\frac{\lambda}{\mu}\right)^{So}\leq m_\theta. \tag{3.25}$$

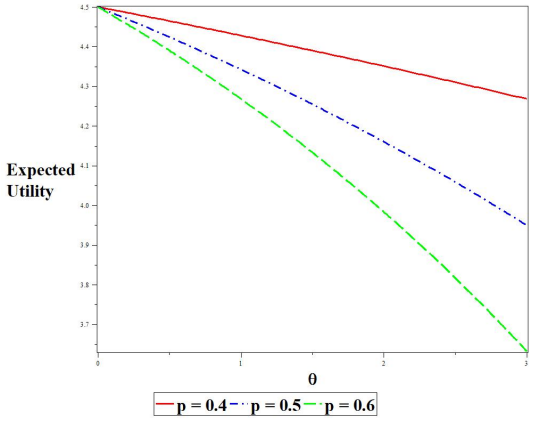

Figure 3.3: The effect of θ on the expected utility for $S = 1$.

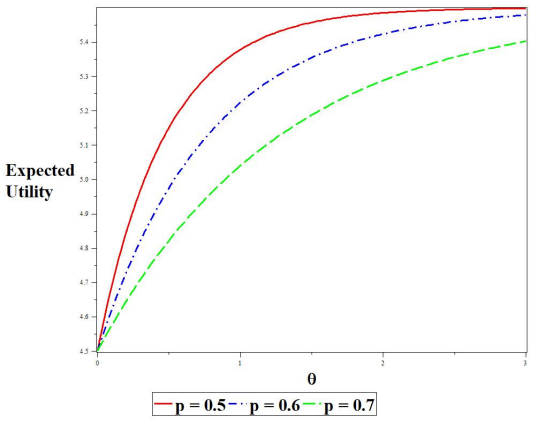

Figure 3.4: The effect of θ on the expected utility for $S = 5$.

ii. The strategy $\overline{p}$=0 is a dominant strategy if

$$m_\theta \leq 0. \tag{3.26}$$

The first condition given by (3.25) shows that if the mean lead-time in the DC is less than the weighed-gain-loss ratio, all customers will join and $\overline{p}$=1 is a dominant strategy. In contrast, the second condition given by (3.26) states that no customer will join the queue and a dominant strategy is $\overline{p}$=0, if the weighed-gain-loss ratio is smaller than zero. We can conclude that partial join strategy takes place if the weighed-gain-loss ratio is between the two bounds given by (3.25) and (3.26).

3.2.2 Risk Aversion Degree Effect

In this section, we analyze the effect of θ on the customer behavior and decision. First, we investigate the effect of θ on the customer's expected utility function.

Lemma 5 *For a given p, the customer's expected utility function is concave in θ.*

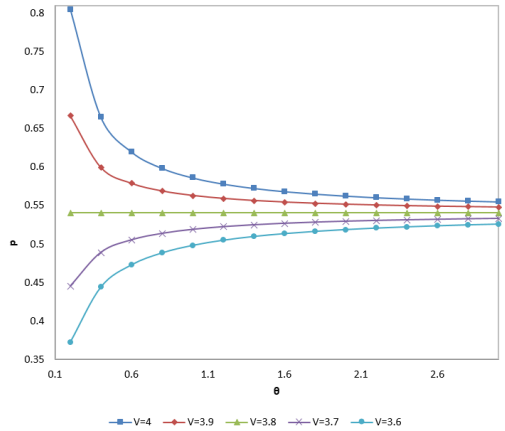

Figure 3.5: The effect of θ on $\overline{p}$ for $c = 3.8$.

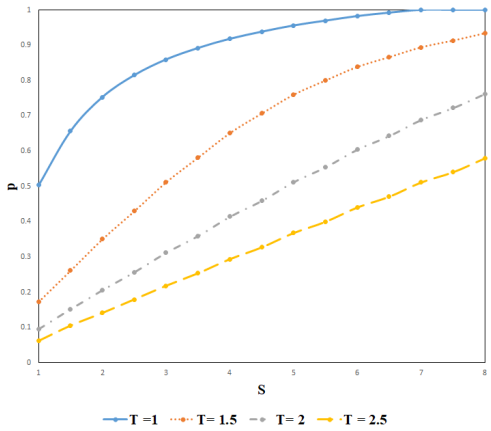

Figure 3.6: The effect of S on $\overline{p}$.

Figures 3.3 and 3.4 show the effect of θ on the expected utility function in red when $p = 0.4$, in blue when $p = 0.5$, and in green when $p = 0.6$ for two different base-stock levels. Lemma 5 indicates that increases in the benefit gained by a customer, declines when the risk aversion degree rises. In other words, when customers become more risk averse, the marginal benefit due to an increase in θ declines. Also, since $\frac{\partial U(p,Pr)}{\partial \theta} = -c\ \bar{W}^{\theta}(p,S)\ \ln(\bar{W}(p,S))$, the customer's expected utility can be increasing or decreasing with respect to θ based on the base-stock level at the DC, as shown in Figure 3.3 when $S = 1$ and in Figure 3.4 when $S = 5$.

Proposition 12 *If the weighed-gain-loss ratio $m_\theta > 1$, the equilibrium joining probability, $\overline{p}$, is monotone decreasing and if weighed-gain-loss ratio $m_\theta < 1$, the equilibrium joining probability is monotone increasing in θ. In both cases, $\overline{p}$ converges to p' which is the equilibrium joining probability when $R - Pr = c$.*

As illustrated in Figure 3.5, when the value of the product, $V = R - Pr$, increases, the equilibrium joining probability changes from increasing to decreasing with respect to the aversion degree. When the value of product is higher than c, $V > c$, and customers become more impatient (more risk averse), the weighed-gain-loss ratio declines resulting

in a decrease in the joining probability. However, when the value of product is less than c, $V < c$, and customers become more impatient, the weighed-gain-loss ratio rises resulting in an increase in the joining probability. Also, as the risk aversion degree increases, for any level of V, the equilibrium joining probability converges to a certain point p'.

Next, we analyze customer behavior with respect to the supply chain parameters such as the transportation time and the base-stock levels at the DC.

Proposition 13 *The equilibrium joining probability, $\overline{p}$, is increasing in S.*

Increasing the base-stock level at the DC reduces the lead-time, encouraging people to place an order. Also as shown in Figure 3.6, by increasing the base-stock level, the marginal increase in the joining probability is higher for the shorter transportation times. One reason for this is that when the transportation time is low, any changes in the base-stock level has a high impact on the expected lead-time and also on the probability of finding the DC out of stock. Thus, in this case any changes in the base-stock level affect customer decision much more than the case with longer transpiration times. In the following corollary, the effect of transportation time on the joining probability is discussed.

Corollary 7 *$\overline{p}$ is monotone decreasing in T.*

3.2.3 DC as a Stackelberg leader

In this section, we investigate the interaction between customers as the followers and the DC as the leader in a Stackelberg game. We then find the DC optimal expected revenue, price, and the base-stock level. We assume that the DC as the leader, has complete information about customers' expected utility. According to the Stackelberg game as a non-cooperative game, the leader (DC) takes an action first and commits to its strategy to the followers (customers). Given the leader's decision, the followers behaves as self-maximization customers, then simultaneously make their own decision.

In this system, the DC chooses the price and the base-stock level as its strategy $\sigma = (Pr, S)$, to maximize its expected revenue given in (3.21). Then, given the DC's strategy, customers as followers decide on choosing the probability of placing an order at the DC considering their expected utility function given in (3.22). Customers are self-interested and each strategy σ taken by the DC leads to an equilibrium joining probability (if any) chosen by customers. Therefore, the objective of the DC is to find an optimal strategy σ^* which leads to the Nash equilibrium for customers such that no customer will be better off by unilaterally changing his strategy. To do so, the DC should consider the best response function of customers for any given strategy σ and set the price and the base-stock level according to this best response function. Since the DC plays as a monopolist, it can charge a price Pr^* to gain total surplus, i.e. $U(p, Pr^*) = 0$. Therefore, from (3.22) the optimal price is obtained as:

$$Pr^* = R - c\, \bar{W}^{\theta}(p, S). \tag{3.27}$$

Knowing the customer response function, the DC can substitute the price Pr^* and modify its expected revenue function given by (3.21) as

$$\hat{\Delta}(S, \lambda p, f(\cdot)) = \lambda p \left(R - \left(c\, \bar{W}^{\theta}(p, S) + h\, W^{P}(S, \lambda p, f(\cdot))\right)\right). \tag{3.28}$$

Therefore, according to (3.28), the strategy of the DC as a Stackelberg leader is to find the optimal base-stock level and the optimal joining probability (instead of price) (p^*, S^*) which maximizes the expected revenue function defined in (3.28). We demonstrate that for the optimal base-stock level S^*, the price $Pr^* = R - c\, \bar{W}^{\theta}(p, S)$ can induce the optimal joining probability p^*, i.e. customers will choose p^* as the Nash equilibrium, and the DC can extract the total surplus (the sum of all customers' surplus).

To obtain the optimal DC set (p^*, S^*), we first analyze the modified expected revenue function given by (3.28) when the joining probability is exogenous.

Lemma 6 *If the joining probability is exogenous, the modified expected revenue function has a maximum in S.*

The DC charges a price to obtain total surplus. According to Lemma 6, if the effective arrival rate to the DC is fixed, the DC can find a base-stock level which maximizes the

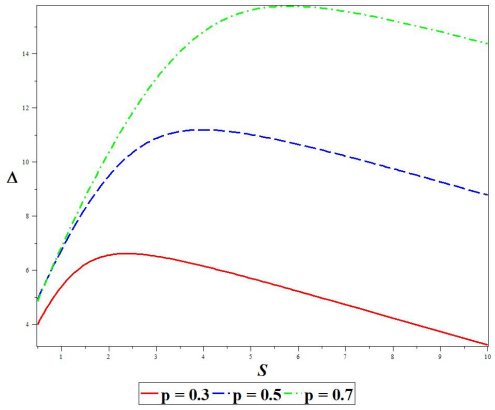

Figure 3.7: The effect of S on the modified expected revenue function for given p.

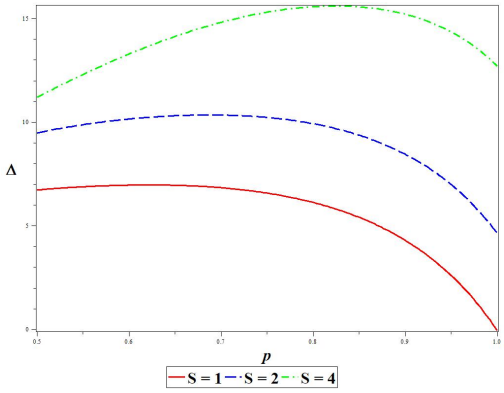

Figure 3.8: The effect of p on the modified expected revenue function for given S.

expected revenue function (Figure 3.7). If there is no price charged, the DC's expected cost (revenue) is increasing (decreasing) with respect to the base-stock level. However, since the higher base-stock level gives a higher surplus to customers and therefore more gains to the DC when a price is charged, there would be a unique maximum base-stock level for the DC (Figure 3.7).

Next, we investigate the behavior of the modified expected revenue function with respect to the joining probability when the base-stock level at the DC is given. This helps us to prove the existence of the optimal set (p^*, S^*) for the DC.

Lemma 7 *If the base-stock level at the DC is exogenous, the modified expected revenue function has a maximum in p.*

If the DC has no control on the base-stock level and only charges a price to control the effective arrival rate, there exists a joining probability which maximizes the expected revenue (Figure 3.8). This happens because the joining probability has two opposing effects on the modified expected revenue function; when the joining probability increases, according to Lemma 3, the inventory holding cost at the DC decreases, however, according to (3.23)

and (3.27), the charged price declines. Therefore, there should be a trade-off between these two opposing forces in order to find the optimal probability and price.

Now assume that the DC can control the arrival rate by choosing the optimal base-stock level and price. The DC knows the best response function of customers and gains total surplus by charging a price. The firm suffers due to the inventory holding cost but it benefits due to the charging price. Therefore, the DC is interested in choosing the optimal policy to maximize its expected revenue. In the following theorem we develop the optimal policy for the DC as the Stackelberg leader.

Theorem 5 *If the DC operates as a Stackelberg leader, there exists an optimal set* (p^*, S^*) *with corresponding* Pr^* *which maximizes the modified expected revenue function.*

According to Theorem 5, we can ensure that there always exists an optimal policy for the DC as the leader. The important fact here is that the optimal set (p^*, S^*) is equivalent to the optimal decision set (Pr^*, S^*). Since the DC gains total surplus by charging a price, it always keeps the joining probability at the equilibrium point. In this case, instead of choosing the optimal set (Pr^*, S^*), the DC can find the optimal set (p^*, S^*) and by charging a price equal to the expected utility of a joining customer, the equilibrium joining probability coincides with the optimal joining probability.

3.2.4 Observations

In this section, we discuss some interesting observations that are based on our numerical examples. We vary the parameters as follows: the production rate $\mu = 1$, holding cost at the DC $h \in \{1, 2, 3, 4\}$, deterministic transportation time $T \in [0, 5]$, the product reward $R \in [1, 5]$, waiting cost parameter $c \in [0.1, 5.5]$, the risk aversion degree $\theta \in [1, 3]$, the base-stock level in the warehouse $S_0 \in \{1, 2, 3, 4, 5\}$ and the arrival rate $\lambda \in [0.5, 0.9]$.

Observation 1 *The risk aversion (impatience) degree* θ *has a small effect on the joint optimal* (p^*, S^*) *for short transportation times.*

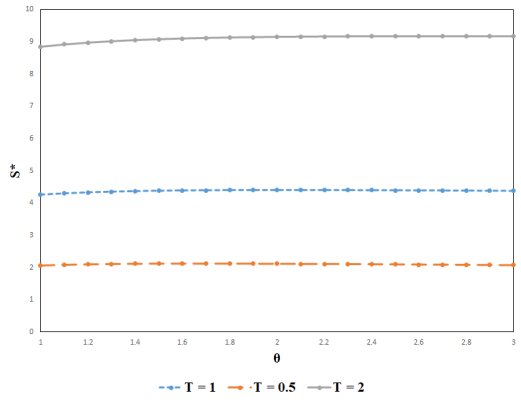

Figure 3.9: The effect of θ on the optimal base-stock.

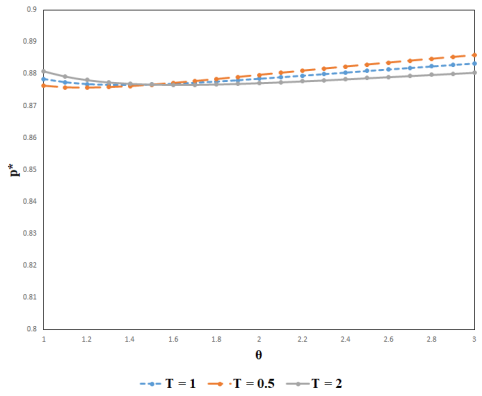

Figure 3.10: The effect of θ on the optimal joining probability.

As shown in Figure 3.9 and 3.10, changes in the optimal base-stock level and joining probability are small with respect to changes in the risk aversion degree. One reason for that is the charged price. Suppose the optimal set is (p^*, S^*) for a certain risk aversion degree; when the risk aversion degree changes, the waiting cost changes and customers start altering their joining probability. In this case, the DC can either change the base-stock level, price or both. Numerical results demonstrate that the DC tends to maintain the inventory cost at the same level, and to control the price in response to the risk aversion degree changes. In this case, the DC is more interested in avoiding large changes in arrival rate, resulting in small changes in the base-stock level. The DC can keep the base-stock level at the same level by keeping the arrival rate at the same level using the charged price. As discussed later, this observation helps us to find a good approximation for the base-stock level and price for all levels of risk aversion degrees.

Lemma 8 *If the risk aversion degree is equal to 1, the optimal base-stock level at the DC is given by the solution of* $W^Q(S-1,\lambda p, f(\cdot)) = 0$ *where,*

$$W^Q(S-1,\lambda p, f(\cdot)) = \frac{hZa(S,\lambda p, f(\cdot)) - c(1 - Za(S,\lambda p, f(\cdot)))^2}{\lambda pc(Za(S,\lambda p,f(\cdot)) - Za(S-1,\lambda p,f(\cdot)))}. \tag{3.29}$$

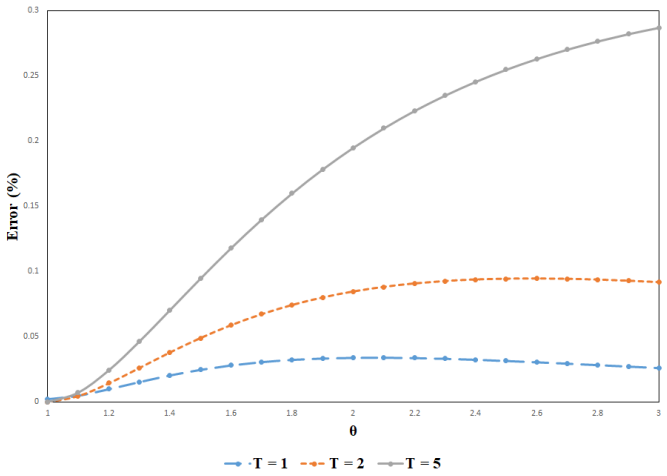

Figure 3.11: The relative error between the exact and approximation method.

Using (3.29) we can approximate the base-stock level for all values of the risk aversion degree. First, we obtain the optimal base-stock level, denoted by $\hat{S}^*$ when $\theta = 1$; then, for all other values of the risk aversion degree, we set the optimal base-stock level at $\hat{S}^*$. Since the optimal base-stock level is given for all θ, it is sufficient to find the optimal joining probability. Using $(p^*, \hat{S}^*)$, the DC sets a price to obtain all customers' surplus and force customers to choose p^* as the equilibrium joining probability. As shown in Figure 3.11, the gap in the expected revenue due to using this approximation instead of the exact value is lower than 1% in most cases even when the risk aversion degree increases. The importance of this result is that when the risk aversion degree changes, the DC can only manage the price to avoid losing expected revenue. In other words, it is not necessary for the DC to change the base-stock level and inventory holding facilities, while looking for the appropriate price.

