JOINT MARKETING AND OPERATIONAL DECISIONS IN SERVICE/MAKE-TO-ORDER SYSTEMS

A Ph.D. Dissertation
by
Ata Jalili Marand

Aarhus BSS, Aarhus University
Department of Economics and Business Economics

August 2018
Joint Marketing and Operational Decisions in Service/Make-To-Order Systems

by

Ata Jalili Marand

A Ph.D. dissertation submitted to the
School of Business and Social Sciences, Aarhus University,
in partial fulfillment of the Ph.D. degree in Economics and Business

Main supervisor: Associate Professor Hongyan Li
Co-supervisor: Professor Anders Thorstenson

August 2018
“I’ve made the most important discovery of my life. It’s only in the mysterious equation of love that any logical reasons can be found.” — John Nash
Contents

Summary iv

Resumé (Danish summary) vi

Acknowledgments viii

1 Introduction 1

1.1 Service and MTO Systems 2

1.2 Main Bodies of Literature 3

1.2.1 Price and Time Trade-off in Service Systems 3

1.2.2 Time and Reliability Trade-off in Service Systems 4

1.2.3 Service-Inventory Systems 4

1.2.4 Bounded Rationality 5

1.3 Research Methodology 7

1.3.1 Fractional programming 7

1.3.2 Game theory 8

1.4 Structure and Contributions of the Dissertation 9

1.5 Managerial Insights 11

1.6 Future Research 12

2 Joint Inventory Control and Pricing in a Service-Inventory System 14

2.1 Introduction 16

2.2 Literature Review 18

2.2.1 Inventory Control Related Issues in Service Systems 18

2.2.2 Marketing Issues in Service Systems 20

2.3 Problem Description and Modelling 21

2.3.1 Problem Description 21

2.3.2 Model Formulation 25

2.4 Solution Methods 27

2.4.1 Structural Properties 27
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.2 Solution Algorithms</td>
<td>28</td>
</tr>
<tr>
<td>2.5 Extensions</td>
<td>31</td>
</tr>
<tr>
<td>2.5.1 Fill-Rate Constraint</td>
<td>31</td>
</tr>
<tr>
<td>2.5.2 Service-Reliability Constraint</td>
<td>32</td>
</tr>
<tr>
<td>2.5.3 Solving the Constrained Problem</td>
<td>32</td>
</tr>
<tr>
<td>2.6 Numerical Study and Managerial Insights</td>
<td>33</td>
</tr>
<tr>
<td>2.6.1 Performance of the Proposed Algorithms</td>
<td>33</td>
</tr>
<tr>
<td>2.6.2 Problem ((P_{Q-p}))</td>
<td>34</td>
</tr>
<tr>
<td>2.6.3 Problem ((P_{r-p}))</td>
<td>38</td>
</tr>
<tr>
<td>2.7 Conclusion and Future Research</td>
<td>39</td>
</tr>
<tr>
<td>Appendices</td>
<td>41</td>
</tr>
<tr>
<td>2.A Proofs</td>
<td>41</td>
</tr>
<tr>
<td>2.B Extensions</td>
<td>43</td>
</tr>
<tr>
<td>3 Pricing and Inventory Control in Service-Inventory Systems with Boundedly Rational Customers</td>
<td>45</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Literature Review</td>
<td>48</td>
</tr>
<tr>
<td>3.2.1 Pricing in Service Systems</td>
<td>48</td>
</tr>
<tr>
<td>3.2.2 Pricing in Inventory Systems</td>
<td>49</td>
</tr>
<tr>
<td>3.2.3 Service-Inventory Systems</td>
<td>50</td>
</tr>
<tr>
<td>3.2.4 Bounded Rationality in Service Systems</td>
<td>51</td>
</tr>
<tr>
<td>3.3 Problem Description</td>
<td>52</td>
</tr>
<tr>
<td>3.4 Model Formulation</td>
<td>53</td>
</tr>
<tr>
<td>3.4.1 Performance Measures of the System</td>
<td>53</td>
</tr>
<tr>
<td>3.4.2 Rational Customers’ Choice Behavior</td>
<td>55</td>
</tr>
<tr>
<td>3.4.3 Boundedly Rational Customers’ Choice Behavior</td>
<td>56</td>
</tr>
<tr>
<td>3.4.4 Service Provider’s Optimization Problem</td>
<td>58</td>
</tr>
<tr>
<td>3.4.5 Structural Properties and Solution Method</td>
<td>59</td>
</tr>
<tr>
<td>3.5 Numerical Study and Managerial Insights</td>
<td>60</td>
</tr>
<tr>
<td>3.5.1 The Inventory Control Decisions: Endogenous vs. Exogenous</td>
<td>61</td>
</tr>
<tr>
<td>3.5.2 The Impact of the Bounded Rationality</td>
<td>63</td>
</tr>
<tr>
<td>3.5.3 The Impact of the Inventory Control Costs</td>
<td>64</td>
</tr>
<tr>
<td>3.6 Conclusion and Future Research</td>
<td>67</td>
</tr>
<tr>
<td>Appendices</td>
<td>69</td>
</tr>
<tr>
<td>3.A Proofs</td>
<td>69</td>
</tr>
<tr>
<td>4 Quandary of Service Logistics: Fast or Reliable?</td>
<td>71</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>----</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>73</td>
</tr>
<tr>
<td>4.2 Literature Review</td>
<td>75</td>
</tr>
<tr>
<td>4.2.1 Empirical Studies</td>
<td>75</td>
</tr>
<tr>
<td>4.2.2 Analytical Studies</td>
<td>76</td>
</tr>
<tr>
<td>4.3 Homogeneous Customers</td>
<td>79</td>
</tr>
<tr>
<td>4.3.1 Problem Description</td>
<td>79</td>
</tr>
<tr>
<td>4.3.2 Mathematical Formulation and Analysis</td>
<td>80</td>
</tr>
<tr>
<td>4.3.3 Sensitivity Analysis and Managerial Insights</td>
<td>85</td>
</tr>
<tr>
<td>4.4 Heterogeneous Customers</td>
<td>91</td>
</tr>
<tr>
<td>4.4.1 Problem Description and Mathematical Formulation</td>
<td>91</td>
</tr>
<tr>
<td>4.4.2 Sensitivity Analysis and Managerial Insights</td>
<td>93</td>
</tr>
<tr>
<td>4.4.3 Numerical Example</td>
<td>96</td>
</tr>
<tr>
<td>4.5 Conclusion and Future Research</td>
<td>98</td>
</tr>
</tbody>
</table>

**Appendices**

<table>
<thead>
<tr>
<th>4.A Proofs</th>
<th>100</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>5 Price, Time, and Reliability Competition for Service Delivery</th>
<th>111</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>112</td>
</tr>
<tr>
<td>5.2 Literature Review</td>
<td>114</td>
</tr>
<tr>
<td>5.3 The Competition Game and Analysis</td>
<td>118</td>
</tr>
<tr>
<td>5.3.1 Problem Description</td>
<td>118</td>
</tr>
<tr>
<td>5.3.2 The Best Response Problem</td>
<td>121</td>
</tr>
<tr>
<td>5.3.3 The Equilibrium Analysis</td>
<td>123</td>
</tr>
<tr>
<td>5.4 Numerical Study and Managerial Insights</td>
<td>124</td>
</tr>
<tr>
<td>5.4.1 Monopoly Market</td>
<td>125</td>
</tr>
<tr>
<td>5.4.2 Comparison: The PTR and TR Competitions</td>
<td>129</td>
</tr>
<tr>
<td>5.4.3 Comparison: The PTR and PT Competitions</td>
<td>132</td>
</tr>
<tr>
<td>5.4.4 The MCI Model</td>
<td>133</td>
</tr>
<tr>
<td>5.5 Conclusion and Future Research</td>
<td>135</td>
</tr>
</tbody>
</table>

**Appendices**

| 5.A Proofs                                       | 136 |

**Bibliography**

|   | 143 |
Summary

This dissertation consists of four self-contained research papers that address different problems regarding the marketing and operational decisions in service/make-to-order systems. The first two papers consider the pricing and inventory control decisions in a service system with an attached inventory, or the so-called service-inventory system. In a service-inventory system, both an on-hand inventory item and a positive service time are required to fulfill the customer demands. The last two papers address the impact of the integrated pricing, delivery time, and delivery reliability level decisions on the customer’s choice behavior and on the profitability of a profit/revenue-maximizing service provider.

Joint Inventory Control and Pricing in a Service-Inventory System. The first paper addresses the joint inventory control and pricing decisions for a service-inventory system. It is assumed that the customers’ inter-arrival times, service times, and inventory replenishment lead times are independently and exponentially distributed. The service system is modeled as an $M/M/1$ queue. Moreover, it is assumed that the single-item inventory is continuously reviewed under the well-known $(r, Q)$ policy, and that customers arriving during stock-out periods are lost. The demand is modeled as a linear function of the price. We make the following contributions. We integrate pricing and inventory control decisions in the service-inventory system. The problem is formulated and analyzed as a fractional programming problem, and structural properties are explored for the model. Two solution algorithms are proposed: The first one provides optimal solutions, and the second one is efficient even for large problems. Together with the algorithms, the impact of the integrated pricing and inventory control decisions on the overall system performance is investigated.

Pricing and Inventory Control in Service-Inventory Systems with Boundedly Rational Customers. The second paper is built on the first paper. In the second paper, both the pricing and the inventory control decisions are made endogenously, while in the first paper, the pricing and order quantity (reorder point) decisions are made endogenously for an exogenously given reorder point (order quantity). We also employ
a reward-cost function that depends on the service price and the expected waiting time in the system to capture the individual customers’ choice behavior. The problem is considered under the assumptions of either complete rationality or bounded rationality of the customers. We propose an integrated framework to analyze the customers’ choice behavior and to optimize a profit-maximizing service provider’s pricing, reorder point, and order quantity decisions. Our results suggest that the integration of the pricing and inventory control decisions considerably contributes to the service provider’s profitability. Moreover, we show that the customers’ bounded rationality level significantly influences the service provider’s optimal pricing and inventory control decisions.

Quandary of Service Logistics: Fast or Reliable? The third paper addresses the interactive impact of the price, delivery time, and delivery-reliability level on the equilibrium behavior of rational customers and the optimal decisions of a revenue-maximizing service provider. The customers’ sensitivity to the delivery-reliability level is characterized by an increasing concave service value function. We model the operations of the service provider as an $M/M/1$ queue. Two cases are investigated: homogeneous customers and heterogeneous customers. For the homogeneous customers case, we analytically characterize the service provider’s optimal price, delivery-time, and delivery-reliability level decisions. Our results suggest that the service provider has to increase the delivery-reliability level at the expense of a longer delivery time when the customers become more sensitive to the delivery-reliability level. However, the optimal price may either increase or decrease depending on a benchmark value for the delivery-reliability level. For the heterogeneous customers case, our results suggest that when the potential arrival rate is sufficiently large, the service provider always benefits from markets with higher customer heterogeneity levels.

Price, Time, and Reliability Competition for Service Delivery. The last paper extends the results of the previous paper to a competitive setting. We model the problem as the competition among an arbitrary number of profit-maximizing service providers that face boundedly rational customers who can choose to buy the service from one of the service providers or to balk. The existence of a unique Nash equilibrium is proved, and a simple iterative method that converges to this equilibrium is proposed to solve the competition problem. Our results suggest that a service provider with a higher capacity level is not always better off in a market which is more sensitive to the delivery reliability, and that even a firm with a lower capacity level may benefit from a market which is more sensitive to the promised delivery time.


**Pricing and Inventory Control in Service-Inventory Systems with Boundedly Rational Customers.** Den anden artikel bygger oven på den første, og i den anden artikel bliver beslutningerne omkring såvel pris som lagerstyring taget endogent, hvorimod i den første artikel blev beslutninger om pris og ordrestørrelse (genbestillingspunkt) taget endogent for et eksogent bestemt genbestillingspunkt (ordrestørrelse). Desuden bruger vi en belønnings-/omkostningsfunktion, der afhænger af serviceprisen og den forventede ventetid i systemet for på denne måde at bestemme kundernes adfærd og for at optimere en profitsøgende serviceleverandørs beslutninger om pris, genbestillingspunkt og
ordrestørrelse. Vi analyserer problemet både under den antagelse, at kunderne er fuldstændigt rationelle og under den antagelse, at de er begrænset rationelle. Vores resultater viser, at serviceleverandørens fortjeneste øges, når pris- og lagerstyringsbeslutninger integreres. Endvidere påviser vi, at graden af kundernes begrænsede rationalitet i høj grad har indflydelse på serviceleverandørens optimale pris- og lagerstyringsbeslutninger.

Quandary of Service Logistics: Fast or Reliable? Den tredje artikel ser på, hvordan samspillet mellem pris, leveringstid og leveringspålidelighed påvirker rationelle kunders adfærd og de optimale beslutninger, som en udbytteoptimerende serviceleverandør bør tæffe. Kundernes følsomhed i forhold til overholdelsen af leveringstid er karakteriseret ved en stigende konkav funktion. Vi bruger en M/M/1-kø til at modellere serviceleverandørens situation. Vi analyserer først serviceleverandørens optimale beslutninger i forhold til pris, leveringstid og overholdelse af leveringstid i tilfælde af homogene kunder, og vores resultater tyder på, at jo mere følsomme kunderne er over for overholdelsen af leveringstiden, jo mere bliver serviceleverandøren nødt til at sætte sin leveringstid op. Men vi finder samtidig, at leverandøren kan øge eller være tvunget til at sænke den optimale pris afhængigt af benchmark niveauet for leveringspålidelighed. I tilfælde af heterogene kunder viser vores resultater, at serviceleverandøren altid er bedst stillt i et marked med stor heterogenitet blandt kunderne forudsat, at der er et tiltrækkeligt stort antal kunder.

Price, Time, and Reliability Competition for Service Delivery. Den sidste artikel sætter problemet fra forrige artikel ind i et marked med konkurrence. Vi modellerer problemet som konkurrence mellem et vilkårligt antal profitoptimerende serviceleverandører, der står over for begrænset rationelle kunder, som kan vælge at købe servicen fra én af serviceleverandørerne eller at trække sig. Vi påviser eksistensen af en unik Nash-ligevægt, og vi udvikler en simpel, iterativ metode til at bestemme denne ligevægt. Vores resultater tyder på, at en serviceleverandør, der har høj kapacitet, ikke nødvendigvis er bedst placeret i et marked, hvor kundernes vægtning af leveringspålidelighed er høj, og vi påviser, at en leverandør, der har en lavere kapacitet, kan få fordele i et marked, hvor kunderne vægter den lovede leveringstid højt.
The following lines are devoted to thank the people who have in some way helped me to shape this dissertation. What is written here is just what I could partly translate into words, and in no sense can express my true gratitude to these tremendous people.

First and foremost, I would like to thank my supervisor Associate Professor Hongyan Li for her continuous guidance throughout my Ph.D. studies. This dissertation could not have been completed without her supervision, contribution, and encouragement. Besides many other things, I learned from her to have courage and express my idea regardless of what other people may think of that; a lifelong lesson. Doing my Ph.D. under her supervision was a great experience.

I am also very grateful to my co-supervisor Professor Anders Thorstenson. He is an encouraging, supportive, and kind professional. His smart, constructive, and insightful comments were vital for the development of this work. I learned from him that wrong is wrong even though many others do it. I have been very pleased with our collaboration.

During my Ph.D., I was fortunate to spend about 3 months at Linköping University, where I visited Professor Ou Tang. I would also like to thank him for being generous with his time and for fruitful discussions that we had. It was a pleasure for me to co-author a paper with him.

I offer my gratitude to the Department of Economics and Business Economics at Aarhus University and especially to the members of the Cluster for Operations Research, Analytics, and Logistics (CORAL) for providing such a great research environment. I would like to thank Professor Lars Relund Nielsen and Professor Kim Allan Andersen, the current and former heads of the section, for their support during my enrolment as a Ph.D. student at Aarhus University.

Many administrative staff have helped me in various ways during the past 3 years. My sincere gratitudes go to all of them, and in particular to: Betina Sørensen, Ingrid Lautrup, Charlotte Sparrevoohn, and Susanne Christensen. I am also very thankful to Karin Vinding, Betina Sørensen, and Malene Vindfeldt Skals for proofreading different chapters of this dissertation.

I was lucky to get to know Associate Professor Leonidas Enrique de la Rosa and to
have him as my mentor in the second year of my Ph.D. He is a great man with whom I could talk about almost everything. I would like to thank him for his time and for our deep and sincere discussions.

My parents, Maman Fariba and Baba Rasoul, have always believed in me, supported me, and motivated me to follow my heart. I am very grateful and indebted to them. I would also like to thank my sister, Sara, for being so close and encouraging even though we are so far away from each other. I am also very thankful to Maman Esmat and Baba Ahad for supporting me with their inspiring ideas and kind words. My heartfelt gratitudes also go to Farhad and Ulla whose hands I have always felt on my shoulders.

Last but by no means least, I have to thank my awesome wife Samare for her love, forbearance, and unconditional support. Thank you so much dear.

Ata Jalili Marand

August 2018
CHAPTER 1

Introduction
This dissertation addresses marketing and operational issues in service and make-to-order (MTO) systems. Four self-contained research papers, focusing on different problems, are presented in four chapters. This introductory chapter aims at presenting core concepts of the dissertation, distinctions of the different chapters, and main modeling approaches adopted to solve the problems.

1.1 Service and MTO Systems

The service systems and MTO systems share an important feature: an arriving customer cannot be immediately satisfied and there is a time, i.e., a service/processing time, between the customer order placement and the time when the service/product is delivered to the customer. The service/processing time is inevitable in many cases. For example, in consultancy services, the service process starts when the customer clarifies his problem and his request. The customer's participation in consultancy and the consulting time are both indispensable. Another example can be found in transportation services where there is an inevitable lapse of time between the customer's request for a service (e.g., requesting a parcel to be transported) until the time when that service is delivered (e.g., the parcel is delivered at its destination). When it comes to manufacturing systems, a similar time lapse occurs when a firm customizes a product based on the customer's specifications.

When considered under stochastic customer inter-arrival time and service time assumptions and given that resources are limited, e.g., limited processing capacity and/or inventory, the service and MTO systems can be modeled as queueing systems. In fact, the stochastic features and resource restrictions cause the formation of queues, and hence, the problems in the service and MTO systems have primarily been analyzed by applying queueing theory. Hassin (2016) categorizes queueing theory research based on the number of decision makers as the (i) performance analysis, (ii) optimal design, and (iii) analysis of choice behavior.

In the first category, there is no decision maker. The main goal of the studies in this category is to derive the performance measures of the queueing systems under different assumptions regarding the customers' inter-arrival times, service times, number of servers, queue regime, etc (see Table 2.1 for some studies in this category). In the second category, it is assumed that a central decision maker designs the system. The difference between the first and second category is the fact that some of the parameters are subject to decision and that these decisions are made with explicit attention to economic considerations (see, e.g., Stidham, 2009). In the last category, there are at least two decision makers, i.e., the service provider(s) and the customer(s).

Chapter 2 falls within the second category, while Chapters 3-5 consider the choice
behavior of customers as welfare-maximizing independent decision makers and thus fall within the last category. The behavior of an individual customer is affected by the other customers and also by the service providers’ decisions. The interactions between the customers’ and the service providers’ decisions result in an aggregate equilibrium pattern of behavior (Hassin and Haviv, 2003). This equilibrium refers to a state where players cannot unilaterally improve their payoffs. The existence, uniqueness, and characterization of the equilibrium is the main concern of the research in this stream.

1.2 Main Bodies of Literature

There are several different streams of the literature that are related to the problems considered in this dissertation, e.g., the problems regarding the operations and manufacturing interface (e.g., Tang, 2010), the service supply chain management (e.g., Wang et al., 2015), the time-based competition (e.g., Stalk and Hout, 1990), the due date management (e.g., Keskinocak and Tayur, 2004), the pricing in inventory systems (e.g., Chan et al., 2004), etc. We introduce each of these streams in the relevant chapters. Here, we focus on the most related bodies of the literature in order to depict an overall image of the contributions made in each chapter.

1.2.1 Price and Time Trade-off in Service Systems

The use of pricing, or the so-called levying tolls, as a way to regulate the service systems dates back to the seminal work by Naor (1969). He proposes the first mathematical model of a service system that incorporates the rational customers’ decisions. He assumes that the system is observable, and that customers can see the queue length before making their joining/balking decisions. To make the joining/balking decisions, the customers evaluate a cost-reward function, also referred to as utility function. The utility function consists of two parts: the service value, i.e., the reward that the customer gains by the completion of the service, and the full price of the service, i.e., the cost incurred by the customer for being served. The full price itself consists of two parts: the direct price, i.e., the monetary fee of the service, and the waiting cost, i.e., the cost that the customer incurs while waiting for the service, which is often proportional to the delivery time. Naor concludes that the pricing is a tool by which the social planner can reach the maximum social welfare. Later, the paper by Naor is extended to an unobservable system by Edelson and Hilderbrand (1975). The problem has been continuously investigated until recently (see, e.g., Hassin and Haviv, 2003; Hassin, 2016, for two comprehensive literature reviews).

In a service system modeled as a queueing system, the pricing decision and the waiting
time of the customers in the system, i.e., the delivery time, are linked to each other. When the customers are price and time sensitive, increasing the price reduces the customer arrival rate. As a result of the reduced arrival rate, the congestion level in the system decreases, and consequently the expected waiting time is reduced. The decrease in the waiting time, i.e., the waiting cost, makes the service more attractive for the customers. Thereby, the arrival rate and consequently the congestion level increase. The price and time trade-off is the core of the service pricing literature. This trade-off is also a key consideration in different chapters of this dissertation.

1.2.2 Time and Reliability Trade-off in Service Systems

The service pricing literature dominantly views the delivery time as the only delivery performance measure that influences customers’ utility. However, there is abundant empirical evidence to support the claim that customers’ perceived utility of the service is also affected by the delivery reliability (see, e.g., Kelley et al., 1993; Lee and Whang, 2001; Rao et al., 2011, 2014). It has been shown that the delivery reliability directly impacts customer satisfaction (Rosenzweig et al., 2003) and contributes to the profitability of the service provider by allowing him to charge premium prices (Peng and Lu, 2017). In fact, the delivery time and the delivery reliability are acknowledged as two top-level delivery performance measures.

Coupled with the delivery reliability, the above-mentioned price-time trade-off becomes even more complicated. The reason is the tight connection between the delivery time and the delivery reliability. Assume, for instance, a service system with a fixed capacity and exogenously determined price. In this case, the delivery time and the delivery reliability interact in the following way: a shorter delivery time compels the service provider to lower the delivery reliability, and a longer delivery time lets the service provider to deliver a better reliability performance. The time and reliability trade-off is scarcely addressed in the service pricing literature (see, e.g., Boyaci and Ray, 2006; Ho and Zheng, 2004; Shang and Liu, 2011). In Chapters 4 and 5, we add to existing service pricing literature by incorporating the price-time-reliability trade-offs.

1.2.3 Service-Inventory Systems

A service system with an attached inventory, or a so-called service-inventory system, can be defined as a system in which both an on-hand inventory item and a positive service/processing time are required to fulfill each customer’s demand. The operations of the system can be described as follows. Imagine a system at which the customers arrive one by one. Each customer requires a service that takes a positive time. The service
provider needs to keep inventories of an individual item to be able to serve the customer. The item may be a product to be customized based on the customer’s requirements before it can be delivered, e.g., a computer with customized configuration, or materials necessary for conducting the service, e.g., blood bags for a surgery. By the completion of the service, the customer leaves the system and the inventory level drops by one. The inventory is supplied from an external supplier, and the replenishment process is also time consuming.

The term service-inventory systems first appears in the work by Schwarz et al. (2006), but the earliest published study in this area dates back to Sigman and Simchi-Levi (1992). The related literature can be categorized according to the assumptions made about customers’ inter-arrival times, service times, replenishment lead times, waiting hall capacity, replenishment policy, how demands during stock-out periods are treated, etc. The readers are referred to Krishnamoorthy et al. (2011) for a recent literature review. Table 2.1 summarizes the literature based on the attributes of the works.

A service-inventory system is a combination of a service system and an inventory system. As an integrated framework, the service-inventory system is investigated to explore the impact of the inventory system’s operational efficiency as well as related decisions on the service system’s performance and vice versa. Service-inventory systems are more general compared to classical inventory models, and they can better capture the properties of MTO systems. The service-inventory systems have attracted more attention over the past decades as a result of the increased focus on integrated supply chain management.

Based on the classification of Hassin (2016), the majority of the literature about service-inventory systems falls into the first category, i.e., the performance analysis. The aim of these studies is to calculate the performance measures of the service system (e.g., the average queue length and waiting time) and the inventory system (e.g., average inventory level and fill rate) in terms of the complex integrated system parameters. Optimization problems with economic considerations are rarely addressed in this stream. In Chapter 2, we incorporate the pricing decision into service-inventory systems. To the best of our knowledge, we provide the first research on this topic. Furthermore, we extend the results to also consider the service delivery time decision and bounded rationality of customers in Chapter 3.

1.2.4 Bounded Rationality

The term bounded rationality is first introduced by Simon (1957). He uses this term to describe the behavior of the human decision makers who search over alternatives until they find satisfactory, but not necessarily optimal solutions (Simon, 1955). This behavior
reflects the human decision maker’s inconsistency in comparative judgments. Certain phenomena, e.g., within the fields of economics, finance, and marketing, are noticed but cannot be captured by the rationality assumption in a satisfactory way (Spiegler, 2011). Despite these findings in various fields, a behavioral perspective has largely been absent in the field of operations management (Gino and Pisano, 2008).

The traditional operations management literature assumes that the customers are either fully rational or can be induced to behave rationally, and it ignores their cognitive biases and the inconsistency in their preferences. The theoretical and practical implications of incorporating behavioral and cognitive factors into the operations management models are explored by Boudreau et al. (2003), Bendoly et al. (2006), and Gino and Pisano (2008). Gino and Pisano (2008) emphasize the importance of the departures from the rationality assumption in the operations management models and theories. Their arguments question the robustness of the existing managerial results found under the rationality assumption in the operations management field.

Recently, the operations management literature has been complemented by another body of literature that explicitly takes the bounded rationality into account. Ren and Huang (2017) summarize the various approaches employed to model the bounded rationality in the operations management literature, such as the logit choice model, anecdotal reasoning, cognitive hierarchy, hyperbolic discounting, and reference dependence and loss aversion. The logit choice model is one of the models most widely used to capture the customers’ choice behavior (see, e.g., Strauss et al., 2018, for more details). It is based on the stochastic choice rule framework. According to Luce (2005), the human decision maker’s inconsistency in comparative judgments usually results in probabilistic choices, where better alternatives are chosen more often. In their recent literature review, Ren and Huang (2017) list only two papers with queueing models that use the logit choice model, i.e., Huang et al. (2013b) and Li et al. (2016b).

Huang et al. (2013b) are the first to study the bounded rationality in service systems, where customers are sensitive to price and time. They assume that customers may lack the capability to accurately calculate their expected waiting time before making their joining/balking decisions. They study the impact of the bounded rationality on both observable and unobservable systems from both a revenue maximizing service provider and a social planner’s points of views. They conclude that ignoring the bounded rationality may result in significant revenue and welfare losses. We adapt their approach in Chapters 3 and 5 to account for the customers’ bounded rationality in service-inventory systems and service competition, respectively.
1.3 Research Methodology

In this dissertation, we adopt a quantitative approach to analyze different operational and marketing problems in the context of service and MTO systems. We use analytical methods and numerical simulation to obtain results. Among the other theories and methods, fractional programming and game theory are predominantly used in different chapters. In Chapters 2 and 3, we formulated the problems as fractional programming problems that are solved by the means of a parametric fractional programming method. Furthermore, as mentioned in Section 1.1, the problems investigated in Chapters 3-5 model and analyze the interactions of multiple decision-makers, i.e., service provider(s) and customer(s). Game theory provides a powerful tool for analyzing such situations. We briefly describe these methods below.

1.3.1 Fractional programming

Fractional programming deals with optimization problems where the objective function appears as a ratio of functions, such as cost/profit/revenue divided by time, cost divided by volume, or profit/revenue divided by capital. The examples can be found in various fields of engineering, economics, finance, etc. Different methods have been developed to solve the fractional programming problems with different characteristics, e.g., linear or nonlinear functions. Here, we are interested in the solution methods designed for the optimization problems in which the objective function appears as the ratio of nonlinear functions. Different methods are introduced to deal with such problems, e.g., the change of variable method, the linearization method, and the parametric method. An algorithm based on the parametric method is used in Chapters 2 and 3. Therefore, we briefly describe the parametric method below (refer to Stancu-Minasian, 2012, for detailed discussions about the other methods).

Consider the following fractional programming problem:

\[
\max_x q(x) = \frac{n(x)}{d(x)}
\]

\[s.t. \ x \in \Theta_x,\]

where \( q, n, \) and \( d \) are nonlinear functions of \( x \), \( x \) is the vector of decision variables, and \( \Theta_x \) defines the nonempty compact feasible space. The parametric method turns this fractional programming problem into the following parametric fractional programming
problem, for a given real-valued $\gamma$:

$$
\max_x q^\gamma(x) = n(x) - \gamma d(x)
$$

s.t. $x \in \Theta_x$.

Define $f(\gamma)$ as the optimal value of the objective function of the parametric fractional programming problem above, i.e., $f(\gamma) = \max\{q^\gamma(x) = n(x) - \gamma d(x) \mid x \in \Theta_x\}$. It has been shown that an optimal solution to the original problem is also an optimal solution to the parametric fractional programming problem if $f(\gamma) = 0$ (Stancu-Minasian, 2012). In fact, a $\gamma$ solving $f(\gamma) = 0$ is equal to the optimal value of the original problem. This way, solving the original fractional programming problem is reduced to solving the nonlinear equation $f(\gamma) = 0$. In Chapters 2 and 3, we use the Dinkelbach method (Dinkelbach, 1967), to solve $f(\gamma) = 0$. The Dinkelbach method is an iterative procedure. Based on the fact that $f(\gamma)$ is strictly decreasing in $\gamma$ (see, e.g., Lemma 3 in Dinkelbach, 1967), the Dinkelbach method starts with a $\gamma = \gamma_0$ (a lower bound for the original problem), and in each iteration, increases $\gamma$ to generate a sequence that leads to $f(\gamma) = 0$.

### 1.3.2 Game theory

Game theory is a powerful tool for analyzing situations where multiple decision-makers, or so-called players, interact, and each player’s decisions potentially influence other players’ decisions and payoffs. After Morgenstern and von Neumann (1944) who established the foundation of modern game theory, the game theory has been developing in various areas. According to Wang and Parlar (1989), the game-theoretic models can be classified based on the number of players (e.g., two-person games, three-person games, and $n$-person games), nature of payoff functions (e.g., zero-sum games and nonzero-sum games), nature of pre-play negotiations (e.g., cooperative games and non-cooperative games), number of strategies (e.g., finite games and infinite games), state of information available to each player (e.g., complete information games and incomplete information games), and the involvement of time (e.g., static games and dynamic games). In this subsection, we focus on $n$-player, static, non-cooperative games with complete information. The material in this subsection is predominantly from Cachon and Netessine (2006).

A game in its normal form is defined by (i) a set of players indexed by $i = 1, ..., n$, where $n$ is the number of players, (ii) each player’s set of available strategies or strategy space denoted by $X_i$, and (iii) each player’s payoff denoted by $\pi_i : X_1 \times X_2 \times ... \times X_n \rightarrow \mathbb{R}$ that is a function of the strategy profile of the $n$ players, i.e., $(x_1, ..., x_n)$. A strategy corresponds to actions that each player takes in the game. In the normal form, all players choose their strategies simultaneously. As an alternative, the players may make their decisions
sequentially or in turns that is known as the extensive form. A player can choose either to play a pure strategy where he chooses a particular strategy or to play a mixed strategy where he randomly chooses from a set of strategies. In other words, a mixed strategy for player $i$ is a probability distribution over his strategy space $X_i$. Both pure and mixed strategies are considered in this dissertation. We focus on the normal form games and the games where the players’ strategies and payoffs are common knowledge to all players, i.e., complete information. The game is played such that each player chooses a strategy to maximize his expected payoff, and all players make their choices simultaneously.

Given the decisions of the other players, the best choice of player $i$ can be found by solving the best response problem. Suppose the strategies of all player’s other than player $i$ is denoted by $x_{-i}$. Player $i$’s best response, denoted by $x^*_i$, maximizes player $i$’s payoff $\pi_i(x_i|x_{-i})$, i.e., $x^*_i = \arg \max_{x_i} \pi_i(x_i|x_{-i})$. Based on the concept of the best response, the concept of Nash equilibrium can be defined. A Nash equilibrium denoted by $(x^*_1, ..., x^*_n)$ is an outcome of the game in which $x^*_i$ is a best response to $x^*_{-i}$ for all $i = 1, ..., n$. In other words, a Nash equilibrium solution to a non-cooperative game is a vector consisting of the individual players’ decisions such that no player has an incentive to unilaterally deviate from the equilibrium solution.

1.4 Structure and Contributions of the Dissertation

The remainder of this dissertation consists of four self-contained chapters, each presented as a journal paper. These chapters all address specific marketing and/or operational issues in service and MTO systems. The first two chapters consider a service-inventory system and propose integrated frameworks for addressing the pricing and inventory control decisions in such a system. The other two papers address the impact of the integrated pricing, delivery time, and delivery reliability decisions on the choice behavior of customers and on the profitability of a revenue/profit-maximizing service provider in monopolistic and competitive markets, respectively. We also incorporate the notion of the bounded rationality in Chapters 3 and 5. Figure 1.1 shows the positioning of each chapter in the related literature. We briefly describe each chapter below.

Chapter 2 addresses the joint inventory control and pricing decisions for a service-inventory system. It is assumed that the customers’ inter-arrival times, service times, and inventory replenishment lead times are independently and exponentially distributed. The service system is modeled as an $M/M/1$ queue. Moreover, it is assumed that the inventory of an individual item is continuously reviewed under the well-known $(r, Q)$ policy, and that customers arriving during stock-out periods are lost. Additionally, the demand is modeled as a linear function of price. We integrate pricing and inventory control
decisions in the service-inventory system. The problem is formulated and analyzed as a fractional programming problem, and structural properties are explored for the model. Two solution algorithms are proposed: The first one provides optimal solutions, and the second one is efficient even for large problems. Using the algorithms, the impact of the integrated pricing and inventory control decisions on the overall system performance is investigated. We observe that at a high fill rate, the optimal order quantity can be increasing in the reorder point. Moreover, for an active fill-rate constraint, reducing the mean replenishment lead time results in more customers lost on average.