Observation 2 *While the transportation time, T, increases, the optimal expected revenue and price are decreasing.*

As the transportation time increases, the lead-time increases. To respond to this change, the DC increases the optimal base-stock level reducing the chance of being out of stock

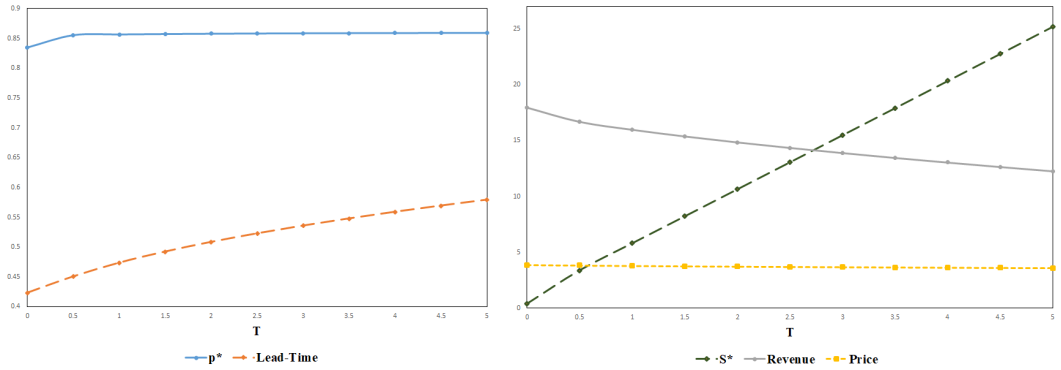

Figure 3.12: The effect of T on the p^*, S^*, optimal expected revenue, price and lead-time.

(Figure 3.12). However, it causes an increase in the inventory holding cost. Numerical results show that an increase in the base-stock level cannot completely compensate the effect of the transportation time. This forces the DC to slightly decrease the price to encourage people to join the system and place an order, causing a decline in the expected revenue (see Figure 3.12). Since the DC cannot compensate the effect of transportation time by controlling its (Pr, S), it is beneficial for the supply chain planner to expedite the transportation time much as possible.

Observation 3 *Depending on the value of c, the optimal joining probability, price and expected revenue can be increasing, decreasing nor neither with respect to the risk aversion degree. Also, the value of c has very low effect on the optimal base-stock behavior.*

As illustrated in Figure 3.13, the optimal base-stock level is smoothly increasing in the risk aversion degree and also the parameter c does not change this behavior. However, when the parameter c varies, the optimal joining probability shows different behaviors with respect to θ (Figure 3.14). With low levels of the parameter c, customers are less interested in joining the system when the risk aversion degree increases (see Figure 3.14). For high values of the parameter c, by increasing the risk aversion degree, the joining probability and expected revenue start increasing (consistent with Proposition 12). In short, we can conclude that the behavior of the joining probability, price and expected revenue, highly

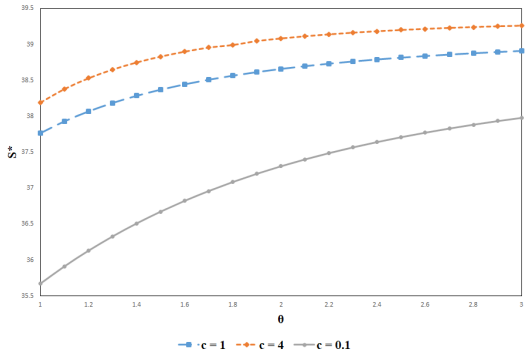

Figure 3.13: The effect of θ on the optimal base-stock with different values of the parameter c.

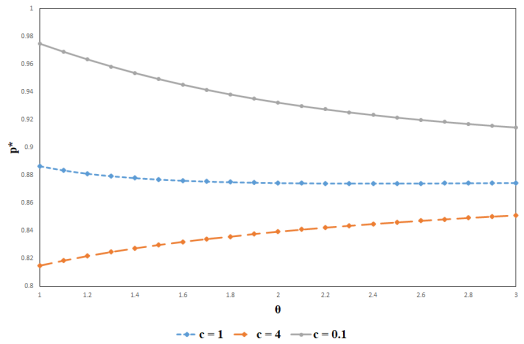

Figure 3.14: The effect of θ on the optimal joining probability with different values of the parameter c.

depends on the value of c (Figure 3.15). While the parameter c increases, the DC applies its control policy, (Pr, S), to manage the arrival rate. As shown in Figure 3.16, when the parameter c increases, the DC smoothly increases the base-stock level to reduce the lead-time and lower the price to compensate the customer losses imposed by the waiting cost. Numerical results demonstrate that the DC cannot completely compensate the effect of the waiting cost changes using (Pr^*, S^*) control policy, and consequently loses some revenue.

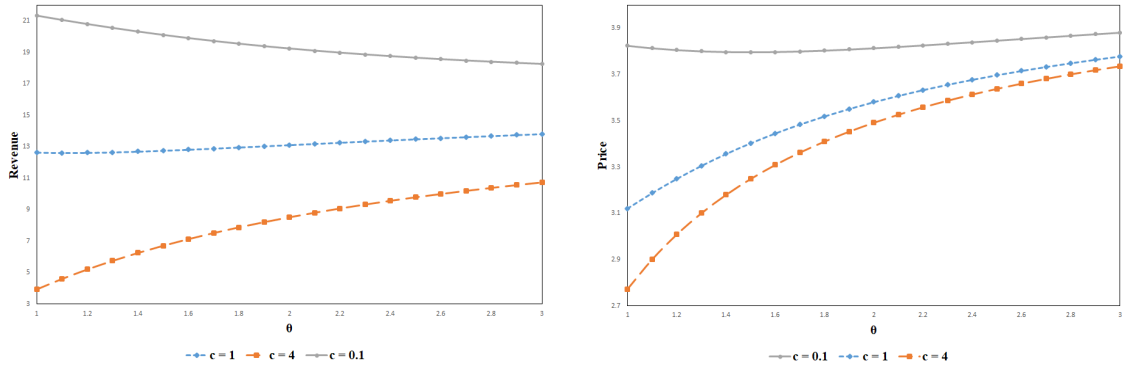

Figure 3.15: The effect of θ on the optimal expected revenue and price with different values of the parameter c.

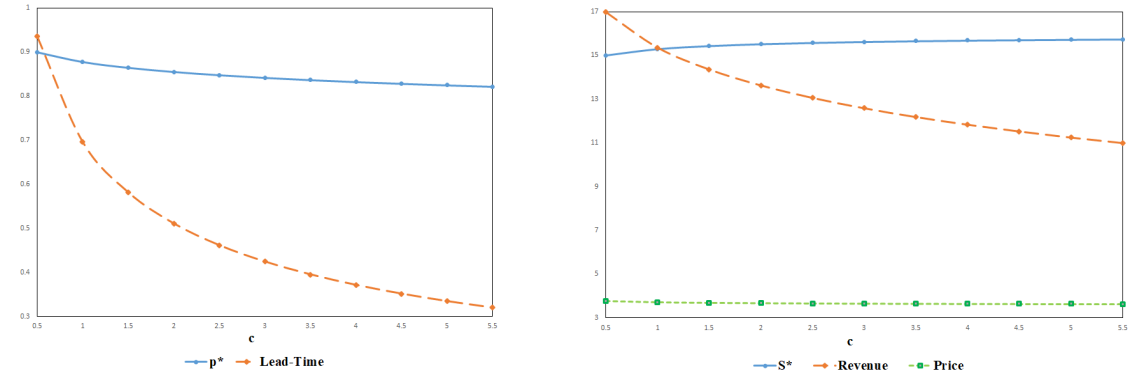

Figure 3.16: The effect of c on the optimal expected revenue, p^*, S^*, price and lead-time.

Chapter 4

Conclusions and Future Research

In this study we analyzed customer behavior and system manager's strategy in two different settings: (1) customers are served in a service system, or (2) they receive a product in a supply chain.

In the first model, we analyzed customer decisions regarding balking and reneging from a queue. To capture behavioral factors, we employed a non-linear concave-convex waiting cost function. To the best of our knowledge, this research is the first attempt to analyze customer decisions with respect to mixed-risk behavior in queue. We also studied the socially optimal strategy and focused on characterizing a control policy to eliminate the negative effects of customer externalities by proposing a joint abandonment threshold and pricing mechanism. Using the abandonment threshold policy, we demonstrated that customers choose this threshold as a Nash equilibrium, i.e. under this threshold no one will be better off by choosing a reneging time less than this abandonment threshold. In order to eliminate negative externalities imposed by customers, we proved that the planner must charge a fee as well as applying an abandonment threshold. We determined how a revenue maximizer can induce the socially optimal behavior and achieve the optimal profit by adopting an appropriate joint entrance-fee/abandonment-threshold mechanism. Our results showed that customer anticipation of the delay has considerable effects on her strategic behavior. More broadly, we illustrated the critical impact of an anticipation point

on the social welfare and expected revenue. From the firm's perspective, customers with a moderate anticipation point provide a higher expected revenue.

Our main contributions to the customer strategic behavior in a service system literature are as follows:

1. We considered a non-linear waiting cost function with mixed-risk behavior and an anticipation point. In this context we fully characterized the threshold structure of customer's decision.

2. We fully characterized the socially optimal balking and abandonment strategy and determined when partial or fully join strategy is optimal.

3. We demonstrated that pricing mechanism cannot induce the socially optimal behavior, and therefore we introduce a joint pricing and abandonment threshold mechanism to capture total surplus. We showed how this mechanism changes customer behavior in joining/reneging inducing a socially optimal behavior. We numerically illustrated that the firm's expected revenue is much greater when customers have a moderate anticipation point compared to an excessively high anticipation point.

Our model and results provide theoretical support for queue control policies. The findings also suggest that it is critical for firms to understand customer behavior with respect to delay cost and anticipation point. We demonstrated how an organization can achieve optimal profit by adopting an appropriate abandonment threshold and pricing mechanism. Our results showed that customer anticipation of the delay has considerable effects on self-maximization and socially optimal behavior.

In the second model, considering a two-echelon production-inventory system with endogenous lead-times and the base-stock policies, we obtained a closed-form expression to find the optimal base-stock level at the DC, when the arrival rate is exogenous. Then, we investigated customer behavior and the supply chain manager's strategy considering the endogenous arrival rate scenario. We considered a two-echelon production inventory system with a single DC and a manufacturer operating from a warehouse, in which the DC

acts as a Stackelberg leader and customers are the followers. Customers who are impatient and risk averse arrive at the DC according to a Poisson process. We assumed that risk averse customers use strategy (p) where they place an order with probability p. To the best of our knowledge, there is no study that analyzes the strategy of supply chain holder and equilibrium behavior of customers with respect to different degrees of risk aversion (impatience) in a multi-echelon inventory system.

Our main contributions to the customer strategic behavior in a production inventory system literature are as follows:

1. We investigated the effect of the risk aversion degree on the customer's expected utility function, equilibrium joining probability and optimal expected revenue. We demonstrated that for any level of risk aversion , there is an optimal policy, i.e. optimal price and base-stock level, for the DC.

2. We also showed that the price has a dominant effect on customer decision and the optimal expected revenue when the risk aversion degree changes. It means that for different levels of impatience, the price is a principal factor in controlling the system while the base-stock level can be kept unchanged.

This thesis can stimulate further research. There are many directions that can be explored in the future considering either service systems or production systems:

1. Service Systems

   In our research we consider a service system with a single service provider where rational homogeneous customers decide whether to join the queue and how long to wait before reneging. However, in future research, homogeneity assumption can be relaxed to consider heterogeneous customers. In this case, we may consider the following deviations: the anticipation time β and the service value V can follow certain distributions over the population. It means that customers have different anticipation toward delay and the server provides different values for customers.

Another direction is related to the service provider. First of all, the assumption on the service time distribution can be relaxed and we can assume that the service times are generally distributed. Moreover, we may consider server vacation times which are periods of time that the server is away from the queue. Also, assuming a finite waiting room with multiple servers can be other possible directions. We think these extensions are important areas of further research from both practical and theoretical aspects.

2. Production Systems

   In this research we analyze the optimal strategy of a single centralized decision maker (DC) who plays as a Stackelberge leader in a two-echelon production inventory system and provides a single product to homogeneous customers. Our model can be extended to include multiple products in further investigation. In addition, we can assume that customers have different risk aversion degrees and they also gain different values from receiving the product. In particular, in order to study the effect of customer heterogeneity on the supply chain planner's strategy, we can assume that the product reward R and the risk aversion degree θ follow certain distributions.

   Furthermore, the customer's strategy is a critical issue worth considering. In our research we assume that customers decide whether to join the system and place an order, however, future research can assume that customers are impatient and may cancel the order after waiting for some time. Therefore, future work may seek the effect of the both backorder and lost sale costs on the supply chain manager's strategy. Also, we can consider a decentralized supply chain where the DC and the manufacturer make decision based on their own benefit. In this case, future research question can focus on investigating appropriate coordination mechanisms between the manufacturer and the DC considering risk averse customers.

References

[1] Abouee-Mehrizi, H., O. Berman, H. Shavandi, and A. G. Zare (2011). An exact analysis of a joint production-inventory problem in two-echelon inventory systems. *Naval Research Logistics (NRL) 58*(8), 713–730.

[2] Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, Volume 55. Courier Corporation.

[3] Adiri, I. and U. Yechiali (1974). Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research 22*(5), 1051–1066.

[4] Afèche, P. (2013). Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management 15*(3), 423–443.

[5] Afèche, P., O. Baron, and Y. Kerner (2013). Pricing time-sensitive services based on realized performance. *Manufacturing & Service Operations Management 15*(3), 492–506.

[6] Afèche, P. and N. Sanajian (2013). Competition among risk-averse newsvendors. *Working paper*.

[7] Afèche, P. and V. Sarhangian (2015). Rational abandonment from priority queues: Equilibrium strategy and pricing implications. *Working paper*.

[8] Agrawal, V. and S. Seshadri (2000). Impact of uncertainty and risk aversion on price and order quantity in the newsvendor problem. *Manufacturing & Service Operations Management 2*(4), 410–423.

[9] Akan, M., B. ş. Ata, and T. Olsen (2012). Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research 60*(6), 1505–1519.

[10] Ata, B. and T. L. Olsen (2009). Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Operations Research 57*(3), 753–768.

[11] Axsäter, S. (1990). Simple solution procedures for a class of two-echelon inventory problems. *Operations Research 38*(1), 64–69.

[12] Barrer, D. (1957). Queuing with impatient customers and ordered service. *Operations Research 5*(5), 650–656.

[13] Berman, O. and J. A. Schnabel (1986). Mean-variance analysis and the single-period inventory problem. *International journal of systems science 17*(8), 1145–1151.

[14] Cachon, G. P. and P. Feldman (2011). Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management 13*(2), 244–260.

[15] Carmon, Z. and D. Kahneman (1996). The experienced utility of queuing: real time affect and retrospective evaluations of simulated queues. Technical report, Working paper, Duke University.

[16] Carmon, Z., J. G. Shanthikumar, and T. F. Carmon (1995). A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Science 41*(11), 1806–1815.

[17] Chen, X., M. Sim, D. Simchi-Levi, and P. Sun (2007). Risk aversion in inventory management. *Operations Research 55*(5), 828–842.

[18] Erlichman, J. and R. Hassin (2015). Strategic overtaking in a monopolistic m/m/1 queue. *IEEE Transactions on Automatic Control 60*(8), 2189–2194.

[19] Federgruen, A. (1993). Centralized planning models for multi-echelon inventory systems under uncertainty. *Handbooks in operations research and management science 4*, 133–173.

[20] Finkelstein, S. R., N. Liu, B. Jani, and D. Rosenthal (2014). Risk-seekers prefer speedy access to care over quality of care: The role of risk attitudes in health care utilization. *Working paper*.

[21] Guo, P. and P. Zipkin (2007). Analysis and comparison of queues with different levels of delay information. *Management Science 53*(6), 962–970.

[22] Guo, P. and P. Zipkin (2009). The impacts of customers'delay-risk sensitivities on a queue with balking. *Probability in the engineering and informational sciences 23*(03), 409–432.

[23] Hassin, R. (1985). On the optimality of first come last served queues. *Econometrica 53*(1), 201–02.

[24] Hassin, R. (1995). Decentralized regulation of a queue. *Management Science 41*(1), 163–173.

[25] Hassin, R. (2016). *Rational queueing*. CRC press.

[26] Hassin, R. and M. Haviv (1995). Equilibrium strategies for queues with impatient customers. *Operations Research Letters 17*(1), 41–45.

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

[28] Hassin, R. and A. Koshman (2015). Optimal control of a queue with high-low delay announcements: The significance of the queue. *Working paper*.

[29] Haviv, M. and Y. Ritov (2001). Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems 38*(4), 495–508.

[30] He, Q.-M., E. M. Jewkes, and J. Buzacott (2002). Optimal and near-optimal inventory control policies for a make-to-order inventory–production system. *European Journal of Operational Research 141*(1), 113–132.

[31] Janakiraman, N., R. J. Meyer, and S. J. Hoch (2011). The psychology of decisions to abandon waits for service. *Journal of Marketing Research 48*(6), 970–984.

[32] Kumar, A., M. A. Killingsworth, and T. Gilovich (2014). Waiting for merlot: Anticipatory consumption of experiential and material purchases. *Psychological Science 25*(10), 1924–1931.

[33] Kumar, S. and R. S. Randhawa (2010). Exploiting market size in service systems. *Manufacturing & Service Operations Management 12*(3), 511–526.

[34] Liu, N., S. R. Finkelstein, M. E. Kruk, and D. Rosenthal (2017). When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science*.

[35] Mandelbaum, A. and N. Shimkin (2000). A model for rational abandonments from invisible queues. *Queueing Systems 36*(1), 141–173.

[36] Mandelbaum, A. and S. Zeltyn (2013). Data-stories about (im) patient customers in tele-queues. *Queueing Systems 75*(2-4), 115–146.

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

[38] Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology 29*(1), 82–105.

[39] Shimkin, N. and A. Mandelbaum (2004). Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems 47*(1), 117–146.

[40] Simchi-Levi, D. and Y. Zhao (2007). Three generic methods for evaluating stochastic multi-echelon inventory systems. Technical report, Working paper, Massachusetts Institute of Technology, Cambridge.

[41] Sun, W. and S. Li (2012). Customer threshold strategies in observable queues with partial information of service time. *Information Computing and Applications*, 456–462.

[42] Svoronos, A. and P. Zipkin (1991). Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Science 37*(1), 68–83.

[43] Tekin, M. and S. Özekici (2015). Mean-variance newsvendor model with random supply and financial hedging. *IIE Transactions 47*(9), 910–928.

[44] Wang, K., N. Li, and Z. Jiang (2010). Queueing system with impatient customers: A review. In *Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on*, pp. 82–87. IEEE.

[45] Wang, Q. (2011). Control policies for multi-echelon inventory systems with stochastic demand. In *Supply chain coordination under uncertainty*, pp. 83–108. Springer.

[46] Yechiali, U. (1971). On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research 19*(2), 349–370.

[47] Yechiali, U. (1972). Customers' optimal joining rules for the GI/M/s queue. *Management Science 18*(7), 434–443.

[48] Zare, A. G., H. Abouee-Mehrizi, and O. Berman (2017). Exact analysis of the (R, Q) inventory policy in a two-echelon production–inventory system. *Operations Research Letters 45*(4), 308–314.

[49] Zipkin, P. H. (2000). *Foundations of inventory management*, Volume 2. McGraw-Hill New York.

[50] Zohar, E., A. Mandelbaum, and N. Shimkin (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science 48*(4), 566–583.

APPENDICES

Appendix A

Proofs

A.1 Control Mechanism in a Queue with Joint Balking and Reneging Strategy

A.1.1 Proof of Proposition 1

Recall that $G_T^p(x)$ is the distribution of offered waiting time (including service time). An arriving customer wants to maximize her expected utility by choosing an appropriate reneging time regarding the offered waiting time. When all other customers follow the T strategy, the expected utility function given in (2.4) can be rewritten as:

$$U_T(p,\tau) = VG_T^p(\tau) - \int_0^\tau C(x)\ g_T^p(x)\,dx - C(\tau)\ (1 - G_T^p(\tau)). \tag{A.1}$$

Applying integration by parts, we get

$$U_T(p,\tau) = VG_T^p(\tau) - \int_0^\tau \frac{\partial C(x)}{\partial x}(1 - G_T^p(x))dx. \tag{A.2}$$

According to (A.2), the first derivative of the expected utility with respect to the customer's reneging time τ is:

$$\frac{\partial U_T(p,\tau)}{\partial \tau} = V\frac{\partial G_T^p(\tau)}{\partial \tau} - \frac{\partial C(\tau)}{\partial \tau}(1 - G_T^p(\tau)). \tag{A.3}$$

Therefore, local extremums occur where $\frac{\partial U_T(p,\tau)}{\partial \tau} = 0$ for $\tau > 0$, which results in $\frac{\frac{\partial G_T^p(\tau)}{\partial \tau}}{(1-G_T^p(\tau))} = \frac{\frac{\partial C(\tau)}{\partial \tau}}{V}$.

A.1.2 Proof of Corollary 1

According to (2.5), since the hazard rate function is bounded by μ, i.e., $h_T(\tau) \leq \mu$, and the cost-reward ratio is monotone increasing for all $\tau \geq \beta$, and $\lim_{\tau\to\infty} \gamma(\tau) = \infty$, we can conclude that when $\tau \to \infty$, the hazard rate function is less than the cost-reward ratio function (the cost-reward ratio will never cross the hazard rate function at a point greater than μ) indicating that each arriving customer will eventually renege the system after spending a certain amount of time. The best scenario for an arriving customer is when no one is in the system and the hazard rate is equal to μ. In this case, an arriving customer will renege the system when the cost-reward ratio function passes the hazard rate function from below. Therefore, the maximum abandonment threshold is where $\gamma(\tau) = \mu$. Using (2.5) we get the maximum reneging time given in (2.7).

A.1.3 Proof of Proposition 2

We investigate the roots of $h_T(\tau) = \gamma(\tau)$ to find the local extremums of an individual customer's expected utility function. According to (2.5), since the the cost-reward ration function is monotone decreasing before reaching $\tau = \beta$ and monotone increasing in τ after passing $\tau = \beta$, we investigate the roots of $h_T(\tau) = \gamma(\tau)$, when $\tau \leq \beta$ and $\tau > \beta$, respectively. Since the cost-reward ratio function is monotone decreasing in $\tau < \beta$, without loss of generality, we can assume that T is always greater than β. A candidate for a local extremum is a point τ where $h_T(\tau) = \gamma(\tau)$, i.e.,

$$\begin{cases} \frac{\mu\,(\mu-\lambda\,p)}{\mu-\lambda\,p\mathrm{e}^{-(\mu-\lambda\,p)(T-\tau)}} = \frac{c(\tau-\beta)^{2\,\alpha-2}(2\,\alpha-1)}{V}, & \tau < T \\ \mu = \frac{c(\tau-\beta)^{2\,\alpha-2}(2\,\alpha-1)}{V}, & T \leq \tau \end{cases} . \tag{A.4}$$

1. Suppose that $\tau \leq \beta$. Note that the cost-reward function is monotone decreasing in this interval and $\gamma(\beta) = 0$. Since according to (2.5), $h_T(\tau)$ is a non-decreasing

positive function, if $h(0) < \gamma(0)$ we can conclude that $h_T(\tau) = \gamma(\tau)$ has exactly one root at which the sign of $h_T(\tau) - \gamma(\tau)$ is changed from negative to positive. This means that τ is a local minimum point. Solving $h(0) < \gamma(0)$, we get

$$V < \frac{\left(\mu - \lambda\, p \mathrm{e}^{-(\mu-\lambda\, p)T}\right) c\,(2\,\alpha - 1)\,\beta^{2\,\alpha-2}}{\mu\ (\mu - \lambda\, p)}. \tag{A.5}$$

We denote the right-hand side of (A.5) by $\bar{V}_1(\beta, p, T)$.

2. Suppose that $\tau > \beta$. The cost-reward ratio is monotone increasing in this interval. Since the hazard rate function is a positive non-decreasing function and it is constant after time T, i.e. $h(T) = \mu$, then $h_T(\tau) = \gamma(\tau)$ can have one or more roots in this interval (note that $\gamma(\beta) = 0$ and $\lim_{\tau\to\infty} \gamma(\tau) = \infty$). Reconsidering (A.4),

$$1 - \frac{V\,(\mu - \lambda\, p)\,(\tau - \beta)^{-2\,\alpha+2}}{c\,(2\,\alpha - 1)} = \begin{cases} \frac{\lambda\, p \mathrm{e}^{-(\mu-\lambda\, p)(T-\tau)}}{\mu} & \beta \le \tau < T \\ \lambda p/\mu & T \le \tau \end{cases}. \tag{A.6}$$

The left hand side of (A.6), denoted by $I_1(\tau)$, is concave and monotone increasing for $\tau \ge \beta$ and the right hand side, denoted by $I_2(\tau)$, is convex and monotone increasing positive function for $\tau \le T$ bounded by $\lambda p/\mu$, and it is constant after time T. Also, we have,

$$I_1(\tau) = \begin{cases} -\infty, & \tau \to \beta \\ 1 & \tau \to \infty \end{cases},$$

and

$$I_2(\tau) = \begin{cases} \frac{\lambda\, p \mathrm{e}^{-(\mu-\lambda\, p)(T-\beta)}}{\mu}, & \tau \to \beta \\ \lambda p/\mu \le 1 & \tau \to \infty \end{cases}. \tag{A.7}$$

Therefore, we can conclude that $I_1(\tau) = I_2(\tau)$ has at least one root and at most three roots for $\tau \ge \beta$. In this case, if $I_1(T) > I_2(T)$ (which means $h_T(T) < \gamma(T)$), then (A.6) has at most one root, resulting in changing the sign of $h_T(\tau) - \gamma(\tau)$ from positive to negative (local maximum point). However, if $I_1(T) \le I_2(T)$ (which means $h_T(T) \ge \gamma(T)$), then (A.6) has one or three roots (note that $\lambda p/\mu \le 1$), i.e., the cost-reward ratio function crosses the hazard rate function at one or three

points. In this case, considering $I_1(T) \leq I_2(T)$ and using (A.7), if $I_1(\tau)$ crosses $I_2(\tau)$ at exactly one point, this point must be greater than T, resulting in changing the sign of $h_T(\tau) - \gamma(\tau)$ from positive to negative (local maximum point); otherwise, in the case of having three roots, at the first ($\tau < T$) and last point ($\tau \geq T$), the sign of $h_T(\tau) - \gamma(\tau)$ is changed from positive to negative (local maximum point). Solving $h(T) < \gamma(T)$, we get $V < \bar{V}_2(\beta, T)$, where

$$\bar{V}_2(\beta, T) = \frac{c\,(2\,\alpha - 1)\,(T - \beta)^{2\,\alpha - 2}}{\mu}.$$

A.1.4 Proof of Corollary 2

Note that for any $\tau \geq T$ the hazard rate function is equal to μ. According to the proofs of Proposition 2 and Corollary 1, the results are immediate.

A.1.5 Proof of Proposition 3

1. Suppose p is given. In this case the optimal T for the social welfare function $\lambda p U_T(p, T)$, is the same for $U_T(p, T)$. Therefore, instead of the social welfare function, we can consider the expected utility function in our analysis. Based on (A.2),

$$\frac{\partial U_T(p, T)}{\partial T} =$$

$$V \frac{\partial G_T^p(T)}{\partial T} + c \int_0^T (2\alpha - 1)\,(x - \beta)^{2\alpha - 2} \left(\frac{\partial G_T^p(x)}{\partial T}\right) dx - c(2\alpha - 1)\,(T - \beta)^{2\alpha - 2}(1 - G_T^p(T)). \tag{A.8}$$

Considering (2.2),

$$\frac{\partial G_T^p(x)}{\partial T} = -\frac{\mu\,\left(1 - e^{-(\mu - \lambda\,p)x}\right)\lambda\,p\,(\mu - \lambda\,p)\,e^{-(\mu - \lambda\,p)T}}{\left(\mu - \lambda\,p e^{-(\mu - \lambda\,p)T}\right)^2},$$

$$\frac{\partial G_T^p(T)}{\partial T} = \frac{\mu\,(\mu - \lambda\,p)^2\,e^{-(\mu - \lambda\,p)T}}{\left(\mu - \lambda\,p e^{-(\mu - \lambda\,p)T}\right)^2}. \tag{A.9}$$

Therefore,

$$\frac{\partial U_T\,(p,T)}{\partial T} = F_1(T)\,(F_2(T) - F_3(T)), \tag{A.10}$$

where

$$\begin{aligned}
&F_1(T) = \frac{\mu\,(\mu - \lambda\,p)\,e^{-(\mu-\lambda\,p)T}}{(\mu - \lambda\,pe^{-(\mu-\lambda\,p)T})^2},\\
&F_2(T) = \\
&\lambda pc\left(\int_0^T (2\alpha-1)(x-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\,p)x}dx + \frac{(2\alpha-1)}{\mu}(T-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\,p)T}\right),\\
&F_3(T) = \lambda cp\left((T-\beta)^{2\alpha-1} + \beta^{2\alpha-1}\right) + c(2\alpha-1)(T-\beta)^{2\alpha-2} - V\,(\mu - \lambda\,p)\,.
\end{aligned} \tag{A.11}$$

Note that $F_1(T) > 0$ since $\mu > \lambda p$. Therefore, the behavior (sign and number of extreme points) of $\frac{\partial U_T(p,T)}{\partial T}$ is derived by $F_2(T) - F_3(T)$. We next investigate the behavior of $F_2(T) - F_3(T)$. Considering (A.11), we have,

$$\begin{aligned}
\frac{dF_2(T)}{dT} &= \frac{\lambda cp}{\mu}\; e^{(-\mu+\lambda\,p)T}(2\alpha-1)(T-\beta)^{2\alpha-3}\,(2\alpha-2+\lambda\,p(T-\beta))\,,\\
\frac{dF_3(T)}{dT} &= c(2\alpha-1)(T-\beta)^{2\alpha-3}(2\alpha-2+\lambda p(T-\beta)).
\end{aligned} \tag{A.12}$$

Note that since $\frac{dF_2(T)}{dT} = \frac{\lambda p}{\mu}\; e^{(-\mu+\lambda\,p)T}\frac{dF_3(T)}{dT}$ and $\mu > \lambda p$, then

$$|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|. \tag{A.13}$$

Also, $\frac{dF_2(T)}{dT} = \frac{dF_3(T)}{dT} = 0$ if and only if $T = \beta - \frac{2\,(\alpha-1)}{\lambda\,p}$ or $T = \beta$. We denote these two points by T_1 and T_2, respectively, i.e., $T_1 = \beta - \frac{2\,(\alpha-1)}{\lambda\,p}$ and $T_2 = \beta$. Considering that α is an integer greater than one, T_1 and T_2 are the local maximum and minimum points, respectively, of both $F_2(T)$ and $F_3(T)$. To characterize the behavior of $\frac{\partial U_T(p,T)}{\partial T}$, we need to identify the conditions under which $F_2(T)$ and $F_3(T)$ cross each other (due to A.10). We only need to examine the relative position of $F_1(T)$ and $F_2(T)$ at $T = 0$, $T = T_1$, and $T = T_2$ (see Figures A.1 and A.2) since $|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|$, $\frac{dF_2(T)}{dT} = \frac{dF_3(T)}{dT} = 0$ at T_1 and T_2, and $\lim_{T\to\infty} F_2(T) < \lim_{T\to\infty} F_3(T)$ (based on A.11).

We next examine $F_1(T)$ and $F_2(T)$ at $T = 0$, $T = T_1$, and $T = T_2$.

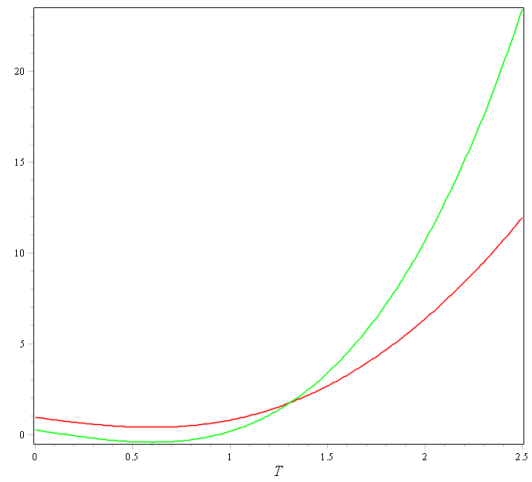

Figure A.1: The behavior of $F_2(T)$ (red curve) and $F_3(T)$ (green curve) in T when $\beta < \frac{2\,(\alpha+1)}{\lambda\,p}$.

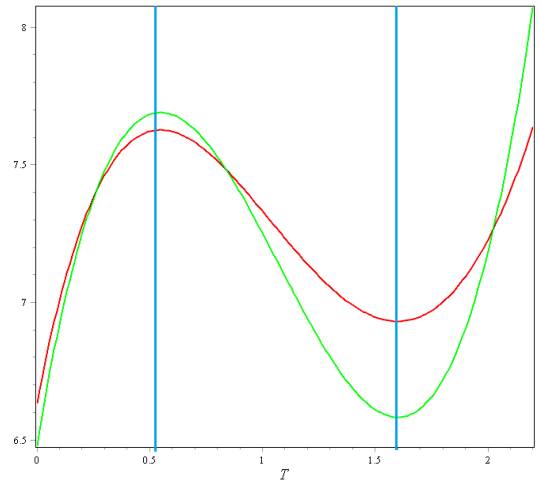

Figure A.2: The behavior of $F_2(T)$ (red curve) and $F_3(T)$ (green curve) in T when $\beta > \frac{2\,(\alpha+1)}{\lambda\,p}$.

(i) Consider $T = 0$:

$$F_2(0) = \frac{\lambda\,pc\,(2\,\alpha-1)\,\beta^{2\,\alpha-2}}{\mu},$$
$$F_3(0) = c\,(2\,\alpha-1)\,\beta^{2\,\alpha-2} - V\,(\mu-\lambda\,p)\,. \tag{A.14}$$

Therefore, $F_2(0) < F_3(0)$ if $V < \frac{c(2\,\alpha-1)\beta^{2\,\alpha-2}}{\mu}$. Let

$$V_2(\beta) = \frac{c\,(2\,\alpha-1)\,\beta^{2\,\alpha-2}}{\mu}. \tag{A.15}$$

(ii) Consider $T = T_2 = \beta$:

$$F_2(\beta) = \lambda\,pc\int_0^{\beta}(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{-(\mu-\lambda\,p)x}\,\mathrm{d}x,$$
$$F_3(\beta) = \lambda\,pc\beta^{2\,\alpha-1} - V\,(\mu-\lambda\,p)\,. \tag{A.16}$$

Therefore, $F_2(\beta) < F_3(\beta)$ if

$$V < \frac{\lambda\,pc\left(\int_0^{\beta}(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{(\lambda\,p-\mu)x}\,\mathrm{d}x - \beta^{2\,\alpha-1}\right)}{\lambda\,p-\mu}.$$

Let

$$V_1(\beta, p) = \frac{\lambda\, pc\beta^{2\,\alpha-1}\left(1 - M\left(1,\, 2\,\alpha,\, -\left(\mu - \lambda\, p\right)\beta\right)\right)}{\mu - \lambda p}, \tag{A.17}$$

where $M(a, b, z)$ is is the confluent hypergeometric function of the first kind defined as

$$M\left(1,\, 2\,\alpha,\, \left(\lambda\, p - \mu\right)\beta\right) = \beta^{1-2\,\alpha}\int_0^{\beta}(2\,\alpha - 1)\left(x - \beta\right)^{2\,\alpha-2}\mathrm{e}^{-(\mu-\lambda\, p)x}\,\mathrm{d}x. \tag{A.18}$$

(iii) Consider $T = T_1 = \beta - 2\,\frac{\alpha-1}{\lambda\, p}$:

$$\begin{aligned}
F_2(\beta - 2\,\frac{\alpha-1}{\lambda\, p}) &= \lambda pc\left(\int_0^{\beta-2\,\frac{\alpha-1}{\lambda\, p}}(2\,\alpha-1)\left(x-\beta\right)^{2\,\alpha-2}\mathrm{e}^{-(\mu-\lambda\, p)x}\,\mathrm{d}x\right)\\
&+ \lambda pc\left(\frac{2\,\alpha-1}{\mu}\left(-2\,\frac{\alpha-1}{\lambda\, p}\right)^{2\,\alpha-2}\mathrm{e}^{\left(\beta-2\,\frac{\alpha-1}{\lambda\, p}\right)-(\mu-\lambda\, p)}\right),\\
F_3(\beta - 2\,\frac{\alpha-1}{\lambda\, p}) &= \lambda\, pc\left(\left(-2\,\frac{\alpha-1}{\lambda\, p}\right)^{2\,\alpha-1} + \beta^{2\,\alpha-1}\right)\\
&+ c\,(2\,\alpha-1)\left(-2\,\frac{\alpha-1}{\lambda\, p}\right)^{2\,\alpha-2} - V\left(\mu - \lambda\, p\right).
\end{aligned}$$