Chapter 3 is built on Chapter 2. In Chapter 3, both the pricing and the inventory control decisions are made endogenously, while in the first paper, the pricing and order quantity (reorder point) decisions are made endogenously for an exogenously given reorder point (order quantity). We also employ a reward-cost function that depends on the service price and the expected waiting time in the system to capture each customer’s choice behavior. The problem is considered under the assumptions of either complete rationality or bounded rationality of the customers. We propose an integrated framework to ana-
lyze the customers’ choice behavior and to simultaneously optimize a profit-maximizing service provider’s pricing, reorder point, and order quantity decisions. Our results suggest that the integration of the pricing and inventory control decisions may substantially contribute to the service provider’s profitability. Moreover, we show that the customers’ bounded rationality level significantly influences the service provider’s optimal pricing and inventory control decisions.

Chapter 4 addresses the interactive impact of the price, delivery time, and delivery-reliability level on the equilibrium behavior of rational customers and the optimal decisions of a revenue-maximizing service provider. The customers’ sensitivity to the delivery-reliability level is characterized by an increasing concave service value function. We model the operations of the service provider as an $M/M/1$ queue. Two cases are investigated: homogeneous customers and heterogeneous customers. For the homogeneous customers case, we analytically characterize the service provider’s optimal price, delivery time, and delivery-reliability level decisions. Our results suggest that the service provider has to increase the delivery-reliability level at the expense of a longer delivery time when the customers become more sensitive to the delivery-reliability level. However, the optimal price may either increase or decrease depending on a benchmark value for the delivery-reliability level. For the heterogeneous customers case, our results suggest that when the potential arrival rate is sufficiently large, the service provider always benefits from markets with higher customer heterogeneity levels.

Chapter 5 extends the results of Chapter 4 to a competitive setting. We model the problem as the competition among an arbitrary number of profit-maximizing service providers that face boundedly rational customers who can choose to buy the service from one of the service providers or to balk. The existence of a unique Nash equilibrium is proved, and a simple iterative method that converges to this equilibrium is proposed to solve the competition problem. Our results suggest that a service provider with a higher capacity level is not always better off in a market which is more sensitive to the delivery reliability; and a firm with a lower capacity level may even benefit from a market which is more sensitive to the promised delivery time.

1.5 Managerial Insights

We summarize the overall managerial insights of the dissertation in this section. Each chapter contains detailed discussions in this regard. As mentioned in the previous section, the first two papers focus on the service-inventory systems. In these two papers, we try to highlight the importance of adopting an integrated framework for studying the service-inventory systems. Our results suggest that a disintegrated decision making procedure
can lead to substantial profit losses. In the last two papers, we shift our focus from the service-inventory systems to the service systems. Based on the empirical results revealing the existing impact of the price, delivery time, and delivery reliability on customers’ choice behavior, we propose an analytical model to capture the interactions between these three factors on customers’ choice behavior and service providers’ optimal decisions. Our models can explain some of the observations made in empirical studies and can serve as an starting point for addressing the issues raised in more complicated real-world systems. In these two papers, we show that ignoring the reliability sensitivity of the customers can result in considerable profit losses in both monopolistic and competitive markets. Moreover, in the second and last papers, we consider bounded rational customer. The analysis made in these two papers enables us to capture human decision maker’s inconsistency in comparative judgments. We show that marginalizing the impact of the customers’ bounded rationality can significantly affect the optimal decisions of the service provider and lead to high profit losses. Our results suggest that having a clear understanding of customers’ cognitive abilities and sensitivities to different factors is critical in the decision making process and its importance should be well understood by the managers.

1.6 Future Research

The overall theme of this dissertation is operations and marketing related decisions in service/MTO systems. We have addressed the joint inventory and pricing optimization, joint service delivery and pricing optimization, and service delivery and pricing competition in service/MTO systems. While numerous research results and managerial insights are explored, the study can still be further developed in different ways. At the end of each chapter, a list of possible future research directions is mentioned. We briefly clarify some of them in this section.

In a service system, the pricing decision is highly dependent on the time performance of the system, and the time performance is tightly linked to the availability of capacity and efficiency of operational activities. We consider operational decisions, i.e., inventory control decisions, in Chapters 2 and 3. It is shown how an integrated approach contributes to the profitability of service providers. However, the capacity decision and capacity cost and their impact on service providers’ profitability is not investigated. Integrating the capacity decision, as a strategic decision, with marketing and operational decisions is worth further study. In addition, we focus on a given inventory control policy, i.e., the \((r, Q)\) policy, and contribute to the literature by integrating the pricing and inventory control decisions in service-inventory systems. It would also be interesting to compare...
the results with a sequential decision making strategy: to first find the optimal inventory control policy for such systems and then integrate it with the pricing decision.

Chapters 4 and 5 consider the time and reliability trade-off in service systems. In Chapter 5, it is assumed that the customers’ utility is linearly dependent on the price, delivery time, and delivery reliability. Chapter 4 relaxes the linear dependency of the utility function to the delivery reliability. In either chapter, the time and reliability are assumed to be separable similar to the related literature (see, e.g., Ho and Zheng, 2004; Boyaci and Ray, 2006; Shang and Liu, 2011). However, the separability assumption may not satisfactorily reflect the tight relation between the time and reliability. Non-linear utility functions are worth to be considered and analyzed in future research.

The service pricing literature predominantly assumes that the customers are rational and their utility can be modeled as a function of price and time (see, e.g., Hassin, 2016). In Chapter 4, we also make a similar assumption and extend the literature by letting the utility function be dependent not only on the price and time, but also on the reliability. However, ignoring all other aspects of human beings’ decision-making process and trying to capture their rational behavior only through their sensitivity to price, time, and reliability is in contradiction with the rationality assumption. Therefore, in order to compensate the limitations of stylized models and classic analytical research methods in operations research areas, other methods, e.g., empirical methods or big data analysis, may also be needed to better capture customers’ rational behavior.
CHAPTER 2

Joint Inventory Control and Pricing in a Service-Inventory System

History: This chapter is based on a paper accepted in the International Journal of Production Economics, July 2017. It is also presented at the 19th International Symposium on Inventory Control, September 2016, Budapest, Hungary.
Joint Inventory Control and Pricing in a Service-Inventory System

Ata Jalili Marand, Hongyan Li, and Anders Thorstenson

Cluster for Operations Research, Analytics, and Logistics,
Department of Economics and Business Economics,
Aarhus University, Denmark

Abstract

This study addresses joint inventory control and pricing decisions for a service-inventory system. In such a system both an on-hand inventory item and a positive service time are required to fulfill customer demands. The service-inventory system also captures the main features of the classical inventory systems with a positive processing time, e.g., make-to-order systems. In this study the service-inventory system is modeled as an $M/M/1$ queue in which the customer arrival rate is price dependent. The single-item inventory is continuously reviewed under an $(r,Q)$ policy. The replenishment lead times of the inventory are exponentially distributed. Furthermore, customers arriving during stock-out periods are lost. The stochastic customers’ inter-arrival times, service times, and inventory replenishment lead times cause the high complexity of the problem and the difficulty in solving it. The aim of this study is to formulate the problem and solve it to optimality. We make three main contributions: (1) We integrate inventory control and pricing in the service-inventory system. The problem is formulated and analyzed as a fractional programming problem, and structural properties are explored for the model. (2) Two solution algorithms are proposed. The first one provides optimal solutions, while the second one is more efficient. (3) The impact of the integrated inventory control and pricing decisions on the overall system performance is investigated. We compare the solutions of the models both with and without fill-rate and service-reliability constraints and report the main interesting managerial insights.

Keywords: Service-inventory system; Operations-marketing interface; Inventory control; Pricing.
2.1 Introduction

Operations and marketing are key functional areas with internal and external focus, respectively, which contribute to the success of a firm (Tang, 2010). The need for alignment of operations and marketing incentives has been widely emphasized. For example, Otley (2002) discusses how dividing a firm into independent units in order to measure its performance in accounting terms would lead to suboptimal system performance. Furthermore, Hausman et al. (2002) empirically show that a successful business strategy implementation increasingly depends on the ability of operations and marketing to work together harmoniously.

Inventory control and pricing are prominent representatives of operations and marketing aspects, respectively, and interact with each other in many cases. For instance, blood bags are required for conducting cosmetic surgeries at hospitals, and cosmetic surgery demand is highly influenced by price (Connell, 2006; Krieger and Lee, 2004). As another example, car oil change service requires inventories of engine oil. The fact that a lower price for such a service brings more customers to the service center, intensifies the importance of the alignment of inventory control and pricing decisions in this system. Moreover, McDonald’s price-sensitive customers cannot be served without beef patties, bread, and cheeses kept in inventories. As an illustration from the electronic product sector, consider a computer assembly firm (e.g., Dell computers) that keeps inventories of semi-finished parts and main components in order to be able to customize and deliver customer demands within a guaranteed delivery time. One of the main parameters influencing PC purchasers’ decisions is the price of the product (Malerba et al., 1999). Hence, it is crucial to understand the importance of integrating operations and marketing decisions in such systems.

The above-mentioned examples share a distinct feature: Both an on-hand inventory and a positive processing/service time are required to fulfill a customer’s demand. Systems of this type are called service-inventory systems. Service-inventory systems represent service systems with an attached inventory as well as inventory systems with a positive processing/service time, e.g., make-to-order systems. Contrary to classical inventory systems, where the time for processing inventory items is assumed to be negligible, this time is regarded as an imperative and integral part of the demand fulfillment process in the service-inventory systems. Even in a pure inventory system, where there is no production stage, items in inventory require some time for retrieval, preparation, packing, and loading (Saffari et al., 2011). A positive processing/service time requires extension of the study from pure inventory systems to service-inventory systems modeled as queueing systems. Compared to classical inventory control models, these models are more general
and capture more realistic features of many systems in both manufacturing and service industries.

Given the rise in product customization, firms are increasingly turning to service as the most promising means of differentiation (Rust and Chung, 2006). Companies have realized that production in itself is insufficient to maintain competitiveness. The fact that a substantial portion of the GDP in big economies comes from the service sector reveals the service-oriented growth of the world economy. This increasingly important role of service, especially in developed economies, has led to a shift of emphasis in marketing science research from traditional product marketing studies to service marketing studies. A stream of literature has emerged that considers marketing issues in service systems. In this stream, researchers study the impact of marketing decisions, e.g., price and lead-time quotations, on the overall performance of service systems. In spite of the rich literature on marketing issues in service systems, such issues have not been considered in the stochastic service-inventory systems. In this study we aim at filling this gap by incorporating the pricing decision in a service-inventory system.

We consider a service-inventory system modeled as an $M/M/1$ queue in which the customer arrival rate is price dependent. The single-item inventory is continuously reviewed and replenished using the well-known $(r, Q)$ policy. The inventory replenishment lead times are exponentially distributed. Moreover, customers arriving during stock-out periods are lost. In this setting, we aim at maximizing the firm’s profit by making pricing and inventory decisions. We make three main contributions. First, we integrate inventory control and pricing decisions in a service-inventory system. We formulate the problem as a fractional programming problem and explore some structural properties. Second, based on the explored structural properties, we propose two solution algorithms and prove convergence to the optimal solution for one of them. Third, we study the impact of the integrated inventory control and pricing decisions on the overall system performance when the problem is solved with and without fill-rate and service-reliability constraints. The results show that, at a high fill rate, the optimal order quantity is increasing in the reorder point, while the direction of the change is opposite at a low fill rate. Moreover, for an active fill-rate constraint, reducing the mean replenishment lead time results in more lost sales.

The remainder of this paper is organized as follows: Section 2.2 reviews the related literature. In Section 2.3, the problem under investigation is described and formulated. Section 2.4 is devoted to analyzing the structural properties and specification of the solution methods used in this paper. Some model extensions are considered in Section 2.5. Numerical results and managerial insights are presented in Section 2.6. Finally, we conclude this study in Section 2.7.
2.2 Literature Review

This study is closely related to two streams of literature: Inventory control related issues in service systems and marketing issues in service systems. In the following subsections, we review representative and closely related studies from these two perspectives.

2.2.1 Inventory Control Related Issues in Service Systems

The principal focus of the studies in this stream is on finding the structural properties of service systems with attached inventories through queue analytic discussions. Compared to the classical inventory control problems, the distinguishing feature of the inventory problem in the context of service systems is the required processing/service time prior to satisfying a demand. The processing/service time, that implies the server has limited capacity, causes a queue of demands. Therefore, most of the literature in this stream aims at dealing with this problem as a queueing problem and thus at finding time-dependent and steady-state distributions for system states.

The first paper published in this area seems to be the one by Sigman and Simchi-Levi (1992). Subsequently, a series of papers by Berman and his fellow scholars (e.g., Berman et al., 1993; Berman and Kim, 1999; Berman and Sapna, 2000; Berman and Kim, 2001; Berman and Sapna, 2001; Berman and Kim, 2004) have developed the area. In general, the related literature can be categorized according to the assumptions about customers’ inter-arrival times, service times, replenishment lead times, waiting hall capacity, and how demands during stock-out periods are treated.

Poisson arrivals and exponentially distributed service times are commonly assumed in the literature. (e.g., Berman and Kim, 1999, 2001; Berman and Sapna, 2001; Berman and Kim, 2004; Schwarz et al., 2006; Schwarz and Daduna, 2006; Saffari et al., 2011; Zhao and Lian, 2011; Saffari et al., 2013; Nair et al., 2015; Krishnamoorthy et al., 2015). However, there are a few exceptions: Sigman and Simchi-Levi (1992) and Berman and Sapna (2000) consider Poisson arrivals and a general distribution of service times. A compound Poisson process for arrivals and exponentially distributed service times are assumed in Sivakumar et al. (2006). Having Poisson arrivals, Manuel et al. (2008) assume a phase-type distribution for service times. In addition, both of the latter studies assume that inventory items are perishable.

Except for a few studies (e.g., Berman and Kim, 1999; Berman and Sapna, 2000; Nair et al., 2015) that assume zero replenishment lead times, positive replenishment lead times are considered in most of the related studies. For instance, the replenishment lead-time distribution is assumed to be Erlang in Berman and Kim (2001) and Berman and Kim...
Berman and Kim (2004) also consider the case of exponentially distributed replenishment lead times. Saffari et al. (2011) assume that the replenishment lead times follow a mixed exponential distribution which is a weighted summation of two exponential distributions with different parameters. The model is applicable when there are two suppliers with different exponentially distributed lead times and the probability of availability of a supplier is equal to its given weight. Furthermore, Saffari et al. (2013) find closed-form solutions for generally distributed replenishment lead times. Beside the aforementioned examples, the exponentially distributed replenishment lead-time assumption is dominant in this stream.

Other inventory related attributes can be used for categorizing the related literature: Lost sales and backorders. Demands during stock-out periods are either lost or backordered. It is worth mentioning that there is a difference between the meanings of backorder and lost sale in the service-inventory systems and in the classical inventory systems. In the service-inventory systems, even if there is on-hand inventory, demands cannot be satisfied immediately, because arriving customers are always backlogged when the server is busy. However, this is not called a backorder in the literature related to service-inventory systems. What distinguishes lost sales from backorders is the way that demands during stock-out periods are treated. If, during a stock-out period, a customer (demand) is allowed to join the system, then it is backordered, otherwise it is lost. Based on this terminology, both lost sale and backorder cases are treated in the literature. For instance, Sivakumar et al. (2006); Schwarz and Daduna (2006); Manuel et al. (2008), and Zhao and Lian (2011) consider the backordering case, and Schwarz et al. (2006); Saffari et al. (2011, 2013); Nair et al. (2015), and Krishnamoorthy et al. (2015) take lost sales into account.

The model considered in our study is similar to that of Schwarz et al. (2006). They study a service system with an attached inventory in which customers arrive according to a Poisson process, and service times and replenishment lead times are exponentially distributed. The system is considered under both limited and unlimited waiting room capacity. They discuss the \((r, Q)\) and \((r, S)\) policies and a general randomized order policy with reorder point \(r = 0\). For the latter case, they conclude that the optimal policy is obtained when the order quantity is deterministic. Joint distribution of the number of customers in the system and the inventory level is derived in product form. They seem to be the first to find closed-form solutions to service facilities with attached inventories and stochastic lead times, under the assumption of lost sales. Two other studies generalize Schwarz et al. (2006): First, Saffari et al. (2013) relax the exponentially distributed replenishment lead-time assumption. Second, Krishnamoorthy et al. (2015) assume that each customer demands an inventory item with a specific probability after the completion
of service. Krishnamoorthy et al. (2011) provide a review of studies related to inventory systems with positive service times.

Given the large body of literature, we use Table 2.1 to summarize the studies quoted above. Focusing on the analytics of service-inventory systems, none of the studies have considered pricing decisions for such systems. We aim at contributing to the development of this stream by incorporating pricing in a service-inventory system.

2.2.2 Marketing Issues in Service Systems

A dominant assumption in marketing studies of service systems is the time sensitivity of customers. According to the way that the time sensitivity is addressed, the related literature can be divided into two groups. One group deals with customers whose valuation of service is a function of the offered price (posted price) and the waiting cost, aggregated into the full price. Customers in this group are assumed to be able to evaluate the expected steady-state waiting time that they would experience and convert it to a waiting cost. This assumption is well-addressed in the literature (e.g., Mendelson, 1985; Mendelson and Whang, 1990; Dewan and Mendelson, 1990; Stidham, 1992; Afche and Mendelson, 2004). The other group of studies considers an aggregate market-level demand function that is decreasing in the price and quoted lead time. Quoting a lead time brings an additional constraint to the problem in order to ensure that the quoted lead time is met with a specific reliability level (a so-called service-reliability constraint). Our study is closely related to the second group. A series of the related studies is reviewed in more details below.

Modeling a firm’s operations as an $M/M/1$ queue, Palaka et al. (1998) examine the lead time setting, capacity utilization, and pricing decisions for a firm serving customers who are sensitive to both price and quoted lead time. The customer arrival rate is modeled as a linear function of price and quoted lead time. The customer arrival rate is modeled as a power function of price and quoted lead time. Later, So (2000) extends So and Song (1998) by considering multiple service providers competing for customers in a market. He assumes that each service provider’s share of the market is determined by a multiplicative competitive interaction model. Ray and Jewkes (2004) assume that the price itself is a function of the quoted lead time. Believing that customers are willing to pay a price premium for shorter delivery times, they investigate the relationship between price and quoted lead time.

Some studies address pricing and lead-time quotation in a decentralized supply chain context. For instance, Liu et al. (2007) consider a decentralized supply chain consisting of one supplier and one retailer. A Stackelberg game model is developed to capture
these two members’ interaction when the supplier is the leader and the retailer is the follower. Pekgün et al. (2008) study the impact of decentralization of the marketing and production departments of a make-to-order system, where the pricing decision is made by the marketing department and the lead-time decision is made by the production department. The problem is formulated as a Stackelberg game with two alternative decision-making sequences. They analyze the inefficiencies created by the decentralization of the price and lead-time decisions. Xiao and Qi (2012) consider a two-stage supply chain in which the demand is linearly dependent on the price, quoted lead time, and reliability of service, similar to Boyaci and Ray (2006) for a segmented market.

Furthermore, some studies consider two customer classes (e.g., Boyaci and Ray, 2003, 2006; Teimoury et al., 2011; Jayaswal et al., 2011; Zhao et al., 2012; Jayaswal and Jewkes, 2016). These studies segment the market based on customers’ heterogeneity in price and lead-time sensitivity. Table 2.2 summarizes and categorizes the literature related to this stream. As one may notice, inventory decisions are not addressed in this stream. Hence, we aim at filling this gap by incorporating inventory decisions in service systems.

It is worth mentioning that the integrated inventory-pricing decision has been extensively studied (see Chan et al., 2004, for a comprehensive review of early contributions). However, the stochastic replenishment lead times and processing/service times along with stochastic inter-arrival times have not previously been considered in the literature. Moreover, most of the related literature is centered on a dynamic pricing and order quantity setting. To the best of our knowledge, marketing issues have not been investigated in the context of service-inventory systems. In this study, we incorporate the pricing decision in such systems. In all of the above-mentioned studies, the customers’ sensitivity to delay has an explicit impact on the arrival rate. In our study, however, the arrival of customers is not directly influenced by the waiting time or the quoted lead time. We assume that the arrival rate is only dependent on the price. This is reasonable when there is a standard industry lead time that is guaranteed with a high reliability level by the system. We further address this standard lead time in Section 2.5.

2.3 Problem Description and Modelling

2.3.1 Problem Description

We consider a service-inventory system as an $M/M/1$ queue in which customers arrive according to a pure Poisson process with the arrival rate $\lambda$. We assume that the system employs a uniform pricing strategy to attract the customers. The arrival rate is price-dependent and assumed to be a decreasing linear function of the price, i.e., $\lambda = \alpha - \beta p$, 


<table>
<thead>
<tr>
<th>Literature</th>
<th>Arrival</th>
<th>Service</th>
<th>Lead-time</th>
<th>Stock-out</th>
<th>Waiting hall</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman et al. (1993)</td>
<td>Constant</td>
<td>Constant/</td>
<td>Zero</td>
<td>Backorder</td>
<td>Capacitated</td>
<td>-</td>
</tr>
<tr>
<td>Sigman and Simchi-Levi (1992)</td>
<td>Poisson</td>
<td>General</td>
<td>Exponential</td>
<td>Backorder</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
<tr>
<td>Berman and Kim (1999)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Zero</td>
<td>Backorder</td>
<td>Uncapacitated/ Capacitated</td>
<td>-</td>
</tr>
<tr>
<td>Berman and Sapna (2000)</td>
<td>Poisson</td>
<td>General</td>
<td>Zero</td>
<td>Backorder</td>
<td>Capacitated</td>
<td>-</td>
</tr>
<tr>
<td>Berman and Kim (2001)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Erlang</td>
<td>Backorder</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
<tr>
<td>Berman and Sapna (2001)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Lost-sale</td>
<td>Capacitated</td>
<td></td>
<td>State dependent service rate</td>
</tr>
<tr>
<td>Berman and Kim (2004)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Exponential/Erlang</td>
<td>Backorder</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
<tr>
<td>Sivakumar et al. (2006)</td>
<td>Compound</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Backorder</td>
<td>Capacitated</td>
<td>Perishable product with exponential lifetime</td>
</tr>
<tr>
<td>Schwarz et al. (2006)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Lost-sale</td>
<td>Uncapacitated/Capacitated</td>
<td>-</td>
</tr>
<tr>
<td>Schwarz and Daduna (2006)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Backorder</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
<tr>
<td>Manuel et al. (2008)</td>
<td>Poisson</td>
<td>Phase-type</td>
<td>Exponential</td>
<td>Backorder</td>
<td>Capacitated</td>
<td>Perishable product with exponential lifetime/Orbit of rejected customers</td>
</tr>
<tr>
<td>Saffari et al. (2011)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Mixed exponential</td>
<td>Lost-sale</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
<tr>
<td>Zhao and Lian (2011)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Backorder</td>
<td>Uncapacitated</td>
<td>- Two classes of customers</td>
</tr>
<tr>
<td>Saffari et al. (2013)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>General</td>
<td>Lost-sale</td>
<td>Uncapacitated</td>
<td>- Inventory items are considered as servers (multi-server)</td>
</tr>
<tr>
<td>Nair et al. (2015)</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Zero</td>
<td>Lost-sale</td>
<td>Uncapacitated</td>
<td>- Customers demand a product with a specific probability</td>
</tr>
<tr>
<td>Krishnamoorthy et al. (2015)</td>
<td>Poisson</td>
<td>Exponential/General</td>
<td>Exponential</td>
<td>Lost-sale</td>
<td>Uncapacitated</td>
<td>- Pricing decision is incorporated</td>
</tr>
<tr>
<td>Our study</td>
<td>Poisson</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Lost-sale</td>
<td>Uncapacitated</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.1: Literature related to inventory issues in service systems
<table>
<thead>
<tr>
<th>Literature</th>
<th>Demand model</th>
<th>Customer class</th>
<th>Decisions</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palaka et al. (1998)</td>
<td>Linear(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT,Cap</td>
<td>Three sets of decisions are considered: Price-lead time, lead time-capacity, and price-lead time-capacity</td>
</tr>
<tr>
<td>So and Song (1998)</td>
<td>Power(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT,Cap</td>
<td>Competition among service providers</td>
</tr>
<tr>
<td>So (2000)</td>
<td>MCI(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT,Cap</td>
<td>Dedicated capacity</td>
</tr>
<tr>
<td>Boyaci and Ray (2003)</td>
<td>Linear(Pr,LT)</td>
<td>Two</td>
<td>Pr,LT,Cap</td>
<td>Market price considered as a function of the quoted lead time</td>
</tr>
<tr>
<td>Ray and Jewkes (2004)</td>
<td>Linear(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT,Cap</td>
<td>Dedicated capacity. Three sets of decisions are considered: Price-lead time, lead time-service reliability, and price-lead time-service reliability</td>
</tr>
<tr>
<td>Boyaci and Ray (2006)</td>
<td>Linear(Pr,LT,SR)</td>
<td>Two</td>
<td>Pr,LT,SR</td>
<td>Decentralized supply chain consisting of one supplier (wholesale price and lead time decisions) and one retailer (retail price) interacting in a Stackelberg game Decentralized marketing department (pricing decision) and production department (lead time decision) interacting in a Stackelberg game</td>
</tr>
<tr>
<td>Liu et al. (2007)</td>
<td>Linear(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT</td>
<td>Decentralized marketing department (pricing decision) and production department (lead time decision) interacting in a Stackelberg game</td>
</tr>
<tr>
<td>Pekgûn et al. (2008)</td>
<td>Linear(Pr,LT)</td>
<td>Single</td>
<td>Pr,LT</td>
<td>Decentralized marketing department (pricing decision) and production department (lead time decision) interacting in a Stackelberg game</td>
</tr>
<tr>
<td>Teimoury et al. (2011)</td>
<td>Linear(Pr,LT)</td>
<td>Two</td>
<td>Pr,LT,Cap</td>
<td>Dedicated/shared capacity</td>
</tr>
<tr>
<td>Jayaswal et al. (2011)</td>
<td>Linear(Pr,LT)</td>
<td>Two</td>
<td>Pr,LT,Cap</td>
<td>Dedicated/shared capacity</td>
</tr>
<tr>
<td>Zhao et al. (2012)</td>
<td>Linear(Pr,LT)</td>
<td>Two</td>
<td>Pr,LT,Cap</td>
<td>Dedicated capacity. Two cases are considered and compared: Uniform and differentiated pricing and lead-time quotation Two-stage supply chain coordination via all-unit quantity discount</td>
</tr>
<tr>
<td>Xiao and Qi (2012)</td>
<td>Linear(Pr,LT,SR)</td>
<td>Single</td>
<td>Pr,LT,SR</td>
<td>Dedicated/shared capacity. Competition among service providers</td>
</tr>
<tr>
<td>Jayaswal and Jewkes (2016)</td>
<td>Linear(Pr,LT)</td>
<td>Two</td>
<td>Pr,LT,Cap</td>
<td>A standard industry lead time is announced to the customers. Inventory decisions are included</td>
</tr>
<tr>
<td>Our study</td>
<td>Linear(Pr)</td>
<td>Single</td>
<td>Pr,Q/ Pr,R</td>
<td>A standard industry lead time is announced to the customers. Inventory decisions are included</td>
</tr>
</tbody>
</table>

Pr = price, LT = lead time, Cap = capacity, SR = service-reliability level, Q = order quantity, R = reorder point

Table 2.2: Literature related to marketing issues in service systems with quoted lead time-sensitive customers
where $\alpha$ is the potential market scale, $\beta$ is the price sensitivity, and $p$ is the product/service price. There is one server serving all customers under the first-come-first-served rule, and one customer is served at a time. Service times follow an exponential distribution with the processing rate $\mu$. When the server is busy, newly arrived customers form a queue in an unlimited waiting hall. Serving each customer requires one inventory item which is removed from the inventory when the service is completed. The inventory is controlled using a continuous review system under the $(r,Q)$ policy, where $r$ is the reorder point and $Q$ is the replenishment order quantity. The inventory replenishment lead times follow an exponential distribution with the parameter $\vartheta$. We assume $0 \leq r < Q$ to avoid excessive reorders by only allowing at most one outstanding replenishment order at a time. As long as there is a positive on-hand inventory, arriving customers are allowed to join the system. However, as soon as a stock out occurs, all arriving customers are rejected (lost). Moreover, we assume that the customers who have already entered the system are patient and will wait for their turns. By the arrival of an outstanding reorder, the service process starts again. Later, we consider a standard industry lead time announced to customers. In that case, it is reasonable to assume that customers, knowing that their waiting times will not exceed a pre-specified time (with a high reliability), are willing to wait.

A combined inventory and service system, such as the one introduced in the previous paragraph, is complex. However, Schwarz et al. (2006), Saffari et al. (2013) (the case with exponentially distributed replenishment lead time), and Krishnamoorthy et al. (2015) (the case where customers require the product with probability one) study the same system analytically to derive the performance measures. They study the joint process of the number of customers in the system and the inventory level. The ergodicity condition and joint distribution of this queueing-inventory process are found. They show that the system is ergodic if and only if $\lambda < \mu$.

The joint distribution of the number of customers in the system and the inventory level is $Pr(i,j) = Pr_n(i)Pr_k(j)$ where $Pr_n(i)$ denotes the probability that the number of customers in the system ($n$) is equal to $i$, and $Pr_n(i) = (1 - \lambda/\mu)(\lambda/\mu)^i$. The limiting distribution of the number of customers in the system coincides with that of the $M/M/1/\infty$ queue with arrival rate $\lambda$ and service rate $\mu$. Similarly, $Pr_k(j)$ denotes the probability that the inventory level ($k$) is equal to $j$, and $Pr_k(j) = E(t_j)/E(\tau)$ where $E(t_j)$ is the expected time that the inventory level equals $j$ during each cycle, and $E(\tau) = \frac{Q}{\lambda} + \frac{1}{\vartheta}(\frac{\lambda}{\lambda + \vartheta})^{r}$ is the expected cycle length. Each cycle starts when the inventory level reaches $r$ and ends when the next cycle starts.

The reason why the distribution of the number of customers in the system and the distribution of the inventory level are independent is discussed in Saffari et al. (2013).
Briefly, due to the lost-sales assumption when the inventory level reaches zero, neither the number of customers in the system nor the inventory level changes until the arrival of the next replenishment order. In other words, the system freezes during stock-out periods. Omitting these freezing periods, the system with remaining working periods is identical to an $M/M/1$ queueing system with inventory and a zero replenishment lead time. The distribution of the number of customers in this system is also identical to that of an $M/M/1$ queueing system which is independent of the inventory level.

During the stock-out periods some customers are lost. Hence, $\lambda$ is not the effective arrival rate. We define $\lambda_e = \lambda(1 - Pr_k(0))$, where $Pr_k(0) = \frac{1}{3w + \frac{1}{\theta}}$, as the number of customers joining the system per time unit or equivalently the effective arrival rate. Similarly, during stock-out periods, no service is conducted. Thereby, the effective service rate can be defined as $\mu_e = \mu(1 - Pr_k(0))$. Our analytic results for the described queueing-inventory process are consistent with those of Schwarz et al. (2006), Saffari et al. (2013), and Krishnamoorthy et al. (2015). To avoid unnecessary replication, we refer to these articles for further details. Instead, we focus on the integration of pricing and inventory control in this system. Hence, based on this setting, we aim at maximizing the firm’s profit by deciding on price ($p$) and order quantity ($Q$) (reorder point ($r$)) for a given reorder point (order quantity). Decisions are considered in pairs, particularly ($Q, p$) and ($r, p$). The notation used in this article is listed in Table 2.3. The mathematical formulation of the described problem is presented in the next subsection.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>arrival rate as a function of price</td>
<td>$S$</td>
<td>loss of goodwill cost per lost sale</td>
</tr>
<tr>
<td>$\lambda_e$</td>
<td>effective arrival rate</td>
<td>$f$</td>
<td>fill rate</td>
</tr>
<tr>
<td>$\mu$</td>
<td>service rate</td>
<td>$w$</td>
<td>standard industry lead time</td>
</tr>
<tr>
<td>$\mu_e$</td>
<td>effective service rate</td>
<td>$s$</td>
<td>targeted service-reliability level</td>
</tr>
<tr>
<td>$\theta$</td>
<td>replenishment lead time parameter</td>
<td>$k$</td>
<td>a constant equal to $\frac{1}{w} \ln \left( \frac{1}{1-s} \right)$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>market scale</td>
<td>$Q$</td>
<td>warehouse capacity</td>
</tr>
<tr>
<td>$\beta$</td>
<td>price sensitivity</td>
<td>$r$</td>
<td>reorder point (decision variable)</td>
</tr>
<tr>
<td>$H$</td>
<td>unit inventory holding cost per unit time</td>
<td>$Q$</td>
<td>order quantity (decision variable)</td>
</tr>
<tr>
<td>$A$</td>
<td>fixed ordering cost</td>
<td>$p$</td>
<td>product/service price (decision variable)</td>
</tr>
<tr>
<td>$N$</td>
<td>unit service cost</td>
<td>$\Pi$</td>
<td>average profit per time unit</td>
</tr>
<tr>
<td>$M$</td>
<td>unit purchase cost</td>
<td>$\Pi^c$</td>
<td>average profit per cycle</td>
</tr>
<tr>
<td>$C$</td>
<td>waiting cost per unit time for each customer</td>
<td>$T$</td>
<td>cycle length</td>
</tr>
</tbody>
</table>

Table 2.3: Notations

### 2.3.2 Model Formulation

In this section, we first formulate the revenue and the cost terms separately. Then, we add them to obtain the integrated objective function which is the average profit per time unit.
Marginal contribution: Let $\pi = p - (M + N)$ be the marginal contribution of selling each unit of the product. The average marginal contribution is

$$MC = \pi \lambda_e = \frac{\pi Q}{\frac{Q}{\lambda} + \frac{1}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r}.$$ (2.1)

Inventory holding cost: According to Schwarz et al. (2006), the average inventory level per time unit is equal to $\sum_{j=0}^{Q+r} j Pr_k(j) = \frac{Q^2 + (2r+1)Q + \frac{Q}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r - 1}{Q + \frac{1}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r}$. Thus, the inventory holding cost per time unit is

$$IC = H \sum_{j=0}^{Q+r} j Pr_k(j) = H \frac{Q^2 + (2r+1)Q + \frac{Q}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r - 1}{Q + \frac{1}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r}. \quad (2.2)$$

Ordering cost: The system incurs a fixed cost $(A)$ each time it makes a replenishment order. In each cycle, one replenishment order is made. Thus, the ordering cost per time unit is

$$OC = \frac{A}{E(\tau)} = \frac{A}{\frac{Q}{\lambda} + \frac{1}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r}. \quad (2.3)$$

Loss of goodwill cost: We assume that lost sales may negatively affect the image of the firm in the long run. Hence, we consider a loss of goodwill cost per unit $(S)$ for the lost sales. Customers arrive with rate $\lambda$ and join the system with rate $\lambda_e$. Therefore, customers are lost with rate $\lambda - \lambda_e = \lambda Pr_k(0)$. Thus, the average loss of goodwill cost is

$$LC = S(\lambda - \lambda_e) = \frac{S\lambda \left(\frac{\lambda}{\lambda + \rho}\right)^r}{Q + \frac{1}{\rho} \left(\frac{\lambda}{\lambda + \rho}\right)^r}. \quad (2.4)$$

Waiting cost: The average number of customers in the queue per time unit in steady state is $\sum_{i=1}^{\infty} (i - 1) Pr_n(i) = \frac{\lambda^2}{\mu(\mu - \lambda)}$. Thus, with a unit waiting cost rate $C$, the average waiting cost per time unit is

$$WC = C \sum_{i=1}^{\infty} (i - 1) Pr_n(i) = \frac{C\lambda^2}{\mu(\mu - \lambda)}. \quad (2.5)$$

Objective function: The long-run average profit function is

$$\Pi = MC - IC - OC - LC - WC. \quad (2.6)$$

Assume that $Q$, $r$, and $p$ are defined on the following intervals, respectively: $\Theta_Q = [\underline{Q}, \overline{Q}]$, $\Theta_r = [\underline{r}, \overline{r}]$, and $\Theta_p = [\underline{p}, \overline{p}]$. Based on this formulation, the problems $(P_{Q-r-p})$ and
(P_{r-p}) are defined as

\[(P_{Q-p}) : \max_{(Q,p)} \{\Pi_{(Q,p)} = \Pi | (Q,p) \in \Theta_Q \times \Theta_p\}\]  \hspace{1cm} (2.7)

\[(P_{r-p}) : \max_{(r,p)} \{\Pi_{(r,p)} = \Pi | (r,p) \in \Theta_r \times \Theta_p\}\].  \hspace{1cm} (2.8)

The variables \(r\) in problem \((P_{Q-p})\), and \(Q\) in problem \((P_{r-p})\) are treated as exogenous variables. The order quantity and reorder point are integers because items are removed from the inventory one by one. However, in order to facilitate the mathematical analysis, we consider them as real values. Thus, \(\Pi\) is continuous and differentiable with respect to all decision variables. Since the feasible space is compact and convex, an optimal solution exists. Moreover, since each variable belongs to an interval in \(\mathbb{R}\), the constraint qualification condition is always met. It is pointless to study the pricing decision of systems with no positive profit in practice. Therefore, without loss of practicality, we assume that there exists at least one feasible point at which the objective value is positive. Because of the continuity of \(\Pi\), it would be an area rather than a single point. In the next section, we introduce some structural properties of these problems and propose two solution methods for them.