Therefore, $F_2(\beta - 2\,\frac{\alpha-1}{\lambda\, p}) < F_3(\beta - 2\,\frac{\alpha-1}{\lambda\, p})$ if

$$\begin{aligned}
V < \quad & -\frac{c}{\mu\, -(\mu - \lambda\, p)}\\
& \left(-\left(2\,\alpha-1\right)\left(\lambda\, p\mathrm{e}^{\left(\beta-2\,\frac{\alpha-1}{\lambda\, p}\right)-(\mu-\lambda\, p)} - \mu\right)\left(\frac{-2\,\alpha+2}{\lambda\, p}\right)^{2\,\alpha-2}\right.\\
& \left.+p\left(\left(\frac{-2\,\alpha+2}{\lambda\, p}\right)^{2\,\alpha-1} - J + \beta^{2\,\alpha-1}\right)\mu\,\lambda\ \right),
\end{aligned} \tag{A.19}$$

where,

$$J = \int_0^{\beta-2\,\frac{\alpha-1}{\lambda\, p}}(2\,\alpha-1)\left(x-\beta\right)^{2\,\alpha-2}\mathrm{e}^{(\lambda\, p-\mu)x}\,\mathrm{d}x =$$

$$\mathrm{e}^{-(\mu-\lambda\,p)\beta}\int_{2\,\frac{\alpha-1}{\lambda\,p}}^{\beta}(2\,\alpha-1)\,y^{2\,\alpha-2}\mathrm{e}^{--(\mu-\lambda\,p)y}\,\mathrm{d}y. \tag{A.20}$$

Recall that $\Gamma(\alpha,z)$ denotes the incomplete Gamma function given by

$$\Gamma(\alpha,z)=\int_{z}^{\infty}t^{\alpha-1}e^{-t}dt, \tag{A.21}$$

where $\Gamma(\alpha)=\Gamma(\alpha,0)$. Considering the definition of $\Gamma(\alpha,z)$ (see e.g., Abramowitz and Stegun, 1964), we get

$$\int_{0}^{-a}y^{2\,\alpha-2}\mathrm{e}^{-y}\,\mathrm{d}y=-a\left(-\frac{\Gamma\,(2\,\alpha)}{(2\,\alpha-1)\,a}+\frac{(-a)^{2\,\alpha}\,\mathrm{e}^{a}}{(2\,\alpha-1)\,a^{2}}+\frac{\Gamma\,(2\,\alpha,-a)}{(2\,\alpha-1)\,a}\right). \tag{A.22}$$

Let $k_1=2\,\frac{(\alpha-1)(\mu-\lambda\,p)}{\lambda\,p}$ and $k_2=\beta\,\left(\mu-\lambda\,p\right)$. Substituting (A.22) in (A.20), we get

$$\begin{aligned}
J &= \frac{\mathrm{e}^{-(\mu-\lambda\,p)\beta}\left({k_1}^{2\,\alpha}\mathrm{e}^{k_1}k_2-\left({k_2}^{2\,\alpha}\mathrm{e}^{k_2}+k_2\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)\right)k_1\right)}{-\left(\mu-\lambda\,p\right)^{2\,\alpha-1}k_1k_2}\\
&= \frac{{k_1}^{2\,\alpha}\mathrm{e}^{k_1-k_2}k_2-k_1{k_2}^{2\,\alpha}-\mathrm{e}^{-k_2}k_2\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)k_1}{-\left(\mu-\lambda\,p\right)^{2\,\alpha-1}k_2k_1}\\
&= \beta^{2\,\alpha-1}-\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-1}\mathrm{e}^{k_1-k_2}+\frac{\mathrm{e}^{-k_2}\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)}{\left(\mu-\lambda\,p\right)^{2\,\alpha-1}}.
\end{aligned} \tag{A.23}$$

Substituting (A.23) in (A.19), we get $F_2(\beta - 2\,\frac{\alpha-1}{\lambda\,p}) < F_3(\beta - 2\,\frac{\alpha-1}{\lambda\,p})$ if

$$
\begin{aligned}
V \;<\; & -\frac{c\,(2\,\alpha-1)\left(\lambda\,p\mathrm{e}^{k_1-k_2}-\mu\right)}{\mu\,(\mu-\lambda\,p)}\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-2} \\
& + \frac{cp\lambda}{\mu-\lambda\,p}\left(-\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-1}+\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-1}\mathrm{e}^{k_1-k_2}\right. \\
& \left.-\frac{\mathrm{e}^{-k_2}k_2\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)}{\beta\,(\mu-\lambda\,p)^{2\,\alpha}}\right) \\
= \; & \frac{c}{\mu\,(\mu-\lambda\,p)} \\
& \left(-\left(2\,\alpha-1\right)\lambda\,p\mathrm{e}^{k_1-k_2}\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-2}+\mu\,\left(2\,\alpha-1\right)\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-2}\right) \\
& +\frac{c}{\mu\,(\mu-\lambda\,p)} \\
& \left(-p\mu\,\lambda\,\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-1}+\right. \\
& \left.p\mu\,\lambda\,\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-1}\mathrm{e}^{k_1-k_2}-\frac{p\mathrm{e}^{-k_2}k_2\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)\mu\,\lambda}{\beta\,(\mu-\lambda\,p)^{2\,\alpha}}\right) \\
= \; & \frac{c\left(\left(2\,\left(\alpha-1\right)\left(-\lambda\,p+\mu\right)-\lambda\,p\right)\mathrm{e}^{k_1-k_2}+\mu\right)}{\mu\,(\mu-\lambda\,p)}\left(2\,\frac{\alpha-1}{\lambda\,p}\right)^{2\,\alpha-2} \\
& -\frac{c\lambda\,p\mathrm{e}^{-k_2}\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)}{(\mu-\lambda\,p)^{2\,\alpha}} \\
= \; & c\left(\frac{\beta}{k_2}\right)^{2\,\alpha-1} \\
& \left(\frac{{k_1}^{2\,\alpha-2}\left(\lambda\,p\left(k_1-1\right)\mathrm{e}^{k_1-k_2}+\mu\right)}{\mu}-\frac{\lambda\,p\beta\,\mathrm{e}^{-k_2}\left(\Gamma\left(2\,\alpha,-k_2\right)-\Gamma\left(2\,\alpha,-k_1\right)\right)}{k_2}\right).
\end{aligned}
\tag{A.24}
$$

Let $V_3(\beta,p) = c\left(\frac{\beta}{k_2}\right)^{2\,\alpha-1}\left(\frac{{k_1}^{2\,\alpha-2}\left(\lambda\,p(k_1-1)\mathrm{e}^{k_1-k_2}+\mu\right)}{\mu}-\frac{\lambda\,p\beta\,\mathrm{e}^{-k_2}\left(\Gamma(2\,\alpha,-k_2)-\Gamma(2\,\alpha,-k_1)\right)}{k_2}\right)$. We next examine the behavior of $U_T\,(p,T)$ considering two cases: 1) $T_1 \leq 0$, 2) $T_1 > 0$. Recall that T_1 and T_2 are the local maximum and minimum points, respectively, of both $F_2(T)$ and $F_3(T)$.

(a) Suppose $T_1 \leq 0$, i.e., $\beta < \frac{2\,(\alpha+1)}{\lambda\, p}$.

- Considering that T_2 is the global minimum of both $F_2(T)$ and $F_3(T)$ as well as the fact that $|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|$ (due to (A.13)), if $F_2(0) < F_3(0)$ and $F_2(\beta) > F_3(\beta)$, the equation $F_2(T) - F_3(T) = 0$ has two roots such that it is positive between these roots and negative otherwise. Note that $F_2(0) < F_3(0)$ and $F_2(\beta) > F_3(\beta)$ if and only if $V_1(\beta, p) < V < V_2(\beta)$ due to (i) and (ii). Thus, considering (A.10), we conclude that if $V_1(\beta, p) < V < V_2(\beta)$, then $U_T\,(p, T)$ is bimodal and decreasing at $T = 0$.
- On the other hand, if $V > V_2(\beta)$, then $F_2(0) > F_3(0)$. Therefore, according to (A.12) and (A.13), $F_2(T) - F_2(T) = 0$ has exactly one positive root at which the function is decreasing. Considering (A.11), we conclude that if $V > V_2(\beta)$, $U_T\,(p, T)$ is unimodal in T (≥ 0) and has a positive maximum point.
- If $V < V_2(\beta)$ and $V < V_1(\beta, p)$ or equivalently $F_2(0) > F_3(0)$ and $F_2(\beta) < F_3(\beta)$ (based on (i) and (ii)), then the equation $F_2(T) - F_3(T)$ does not have a nonnegative real-value solution. This is due to $|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|$. Therefore if $V < \min\,(V_2(\beta), V_1(\beta, p))$, according to (A.11) the expected utility function $U_T\,(p, T)$ is monotone decreasing in T.

(b) Suppose that T_1 and T_2 are positive.

- If $V_1(\beta, p) < V < V_2(\beta)$ or equivalently $F_2(0) < F_3(0)$ and $F_2(\beta) > F_3(\beta)$, according to (A.12) and (A.13), $F_2(T) - F_3(T) = 0$ has two roots; thus, $U_T\,(p, T)$ is bimodal and decreasing at $T = 0$.
- If $V_2(\beta, p) < V < V_3(\beta, p)$ and $V_1(\beta, p) < V$ or equivalently $F_2(0) > F_3(0)$, $F_2(\beta - \frac{2\,(\alpha+1)}{\lambda\, p}) < F_3(\beta - \frac{2\,(\alpha+1)}{\lambda\, p})$, and $F_2(\beta) > F_3(\beta)$, then according to (A.12) and (A.13), $F_2(T) - F_3(T) = 0$ has three roots. Based on (A.11), $U_T\,(p, T)$ has two positive maximum points and one minimum point.
- When $V_3(\beta, p) < V$ or $V_2(\beta, p) < V < V_1(\beta, p)$, $F_2(\beta - \frac{2\,(\alpha+1)}{\lambda\, p}) > F_3(\beta - \frac{2\,(\alpha+1)}{\lambda\, p})$ or $F_2(0) > F_3(0)$ and $F_2(\beta) < F_3(\beta)$ due to (i), (ii) and (iii). Then, according

to (A.12), $F_2(T) - F_3(T)$ has one root resulting in $U_T\,(p, T)$ to be unimodal in T and there is a positive maximum point.

- If $V < V_2(\beta)$ and $V < V_1(\beta, p)$ or equivalently $F_2(0) > F_3(0)$ and $F_2(\beta) < F_3(\beta)$, similar to the case $T_1 \leq 0$, $U_T\,(p, T)$ is monotone decreasing in T.

2. Before proving $V_1(\beta, p)$ and $V_3(\beta, p)$ are increasing in p and β, we provide a lower and an upper bound for the confluent hypergeometric function $M\,(1,\, 2\,\alpha,\, -\,(\mu - \lambda\,p)\,\beta)$.

Lemma 9 *For* $M\,(1,\, 2\,\alpha,\, (\lambda\,p - \mu)\,\beta)$ *we have,*

$$\frac{2\,\alpha - 1}{(\mu - \lambda p)\,\beta + 2\,\alpha - 1} < M\,(1,\, 2\,\alpha,\, -\,(\mu - \lambda\,p)\,\beta) < \frac{2\,\alpha}{(\mu - \lambda p)\,\beta + 2\,\alpha}. \tag{A.25}$$

Proof.

Using the definition of $M\,(1,\, 2\,\alpha,\, -\,(\mu - \lambda\,p)\,\beta)$ given in (A.18) we have,

$$\frac{\partial M\,(1,\, 2\,\alpha,\, -\,(\mu - \lambda\,p)\,\beta)}{\partial p} =$$

$$\frac{\lambda\,\left((-\beta\,\lambda\,p + \beta\,\mu + 2\,\alpha - 1)\;\; M\,(1,\, 2\,\alpha,\, (\lambda\,p - \mu)\,\beta) - 2\,\alpha + 1\right)}{\mu - \lambda\,p} > 0. \tag{A.26}$$

The last inequality is due to $\mu > \lambda p$. Therefore, the lower bound is obtained as,

$$\frac{2\,\alpha - 1}{(\mu - \lambda p)\,\beta + 2\,\alpha - 1} < M\,(1,\, 2\,\alpha,\, -\,(\mu - \lambda\,p)\,\beta).$$

We next obtain the upper bound. Note that for any $x > 0$ we have $e^{-x} < 1/(x+1)$. Thus,

$$\int_0^\beta (2\,\alpha - 1)\,(x - \beta)^{2\,\alpha - 2}\,((\mu - \lambda p)x + 1)\,\mathrm{e}^{-(\mu - \lambda\,p)x}\,\mathrm{d}x <$$

$$\int_0^\beta (2\,\alpha - 1)\,(x - \beta)^{2\,\alpha - 2}\,\mathrm{d}x = \beta^{2\alpha - 1}. \tag{A.27}$$

Multiplying both sides by $\lambda\beta^{-(2\alpha-1)}$ we get,

$$\begin{aligned}&\lambda\beta^{-(2\alpha-1)}\left(\int_0^{\beta}(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,((\mu-\lambda p)x)\,\mathrm{e}^{-(\mu-\lambda\,p)x}\,\mathrm{d}x\right.\\&\left.+\int_0^{\beta}(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{-(\mu-\lambda\,p)x}\,\mathrm{d}x\right)\\&<\lambda.\end{aligned}$$

Therefore, considering (A.18), we get

$$\frac{\partial M\,(1,\,2\,\alpha,\,-\,(\mu-\lambda\,p)\,\beta)}{\partial p}+\frac{\lambda}{\mu-\lambda p}M\,(1,\,2\,\alpha,\,-\,(\mu-\lambda\,p)\,\beta)<\frac{\lambda}{\mu-\lambda p}. \tag{A.28}$$

Substituting (A.26) in (A.28), we get

$$\frac{\lambda\,\left((-\beta\,\lambda\,p+\beta\,\mu+2\,\alpha-1)\;\;M\,(1,\,2\,\alpha,\,(\lambda\,p-\mu)\,\beta)-2\,\alpha+1\right)}{\mu-\lambda\,p}<$$

$$\frac{\lambda}{\mu-\lambda p}\left(1-M\,(1,\,2\,\alpha,\,-\,(\mu-\lambda\,p)\,\beta)\right).$$

Therefore,

$$M\,(1,\,2\,\alpha,\,-\,(\mu-\lambda\,p)\,\beta)<\frac{2\,\alpha}{(\mu-\lambda p)\,\beta+2\,\alpha}.\ \blacksquare$$

(i) We first prove that $V_1(\beta,p)$ is increasing in p. Considering (A.17), we get

$$\frac{\partial V_1(\beta,p)}{\partial p}=(\lambda\beta^{2\,\alpha-1}c)$$

$$\frac{\left((p^2\beta\,\lambda^2-p\,(\beta\,\mu+2\,\alpha-1)\,\lambda-\mu)\,M\,(1,\,2\,\alpha,\,-\,(\mu-\lambda\,p)\,\beta)+p\,(2\,\alpha-1)\,\lambda+\mu\right)}{-\,(\mu-\lambda\,p)^2}. \tag{A.29}$$

Note that $(p^2\beta\,\lambda^2-p\,(\beta\,\mu+2\,\alpha-1)\,\lambda-\mu)$ is the only term in (A.29) that can be negative. Substituting the upper bound given in Lemma 9 in (A.29), we get

$$\frac{\partial V_1(\beta,p)}{\partial p}>\frac{\beta^{2\,\alpha}c\lambda}{(\mu-\lambda\,p)\,\beta+2\,\alpha}>0.$$

Thus, $V_1(\beta,p)$ is monotone increasing in p.

(ii) We next prove that $V_3(\beta,p)$ is increasing in p when $T_1=\beta-\frac{2(\alpha-1)}{\lambda p}>0$.