2.4 Solution Methods

The stochastic inter-arrival times, processing/service times, and replenishment lead times make the problem complex and difficult to solve. However, we can still find closed form expressions for revenue and cost terms and other performance measures. “In spite of having these explicit functions at hand, it is not immediate to prove structural properties of these functions, e.g., convexity or concavity. Even for the much simpler case of having no explicit service times and service constraints in the system’s behaviour the application of calculus methods to establish these properties for the cost functions in an analytical sense turns out to be impractical” (Schwarz et al., 2006). Moreover, it is not difficult to come up with numerical examples for which even separate concavity of the objective function is violated. Despite these facts, we identify structural properties for the objective functions of problems \((P_{Q-p})\) and \((P_{r-p})\). Later, we exploit these properties to develop two solution algorithms.

2.4.1 Structural Properties

Let \(\Pi^c\) denote each cycle’s expected profit. The profit function equals \(\Pi = \frac{\Pi^c}{E(\tau)}\). Our primary problems \((P_{Q-p})\) and \((P_{r-p})\) can be rewritten as fractional programming models.
Based on concave fractional programming results, the following propositions are explored (for some key results of concave fractional programming refer to Schaible, 1983). For the proofs, refer to Appendix 2.A.

**Proposition 2.1.** $\Pi$ is strictly pseudo-concave with respect to $Q$ on $\Theta_Q$. The unique Optimal $Q$ is either an extreme point ($\underline{Q}$ or $\overline{Q}$) or a point satisfying the first-order condition with respect to $Q$.

**Proposition 2.2.** $\Pi$ is strictly pseudo-concave with respect to $p$ on $[\underline{p}, \overline{p}] \subseteq \Theta_p$. The unique Optimal $p$ is either an extreme point ($\underline{p}$ or $\overline{p}$) or a point satisfying the first-order condition with respect to $p$.

**Proposition 2.3.** $\Pi$ is strictly pseudo-concave with respect to $r$ on $[\underline{r}, \overline{r}] \subseteq \Theta_r$. The unique Optimal $r$ is either an extreme point ($\underline{r}$ or $\overline{r}$) or a point satisfying the first-order condition with respect to $r$.

The strict pseudo-concavity on $\mathbb{R}$ implies uni-modality. Propositions 2.1-2.3 signify that KKT conditions are necessary and sufficient for finding the unique solution to the problem with respect to each of the variables. In this study, however, we aim at integrating pricing and inventory decisions. In the next subsections, two solution algorithms for the joint decisions are proposed.

### 2.4.2 Solution Algorithms

In this subsection, two solution algorithms are proposed based on the results of parametric fractional programming and the properties found in Subsection 2.4.1. The two algorithms are applicable for both problems ($P_{Q-p}$) and ($P_{r-p}$). However, in the following subsections, in order to avoid replication, we only describe the solution algorithms regarding problem ($P_{Q-p}$). The same arguments hold for problem ($P_{r-p}$).

**Parametric Fractional Programming (PFP) Algorithm**

The PFP algorithm is based on parametric fractional programming (see, e.g., Crouzeix and Ferland, 1991; Stancu-Minasian, 2012). The main idea is to convert the fractional programming problem to a parametric fractional programming problem and to solve a non-linear equation in order to find the solution to the primary problem. Consider parametric problem ($P_\gamma$) for a $\gamma \in \mathbb{R}$ as

$$ (P_\gamma) : \max_{Q,p} \{ f_\gamma(Q, p) = \Pi^c - \gamma E(\tau) \mid (Q, p) \in \Theta_Q \times \Theta_p \} . \tag{2.9} $$
Let $F(\gamma)$ be the optimal value of the objective function of problem $(P_\gamma)$. It can be proven that if $F(\gamma) = 0$, then an optimal solution to problem $(P_\gamma)$ is also an optimal solution to problem $(P_{Q-p})$ (see Stancu-Minasian, 2012, for details). To solve this non-linear equation, we use the Dinkelbach method (Dinkelbach, 1967) as shown in Algorithm 2.1. For a given $\gamma > 0$, $f_\gamma(Q, p)$ is separately concave. It is clear that $(P_\gamma)$ has an interior optimal solution if the system of non-linear equations $(\frac{\partial f_\gamma(Q, p)}{\partial Q} = 0, \frac{\partial f_\gamma(Q, p)}{\partial p} = 0)$ has a solution on $\Theta_Q \times \Theta_p$. The first-order condition yields the optimal $Q$ as $Q_{(p, \gamma)} = \left(\frac{\lambda}{\pi} - \frac{H(2n+1)}{2\lambda} + \frac{H}{\sqrt{\lambda+\theta}} \right)^{-1} - \frac{C\lambda \mu (\mu - \lambda) - \gamma}{\mu (\mu - \lambda) - \gamma}$. Substituting $Q_{(p, \gamma)}$ in $\frac{\partial f_\gamma(Q, p)}{\partial p} = 0$, a polynomial equation with respect to $p$ should be solved for which different numerical methods have been developed. Hence, for the given $\gamma$, the optimal price, i.e., $p_{(\gamma)}$, can be found. Having the optimal price, the corresponding optimal order quantity, i.e., $Q_{(p, \gamma)}$, can be calculated. If the solution to the system of equations is not feasible, the optimal solution is on the boundaries. Fixing one of the variables at its extreme, the other one can easily be found by the first-order condition (Propositions 2.1-2.2).

We start with a $\gamma = \gamma_0$, which is a lower bound for the primary problem, e.g., $\gamma_0 = 0$ (recall that we have assumed that there is a feasible point at which the objective value

\begin{algorithm}
\caption{Pseudo-code of the PFP algorithm}
\begin{algorithmic}
\State Properly set $\epsilon_0$ and $\Delta\gamma$;
\State Set $\gamma_1 = 0$;
\State Set $i = 1$;
\While{$\Delta\gamma > \epsilon_0$}
\State $p_i = p_{(\gamma_i)}$;
\State $Q_i = Q_{(p_i, \gamma_i)}$;
\If{$(Q_i, p_i) \notin \Theta_Q \times \Theta_p$}
\State $p_{\text{temp}1} = \arg\max\{f_\gamma(Q, p)\}$;
\State $Q_{\text{temp}1} = \arg\max\{f_\gamma(Q, p)\}$;
\State $(Q_i, p_i) = \arg\max\{f_\gamma(Q, p_{\text{temp}1}), f_\gamma(Q, p_{\text{temp}2}), f_\gamma(Q_{\text{temp}1}, p), f_\gamma(Q_{\text{temp}2}, p)\}$;
\EndIf
\If{$f_\gamma(Q_i, p_i) = 0$}
\State Terminate While cycle.
\Else
\If{$f_\gamma(Q_i, p_i) > 0$}
\State $\gamma_{i+1} = \Pi(Q_i, p_i)$;
\State $i = i + 1$;
\EndIf
\EndIf
\State $\Delta\gamma = \gamma_i - \gamma_{i-1}$;
\EndWhile
\State Set $(Q^*, p^*) = (Q_i, p_i)$.
\end{algorithmic}
\end{algorithm}

\footnote{For problem $(P_{r-p})$, we have $r_{(p, \gamma)} = \frac{1}{z} \ln \frac{x}{y}$ in which $x = \frac{kQ}{\lambda}$, $y = -\frac{kQ}{\gamma} - \frac{S}{\alpha} - \frac{C\lambda^2}{\mu (\mu - \lambda)} - \frac{\gamma}{\pi}$, and $z = \ln \frac{\lambda}{\lambda+\theta}$.}
is positive). Knowing that \(F(\gamma)\) is strictly decreasing in \(\gamma\) (Stancu-Minasian, 2012), \(\gamma\) is updated in each step to generate a sequence \(\{\gamma_n\}\) leading to \(F(\gamma) = 0\). Now, consider Lemma 2.1.

**Lemma 2.1.** If \(f_\gamma(Q, p) > 0\), then \(\gamma' = \Pi_{(Q, p)} > \gamma\).

In each step, \(\gamma\) is updated according to Lemma 2.1. For instance, in step \(n\), if \(f_\gamma(Q_n, p_n) > 0\), then we set \(\gamma_{n+1} = \Pi_{(Q_n, p_n)}\). Otherwise the algorithm is terminated. Consider Proposition 2.4.

**Proposition 2.4.** The PFP algorithm monotonically converges to the optimal solution.

### Simple Iterative (SI) Algorithm

A simple iterative algorithm is proposed in this subsection according to the results in Subsection 2.4.1. The idea behind this algorithm is based on the joint uni-modality of the objective function. Our observations of a large number of numerical examples suggest that the objective function is jointly uni-modal with respect to \(Q\) and \(p\) although we have not been able to prove it. The SI algorithm works as follows: For a given \(Q_0\) (if there exists a feasible \(p\) at which the objective value is positive), there must be one \(p_1 \in \Theta_p\) for which \(\Pi_1 = \Pi_{(Q_0, p_1)} \geq \Pi_{(Q_0, p)}\) for any \(p \in \Theta_p\) (and \(\Pi_1 > \Pi_{(Q_0, p)}\) for any \(p \in \Theta_p \setminus \{p_1\}\)) due to Proposition 2.2. Now having \(p_1\), there must be one \(Q_1 \in \Theta_Q\) for which \(\Pi_2 = \Pi_{(Q_1, p_1)} \geq \Pi_{(Q, p_1)}\) for any \(Q \in \Theta_Q\) (and \(\Pi_2 > \Pi_{(Q, p_1)}\) for any \(Q \in \Theta_Q \setminus \{Q_1\}\)) due to Proposition 2.1. This implies that \(\Pi_2 \geq \Pi_1\). Following this procedure, the objective function never gets worse, i.e., \(\Pi_i \geq \Pi_{i-1}\) for all integer \(i\). This procedure can be run as long as the difference between the objective values of two successive steps is greater than a pre-specified tolerance, i.e., \(\epsilon_0\), which can be chosen appropriately small. The pseudo-code of the algorithm is presented in Algorithm 2.2. Despite the improvement feature of this algorithm, there is no proof of convergence to the global solution. As \(\Pi_i\)
is an increasing sequence and the problem is bounded, it converges to some number, say \( \hat{\Pi} \). It is still an open question whether \( \hat{\Pi} = \Pi^* \).

Using the proposed algorithms, both problems \((P_{Q-p})\) and \((P_{r-p})\) can be solved. In the next section, we extend the models to include fill-rate and service-reliability constraints, and investigate how these extended problems can be solved.

## 2.5 Extensions

Fill-rate and service-reliability constraints are commonly applied to inventory and service systems, respectively. Since we are dealing with service-inventory systems in this study, both constraints can be included. In this section, we address how these constraints affect the joint inventory and pricing decisions. Before looking into them, we introduce some essential constraints. In order to avoid negative unit marginal contribution we set \( M + N \leq p \). Furthermore, from the demand positivity and ergodicity conditions, i.e., \( 0 \leq \lambda < \mu \), we have \((\alpha - \mu)/\beta < p \leq \alpha/\beta\). Hence, we update \( \bar{p} = \max\{(\alpha - \mu)/\beta + \epsilon, M + N\} \), \( \bar{p} = \alpha/\beta \) where \( \epsilon \) is very small. In addition, we have \( Q = r + 1 \) in problem \((P_{Q-p})\), and \( r = 0 \) and \( \bar{r} = Q - 1 \) in problem \((P_{r-p})\). We also assume that there is an upper bound for the order quantity, i.e., \( \bar{Q} \), which is chosen sufficiently high in order not to influence our results.

### 2.5.1 Fill-Rate Constraint

The fraction of demand that is satisfied by inventory on hand is the so-called fill rate. To satisfy a required fill rate, i.e., \( f \), the following constraint must hold:

\[
\frac{E(\text{demand satisfied per unit time})}{E(\text{total demand per unit time})} = \frac{Q}{Q + \frac{\lambda}{\vartheta}(\frac{\lambda}{\lambda+\vartheta})^r} \geq f. \tag{2.10}
\]

For a high required fill rate, a long replenishment lead time, and a small reorder point, the constraint tends to be active. Whenever the fill-rate constraint is active, the optimal order quantity can be expressed as a function of the arrival rate (or the price), replenishment lead time, reorder point, and required fill rate, i.e., \( Q = \frac{f\lambda}{(1-f)\vartheta} \left(\frac{\lambda}{\lambda+\vartheta}\right)^r \). By substituting \( Q \) in problem \((P_{Q-p})\), the dimension of the decision problem is reduced. Given other parameters, for an active fill rate, the optimal order quantity is decreasing in price, decreasing in reorder point, increasing in mean replenishment lead time, and increasing in fill rate. Likewise, when the fill-rate constraint is active, we have \( r = \frac{\ln((1-f)Q\vartheta) - \ln(f\lambda)}{\ln(\lambda) - \ln(\lambda+\vartheta)} \). By substituting \( r \) in problem \((P_{r-p})\), a decision problem of a lower dimension has to be solved. Analogously, having all other parameters fixed, for an active
fill-rate constraint, the optimal reorder point is decreasing in price, decreasing in order quantity, increasing in mean replenishment lead time, and increasing in fill rate.

2.5.2 Service-Reliability Constraint

We assume that there is an established standard industry lead time that the system assures with a high probability. The queueing system of this study has the effective arrival rate $\lambda_e$, and the effective service rate $\mu_e$. As a result of $M/M/1$ queue analytics, to assure that the steady-state actual waiting time of a customer will not exceed a quoted lead time, i.e., $w$, at least with a probability of $s$, the following constraint must hold:

$$P(\text{waiting time} < w) = 1 - e^{-(\mu_e - \lambda_e)w} \geq s,$$  

(2.11)

which can be rewritten as

$$\frac{Q(\mu - \lambda)}{Q + \frac{1}{w}} \geq \frac{1}{w \ln \left( \frac{1}{1 - s} \right)}.$$  

(2.12)

Note that, even for $G/G/c$ queues, the tail of the waiting time distribution for high levels of service reliability can be accurately approximated by an exponential distribution (Abate et al., 1995, 1996).

For a short quoted lead time with a high reliability, a long replenishment lead time, a small reorder point, and a low service rate, the service-reliability constraint tends to be active. In this case, the optimal order quantity equals

$$Q = \frac{k \lambda}{\varphi(\mu - \lambda - k)(\lambda + \varphi)} \left( \frac{\lambda}{\lambda + \varphi} \right)^r \text{ where } k = \frac{1}{w \ln \left( \frac{1}{1 - s} \right)}.$$  

By substituting $Q$ in problem $(P_{Q-p})$, the dimension of the decision problem is reduced. Fixing all other parameters, for an active service-reliability constraint, the optimal order quantity is decreasing in price, decreasing in reorder point, increasing in mean replenishment lead time, decreasing in service rate, and decreasing in quoted lead time. Similarly, in problem $(P_{r-p})$ when the service-reliability constraint is active, the optimal reorder point equals

$$r = \frac{\ln (Q \varphi(\mu - \lambda - k)) - \ln (k \lambda)}{\ln (\lambda) - \ln (\lambda + \varphi)}.$$  

By substituting $r$ in problem $(P_{r-p})$, a problem of lower dimension has to be solved. Correspondingly, for a fixed set of parameters and an active service-reliability constraint, the optimal reorder point is decreasing in price, decreasing in order quantity, increasing in mean replenishment lead time, and decreasing in quoted lead time.

2.5.3 Solving the Constrained Problem

This subsection explains how the fill-rate and service-reliability constraints can be applied to problem $(P_{Q-p})$. The same discussion holds for problem $(P_{r-p})$. Let $(\bar{Q}, \bar{p})$ represent
a solution found by the solution algorithms proposed in Subsection 2.4.2, and let \((\hat{Q}, \hat{p})\) be the intersection point of the fill-rate and service-reliability constraints. Though unimodality is not guaranteed in general, numerical examples always show uni-modality. If the uni-modality assumption holds, what follows can be justified. Solving the problem subject to the fill-rate and service-reliability constraints, if point \((\tilde{\theta}Q, \tilde{\theta}p)\) is not feasible, then an interior solution is either

\[
Q(p) = \frac{k\lambda}{(\mu - \lambda - k)\theta}(\frac{\lambda}{\lambda + \theta})^r \quad \text{and} \quad p = \text{arg}\left\{\frac{\partial \Pi(Q(p),p)}{\partial p} = 0\right\} \quad \text{on intervals} \quad [Q, \bar{Q}] \quad \text{and} \quad [\min\{\bar{p}, \max\{\tilde{\theta}p\}\}, \bar{p}].
\]

Otherwise, the optimal solution is on the boundaries. It can also be concluded that if only the fill-rate constraint is applied and violated at point \((\tilde{\theta}Q, \tilde{\theta}p)\), then an interior solution is

\[
Q(p) = \frac{k\lambda}{(\mu - \lambda - k)\theta}(\frac{\lambda}{\lambda + \theta})^r \quad \text{and} \quad p = \text{arg}\left\{\frac{\partial \Pi(Q(p),p)}{\partial p} = 0\right\} \quad \text{on intervals} \quad [Q, \bar{Q}] \quad \text{and} \quad [\min\{\bar{p}, \max\{\tilde{\theta}p\}\}, \bar{p}].
\]

Clearly, if the solutions found are not interior points, then the optimal solution is on the boundaries. For more details, see Appendix 2.B.

2.6 Numerical Study and Managerial Insights

In this section, we use the algorithms introduced in Section 2.4.2 to solve a series of problem instances. The algorithms are coded in Matlab. First, we use the parameters of the numerical instances in Saffari et al. (2013) to test the effectiveness and efficiency of the proposed algorithms. Then, a sensitivity analysis is carried out regarding each key parameter. The initial parameters are \(\alpha = 200, \beta = 4, \theta = 1, \mu = 50, \omega = 0.25, M = 1, N = 2, H = 1, A = 200, C = 25, S = 50, \bar{Q} = 1000\) and \(r = 10\). The problem is solved for a range of fill rates and service-reliability levels.

2.6.1 Performance of the Proposed Algorithms

Table 2.4 shows the performance of the two algorithms for a number of problem instances with different reorder points. Although the gap between the solutions found is zero, the time performance of the SI algorithm strictly dominates that of the PFP algorithm. In other words, although both algorithms are effective, the SI algorithm is more efficient. From the run time columns, we can also conclude that whereas the run times of the PFP algorithm are fluctuating for different reorder points, those of the SI algorithm are consistently increasing in the reorder point. Thus, we use the SI algorithm to solve problem \((P_{Q-p})\).

Using the proposed algorithms in each step of solving problem \((P_{Q-p})\), a polynomial
The equation should be solved for which efficient numerical methods are developed. However, in each step of solving problem \((P_r - p)\), a fractional equation with logarithmic terms in both numerator and denominator should be solved. Due to this fact and in spite of the effectiveness of the proposed algorithms for solving problem \((P_r - p)\), they are not efficient. Therefore, the results for problem \((P_r - p)\) are obtained through complete enumeration.

Our focus is on the impact of the mean replenishment lead time, service rate, fill rate, and service-reliability level on the optimal solution and the performance measures of the service-inventory system described in this paper. In order to investigate the impact of the fill-rate and the service-reliability levels, we fix one of them at a time and study the impact of varying the other constraint. The main interesting findings of the sensitivity analysis are presented in the following two subsections.

### 2.6.2 Problem \((P_Q - p)\)

#### Reorder Point

For a fixed price and an active fill-rate constraint, the optimal order quantity is decreasing in the reorder point, i.e., \(\frac{\partial Q}{\partial r} = \frac{\lambda}{\sigma(1-f)} \ln\left(\frac{\lambda}{\lambda + \sigma}\right) < 0\). However, as shown in Figure 2.1a, while the optimal order quantity is decreasing in the reorder point for low fill rates, it is increasing in the reorder point for high fill rates (98% in this case). This effect is caused by incorporating the pricing decision in the inventory problem.

**Observation 2.1.** In problem \((P_Q - p)\), the optimal order quantity is increasing in the reorder point when the required fill rate is sufficiently high.

We also see that the optimal price is decreasing in the reorder point and increasing in the fill rate as shown in Figure 2.1b. Increasing the price reduces the arrival rate.
The number of unsatisfied demands thereby decreases and it becomes possible to meet a higher fill rate. Furthermore, Figure 2.1b shows that while the pricing decision stays unchanged for different reorder points with low fill rates, it plays a role in the system regularization for higher fill rates. In other words, the pricing decision is more important in the presence of a high fill-rate constraint. Since fulfilling high fill rates is more crucial in competitive environments, the significance of the current study should be particularly understood in the context of competitive markets.

**Observation 2.2.** In problem \((P_{Q-p})\), at high fill rates, the impact of the reorder point on the optimal order quantity and pricing decisions, and thereby on the system’s profitability and performance, is more significant.

When we solve the problem subject to a service-reliability constraint, we get a steep increase in the optimal price needed to decrease the arrival rate, and hence reduce the number of customers in the queue and their waiting times (Figure 2.2b). In this figure, however, the price seems to be invariant with respect to changes in the reorder point. Moreover, the optimal order quantity is also decreasing in both the service-reliability level and the reorder point as shown in Figure 2.2a. Additionally, the optimal profit is decreasing in both the fill rate and the service-reliability level.

**Observation 2.3.** In problem \((P_{Q-p})\), in the presence of a service-reliability constraint, the optimal price is invariant with respect to the reorder point.
Service reliability level \((s)\)

Optimal order quantity \((Q)\)

\(r = 50\)
\(r = 40\)
\(r = 30\)
\(r = 20\)
\(r = 10\)
\(r = 0\)

(a)

Optimal price \((p)\)

\(r = 50\)
\(r = 40\)
\(r = 30\)
\(r = 20\)
\(r = 10\)
\(r = 0\)

(b)

Figure 2.2: The impact of the reorder point and service reliability on (a) the optimal order quantity and (b) the optimal price

**Replenishment Lead Time**

The average number of customers lost is closely related to the average length of the stock-out periods. Among other parameters, the length of the stock-out period is dependent on the replenishment lead time. Intuitively, it is expected that as the replenishment lead time decreases, the stock-out period becomes shorter, and consequently the number of customers lost decreases. The decreasing line in Figure 2.3 supports this interpretation. However, as soon as the fill-rate constraint becomes active, the average number of customers lost increases as the mean replenishment lead time decreases. In other words, a lower mean replenishment lead time results in a higher average number of customers lost. For an active fill-rate constraint, the average number of customers lost is equal to \((1 - f)\lambda\). A shorter replenishment lead time causes the optimal price to decrease in order to increase the inflow of customers to the system \((\lambda)\). Consequently, a greater number of customers is lost. A higher fill rate also contributes to less lost customers. A shorter replenishment lead time also leads to higher profitability and system utilization. Hence, although an investment for shortening the replenishment process would yield a higher profit for the firm, the firm may end up losing more customers on average.

**Observation 2.4.** For an active fill-rate constraint, the average number of lost customers is decreasing in the mean replenishment lead time.

**Service Rate**

The variation of the system utilization with respect to the service rate for different fill rates and service-reliability levels is shown in Figures 2.4a and 2.4b, respectively. These
Parameter of replenishment lead time distribution ($\vartheta$)

Figure 2.3: The impact of the mean replenishment lead time on the number of customers lost.

Observation 2.5. The system utilization is increasing in the service rate at high service-reliability levels (when the service-reliability constraint is active), whereas it is decreasing at high fill rates (when the fill-rate constraint is active).

Figure 2.4: (a) The impact of the service rate and fill rate on the system utilization, and (b) the impact of the service rate and service reliability on the system utilization.
2.6.3 Problem \((P_{r-p})\)

Order Quantity

Figures 2.5a and 2.5b show how the optimal reorder point and the optimal price, respectively, vary with respect to the order quantity for different service-reliability levels. Figure 2.5b shows that the optimal price is increasing in the service-reliability level. Moreover, for an active service-reliability constraint, the optimal price is invariant to the order quantity level. From Figure 2.5a we can conclude that the optimal reorder point is decreasing in the order quantity at high service-reliability levels, e.g., 99.5%.

![Figure 2.5: The impact of the order quantity and service reliability level on (a) the optimal reorder point and (b) the optimal price](image)

The optimal prices obtained by solving problem \((P_{r-p})\) subject to a fill-rate constraint are presented in Figure 2.6a. This figure shows that although the optimal price is increasing in the fill rate for low order quantities, setting the order quantity too high results in a lower optimal price, or equivalently a higher demand.

Figure 2.6b shows the optimal profit with respect to the fill rate for different order quantities. From this figure we can conclude that the profit decreases as the fill rate increases for a low order quantity. Nevertheless, setting the order quantity at a high value makes the profit invariant to the fill rate. Furthermore, a too high order quantity results in a higher inventory holding cost, which reduces the profit.

Replenishment Lead Time

In problem \((P_{r-p})\), reducing the mean replenishment lead time has the same effect on the number of customers lost as it has in problem \((P_{Q-p})\) (Observation 2.4). Although
Figure 2.6: The impact of the order quantity and fill rate on (a) the optimal price and (b) the optimal profit

The probability of a stock out decreases as the mean replenishment lead time decreases, more customers are lost with a shorter replenishment lead time. Again, this is due to the pricing effect.

**Service Rate**

The system utilization exhibits a similar behavior in problem \((P_r-p)\) as it does in problem \((P_Q-p)\) (Observation 2.5). As long as the service-reliability constraint is active, the system utilization dramatically increases in the service rate. However, for an active fill-rate constraint, the system utilization is decreasing in the service rate.

### 2.7 Conclusion and Future Research

In this study we address the integration of pricing and inventory decisions in a service-inventory system with stochastic customer inter-arrival times, processing/service times, and replenishment lead times. This system represents service systems with an attached inventory as well as inventory systems with a positive service time. The applications of the current study span a wide range of manufacturing and service industries. Specifically in this study, the system is modeled as an \(M/M/1\) queue with a price-dependent customer arrival rate and an exponentially distributed inventory replenishment lead time. The inventory is continuously reviewed and replenished using the well-known \((r,Q)\) policy. During stock-out periods, arriving customers are lost. The resulting complex problem is modeled as a fractional programming problem. The structural properties are explored for the model, and based on them two solution algorithms are proposed. Solving the problem...
subject to fill-rate and service-reliability constraints, the impact of the integrated pricing and inventory decisions on the profitability and performance of the described system is investigated. We observe that at a high fill rate, the optimal order quantity can be increasing in the reorder point. Moreover, for an active fill-rate constraint, reducing the mean replenishment lead time results in more customers lost on average. Although the model studied in this paper is a simplified version of real world cases, it gives useful managerial insights about more complex systems and can be used as a starting point for more complicated settings.

The study can be extended by examining other queueing systems and other distributions for inventory replenishment lead time. One may also consider multiple items meant for final products or for serving a customer. Additionally, substituting the lost-sale assumption with backordering could be an interesting challenge. Because there is no closed-form expression for the system performance in the backordering case, a deep simulation-based numerical study should be conducted to reveal the impact of the pricing decision on such systems. Other operational decisions, like capacity level setting, can also be integrated in the current model. Moreover, it might be even more realistic to consider both price- and time-sensitive customers, in which case the market can also be segmented based on the customers' sensitivity to price and quoted lead time. However, all these variations augment the complexity and may lead to analytical intractability.
Appendix

2.A Proofs

It is worth mentioning that since it is pointless to study pricing decisions of systems with no positive profit in practice, the following assumption is made:

Assumption 2.1. There is at least one feasible point that satisfies $\Pi > 0$.

Because of the continuity of $\Pi$, it would be an interval rather than a single point.

Proof of Proposition 2.1. Since $\Pi^c$ is strictly concave in $Q$, i.e., $\frac{\partial^2 \Pi^c}{\partial Q^2} = -\frac{H}{\lambda} < 0$, and $E(\tau)$ is affine with respect to $Q$, condition K in Schaible (1983) is met, and therefore, the problem is a concave fractional programming problem. According to Proposition 2 in Schaible (1983), $\Pi$ is strictly pseudo-concave. Furthermore, based on Proposition 3 in Schaible (1983) and the fact that the problem has a solution, a point satisfying KKT conditions that yields the best objective value is the unique optimal $Q$. Because of linearity of the constraints, this point is either an extreme point or a point satisfying the first order condition with respect to $Q$. □

Proof of Proposition 2.2. Due to Assumption 2.1 and the strict positivity of $E(\tau)$, there is an area within which $\Pi^c > 0$. From

$$\frac{\partial^2 \Pi^c}{\partial p^2} = -\frac{HQ\beta^2}{\lambda^3} \left( Q + 1 + 2\lambda \left( 1 - g^r \right) + \frac{\partial \tau (r + 1) g^r}{\partial \lambda} \right) - \frac{S\partial \tau (r + 1) g^r}{\lambda^3} - \frac{2CQ\beta^2}{\sigma (\mu - \lambda)^3} - \frac{2C\mu g^r}{\sigma (\mu - \lambda)^3} < 0,$$

where $g = \frac{\lambda}{\lambda + \mu}$, we see that $\Pi^c$ is strictly concave in $p$. Thus, there must be at most two points, i.e., $p_1$ and $p_2$, for which it is the case that $\Pi^c > 0$ for $p > p_1$, or $p < p_2$, or $p_1 < p < p_2$. $p_1$ and $p_2$ can be positive and negative infinities, respectively. Let us define a new feasible area for $p$ in our problem:

$$\max \{ p, p_1 \} = p' \leq p \leq \overline{p} = \min \{ \overline{p}, p_2 \}. \quad (2.13)$$
It is obvious that any \( p \in [p', \overline{p}] \) satisfies \([p, \overline{p}]\). In addition, from

\[
\frac{\partial^2 E(\tau)}{\partial p^2} = \frac{2\beta^2(Q - rg^{r+1})}{\lambda^3} + \frac{\partial r(r + 1)\beta^2 g^{r+2}}{\lambda^4} > 0,
\]

we see that \( E(\tau) \) is strictly convex with respect to \( p \). Due to the strict concavity and positivity of \( \Pi^c \) on \([p', \overline{p}]\), and the strict convexity and positivity of \( E(\tau) \), condition K in Schaible (1983) is satisfied, and the problem is a concave fractional programming problem in the newly defined area. According to Proposition 1 in Schaible (1983), \( \Pi \) is strictly pseudo-concave. Furthermore, based on Proposition 3 in Schaible (1983) and the fact that the problem has a solution, a point satisfying KKT conditions that yields the best objective value is the unique optimal \( p \). Because of the linearity of the constraints, this point is either an extreme point \( (p' \text{ or } \overline{p}) \) or a point satisfying the first order condition with respect to \( p \). If \([p', \overline{p}]\) is bounded by \( p_1 \) or \( p_2 \), then there is another point in this area for which \( \Pi(p) > \Pi(p_i), \ i = 1, 2 \). Thus, \( p_i, \ i = 1, 2 \), cannot be an optimal solution. Thereby, only when \([p', \overline{p}]\) is bounded by \( p \) or \( \overline{p} \), an optimal solution can happen at extreme points.

**Proof of Proposition 2.3.** Due to Assumption 2.1 and the strict positivity of \( E(\tau) \), there is an area within which \( \Pi^c > 0 \). From

\[
\frac{\partial^2 \Pi^c}{\partial r^2} = -\left(\frac{HQ}{L} + \frac{S\lambda}{L} + \frac{C\lambda^2}{\partial^2(\mu - \lambda)}\right)(\ln g)^2 g^r < 0,
\]

where \( g = \frac{\lambda}{\lambda + \vartheta} \), we see that \( \Pi^c \) is strictly concave in \( r \). Thus, there must be at most two points, i.e., \( r_1 \) and \( r_2 \), for which it is the case that \( \Pi^c > 0 \) for \( r > r_1 \), or \( r < r_2 \), or \( r_1 < r < r_2 \). \( r_1 \) and \( r_2 \) can be positive and negative infinities, respectively. Let us define a new feasible area for \( r \) in our problem:

\[
\max\{r, r_1\} = r' \leq r \leq \tau = \min\{r, r_2\}.
\]

(2.15)

It is obvious that any \( r \in [r', \tau'] \) satisfies \([r, \tau]\). In addition, from

\[
\frac{\partial^2 E(\tau)}{\partial r^2} = \frac{1}{\vartheta}(\ln g)^2 g^r > 0,
\]

we see that \( E(\tau) \) is strictly convex with respect to \( r \). Due to the strict concavity and positivity of \( \Pi^c \) on \([r', \tau']\), and the strict convexity and positivity of \( E(\tau) \), condition K in Schaible (1983) is satisfied and the problem is a concave fractional programming problem in the newly defined area. According to Proposition 1 in Schaible (1983), \( \Pi \) is strictly pseudo-concave. Furthermore, based on Proposition 3 in Schaible (1983) and the fact that the problem has a solution, a point satisfying KKT conditions that yields the best
objective value is the unique optimal \( r \). Because of the linearity of the constraints, this point is either an extreme point (\( r' \) or \( r'' \)) or a point satisfying the first order condition with respect to \( r \). If \([r', r'']\) is bounded by \( r_1 \) or \( r_2 \), then there is another point in this area for which \( \Pi(r') > \Pi(r_i), \ i = 1, 2 \). Thus, \( r_i, i = 1, 2 \), cannot be an optimal solution. Thereby, only when \([r', r'']\) is bounded by \( r' \) or \( r'' \), an optimal solution can happen at extreme points. 

**Proof of Lemma 2.1.** By definition we have \( f_s(Q, p) = \Pi - \gamma E(\tau) = (\Pi - \gamma)E(\tau) > 0 \). Due to the positivity of the expected cycle length \( E(\tau), \Pi > \gamma \) and then \( \gamma' > \gamma \).

**Proof of Proposition 2.4.** This proof follows the proof of Proposition 1 in (Chao and Zhou, 2006). Since \( \gamma_n \) is an increasing sequence and the problem is bounded, it converges to a number like \( \gamma \). We need to show that it converges to the optimal value of the problem, or equivalently to \( \gamma^* \). If this is not true, then \( F(\gamma) > 0 \). From Lemma 1, it follows that \( \gamma' = \Pi(Q_n(\gamma), p_n(\gamma)) > \gamma \). Since \( \Pi(Q_n(\gamma), p_n(\gamma)) \) is a continuous function of \( \gamma \), for any \( \epsilon > 0 \) such that \( \gamma' - \epsilon > \gamma \), there exists a positive integer \( N \) such that when \( n > N \), it is the case that

\[
\gamma_{n+1} = \Pi(Q_n(\gamma), p_n(\gamma)) \geq \Pi(Q_n(\gamma), p_n(\gamma)) - \epsilon = \gamma' - \epsilon > \gamma. \tag{2.17}
\]

This apparently contradicts \( \gamma_{n+1} \leq \gamma \). Therefore, as \( n \to \infty \) it is the case that \( \gamma_n \to \gamma^* \) and \( f(\gamma_n) \to 0 \).

## 2.B Extensions

Consider the fill-rate and service-reliability constraints introduced in Section 2.5. We have \( Q_f(p) = \frac{f}{(1-f)\theta}(\frac{\lambda}{\theta + \theta})^r \) for the fill-rate constraint, and \( Q_s(p) = \frac{k\lambda}{(\mu - \lambda - k)\theta}(\frac{\lambda}{\theta + \theta})^r \) for the service-reliability constraint. Both \( Q_f \) and \( Q_s \) are decreasing in \( p \), i.e., \( \frac{\partial Q_f}{\partial p} \leq 0 \) and \( \frac{\partial Q_s}{\partial p} \leq 0 \). From \( Q_f = Q_s \), we have \( \mu - \lambda = \frac{k}{f} \) by which the intersection point can be found as \((\hat{Q}, \hat{p})\) where \( \hat{Q} = \frac{f(\mu - k/f)}{(1-f)\theta}(\frac{\mu - k/f}{\mu - k/f + \theta})^r \) and \( \hat{p} = \frac{\theta - k/f}{\theta} \). Now, consider \( \hat{p} < \hat{p} \). It is clear that at \( \hat{p}, \mu - \lambda < \frac{k}{f} \). Furthermore, to have a real positive \( Q_s, k < \mu - \lambda \) should hold. At \( \hat{p}, \mu - \lambda = \frac{\theta k}{f} \) is clearly the case for some \( \theta \in (f, 1) \), which satisfies \( k < \mu - \lambda < \frac{k}{f} \). By substituting \( \mu - \lambda \) with \( \frac{\theta k}{f} \), we get

\[
Q_s(\hat{p}) = \frac{k\lambda}{(\theta k/f - k)\theta}(\frac{\lambda}{\theta + \theta})^r = \frac{f\lambda}{(\theta - f)\theta}(\frac{\lambda}{\theta + \theta})^r > Q_f(\hat{p}).
\]

Knowing that both \( Q_f(p) \) and \( Q_s(p) \) are decreasing, we can conclude that for a \( p < \hat{p} \) \((p > \hat{p})\), it is the case that \( Q_s > Q_f ) \((Q_s < Q_f)\). Figure 2.B.1 schematically depicts the two constraints on \( p-Q \) plane. Any point below each curve violates the corresponding constraint.
Taking all the above-mentioned discussions into consideration, the $p$-$Q$ plane can be partitioned into four regions as shown in Figure 2.B.1: Any solution in the region (1) satisfies both constraints; (2) violates both constraints; (3) satisfies the fill-rate constraint, but violates the service-reliability constraint; (4) satisfies the service-reliability constraint, but violates the fill-rate constraint. Restricting ourselves to the feasible range of $p$, we can conclude that, when solving the problem subject to both constraints, on interval $[p, \max\{p, \min\{p, \hat{p}\}\}]$ the fill-rate constraint is redundant, and on interval $[\min\{p, \max\{p, \hat{p}\}\}, \hat{p}]$ the service-reliability constraint is redundant. Note that these intervals are determined by known values.
CHAPTER 3

Pricing and Inventory Control in Service-Inventory Systems with Boundedly Rational Customers

**History:** This chapter is based on a paper submitted to the journal of *Computers & Industrial Engineering*. An earlier version of this chapter is presented at the *GOR 2017 Conference*, September 2017, Berlin, Germany.
Pricing and Inventory Control in Service-Inventory Systems with Boundedly Rational Customers

Ata Jalili Marand

Cluster for Operations Research, Analytics, and Logistics, Department of Economics and Business Economics, Aarhus University, Denmark

Abstract

A service-inventory system can be defined as a service system with an attached inventory where both an on-hand inventory item and a positive processing time are required to satisfy a customer’s demand. This study addresses the integration of pricing and inventory control decisions in a service-inventory system with boundedly rational customers. We consider a service system, modeled as an $M/M/1$ queue, with an attached inventory that is continuously reviewed under the $(r, Q)$ policy. The inventory replenishment lead time is assumed to be exponentially distributed. Moreover, it is assumed that the customers arriving during stock-out periods are lost. Upon their arrivals, the price- and time-sensitive customers make a choice between buying the service or balking. Under both rationality and bounded rationality assumptions, we model and analyze the customers’ choice behavior and a profit-maximizing service provider’s optimal pricing, reorder point, and order quantity decisions. Our results suggest that the integration of the pricing and inventory control decisions considerably contributes to the service provider’s profitability, particularly, when customers are highly time sensitive or when the expected replenishment lead time is long. Moreover, we show that the customers’ bounded rationality level significantly influences the service provider’s optimal pricing and inventory control decisions. The impact of the inventory control costs on the service provider’s profitability is studied by numerical examples, and the main managerial insights are reported.

Keywords: Service-inventory system; Operations-marketing interface; Inventory control; Pricing; Bounded rationality.
3.1 Introduction

A service system with an attached inventory, or a so-called service-inventory system, differs from both the classical service systems (see, e.g., Hassin and Haviv, 2003; Hassin, 2016) and the classical inventory systems (see, e.g., Bijvank and Vis, 2011; Glock et al., 2014; Andriolo et al., 2014). In a service-inventory system, an inventory item needs to be processed according to each customer’s preferences before it can be delivered to the customer. Given a limited processing capacity, the processing time leads to a queue of customer demands. In the classical inventory systems, however, a customer’s demand would be immediately satisfied if there is enough on-hand inventory. The inventory item does not require further processing or, if it does, it is assumed that the processing time is zero, which may indicate an unlimited processing capacity. In the classical service systems, it is assumed that no physical item is required for satisfying the customers or, if an item is needed, the impact of the inventory on the system’s operations is negligible. However, in a service-inventory system, both an on-hand inventory item and a positive processing time are required to fulfill a customer’s demand. Service-inventory systems can capture the properties of the well-known make-to-order systems.

Examples of companies operating under make-to-order systems are abundant. For instance, consider Dell computers. Dell’s traditional direct-to-customer channel allows the customers to configure products which are subsequently delivered to their doorsteps by Dell. Although Dell transformed its business model in 2017, about 35% of Dell’s units are still sold through the direct channel (Martin et al., 2014). Due to the high demand uncertainty, instead of stocking the finished goods, Dell keeps inventories of semi-finished parts and main product components, and customizes them based on customers’ requirements (Simchi-Levi et al., 2013). As another example, IBM, one of the largest providers of server computers, stocks the components with long manufacturing lead times and configures these components in response to the customer demands (Cheng et al., 2002).

In this study, we consider a service provider serving price- and time-sensitive customers. The service provider’s operations are modeled as a service-inventory system. It is assumed that the customers’ inter-arrival times, processing times, and inventory replenishment lead times are independently and exponentially distributed. Moreover, the inventory is assumed to be continuously reviewed under the \((r,Q)\) policy. Customers arriving during stock-out periods are assumed to be lost. We model and analyze the customers’ choice behavior in response to the service provider’s decisions under two assumptions: the customers are (i) rational or (ii) boundedly rational. The objective is to integrate the service provider’s pricing and inventory control decisions and to maximize
its profit.

We make the following contributions. First, to the best of our knowledge, the choice behavior of price- and time-sensitive customers in a service-inventory system has not been investigated in an integrated framework. We propose an integrated framework for studying the service provider’s optimal pricing and inventory decisions while taking the customers’ price and time sensitivities into account. Second, we explicitly model the customers’ bounded rationality in order to account for their limited ability to assess their expected utility. Furthermore, we employ a parametric fractional programming method to solve the problem and explore structural properties for the model to facilitate the solution procedure. At last, we investigate the impact of the integrated decision making, the bounded rationality level, and the inventory control costs on the customers’ choice behavior and the service provider’s optimal decisions and profitability. The numerical results show that integrating the pricing and inventory control decisions significantly improves to the service provider’s profitability. We also observe that the service provider’s profit and both pricing and inventory control decisions are very sensitive to the variations in the bounded rationality level.

The remainder of the paper is organized as follows. Section 3.2 reviews the most related literature. Section 3.3 describes the problem under consideration. In Section 3.4, we mathematically formulate the problem. In Subsection 3.4.1, the performance measures of the system are presented. Then, in Subsections 3.4.2 and 3.4.3, we investigate the equilibrium behavior of the customers under rationality and bounded rationality assumptions, respectively. The service provider’s optimization problem and the solution method are discussed in Subsections 3.4.4 and 3.4.5. Section 3.5 is devoted to numerical results and main managerial insights. We conclude the paper in Section 3.6.

3.2 Literature Review

This study is closely related to four streams of literature: pricing in service systems, pricing in inventory systems, service-inventory systems, and bounded rationality in service systems. In this section, we review the literature in each stream.

3.2.1 Pricing in Service Systems

The research in this stream mainly focuses on the trade-off between price and delivery time (i.e., expected waiting time). In fact, it is the congestion that makes pricing so important for service systems. When customers are price and time sensitive, the service provider can reduce the arrival rate in a congested system by increasing the price and in
this way make it possible for itself to serve the customers in a shorter time. But a shorter waiting time, in return, attracts more customers and leads to a higher congestion level. This makes the service less attractive for customers.

The use of price as a tool for controlling the congestion level in service systems dates back to seminal works by Naor (1969) and Edelson and Hilderbrand (1975) for observable and unobservable systems, respectively. Both Naor (1969) and Edelson and Hilderbrand (1975) assume that customers are price and time sensitive and rational. They construct a reward-cost function, or a so-called utility function, to capture individual customers’ choice behavior. These canonical studies have been a source of inspiration for most of the more recent research. See Hassin and Haviv (2003) and Hassin (2016) for two comprehensive reviews on the pricing in service systems.

There is another body of literature in this stream that focuses on the pricing, lead-time quotation, and capacity level decisions in service and make-to-order systems. These studies employ a market-level aggregate demand function rather than a utility function for analyzing individual customers’ equilibrium behavior. In an early study in this stream, Palaka et al. (1998) consider a make-to-order system modeled as an $M/M/1$ queue and aim at maximizing the profit through optimal pricing, lead-time quotation, and capacity expansion decisions. The literature has developed in different directions later on. For instance, it has been extended to competition models (e.g., So, 2000; Liu et al., 2007; Pekgün et al., 2008; Jayaswal and Jewkes, 2016), models with segmented markets (e.g., Boyaci and Ray, 2003, 2006; Teimoury et al., 2011; Jayaswal et al., 2011; Zhao et al., 2012; Jayaswal and Jewkes, 2016), models considering decentralized supply chains (e.g., Liu et al., 2007; Pekgün et al., 2008), and models with a lead-time dependent price (e.g., Ray and Jewkes, 2004).

In spite of their insightful results, the studies mentioned above all assume that the impact of inventory control decisions and parameters on the system is negligible. We contribute to this stream of literature by considering a service system with an attached inventory and by integrating the pricing and inventory decisions in such a system.

### 3.2.2 Pricing in Inventory Systems

Whitin (1955) is the first to incorporate the pricing decision to the inventory problems. In a newsvendor setting, he aims at simultaneously determining the optimal selling price and order quantity. Later on, the literature has branched out in different directions based on the assumptions made on, e.g., the time horizon (single/multiple, finite/infinite), pricing decision (static/dynamic), demand type (deterministic/stochastic), restocking (every period/once), and number of products (single/multiple). For instance, Ray et al. (2010) consider two periodic review inventory models with either a time-independent or time-
dependent backordering cost. They assume that the demand is price dependent and stochastic. They show that the model with time-dependent backordering cost generally results in longer review periods and lower retail prices. Considering a firm that sells a single product over a horizon of \( n \) periods, Lu et al. (2014) study the optimal pricing and replenishment policy when the firm employs a dynamic dual pricing strategy, i.e., unit selling price and quantity discounts. Assuming a stochastic price-dependent demand, Chen and Simchi-Levi (2004) analyze a finite horizon, single product, periodic review model to simultaneously find an inventory policy and a pricing strategy that maximize the expected profit over a finite horizon. Refer to Chan et al. (2004) for a comprehensive review.

These studies primarily focus on dynamic pricing and lot sizing. Moreover, to the best of our knowledge, integration of pricing, reorder point, and order quantity decisions has not been considered previously in service-inventory systems with stochastic inter-arrival times, stochastic processing times, and stochastic inventory replenishment lead times when the customers are price and time sensitive. We aim at filling this gap and contributing to this stream of literature.

### 3.2.3 Service-Inventory Systems

Sigman and Simchi-Levi (1992) are the first to study a service-inventory system. Their work is followed by a series of publications by Berman (e.g., Berman et al., 1993; Berman and Kim, 1999; Berman and Sapna, 2000; Berman and Kim, 2001; Berman and Sapna, 2001; Berman and Kim, 2004), and later by Schwarz et al. (2006), Schwarz and Daduna (2006), Saffari et al. (2011), Saffari et al. (2013), Zhao and Lian (2011), etc. The literature in this stream can be categorized by the assumptions made about the customers’ inter-arrival time distribution, service time distribution, replenishment lead time distribution, waiting hall capacity, and by how demands during stock-out periods are treated. See Krishnamoorthy et al. (2011) for a comprehensive review on service-inventory systems, and Table 1 in Jalili Marand et al. (2017a) for a categorized literature review. Studies in this stream mainly focus on deriving performance measures of service-inventory systems under different assumptions, and they rarely consider optimization problems with economic considerations in such systems.

The study by Schwarz et al. (2006) is the most related work to ours in this stream. They consider an inventory system with a Poisson customer arrival process, and independently and exponentially distributed processing times and inventory replenishment lead times. They assume that the inventory is continuously reviewed under the \((r,Q)\) policy and derive closed-form solutions to the joint distribution of the number of customers in the system and the inventory level. They state the optimization problem for finding the
order quantity and reorder point, but they leave the structural properties of the resulting problem for future research. We consider a similar system and extend Schwarz et al. (2006) by integrating the pricing and inventory control decisions when customers are price and time sensitive and boundedly rational.

There are two studies that address the pricing decision in service-inventory systems: Jalili Marand et al. (2017a) and Wang and Zhang (2017). Jalili Marand et al. (2017a) consider a system similar to Schwarz et al. (2006) and aim at maximizing the system-wide profit by making pricing and order quantity (reorder point) decisions for a given reorder point (order quantity) subject to a service-reliability constraint and a fill-rate constraint. They employ an aggregate demand function that is linearly dependent on the service price, but not on the service time. Although they allow their objective function to include a waiting cost, they assume that the customers are not time sensitive. They explore the structural properties of the problem and employ a fractional programming technique to solve the resulting non-linear model to optimality. Our work differs from their study in three respects: (i) instead of using an aggregate demand function, we derive the demand from the individual customers’ choice behavior, (ii) we assume that the customers are both price and time sensitive, and (iii) we simultaneously decide on price, reorder point, and order quantity.

Wang and Zhang (2017) also consider a system similar to that in Schwarz et al. (2006). The problem is considered under two different information level assumptions, i.e., a fully unobservable system and a partially observable system. They investigate the rational customers’ individually and socially optimal strategies, and further consider the optimal pricing issue that maximizes the service provider’s profit. However, they assume that the order quantity and reorder point are exogenously set and, for simplification, they force the reorder point to be equal to zero. They conclude that withholding some system information from customers may be beneficial for the profit-maximizing service provider. Our work differs from the study by Wang and Zhang (2017), since (i) we determine the inventory decisions, i.e., the reorder point and order quantity, endogenously, and (ii) we also study the impact of the customers’ bounded rationality on the service provider’s optimal decisions.

### 3.2.4 Bounded Rationality in Service Systems

The bounded rationality assumption is first incorporated to the classical service pricing literature by Huang et al. (2013b). They assume that the customers are not capable of accurately estimating their expected waiting time and employ a multinomial logit (MNL) model to capture the customers’ choice behavior. They investigate the problem from both a social planner’s and also a profit-maximizing service provider’s view points and for both
observable and unobservable systems. For unobservable systems, they conclude that the profit-maximizing service provider may benefit from the customers’ bounded rationality if the bounded rationality level is sufficiently high. Moreover, they show that ignoring the customers’ bounded rationality can lead to significant profit losses. Huang et al. (2013b)’s work is extended to the context of customer-intensive services by Li et al. (2016b) and Li et al. (2017). See Ren and Huang (2017) for recent developments in this area. To the best of our knowledge, the impact of price- and time-sensitive customers’ bounded rationality on a profit-maximizing service provider’s optimal decisions has not been previously addressed in the context of service-inventory systems. We contribute to this stream of research by filling this gap.

3.3 Problem Description

Consider a service-inventory system serving a homogeneous population of customers. Customers arrive at the service-inventory system according to a Poisson process with rate $\Lambda$ (potential arrival rate). Customers are served one by one and according the the first-in-first-out rule. Serving each customer requires an exponentially-distributed service time with parameter $\mu$ and one item from the inventory. The system therefore needs to keep an inventory of the required item. It is assumed that the inventory is continuously reviewed under the $(r,Q)$ policy, i.e., when the inventory level reaches $r$, an order of size $Q$ is made to replenish the inventory. We assume $0 \leq r < Q$ to avoid excessive reorders by not allowing more than one outstanding replenishment order at a time. The system incurs an inventory holding cost, $h$, per unit of inventory per time unit and a fixed inventory ordering cost, $a$, each time it makes an order.

In order to take the supply side uncertainties into account, we assume that the replenishment lead times are exponentially distributed with parameter $\eta$. When the replenishment lead times are positive, stock outs may occur. We assume that customers arriving during stock-out periods are rejected (the lost-sales assumption). This procedure lets the service provider avoid long queues when there is no item available and enables him to deliver a superior time performance. During the stock-out periods, the service provider is idle. When a replenishment order arrives, however, the service providers starts to serve the customers again.

It is assumed that the inventory control policy, order quantity, reorder point, replenishment lead time distribution, service time distribution, and the potential arrival rate are common knowledge. Upon their arrivals, customers decide either to buy the service or to balk. Customers make their decisions based on a utility function that is dependent on the service price and the expected waiting cost of each customer. If a customer decides
to buy the service and the inventory is not empty, the customer will be admitted to the system; otherwise, if there is no inventory on-hand, the customer will be rejected. We assume that the system incurs a loss of goodwill cost for each customer rejected, $l$. Moreover, it is assumed that the customers who have already been admitted are patient and will wait for their turns. Figure 3.1 schematically depicts the system under consideration.

![Figure 3.1: System view: (a) the server is busy, (b) the server is idle due to a stock out, (c) the server is idle due to an empty line.](image)

**3.4 Model Formulation**

The performance measures of the system described in this section are derived by Schwarz et al. (2006). It should be mentioned that while their focus is on deriving the performance measures of such a system, our focus is on analyzing the customers’ equilibrium behavior and the service provider’s optimization problem. We adopt their results to conduct our study.

**3.4.1 Performance Measures of the System**

Let $\lambda$ denote the equilibrium arrival rate, i.e., the arrival rate of the customers who decide to buy the service. According to Schwarz et al. (2006), the system is ergodic if and only if $\lambda < \mu$. Moreover, the joint distribution of the number of customers in the system and the inventory level is of product form as $P(i, j) = P_n(i)P_k(j)$, where $P_n(i)$ is the distribution of the number of customers, i.e., $n$, in the system, and $P_k(j)$ is the distribution of the inventory level, i.e., $k$. In other words, $P_n(i) = Pr(n = i)$ and $P_k(j) = Pr(k = j)$. From Schwarz et al. (2006), we have $P_n(i) = (1 - \lambda/\mu)(\lambda/\mu)^i$, and $P_k(j) = E(t_j)/E(\tau)$, where
$E(t_j)$ is the expected time that the inventory level equals $j$ during each cycle, and

$$E(\tau) = \frac{Q}{\lambda} + \frac{1}{\eta} \left( \frac{\lambda}{\lambda + \eta} \right)$$  \hspace{1cm} (3.1)$$

is the expected cycle length. Each cycle starts when the inventory level reaches $r$ and ends when the next cycle starts. In addition, the average inventory level equals

$$\sum_{k=0}^{Q+r} k Pr_k(k) = \frac{Q^2+(2r+1)Q}{2\lambda} + \frac{Q}{\eta} \left( \frac{\lambda}{\lambda + \eta} \right) - 1. \hspace{1cm} (3.2)$$

Due to stock outs, some customers may be rejected. $P_k(0)$ denotes the probability that a customer finds the inventory out of stock. According to Schwarz et al. (2006), we have

$$P_k(0) = \frac{1}{\eta} \left( \frac{\lambda}{\lambda + \eta} \right) \frac{Q}{\lambda + 1} \frac{\lambda}{\lambda + \eta}. \hspace{1cm} (3.3)$$

We define $\lambda_{eff} = \lambda(1 - P_k(0))$ as the effective arrival rate, i.e., the arrival rate of the customers that are admitted to the system (See Figure 3.2). During stock-out periods, no service is conducted. Thus, the effective service rate can also be defined as $\mu_{eff} = \mu(1 - P_k(0))$. Furthermore, the expected (steady-state) waiting time in the system equals

$$w(\lambda, r, Q) = \frac{1}{\mu_{eff} - \lambda_{eff}} = \frac{1}{\mu - \lambda} (1 + \frac{\lambda}{\eta Q} (\frac{\lambda}{\lambda + \eta})^r). \hspace{1cm} (3.4)$$

We refer the reader to Schwarz et al. (2006) for detailed discussions. In the following subsections, we analyze the customers’ choice behavior and the service provider’s optimization problem under the rationality and bounded rationality assumptions.

---

**Figure 3.2:** The potential arrival rate, equilibrium arrival rate, and effective arrival rate in the described service-inventory system
3.4.2 Rational Customers’ Choice Behavior

In this subsection, we analyze the customers’ choice behavior for a given price, \( p \), order quantity, \( Q \), and reorder point, \( r \), assuming that the customers are fully rational. Following the service pricing literature, we assume that each customer’s utility from buying the service is linearly decreasing in the price and the expected waiting cost as

\[
U(\lambda, p, r, Q) = v - p - c_w w(\lambda, r, Q),
\]

(3.5)

where \( v \) is the service value, i.e., the reward that a customer receives after the service completion, \( c_w \) is the waiting cost rate of each customer, and \( w(\lambda, r, Q) \) is defined by Eq. (3.4). \( c_w \) is a measure of customers’ time sensitivity. Without loss of generality, we normalize the utility of a customer that is not admitted to the system to zero. The inventory policy, order quantity, reorder point, replenishment lead time distribution, service time distribution, and the potential arrival rate are common knowledge. Customers are assumed to be rational such that they decide to buy the service if their actual (expected) utility is non-negative, i.e., \( U(\lambda, p, r, Q) \geq 0 \). The rationality assumption also indicates that the customers are capable of accurately computing the expected waiting time in the system.

Since the choice of each customer affects the waiting time experienced by other customers, a game theoretic approach is commonly used to analyze this situation (Armony and Haviv, 2003). In this stage, we assume that the price and inventory control decisions are given, and the competition is among the customers to maximize their utilities defined by Eq. (3.5). Let \( q(p, r, Q) \) denote each customer’s equilibrium probability of buying the service and \( \lambda = q\Lambda \) denote the corresponding equilibrium arrival rate. According to Chen and Frank (2004), three cases may occur:

(i) If \( U(0, p, r, Q) \leq 0 \), i.e., \( p \geq v - c_w w(0, r, Q) \), then no customer will get a non-negative utility even if no other customer buys the service. In this case, buying with probability \( q(p, r, Q) = 0 \) (balking with probability one) is the dominant equilibrium strategy, and the equilibrium arrival rate equals 0.

(ii) If \( U(\Lambda, p, r, Q) \geq 0 \), i.e., \( p \leq v - c_w w(\Lambda, r, Q) \), then all of the customers will gain a non-negative utility even if they all buy the service. In this case, buying with probability \( q(p, r, Q) = 1 \) is the dominant equilibrium strategy, and the equilibrium arrival rate equals \( \Lambda \).

(iii) If \( U(\Lambda, p, r, Q) < 0 < U(0, p, r, Q) \), i.e., \( v - c_w w(\Lambda, r, Q) < p < v - c_w w(0, r, Q) \), then each customer plays a mixed strategy at equilibrium such that \( q(p, r, Q) = \ldots \)
\[ \lambda \in (0, 1), \text{ where } \lambda \text{ solves } U(\lambda, p, r, Q) = 0 \text{ that results in } \]

\[ p(\lambda, r, Q) = v - c_w w(\lambda, r, Q), \tag{3.6} \]

where \( w(\lambda, r, Q) \) is defined by Eq. (3.4).

Note that it is assumed that \( v - c_w w(0, r, Q) = v - \frac{c_w}{\mu} > 0 \) to rule out the trivial case where no customer decides to buy the service. Moreover, we assume \( \Lambda \geq \mu \) to rule out the trivial case where all of the customers decide to buy the service. It is also noteworthy that some customers who decide to buy the service are not admitted. Customers find the system out of stock with probability \( P_k(0) \), and as mentioned before, the customers arriving during stock-out periods are rejected. Thus, a customer is served with probability \( q(p, r, Q)(1 - P_k(0)) \), and is not served with probability \( 1 - q(p, r, Q)(1 - P_k(0)) \). The effective arrival rate therefore equals

\[ \lambda_{eff}(\lambda, r, Q) = \Lambda q(p, r, Q)(1 - P_k(0)) = \lambda(1 - P_k(0)). \tag{3.7} \]

Figure 3.2 summarizes this part.

Based on the equation \( U(\lambda, p, r, Q) = 0 \), it is easy to show that the equilibrium arrival rate is decreasing in \( p \), and it is increasing in \( r, Q \), and \( \eta \). It is intuitive that the customers will be more attracted to a cheaper service. Moreover, increasing \( r, Q \), or \( \eta \) reduces the stock-out probability, which consequently decreases the customers’ expected waiting time. As a result, the customers incur a lower waiting cost rate and decide to buy the service with a higher probability.

### 3.4.3 Boundedly Rational Customers’ Choice Behavior

In this subsection, we assume that the customers are boundedly rational such that they are not capable of accurately estimating the expected waiting time. This assumption is more critical in the service-inventory system described in this study, since the expected waiting time follows a complex expression as seen in Eq. (3.4). Similar to the approach by Huang et al. (2013b), the noisy waiting cost estimation is characterized by a random error term such that

\[ U(\lambda, p, r, Q, \epsilon) = U(\lambda, p, r, Q) + \epsilon, \tag{3.8} \]

where \( U \) denotes each customer’s perceived utility from buying the service, and \( U \) is defined similar to Eq. (3.5).

It is assumed that \( \epsilon \) is a Gumbel-distributed random variable with mean \( E(\epsilon) = 0 \) and variance \( Var(\epsilon) = \frac{\theta^2}{6} \), where \( \theta > 0 \) is the scale parameter of the corresponding logistic
distribution (see, e.g., Talluri and Van Ryzin, 2006, for more detailed discussions). In fact, $\theta$ reflects the customers’ bounded rationality level. As the parameter $\theta$ approaches zero, the customers become fully rational and the model equals that of Subsection 3.4.2. Conversely, the customers become completely irrational as the parameter $\theta$ approaches positive infinity.

We use the MNL model to characterize the customers’ choice behavior. The MNL model is one of the most widely used choice models in revenue management (see, e.g., Strauss et al., 2018, for a recent review). With a slight abuse of notation, let $q(p, r, Q)$ denote the fraction of customers that decide to buy the service at a given price, $p$, order quantity, $Q$, and reorder point, $r$, and under the bounded rationality assumption. According to the MNL model, we have

$$q(p, r, Q) = \frac{e^{U(\lambda, p, r, Q)/\theta}}{e^{U(\lambda, p, r, Q)/\theta} + 1}, \quad (3.9)$$

where $\lambda$ denotes the corresponding equilibrium arrival rate such that $q(p, r, Q) = \frac{\lambda}{\Lambda} \in (0, 1)$.

Equating the right-hand side of Eq. (3.9) and $\frac{\lambda}{\Lambda}$ results in

$$U(\lambda, p, r, Q) = \theta \ln\left(\frac{\lambda}{\Lambda - \lambda}\right). \quad (3.10)$$

From Eq. (3.10), the customers get a positive (negative) utility from buying the service when the arrival rate is greater (less) than $\frac{\lambda}{\Lambda}$, and they get a zero utility when the arrival rate is equal to $\frac{\lambda}{2}$. Recall that the customers get a zero utility at equilibrium under the rationality assumption. Based on Eq. (3.10) and Eq. (3.5), we have

$$p(\lambda, r, Q) = v - c_w w(\lambda, r, Q) - \theta \ln\left(\frac{\lambda}{\Lambda - \lambda}\right), \quad (3.11)$$

where $w(\lambda, r, Q)$ is defined by Eq. (3.4).

Similar to the rational customers case, some of the customers that decide to buy the service are not admitted to the system. A customer arriving during a stock-out period is rejected. This happens with probability $P_k(0)$. Thus, a customer is served with probability $q(p, r, Q)(1 - P_k(0))$, and is not served with probability $1 - q(p, r, Q)(1 - P_k(0))$. Therefore, the effective arrival rate equals

$$\lambda_{eff}(\lambda, r, Q) = \Lambda q(p, r, Q)(1 - P_k(0)) = \lambda(1 - P_k(0)). \quad (3.12)$$
3.4.4 Service Provider’s Optimization Problem

In this subsection, we construct the service provider’s optimization problem. The service provider’s profit rate equals the revenue rate minus the inventory control cost rates and the loss of goodwill cost rate. The inventory control cost rates include the inventory holding cost rate and the (fixed) ordering cost rate. Let \( I(\lambda, r, Q) \) denote the expected inventory level and \( T(\lambda, r, Q) \) denote the expected cycle length as functions of \( \lambda \), \( r \), and \( Q \). The functions \( I(\lambda, r, Q) \) and \( T(\lambda, r, Q) \) are defined by Eq. (3.2) and Eq. (3.1), respectively. The service provider’s optimization problem can be modeled as

\[
\begin{align*}
\max_{(\lambda, r, Q)} \Pi(\lambda, r, Q) &= p(\lambda, r, Q)\lambda_{ef}(\lambda, r, Q) - hI(\lambda, r, Q) - \frac{a}{T(\lambda, r, Q)} - l(\lambda - \lambda_{ef}(\lambda, r, Q)) \\
\text{s.t.} \quad 0 &\leq r < Q \\
0 &\leq \lambda < \mu \\
0 &\leq p(\lambda, r, Q),
\end{align*}
\]  

(3.13)

(3.14)

(3.15)

(3.16)

where \( p(\lambda, r, Q) \) and \( \lambda_{ef}(\lambda, r, Q) \) are defined by Eq. (3.6) and Eq. (3.7) (Eq. (3.11) and Eq. (3.12)) under the rationality (bounded rationality) assumption, respectively. The objective function (3.13) consists of four terms: the revenue rate, inventory holding cost rate, ordering cost rate, and loss of goodwill cost rate. It can be rewritten as

\[
\Pi(\lambda, r, Q) = \frac{v\lambda - \frac{\eta\lambda}{\mu - \lambda}(1 + \frac{\lambda}{\eta Q}(\frac{\lambda}{\lambda + \eta})^r) - \theta\lambda\ln(\frac{\lambda}{\lambda + \eta}) - h(\frac{1}{2}(Q + 2r + 1) + \frac{\lambda}{\eta}\left((\frac{\lambda}{\lambda + \eta})^r - 1\right)) - \frac{a\lambda}{Q} - \frac{\lambda^2}{\eta Q}(\frac{\lambda}{\lambda + \eta})^r}{1 + \frac{\lambda}{\eta Q}(\frac{\lambda}{\lambda + \eta})^r},
\]

(3.17)

where \( \theta = 0 \) under the rationality assumption, and \( \theta > 0 \) under the bounded rationality assumption. Constraints (3.14) ensure that there is at most one outstanding replenishment order at a time and that the order quantity and reorder point are non-negative. Constraints (3.15) ensure the non-negativity of the equilibrium arrival rate and that the ergodicity condition is met. Constraint (3.16) ensures that the price is non-negative. It is noteworthy that although \( Q \) and \( r \) take integer values, in order to facilitate the analysis, we assume that they are continuous decision variables. Moreover, without loss of practicality, we assume that the service provider is profitable at optimality.
3.4.5 Structural Properties and Solution Method

To solve the service provider’s optimization problem, we convert it to a parametric fractional programming problem and solve a nonlinear equation in order to find the optimal solution to the original problem. Consider the following parametric fractional programming problem, for a $\gamma \in \mathbb{R}$,

$$
\max_{(\lambda, r, Q)} \Pi^\gamma(\lambda, r, Q) = \Pi^N - \gamma \Pi^D
$$

s.t. (3.14) - (3.16),

where $\Pi^N$ and $\Pi^D$ are the numerator and denominator of the objective function (3.17), respectively. Let $F(\gamma)$ be the optimal value of the objective function of the parametric fractional programming problem. It has been shown that an optimal solution to the original problem is also an optimal solution to the parametric fractional programming problem if $F(\gamma) = 0$ (see, e.g., Stancu-Minasian, 2012, for more details). In fact, a $\gamma$ that solves $F(\gamma) = 0$ is equal to the optimal value of the original problem. We employ a generalization of the Newton-Raphson method, i.e., the Dinkelbach method (Dinkelbach, 1967), to solve this nonlinear equation. In the following, we briefly describe the Dinkelbach method.