Let $\hat{V}(\beta,p,T)$ denote the value of service (V) at which $F_2(T)=F_3(T)$. Then,

$$\begin{aligned}\hat{V}(\beta,p,T) &= \frac{c\,(2\,\alpha-1)\left(\lambda\,p\mathrm{e}^{-(\mu-\lambda\,p)T}-\mu\right)(T-\beta)^{2\,\alpha-2}}{\mu\,-(\mu-\lambda\,p)} \\ &+ \frac{cp\lambda\left(\int_0^T(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{(\lambda\,p-\mu)x}\,\mathrm{d}x-(T-\beta)^{2\,\alpha-1}-\beta^{2\,\alpha-1}\right)}{\lambda\,p-\mu}.\end{aligned} \tag{A.30}$$

According to (A.11), $V_3(\beta,p)$ denotes the value of the service (V) at which $F_2(T_1)$ $=F_3(T_1)$ where $T_1=\beta-\frac{2(\alpha-1)}{\lambda p}$. Therefore,

$$\frac{\partial V_3(\beta,p)}{\partial p}=\frac{\partial\hat{V}(\beta,p,T)}{\partial p}+\frac{\partial\hat{V}(\beta,p,T)}{\partial T}\frac{\partial T}{\partial p}|_{T=\beta-\frac{2(\alpha-1)}{\lambda p}}. \tag{A.31}$$

The second term in (A.31) is zero based on (A.12),

$$\frac{dF_2(T)}{dT}|_{T=\beta-\frac{2(\alpha-1)}{\lambda p}}=\frac{dF_3(T)}{dT}|_{T=\beta-\frac{2(\alpha-1)}{\lambda p}}=0.$$

Therefore, $\frac{\partial V_3(\beta,p)}{\partial p}=\frac{\partial\hat{V}(\beta,p,T)}{\partial p}$;

$$\frac{\partial\hat{V}(\beta,p,T)}{\partial p}=\frac{\lambda\,c\,(-Z_1(p,T)+Z_2(p,T))}{\mu\,-(\mu-\lambda\,p)^2}, \tag{A.32}$$

where

$$\begin{aligned}Z_1(p,T) &= \mu^2\int_0^T(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{-(\mu-\lambda\,p)x}\,\mathrm{d}x-\mu\,p \\ &-(\mu-\lambda\,p)\int_0^T(2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\lambda\,x\mathrm{e}^{-(\mu-\lambda\,p)x}\,\mathrm{d}x, \\ Z_2(p,T) &= \mu\left((2\,\alpha-1)\,(T-\beta)^{2\,\alpha-2}+\mu\left((T-\beta)^{2\,\alpha-1}+\beta^{2\,\alpha-1}\right)\right) \\ &-(T-\beta)^{2\,\alpha-2}\,(\mu+T\lambda\,p\,(-\lambda\,p+\mu))\,(2\,\alpha-1)\,\mathrm{e}^{T-(\mu-\lambda\,p)}.\end{aligned} \tag{A.33}$$

Thus, $\frac{\partial\hat{V}(p,T)}{\partial p}>0$ if $Z_1(p,T)<Z_2(p,T)$. Note that $Z_1(p,0)=Z_2(p,0)=0$. Then, to prove $Z_1(p,T)<Z_2(p,T)$, we will show that $\frac{\partial Z_1(p,T)}{\partial T}<\frac{\partial Z_2(p,T)}{\partial T}$ in the interval

$[0, \beta - \frac{2(\alpha-1)}{\lambda p}]$. Considering (A.33), we get

$$\begin{aligned}
&\frac{\partial Z_1(p,T)}{\partial T} = -\left(2\,\alpha-1\right)\left(-\beta+T\right)^{2\,\alpha-2}\,\mathrm{e}^{-(\mu-\lambda\,p)T}\left(p\left(\lambda\,p-\mu\right)\lambda\,T-\mu\right)\mu,\\
&\frac{\partial Z_2(p,T)}{\partial T} = \left(T-\beta\right)^{2\,\alpha-3}\mu\,\left(\left(2\,\alpha-1\right)\left(2\,\alpha-2\right)+\left(2\,\alpha-1\right)\mu\,\left(T-\beta\right)\right)\\
&-\left(T-\beta\right)^{2\,\alpha-3}\,\mathrm{e}^{T(\lambda\,p-\mu)}\\
&\left(\left(2\,\alpha-2\right)\left(\mu+T\lambda\,p\left(-\lambda\,p+\mu\right)\right)\left(2\,\alpha-1\right)+\lambda\,p\left(T-\beta\right)\left(-\lambda\,p+\mu\right)\left(2\,\alpha-1\right)\right.\\
&\left.+\left(T-\beta\right)\left(\mu+T\lambda\,p\left(-\lambda\,p+\mu\right)\right)\left(2\,\alpha-1\right)-\left(\mu-\lambda\,p\right)\right).
\end{aligned} \tag{A.34}$$

Therefore,

$$\begin{aligned}
&\frac{\partial Z_2(p,T)}{\partial T}-\frac{\partial Z_1(p,T)}{\partial T} = \left(T-\beta\right)^{2\,\alpha-3}\left(2\,\alpha-1\right)\\
&\left(\left(\left(-Tp^2\left(T-\beta\right)\lambda^2-2\,p\left(\alpha\,T-\beta\right)\lambda-2\,\alpha+2\right)\mu\right.\right.\\
&\left.+\left(pT\left(T-\beta\right)\lambda+2\,\alpha\,T-T-\beta\right)\lambda^2p^2\right)\mathrm{e}^{T(\lambda\,p-\mu)}\\
&\left.+\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)\right).
\end{aligned}$$

Considering $T < \beta$, we conclude that $\frac{\partial Z_2(p,T)}{\partial T} - \frac{\partial Z_1(p,T)}{\partial T} > 0$ if and only if

$$\begin{aligned}
&\left(-Tp^2\left(T-\beta\right)\lambda^2-2\,p\left(\alpha\,T-\beta\right)\lambda-2\,\alpha+2\right)\mu\\
&+\left(\left(pT\left(T-\beta\right)\lambda+2\,\alpha\,T-T-\beta\right)\lambda^2p^2\right)\\
&+\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)\mathrm{e}^{-T(\lambda\,p-\mu)}<0,
\end{aligned}$$

or equivalently if and only if

$$\begin{aligned}
&\mathrm{e}^{T(\mu-\lambda\,p)} >\\
&-\frac{\left(-Tp^2\left(T-\beta\right)\lambda^2-2\,p\left(\alpha\,T-\beta\right)\lambda-2\,\alpha+2\right)\mu}{\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)}\\
&-\frac{\left(pT\left(T-\beta\right)\lambda+2\,\alpha\,T-T-\beta\right)\lambda^2p^2}{\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)}.
\end{aligned} \tag{A.35}$$

To prove the above inequality, we first show that the right hand side of (A.35) is increasing in β, and then demonstrate that the inequality holds even when $\beta \to \infty$. Let

$$Z_3(p,T) =$$

$$-\frac{\left(-Tp^2\left(T-\beta\right)\lambda^2-2\,p\left(\alpha\,T-\beta\right)\lambda-2\,\alpha+2\right)\mu}{\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)}$$
$$-\frac{\left(pT\left(T-\beta\right)\lambda+2\,\alpha\,T-T-\beta\right)\lambda^2p^2}{\mu\,\left(2\,\alpha-2+\mu\,\left(T-\beta\right)\right)}. \tag{A.36}$$

Then,

$$\frac{\partial Z_3(p,T)}{\partial\beta}=2\,\frac{\left(p\lambda-\mu\right)^2\left(\alpha-1\right)\left(pT\lambda+1\right)}{\mu\,\left(2\,\alpha-2+\mu\,\left(-\beta+T\right)\right)^2}>0.$$

Therefore, the right hand side of (A.35) is less than or equal to

$$\lim_{\beta\to\infty} Z_3(p,T)=-\frac{\lambda\,\left(Tp^2\lambda^2+\left(-T\mu+1\right)p\lambda-2\,\mu\right)p}{\mu^2}.$$

Considering that $\mathrm{e}^{-T-(\mu-\lambda\,p)}>1+\left(-\lambda\,p+\mu\right)T$, we get

$$\mathrm{e}^{-T-(\mu-\lambda\,p)}>1+\left(-\lambda\,p+\mu\right)T>-\frac{\lambda\,\left(Tp^2\lambda^2+\left(-T\mu+1\right)p\lambda-2\,\mu\right)p}{\mu^2}>Z_3(p,T),$$

which completes the proof that $V_3(\beta,p)$ is increasing in p.

(iii) Next, we prove that $V_1(\beta,p)$ and $V_3(\beta,p)$ are increasing in β. Similar to (A.31)

$$\frac{\partial V_1(\beta,p)}{\partial\beta}=\frac{\partial\hat{V}(\beta,p,T)}{\partial\beta}|_{T=\beta}+\frac{\partial\hat{V}(\beta,p,T)}{\partial T}\frac{\partial T}{\partial\beta}|_{T=\beta}, \tag{A.37}$$

$$\frac{\partial V_3(\beta,p)}{\partial\beta}=\frac{\partial\hat{V}(\beta,p,T)}{\partial\beta}|_{T=\beta-\frac{2(\alpha-1)}{\lambda p}}+\frac{\partial\hat{V}(\beta,p,T)}{\partial T}\frac{\partial T}{\partial\beta}|_{T=\beta-\frac{2(\alpha-1)}{\lambda p}}. \tag{A.38}$$

Similar to (A.31) and based on (A.12), the second terms in (A.37) and (A.38) are zero. Therefore, to show that $V_1(\beta,p)$ and $V_3(\beta,p)$ are increasing in β, we demonstrate that $\frac{\partial\hat{V}(\beta,p,T)}{\partial\beta}$ is increasing in β for $T\le\beta$. Considering (A.30), we get

$$\frac{\partial\hat{V}(\beta,p,T)}{\partial\beta}=(A_1)(A_2+A_3),$$

where

$$\begin{aligned} A_1 &= \frac{c}{(T-\beta)\,\beta\,\mu\,(\lambda\,p-\mu)}, \\ A_2 &= \beta\left(\mu-\lambda\,p e^{T(\lambda\,p-\mu)}\right)(2\,\alpha-2)\,(2\,\alpha-1)\,(T-\beta)^{2\,\alpha-2}, \\ A_3 &= (2\alpha-1)\,p\mu\lambda(T-\beta) \\ &\quad \left(\beta\,(T-\beta)^{2\alpha-2}-\beta^{2\alpha-1}-\beta\int_0^T(2\alpha-2)\,e^{(\lambda p-\mu)x}\,(x-\beta)^{2\alpha-3}\,dx)\right). \end{aligned}$$

Since $\mu > \lambda p$, $A_1 > 0$ and $A_2 > 0$ for $T \leq \beta$. Also, $A_3 \geq 0$ due to $T \leq \beta$. Thus, $\frac{\partial \hat{V}(\beta,p,T)}{\partial \beta}$ is increasing in β for $T \leq \beta$.

(iv) From (A.15), it is straightforward to see that $V_2(\beta)$ is increasing in β.

A.1.6 Proof of Corollary 3

Since $V_1(\beta,p)$ and $V_3(\beta,p)$ are increasing in p, the results are the conclusions of Proposition 3.

- Here, we show that $V_2(\beta)$ and $V_1(\beta,p)$ intersect at p_e. According to part (iii) in the proof of Proposition 3, $V_3(\beta,p)$ is defined for $\beta \geq \frac{2\,(\alpha-1)}{\lambda\,p}$ such that for $V \geq V_3(\beta,p)$, we have $F_2(\beta-\frac{2\,(\alpha-1)}{\lambda\,p}) \geq F_3(\beta-\frac{2\,(\alpha-1)}{\lambda\,p})$. Since $|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|$ (based on (A.13)), we have $F_2(0) > F_3(0)$ and $F_2(\beta) > F_3(\beta)$. Therefore, according to the definition of $V_1(\beta)$ and $V_2(\beta)$ given in parts (i) and (ii) of the proof of Proposition 3 we have $V_3(\beta,p) > V_2(\beta)$ and $V_3(\beta,p) > V_1(\beta,p)$. Moreover, $V_1(\beta,0) = 0$ and $V_1(\beta,p)$ is increasing in p. Since $V_2(\beta)$ is independent of p, $V_2(\beta)$ and $V_1(\beta,p)$ intersect at most at one point called p_e.
- According to the proof of Proposition 3, the results clearly follow.

A.1.7 Proof of Corollary 4

According to the proof of Proposition 3, $F_2(T)$ and $F_3(T)$ cross each other at most three times. Since T_1 and T_2 are the maximum and minimum points, respectively, of both $F_2(T)$

and $F_3(T)$, and $|\frac{dF_2(T)}{dT}| < |\frac{dF_3(T)}{dT}|$ due to (A.13), then the first point at which $F_2(T)$ and $F_3(T)$ cross each other is less than or equal to T_1. Similarly, the third point at which $F_2(T)$ and $F_3(T)$ cross each other is greater than or equal to T_2. Note that the second point at which $F_2(T)$ and $F_3(T)$ intersect is a minimum point of $U_T(p,T)$.

A.1.8 Proof of Theorem 1

We provide the proof in two parts; in the first part, we show that for any given abandonment threshold T, there exists a unique optimal joining probability which maximizes the social welfare function. Then, we prove that an optimal pair (p^*, T^*) always exists.

1. We show that the social welfare function $W(p,T)$ is either unimodal or increasing in p for $p \in [0,1]$. Note that

$$\frac{\partial W(p,T)}{\partial p} = \lambda \left(U_T(p,T) + p \frac{\partial U_T(p,T)}{\partial p} \right).$$

Considering (A.2), we get

$$\frac{\partial W(p,T)}{\partial p} = V\lambda \left(G_T^p(T) + p \frac{\partial G_T^p(T)}{\partial p} \right) - c\lambda \int_0^T \left((2\alpha-1)(x-\beta)^{2\alpha-2} \right) \left(1 - G_T^p(x) - p \frac{\partial G_T^p(x)}{\partial p} \right) dx. \tag{A.39}$$

Using (2.2), we have

$$\frac{\partial W(p,T)}{\partial p} = \left(\frac{\lambda\mu}{\left(\lambda p \mathrm{e}^{(\lambda p-\mu)T} - \mu\right)^2} \right) V \left(\left(T\lambda^2 p^2 - T\lambda\mu p - \mu \right) \mathrm{e}^{-(\mu-\lambda p)T} + \mu \right) - \left(\frac{c\lambda}{\left(\lambda p \mathrm{e}^{(\lambda p-\mu)T} - \mu\right)^2} \right) \left(\int_0^T \left((2\alpha-1)(x-\beta)^{2\alpha-2} \right) (F_5(x,p)) dx \right) + \left(\frac{c\lambda}{\left(\lambda p \mathrm{e}^{(\lambda p-\mu)T} - \mu\right)^2} \right) \left(\lambda p\mu \left(\lambda pT + 2 \right) \mathrm{e}^{(-\mu+\lambda p)T} \left((T-\beta)^{2\alpha-1} + \beta^{2\alpha-1} \right) \right), \tag{A.40}$$

where,

$$F_5(x,p) = \left(\mathrm{e}^{-(\mu-\lambda p)T}\right)^2 \lambda^2 p^2 + p\left(\lambda p\,(T-x)\,\mathrm{e}^{-(\mu-\lambda p)x}\right)\lambda\,\mu\,\mathrm{e}^{-(\mu-\lambda p)T} + \mathrm{e}^{(\lambda p-\mu)x}\mu^2\,(\lambda p x+1)\,. \tag{A.41}$$

Therefore, to prove that $\frac{\partial W(p,T)}{\partial p}$ has at most one extreme point for $p \in [0,1]$, it is sufficient to show that equation $F_6(p) - F_7(p) = 0$ has at most one root for $p \in [0,1]$ where,

$$\begin{aligned} F_6(p) &= c\lambda\, p\mu\,(\lambda pT+2)\,\mathrm{e}^{(-\mu+\lambda p)T}\left((T-\beta)^{2\alpha-1}+\beta^{2\alpha-1}\right) + V\mu T\lambda^2 p^2 \mathrm{e}^{-(\mu-\lambda p)T}, \\ F_7(p) &= c\int_0^T \left(\,(2\alpha-1)(x-\beta)^{2\alpha-2}\right)(F_5(x,p))dx + \\ & V\mu\left((T\lambda\,\mu\,p+\mu)\,\mathrm{e}^{-(\mu-\lambda p)T}-\mu\right). \end{aligned} \tag{A.42}$$

(a) Since $T \geq 0$, we have $(T-\beta)^{2\alpha-1} + \beta^{2\alpha-1} > 0$; consequently, $F_6(p)$ is convex and monotone increasing in p.

(b) Since $x \leq T$, $F_5(x,p)$ given in (A.41) is convex and monotone increasing in p. Therefore, the first term in $F_7(p)$, $c\int_0^T \left((2\alpha-1)(x-\beta)^{2\alpha-2}\right)(F_5(x,p))dx$, is convex and monotone increasing in p. Moreover, the second term in $F_7(p)$ is also is convex and monotone increasing in p since

$$\frac{\partial\left(V\mu\left((T\lambda\,\mu\,p+\mu)\,\mathrm{e}^{-(\mu-\lambda p)T}-\mu\right)\right)}{\partial p} = V\mu^2\lambda^2 T\mathrm{e}^{(-\mu+\lambda p)T}\,(2+\lambda pT) > 0.$$

Therefore, both $F_6(p)$ and $F_7(p)$ are convex and monotone increasing in p. This means that equation $F_6(p) - F_7(p) = 0$ has at most two roots since any two monotone increasing convex functions (of class C^1) intersect at most at two points. Note that at $p = \frac{\mu}{\lambda}$ (> 1) we have

$$F_6(p) = F_7(p) = c\lambda\,(2+\mu\,T)\,\mu^2\left((T-\beta)^{2\alpha-1}+\beta^{2\alpha-1}\right) + V\mu^3\lambda\,T.$$

Therefore, $F_6(p) - F_7(p) = 0$ has at most one root for $p \in [0,1]$. When p goes to zero, $F_6(p)$ goes to zero while $F_7(p)$ approaches $-\mu U_T\,(p,T)$ at $p = 0$,

$$F_7(0) = \left(c\int_0^T (2\,\alpha-1)\,(x-\beta)^{2\,\alpha-2}\,\mathrm{e}^{-\mu\,x}\mu dx - V\left(1-\mathrm{e}^{-T\mu}\right)\right)\mu,$$

which is non-positive (considering the condition that at least one customer enters the system when the system is empty). Therefore, if $F_6(p)$ and $F_7(p)$ intersect at point $p^* < 1$, the welfare is maximized at p^*. Otherwise, the welfare is non-decreasing in $p \in [0,1]$ and $p = 1$ is the optimal joining probability from the social maximizer perspective.

2. Note that the social welfare function is bounded and continuous in its domain. We proved that for any given abandonment threshold, there exists a unique optimal joining probability p^* which maximizes the social welfare function. Also, based on Proposition 3, for any given joining probability, there exists an optimal abandonment threshold T^* which maximizes the social welfare function. Thus, there exists an optimal pair (p^*, T^*) that maximizes the social welfare function.

   Moreover, according to Proposition 3, if $V < \min(V_1(\beta,1), V_2(\beta))$, the social welfare function is negative for all $T \geq 0$. Since according to Proposition 3, the function $V_1(\beta,p)$ is monotone increasing in p, we can conclude that when $V < \min(V_1(\beta,1), V_2(\beta))$, a joining probability less than one should be chosen to ensure $V \geq \min(V_1(\beta,p), V_2(\beta))$.

A.1.9 Proof of Proposition 4

Let $\hat{U}_T(p,T)$ denote the customer's expected utility function under the entrance/service fee mechanism when she will renege the system after T time units. Then, using (A.2) we have,

$$\lambda p \hat{U}_T(p,T) = \lambda p \left((V-\theta_s) G_T^p(T) - c \int_0^T \big((2\alpha-1)(x-\beta)^{2\alpha-2} \big)(1-G_T^p(x))dx - \theta_e \right). \tag{A.43}$$

Recall that $W(p,T) = \lambda p U_T(p,T)$, and the total expected revenue function is $\Phi(p,T,\theta) = \lambda p\, \theta_s G_T^p(T) + \lambda p\, \theta_e$. Considering (A.43) and the definition of $W(p,T)$ we have $\lambda p\, \hat{U}(p,T) = W(p,T) - \Phi(p,T,\theta)$. Therefore, $\Phi(p,T,\theta) = W(p,T) - \lambda p \hat{U}_T(p,T)$.

Considering that under the optimal action the firm fully extracts the total surplus by an entrance fee, i.e. $\hat{U}_T(p,T)=0$, we get $\Phi(p,T,\theta)=W(p,T)$. Therefore, the optimal pair (p^*,T^*) which maximizes the social welfare function also maximizes the expected revenue function.

Now, we want to show under which conditions this pricing mechanism induces the socially optimal behavior (p^*,T^*). Using an entrance-fee/service-fee mechanism, the planner charges a service fee to control customers' reneging time; meanwhile it charges an entrance fee to fully extract the total surplus (the sum of all customers' surplus) and control the arrival rate.