The Dinkelbach method starts with a $\gamma = \gamma_0$, which is a lower bound for the original problem. Recall that, we have assumed that the service provider is profitable at optimality. Thus, $\gamma_0 = 0$ can be a starting point. Note that $F(\gamma_0 = 0) > 0$. Based on the fact that $F(\gamma)$ is strictly decreasing in $\gamma$ (see, e.g., Lemma 3 in Dinkelbach, 1967), this method updates $\gamma$ in each step and generates an increasing sequence $\{\gamma_n\}$ leading to $F(\gamma) = 0$. To do so, the parametric fractional programming counterpart of the original problem should be solved for each $\gamma$. In each step, $\gamma$ is updated as follows. Suppose that solving the parametric fractional programming problem for $\gamma_n$ results in $(\lambda_n, r_n, Q_n)$. It has been shown that if $\Pi^\gamma_n(\lambda_n, r_n, Q_n) > 0$, then $\gamma_{n+1} = \Pi(\lambda_n, r_n, Q_n) > \gamma_n$ (see, e.g., Lemma 1 in Jalili Marand et al., 2017a). Obviously, $\gamma$ increases, and consequently $F(\gamma)$ decreases in each step. This iterative procedure continues until $\gamma_n - \gamma_{n-1} < \epsilon$, where $\epsilon$ controls the accuracy of the solution.

Now, consider Proposition 3.1 that explores the structural properties of $\Pi^\gamma(\lambda, r, Q)$.

**Proposition 3.1.** Under the rationality as well as under the bounded rationality assumptions and for any $\gamma \geq 0$, (i) $\Pi^\gamma(\lambda, r, Q)$ is concave with respect to $\lambda$ for a given $r$ and $Q$, and (ii) $\Pi^\gamma(\lambda, r, Q)$ is jointly concave with respect to $r$ and $Q$ for a given $\lambda$.

According to Proposition 3.1 (i), there is a unique arrival rate that maximizes $\Pi^\gamma(\lambda, r, Q)$ given the inventory control decisions. Moreover, for a given arrival rate, there exists a
unique optimizing pair of the reorder point and order quantity, and this pair can be found efficiently using numerical methods. We exploit part (ii) of Proposition 3.1 to solve the problem. For each given $\gamma$, we discretize the $\lambda$-line, and for each $\lambda$ we determine $r$ and $Q$ using Matlab’s fmincon function. The $\lambda$ yielding the highest objective value, i.e., $\Pi^\gamma(\lambda, r, Q)$, and the corresponding $r$ and $Q$ are chosen as the optimal solutions to the parametric fractional programming problem. Then we update $\gamma$ and continue this procedure until we meet the termination criterion. Algorithm 3.1 summarizes the solution method.

**Algorithm 3.1** Pseudo-code of the Dinkelbach method

| Appropriately set $\epsilon$ and $\Delta \gamma$; |
| Set $\gamma_0 = 0$; |
| Set $i = 1$; |
| while $\Delta \gamma > \epsilon$ do |
| For $\gamma_i$, find the optimal solution $(\lambda_i, r_i, Q_i)$ maximizing $\Pi^{\gamma_i}(\lambda, r, Q)$; |
| if $\Pi^{\gamma_i}(\lambda_i, r_i, Q_i) = 0$ then |
| Terminate the while cycle. |
| else if $\Pi^{\gamma_i}(\lambda_i, r_i, Q_i) > 0$ then |
| $\gamma_{i+1} = \Pi(\lambda_i, r_i, Q_i)$; |
| $i = i + 1$; |
| end if |
| Set $\Delta \gamma = \gamma_i - \gamma_{i-1}$; |
| end while |
| Set $(\lambda^*, r^*, Q^*) = (\lambda_i, r_i, Q_i)$. |

**Lemma 3.1.** Under the rationality assumption, $\lambda < \mu - \sqrt{cw/v}$ at optimality.

The right-hand side of the inequality in Lemma 3.1, i.e., $\mu - \sqrt{cw/v}$, is the revenue-maximizing equilibrium arrival rate under the rationality assumption for a pure service system (see Hassin and Haviv, 2003, Chapter 3). Apart from restricting the $\lambda$-line and improving the computational time, Lemma 3.1 implies that the described service-inventory system has a lower equilibrium arrival rate compared to a service system without inventory under the rationality assumption.

### 3.5 Numerical Study and Managerial Insights

In this section, we numerically analyze the impact of (i) the integrated decision making, (ii) the bounded rationality assumption, and (iii) the inventory control costs, i.e., $h$ and $a$, on the customers’ equilibrium behavior and the service provider’s optimal decisions and profitability. We first compare our results with those of Wang and Zhang (2017) who consider a similar system operating under the $(0, Q)$ policy. The order quantity is assumed to be exogenous in their study. Our results reveal that the service provider’s
profitability is significantly improved when the reorder point and order quantity are included as endogenous decision variables and when pricing and inventory control decisions are integrated. Then, we investigate the impact of the customers’ bounded rationality level on the service provider’s profitability. Based on our results, the service provider may benefit from boundedly rational customers, and ignoring the customers’ bounded rationality may result in considerable profit losses. It is noteworthy that, although we try to consider a wide range of parameter values in our sensitivity analysis, the conclusions made in this section are based on a limited set of numerical instances, and due to the complexity of the problem structure and complicated interactions between problem parameters and decision variables, they may be subject to variation for other numerical settings.

3.5.1 The Inventory Control Decisions: Endogenous vs. Exogenous

In this subsection, we investigate the impact of the integrated decision making (endogenous inventory control variables) on the service provider’s profitability. To this end, we compare our results with a model similar to Wang and Zhang (2017) in which the inventory control variables are exogenous and are set as \((r, Q) = (0, 250)\). Figures 3.3a-3.3d show the variations of the profit loss, revenue rate, ordering cost rate, and inventory holding cost rate with respect to the replenishment lead time parameter, i.e., \(\eta\), for different waiting costs. Note that the (expected) replenishment lead time increases, as \(\eta\) decreases.

It is obvious that the service provider loses profit when the inventory control variables are exogenous compared to the case where they are endogenous. Figure 3.3a shows the magnitude of this profit loss. The profit losses in this figure are calculated as \(100 \left( \frac{\Pi_{Endo.} - \Pi_{Exo.}}{\Pi_{Exo.}} \right)\), where \(\Pi_{Endo.}\) is the profit with endogenous inventory control variables, and \(\Pi_{Exo.}\) is the profit with exogenous inventory control variables. Figure 3.3a shows that as \(\eta\) decreases, i.e., the replenishment lead time increases, the profit loss increases. For instance, when \(\eta = 0.05\), the profit loss equals 10.7, 11.5, and 13.3 percent for \(c_w = 0.8\), \(c_w = 1.4\), and \(c_w = 2.0\), respectively. As shown in this figure, the profit loss even exceeds 20 percent, which is significant, for longer replenishment lead times. This emphasizes the importance of an integrated decision making approach in the described service system. Figure 3.3a also shows that the profit loss is higher when the customers are more time sensitive, i.e., \(c_w\) is higher.

Figures 3.3b and 3.3c show that the revenue and ordering cost rates are increasing in \(\eta\), meaning that the revenue and ordering cost rates of the service provider decrease as the replenishment lead time increases. When the inventory control variables are exogenous,
the inventory holding cost rate also decreases as the replenishment lead time increases according to Figure 3.3d. The reason is that the (expected) cycle length is increasing in the replenishment lead time as one may notice from Figure 3.3c. In other words, the service provider keeps the same amount of items for a longer time period and therefore incurs a lower inventory holding cost rate. However, when the inventory control variables are endogenous, the inventory holding cost rate increases as the average replenishment lead time increases, implying that the order quantity is increasing in the replenishment lead time (see Figure 3.3d), and the proportional increase in the order quantity dwarfs that of the cycle length.

Figures 3.3e and 3.3f compares the optimal prices and expected waiting times for the exogenous and endogenous inventory control variables for different waiting costs. The results are calculated as $100\left(\frac{p_{Endo.} - p_{Exo.}}{p_{Exo.}}\right)$ and $100\left(\frac{w_{Endo.} - w_{Exo.}}{w_{Exo.}}\right)$. According to Figure 3.3f, when the inventory control variables are endogenous, the service provider can serve the customers in a shorter time, i.e., decrease their waiting cost, and extract the surplus by charging a higher price as shown in Figure 3.3e.

Figure 3.3: The impact of the integrated decision making on the service provider’s profit, revenue, inventory control cost rates, pricing decision, and the customers’ expected waiting time under the rationality assumption ($\Lambda = 2.5$, $\mu = 2.5$, $v = 5$, $h = 0.01$, $a = 100$, $l = 0.1$, $r = 0$, $Q = 250$)
3.5.2 The Impact of the Bounded Rationality

In this subsection, we investigate the impact of the customers' bounded rationality on the service provider’s optimal decisions and profitability. Figure 3.4a shows the variations of the service provider’s profit with respect to the customers’ bounded rationality level for different time sensitivities, i.e., different values of $c_w$. Recall that the customers are fully rational when $\theta = 0$, and as $\theta$ increases, the customers become less rational. According to Figure 3.4a, when the time sensitivity is sufficiently high ($c_w = 2.0$ in this numerical example), the service provider is always better off in a market where the customers are less rational. However, at lower time sensitivities, the service provider’s profit first decreases and then increases in the customers’ bounded rationality level (see the results for $c_w \leq 1.4$ in Figure 3.4a). This observation is consistent with Huang et al. (2013b) who show that the service provider is better off in markets where the bounded rationality level is sufficiently high.

Figures 3.4b-3.4d show the service provider’s optimal pricing, reorder point, and order quantity decisions, respectively. As the customers become less rational, the service provider increases its price when the customers’ time sensitivity is sufficiently high as shown in Figure 3.4b for $c_w = 2.0$. When customers are less time sensitive, however, the service provider’s optimal price is first decreasing and then increasing in the bounded}

Figure 3.4: The impact of the bounded rationality on the service provider’s optimal decisions and profitability ($\Lambda = 2.5$, $\mu = 2.5$, $\eta = 0.01$, $v = 5$, $h = 0.01$, $a = 100$, $l = 0.1$)
rationality level. In addition, while the optimal price is much affected by the time sensitivity at lower bounded rationality levels, it seems to be invariant to the time sensitivity at higher bounded rationality levels.

Figure 3.4c shows that the optimal reorder point is first decreasing and then increasing in the bounded rationality level. Moreover, Figure 3.4d demonstrates that the optimal order quantity decreases as the customers become less rational. Figures 3.4c and 3.4d also indicate that the variations of the service provider’s optimal reorder point and order quantity decisions with respect to the bounded rationality level can be significant. For instance, the optimal reorder point and order quantity equal \((r, Q) = (106, 243)\) when \(\theta = 0\) (customers are fully rational), \((r, Q) = (93, 221)\) when \(\theta = 0.5\), \((r, Q) = (79, 187)\) when \(\theta = 2\), and \((r, Q) = (82, 173)\) when \(\theta = 4\) for \(c_w = 2\). This emphasizes the fact that customers’ cognitive limitations and choice behavior significantly influence the service provider’s operational decisions.

### 3.5.3 The Impact of the Inventory Control Costs

In this subsection, we investigate the impact of the unit inventory holding cost rate, i.e., \(h\), and the fixed ordering cost, i.e., \(a\), on the service provider’s optimal decisions.

#### The Unit Inventory Holding Cost Rate

Regardless of the bounded rationality level, Figure 3.5a shows that the optimal profit is decreasing in the unit inventory holding cost rate, i.e., \(h\). Wang and Zhang (2017) report a similar behavior for the case with exogenous inventory variables and fully rational customers. Through a numerical example, Wang and Zhang (2017) also conclude that the optimal price is decreasing in the unit inventory holding cost rate. Our results, however, suggest that the optimal price can also be increasing in \(h\). According to Figure 3.5b, when customers are fully rational, i.e., \(\theta = 0\), the optimal price is decreasing in \(h\). But at a high bounded rationality level \((\theta = 1 \text{ or } 2 \text{ in this numerical example})\), the optimal price increases as \(h\) increases. For \(\theta = 0.1\), the optimal price is first increasing and then decreasing in \(h\). Figure 3.5b shows that the customers’ bounded rationality level significantly affects the service provider’s optimal pricing decisions in response to the variations of the unit inventory holding cost rate.

Moreover, as \(h\) increases, both the optimal reorder point and the optimal order quantity decrease as shown in Figures 3.5c and 3.5d. It is intuitive that the service provider decreases the (expected) inventory level in order to decrease the inventory holding cost rate as the unit inventory holding cost rate increases. Furthermore, Figures 3.5e and 3.5f reveal that the equilibrium arrival rate and the effective arrival rate also decrease as \(h\)
increases. A lower arrival rate with a lower variance explains the decrease in the optimal reorder point with respect to \( h \). Figures 3.5g-3.5i show the variations of the revenue rate, the ordering cost rate, and the inventory holding cost rate with respect to \( h \). According to these figures, as \( h \) increases, the revenue rate decreases, the fixed ordering cost rate increases, and the inventory holding cost rate first increases and then decreases.

The Fixed Ordering Cost

Figure 3.6a shows that the service provider’s profit decreases as the fixed ordering cost, i.e., \( a \), increases. Figures 3.6b-3.6d indicate that the revenue rate and the loss of goodwill cost rate vary only slightly with respect to \( a \), and the decreasing profit is caused by the other two cost terms: the fixed ordering cost rate and the inventory holding cost rate.
According to Figures 3.6h and 3.6i, the fixed ordering cost rate and inventory holding cost rate increase more significantly as $a$ increases.

Figures 3.6h and 3.6i also indicate that while the inventory holding cost rate is sensitive to the customers’ bounded rationality level, the ordering cost rate demonstrates less variations in this respect. A similar observation can be made from Figures 3.5h and 3.5i, where the inventory holding cost rate is more sensitive to the bounded rationality level compared to the ordering cost rate.

According to Figures 3.6d-3.6f, while the optimal price is rather sensitive to the variations of $a$, the variations of the optimal reorder point and order quantity are more significant. Figure 3.6e (Figure 3.6f) shows that the optimal reorder point (order quantity) is decreasing (increasing) in the fixed ordering cost. In other words, when the fixed ordering cost varies, the service provider tends to reach a new optimal solution by changing the
inventory control decisions rather than the pricing decision.

3.6 Conclusion and Future Research

In this study, we address the integration of a profit maximizing service provider’s pricing and inventory control decisions in a service-inventory system with boundedly rational customers. The system is modeled as an $M/M/1$ queue, and the inventory is assumed to be continuously reviewed under the $(r, Q)$ policy. We take the supply-side uncertainties into account by assuming that the inventory replenishment lead times are positive and exponentially distributed. Moreover, it is assumed that the customers arriving during stock-out periods are lost. We study the customers’ choice behavior and the service provider’s optimal decisions under the rationality assumption as well as under the bounded rationality assumption. Furthermore, we numerically investigate the impact of (i) the integrated decision making, (ii) the bounded rationality level, and (iii) the inventory control costs on the customers’ choice behavior and the service provider’s optimal decisions and profitability. We summarize the main findings below.

(i) Our results suggest that integrating the pricing and inventory control decisions significantly contributes to the service provider’s profitability. Particularly, the profit loss due to exogenous inventory control decisions is higher when the replenishment lead time is longer or when the customers are more time sensitive. Moreover, when the inventory control decisions are made endogenously, the service provider can serve the customers in a shorter time and charge them a premium price for the superior delivery performance.

(ii) We observe that the service provider is better off when the bounded rationality level is sufficiently high. The numerical analysis shows that both the pricing and the inventory control decisions are highly sensitive to the bounded rationality level. We also observe that, at high bounded rationality levels, the impact of the customer’s time-sensitivity level on the optimal pricing decision and the optimal (expected) waiting time diminishes.

(iii) According to our results, the optimal pricing decision is increasing (decreasing) in the unit inventory holding cost rate at sufficiently high (low) bounded rationality levels. We also observe that as the fixed ordering cost increases, the service provider tends to reach a new optimal solution by changing the inventory control decisions rather than the pricing decision.

The current study can be extended in different directions. For instance, one may
consider a more general queueing system, e.g., a $G/G/k$ queue. Examining other distributions for the replenishment lead time is also interesting. In this paper, we assume that the customers arrive one by one, and each customer requires one unit from the inventory. Although these assumptions enable us to study the integrated pricing and inventory control of service-inventory systems as described in this paper, it is more realistic to consider the cases where the customers arrive in groups and serving them requires more than one item. A competition problem can also be define to study the interaction between multiple service providers in a competitive market. Here, we assume that the inventory is monitored under the $(r, Q)$ policy, although it is not the optimal policy. Among the other reasons, we make this assumption because the $(r, Q)$ policy is easily implementable in practice. Relaxing this assumption is also a potential research direction. Furthermore, the joint pricing and inventory control decision making under the assumption of an observable queue is worth further investigation.
3.A Proofs

Proof of Proposition 3.1. As described in Subsection 3.4.5, we set $\gamma_0 = 0$, and we know that $\gamma$ increases in each step. Thus, we have $\gamma \geq 0$.

(i) For a given $r$ and $Q$, $\Pi^\gamma$ is strictly concave with respect to $\lambda$ because

\[
\frac{\partial^2 \Pi^\gamma}{\partial \lambda^2} = -\frac{2c_w}{(\mu - \lambda)^3}(1 + \frac{\lambda g^r}{\eta Q}) - \frac{2c_w \mu}{\eta Q(\mu - \lambda)^2}(g^r + \frac{r \eta g^r + 1}{\lambda}) - \frac{c_w \lambda}{\eta Q(\mu - \lambda)} \frac{r(r + 1)\eta g^r + 2}{\lambda^3} \\
- \theta(\frac{\Lambda}{\Lambda (\Lambda - \lambda)} + \frac{\Lambda}{(\Lambda - \lambda)^2}) - \frac{hr(r + 1)\eta g^r + 2}{\lambda^3} - \frac{\gamma r(r + 1)\eta g^r + 2}{Q \lambda^3} < 0,
\]

where $g = \frac{1}{\lambda + \eta}$.

(ii) For a given $\lambda$, $\Pi^\gamma$ is jointly concave with respect to $r$ and $Q$ because

\[
\frac{\partial^2 \Pi^\gamma}{\partial Q^2} = -\frac{2c_w \lambda^2 g^r}{\eta Q^3(\mu - \lambda)} - \frac{2a \lambda}{Q^3} - \frac{2l \lambda^2 g^r}{\eta Q^3} - \frac{2\gamma \lambda g^r}{\eta Q^3} < 0,
\]

\[
\frac{\partial^2 \Pi^\gamma}{\partial r^2} = -\frac{c_w \lambda^2 g^r}{\eta Q(\mu - \lambda)} - \frac{h \lambda g^r}{\eta} - \frac{l \lambda^2 g^r}{\eta Q} - \frac{\gamma \lambda g^r}{\eta Q} < 0,
\]

and

\[
\frac{\partial^2 \Pi^\gamma}{\partial Q^2} \frac{\partial^2 \Pi^\gamma}{\partial r^2} - \left(\frac{\partial^2 \Pi^\gamma}{\partial Q \partial r}\right)^2 = \frac{X^2(\ln g)^2}{Q^2} + \frac{2h \lambda g^r(\ln g)^2 X}{\eta Q^2} + \frac{2a \lambda (\ln g)^2 X}{Q^3} + \frac{2ah \lambda^2 g^r(\ln g)^2 X}{\eta Q^3} > 0,
\]

where $g = \frac{1}{\lambda + \eta}$ and $X = \frac{c_w \lambda^2 g^r}{\eta Q(\mu - \lambda)} + \frac{l \lambda^2 g^r}{\eta Q} + \frac{\gamma \lambda g^r}{\eta Q} > 0$.

Proof of Lemma 3.1. Under the rationality assumption, the first order condition with
respect to \( \lambda \) results in

\[
v - \frac{c_w \mu}{(\mu - \lambda)^2} - \frac{c_w \mu \lambda g^r}{\eta Q (\mu - \lambda)^2} - \frac{a}{Q} \frac{\lambda g^r}{\eta Q} - \frac{1}{\eta Q} \left( \frac{c_w \lambda}{\eta_Q (\mu - \lambda)} + \frac{h}{\eta} + \frac{l \lambda}{\eta Q} + \frac{\gamma}{\eta Q} \right) \left( g^r + \frac{r \eta g^{r+1}}{\lambda} \right) = 0.
\]

(3.18)

The last four terms on the left-hand side of Eq. (3.18) are negative. Therefore, in order to have a solution to Eq. (3.18), we have to have \( v - \frac{c_w \mu}{(\mu - \lambda)^2} > 0 \) or equivalently \( \lambda < \mu - \sqrt{\frac{c_w \mu}{v}} \).

\( \square \)
CHAPTER 4

Quandary of Service Logistics: Fast or Reliable?

History: This chapter is based on a paper that is under revision for the European Journal of Operational Research. The work is done during my visit at Linköping University from November 2017 to February 2018. It is also presented at the European Logistics Association PhD Workshop, June 2018, Naples, Italy.
Quandary of Service Logistics: Fast or Reliable?

Ata Jalili Marand†, Ou Tang*, and Hongyan Li†

†Cluster for Operations Research, Analytics, and Logistics, Department of Economics and Business Economics, Aarhus University, Denmark

*Division of Production Economics, Department of Management and Engineering, Linköping University, Sweden

Abstract

Delivery time and delivery reliability are two top-level measures of delivery performance, and they both influence customers’ perception of service value. However, the classical queue-pricing literature emphasizes the former and ignores the latter. In order to bridge the gap between research and practice, this study addresses the interactive impact of price, delivery time, and delivery-reliability level on the equilibrium behaviour of rational customers and the optimal decisions of a revenue-maximizing service provider. We assume that the customers’ sensitivity to the delivery-reliability level is characterized by an increasing concave service value function. We model the operations of the service provider as an $M/M/1$ queue. Two cases are investigated: homogeneous customers and heterogeneous customers. For the homogeneous customers case, we analytically characterize the service provider’s optimal price, delivery-time, and delivery-reliability-level decisions. We show how the service provider’s decisions on whether to provide faster or more reliable service are affected when the different problem parameters are subject to variation. For instance, when customers become more sensitive to the delivery-reliability level, the service provider increases the delivery-reliability level at the expense of a longer delivery time. However, the optimal price may either increase or decrease depending on a benchmark value for the delivery-reliability level. For the heterogeneous customers case, our results suggest that when the potential arrival rate is sufficiently high, the service provider always benefits from markets with higher customer heterogeneity levels.

Keywords: OR in service industries; Revenue management; Pricing; Reliability; Queueing.
4.1 Introduction

There is sufficient empirical evidence to accentuate the significance of a superior delivery performance for incentivizing customers to buy and pay more. Two top-level dimensions of the delivery performance are emphasized in the literature: delivery time (speed) and delivery reliability (Handfield and Pannesi, 1992; Morash et al., 1996; Peng and Lu, 2017). The delivery reliability is measured as the rate of on-time deliveries, i.e., deliveries before or on the promised delivery time. The delivery time and delivery reliability are also underscored as the most important customer service elements in the marketing literature (Gronroos, 1988; Ballou, 2007).

Delivery performance is a key consideration in both B2C and B2B contexts. For instance, in a B2C context, comparison-shopping websites, e.g., shopzilla.org and bizrate.com, reveal that customer complaints often stem from order fulfilment failures in on-line retailing environments, and these failures bring about negative reactions of the customers (Rao et al., 2011). The failures concerning late deliveries are viewed as controllable and are hence judged more harshly by the customers (Kelley et al., 1993). Thus, consistent on-time delivery is a key element in determining the success of retailers in the e-commerce era (Lee and Whang, 2001). A superior delivery performance also plays an important role in customer satisfaction in B2B contexts: shorter delivery times allow business customers to speed up their operations, and more reliable deliveries enable them to accurately plan and coordinate their manufacturing activities (Peng and Lu, 2017). Moreover, selecting a fast and reliable supplier is particularly important when the buyer envisions a long-term relationship with the supplier (Benjaafar et al., 2007).

Beside delivery performance, the price of the service/product is another determinative factor influencing customers’ purchase behavior. Considering the price as a tool by which the congestion level can be controlled in congested systems dates back to Naor (1969) and Edelson and Hilderbrand (1975) for observable and unobservable systems, respectively. The interaction between the price and delivery time, therefore, has been widely addressed in the queue-pricing literature (see, e.g., Hassin and Haviv, 2003; Hassin, 2016). However, the delivery performance measure is often restricted to the delivery time, whereas the impact of the delivery reliability is ignored. In a congested system with price- and time-sensitive customers, when the delivery-reliability level is exogenously determined, a higher price reduces the throughput of the system and facilitates reduced service/product delivery times. But the practice could be more complicated. A shorter delivery time attracts more customers and leads to a higher congestion which makes the service less attractive for the upcoming customers. Studying the trade-off between price and delivery time has long been the primary focus of the queue-pricing studies.
Considering the delivery-reliability level along with the price and delivery time as endogenous decision variables makes the above-mentioned trade-off even more complex. This is because the delivery time and delivery-reliability level are tightly linked together in congested systems. When the customers’ perception of service value is also positively correlated with the delivery-reliability level and the price and capacity level are fixed, the delivery time and delivery-reliability level are substitutable decisions; the service provider has to make a trade-off between providing the customers with a shorter delivery time or a more reliable delivery time. However, given that the price can also be used to control the congestion level, the substitution relationship between the delivery time and delivery-reliability level decisions may no longer hold. Therefore, it is necessary to look profoundly into the interactive effects between the price, delivery time, and delivery-reliability level.

Although both empirical results (Baker et al., 2001; Peng and Lu, 2017) and analytical studies (So, 2000; Shang and Liu, 2011) stress the importance of an analytical framework to study the interactions between price, delivery time, and delivery-reliability level simultaneously, this issue is not addressed in depth in the literature. In this study, we aim at bridging this gap by proposing a model to capture such interactions. We make the following main contributions: We formulate and analyze the price-, delivery-time-, and delivery-reliability-sensitive customers’ equilibrium behavior and the revenue-maximizing service provider’s optimal decisions. Furthermore, we consider two cases: homogeneous and heterogeneous customers. In either case, we present insightful analytical results to highlight the importance of an integrated framework for addressing such problems.

For the case with homogeneous customers, we determine a feasible region for the delivery-reliability level and show that the service provider is better off when guaranteeing an intermediate delivery-reliability level instead of setting it close to the two ends of its feasible region. Our results also suggest that the service provider is better off in markets where the reliability sensitivity is either high or low rather than being an intermediate value. Regarding the heterogeneous customers case, our results show that as the customer heterogeneity level decreases, the optimal delivery performance declines (improves), and surprisingly the service provider’s optimal revenue increases (decreases), if the current equilibrium arrival rate is sufficiently high (low). At an intermediate equilibrium arrival rate, however, the service provider may either regain or lose revenue if the customer heterogeneity level decreases, but in either case the service provider charges a lower price and offers a slower and less reliable service.

The remainder of this paper is organized as follows. Section 4.2 reviews the related studies. The homogeneous customers case is addressed in Section 4.3. The main analytical results for the homogeneous customers case are presented in Subsection 4.3.3. In Section 4.4, we present the problem formulation and the analytical results regarding the
heterogeneous customers case. Finally, Section 4.5 concludes the paper.

4.2 Literature Review

The related literature can be categorized into two main streams: empirical studies and analytical studies. The first stream of studies empirically investigates the impact of delivery performance on customers’ perception of service/product (hereafter, service) value and their subsequent purchase behavior. The second stream, however, focuses on the analytical studies to investigate the interactions among delivery performance measures (and in some cases price) on customers’ equilibrium behavior and on optimal decisions of a social welfare/profit/revenue-maximizing service provider/manufacturer (hereafter, service provider). We review the most related studies in each stream below.

4.2.1 Empirical Studies

Empirical evidence has proven that the delivery time and delivery reliability are two measures of delivery performance that are important for both customers and service providers. In an early study, Kelley et al. (1993) investigate the specific types of failures occurring in a retail setting. They emphasize that consistent on-time delivery would lead to higher perceived value by the customers and would increase the customer willingness to pay. Rosenzweig et al. (2003) discuss the impact of the delivery reliability on the customer satisfaction and show that the delivery reliability directly impacts customer satisfaction.

According to Rao et al. (2014), there is statistically significant evidence suggesting that the delivery reliability mitigates the risk of product returns. Considering the demand-oriented capabilities, Morash et al. (1996) show that the delivery reliability is rated highest in CEO-perceived importance. Lee and Whang (2001) point up the on-time delivery as the most critical performance measure for both the on-line and off-line businesses of the firms engaged in e-commerce.

Studying the marketing and operations interface in relation to customer value, Sawhney and Piper (2002) discuss that lapses in on-time delivery can result in the loss of repeat business. They suggest that the firms can improve their on-time delivery performance by increasing the quality and speed of their marketing and operations interface. They state that when a firm’s operations make informed decisions based on explicit knowledge of customer needs and market conditions, and its marketing makes informed decisions based on explicit knowledge of capacity availability and technological constraints, then the firm tends to deliver more value to their customers.
Rao et al. (2011) empirically examine the relationship between order fulfillment delays and subsequent shopping behavior for previously loyal customers in an on-line retailing environment. They infer that the failures in keeping the delivery time promises can be detrimental. They use the data from a moderate-sized on-line retailer of printed materials and conclude that the delivery failures negatively affect the order frequency, order size, and customer anxiety. They also suggest that operations managers should endeavor to minimize not only the occurrence of the late deliveries, but also the length of delays when they are inevitable.

In a recent empirical study, Peng and Lu (2017) analyze the transaction data collected from a heating, ventilation, and air conditioning control product supply chain to study the interactions between delivery performance measures and price, and they examine the effect of the delivery performance on future customer transaction quantities. They take three dimensions of the delivery performance, i.e., delivery speed, on-time delivery rate (delivery-reliability level), and delivery date inaccuracy (late or early delivery), into consideration. According to this study, the price is affected by the on-time delivery rate and delivery speed, but not by delivery inaccuracy. They conclude that early deliveries do not appear to negatively affect the transaction quantity. Their results suggest that the delivery-reliability level contributes to profitability by allowing the firms to charge premium prices.

Although the empirical evidence emphasizes the existence of interactions between price, delivery time, and delivery reliability, there is no integrated framework to analytically study these interactions. We contribute to this stream of the literature by analytically studying the interactions between price, delivery time, and delivery reliability to understand the customers’ equilibrium behavior and the firms’ optimal decisions.

4.2.2 Analytical Studies

As shown in the previous subsection, empirical results demonstrate that decisions on price, delivery time, and delivery reliability do indeed interact. The importance of an integrated framework to study the interactions between these three decisions is also emphasized in the literature (see, e.g., So, 2000; Shang and Liu, 2011). However, only a few studies have addressed this issue analytically (see, e.g., Boyaci and Ray, 2006; Xiao and Qi, 2016). Even the interaction between delivery time and delivery reliability (with an exogenous price) is rarely addressed in the queue-pricing literature (see, e.g., Ho and Zheng, 2004; Shang and Liu, 2011). We review the studies that are closely related to our work below. See Table 4.1 for a summary of the key literature.

Boyaci and Ray (2006) consider the optimal differentiation strategy in terms of prices, delivery times, and delivery-reliability levels of a profit-maximizing firm that sells two
variants of a product to two classes of customers in a capacitated environment. The demand for each product is assumed to be linearly dependent on both products’ prices, delivery times, and delivery-reliability levels. They consider three cases: (i) price-and-time-based differentiation (with exogenous delivery-reliability level), (ii) time-and-reliability-based differentiation (with exogenous price), and (iii) price–time-and-reliability-based differentiation. For the latter case, they consider two scenarios: when the price, delivery time, and delivery reliability are (i) exogenous for one customer class, and when they are (ii) endogenous for both customer classes. The firm’s differentiation strategy for both scenarios is numerically solved due to intractability of the model.

In a two-stage supply chain with one supplier and one manufacturer, Xiao and Qi (2016) study the supplier’s and manufacturer’s equilibrium decisions in the supply chain under an all-unit quantity discount contract. In their model, the demand function is linearly dependent on the price, delivery time, and delivery-reliability level. They consider four cases regarding whether the delivery time, delivery reliability, and the manufacturer’s capacity are endogenous, and whether the manufacturer’s production cost is its private information. They conclude that an all-unit quantity discount scheme can coordinate the supply chain in most cases.

Our study differs from Boyaci and Ray (2006) and Xiao and Qi (2016) in two respects: First, we define a general service value function to capture the dependency between customers’ valuation of service and the reliability level, whereas they assume a linear price-, time-, and reliability-dependent demand function. In other words, while they employ a market-level aggregate demand model, we employ a quite different approach and construct the demand directly from the consumer utility function. Second, Boyaci and Ray (2006) and Xiao and Qi (2016) focus on the product differentiation strategy and supply chain coordination, respectively, whereas we analyze both the customers’ equilibrium behavior and the service provider’s optimal price, delivery time, and delivery-reliability level decisions.

Ho and Zheng (2004) seem to be the first to assume that the customer utility is a function of both delivery time and delivery-reliability level. This relationship is mod-

<table>
<thead>
<tr>
<th>Literature</th>
<th>Price</th>
<th>Delivery time</th>
<th>Delivery reliability</th>
<th>Customer choice behavior analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority of the classical queue-pricing literature</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Ho and Zheng (2004)</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Boyaci and Ray (2006)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Shang and Liu (2011)</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Anand et al. (2011)</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Xiao and Qi (2016)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Our Study</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of the related analytical studies
eled as a linear function. They consider the competition between two service providers, modeled as $M/M/1$ queues, and employ a multinomial logit (MNL) model to capture each service provider’s market share. The fact that they employ the MNL model and do not consider any constraint on the utility function makes it reasonable to infer that the customers are assumed to be boundedly rational. Having assumed that the market price is exogenous, each service provider aims at maximizing its market share. They show the existence of a unique Nash equilibrium. Moreover, they show that the game becomes a prisoner’s dilemma when the service providers also compete in capacity.

Shang and Liu (2011) extends Ho and Zheng (2004) to a two stage game among $N$ service providers, where the service providers compete in terms of capacity in the first stage and in terms of delivery time and delivery reliability in the second stage. When the capacities are fixed, they also show the existence of a unique Nash equilibrium. In addition, they find sufficient conditions for the existence of an interior equilibrium and a characterization of boundary equilibria in the first stage. Shang and Liu (2011) also assume that the price is exogenous and employ an MNL model. Also in their model, it is implicitly assumed that the customers are boundedly rational.

Our study also differs from Ho and Zheng (2004) and Shang and Liu (2011) in three perspectives. First, we assume that the customers are fully rational. It implies that each customer would choose to purchase the service/product only if her actual utility is greater than a threshold value. Second, we let the service value be an increasing concave function of the delivery-reliability level with a general shape. This implies that the incremental loss of value is higher at the lower delivery-reliability levels. Third, we consider the price as an endogenous decision variable.

In the context of customer-intensive service systems, Anand et al. (2011) let the service value be dependent on the service speed such that a lower service speed implies a higher value. Even though a slower service is valued higher by the customers, it would result in more congestion and negatively influence the utility. They show that the service provider would choose an intermediate service speed, which reflects this trade-off. The paper by Anand et al. (2011) is extended to the multiple customer classes and competition problems by Ni et al. (2013) and Li et al. (2016b), respectively.

Our model differs from that of Anand et al. (2011). While service speed is a decision variable in Anand et al. (2011), we assume that it is fixed, and it cannot be adjusted in the short run, because it is mainly determined at the strategic level. However, we let the service value be dependent on the delivery-reliability level. Our model shows how a service provider with fixed operational properties, i.e., fixed capacity level (service speed), has to decide on the service price, delivery time, and delivery-reliability level when customers are sensitive to these decisions.
To the best of our knowledge, there is no study in the queue-pricing literature that focuses on the equilibrium behavior of the price-, delivery-time-, and delivery-reliability-sensitive customers and assumes the customers to be rational. We focus on modeling and analyzing the customers’ equilibrium behavior and on optimizing the service provider’s price, delivery time, and delivery-reliability-level decisions. We consider two cases: homogeneous and heterogeneous customers.

4.3 Homogeneous Customers

4.3.1 Problem Description

We consider a revenue-maximizing service provider serving homogeneous, rational, and self-interested customers. The customers are assumed to be price, delivery time, and delivery-reliability sensitive. Moreover, it is assumed that the customers cannot observe the queue. Thus, they base their joining/balking decisions on the announced price and the promised delivery time and delivery-reliability level. Similar to the classical queue-pricing literature, we assume that the customer utility is linearly decreasing in the service price, \( p \), and promised delivery time, \( w \). We also let the service value be a function of the delivery-reliability level, \( r \), measured as the on-time-delivery rate, i.e., the percentage of deliveries before and on the promised delivery time. The utility function is defined as

\[
U(p, w, r) = V(r) - p - c_w w, \tag{4.1}
\]

where \( V(r) : [0, 1] \rightarrow \mathbb{R} \) is the service value function, and \( c_w \) is the waiting cost rate of customers. We impose two conditions on the service value function: (i) it is increasing in the delivery-reliability level, i.e., \( \frac{\partial V(r)}{\partial r} > 0 \), and (ii) the incremental value gained from additional reliability diminishes, i.e., \( \frac{\partial^2 V(r)}{\partial r^2} < 0 \).

We assume that the customers are rational. When their utility is greater (less) than a threshold value, they do purchase (balk), and if it is equal to the threshold value, the customers are indifferent between joining and balking. Without loss of generality, we normalize this threshold value to zero. Customers act independently and aim at maximizing their welfares. Given the price, delivery time, and delivery-reliability level, their decisions result in an aggregate equilibrium pattern of behavior called the customers’ equilibrium behavior in the queue-pricing literature (Hassin and Haviv, 2003).

The service operations of the service provider are modeled as an \( M/M/1 \) queue, similar to the most of related studies. Let \( \Lambda \) denote the potential arrival rate, \( \lambda \) denote the (effective) arrival rate, and \( \mu \) denote the service rate which is fixed and cannot be adjusted in the short run. We measure the delivery time by the sojourn time.
An M/M/1 queue, the sojourn time is exponentially distributed (Asmussen, 2008). The probability that the service time \( t \) does not exceed the promised delivery time \( w \), is 
\[
Pr(t \leq w) = 1 - e^{-(\mu - \lambda)w}.
\]
To meet the delivery-reliability level \( r \), the service provider requires that 
\[
Pr(t \leq w) = 1 - e^{-(\mu - \lambda)w} \geq r.
\]
Thus, to provide a service with the reliability level \( r \), the service provider has to promise a delivery time equal to or greater than 
\[
\frac{1}{\mu - \lambda} \ln \left( \frac{1}{1 - r} \right).
\]
Since the utility function is decreasing in the delivery time, i.e., 
\[
\frac{\partial U(p,w,r)}{\partial w} = -c_w < 0,
\]
the service provider is better off setting 
\[
w(\lambda, r) = \frac{1}{\mu - \lambda} \ln \left( \frac{1}{1 - r} \right).
\]
Moreover, it is assumed that the customers are able to accurately estimate their waiting time in the system given the reliability level \( r \) and the arrival rate \( \lambda \). Substituting \( w(\lambda, r) \) in Eq. (4.1), the utility function can be rewritten in terms of the price, arrival rate and delivery reliability level as
\[
U(p, \lambda, r) = V(r) - p - c_w w(\lambda, r),
\]
where
\[
w(\lambda, r) = \frac{1}{\mu - \lambda} \ln \left( \frac{1}{1 - r} \right).
\]
The promised delivery time \( w(\lambda, r) \) is strictly increasing in the delivery-reliability level, \( r \), and the arrival rate, \( \lambda \), and is strictly decreasing in the service rate, \( \mu \). Based on Eq. (4.2), we first analyze the customers’ equilibrium behavior and then determine the optimal price and delivery-reliability level.

### 4.3.2 Mathematical Formulation and Analysis

In this subsection, we formulate and analyze the described problem. The analysis includes four steps. First, we find a feasible region for the delivery-reliability level. Then, we analyze the customers’ equilibrium behavior when the service provider’s decisions are given. Having the customers’ equilibrium behavior, we can analyze the service provider’s problem. First for a given delivery-reliability level we find the service provider’s optimal pricing decision, and then we find the optimal delivery-reliability level of the service provider. To summarize we have the following steps:

(i) **Service-value analysis**: We analyze the impact of the delivery-reliability level on the customers’ valuation of service. This analysis lets us define a feasible range for the delivery-reliability level.

(ii) **Equilibrium arrival rate**: We study the customers’ equilibrium behavior for a given price and delivery-reliability level. This step gives the customers’ equilibrium arrival rate as a function of price and delivery-reliability level.
(iii) **Optimal price:** We find the optimal price for a given delivery-reliability level. This step gives us the optimal price and the corresponding equilibrium arrival rate as a function of the delivery-reliability level.

(iv) **Optimal delivery-reliability level:** The previous step reduces the original problem to a single-variable optimization problem. This step gives us the optimal delivery-reliability level and the corresponding optimal price and the equilibrium arrival rate.

**Service-Value Analysis**

In this subsection, we analyze the relationship between the delivery-reliability level and the customers’ perception of service value. In the utility function (4.2), the first term, \( V(r) \), is increasing and the last term, \( -c_w w(\lambda, r) \), is decreasing in the delivery-reliability level, \( r \). These conflicting effects are the core of our analysis in this subsection. We assume that the price and the congestion level of the system are at their lowest levels, i.e., \( p = 0 \) and \( \lambda = 0 \). Our aim is to answer the following question: When a customer does not need to pay and wait in queue for the service, is there any delivery-reliability level for which the service is worthy enough for the customer to wait for during the service time?

Upon her arrival, a customer would choose to join the service only if she obtains a non-negative utility. \( U(0, 0, r) \) is concave with respect to \( r \), i.e., \( \frac{\partial^2 U(0,0,r)}{\partial r^2} = \frac{\partial^2 V(r)}{\partial r^2} - \frac{c_w}{\mu(1-r)^2} < 0 \), and \( \frac{\partial U(0,0,r)}{\partial r} \to -\infty \) when \( r \to 1 \), and \( \frac{\partial U(0,0,r)}{\partial r} \to \frac{\partial V(r)}{\partial r} \big|_{r=0} - \frac{c_w}{\mu} \) when \( r \to 0 \). When \( \frac{\partial V(r)}{\partial r} \big|_{r=0} \leq \frac{c_w}{\mu} \), the impact of the waiting cost, \( -c_w w(\lambda, r) \), dominates the impact of service value, \( V(r) \), on their domain, and the above-mentioned conflicting behavior does not clearly appear. As a result, \( U(0,0,r) \) becomes a decreasing function with respect to \( r \) when \( \frac{\partial V(r)}{\partial r} \big|_{r=0} \leq \frac{c_w}{\mu} \).

However, when \( \frac{\partial V(r)}{\partial r} \big|_{r=0} > \frac{c_w}{\mu} \), \( U(0,0,r) \) first increases and then decreases with respect to \( r \). This implies that, at lower (higher) delivery-reliability levels, the increase (decrease) in service value (waiting cost) overweights the decrease (increase) in waiting cost (service value) as the delivery-reliability level increases. Let \( r_1 \) and \( r_2 \), \( r_1 < r_2 \), denote the delivery-reliability levels at which \( U(0,0,r) \) crosses the \( r \)-axis. In fact, \( r_1 \) and \( r_2 \) are two threshold values that a delivery reliability level below \( r_1 \) or above \( r_2 \) is not of customers’ interest at all even if the customers pay no other cost except for the time they wait to be served. Although both cases can be analyzed similarly, we are more interested in the second case, i.e., \( \frac{\partial V(r)}{\partial r} \big|_{r=0} > \frac{c_w}{\mu} \), and focus on this case throughout the rest of the paper. Apparently, customers will choose to join the service provider if \( U(0,0,r) \) is non-negative for some values of delivery-reliability level. Thus, we assume that there exists a \( r_b \in (r_1, r_2) \), such that \( U(0,0,r_b) \geq 0 \), since otherwise the problem becomes trivial. Figure 4.1 depicts
$U(0,0,r)$ with respect to $r$ for the two above-mentioned cases.

When $\frac{\partial V(r)}{\partial r}|_{r=0} > \frac{c_w}{\mu}$, the delivery-reliability level line is partitioned into three parts. In region 1 ($0 \leq r < r_1$), the delivery-reliability level is too low. Although it results in a short delivery time, the customers’ gain from the lower waiting cost does not compensate for the value loss that results in a negative utility. Similarly in region 3 ($r_2 < r \leq 1$), although the value of the service is high, it does not compensate for the high waiting cost. It is only in region 2 ($r_1 \leq r \leq r_2$) that the customers gain a non-negative utility. Thus, the feasible region for the delivery-reliability level can be tightened to $r \in [r_1, r_2]$. The assumption $U(0,0,r_b) \geq 0$ ensures that the interval $[r_1, r_2]$ is well defined. In fact, we have $V(r) > \frac{c_w}{\mu} \ln\left(\frac{1}{1-r}\right)$ for any $r \in (r_1, r_2)$, and $V(r) = \frac{c_w}{\mu} \ln\left(\frac{1}{1-r}\right)$ for $r \in \{r_1, r_2\}$. Lemma 4.1 shows how the problem parameters influence the feasible region of the delivery-reliability level. All the proofs are provided in Appendix 4.A.

**Lemma 4.1.** The interval $[r_1, r_2]$ becomes wider ($r_1$ decreases and $r_2$ increases) as $\mu$ increases or $c_w$ decreases.

**Equilibrium Arrival Rate**

In this subsection, we analyze the customers’ equilibrium behavior for a given price, $p$, and delivery-reliability level, $r$. The potential arrival rate, service rate, price, delivery-reliability level, and waiting cost rate are common knowledge. Since the choice of any customer affects the delivery performance experienced by other customers, a game theoretic approach is commonly used to analyze this situation (Armony and Haviv, 2003). Let $q^e(p,r)$ denote each customer’s equilibrium probability of joining and $\lambda^e(p,r)$ denote

\[\text{Figure 4.1: The impact of delivery-reliability level on the customer utility}\]

\[\text{Note that when } \frac{\partial V(r)}{\partial r}|_{r=0} \leq \frac{c_w}{\mu}, \text{ we have } r_1 = 0.\]
the corresponding equilibrium arrival rate. According to Chen and Frank (2004), three cases may occur:

(i) If \( V(r) \geq p + c_w w(\Lambda, r) \), then all of the customers will gain a non-negative utility even if they all join. In this case, joining with probability \( q^e(p, r) = 1 \) is the dominant equilibrium strategy.

(ii) If \( V(r) \leq p + c_w w(0, r) \), then any customer will get a negative utility even if no other customer joins. In this case, joining with probability \( q^e(p, r) = 0 \) (balking with probability one) is the dominant equilibrium strategy.

(iii) If \( p + c_w w(0, r) < V(r) < p + c_w w(\Lambda, r) \), then each customer plays a mixed strategy at equilibrium such that \( q^e(p, r) = \frac{\lambda^e(p, r)}{\Lambda} \in (0, 1) \), where \( \lambda^e(p, r) \) solves \( V(r) = p + c_w w(\lambda^e(r, p), r) \).

Lemma 4.2 characterizes the equilibrium arrival rate.

**Lemma 4.2.** For a given price \( p \) and delivery-reliability level \( r \), such that \( r \in [r_1, r_2] \) and \( p < V(r) \), the customer equilibrium arrival rate is characterized as

\[
\lambda^e(p, r) = \begin{cases} 
\Lambda & \text{if } \Lambda \leq \tilde{\lambda}(p, r), \\
\tilde{\lambda}(p, r) & \text{if } 0 \leq \tilde{\lambda}(p, r) < \Lambda, \\
0 & \text{if } \tilde{\lambda}(p, r) < 0,
\end{cases}
\]  

(4.4)

where \( \tilde{\lambda}(p, r) = \mu - \frac{c_w}{V(r) - p} \ln \left( \frac{1}{1 - r} \right) \).

Lemma 4.2 implies that, when the market size \( \Lambda \) is low, i.e., \( \Lambda \leq \tilde{\lambda}(p, r) \), all customers will join the service. In this case, the equilibrium arrival rate would be independent of the price and the delivery-reliability level. However, the equilibrium arrival rate is uniquely determined as a function of \( p \) and \( r \) when the market size \( \Lambda \) is large enough, i.e., \( \tilde{\lambda}(p, r) < \Lambda \).

**Optimal Pricing Decision**

Based on the customers’ equilibrium behavior, the service provider’s optimization problem is modeled as follows.

\[
\max_{p, r} R(p, r) = p \lambda^e(p, r) \\
\text{s.t. } 0 \leq p < V(r), \\
\quad r_1 \leq r \leq r_2.
\]
The objective is to maximize the revenue rate. To solve this optimization problem, we first determine the optimal pricing decision as a function of the delivery-reliability level and then find the optimal delivery-reliability level. The results are presented in Lemma 4.3.

Lemma 4.3. For a given delivery-reliability level \( r \), such that \( r \in [r_1, r_2] \), the optimal pricing decision of the service provider is

\[
p(r) = \begin{cases} 
V(r) - \frac{c_w}{\mu - \Lambda} \ln\left(\frac{1}{1-r}\right) & \text{if } \Lambda \leq \hat{\lambda}(r), \\
V(r) - \frac{c_w}{\mu - \hat{\lambda}(r)} \ln\left(\frac{1}{1-r}\right) & \text{if } 0 \leq \hat{\lambda}(r) < \Lambda,
\end{cases}
\]

where \( \hat{\lambda}(r) = \mu - \sqrt{\frac{c_w\mu}{V(r)} \ln\left(\frac{1}{1-r}\right)} \). Moreover, the corresponding equilibrium arrival rate equals

\[
\lambda^e(r) = \begin{cases} 
\Lambda & \text{if } \Lambda \leq \hat{\lambda}(r), \\
\hat{\lambda}(r) & \text{if } 0 \leq \hat{\lambda}(r) < \Lambda,
\end{cases}
\]

and the corresponding revenue function equals

\[
R(r) = \begin{cases} 
\Lambda(V(r) - \frac{c_w}{\mu - \Lambda} \ln\left(\frac{1}{1-r}\right)) & \text{if } \Lambda \leq \hat{\lambda}(r), \\
(\sqrt{\mu V(r)} - \sqrt{c_w \ln\left(\frac{1}{1-r}\right)})^2 & \text{if } 0 \leq \hat{\lambda}(r) < \Lambda.
\end{cases}
\]

Optimal Delivery-Reliability Level Decision

Proposition 4.1 characterizes the service provider’s optimal delivery-reliability level and price decisions and the corresponding equilibrium arrival rate.

Proposition 4.1. Define \( r_{\text{full}} \) and \( r_{\text{part}} \) as

\[
r_{\text{full}} = \arg\left\{ \frac{\partial}{\partial r}(V(r) - \frac{c_w}{\mu - \Lambda} \ln\left(\frac{1}{1-r}\right)) = 0 \right\},
\]

\[
r_{\text{part}} = \arg\left\{ \frac{\partial}{\partial r}(\sqrt{\mu V(r)} - \sqrt{c_w \ln\left(\frac{1}{1-r}\right)}) = 0 \right\}.
\]

(i) If \( \Lambda > \hat{\lambda}(r_{\text{part}}) \), the optimal delivery-reliability level equals \( r^* = r_{\text{part}} \) as determined by Eq. (4.9), the corresponding equilibrium arrival rate equals \( \lambda^e = \mu - \sqrt{\frac{c_w\mu}{V(r^*)} \ln\left(\frac{1}{1-r^*}\right)} \), and the optimal price equals \( p^* = V(r^*) - \frac{c_w}{\mu - \lambda^e} \ln\left(\frac{1}{1-r^*}\right) \).

(ii) If \( \Lambda \leq \hat{\lambda}(r_{\text{part}}) \), the optimal delivery-reliability level equals \( r^* = r_{\text{full}} \) as determined by Eq. (4.8), the corresponding equilibrium arrival rate equals \( \lambda^e = \Lambda \), and the
The service provider serves either all the customers, which is defined as full market coverage, or just a part of the customers, which is defined as partial market coverage. These two terms will be used for the forthcoming analysis. According to Proposition 4.1 (i), when the potential arrival rate is sufficiently high, i.e., \( \Lambda > \hat{\lambda}(r_{\text{part}}) \), the service provider only serves some of the arriving customers at optimality (partial market coverage). In this case, the optimal delivery-reliability level, price, and delivery time are independent of the potential arrival rate \( \Lambda \). Thus, decreasing the potential arrival rate from any high value, e.g., infinity, down to the threshold value \( \hat{\lambda}(r_{\text{part}}) \) does not influence the optimal decisions of the service provider. According to Proposition 4.1 (ii), however, when the potential arrival rate is sufficiently low, i.e., \( \Lambda \leq \hat{\lambda}(r_{\text{part}}) \), the service provider serves all of the arriving customers at optimality (full market coverage). In this case, the optimal delivery-reliability level, price, and delivery time are all dependent on the potential arrival rate.

**4.3.3 Sensitivity Analysis and Managerial Insights**

The following three propositions further elaborate the results presented in Proposition 4.1. Proposition 4.2 elaborates the relationships between the price, equilibrium arrival rate, and the optimal delivery-reliability level for the partial-market coverage case.

**Proposition 4.2.** In the case of partial market coverage, the price \( p(r) \) is a strictly pseudo-concave and unimodal function with respect to \( r \) over the interval \([r_1, r_2]\) and reaches its maximum at \( r = r_p \), such that \( r_p \in (r_1, r_2) \) and \( r_p > r^* = r_{\text{part}} \). Moreover, the equilibrium arrival rate \( \lambda^*(r) \) is a pseudo-concave and unimodal function with respect to \( r \) over the interval \([r_1, r_2]\) and reaches its maximum at \( r = r_\lambda \), such that \( r_\lambda \in (r_1, r_2) \) and \( r_\lambda < r^* = r_{\text{part}} \). Thus, \( r^* \in (r_\lambda, r_p) \).

While Subsection 4.3.2 explains how the feasible region for the delivery-reliability level should be restricted to \([r_1, r_2] \subset [0, 1]\), Proposition 4.2 shows that the optimal delivery-reliability level belongs to an even narrower interval, i.e, \( r^* \in (r_\lambda, r_p) \subset [r_1, r_2] \). Figure 4.2 schematically depicts the relationships between \( r^*, r_\lambda, \) and \( r_p \), and it partitions the interval \([r_1, r_2]\) (region 2 in Figure 4.1) into three parts. Next, we explain the behavior of the price and equilibrium arrival rate in each of these parts. Our analysis in this part is similar to that of Anand et al. (2011) for a service-speed dependent service value function.

Figure 4.2 shows that by increasing the delivery-reliability level from \( r_1 \) to \( r_\lambda \) (part 1), the service provider can increase both the price and the equilibrium arrival rate. In the interval \([r_1, r_\lambda]\), as the delivery-reliability level increases, the service value, \( V(r) \), also

\[
\text{optimal price equals } p^* = V(r^*) - \frac{c w}{\mu - \Lambda} \ln(\frac{1}{1-r^*}).
\]
increases, and the increase of the service value outweighs the increase of the waiting cost, \( c_w \lambda(r, r) \). In contrast to the interval \([r_1, r_\lambda]\) (part 1), as the delivery-reliability level increases, both the price and the equilibrium arrival rate decrease in the interval \([r_p, r_2]\) (part 3). This implies that the increase in the waiting cost strictly dominates the increase in the service value. Therefore, from the service provider’s point of view, \( r_\lambda \) performs better than any other delivery-reliability level in \((r_1, r_\lambda)\), and \( r_p \) performs better than any other delivery-reliability level in \((r_p, r_2)\).

Despite that, as the delivery-reliability level increases from \( r_\lambda \) to \( r^* \), although the equilibrium arrival rate decreases, the service provider gains a higher revenue by charging a higher price for a more reliable service. However, as the delivery-reliability level increases from \( r^* \) to \( r_p \), the decrease in the equilibrium arrival rate outweighs the increase in the price, and subsequently the revenue decreases. Thus, the optimal delivery-reliability level is located within the part 2, i.e., \( r^* \in (r_\lambda, r_p) \). This implies that the service provider must choose an intermediate delivery-reliability level rather than an extreme value.

Based on the analysis above, it is also interesting to analyze the impact of the problem parameters on the customers’ equilibrium behavior and the optimal decisions of the service provider. Therefore, we conduct a sensitivity analysis subject to a specific service value function as

\[
V(r) = v_b + c_r \left( \frac{1}{r_b} - \frac{1}{r} \right),
\]

where \( v_b \) is the benchmark net value of the service, \( r_b \) is the benchmark delivery-reliability level, \( r_b \in (0, 1) \), and \( c_r \) is the customers’ sensitivity to the delivery-reliability level (the reliability sensitivity in short). The service value function in Eq. (4.10) satisfies the two conditions imposed in Subsection 4.3.1, i.e., it is an increasing and concave function of the delivery-reliability level. The service value equals the benchmark net value when the
service provider provides his service with the benchmark delivery-reliability level, i.e., \( r = r_b \). When \( c_r = 0 \), the service value function is independent of the delivery-reliability level, i.e., \( V(r) = v_b \), as it is in the classical queue-pricing literature. However, as \( c_r \) increases, the service value becomes more sensitive to deviations from the benchmark delivery-reliability level. It is clear that any delivery-reliability level above the benchmark results in a service value greater than the benchmark service value, and vice versa.

Proposition 4.3 elaborates the sensitivity of the service provider’s optimal decisions to the potential arrival rate in the full market coverage case.

**Proposition 4.3.** (i) In the full market coverage case, the optimal delivery-reliability level and price are decreasing, and the optimal delivery time is increasing in the potential arrival rate, \( \Lambda \). (ii) In the full market coverage case compared to the partial market coverage case, the service provider charges a higher price and offers a faster and more reliable service.

As the potential arrival rate, \( \Lambda \), decreases from the threshold value \( \hat{\lambda}(r_{\text{part}}) \) to zero, the service provider’s residual capacity, i.e., \( \mu - \lambda^* \), increases. Consequently, the service provider exploits the higher residual capacity to improve his delivery performance by increasing the delivery-reliability level and shortening the delivery time as stated in Proposition 4.3 (i). Moreover, he charges the customers a higher price for his superior delivery performance.

Proposition 4.3 (ii) additionally concludes that, for any sufficiently low potential arrival rate, i.e., \( \Lambda \leq \hat{\lambda}(r_{\text{part}}) \), the optimal price and delivery-reliability level are higher, and the optimal delivery time is lower compared to the case when the potential arrival rate is sufficiently high, i.e., \( \Lambda > \hat{\lambda}(r_{\text{part}}) \). This can be explained as follows: In the partial market coverage case, the service provider has a lower residual capacity compared to the full market coverage case. The lower residual capacity brings about an inferior delivery performance, i.e., a lower delivery-reliability level and a slower service, which also forces the service provider to charge a lower price for an inferior delivery performance.

Proposition 4.4 elaborates the sensitivity of the service provider’s optimal decisions to the problem parameters.

**Proposition 4.4.** The optimal decisions of the service provider are sensitive to the problem parameters as reported in Table 4.2.

**Sensitivity to** \( c_r \). According to Proposition 4.4, in the full market coverage case, as the customers become more sensitive to the delivery-reliability level, i.e., as \( c_r \) increases, the service provider provides a more reliable service and also increases the delivery time...
Figure 4.3: Sensitivity of the service provider’s optimal decisions to the problem parameters in the partial market coverage case

(a) $\mu = 4.0$, $v_b = 4.0$, $\Lambda > \mu$

(b) $\mu = 4.0$, $v_b = 4.0$, $\Lambda > \mu$

(c) $\mu = 4.0$, $v_b = 4.0$, $\Lambda > \mu$

(d) $\mu = 4.0$, $v_b = 4.0$, $\Lambda > \mu$

(e) $\mu = 4.0$, $v_b = 4.0$, $\Lambda > \mu$

(f) $r_b = 0.90$, $v_b = 4.0$, $\Lambda = 10$

(g) $\mu = 4.0$, $c_r = 5.5$, $c_w = 4.5$, $\Lambda > \mu$ (h) $\mu = 4.0$, $c_r = 5.5$, $c_w = 4.5$, $\Lambda > \mu$
Table 4.2: Sensitivity of the service provider’s optimal decisions to the problem parameters

to be able to meet the delivery-reliability level. However, the service provider’s pricing strategy depends on the benchmark delivery-reliability level, \( r_b \). If the optimal delivery-reliability level is less (greater) than the benchmark delivery-reliability level, i.e., \( r^* < r_b \) (\( r^* > r_b \)), then the service provider decreases (increases) the price as the reliability sensitivity increases. In other words, when the reliability performance of the service provider is below (above) the market benchmark, then he charges a more reliability-sensitive customer a lower (higher) price.

As the reliability sensitivity increases, it is obvious that the optimal revenue also changes according to the reliability performance of the service provider compared to the market benchmark. It can be shown that the condition \( r^* < r_b \) (\( r^* > r_b \)), under the full market coverage case in Table 4.2, can be translated to \( c_r < \frac{c_w}{\mu - \Lambda} \frac{r_b^2}{1 - r_b} \) (\( c_r > \frac{c_w}{\mu - \Lambda} \frac{r_b^2}{1 - r_b} \)). Thus, if the reliability sensitivity, \( c_r \), is less (greater) than the threshold value \( \frac{c_w}{\mu - \Lambda} \frac{r_b^2}{1 - r_b} \), then the optimal revenue is decreasing (increasing) in \( c_r \).

For the partial market coverage case, Proposition 4.4 shows that the delivery-reliability level is increasing in \( c_r \) similar to the full market coverage case. However, we do not have analytical results for the change directions of the other decisions variables. Figures 4.3a-4.3c show the sensitivity of the optimal price, delivery time, and the equilibrium arrival rate with respect to the reliability sensitivity\(^2\). According to Figure 4.3b, as the customers become more sensitive to the delivery-reliability level, the service provider quotes a longer delivery time. This observation is intuitive, since a longer delivery time would be more easily met at a higher delivery-reliability level.

Figures 4.3a and 4.3c show that the optimal price and the equilibrium arrival rate first decrease and then increase as \( c_r \) increases. Thus, the price has a behavior similar to the full market coverage case. As a result of the interaction between the optimal price and the equilibrium arrival rate, the optimal revenue also first decreases and then increases as \( c_r \) increases. To be more precise, Propositions 4.4 states that as long as the reliability performance of the service provider is below (above) the market benchmark, the service

\(^2\)Our numerical results for a wide range of problem parameters reveal the same behavior as presented in Figure 4.3 for some sample parameter sets.
provider is worse off (better off) in a more reliability-sensitive market. It can similarly be shown that the condition \( r^* < r_b \) \((r^* > r_b)\), under the partial market coverage case in Table 4.2, can be translated to \( c_r < \frac{c_w}{\mu - \lambda(r_b)} \frac{r_b^2}{1-r_b} \) \((c_r > \frac{c_w}{\mu - \lambda(r_b)} \frac{r_b^2}{1-r_b})\). Thus, as long as \( c_r < \frac{c_w}{\mu - \lambda(r_b)} \frac{r_b^2}{1-r_b} \), the revenue is decreasing with respect to \( c_r \), and it is increasing for \( c_r > \frac{c_w}{\mu - \lambda(r_b)} \frac{r_b^2}{1-r_b} \). This conclusion is consistent with that of the full market coverage case.

**Sensitivity to** \( c_w \). According to Proposition 4.4, in both the full market coverage and partial market coverage cases, as the waiting cost rate, \( c_w \), increases, the service provider quotes a shorter delivery time and decreases the delivery-reliability level in order to be able to meet the shortened quoted delivery time. Moreover, the service provider regulates the queue by a lower price. Figure 4.3e shows that the equilibrium arrival rate also decreases as the waiting cost rate increases. Proposition 4.4 additionally states that the service provider is better off in markets where the waiting cost rate is lower.

**Sensitivity to** \( \mu \). As the capacity level, \( \mu \), increases, the service provider has the capability to offer a faster and also more reliable service. He also charges a premium price for the superior delivery performance. These results hold for both the full market coverage and partial market coverage cases. Moreover, these results are consistent with the empirical evidence in Peng and Lu (2017), that suggests that the price is positively influenced by a more reliable and faster service. Figure 4.3f shows the impact of the capacity level on the equilibrium arrival rate in the partial market coverage case. According to this figure, the equilibrium arrival rate is increasing in the capacity level. It is intuitive that a higher capacity level enables the service provider to serve more customers at a higher price and make more revenue. When the extra capacity comes with a cost, of course, the results would be different.

**Sensitivity to** \( v_b \) and \( r_b \). In the full market coverage case, the delivery-reliability level and delivery time are invariant with respect to the benchmark net value of service, \( v_b \), and the benchmark delivery-reliability level, \( r_b \). The service provider, however, charges a higher price in a market with a higher \( v_b \). Because as \( v_b \) increases, the service value also increases for a given delivery-reliability level, i.e., \( \frac{\partial V(r)}{\partial v_b} = 1 > 0 \). Moreover, as \( r_b \) increases, the service value decreases for a given delivery-reliability level, i.e., \( \frac{\partial V(r)}{\partial r_b} = -\frac{c_r}{r_b^2} < 0 \). As a result, the service provider decreases the price to compensate for the service-value loss.

In the partial market coverage case, however, the service provider lowers the delivery performance, i.e., reduces the delivery-reliability level and increases the delivery time, as \( v_b \) increases. In contrast, as \( r_b \) increases, the service provider enhances the delivery performance by offering a more reliable and faster service.
The variations in the price with respect to \( v_b \) and \( r_b \) are shown in Figures 4.3g and 4.3h respectively. According to these figures and Table 4.2, as \( v_b \) increases, the service provider charges a premium price even for an inferior delivery performance. Conversely, as \( r_b \) increases, the service provider charges a lower price even for a superior delivery performance. This can be explained as follows: As the benchmark net value increases from \( v_b \) to \( v_b' \), or the benchmark delivery-reliability level decreases from \( r_b \) to \( r_b' \), the service value \( V(r) \) increases for the previously optimal delivery-reliability level (the optimal delivery-reliability level for \( v_b \) or \( r_b \)). Consequently, a positive surplus remains for the customer. In order to extract this surplus, the service provider charges a higher price, increases the delivery time, and decreases the delivery-reliability level.

Figure 4.3g and Table 4.2 also imply that, as \( v_b \) increases, the revenue of the service provider increases, since both the optimal price and the equilibrium arrival rate increase. On the other hand, according to Figure 4.3h and Table 4.2, as \( r_b \) increases, both the optimal price and the equilibrium arrival rate decrease, and this results in a lower revenue for the service provider. We therefore conclude that the service provider is better off in markets in which the benchmark net value of service is high, or the benchmark delivery-reliability level is low.

### 4.4 Heterogeneous Customers

#### 4.4.1 Problem Description and Mathematical Formulation

In this section, we assume that the customers are heterogeneous in their benchmark net values, \( v_b \). The customer heterogeneity can also be captured through other modeling assumptions, for instance, by assuming that the customers have different benchmarks for the delivery-reliability level, \( r_b \). Results presented in this section will hold with slight adjustments for this case as well. Littlechild (1974) is the first to study customer heterogeneity in service value in an unobservable queuing system. However, he does not consider the reliability sensitivity of the customers.

We assume that \( v_b \) is a random variable with cumulative distribution function \( F(\cdot) \) such that \( v_b \in [\underline{v}_b, \overline{v}_b] \), and \( 0 \leq \underline{v}_b < \overline{v}_b < \infty \). Moreover, \( F \) is assumed to be continuous and differentiable. Let \( \overline{F} = 1 - F \) and \( \overline{F}^{-1} \) be the inverse of \( \overline{F} \). We aim at looking into the optimization problem of a revenue-maximizing service provider. We go through the following steps. (i) We find the optimal price as a function of the equilibrium arrival rate and the delivery-reliability level. Then, (ii) we find the optimal delivery-reliability level for a given equilibrium arrival rate. At last, (iii) we find the equilibrium arrival rate that maximizes the service provider’s revenue.
At equilibrium, only the customers who get a non-negative utility would buy the service. Thus, for a given price and delivery-reliability level, the equilibrium arrival rate equals

\[ \lambda^e = \Lambda \Pr(U \geq 0) = \Lambda \Pr(v_b + c_r \left( \frac{1}{r_b} - \frac{1}{r} \right) - p - c_w w(\lambda^e, r) \geq 0) = \Lambda \Pr(v_b + c_r \left( \frac{1}{r_b} - \frac{1}{r} \right), \]

that results in

\[ F^{-1}(\frac{\lambda^e}{\Lambda}) = p + c_w w(\lambda^e, r) - c_r \left( \frac{1}{r_b} - \frac{1}{r} \right). \] (4.11)

From Eq. (4.11), the price can be expressed as a function of the equilibrium arrival rate and delivery-reliability level, as

\[ p(\lambda^e, r) = F^{-1}(\frac{\lambda^e}{\Lambda}) - c_w w(\lambda^e, r) + c_r \left( \frac{1}{r_b} - \frac{1}{r} \right). \] (4.12)

The service provider’s optimization problem is modeled as follows.

\[
\max_{\lambda^e, r} R(\lambda^e, r) = \lambda^e \mu \\
\text{s.t. } 0 \leq \lambda^e < \mu, \\
\lambda^e \leq \Lambda, \\
0 \leq r \leq 1.
\]

\( R(\lambda^e, r) \) is concave with respect to the delivery-reliability level \( r \), i.e., \( \frac{\partial^2 R(\lambda^e, r)}{\partial r^2} = -\frac{c_w \lambda^e}{(\mu - \lambda^e)(1 - r)^2} - \frac{2c_w \lambda^e}{r^3} < 0 \). Thus, we use the first-order condition to find the delivery-reliability level as a function of the equilibrium arrival rate. The delivery-reliability level equals

\[ r(\lambda^e) = -\frac{c_r (\mu - \lambda^e)}{2c_w} + \sqrt{\left(\frac{c_r (\mu - \lambda^e)}{2c_w}\right)^2 + \frac{c_r (\mu - \lambda^e)}{c_w}}, \] (4.13)

for \( \lambda^e \in [0, \mu] \). It is clear that \( r(\lambda^e) \in (0, 1) \), thus the constraint over the delivery-reliability level is satisfied.