Since $W(p,T)=\lambda p U_T(p,T)$, the optimal T^* also maximizes $U_T(p,T)$. According to (A.3),

$$\frac{\partial U_T(p,T)}{\partial T}=V\frac{\partial G_T^p(T)}{\partial T}-\frac{\partial C(T)}{\partial T}(1-G_T^p(T))+\int_0^T\frac{\partial C(x)}{\partial x}(\frac{\partial G_T^p(x)}{\partial T})dx. \tag{A.44}$$

Using (2.3), we have,

$$\frac{dP_0}{dT}=\lambda p e^{-(\mu-\lambda p)T}P_0^2=-RP_0, \tag{A.45}$$

where $R=\lambda p e^{-(\mu-\lambda p)T}P_0$. Therefore, considering (2.2) we get

$$\begin{aligned}\frac{dG_T^p(x)}{dT}&=-RG_T^p(x),\\ \frac{dG_T^p(T)}{dT}&=-RG_T^p(T)+\frac{\partial G_T^p(t)}{\partial t}|_{t=T}.\end{aligned} \tag{A.46}$$

Now, according to (A.44), (A.45) and (A.46) we have:

$$\begin{aligned}\frac{\partial U_T(p,T)}{\partial T}=V(-RG_T^p(T)+\frac{\partial G_T^p(t)}{\partial t}|_{t=T})-\frac{\partial C(T)}{\partial T}(1-G_T^p(T))\\ -\int_0^T\frac{\partial C(x)}{\partial x}(RG_T^p(x))dx,\end{aligned}$$

which results in

$$\frac{\partial U_T(p,T)}{\partial T}=V\frac{\partial G_T^p(t)}{\partial t}|_{t=T}-\frac{\partial C(T)}{\partial T}(1-G_T^p(T))-VRG_T^p(T)$$

$$+\int_0^T \frac{\partial C(x)}{\partial x}(R - RG_T^p(x) - R)dx.$$

Thus,

$$\frac{\partial U_T(p,T)}{\partial T} =$$
$$V\frac{\partial G_T^p(t)}{\partial t}|_{t=T} - \frac{\partial C(T)}{\partial T}(1 - G_T^p(T))$$
$$-R\left(VG_T^p(T) - \int_0^T \frac{\partial C(x)}{\partial x}(1 - G_T^p(x))dx\right) - RC(T),$$

and finally using (A.2),

$$\frac{\partial U_T(p,T)}{\partial T} = V\frac{\partial G_T^p(t)}{\partial t}|_{t=T} - \frac{\partial C(T)}{\partial T}(1 - G_T^p(T)) - R(U_T(p,T) + C(T)). \tag{A.47}$$

Recall that according to (A.3),

$$\frac{\partial U_T(p,\tau)}{\partial \tau} = V\frac{\partial G_T^p(\tau)}{\partial \tau} - \frac{\partial C(\tau)}{\partial \tau}(1 - G_T^p(\tau)). \tag{A.48}$$

Consequently rewriting (A.47), we end up with,

$$\frac{\partial U_T(p,T)}{\partial T} = \frac{\partial U_T(p,t)}{\partial t}|_{t=T} - R(U_T(p,T) + C(T)). \tag{A.49}$$

Therefore, if T^* is the socially optimal reneging time, then according to (A.49) we must have

$$U_{T^*}(p,T^*) + C(T^*) = \frac{\frac{\partial U_{T^*}(p,t)}{\partial t}|_{t=T^*}}{R}. \tag{A.50}$$

Now consider an arriving customer decision. According to (A.43) and using (A.3),

$$\frac{\partial \hat{U}_T(p,\tau)}{\partial \tau} = (V - \theta_s)\frac{\partial G_T^p(\tau)}{\partial \tau} - \frac{\partial C(\tau)}{\partial \tau}(1 - G_T^p(\tau)). \tag{A.51}$$

Therefore, we have,

$$\frac{\partial \hat{U}_T(p,\tau)}{\partial \tau} = V\frac{\partial G_T^p(\tau)}{\partial \tau} - \frac{\partial C(\tau)}{\partial \tau}(1 - G_T^p(\tau)) - \theta_s\frac{\partial G_T^p(\tau)}{\partial \tau} = \frac{\partial U_T(p,\tau)}{\partial \tau} - \theta_s\frac{\partial G_T^p(\tau)}{\partial \tau}. \tag{A.52}$$

Suppose that all other customers choose T^* as a reneging time; therefore, a necessary condition for an arriving individual customer to choose T^* as the only reneging time is that (A.52) has only one root in τ equal to T^* where the sign of $\frac{\partial \hat{U}_T(p,\tau)}{\partial \tau}$ changes from positive to negative. If so, we have

$$\frac{\partial U_T(p,\tau)}{\partial \tau}|_{\tau=T^*} = \theta_s \frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau=T^*}. \tag{A.53}$$

According to Proposition (1), we have $h_{T^*}(T^*) = \mu$. Therefore, using (A.48) and (A.53) we get

$$\theta_s = V - \frac{\frac{\partial C(\tau)}{\partial \tau}|_{\tau=T^*}}{\mu} = V(1 - \frac{\gamma(T^*)}{\mu}). \tag{A.54}$$

Using this service fee and according to (A.54) and (A.51),

$$\frac{\partial \hat{U}_T(p,\tau)}{\partial \tau} = (\frac{\frac{\partial C(\tau)}{\partial \tau}|_{\tau=T^*}}{\mu})\frac{\partial G_T^p(\tau)}{\partial \tau} - \frac{\partial C(\tau)}{\partial \tau}(1 - G_T^p(\tau)). \tag{A.55}$$

Since all customers are homogeneous, reneging at T^* will be an equilibrium point for all, if for any arriving customer, the best response is reneging at time T^*, i.e. $\frac{\partial \hat{U}_T(p,\tau)}{\partial \tau}$ must have only one root equal to T^*; therefore, using (A.55) and according to Proposition 2 for $\tau > \beta$, the equation $(\frac{\frac{\partial C(\tau)}{\partial \tau}|_{\tau=T^*}}{\mu})h_{T^*}(\tau) = V\gamma(\tau)$ must have only one root equal to T^*.

Note that in this case, the necessary condition for not reneging at $\tau = 0$ is that the customer's expected utility must be greater than zero at T^*; otherwise, reneging at $\tau = 0$ is the best response for an arriving individual customer. Thus, we next find a condition to avoid this situation. Considering (A.50) and replacing $U_{T^*}(p,T^*)$ with $\hat{U}_{T^*}(p,T^*) + \theta_e + \theta_s G_{T^*}^p(T^*)$, we have

$$\hat{U}_{T^*}(p,T^*) + \theta_e + \theta_s G_{T^*}^p(T^*) + C(T^*) = \frac{\frac{\partial U_{T^*}(p,t)}{\partial t}|_{t=T^*}}{R},$$

and using (A.53),

$$\hat{U}_{T^*}(p,T^*) = \theta_s \left(\frac{\frac{\partial G_{T^*}^p(\tau)}{\partial \tau}|_{\tau=T^*}}{R} - G_{T^*}^p(T^*) \right) - C(T^*) - \theta_e.$$

Then using (A.46),

$$\hat{U}_{T^*}(p,T^*) = \theta_s \left(\frac{RG^p_{T^*}(T^*) + \frac{\partial G^p_T(T)}{\partial T}|_{\tau=T^*}}{R} - G^p_{T^*}(T^*) \right) - C(T^*) - \theta_e$$

$$= \theta_s \frac{\frac{\partial G^p_T(T)}{\partial T}|_{\tau=T^*}}{R} - C(T^*) - \theta_e,$$

and according to (A.54) we can conclude that

$$\hat{U}_{T^*}(p,T^*) = (V - \frac{C'(T^*)}{\mu}) \frac{\frac{\partial G^p_T(T)}{\partial T}|_{\tau=T^*}}{R} - C(T^*) - \theta_e. \tag{A.56}$$

Recall that $R = \lambda p e^{-(\mu-\lambda p)T} P_0$. Therefore using (2.2) and (2.3) we can rewrite (A.56) as follows

$$\hat{U}_{T^*}(p,T^*) = (V - \frac{C'(T^*)}{\mu}) \frac{P_0}{\rho} - C(T^*) - \theta_e, \tag{A.57}$$

where $\rho = \lambda p/\mu$. Recall the necessary condition for an arriving customer to not renege at time zero is that the expected utility function obtained in (A.57) must be positive for $\theta_e = 0$. Since the cost-reward ratio function is $\gamma(T) = \frac{C'(T)}{V}$, according to (A.57) we obtain the following condition

$$\gamma(T^*) < \mu - \frac{\lambda p}{V P_0} C(T^*).$$

Therefore, if the optimal entrance fee given by (A.54) can induce socially optimal reneging strategy, i.e. the discussed two conditions hold, then the firm collects the total surplus by charging an entrance fee equal to the customer willingness to pay which is obtained using (A.57).

However, under the entrance-fee/service-fee mechanism, if the discussed two conditions do not hold, we can not guarantee that this policy induces the socially optimal behavior and the firm may not gain a profit equal to the maximum social welfare.

A.1.10 Proof of Proposition 5

We want to prove that for a given p if all other customers follow the (p, T^*) strategy then, for an arriving customer (considering the offered waiting time), staying in the system until

T^* is the best response. First, we prove that the customer's expected utility is increasing in time at time T^*. Second, we prove that there is no positive local maximum point less than T^* for the customer's expected utility if all follow the T^* policy.

1. According to (A.2), we can conclude that $U_T(p,T)+C(T) \geq VG_T^p(T)$. Also, T^* is the solution of $\frac{\partial U_T(p,T)}{\partial T}=0$. Therefore, considering (A.49), $\frac{\partial U_T(p,\tau)}{\partial \tau}|_{\tau=T^*}>0$, which means the expected utility function is increasing at $\tau=T^*$.

2. Now we want to prove that there is no local maximum for an arriving customer's expected utility function before T^*. Since the expected utility function is increasing at $\tau=T^*$, so if there is a local maximum point before T^*, denoted by τ_2, there should be also a local minimum point between τ_2 and T^*. First we prove the following lemma,

   Lemma 10 *Suppose that all customers follow T^*. If there is local minimum point τ_3 for an arriving customer's expected utility function before T^*, then $T^*<2\tau_3$.*

   Proof.

   It is clear that $x+1\leq e^x$, which results in $1-e^{-x}\geq 1-\frac{1}{x+1}$. Since $2\alpha-2\geq 0$ and $\frac{\lambda p}{\mu}<1$,

   $$1-\frac{\lambda p}{\mu}e^{-x}>1-e^{-x}>\frac{x}{x+2\alpha-2}.$$

   Substituting x by $(\mu-\lambda p)x$ we have

   $$\frac{\mu-\lambda p e^{-(\mu-\lambda p)x}}{\mu}>\frac{(\mu-\lambda p)x}{(\mu-\lambda p)x+2\alpha-2},$$

   and

   $$\frac{(\mu-\lambda p)\mu}{\mu-\lambda p e^{-(\mu-\lambda p)x}}<\mu-\lambda p+\frac{2\alpha-2}{x}.$$

   Assuming $x>\beta$, we can conclude that $\mu-\lambda p+\frac{2\alpha-2}{x}<\mu-\lambda p+\frac{2\alpha-2}{x-\beta}$; consequently we have

   $$\frac{(\mu-\lambda p)\mu}{\mu-\lambda p e^{-(\mu-\lambda p)x}}<\mu-\lambda p+\frac{2\alpha-2}{x-\beta}. \tag{A.58}$$

Now suppose that for an abandonment threshold T, the customer's expected utility function has a minimum point τ_3 greater than β such that $\frac{\partial U_T(p,\tau)}{\partial \tau}|_{\tau=\tau_3} = 0$. In other words, according to (A.3),

$$V\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3} = \frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}(1 - G_T^p(\tau_3)). \tag{A.59}$$

Also since τ_3 is a local minimum point, we have $\frac{\partial^2 U_T(p,\tau)}{\partial \tau^2}|_{\tau=\tau_3} > 0$; therefore, according to (A.3),

$$V\frac{\partial^2 G_T^p(\tau)}{\partial \tau^2}|_{\tau_3} > \frac{\partial^2 C(\tau)}{\partial \tau^2}|_{\tau_3}(1 - G_T^p(\tau_3)) - \frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3}. \tag{A.60}$$

Also according to (2.2), $\frac{\partial^2 G_T^p(\tau)}{\partial \tau^2}|_{\tau_3} = -(\mu - \lambda p)\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3}$. Using (A.60),

$$V(\mu - \lambda p)\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3} < \frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3} - \frac{\partial^2 C(\tau)}{\partial \tau^2}|_{\tau_3}(1 - G_T^p(\tau_3)),$$

and using (A.59),

$$(\mu - \lambda p)\frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}(1 - G_T^p(\tau_3)) < \frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3} - \frac{\partial^2 C(\tau)}{\partial \tau^2}|_{\tau_3}(1 - G_T^p(\tau_3)).$$

Since $\tau_3 > \beta$, then $\frac{\partial C(\tau)}{\partial \tau}|_{\tau_3} > 0$ and we have

$$(\mu - \lambda p) < \frac{\frac{\partial G_T^p(\tau)}{\partial \tau}|_{\tau_3}}{1 - G_T^p(\tau_3)} - \frac{\frac{\partial^2 C(\tau)}{\partial \tau^2}|_{\tau_3}}{\frac{\partial C(\tau)}{\partial \tau}|_{\tau_3}},$$

and using (2.5), we get

$$(\mu - \lambda p) + \frac{2\alpha - 2}{\tau_3 - \beta} < h(\tau_3). \tag{A.61}$$

Now suppose that $T^* > 2\tau_3$. Using (2.5),

$$h(\tau_3) = \frac{(\mu - \lambda p)\mu}{\mu - \lambda p e^{-(\mu-\lambda p)(T-\tau_3)}} < \frac{(\mu - \lambda p)\mu}{\mu - \lambda p e^{-(\mu-\lambda p)\tau_3}}.$$

Using (A.61) we conclude that

$$\mu - \lambda p + \frac{2\alpha - 2}{\tau_3 - \beta} < \frac{(\mu - \lambda p)\mu}{\mu - \lambda p e^{-(\mu-\lambda p)\tau_3}},$$

which contradicts with (A.58); therefore, we conclude that $T^* < 2\tau_3$.■

Next we prove that under abandonment threshold T^*, there is no other local maximum for the customer's expected utility function. Suppose that the customer's expected utility function has a local maximum point τ_2 and consequently a local minimum point $\tau_2 < \tau_3 < T^*$. Therefore $\frac{\partial U_T(p,\tau)}{\partial \tau}|_{\tau=\tau_3} = 0$. As proven in Lemma 10, $T^* \leq 2\tau_3$. Therefore, using (A.3),

$$V\left(-\lambda\, p + \mu\right) = c\left(\tau_3 - \beta\right)^{2\,\alpha-2}\left(2\,\alpha - 1\right)\left(1 - \frac{\lambda\, p \mathrm{e}^{-(\mu-\lambda\, p)(T^*-\tau_3)}}{\mu}\right)$$
$$< c(2\alpha-1)(\tau_3-\beta)^{2\alpha-2}\left(1 - \frac{\lambda p}{\mu}e^{-(\mu-\lambda p)\tau_3}\right).$$

Therefore,

$$\lambda pc\frac{(2\alpha-1)}{\mu}(\tau_3-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\, p)\tau_3} < c(2\alpha-1)(T-\beta)^{2\alpha-2} - V\left(\mu-\lambda\, p\right).$$

Adding the same term to the both sides,

$$\lambda pc\int_0^{\tau_3}(2\alpha-1)(x-\beta)^{(2\alpha-2)}dx + \lambda pc\frac{(2\alpha-1)}{\mu}(\tau_3-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\, p)\tau_3} <$$
$$\lambda cp\left((\tau_3-\beta)^{2\alpha-1} + \beta^{2\alpha-1}\right) + c(2\alpha-1)(\tau_3-\beta)^{2\alpha-2} - V\left(\mu-\lambda\, p\right),$$

and clearly,

$$\lambda pc\int_0^{\tau_3}(2\alpha-1)(x-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\, p)x}dx + \lambda pc\frac{(2\alpha-1)}{\mu}(\tau_3-\beta)^{(2\alpha-2)}e^{-(\mu-\lambda\, p)\tau_3} <$$
$$\lambda cp\left((\tau_3-\beta)^{2\alpha-1} + \beta^{2\alpha-1}\right) + c(2\alpha-1)(\tau_3-\beta)^{2\alpha-2} - V\left(\mu-\lambda\, p\right).$$

Therefore, according to (A.11) and the proof of Proposition 3, we can conclude that $F_3(\tau_3) > F_2(\tau_3)$. Since at T^* the sign of $F_2(T^*) - F_3(T^*)$ is changed from positive to negative, we can conclude that according to the proof of Proposition 3, T^* must be less than τ_3 (which is greater than β) which contradicts with our assumption. Consequently, we can conclude that there is no extremum point for the customer's expected utility function for all $\beta < t < T^*$.

A.1.11 Proof of Proposition 6

We first show that $\frac{\partial G_T^p(x)}{\partial p} \leq 0$. According to (2.2), for $x \leq T$ we have:

$$\frac{\partial G_T^p(x)}{\partial p} = -\frac{e^{-(\mu-\lambda p)x}\mu\left(\left((1+p\,(T-x)\,\lambda)-(1+\lambda\,pT)e^{(\mu-\lambda p)x}\right)e^{-(\mu-\lambda p)T}+x\mu\right)\lambda}{\left(\mu-\lambda\,pe^{-(\mu-\lambda p)T}\right)^2}. \tag{A.62}$$

Therefore, $\frac{\partial G_T^p(x)}{\partial p} \leq 0$ is equivalent to

$$H_1(p) = \left((1+p\,(T-x)\,\lambda)-(1+\lambda\,pT)e^{(\mu-\lambda p)x}\right)e^{-(\mu-\lambda p)T}+x\mu \geq 0. \tag{A.63}$$

We next prove that $H_1(p) \geq 0$. According to (A.63):

$$\frac{\partial H_1(p)}{\partial p} = \lambda\left(p\lambda\,T^2+(2-p\lambda\,x)\,T-x\right)\left(1-e^{(\mu-\lambda p)x}\right)e^{-(\mu-\lambda p)T}. \tag{A.64}$$

Note that $(p\lambda\,T^2+(2-p\lambda\,x)\,T-x) > 0$ since $x \leq T$ which results in $\frac{\partial H_1(p)}{\partial p} \geq 0$ for $p \in [0,1]$, $\mu > \lambda p$. Therefore, $H_1(p)$ gets its minimum value at $p=0$ which is $\left(1-e^{-\mu T}\right)+x\mu \geq 0$. Therefore, $G_T^p(x)$ is monotone decreasing in p. Thus, according to (A.2), the expected utility function $U_T\,(p,T)$ is monotone decreasing in p. Since customers are self interested, they join the queue until the expected utility function is exactly equal to zero, i.e. $U_T\,(\overline{p},T)=0$.

A.1.12 Proof of Corollary 5

The equilibrium joining probability is reached when $U_T\,(p,T) = 0$. Therefore, $\frac{\partial \bar{p}}{\partial \beta} = -\frac{\frac{\partial U_T(p,T)}{\partial \beta}}{\frac{\partial U_T(p,T)}{\partial p}}$. Since the expected utility function is monotone decreasing in p, the behavior of $\frac{\partial \bar{p}}{\partial \beta}$ is the same as $\frac{\partial U_T(p,T)}{\partial \beta}$. According to (A.2):

$$U_T\,(p,T) = VG_T^p\,(T) - c\int_0^T\left((2\alpha-1)(x-\beta)^{2\alpha-2}\right)(1-G_T^p\,(x))dx. \tag{A.65}$$

Therefore,

$$\frac{\partial^2 U_T\,(p,T)}{\partial \beta^2} = -c(2\alpha-1)(2\alpha-2)(2\alpha-3)\int_0^T (x-\beta)^{2\alpha-4}(1-G_T^p\,(x))dx. \tag{A.66}$$

Since α is a positive integer, $\frac{\partial^2 U_T(p,T)}{\partial \beta^2}$ given in (A.66) is non-positive and the expected utility function is unimodal in β.

A.1.13 Proof of Proposition 7

1. As proven in Proposition 6, for a given T, the expected utility function is decreasing in p; based on Proposition 3, for a given p there is an optimal abandonment threshold T^* which maximizes the social welfare function and also according to Proposition 5, customers choose this abandonment threshold as the equilibrium one. Note that if the expected utility function is positive, there is an incentive for customers to join the queue. Therefore, the joining probability increases until the expected utility function reaches zero or the joining probability is equal to 1; thus, one of the following two cases must occur:

$$\frac{\partial U_T(\bar{p},T)}{\partial T}|_{T=T^*}=0, \quad U_T(\bar{p},T^*)=0, \tag{A.67}$$

or

$$\frac{\partial U(1,T^*)}{\partial T}|_{T=T^*}=0, \quad U(\bar{p},T^*)\geq 0. \tag{A.68}$$

2. According to Proposition 3, when $V_3(\beta,1)\leq V$ or $V_2(\beta)<V<V_1(\beta,1)$, for $p=1$, the expected utility and the social welfare functions at T^* are non-negative, and condition (A.68) holds.

3. Also, based on Proposition 3, when $V<\min(V_1(\beta,1),V_2(\beta))$ with $p=1$, the social welfare function is negative for all values of T; otherwise, there exists an optimal abandonment threshold if $V\geq \min(V_1(\beta,p),V_2(\beta))$ for a given p. Note that for $p=0$, $\min(V_1(\beta,0),V_2(\beta))=0$ and also $V_1(\beta,p)$ is increasing in p (Proposition 3); therefore, we conclude that for any given V there exists a $p<1$ such that $V\geq \min(V_1(\beta,p),V_2(\beta))$; it means that there exists $(\bar{p},T^*)$ such that condition (A.67) holds .

A.1.14 Proof of Theorem 2

The optimal expected revenue of the service provider is bounded by the maximum social. If the two following conditions hold, the planner can extract total surplus equal to the maximum social welfare: 1. The socially optimal behavior is maintained and 2. the server fully collects the total surplus.

According to Proposition 5, the optimal abandonment threshold can induce the socially optimal reneging time of customers. Since according to Proposition 6 the expected utility function is monotone increasing in p for any given abandonment threshold, the firm can charge an entrance fee equal to the expected utility of a joining customer to maintain the joining probability at a certain level. Therefore, using the entrance-fee/abandonment-threshold mechanism, the planner can (1) induce the socially optimal reneging time by choosing T^* as the abandonment threshold, and the optimal joining probability by imposing an entrance fee θ_e, respectively; also (2) the planner fully collects all customers' surplus by charging an entrance fee equal to the expected utility of a joining customer.

A.1.15 Proof of Corollary 6

Since $V_1(\beta, p)$ is decreasing in p, using Theorem 1 and Theorem 2 the results immediately follow.

A.2 Analysis of the Multi-Echelon Production Inventory System with Strategic Customers

A.2.1 Proof of Lemma 1.

Using (3.1):

$$Za\left(S,\lambda, f(\cdot)\right) - Za\left(S-1,\lambda, f(\cdot)\right) = \int_0^\infty f\left(x\right) Q\left(S,\lambda x\right) dx \; - \int_0^\infty f\left(x\right) Q\left(S-1,\lambda x\right) dx =$$

$$\int_0^\infty f(x)\left(Q(S,\lambda x)-Q(S-1,\lambda x)\right)dx=\int_0^\infty f(x)\left(\frac{(\lambda x)^{S-1}}{(S-1)!}e^{-\lambda x}\right)dx=$$

$$\frac{(\lambda)^{S-1}}{(S-1)!}\int_0^\infty e^{-\lambda x}\ f(x)\left(x^{S-1}\right)dx\ . \tag{A.69}$$

Recall $f^*(\cdot)$ is the LT of $f(\cdot)$. Then,

$$E\left((-x)^n e^{-kx}\right)=\frac{d^n}{dk^n}\ E\left(e^{-kx}\right)=\frac{d^n}{dk^n}\ f^*(k)=f^{*(n)}(k). \tag{A.70}$$

Therefore, (A.69) can be written as follows:

$$Za(S,\lambda,f(\cdot))-Za(S-1,\lambda,f(\cdot))=\frac{(\lambda)^{S-1}}{(S-1)!}\ (-1)^{S-1}\ \frac{d^{S-1}}{dk^{S-1}}\ f^*(k)|_{k=\lambda}.$$

Solving $Za(S,\lambda,f(\cdot))$ recursively we get:

$$\left\{\begin{array}{c} Za(S,\lambda,f(\cdot))-Za(S-1,\lambda,f(\cdot))=\frac{(\lambda)^{S-1}}{(S-1)!}\ (-1)^{S-1}\ \frac{d^{S-1}}{dk^{S-1}}\ f^*(k)|_{k=\lambda} \\ Za(S-1,\lambda,f(\cdot))-Za(S-2,\lambda,f(\cdot))=\frac{(\lambda)^{S-2}}{(S-2)!}\ (-1)^{S-2}\ \frac{d^{S-2}}{dk^{S-2}}\ f^*(k)|_{k=\lambda} \\ \vdots \\ Za(2,\lambda,f(\cdot))-Za(1,\lambda,f(\cdot))=\frac{(\lambda)^{1}}{(1)!}\ (-1)^{1}\ \frac{d^{1}}{dk^{1}}\ f^*(k)|_{k=\lambda} \end{array}\right. \Rightarrow$$

$$Za(S,\lambda,f(\cdot))=\sum_{i=1}^{S}(-1)^{i-1}\frac{\lambda^{i-1}}{(i-1)!}\ f^{*(i-1)}(k)|_{k=\lambda}, \tag{A.71}$$

and

$$Za(1,\lambda,f(\cdot))=f^*(\lambda).$$

A.2.2 Proof of Lemma 2.

Applying integration by parts and (3.1) we get:

$$Za(S,\lambda,f(\cdot))=\int_0^\infty f(x)\,Q(S,\lambda x)\,dx=$$

$$F\left(x\right)Q\left(S,\lambda x\right)|_0^\infty+\int_0^\infty F\left(x\right)\lambda\left(\frac{\left(\lambda x\right)^{S-1}}{\left(S-1\right)!}e^{-\lambda x}\right)dx$$

$$=\frac{\left(\lambda\right)^S}{\left(S-1\right)!}\int_0^\infty e^{-\lambda x}\ F\left(x\right)\left(x^{S-1}\right)dx\ .$$

Using (A.69):

$$Za\left(S,\lambda,f(\cdot)\right)=(-1)^{S-1}\frac{\lambda^S}{\left(S-1\right)!}\ \ \frac{d^{S-1}}{dk^{S-1}}\left(F^*(k)\right)|_{k=\lambda}. \tag{A.72}$$

Now, replacing $f(\cdot)$ with $F(\cdot)$ in Lemma 1, we obtain:

$$Za\left(S,\lambda,F(\cdot)\right)=\sum_{i=1}^{S}(-1)^{i-1}\frac{\lambda^{i-1}}{(i-1)!}\ F^{*(i-1)}(k)|_{k=\lambda}$$

$$=\frac{1}{\lambda}\sum_{i=1}^{S}(-1)^{i-1}\frac{\lambda^{i}}{(i-1)!}\ F^{*(i-1)}(k)|_{k=\lambda}.$$

According to (A.72), the following result can be concluded:

$$Za\left(S,\lambda,F(\cdot)\right)=\frac{1}{\lambda}\sum_{i=1}^{S}Za\left(i,\lambda,f(\cdot)\right).$$

A.2.3 Proof of Theorem 3.

According to (3.2), we have:

$$W^P\left(S,\lambda,f(\cdot)\right)=\int_0^\infty G^P\left(t\right)dt=\int_0^\infty\int_0^\infty f(Q\left(S,\lambda\left(x+t\right)\right)dxdt.$$

Then, according to Fubini's theorem we can interchange the order of the integral and rewrite it as:

$$W^P\left(S,\lambda,f(\cdot)\right)=E_X\left(\int_0^\infty Q\left(S,\lambda\left(x+t\right)\right)\ dt\right),$$

where, $E_X(f(\cdot))$ is the expected value of $f(x)$ with respect to the random variable X. Substituting z for $\lambda(x+t)$, we get:

$$W^P\left(S,\lambda,f(\cdot)\right)=E_X\left(\frac{1}{\lambda}\int_{\lambda x}^\infty Q\left(S,z\right)\ dz\right). \tag{A.73}$$

Taking integration by parts and using (3.1), the following yields:

$$W^P(S,\lambda,f(\cdot)) = \int_0^\infty G^P(t)\ dt = E_x\left[\frac{1}{\lambda}\left(z\ Q(S,z)\,|_{\lambda x}^{\infty} + \int_{\lambda x}^{\infty} z\left(\frac{(z)^{S-1}}{(S-1)!}e^{-z}\right)\ dz\right)\right]$$

$$= E_x\left[\frac{1}{\lambda}\left(-\lambda x\ Q(S,\lambda x) + \int_{\lambda x}^{\infty} S\left(\frac{(z)^{S}}{(S)!}e^{-z}\right)\ dz\right)\right] =$$

$$E_x\left[\frac{1}{\lambda}\left(-\lambda x\ Q(S,\lambda x) + S\ Q(S+1,\lambda x)\right)\right] = \int_0^\infty f(x)\left(\frac{S}{\lambda}Q(S+1,\lambda x) - xQ(S,\lambda x)\right)dx.$$

Applying integration by parts again and using (3.1) leads to

$$W^P(S,\lambda,f(\cdot)) = \int_0^\infty F(x)\left(\left(\frac{S}{\lambda}\right)\frac{(\lambda)(\lambda x)^{S}}{(S)!}e^{-\lambda x} - x\lambda\frac{(\lambda x)^{S-1}}{(S-1)!}e^{-\lambda x} + Q(S,\lambda x)\right)dx.$$

Therefore,

$$W^P(S,\lambda,f(\cdot)) = \int_0^\infty F(x)Q(S,\lambda x)\,dx = Za(S,\lambda,F(\cdot)). \tag{A.74}$$

Now, according to (3.3):

$$W^Q(S,\lambda,f(\cdot)) = \int_0^\infty G^Q(t)\ dt = \int_0^\infty\int_t^\infty f(x)(1-Q(S,\lambda(x-t))\,dx\ dt$$

$$= \int_0^\infty\int_0^x f(x)(1-Q(S,\lambda(x-t))\,dt\ dx.$$

Substituting z for $\lambda(x-t)$,

$$W^Q(S,\lambda,f(\cdot)) = E_X\left(-\frac{1}{\lambda}\int_{\lambda x}^{0}(1-Q(S,z))\ dz\right) = E_X\left(\frac{1}{\lambda}\int_0^{\lambda x}(1-Q(S,z))\ dz\right).$$

Since $\int_0^\infty Q(S,z)\ dz = S$, we have,

$$W^Q(S,\lambda,f(\cdot)) = E_X\left(\frac{1}{\lambda}(\lambda x - \int_0^{\lambda x} Q(S,z)\ dz)\right) = E_X\left(x - \frac{1}{\lambda}(S - \int_{\lambda x}^{\infty} Q(S,z)\ dz\right)$$

$$= E_X\left(x - \frac{S}{\lambda}\right) + E_X\left(\int_{\lambda x}^{\infty} Q(S,z)\ dz\right). \tag{A.75}$$

According to (A.73) and (A.74) and using (A.75), we get:

$$W^Q(S,\lambda,f(\cdot)) = W^P(S,\lambda,f(\cdot)) + \int_0^\infty f(x)\,(x-\frac{S}{\lambda})\,dx =$$

$$W^P(S,\lambda,f(\cdot)) - f^{*(1)}(k)|_{k=0} - \frac{S}{\lambda}$$

$$\Rightarrow W^Q(S,\lambda,f(\cdot)) = Za(S,\lambda,F(\cdot)) - f^{*(1)}(k)|_{k=0} - \frac{S}{\lambda}.$$

A.2.4 Proof of Proposition 8.

According to (3.8) and (3.9),

$$\Pi(S,f(\cdot)) - \Pi(S-1,f(\cdot)) = \lambda\ \left(hW^P(S,\lambda,f(\cdot)) + b\,W^Q(S,\lambda,f(\cdot))\right) -$$

$$\lambda\left(hW^P(S-1,\lambda,f(\cdot)) + b\,W^Q(S-1,\lambda,f(\cdot))\right) =$$

$$\lambda\,h(W^p(S,\lambda,f(\cdot)) - W^P(S-1,\lambda,f(\cdot)) + \lambda\,b\left(W^Q(S,\lambda,u(\cdot)) - W^Q(S-1,\lambda,f(\cdot))\right).$$

Using Theorem 3,

$$\Pi(S,f(\cdot)) - \Pi(S-1,f(\cdot)) = (h+b)\,\lambda\,(Za(S,\lambda,F(\cdot)) - Za(S-1,\lambda,F(\cdot))) - b$$

According to Lemma 2 and using (3.6):

$$\Pi(S,f(\cdot)) - \Pi(S-1,f(\cdot)) = (h+b)Za(S,\lambda,f(\cdot)) - b. \tag{A.76}$$

Therefore, using Lemma 1, we have:

$$(\Pi(S,f(\cdot)) - \Pi(S-1,f(\cdot))) - (\Pi(S-1,f(\cdot)) - \Pi(S-2,f(\cdot))) =$$

$$(h+b)\left(\left(\int_0^\infty f(x)\,Q(S,\lambda x)\,dx\right) - \left(\int_0^\infty f(x)\,Q(S-1,\lambda x)\,dx\right)\right)$$

$$= (h+b)\int_0^\infty f(x)\left(\frac{(\lambda x)^{S-1}}{(S-1)!}e^{-\lambda x}\right)dx \quad \geq 0.$$

Therefore, $\Pi(S,f(\cdot))$ is also convex in S. The optimal base-stock level can be found by solving (A.76).

A.2.5 Proof of Proposition 9.

Using Theorem 3 and replacing $u(\cdot)$ with $f(\cdot)$, the results are obtained.

A.2.6 Proof of Proposition 10.

According to (3.16),

$$\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)=\lambda\ \left(hW^P\left(S,\lambda,u(\cdot)\right)+b\ W^Q\left(S,\lambda,u(\cdot)\right)\right)-$$

$$\lambda\left(hW^P\left(S-1,\lambda,u(\cdot)\right)+b\ W^Q\left(S-1,\lambda,u(\cdot)\right)\right)\ =$$

$$\lambda\ h(W^p\left(S,\lambda,u(\cdot)\right)-\ W^P\left(S-1,\lambda,u(\cdot)\right)\ +\lambda\ b\left(\ W^Q\left(S,\lambda,u(\cdot)\right)-W^Q\left(S-1,\lambda,u(\cdot)\right)\right).$$

Using Proposition 9,

$$\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)=(h+b)\,\lambda\left(Za\left(S,\lambda,U(\cdot)\right)-Za\left(S-1,\lambda,U(\cdot)\right)\right)-b.$$

According to Lemma 2 and using (3.6):

$$\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)=(h+b)Za\left(S,\lambda,u(\cdot)\right)-b. \tag{A.77}$$

First we want to prove that $\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)$ is an increasing function with respect to S. Considering (A.77) and using (3.1), we can elicit the following:

$$\left(\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)\right)-\left(\Pi\left(S-1,S_0,u(\cdot)\right)-\Pi\left(S-2,S_0,u(\cdot)\right)\right)=$$

$$(h+b)\left(\ \left(\int_0^\infty u\left(x\right)Q\left(S,\lambda x\right)dx\ \right)-\ \left(\int_0^\infty u\left(x\right)Q\left(S-1,\lambda x\right)dx\ \right)\right)$$

$$=(h+b)\int_0^\infty u\left(x\right)\left(\frac{\left(\lambda x\right)^{S-1}}{\left(S-1\right)!}e^{-\lambda x}\right)dx\quad\geq 0. \tag{A.78}$$

Therefore, $\Pi\left(S,S_0,u(\cdot)\right)-\Pi\left(S-1,S_0,u(\cdot)\right)$ is increasing in S. Therefore, $\Pi\left(S,S_0,u(\cdot)\right)$ is also convex in S and the optimal base-stock level S^* can be found by solving (A.77).