The delivery time can be expressed in terms of the equilibrium arrival rate as \( w(\lambda^e) = w(\lambda^e, r(\lambda^e)) \) following from Eq. (4.3). Similarly, the price can be expressed as a function of the equilibrium arrival rate by substituting Eq. (4.13) in Eq. (4.12), i.e., \( p(\lambda^e) = p(\lambda^e, r(\lambda^e)) \). As a result, the optimization problem can be reduced to finding the optimal
equilibrium arrival rate as follows:

\[
\max_{\lambda^e} R(\lambda^e) = p(\lambda^e) \lambda^e \\
\text{st. } 0 \leq \lambda^e < \mu, \\
\lambda^e \leq \Lambda.
\]

To further study the above optimization problem, the following definition and assumptions are necessary. We define the marginal value functions as

\[v(\lambda^e) = \frac{F^{-1}(\lambda^e \Lambda)}{\Lambda},\]

for \(\lambda^e \in [0, \Lambda]\). \(v(\lambda^e)\) is the valuation of the marginal customer corresponding to \(\lambda^e\). The properties of \(F\) imply that \(0 \leq v = v(\Lambda) < v(0) = \overline{v} < \infty\) and \(-\infty < \frac{dv(\lambda^e)}{d\lambda^e} < 0\) for \(\lambda^e \in [0, \Lambda]\). Moreover, we assume that \(\lambda^e v(\lambda^e)\) is concave (see Afeche, 2013, for a similar assumption). Proposition 4.5 characterizes the service provider’s optimal solutions and the corresponding equilibrium arrival rate. It implies that the unique solution to the problem can be easily found using numerical methods such as the bisection method.

**Proposition 4.5.** \(R(\lambda^e)\) is concave with respect to \(\lambda^e\) and reaches its maximum at

\[\hat{\lambda} = \arg\{\frac{\partial R(\lambda^e)}{\partial \lambda^e} = 0\},\]

such that \(\hat{\lambda} \in [0, \mu]\).

(i) If \(\Lambda > \hat{\lambda}\), the equilibrium arrival rate equals \(\lambda^* = \hat{\lambda}\), and the optimal delivery-reliability level and price are determined by Eq. (4.13) and Eq. (4.12), respectively.

(ii) If \(\Lambda \leq \hat{\lambda}\), the equilibrium arrival rate equals \(\lambda^* = \Lambda\), and the optimal delivery-reliability level and price are determined by Eq. (4.13) and Eq. (4.12), respectively.

Similar to the case with homogeneous customers, the service provider serves either all of the customers (full market coverage) or just a part of the customers (partial market coverage) depending on whether the potential arrival rate is high or low, i.e., \(\Lambda > \hat{\lambda}\) or \(\Lambda \leq \hat{\lambda}\).

**4.4.2 Sensitivity Analysis and Managerial Insights**

The following propositions and corollary elaborate the results in Proposition 4.5.
Proposition 4.6. Regardless of the distribution of the benchmark net value, the delivery-reliability level $r(\lambda^c)$ is decreasing and concave, the price $p(\lambda^c)$ is decreasing, and the delivery time $w(\lambda^c)$ is increasing with respect to the equilibrium arrival rate $\lambda^c$.

Proposition 4.6 presents important results that are used later in this subsection. According to this proposition, all of the service provider’s decisions are monotone in the equilibrium arrival rate. We can directly use these results to conclude the following corollary.

Corollary 4.1. In the full market coverage case, the optimal delivery-reliability level and price are decreasing, and the optimal delivery time is increasing with respect to the potential arrival rate.

Corollary 4.1 is consistent with Proposition 4.3 for the homogeneous customers case. Similarly, it can be stated that, as the potential arrival rate, $\Lambda$, decreases from the threshold value $\hat{\lambda}$ to zero, the service provider’s residual capacity, i.e., $\mu - \lambda^e*$, increases. The service provider consequently exploits the higher residual capacity and enhances his delivery performance by increasing the delivery-reliability level and shortening the delivery time. He also charges a higher price for a better delivery performance.

To further analyze the impact of the customer heterogeneity on the service provider’s optimal decisions, we assume that the benchmark net value, $v_b$, is uniformly distributed. By a slight abuse of notation, we let $v_b$ stand for the mean of the benchmark net value, $E(v_b)$. We also define $v_b = v_b - \epsilon$ and $\bar{v}_b = v_b + \epsilon$. Thus, the variance of the benchmark net value is defined by a single parameter $\epsilon$, i.e. $Var(v_b) = \frac{\epsilon^2}{\pi}$. Apparently, as $\epsilon$ increases, the customers become more dispersed. The parameter $\epsilon$ can be used as a measure representing the customer heterogeneity level.

Proposition 4.7. Fix $\epsilon$. When the mean of the benchmark net value, $v_b$, decreases, the optimal delivery-reliability level increases, the optimal delivery time and the equilibrium arrival rate decrease, and the service provider gains a lower revenue.

Proposition 4.7 is consistent with Proposition 4.4 for the homogeneous customers case. Proposition 4.7 states that, with the same level of the customer heterogeneity, $\epsilon$, the service provider is worse off in markets where the mean benchmark net value, $E(v_b)$, is higher; the service provider consequently has to improve his delivery performance and also decrease the price. Moreover, even with a better delivery performance and a lower price, he gains a lower revenue. These results hold as long as the customer heterogeneity level is fixed. The next proposition reports the sensitivity of the optimal decisions to the customer heterogeneity level.

94
Fix the mean of the benchmark net value. Suppose that $r^*, p^*, w^*, \lambda^*, \lambda^e$, and $R^*$ ($r^{*'}, p^{*'}, w^{*'}, \lambda^{e*}$, and $R^{*'}$) are the optimal delivery-reliability level, price, delivery time, equilibrium arrival rate, and the optimal revenue corresponding to the customer heterogeneity level $\epsilon (\epsilon')$, respectively, such that $0 < \epsilon' < \epsilon \leq v_b$.

**Proposition 4.8.** Fix $v_b$. Then, if the customer heterogeneity level decreases from $\epsilon$ to $\epsilon'$, the service provider’s optimal decisions, revenue, and the equilibrium arrival rate change as reported in Table 4.3.

| $\lambda^e < \frac{\Lambda}{2}$ | $r^* < r^{*'}$ | n/a | $w^* > w^{*'}$ | $\lambda^e > \lambda^{e*'}$ | $R^* > R^{*'}$ |
| $\lambda^e = \frac{\Lambda}{2}$ | $r^* = r^{*'}$ | $p^* > p^{*'}$ | $w^* = w^{*'}$ | $\lambda^e = \lambda^{e*'}$ | $R^* > R^{*'}$ |
| $\frac{\Lambda}{2} < \lambda^e < \frac{\Lambda}{2}$ | $r^* > r^{*'}$ | $p^* > p^{*'}$ | $w^* < w^{*'}$ | $\lambda^e < \lambda^{e*'}$ | n/a |
| $\lambda^e = \Lambda$ | $r^* = r^{*'}$ | n/a | $w^* < w^{*'}$ | $\lambda^e = \lambda^{e*'}$ | $R^* < R^{*'}$ |

Table 4.3: Impact of the customer heterogeneity level on the optimal decisions of the service provider

Proposition 4.8 is consistent with the results presented in Larsen (1998), Subsection 2.2, except that he takes the delivery-reliability level as an exogenous variable. In a part of his paper, Larsen (1998) investigates the impact of the variance of the net value on the service provider’s revenue and the welfare. He concludes that while an increase in the variance would result in either a higher or a lower revenue, it would always result in a higher welfare. We further elaborate this proposition in the following.

According to Proposition 4.8, when the equilibrium arrival rate is relatively low (high) at optimality, i.e., the market share is less (greater) than $\frac{1}{4}$ ($\frac{1}{2}$), if the customers become less dispersed in terms of the net value, then the equilibrium arrival rate and the optimal revenue decrease (increase), even though the service provider offers a superior (inferior) delivery performance. However, when the optimal arrival rate is at an intermediate level, i.e., the market share is greater than $\frac{1}{4}$ and less than $\frac{1}{2}$, the service provider is either better off or worse off in a market where the customers are less dispersed. In this case, the service provider provides an inferior service performance and charges a lower price.

Proposition 4.8 also enables us to compare the results of the homogeneous and heterogeneous customers cases. While $\epsilon' \to 0$ represents the homogeneous customers case, $\epsilon > 0$ represents the heterogeneous customers case. Similar to our discussion in the previous paragraph, the service provider is better off (worse off) in the homogeneous customers case, if the optimal equilibrium arrival rate is relatively high (low) in the heterogeneous customers case. According to Table 4.3, the service provider may regain or lose when a homogeneous population’s net value dispenses around the same benchmark net value,
if the optimal equilibrium arrival rate is at an intermediate level in the heterogeneous customers case.

It is also interesting to see the impact of the other problem parameters on the service provider’s optimal revenue. The following proposition is dedicated to this purpose.

**Proposition 4.9.** (i) The service provider’s optimal revenue is increasing in the capacity level, $\mu$, and it is decreasing in the waiting cost, $c_w$, and the benchmark delivery-reliability level, $r_b$. Moreover, as the reliability sensitivity, $c_r$, increases, the optimal revenue increases (decreases) if $r^* < r_b$ ($r^* > r_b$). (ii) The service provider’s optimal revenue is increasing in the customer heterogeneity level $\epsilon$, if $\frac{A}{2} \geq \mu$, and it is decreasing (increasing), if $\frac{A}{2} < \mu$ and $\epsilon > \epsilon_{\frac{A}{2}}$ ($\epsilon < \epsilon_{\frac{A}{2}}$), such that $\epsilon_{\frac{A}{2}} = \arg\{\frac{\partial R(\lambda e)}{\partial \lambda e}|_{\lambda e = \frac{A}{2}} = 0\}^3$.

Proposition 4.9 (i) is consistent with Proposition 4.4 for the homogeneous customers case. These results can be explained as we did in Subsection 4.3.3. Part (ii) of Proposition 4.9, however, is dedicated to the heterogeneous customers case. It implies that when the potential arrival rate is sufficiently high, i.e., $\frac{A}{2} > \mu$, the service provider benefits from markets with higher customer heterogeneity levels. However, when the potential arrival rate is sufficiently low, i.e., $\frac{A}{2} < \mu$, the service provider is better off in markets at which the customer heterogeneity level is either lower or higher than the threshold value $\epsilon_{\frac{A}{2}}$. In other words, for $\epsilon > \epsilon_{\frac{A}{2}}$, the service provider’s optimal revenue is increasing in the customer heterogeneity level regardless to the size of the potential arrival rate. However, for $0 \leq \epsilon < \epsilon_{\frac{A}{2}}$, the service provider’s optimal revenue is decreasing in the customer heterogeneity level, only if his capacity level is sufficiently high compared to the potential arrival rate.

### 4.4.3 Numerical Example

In this subsection, we show the importance of the integrated decision-making framework through a set of numerical examples. We compare two cases: (i) exogenous delivery-reliability level and (ii) endogenous delivery-reliability level. In case (i), we assume that the service provider sets the delivery-reliability level equal to the benchmark delivery-reliability level, i.e., $r_b$, and treats it as an exogenous variable. In this case, the service value function equals the benchmark net value, i.e., $V(r) = v_b$, and the sensitivity of the customers to the delivery reliability, i.e., $c_r$, is ignored. Since the delivery-reliability level is exogenous in this case, we use the subscript $exo$ to denote this case. In case (ii), the

\[^{3}\epsilon_{\frac{A}{2}} = v_b - c_w w(\frac{A}{2}) - \frac{c_w A}{2} \frac{\partial w(\lambda e)}{\partial \lambda e}|_{\lambda e = \frac{A}{2}} + c_r (\frac{1}{r_b} - \frac{1}{r(\frac{A}{2})}) + \frac{c_r A}{2 r(\frac{A}{2})} \frac{\partial r(\lambda e)}{\partial \lambda e}|_{\lambda e = \frac{A}{2}}.\]

\[^{4}\text{Note that for } \frac{A}{2} < \mu, \text{the threshold value } \epsilon_{\frac{A}{2}} \text{is not necessarily within the interval } [0, v_b]. \text{Thus, if } \epsilon_{\frac{A}{2}} < 0 \text{ (} \epsilon_{\frac{A}{2}} > v_b \text{), then the optimal revenue is increasing (decreasing) in the customer heterogeneity level.}\]
service provider adopts an integrated approach and treats the delivery reliability as an endogenous decision variable. We use the subscript \(\text{end}\) to denote this case.

Table 4.4 shows the percentage of the gap between the delivery-reliability levels, prices, delivery times, arrival rates, and revenues in the two above-mentioned cases. For instance, the percentage of the revenue gap is calculated as \(100\left(\frac{R_{\text{end}} - R_{\text{exo}}}{R_{\text{exo}}}\right)\), where \(R_{\text{end}}\) and \(R_{\text{exo}}\) are the service provider’s revenues in the endogenous and exogenous cases, respectively. The table shows the results for different customer heterogeneity levels, benchmark delivery-reliability levels, and delivery reliability sensitivities, and for \(\mu = 4, v_b = 4, c_w = 5\) and \(\Lambda = 0.5\). The potential arrival rate, \(\Lambda\), is set relatively low compared to the service rate, \(\mu\), in order to let the service provider reach higher market shares. Note that the customers are homogeneous when \(\epsilon = 0\), and \(\epsilon = 4\) is the maximum customer heterogeneity level for \(v_b = 4\).

<table>
<thead>
<tr>
<th>(\epsilon)</th>
<th>(v_b)</th>
<th>(c_r)</th>
<th>Delivery reliability</th>
<th>Price</th>
<th>Delivery time</th>
<th>Arrival rate</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.875</td>
<td>1</td>
<td>0</td>
<td>-36.4</td>
<td>112.2</td>
<td>-60.9</td>
<td>112.2</td>
</tr>
<tr>
<td>7.5</td>
<td>1</td>
<td>-1.8</td>
<td>1.2</td>
<td>-5.7</td>
<td>0.0</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>5.1</td>
<td>19.5</td>
<td>21.2</td>
<td>0.0</td>
<td>19.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.900</td>
<td>0</td>
<td>-38.1</td>
<td>202.9</td>
<td>-64.6</td>
<td>0.0</td>
<td>202.9</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>1</td>
<td>-4.5</td>
<td>13.1</td>
<td>-14.8</td>
<td>0.0</td>
<td>13.1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.2</td>
<td>6.1</td>
<td>9.4</td>
<td>0.0</td>
<td>6.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.925</td>
<td>0</td>
<td>-39.8</td>
<td>429.1</td>
<td>-67.7</td>
<td>0.0</td>
<td>559.3</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>1</td>
<td>-7.1</td>
<td>44.2</td>
<td>-22.1</td>
<td>24.6</td>
<td>79.7</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-0.4</td>
<td>-0.6</td>
<td>-1.7</td>
<td>11.1</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 2 | 0.875 | 1 | 0 | -35.1 | 27.3 | -58.9 | 33.7 |
| 7.5 | 1 | -0.8 | 0.0 | -2.6 | 0.2 | 0.1 |
| 15 | 5.7 | 8.5 | 24.9 | 6.3 | 15.4 |
| 0.900 | 0 | -36.9 | 37.7 | -62.7 | 45.8 | 100.7 |
| 7.5 | 1 | -3.5 | 1.3 | -11.9 | 2.3 | 3.6 |
| 15 | 2.8 | 2.9 | 12.7 | 1.8 | 4.8 |
| 0.925 | 0 | -38.6 | 54.5 | -66.7 | 65.3 | 155.4 |
| 7.5 | 1 | -6.1 | 6.3 | -21.4 | 8.6 | 15.4 |
| 15 | 0.1 | 0.0 | 0.4 | 0.0 | 0.0 |

| 4 | 0.875 | 1 | 0 | -34.7 | 17.2 | -58.9 | 20.2 |
| 7.5 | 1 | -0.7 | 0.0 | -2.3 | 0.1 | 0.1 |
| 15 | 5.8 | 5.4 | 25.1 | 4.3 | 9.8 |
| 0.900 | 0 | -36.5 | 23.0 | -62.8 | 26.7 | 55.8 |
| 7.5 | 1 | -3.4 | 0.8 | -11.7 | 1.3 | 2.1 |
| 15 | 2.9 | 1.8 | 13.0 | 1.2 | 3.0 |
| 0.925 | 0 | -38.2 | 31.6 | -66.9 | 36.3 | 79.4 |
| 7.5 | 1 | -6.0 | 3.7 | -21.3 | 4.8 | 8.6 |
| 15 | 0.1 | 0.0 | 0.5 | 0.0 | 0.0 |

Table 4.4: The impact of ignoring the reliability decision on the service provider’s optimal decisions and revenue

Apparently, the integrated decision making approach of case (ii) yields a higher revenue compared to case (i) due to the sub-optimality of the latter case. The last column
of Table 4.4 shows the percentage of the revenue gap between the two cases. As one may notice, the gap varies in a wide spectrum from 0.0% percent for \((\epsilon, r_b, c_r) = (4, 0.925, 15)\) to 559.3% for \((\epsilon, r_b, c_r) = (0, 0.925, 1)\). The average revenue gap is 57.8% for this set of examples. These results emphasize the fact that the service provider’s revenue loss may be substantial when the delivery reliability is not set appropriately based on the market characteristics and shed light on the importance of an integrated decision making approach.

4.5 Conclusion and Future Research

The classical queue-pricing literature dominantly reduces the delivery performance to a single measure; the delivery time. However, the delivery reliability as another measure of the delivery performance is foregrounded in the literature. The interactions between price, delivery time, and delivery reliability are well studied in empirical research, and the importance of an integrated framework to analytically study these interactions is emphasized in the literature. In this study, we address the equilibrium behavior of rational, self-interested, price-, delivery-time-, and delivery-reliability-sensitive customers and the optimal decisions of a revenue-maximizing service provider dealing with such customers. We consider two cases: homogeneous and heterogeneous customers. For each case, we characterize the customers’ equilibrium behavior and the service provider’s optimal decisions. In either case, interesting results are obtained.

**Homogeneous customers case.** We find that, in the case of partial market coverage, the optimal reliability level lies within a narrow interval, i.e., \(r^* \in (r_\lambda, r_p) \subseteq [r_1, r_2] \subseteq [0, 1]\), implying that the service provider should offer an intermediate delivery-reliability level rather than an extreme value. Moreover, we show that variations of the reliability sensitivity could have either positive or negative impact on the service provider’s revenue depending on his current delivery-reliability performance compared to the benchmark value. Our results also show that if the service provider can improve his delivery performance in both dimensions (for instance, by increasing the service rate), he can also charge the customers with a higher price. This result is consistent with the empirical results reported in the literature regarding the positive impact of the delivery performance on the price and on the customers’ willingness to pay. Furthermore, based on our findings, the service provider is better off in markets where the waiting cost rate is small, the benchmark net value is high, and the benchmark delivery-reliability level is low.
**Heterogeneous customers case.** Our results are in general consistent with those of the homogeneous customers case. Assuming that the customers’ net values are drawn from a uniform distribution, we show that for a fixed customer heterogeneity level, as the mean of benchmark net value increases, the service provider’s optimal revenue increases, and he charges a higher price for an inferior delivery performance. Furthermore, as the customer heterogeneity level decreases, the optimal delivery performance declines (improves), and surprisingly the service provider’s optimal revenue increases (decreases), if the current equilibrium arrival rate is sufficiently high (low). At an intermediate equilibrium arrival rate, however, the service provider may either gain or lose revenue if the customer heterogeneity level decreases, but in either case he charges a lower price and offers a slower and less reliable service. We determine a threshold value for the customer heterogeneity level above (below) which the service provider’s optimal revenue is increasing (decreasing) in the customer heterogeneity level.

**Future studies.** In this study, we modeled the operations of the service provider as an $M/M/1$ queue. By considering a sufficiently high lower bound for the delivery-reliability level, the distribution of the waiting time can be accurately approximated by the exponential distribution even for arrival and service processes with general distributions (Abate et al., 1996). A problem with such a constraint can be dealt with in a similar way with slight adjustments, but it would also be interesting to see how a more general queueing system, i.e., $G/G/k$, changes the results. Another possible direction is to capture the customer heterogeneity through the reliability sensitivity. The design of the price, delivery time, and delivery reliability menus and determining the optimal scheduling policy for serving such customers would be interesting and also challenging. When the service provider uses a common server to serve different streams of customers, it would be challenging to determine the waiting time distribution of the customers with lower priority. It is also interesting to compare the performance of the different pricing schemes, e.g., fixed- and differentiated-pricing schemes, when the customers are heterogeneous in their reliability sensitivities. Another interesting direction is to study the price, delivery time, and delivery reliability competition among independent service providers.
Appendix

4.A Proofs

**Proof of Lemma 4.1.** $V(r)$ is an increasing concave function of $r$. $c_w w(0,r)$ is an increasing convex function of $r$. Moreover, $c_w w(0,r)$ is strictly increasing in $c_w$ and strictly decreasing in $\mu$. $r_1$ and $r_2$, $r_1 < r_2$, are the two points at which $V(r)$ and $c_w w(0,r)$ cross each other. For $r = r_1$ and $r = r_2$, we have $V(r) = c_w w(0,r)$, and for any $r \in (r_1, r_2)$, $V(r)$ is above $c_w w(0,r)$, i.e., $V(r) > c_w w(0,r)$. It is clear that as $\mu$ increases ($c_w$ decreases), $c_w w(0,r)$ shifts downward and crosses $V(r)$ at new points, $r'_1$ and $r'_2$ such that $r'_1 < r_1$ and $r'_2 > r_2$. Thus, the feasible region of the delivery-reliability level becomes wider. \hfill \Box

**Proof of Lemma 4.2.** For any $r \in (r_1, r_2)$, the service provider can charge customers price $p$, $p < V(r)$, and still leave a non-negative utility. It is well-known that the revenue-maximizing service provider extracts all customer surplus at the equilibrium (Hassin and Haviv, 2003). Thus, we have $U(p, \lambda, r) = 0$, which results in $\tilde{\lambda}(p, r) = \mu - \frac{c_w}{V(r)-p} \ln(\frac{1}{1-r})$. If $\tilde{\lambda}(p, r) < 0$, all of the customers would choose to balk. If the potential arrival rate, $\Lambda$, is greater than $\tilde{\lambda}(p, r)$, then only $\frac{\tilde{\lambda}(p,r)}{\Lambda}$ of the customers would choose to join the service provider. Otherwise, all of the customers would choose to join the service provider. \hfill \Box

**Proof of Lemma 4.3.** According to Lemma 4.2, for $r \in [r_1, r_2]$, we have

$$\lambda^c(p, r) = \begin{cases} 
\Lambda & \text{if } 0 \leq p \leq V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1-r}), \\
\tilde{\lambda}(p, r) & \text{if } V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1-r}) < p \leq V(r) - \frac{c_w}{\mu} \ln(\frac{1}{1-r}), \\
0 & \text{if } V(r) - \frac{c_w}{\mu} \ln(\frac{1}{1-r}) < p,
\end{cases}$$
Finding the maximum of \( r \) for any \( x \) for positive values of \( r \).

**Proof of Proposition 4.1**

Consider the following cases:

(i) \( R(p, r) = p\Lambda \): \( p\Lambda \) reaches its maximum at \( p(r) = V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) \).

(ii) \( R(p, r) = p\hat{\lambda}(p, r) \): It can be shown that \( p\hat{\lambda}(p, r) \) is concave with respect to \( p \), i.e., \( \frac{\partial^2 p\hat{\lambda}(p, r)}{\partial p^2} = -\frac{2c_w V(r)}{(V(r) - p)^2} \ln(\frac{1}{1 - r}) < 0 \). The first-order condition, with respect to \( p \), yields \( p(r) = V(r) - \sqrt{\frac{c_w V(r)}{\mu - \Lambda} \ln(\frac{1}{1 - r})} \). Substituting \( p(r) \) in \( \hat{\lambda}(p, r) \) results in \( \hat{\lambda}(p(r), r) = \hat{\lambda}(r) \).

If \( V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) < p(r) \leq V(r) - \frac{c_w}{\mu} \ln(\frac{1}{1 - r}) \) (or equivalently \( 0 \leq \hat{\lambda}(r) < \Lambda \)), then \( R(p, r) = p\hat{\lambda}(p, r) \) reaches its maximum at \( p(r) = V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1 - r})} \).

Taking into consideration that \( p\Lambda = p\hat{\lambda}(p, r) \) when \( p = V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) \) (or equivalently \( \hat{\lambda}(r) = \Lambda \)), a logical conclusion is that if \( V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) < V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1 - r})} \leq V(r) - \frac{c_w}{\mu} \ln(\frac{1}{1 - r}) \) (or equivalently \( 0 \leq \hat{\lambda}(r) < \Lambda \)), then the optimal price equals \( p(r) = V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1 - r})} \); otherwise, if \( V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1 - r})} \leq V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) \) (or equivalently \( \Lambda \leq \hat{\lambda}(r) \)), then the optimal price equals \( p(r) = V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}) \). The rest of the results presented in the lemma are immediately resulted from the optimal pricing decision.

\( \square \)

**Proof of Proposition 4.1.** First consider the two revenue functions determined in Eq. (4.7):

(i) \( R(r) = \Lambda(V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r})) \): \( \Lambda(V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r})) \) is concave with respect to \( r \), i.e., \( \frac{\partial^2 \Lambda(V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r}))}{\partial r^2} = \frac{\partial^2 V(r)}{\partial r^2} - \frac{c_w}{(\mu - \Lambda)(1 - r)^2} < 0 \), and reaches its maximum at \( r_{full} = \arg\{\frac{\partial}{\partial r}(V(r) - \frac{c_w}{\mu - \Lambda} \ln(\frac{1}{1 - r})) = 0\} \).

(ii) \( R(r) = (\sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1 - r})})^2 \): From Subsection 4.3.2, we know that \( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1 - r})} \geq 0 \) for any \( r \in (r_1, r_2) \), and \( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1 - r})} = 0 \) for \( r \in \{r_1, r_2\} \).

Finding the maximum of \( (\sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1 - r})})^2 \) corresponds to finding the maximum of \( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1 - r})} \), because the square function \( (f(x) = x^2) \) is strictly increasing for positive values \( (x \in \mathbb{R}^+) \). \( \sqrt{\mu V(r)} \) is increasing and concave with respect to \( r \), i.e., \( \frac{\partial\sqrt{\mu V(r)}}{\partial r} = \sqrt{\mu} \frac{V(r)^{-1/2} \partial V(r)}{2} > 0 \) and \( \frac{\partial^2 \sqrt{\mu V(r)}}{\partial r^2} = \sqrt{\mu} \left(\frac{V(r)^{-1/2} \partial^2 V(r)}{2} - \frac{V(r)^{-3/2}}{4} (\partial V(r))^2\right) < 0 \) for any \( r \in (r_1, r_2) \). In addition, \( \sqrt{c_w \ln(\frac{1}{1 - r})} \) is increasing with respect to \( r \), i.e.,
Then, equal the potential arrival rate, and the optimal delivery-reliability level equals $r$ constraint holds. If $0 < \hat{r} < \infty$ to determine the optimal delivery-reliability level, it suffices to check whether this con-

$\hat{r}$ maximizing delivery-reliability level is equal to $\hat{r}$

$V$ is strictly convex and positive, and

Equation (4.8). In this case, the corresponding optimal price equals $r$

Otherwise, if $\hat{r}$

order condition.

As mentioned before, when the delivery-reliability level is not bounded by the avail-

ability of customers (potential arrival rate $\Lambda$), i.e., $0 \leq \hat{\lambda}(r_{\text{part}}) < \Lambda$, the unique revenue-

maximizing delivery-reliability level is equal to $r_{\text{part}}$ as determined by Eq. (4.9). Thus,

to determine the optimal delivery-reliability level, it suffices to check whether this con-

straint holds. If $0 \leq \hat{\lambda}(r_{\text{part}}) < \Lambda$, then the delivery-reliability level is not bounded by the potential

arrival rate, and the optimal delivery-reliability level equals $r^* = r_{\text{part}}$.

According to Lemma 4.3, the corresponding optimal price and equilibrium arrival rate,

then, equal $p^* = V(r^*) - \sqrt{c_w V(r^*) \ln(\frac{1}{1-r^*})}$ and

$\lambda^* = \mu - \sqrt{\frac{c_w}{V(r^*)} \ln(\frac{1}{1-r^*})}$, respectively. Otherwise, if $\hat{\lambda}(r_{\text{part}}) \geq \Lambda$, then the delivery-reliability level is bounded by the potential

arrival rate, and the optimal delivery-reliability level is determined by $r^* = r_{\text{full}}$ as in

Eq. (4.8). In this case, the corresponding optimal price equals $p^* = V(r^*) - \frac{c_w}{\mu \Lambda} \ln(\frac{1}{1-r^*})$, and all the customers would buy the service, i.e., $\lambda^* = \Lambda$ as elaborated in Lemma 4.3.

Proof of Proposition 4.2. First, we show the pseudo-concavity of $p(r)$ and $\lambda^*(r)$. According to Lemma 4.3 for $\hat{\lambda}(r) < \Lambda$, the price equals $p(r) = V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1-r})}$. $V(r) \ln(\frac{1}{1-r})$ is strictly increasing and positive for $r \in (r_1, r_2)$. Moreover, the square root function ($f(x) = \sqrt{x}$) is strictly convex and increasing. Thus, for any $r_1', r_2' \in (r_1, r_2)$, and from $\frac{1}{2} \sqrt{c_w V(r) \ln(\frac{1}{1-r^*})}^{-1/2} (r_2' - r_1') \frac{\partial}{\partial r} (V(r) \ln(\frac{1}{1-r})) \big|_{r = r_1'} \geq 0$, it can be concluded

that $\sqrt{\frac{c_w V(r_2') \ln(1-\frac{1}{1-r^*})}{\mu}} > \sqrt{\frac{c_w V(r_1') \ln(1-\frac{1}{1-r^*})}{\mu}}$, and hence $\sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1-r})}$ is strictly pseudo-

convex with respect to $r$ over the interval $(r_1, r_2)$. Furthermore, $V(r)$ is positive and

strictly concave (and hence strictly pseudo-concave) with respect to $r$ over the interval

$(r_1, r_2)$. We can conclude that $p(r) = V(r) - \sqrt{\frac{c_w V(r)}{\mu} \ln(\frac{1}{1-r})}$ is strictly pseudo-concave with respect to $r$ over the interval $(r_1, r_2)$.

For $\hat{\lambda}(r) < \Lambda$, the equilibrium arrival rate equals $\lambda^*(r) = \mu - \sqrt{\frac{c_w}{V(r)} \ln(\frac{1}{1-r})}$. $\ln(\frac{1}{1-r})$ is strictly convex and positive, and $V(r)$ is strictly concave and positive with respect to

102
r over the interval \((r_1, r_2)\). According to Stancu-Minasian (2012), Chapter 2, Subsection 5.2, \(\frac{c_w}{V(r)} \ln(\frac{1}{1-r})\) is strictly pseudo-convex with respect to \(r\) over the interval \((r_1, r_2)\). In other words, for any \(r_1', r_2' \in (r_1, r_2)\), and from \((r_2' - r_1') \frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right) \mid_{r=r_1'} \geq 0\), it can be concluded that \(\frac{c_w}{V(r)} \ln(\frac{1}{1-r_2'}) > \frac{c_w}{V(r')} \ln(\frac{1}{1-r_1'})\).

Now, consider \(\frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right)\). We have

\[
\frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right) = \frac{1}{2} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right)^{-1/2} \frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right).
\]

Thus, for any \(r_1', r_2' \in (r_1, r_2)\), the condition \((r_2' - r_1') \frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right) \mid_{r=r_1'} \geq 0\) equals to \((r_2' - r_1') \frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right) \mid_{r=r_1'} \geq 0\). Therefore, for any \(r_1', r_2' \in (r_1, r_2)\), and from \((r_2' - r_1') \frac{\partial}{\partial r} \left( \frac{c_w}{V(r)} \ln(\frac{1}{1-r}) \right) \mid_{r=r_1'} \geq 0\), we can conclude that \(\frac{c_w}{V(r_2')} \ln(\frac{1}{1-r_2'}) > \frac{c_w}{V(r_1')} \ln(\frac{1}{1-r_1'})\), and therefore \(\frac{c_w}{V(r)} \ln(\frac{1}{1-r})\) is pseudo-concave. It can then be concluded that \(\lambda^*(r) = \mu - \sqrt{\frac{c_w}{V(r)} \ln(\frac{1}{1-r})}\) is a pseudo-concave function with respect to \(r\) over the interval \((r_1, r_2)\).

Moreover, since \(\lambda^*(r_1) = \lambda^*(r_2) = 0\, \lambda^*(r)\) is also unimodal in \((r_1, r_2)\).

Now, we show the relationships between \(r^*\), \(r_\lambda\), and \(r_p\). Define \(f(r) = \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1-r})}\), \(g(r) = \sqrt{\frac{V(r)}{\mu}}\), and \(h(r) = \sqrt{\frac{V(r)}{\mu}}\), by which we have \(p(r) = f(r)g(r)\) and \(\lambda^*(r) = f(r)h(r)\). \(r^*\) is the unique root of \(\frac{\partial f(r)}{\partial r} = 0\). Furthermore, the maximum points of \(p(r)\) and \(\lambda^*(r)\) are the unique roots of the following equation

\[
\frac{\partial f(r)}{\partial r} + \frac{f(r)}{g(r)} \frac{\partial g(r)}{\partial r} = 0, \tag{4.16}
\]

\[
\frac{\partial f(r)}{\partial r} + \frac{f(r)}{h(r)} \frac{\partial h(r)}{\partial r} = 0, \tag{4.17}
\]

respectively. Since \(f(r), g(r),\) and \(h(r)\) are all positive for \(r \in (r_1, r_2)\), \(\frac{\partial g(r)}{\partial r}\) is positive, i.e., \(\frac{\partial g(r)}{\partial r} = \frac{V(r)^{1/2} \frac{\partial V(r)}{\partial r}}{2 \sqrt{\mu}} > 0\), and \(\frac{\partial h(r)}{\partial r}\) is negative, i.e., \(\frac{\partial h(r)}{\partial r} = -\frac{\sqrt{\mu V(r)^{1/2}} \frac{\partial V(r)}{\partial r}}{2} < 0\), we can conclude that the root of Eq. (4.16), \(r_p\), is greater than \(r^*\), and the root of Eq. (4.17), \(r_\lambda\), is smaller than \(r^*\).