A.2.7 Proof of Lemma 3.

i. First we prove that $Za\left(S,\bar{\lambda},f(\cdot)\right)$ is monotone decreasing in $\bar{\lambda}=\lambda p$. Similar to Abouee et. al (2011), $Za\left(S,\bar{\lambda},f(\cdot)\right)$, the probability of finding the DC not empty, is obtained as follows:

$$Za\left(S,\bar{\lambda},f(\cdot)\right)=\left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}\left(Q\left(S,\bar{\lambda}\,T\right)-\frac{\mathrm{e}^{(\mu-\bar{\lambda})T}\bar{\lambda}^{S}Q\left(S,T\mu\right)}{\mu^{S}}\right)+\left(1-\left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}\right)Q\left(S,\bar{\lambda}\,T\right).$$

Therefore,

$$\frac{\partial Za\left(S,\bar{\lambda},f(\cdot)\right)}{\partial\bar{\lambda}}=$$

$$\frac{1}{\bar{\lambda}\,\mu^{S}}\left(\left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}\bar{\lambda}^{S}Q\left(S,T\mu\right)\left(\bar{\lambda}\,T-S-S_0\right)\mathrm{e}^{(\mu-\bar{\lambda})T}+\frac{\partial Q\left(S,\bar{\lambda}\,T\right)}{\partial\bar{\lambda}}T\bar{\lambda}\,\mu^{S}\right). \tag{A.79}$$

According to definition of $Q\left(S,\bar{\lambda}\,T\right)$ we have:

$$\frac{\partial Q\left(S,\bar{\lambda}\,T\right)}{\partial\bar{\lambda}}=-\frac{\left(\bar{\lambda}\,T\right)^{S-1}\mathrm{e}^{-\bar{\lambda}\,T}}{\Gamma\left(S\right)}. \tag{A.80}$$

Therefore, using (A.79) and (A.80) we get the following:

$$\frac{\partial Za\left(S,\bar{\lambda},f(\cdot)\right)}{\partial\bar{\lambda}}=$$

$$\frac{1}{\Gamma\left(S\right)\bar{\lambda}\,\mu^{S}}\left(\left(\frac{\bar{\lambda}}{\mu}\right)^{S_0}\bar{\lambda}^{S}\Gamma\left(S\right)Q\left(S,T\mu\right)\left(\bar{\lambda}\,T-S-S_0\right)\mathrm{e}^{(\mu-\bar{\lambda})T}-\left(\bar{\lambda}\,T\right)^{S}\mathrm{e}^{-\bar{\lambda}\,T}\mu^{S}\right). \tag{A.81}$$

If $\bar{\lambda}\,T-S-S_0<0$, obviously $\frac{\partial Za\left(S,\bar{\lambda},f(\cdot)\right)}{\partial\bar{\lambda}}<0$. Otherwise, using (A.81), we show that

$$Q\left(S,T\mu\right)<\frac{T^{S}\mu^{S+S_0}}{\bar{\lambda}^{S_0}\Gamma\left(S\right)\left(\bar{\lambda}\,T-S-S_0\right)\mathrm{e}^{T\mu}}, \tag{A.82}$$

considering that $\bar{\lambda}\,T-S-S_0>0$. Since the right hand side of (A.82) is monotone decreasing in $\bar{\lambda}$, it is sufficient to prove that this inequality holds for maximum value of $\bar{\lambda}=\mu$, i.e.,

$$Q\left(S,T\mu\right)<\frac{\left(T\mu\right)^{S}}{\Gamma\left(S\right)\left(T\mu-S-S_0\right)\mathrm{e}^{T\mu}}.$$

Since $\frac{(T\mu)^S}{\Gamma(S)(T\mu-S)e^{T\mu}} < \frac{(T\mu)^S}{\Gamma(S)(T\mu-S-S_0)e^{T\mu}}$, it is sufficient to show that

$$Q\left(S,T\mu\right) < \frac{\left(T\mu\right)^S}{\Gamma\left(S\right)\left(T\mu-S\right)e^{T\mu}}. \tag{A.83}$$

Let $N(S,T\mu)$ denote the right hand side of (A.83), i.e.,

$$N(S,T\mu) = \frac{\left(T\mu\right)^S}{\Gamma\left(S\right)\left(T\mu-S\right)e^{T\mu}}. \tag{A.84}$$

Now, using induction on S, we prove that the inequality in (A.83) holds. For $S=1$, we have $Q(1,\mu T) = e^{-T\mu}$ and $N(1,\mu T) = \frac{T\mu e^{-T\mu}}{(T\mu-1)}$. Therefore, $Q(1,\mu T) < N(1,\mu T)$. Now assume that (A.83) holds for all $i \leq S-1$, and assume that $Q(S-1,\mu T) < N(S-1,\mu T)$. It is easy to show that

$$Q(S,\mu T) = Q(S-1,\mu T) + \frac{\left(T\mu\right)^{S-1}}{\Gamma\left(S\right)e^{T\mu}} <$$

$$\frac{\left(T\mu\right)^{S-1}}{\Gamma\left(S-1\right)\left(T\mu-S+1\right)e^{T\mu}} + \frac{\left(T\mu\right)^{S-1}}{\Gamma\left(S\right)e^{T\mu}} = \frac{\left(T\mu\right)^S}{\Gamma\left(S\right)\left(T\mu-S+1\right)e^{T\mu}} < N(S,T\mu).$$

ii. According to the definition of $W^P\left(S,\bar{\lambda},f(\cdot)\right)$, we have

$$W^P\left(S,\bar{\lambda},f(\cdot)\right) = \frac{1}{\bar{\lambda}}\sum_{i=1}^{S} Za\left(i,\bar{\lambda},f(\cdot)\right).$$

Since $Za\left(S,\bar{\lambda},f(\cdot)\right)$ is also decreasing in $\bar{\lambda}$, we can conclude that $W^P\left(S,\bar{\lambda},f(\cdot)\right)$ is decreasing in $\bar{\lambda}$. Note that according to (A.73) and (A.75), $W^P\left(S,\bar{\lambda},f(\cdot)\right)$ and $W^Q\left(S,\bar{\lambda},f(\cdot)\right)$ can be written as:

$$W^P\left(S,\bar{\lambda},f(\cdot)\right) = E_X\left(\frac{1}{\bar{\lambda}}\int_{\bar{\lambda}x}^{\infty} Q\left(S,z\right)\ dz\right),$$

$$W^Q\left(S,\bar{\lambda},f(\cdot)\right) = E_X\left(\frac{1}{\bar{\lambda}}\int_{0}^{\bar{\lambda}x} \left(1-Q\left(S,z\right)\right)\ dz\right).$$

Therefore, while $W^P\left(S,\bar{\lambda},f(\cdot)\right)$ is decreasing in $\bar{\lambda}$, $W^Q\left(S,\bar{\lambda},f(\cdot)\right)$ is increasing with respect to $\bar{\lambda}$.

A.2.8 Proof of Theorem 4.

Using (3.22),

$$\frac{\partial U\left(p, Pr\right)}{\partial p} = -c\theta \frac{(\bar{W}(p,S))^{\theta-1}\partial \bar{W}(p,S)}{\partial p}.$$

Based on Lemma 3 and according to (3.23), since $1 - Za\left(S,\bar{\lambda}, f(\cdot)\right)$ and $W^{Q}\left(S,\bar{\lambda}, f(\cdot)\right)$ are monotone increasing in $\bar{\lambda} = \lambda p$, we can conclude that $\frac{\partial(\bar{W}(p,S))^{\theta-1}}{\partial p}$ is positive and consequently $\frac{\partial U(p,Pr)}{\partial p}$ is negative. Since $U\left(p, Pr\right)$ is monotone decreasing in p, the equilibrium probability $\bar{p}$ is the solution of $U\left(\bar{p}, Pr\right) = 0$.

A.2.9 Proof of Lemma 4.

According to the definition of $Za\left(S,\bar{\lambda}, f(\cdot)\right)$ (see Zare et al., 2017) we have:

$$Za\left(S,\bar{\lambda}, f(\cdot)\right) - Za\left(S-1,\bar{\lambda}, f(\cdot)\right) = \frac{\left(\bar{\lambda}\right)^{S-1}}{(S-1)!}\int_{0}^{\infty} e^{-\bar{\lambda}x}\, f\left(x\right)\left(x^{S-1}\right) dx \;\; > 0.$$

Therefore, $Za\left(S,\bar{\lambda}p, f(\cdot)\right)$ is monotone increasing in S. Also we have:

$$W^{Q}\left(S,\bar{\lambda}, f(\cdot)\right) - W^{Q}\left(S-1,\bar{\lambda}, f(\cdot)\right) = \frac{1}{\bar{\lambda}}(Za\left(S,\bar{\lambda}, f(\cdot)\right) - 1) < 0. \tag{A.85}$$

Since $Za\left(S,\bar{\lambda}, f(\cdot)\right) \leq 1$, we get $W^{Q}\left(S,\bar{\lambda}, f(\cdot)\right)$ is monotone decreasing in S. According to the definition of the expected utility function given in (3.22) and (3.23), we can conclude that the expected utility function is monotone increasing in the base-stock level of the DC.

A.2.10 Proof of Proposition 11.

i. According to Lemma 4 and Theorem 4, since the expected utility function is monotone increasing in S and monotone decreasing in p, if the DC sets its base-stock level at zero and the expected utility function with fully join strategy is still positive, all customers will join the system independently of DC policy, i.e. using Theorem 3 and (3.23):

$$U(1, Pr) = R - Pr - c\left((1 - Za(0,\bar{\lambda},f(\cdot)))W^{Q}\left(0,\bar{\lambda}, f(\cdot)\right)\right)^{\theta} = R - Pr - c(-f^{*(1)}(k)|_{k=0})^{\theta}.$$

By solving $U(1, Pr) \geq 0$ and using (3.20) to derive the last term in $U(1, Pr)$, the condition is obtained.

ii. According to the definition of $U(p, Pr)$, if $R < Pr$ (which occurs when $m_\theta < 0$), the expected utility function is always negative and no one will join the system independently of the base-stock level at the DC.

A.2.11 Proof of Lemma 5.

According to (3.22):

$$\frac{\partial^2 U(p, Pr)}{\partial \theta^2} = -c\ (\bar{W}(p, S))^\theta\ \ln^2(\bar{W}(p, S)) < 0.$$

A.2.12 Proof of Proposition 12.

According to (3.24), the equilibrium joining probability is obtained by solving $m_\theta - \bar{W}(p, S) = 0$. Let

$$D(p, S, \theta) = m_\theta - (1 - Za(S,\bar{\lambda},f(\cdot)))(W^Q\left(S,\bar{\lambda}, f(\cdot)\right)) = m_\theta - \bar{W}(p, S). \tag{A.86}$$

Therefore,

$$\frac{\partial \bar{p}}{\partial \theta} = -\frac{\frac{\partial D(\bar{p},S,\theta)}{\partial \theta}}{\frac{\partial D(\bar{p},S,\theta)}{\partial \bar{p}}} = \frac{\frac{\partial m_\theta}{\partial \theta}}{\frac{\partial \bar{W}(\bar{p},S)}{\partial \bar{p}}}.$$

According to the proof of Theorem 4, $\bar{W}(p, S)$ is monotone increasing in p and therefore $\frac{\partial \bar{W}(\bar{p},S)}{\partial p} > 0$. Also we have

$$\frac{\partial m_\theta}{\partial \theta} = -\frac{1}{\theta^2}\left(\frac{R - \mathrm{Pr}}{c}\right)^{\theta^{-1}} \ln\left(\frac{R - \mathrm{Pr}}{c}\right).$$

Since $m_\theta = ((R - Pr)/c)^{\frac{1}{\theta}}$, if $\frac{R-\mathrm{Pr}}{c} > 1$ (equivalently $m_\theta > 1$), then $\frac{\partial m_\theta}{\partial \theta}$ and $\frac{\partial \bar{p}}{\partial \theta}$ are less than zero resulting in that the equilibrium joining probability is monotone decreasing with respect to θ. Otherwise, for $\frac{R-\mathrm{Pr}}{c} < 1$ (equivalently $m_\theta < 1$), the equilibrium joining probability is monotone increasing in θ.

A.2.13 Proof of Proposition 13.

Using (A.86),

$$\frac{\partial \bar{p}}{\partial S} = -\frac{\frac{\partial D(p,S,\theta)}{\partial S}}{\frac{\partial D(p,S,\theta)}{\partial p}} = -\frac{\frac{\partial \bar{W}(p,S)}{\partial S}}{\frac{\partial \bar{W}(p,S)}{\partial p}}.$$

According to the proof of Theorem 4, $\frac{\partial \bar{W}(p,S)}{\partial p}$ is always positive. Also from the proof of Lemma 4, we can conclude that $\frac{\partial \bar{W}(p,S)}{\partial S}$ is always negative. Therefore, the equilibrium joining probability is monotone increasing with respect to the base-stock level at the DC.

A.2.14 Proof of Corollary 7.

Using (A.86),

$$\frac{\partial \bar{p}}{\partial T} = -\frac{\frac{\partial D(p,S,\theta)}{\partial T}}{\frac{\partial D(p,S,\theta)}{\partial p}} = -\frac{\frac{\partial \bar{W}(p,S)}{\partial T}}{\frac{\partial \bar{W}(p,S)}{\partial p}}.$$

We first prove that $\frac{\partial \bar{W}(p,S)}{\partial T} > 0$. According to (3.20), we have $\frac{\partial F^*(s)}{\partial T} = \frac{\partial f^*(s)/s}{\partial T} = -sf^*(s)/s = -f^*(s)$; therefore, using Lemma 2 we get

$$\frac{\partial Za(S,\bar{\lambda},f(\cdot))}{\partial T} = -(-1)^{S-1}\frac{\bar{\lambda}^S}{(S-1)!}\ \frac{d^{S-1}}{dk^{S-1}}\left(f^*(k)\right)|_{k=\bar{\lambda}} =$$

$$-\frac{\left(\bar{\lambda}\right)^S}{(S-1)!}\int_0^\infty e^{-\lambda x}\ f\left(x\right)\left(x^{S-1}\right)dx < 0.$$

This means that $1 - Za(S,\bar{\lambda},f(\cdot))$ is increasing in T. Also considering Theorem 3 we have,

$$\frac{\partial W^Q(S,\bar{\lambda},f(\cdot))}{\partial T} = \frac{\partial Za(S,\bar{\lambda},F(\cdot))}{\partial T} + 1. \tag{A.87}$$

Again using Lemma 2,

$$\frac{\partial Za(S,\bar{\lambda},F(\cdot))}{\partial T} = -(-1)^{S-1}\frac{\bar{\lambda}^S}{(S-1)!}\ \frac{d^{S-1}}{dk^{S-1}}\left(F^*(k)\right)|_{k=\bar{\lambda}} = -Za(S,\bar{\lambda},f(\cdot)).$$

Since $Za(S,\bar{\lambda},f(\cdot))$ is the probability of finding the DC not empty, i.e. $Za(S,\bar{\lambda},f(\cdot)) \leq 1$, using (A.87) and (3.24), we have $\frac{\partial W^Q(S,\bar{\lambda},f(\cdot))}{\partial T} \geq 0$ which results in $\frac{\partial \bar{W}(p,S)}{\partial T} > 0$. Finally, according to the proof of Proposition 12 ($\frac{\partial \bar{W}(p,S)}{\partial p} > 0$), we can conclude that $\frac{\partial \bar{p}}{\partial T} < 0$.

A.2.15 Proof of Lemma 6.

According to the proof of Lemma 4, $W^P(S, \lambda p, f(\cdot))$ is monotone increasing in S and $\bar{W}(p, S)$ is monotone decreasing in S. Also note that when $S = 0$, the modified expected revenue function has a finite value, however when the base-stock level goes to infinity, the modified expected revenue function tends to be negative infinity. Consequently using (3.28), the derivative of $\hat{\Delta}(S, \lambda p, f(\cdot))$, i.e., $\frac{\partial \hat{\Delta}(S,\lambda p,f(\cdot))}{\partial S}$, has at least one root indicating the maximum point.

A.2.16 Proof of Lemma 7.

According to Lemma 3, since $W^P(S, \lambda p, f(\cdot))$ is monotone decreasing in p and $\bar{W}(p, S)$ is monotone increasing in p, using (3.28), the equation $\frac{\partial \hat{\Delta}(S,\lambda p,f(\cdot))}{\partial p} = 0$ has at least one root. Also note that when $p \to 0$, the modified expected revenue function has a finite value, however, when the joining probability increases and $\lambda \to \mu$, due to production capacity, the modified expected revenue function tends to negative infinity. Consequently, it ensures that equation $\frac{\partial \hat{\Delta}(S,\lambda p,f(\cdot))}{\partial p} = 0$ has a root indicating the maximum point.

A.2.17 Proof of Theorem 5.

According to Lemmas 6 and 7, for a given p, there always exists an optimal S which maximizes the modified expected revenue function and for a given S there always exists an optimal p which maximizes the modified expected revenue function. Therefore, there exists an optimal set (p^*, S^*) which maximizes the modified expected revenue function.

A.2.18 Proof of Lemma 8.

Using (3.28) and (3.23) and setting $\theta = 1$, we have

$$\hat{\Delta}(S, \lambda p, f(\cdot)) - \hat{\Delta}(S - 1, \lambda p, f(\cdot)) =$$

$$-\lambda p\left(h(W^p(S,\lambda p,f(\cdot))-W^p(S-1,\lambda p,f(\cdot)))\right)-\lambda pc$$

$$\Big(\big((1-Za(S,\lambda p,f(\cdot)))(W^Q\left(S,\lambda p,f(\cdot)\right))-(1-Za(S-1,\lambda p,f(\cdot)))(W^Q\left(S-1,\lambda p,f(\cdot)\right))\big)\Big)$$

Using Theorem 3 and Lemma 2 we have,

$$\lambda p\left(h(W^p(S,\lambda p,f(\cdot))-W^p(S-1,\lambda p,f(\cdot)))\right)=h\left(Za(S,\lambda p,f(\cdot))\right)$$

Therefore, using (A.85),

$$\hat{\Delta}(S,\lambda p,f(\cdot))-\hat{\Delta}(S-1,\lambda p,f(\cdot))=-h\left(Za(S,\lambda p,f(\cdot))\right)-$$

$$c\lambda p\Bigg((1-Za(S,\lambda p,f(\cdot)))(W^Q\left(S-1,\lambda p,f(\cdot)\right)+\frac{Za(S,\lambda p,f(\cdot))-1}{\lambda p})$$
$$-(1-Za(S-1,\lambda p,f(\cdot)))(W^Q\left(S-1,\lambda p,f(\cdot)\right))\Bigg),$$

which results in

$$=-h\left(Za(S,\lambda p,f(\cdot))\right)-c\lambda p$$
$$\left(-\frac{(1-Za(S,\lambda p,f(\cdot)))^2}{\lambda p}-\left(Za(S,\lambda p,f(\cdot))\right)-Za(S-1,\lambda p,f(\cdot)))\right)W^Q\left(S-1,\lambda p,f(\cdot)\right)\right).$$

Now setting $\hat{\Delta}(S,\lambda p,f(\cdot))-\hat{\Delta}(S-1,\lambda p,f(\cdot))=0$, the result is obtained.