**Proof of Proposition 4.3.** Part (i): From Proposition 4.1, we know that \(r_{fall}\) is the intersection point of \(\frac{\partial V(r)}{\partial r}\) and \(\frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) within \((r_1, r_2)\). Moreover, we know that \(\frac{\partial V(r)}{\partial r}\) is decreasing and \(\frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) is increasing in \(r\). We also know that \(\frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) is increasing in \(\lambda\). If \(\lambda\) increases, \(\frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) moves upward, and \(\frac{\partial V(r)}{\partial r}\) and \(\frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) cross each other at a point lower than \(r_{fall}\). Therefore, the optimal reliability level is decreasing in \(\lambda\). In addition, in the full market coverage case, the optimal price equals \(p(r) = V(r) - \frac{c_w}{\mu - \lambda} \ln(\frac{1}{1-r})\). We have \(\frac{\partial p(r)}{\partial \lambda} = \left( \frac{\partial V(r)}{\partial r} - \frac{c_w}{\mu - \lambda} \frac{1}{1-r} \right) \frac{\partial r}{\partial \lambda} - \frac{c_w}{(\mu - \lambda)^2} \ln(\frac{1}{1-r}) < 0\) since \(\frac{\partial V(r)}{\partial r} - \frac{c_w}{\mu - \lambda} \frac{1}{1-r}\) equals zero

103
at optimality.

From Proposition 1, \( \frac{1}{\mu - \Lambda} = \frac{c_r(1-r)}{c_{w}r^2} \) at optimality. Thus, the delivery time equals \( w = \frac{1}{\mu - \Lambda} \ln(\frac{1}{1-r}) = \frac{c_r(1-r)}{c_{w}r^2} \ln(\frac{1}{1-r}) \). We have \( \frac{\partial w}{\partial r} = -\frac{c_r(1-r)^2}{\mu - \Lambda r^2} \ln(\frac{1}{1-r}) + \frac{1}{\mu - \Lambda} < \frac{r}{r^2} < 0 \). The first inequality holds, because \( \ln(x) < x - 1 \). We can conclude that \( \frac{\partial w}{\partial \Lambda} = \frac{\partial w}{\partial r} \frac{\partial r}{\partial \Lambda} > 0 \) because \( \frac{\partial w}{\partial r} \) and \( \frac{\partial r}{\partial \Lambda} \) are both negative.

Part (ii): From Proposition 4.1, we know that \( r_{full} \) is the unique root of the equation \( \frac{\partial V(r)}{\partial r} = 0 \). Therefore, \( \frac{\partial c}{c_{w}} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) is strictly increasing, and \( \frac{\partial c}{c_{w}} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) stands below \( \frac{\partial w}{\partial r} \) at point \( r = r_{part} \). Due to the facts that \( \frac{\partial V(r)}{\partial r} \) is strictly decreasing, and \( \frac{\partial c}{c_{w}} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) is strictly increasing, they cross each other at a point greater than \( r_{part} \), i.e., \( r_{full} \). Therefore, \( r_{part} < r_{full} \). In words, the full-market coverage’s optimal delivery-reliability level is greater than the partial-market coverage’s optimal delivery-reliability level.

From Proposition 4.4, we know that \( r_{part} < r_{full} \). From Proposition 1, we have \( \frac{\partial V(r)}{\partial r} = \frac{\partial c}{c_{w}} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \). In words, \( V(r) - \frac{c_r(1-r)}{c_{w}r^2} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) stands above \( V(r) - \frac{c_r(1-r)}{c_{w}r^2} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) at point \( r = r_{part} \). Moreover, \( V(r) - \frac{c_r(1-r)}{c_{w}r^2} \frac{1}{\mu - \Lambda} \frac{1}{1-r} \) reaches its maximum at \( r = r_{full} \). Hence, \( \frac{\partial V(r)}{\partial r} \) crossing each other at a point greater than \( r_{part} \), i.e., \( r_{full} \). In words, the full-market coverage’s optimal price is greater than the partial-market coverage’s optimal price.

At optimality, \( \frac{1}{\mu - \Lambda} = \frac{c_r(1-r)}{c_{w}r^2} \). Thus, the delivery time equals \( w = \frac{1}{\mu - \Lambda} \ln(\frac{1}{1-r}) = \frac{c_r(1-r)}{c_{w}r^2} \ln(\frac{1}{1-r}) \). We have \( \frac{\partial w}{\partial \Lambda} > 0 \). The first inequality holds, because \( \ln(x) < x - 1 \). Thus, as long as \( c_r \) and \( c_w \) are fixed, a higher delivery-reliability level brings about a shorter delivery time. Since \( r_{part} < r_{full} \), the delivery time is shorter in the full market coverage case compared to the partial market coverage case.

**Proof of Proposition 4.4.** First, we prove the results for the full market coverage case and then for the partial market coverage case.

**Full market coverage.** According to Proposition 4.1, in the full market coverage case, \( r \) is the unique root of the equation \( \frac{\partial V}{\partial r}(V(r) - \frac{c_r(1-r)}{c_{w}r^2} \ln(\frac{1}{1-r})) = 0 \) or equivalently \( \frac{c_r(1-r)}{c_{w}r^2} \ln(\frac{1}{1-r}) = 0 \). Let \( f \) denote the left-hand side of this equation, i.e., \( f = \frac{c_r(1-r)}{c_{w}r^2} - \frac{c_r}{\mu - \Lambda}(\frac{1}{1-r}) \). From the Proof of Proposition 4.1, we have \( \frac{\partial f}{\partial r} < 0 \).

**Sensitivity to \( c_r \):** we have \( \frac{\partial f}{\partial c_r} = \frac{1}{r^2} > 0 \). Thus, \( \frac{\partial r}{\partial c_r} = -\frac{\partial f}{\partial c_r} > 0 \). Consequently, \( w(r) = \frac{1}{\mu - \Lambda} \ln(\frac{1}{1-r}) \). Hence, \( \frac{\partial w}{\partial c_r} = \frac{1}{\mu - \Lambda(1-r)} \ln(\frac{1}{1-r}) > 0 \). Moreover, \( \frac{\partial p}{\partial c_r} = \frac{1}{\mu - \Lambda} \ln(\frac{1}{1-r}) - \frac{c_r}{c_{w}r^2} \frac{1}{1-r} > 0 \). Thus, as long as the optimal delivery-reliability level is greater (less) than the benchmark value, the optimal price is increasing (decreasing) in the reliability sensitivity.

**Sensitivity to \( c_w \):** we have \( \frac{\partial f}{\partial c_w} = -\frac{\partial f}{\partial \Lambda} < 0 \). Thus, \( \frac{\partial r}{\partial c_w} = -\frac{\partial f}{\partial \Lambda} < 0 \). Moreover, we
have \( \frac{\partial w(r)}{\partial r} = \frac{\mu - \lambda}{\mu - \lambda - 1(r)} \frac{\partial r}{\partial w} < 0 \). From \( f = 0 \), we have \( \frac{1}{\mu - \lambda} = \frac{c_r(1 - r)}{c_r r^2} \) at optimality. Thus, the price equals \( p(r) = V(r) - \frac{c_r(1 - r)}{r^2} \ln(\frac{1}{1 - r}) \). We have \( \frac{\partial p(r)}{\partial c_w} = \frac{c_r}{r^2} \frac{\partial r}{\partial c_w} - cr \left( \frac{2 - r}{r^4} \ln(1 - r) + \frac{1}{r^2} \right) \frac{\partial r}{\partial c_w} \). In Proposition 4.3, we showed that \( \left( \frac{2 - r}{r^4} \ln(1 - r) + \frac{1}{r^2} \right) \frac{\partial r}{\partial c_w} < 0 \), thus, \( \frac{\partial p(r)}{\partial c_w} < 0 \).

**Sensitivity to \( \mu \):** we have \( \frac{\partial f}{\partial \mu} = \frac{c_r}{(\mu - \lambda)^2} \ln(\frac{1}{1 - r}) > 0 \). Thus, \( \frac{\partial r}{\partial \mu} = -\frac{\partial f}{\partial \mu} > 0 \). Moreover, we have \( \frac{\partial w(r)}{\partial \mu} = \frac{c_r}{r^2} \frac{\partial r}{\partial \mu} < 0 \). Furthermore, the price equals \( p(r) = V(r) - \frac{c_r(1 - r)}{r^2} \ln(\frac{1}{1 - r}) \) at optimality. Thus, we have \( \frac{\partial p(r)}{\partial \mu} = \frac{c_r}{r^2} \frac{\partial r}{\partial \mu} - cr \left( \frac{2 - r}{r^4} \ln(1 - r) + \frac{1}{r^2} \right) \frac{\partial r}{\partial \mu} \). In Proposition 4.3, we showed that \( \left( \frac{2 - r}{r^4} \ln(1 - r) + \frac{1}{r^2} \right) \frac{\partial r}{\partial \mu} < 0 \), thus, \( \frac{\partial p(r)}{\partial \mu} > 0 \).

**Sensitivity to \( v_b \) and \( r_b \):** Since \( f \) is independent of \( v_b \) and \( r_b \), \( r \) and subsequently \( w(r) \) are invariant with respect to these two parameters. However, we have \( \frac{\partial w(r)}{\partial v_b} = 1 > 0 \), and \( \frac{\partial w(r)}{\partial r_b} = -\frac{c_r}{r_b^2} < 0 \).

It is apparent that the direction of the revenue is positively correlated with that of the price, since in the full market coverage the equilibrium arrival rate remains unchanged for the parameter variations mentioned in Table 4.2.

**Partial market coverage.** Now, we prove the results for the partial market coverage case. According to Proposition 4.1, in the partial market coverage case, \( r \) is the unique root of the \( \frac{d}{dr} \left( \sqrt{\mu V(r)} - \sqrt{c_r \ln(\frac{1}{1 - r})} \right) = 0 \) or equivalently \( \frac{d}{dr} \left( \frac{c_r}{r^2} \ln(\frac{1}{1 - r}) \right) = 0 \). Let \( f \) denote the left-hand side of this equation, i.e., \( f = \frac{c_r}{r^2} - \frac{c_r}{\mu - \lambda(r)} (\frac{1}{1 - r}) = 0 \). Let \( \hat{r} \) be the root of the equation, \( f \), and \( \hat{r} \) satisfies \( \frac{d}{dr} \left( \sqrt{\mu V(r)} - \sqrt{c_r \ln(\frac{1}{1 - r})} \right) = 0 \) and \( \frac{d}{dr} \left( \frac{c_r}{r^2} \ln(\frac{1}{1 - r}) \right) = 0 \), in which \( \sqrt{\frac{\mu V(r)}{r^2}} \) is decreasing at the intersection point. As \( c_r \) increases, \( \sqrt{\frac{\mu V(r)}{r^2}} \) moves upward and \( \sqrt{\frac{c_r \ln(\frac{1}{1 - r})}{r^2}} \) moves downward. Consequently, they cross each other at a larger \( r \). Thus, \( \frac{\partial r}{\partial c_r} > 0 \). Moreover, we have \( \frac{\partial R^*}{\partial c_w} = \frac{1}{2} \left( \sqrt{\mu V(r)} - \sqrt{c_r \ln(\frac{1}{1 - r})} \right)^{-1/2} \frac{\sqrt{\mu V(r)} (\frac{1}{r^2} - \frac{1}{r^2})}{c_r} \) at optimality. Hence, as long as the optimal \( r \) is less (greater) than \( r_b \), the optimal revenue is decreasing (increasing) in \( c_r \).

**Sensitivity to \( c_w \):** The optimal \( r \) is the intersection of \( \frac{\partial V(r)}{\partial r} = \frac{c_r}{r^2} \) and \( \frac{c_r}{\mu - \lambda(r)} (\frac{1}{1 - r}) = \sqrt{\frac{c_r V(r)}{\mu \ln(\frac{1}{1 - r})}} \), or equivalently the intersection of \( \frac{\sqrt{\mu V(r)}}{r^2} \) and \( \sqrt{\frac{c_r \ln(\frac{1}{1 - r})}{r^2}} \), in which \( \frac{\partial V(r)}{\partial r} \) is decreasing at the intersection point. As \( c_w \) increases, \( \frac{\partial V(r)}{\partial r} \) moves upward, but \( \frac{\partial V(r)}{\partial r} \) moves downward and \( \frac{\partial V(r)}{\partial r} \) moves upward and crosses \( \frac{\partial V(r)}{\partial r} \) at a smaller \( r \). Thus, \( \frac{\partial r}{\partial c_w} < 0 \). In the partial market coverage, we have \( w(r) = \frac{1}{\mu - \lambda(r)} \ln(\frac{1}{1 - r}) = \sqrt{\frac{V(r) \ln(\frac{1}{1 - r})}{c_r \mu}} \) for which \( \frac{\partial w(r)}{\partial r} > 0 \). Thus, \( w(r) \) is strictly increasing in \( r \). Increasing \( c_w \) shifts \( w(r) \) down, since \( \frac{\partial w(r)}{\partial c_w} < 0 \). From the previous part, we also know that the optimal delivery-reliability level decreases as \( c_w \) increases. Thus, the corresponding waiting time for an increased \( c_w \) would be lower, \( \frac{\partial w(r)}{\partial c_w} < 0 \). From \( f = 0 \), we know \( \frac{1}{\mu - \lambda(r)} = \frac{c_r(1 - r)}{c_r r^2} \) at optimality.
Thus, the price equals \( p(r) = V(r) - \frac{c_w}{r} \ln(\frac{1}{1-r}) \) at optimality. We have \( \frac{\partial p(r)}{\partial c_w} = \frac{c_w}{r} \frac{\partial r}{\partial c_w} - c_r \left( \frac{2}{r^2} \ln(1-r) + \frac{1}{r} \right) \frac{\partial r}{\partial c_w} \). In Proposition 4.3, we showed that \( \left( \frac{2}{r^2} \ln(1-r) + \frac{1}{r} \right) \frac{\partial r}{\partial c_w} < 0 \), thus, \( \frac{\partial p(r)}{\partial c_w} < 0 \). Moreover, we have \( \frac{\partial R}{\partial c_w} = -\frac{1}{4} \left( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1-r})} \right)^{-1/2} \frac{\ln(\frac{1}{1-r})}{c_w} < 0 \) at optimality.

**Sensitivity to \( \mu \):** Similar to the previous case, as \( \mu \) increases, \( \frac{\partial V(r)}{\partial r} \) does not change, but \( \frac{c_w}{r} \ln(\frac{1}{1-r}) \) moves downward and crosses \( \frac{\partial V(r)}{\partial r} \) at a larger \( r \). Thus, \( \frac{\partial r}{\partial \mu} > 0 \). Moreover, we know \( w(r) = \frac{c_w r}{c_w r^2} \ln(\frac{1}{1-r}) \) at optimality. We know \( \frac{\partial w(r)}{\partial c_w} < 0 \). Thus, we have \( \frac{\partial w(r)}{\partial \mu} = \frac{\partial w(r)}{\partial c_w} \frac{\partial c_w}{\partial \mu} < 0 \). Furthermore, the price equals \( p(r) = V(r) - \frac{c_w(1-r)}{r} \ln(\frac{1}{1-r}) \) at optimality. Thus, we have \( \frac{\partial p(r)}{\partial \mu} = \frac{\partial w(r)}{\partial \mu} = \frac{c_w}{r} \frac{\partial r}{\partial \mu} < 0 \). In Proposition 4.3, we showed that \( \left( \frac{2}{r^2} \ln(1-r) + \frac{1}{r} \right) \frac{\partial r}{\partial \mu} < 0 \), thus, \( \frac{\partial p(r)}{\partial \mu} > 0 \). In addition, we have \( \frac{\partial R}{\partial \mu} = \frac{1}{4} \left( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1-r})} \right)^{-1/2} \frac{\mu V(r)}{c_w} > 0 \) at optimality.

**Sensitivity to \( v_b \):** Through a similar discussion, as \( v_b \) increases, \( \frac{\partial V(r)}{\partial r} \) does not change, but \( \frac{c_w}{\mu - \lambda(r) \ln(\frac{1}{1-r})} \) moves upward and crosses \( \frac{\partial V(r)}{\partial r} \) at a smaller \( r \). Thus, \( \frac{\partial r}{\partial v_b} < 0 \). Moreover, we know \( w(r) = \frac{c_w r}{c_w r^2} \ln(\frac{1}{1-r}) \) at optimality. We know \( \frac{\partial w(r)}{\partial v_b} < 0 \). Thus, we have \( \frac{\partial w(r)}{\partial \mu} = \frac{\partial w(r)}{\partial v_b} \frac{\partial v_b}{\partial \mu} < 0 \). Furthermore, the equilibrium arrival rate equals \( \lambda^e(r) = \mu - \frac{c_w r^2}{c_w r^2} \) at optimality. We have \( \frac{\partial \lambda^e(r)}{\partial v_b} = \frac{c_w r}{c_w r^2} < 0 \). Thus, \( \frac{\partial \lambda^e(r)}{\partial \mu} = \frac{\partial \lambda^e(r)}{\partial v_b} \frac{\partial v_b}{\partial \mu} < 0 \). Similar to the previous cases, we have \( \frac{\partial R}{\partial \mu} = \frac{1}{4} \left( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1-r})} \right)^{-1/2} \frac{\mu V(r)}{c_w} > 0 \) at optimality.

**Sensitivity to \( r_b \):** In a similar way, as \( r_b \) increases, \( \frac{\partial V(r)}{\partial r} \) does not change, but \( \frac{c_w}{\mu - \lambda(r) \ln(\frac{1}{1-r})} \) moves downward and crosses \( \frac{\partial V(r)}{\partial r} \) at a larger \( r \). Thus, \( \frac{\partial r}{\partial r_b} > 0 \). Moreover, we know \( w(r) = \frac{c_w r}{c_w r^2} \ln(\frac{1}{1-r}) \) at optimality. We know \( \frac{\partial w(r)}{\partial r_b} < 0 \). Thus, we have \( \frac{\partial w(r)}{\partial r_b} = \frac{\partial w(r)}{\partial r_b} \frac{\partial r}{\partial r_b} < 0 \). Furthermore, the equilibrium arrival rate equals \( \lambda^e(r) = \mu - \frac{c_w r^2}{c_w r^2} \) at optimality. We have \( \frac{\partial \lambda^e(r)}{\partial r_b} = \frac{c_w r}{c_w r^2} < 0 \). Thus, \( \frac{\partial \lambda^e(r)}{\partial \mu} = \frac{\partial \lambda^e(r)}{\partial r_b} \frac{\partial r}{\partial \mu} < 0 \). Similar to the previous cases, we have \( \frac{\partial R}{\partial \mu} = -\frac{1}{4} \left( \sqrt{\mu V(r)} - \sqrt{c_w \ln(\frac{1}{1-r})} \right)^{-1/2} \frac{\mu V(r)}{c_w} \frac{c_w}{c_w} < 0 \) at optimality.

\[ \blacksquare \]

**Proof of Proposition 4.5.** From Eqs. (4.12)-(4.14), the revenue function equals

\[
R(\lambda^e) = \lambda^e v(\lambda^e) - \frac{c_w \lambda^e}{\mu - \lambda^e} \ln\left(\frac{1}{1-r(\lambda^e)}\right) + c_r \lambda^e \left(\frac{1}{r_b} - \frac{1}{r(\lambda^e)}\right),
\]

(4.18)
in which the first term is assumed to be concave. Thus, if we show that the other two terms are also concave, we can conclude concavity of the revenue function.

Define \( y = \frac{c_r (\mu - \lambda^e)}{c_w} \) (or equivalently \( \lambda^e = \mu - \frac{c_w v_y}{c_w} \)). It is obvious that \( 0 < y \leq \frac{c_r \mu}{c_w} \) for
0 ≤ \lambda^e < \mu. Then, the last two terms can be rewritten as
\[
- \frac{c_w \lambda^e}{\mu - \lambda^e} \ln\left(\frac{1}{1 - r(\lambda^e)}\right) + c_r \lambda^e \left(\frac{1}{r_b} - \frac{1}{r(\lambda^e)}\right) = \frac{c_r \mu}{y} \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2})
\]
\[
- c_w \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2})
\]
\[
+ c_w \left(\frac{c_r \mu}{c_w} - y\right)\left(\frac{1}{r_b} - \frac{1}{-\frac{y}{2} + ((\frac{y}{2})^2 + y)^{1/2}}\right).
\]
(4.19)

The last two terms in Eq. (4.19) are concave with respect to \(y\), since
\[
\frac{\partial^2}{\partial y^2} \left( - c_w \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2}) \right) = - \frac{c_w}{4} (\frac{y}{2} + 1)((\frac{y}{2})^2 + y)^{-3/2} < 0,
\]
\[
\frac{\partial^2}{\partial y^2} \left( c_w \left(\frac{c_r \mu}{c_w} - y\right)\left(\frac{1}{r_b} - \frac{1}{-\frac{y}{2} + ((\frac{y}{2})^2 + y)^{1/2}}\right) \right) = - \frac{c_w}{y^2} (\frac{1}{4} + \frac{1}{x})^{-1/2}
\]
\[
- \frac{c_w}{y^2} (\frac{c_r \mu}{c_w} - y) \left(\frac{1}{4} + \frac{1}{x}\right)^{-3/2} (\frac{1}{4} + \frac{3}{4x}) < 0.
\]
(4.20)

It remains to show that the first term in Eq. (4.19) is concave with respect to \(y\). Regarding the first term, we have
\[
\frac{\partial^2}{\partial y^2} \left( \frac{c_r \mu}{y} \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2}) \right) = \frac{2c_r \mu}{x^3} g(y),
\]
(4.21)
in which \(g(y) = \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2}) + \frac{y}{2}((\frac{y}{2})^2 + y)^{-1/2} + \frac{y^2}{8}(1 + \frac{y}{2})((\frac{y}{2})^2 + y)^{-3/2}\). Since \(\frac{2c_r \mu}{x^3}\) is positive, the sign of \(g(y)\) determines the sign of the second derivative in Eq. (4.21). It can be shown that \(g(y)\) is strictly decreasing in \(y\) for \(y > 0\), i.e., \(\frac{\partial g(y)}{\partial y} = \frac{-y^2}{16} ((\frac{y}{2})^2 + y)^{-5/2} (1 + 2(\frac{y}{2} + 1)^2) < 0\), and \(g(y)\) approaches zero as \(y\) approaches zero, i.e., \(\lim_{y \to 0} g(y) = 0\). Thus, \(g(y) < 0\) for any \(y > 0\). we can conclude that
\[
\frac{\partial^2}{\partial y^2} \left( \frac{c_r \mu}{y} \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2}) \right) = \frac{2c_r \mu}{x^3} g(y) < 0.
\]

Thus, the first term is also concave with respect to \(y\). Since \(\lambda^e\) linearly depends on \(y\), all terms are also concave with respect to \(\lambda^e\). Therefore, the revenue function \(R(\lambda^e)\) is concave with respect to \(\lambda^e\). Moreover, \(R(\lambda^e) \to -\infty\) when \(\lambda^e \to \mu\). Thus, \(\lambda^e\) can uniquely be determined by Eq. (4.15), such that \(\lambda^e < \mu\).

It is obvious that if the solution of Eq. (4.15) is not constrained by the availability of the customers, i.e., potential arrival rate \(\Lambda\), then the service provider would serve all of the customers. However, if \(\tilde{\lambda} < \Lambda\), then the service provider serves only \(\tilde{\lambda} / \Lambda\) of the customers at optimality. \(\square\)
Proof of Proposition 4.6. Define \( y = \frac{c_r(\mu - \lambda^e)}{c_w} \) (or equivalently \( \lambda^e = \mu - \frac{c_c y}{c_r} \)). It is obvious that \( 0 < y \leq \frac{c_c}{c_r} \) for \( 0 \leq \lambda^e < \mu \). The delivery-reliability level as determined by Eq. (4.13) can be rewritten as

\[
 r(y) = -\frac{y}{2} + ((\frac{y}{2})^2 + y)^{1/2}.
\]

We have \( \frac{\partial r(y)}{\partial y} = -\frac{1}{4} + \frac{1}{2}((\frac{y}{2} + 1)((\frac{y}{2})^2 + y)^{-1/2} = \frac{1}{2}(((\frac{y}{2})^2 + y)^{-1/2}((\frac{y}{2} + 1) - ((\frac{y}{2})^2 + y)^{1/2}) \). It is easy to show that \( ((\frac{y}{2} + 1) - ((\frac{y}{2})^2 + y)^{1/2}) > 0 \). Thus, \( \frac{\partial r(y)}{\partial y} > 0 \). Consequently, \( \frac{\partial r(\lambda^e)}{\partial \lambda^e} < 0 \) and the delivery-reliability level is decreasing in the equilibrium arrival rate. Further, we have \( \frac{\partial^2 r(\lambda^e)}{\partial y^2} = \frac{1}{4}(((\frac{y}{2})^2 + y)^{-1/2} - \frac{1}{4}((\frac{y}{2} + 1)^2((\frac{y}{2})^2 + y)^{-3/2} = -\frac{1}{4}((\frac{y}{2})^2 + y)^{-3/2} < 0 \). Thus, \( \frac{\partial^2 r(\lambda^e)}{\partial y^2} < 0 \) and the delivery-reliability level is concave with respect to the equilibrium arrival rate.

By substituting \( y = \frac{c_r(\mu - \lambda^e)}{c_w} \) in the delivery time's formula \( w(\lambda^e) = \frac{1}{\mu - \lambda^e} \ln\left(\frac{1}{1 - r(\lambda^e)}\right) \), it can be expressed as

\[
 w(y) = -\frac{c_r}{c_w} y \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2}.
\]

We have

\[
 \frac{\partial w(y)}{\partial y} = \frac{c_r}{c_w} y \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2) + \frac{c_r}{2c_w} ((\frac{y}{2})^2 + y)^{-1/2} = \frac{c_r}{c_w} y \ln(1 + \frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2) + \frac{y}{2}((\frac{y}{2})^2 + y)^{-1/2) < \frac{c_r}{c_w} y (\frac{y}{2} - ((\frac{y}{2})^2 + y)^{1/2) + \frac{y}{2}((\frac{y}{2})^2 + y)^{-1/2) = \frac{c_r}{c_w} y ((\frac{y}{2})^2 + y)^{-1/2}(((\frac{y}{2})^2 + y)^{1/2 - (\frac{y}{2} + 1)) < 0.
\]

The first inequality holds, because \( \ln(x) < x - 1 \), and the second one holds, because \( ((\frac{y}{2})^2 + y)^{1/2} < (\frac{y}{2} + 1) \). We can conclude that \( \frac{\partial w(\lambda^e)}{\partial \lambda^e} > 0 \).

Following from Eq. (4.12), the optimal price equals

\[
 p(\lambda^e) = \frac{1}{r_b} (\lambda^e - c_w(\lambda^e)) + c_r(\frac{1}{r_b} - \frac{1}{r(\lambda^e)}).
\]

The first term is a decreasing function of \( \lambda^e \) by definition. In the previous paragraph, we showed that \( \frac{\partial w(\lambda^e)}{\partial \lambda^e} > 0 \). Thus, \( -c_w w(\lambda^e) \) is also decreasing in \( \lambda^e \). It remains to show that the last term is also decreasing in \( \lambda^e \). We have \( \frac{\partial}{\partial y} \frac{1}{r(\lambda^e)} = \frac{-1}{(r(\lambda^e))^2} \frac{\partial r(\lambda^e)}{\partial y} < 0 \). Thus, \( \frac{1}{r(\lambda^e)} \) is increasing in \( \lambda^e \), and consequently the last term is also decreasing in \( \lambda^e \). we can conclude that \( \frac{\partial p(\lambda^e)}{\lambda^e} < 0 \).
**Proof of Corollary 4.1.** The proof immediately follows from Proposition 4.6. □

**Proof of Proposition 4.7.** When the benchmark net value is uniformly distributed as described in Subsection 4.4.2, the revenue function equals

\[ R(\lambda^c) = \lambda^c(v_b + \epsilon(1 - \frac{2\lambda^c}{\Lambda})) - \frac{c_w\lambda^c}{\mu - \lambda^c} \ln\left(\frac{1}{1 - r(\lambda^c)}\right) + c_r\lambda^c\left(\frac{1}{r_b} - \frac{1}{r(\lambda^c)}\right), \]

from which it is apparent that the revenue increases as the mean of benchmark net value increases. Moreover, from \( \frac{\partial^2 R(\lambda^c)}{\partial \lambda^c \partial \epsilon} < 0 \) and \( \frac{\partial^2 R(\lambda^c)}{\partial \lambda^c \partial v_b} = 1 > 0 \), we can conclude that \( \frac{\partial \epsilon}{\partial v_b} > 0 \). According to Lemma 4.6, we conclude that \( \frac{\partial w(\lambda^c)}{\partial v_b} = \frac{\partial w(\lambda^c)}{\partial \lambda^c} \frac{\partial \lambda^c}{\partial v_b} > 0 \).

**Proof of Proposition 4.8.** The proof follows from Larsen (1998). Figure 4.A.1 schematically shows the different cases. The equilibrium arrival rate at optimality is the root of

\[ \Lambda \frac{\partial S(\lambda^c)}{\partial \lambda^c} = 0. \]

From Proof of Proposition 4.5, the equilibrium arrival rate is the intersection point of \( -\frac{\partial \epsilon}{\partial \lambda^c} = -v_b - \epsilon(1 - \frac{4\lambda^c}{\Lambda}) \) and \( \frac{\partial S(\lambda^c)}{\partial \lambda^c} \) where \( S(\lambda^c) = -\frac{c_w\lambda^c}{\mu - \lambda^c} \ln\left(\frac{1}{1 - r(\lambda^c)}\right) + c_r\lambda^c\left(\frac{1}{r_b} - \frac{1}{r(\lambda^c)}\right) \) is increasing and \( \frac{\partial S(\lambda^c)}{\partial \lambda^c} \) is decreasing in \( \lambda^c \) as shown in Proof of Proposition 4.5.

From Figure 4.A.1, it is obvious that if \( \lambda^{c*} < \frac{A}{4} \) \( (\lambda^{c*} > \frac{A}{4}) \), then for a lower customer heterogeneity level we have \( \lambda^{c*} > \lambda^{c*'} \) \( (\lambda^{c*} < \lambda^{c*'}) \). Results for the optimal delivery time and delivery-reliability level are also immediately concluded from Proposition 4.6. Proposition 4.6 and Eq. (4.12) together determine the results for the optimal price. Note that the difference between the optimal revenues is determined by the areas between \( -\frac{\partial \epsilon}{\partial \lambda^c} \bigg|_{\epsilon} \) and \( -\frac{\partial \epsilon}{\partial \lambda^c} \bigg|_{\epsilon'} \). In Figure 4.A.1, cases 1-5 correspond to the different cases distinguished in Table 4.3. In cases 1, 2, 4, and 5 as shown in this figure, it is easy
to conclude if the revenue increases or decreases as the customer heterogeneity level decreases. However, in case 2 ($\frac{A}{2} < \lambda^e < \frac{4}{\Lambda}$), it is not straightforward to determine if the change in customer heterogeneity causes the optimal revenue to increase or decrease. For more detailed discussions refer to Larsen (1998).

**Proof of Proposition 4.9.** At optimality we have \( \frac{\partial R(\lambda^e)}{\partial c_r} = -\lambda^e w(\lambda^e) < 0, \frac{\partial R(\lambda^e)}{\partial \mu} = \frac{\omega \lambda^e w(\lambda^e)}{\mu - \lambda^e w(\lambda^e)} > 0, \) and \( \frac{\partial R(\lambda^e)}{\partial r_b} = -\lambda^e \frac{r_b}{r^2} < 0. \) Moreover, we have \( \frac{\partial R(\lambda^e)}{\partial c_r} = \lambda^e \left( \frac{1}{r_b} - \frac{1}{r(\lambda^e)} \right) \) at optimality. Thus, we can conclude that if the optimal delivery-reliability level is greater than the benchmark delivery-reliability level, then \( \frac{\partial R(\lambda^e)}{\partial c_r} > 0; \) otherwise, \( \frac{\partial R(\lambda^e)}{\partial c_r} < 0. \) Regarding the customer heterogeneity level, we have that at optimality \( \frac{\partial R(\lambda^e)}{\partial c} = \lambda^e (1 - \frac{2\lambda^e}{\Lambda}), \) that implies the optimal revenue is increasing (decreasing) in the customer heterogeneity level, if the optimal revenue is less (greater) than \( \frac{A}{2}, \) i.e., the market share is less (greater) than half. Thus, we would like to know at which customer heterogeneity level the market share equals \( \frac{1}{2}, \) if at any point. \( \epsilon = \frac{1}{2} \) determines this value.

\( \square \)
CHAPTER 5

Price, Time, and Reliability Competition for Service Delivery

History: This chapter is based on a paper submitted to the journal of Service Science. It is also presented at the Euro 2018 Conference, July 2018, Valencia, Spain.
Abstract

A firm’s delivery performance may have significant impact on the satisfaction and purchase behavior of its customers. Empirical evidence has shown that customers are willing to pay a higher price for a faster and more reliable service. In this study, we address the interactions between price, promised delivery time, and delivery-reliability level in a competitive setting. We model the problem as competition among an arbitrary number of profit-maximizing firms facing boundedly rational customers who can choose to buy the service from one of the firms or balk. We show the existence of a unique Nash equilibrium and propose a simple iterative method that converges to this equilibrium. Furthermore, we compare our results with those in the existing literature and report interesting managerial insights. Our results suggest that a firm with a higher capacity level is not always better off in a market which is more sensitive to the delivery reliability, and even a firm with a lower capacity level may benefit from a market which is more sensitive to the promised delivery time.

Keywords: Pricing; Reliability; Competition; Queueing; Bounded rationality.

5.1 Introduction

In many industries, firms with better delivery performance gain competitive advantages over their competitors. Previous research has shown that a superior delivery performance entices customers into buying more frequently and paying premium prices (Rao et al., 2011). Dell’s due dates for delivery of products sold on-line, FedEx’s next-day guaranteed delivery, and Domino Pizza’s guaranteed 30-minute services are all examples of promises made to enhance the delivery performance for the end customers. Delivery performance is
also a key consideration for B2B customers. Short promised delivery times allow business customers to speed up their operations, and more reliable deliveries enable them to plan and coordinate their manufacturing activities accurately (Peng and Lu, 2017).

The existing literature highlights two dimensions of delivery performance: promised delivery time (in short, delivery time) and delivery reliability (Handfield and Pannesi, 1992; Morash et al., 1996). Marketing literature also emphasizes that delivery time and delivery reliability are among the most important customer service elements (see, e.g., Gronroos, 1988; Ballou, 2007). In the context of service supply chain management, Cho et al. (2012) defines the delivery reliability as the “ability to perform the promised service dependable and accurately”. In this paper, it is measured as the rate of deliveries on or before the promised delivery time.

From an operational point of view, the delivery time and delivery reliability are tightly linked together. As long as the firm’s capacity is fixed, although a shorter delivery time may be attracting more customers, it would lead to a lower delivery-reliability level (as a result of the increased congestion) and negatively influence the customers’ purchase behavior. The interplay between the delivery time and delivery-reliability level has been studied previously in the literature (e.g., Ho and Zheng, 2004; Boyaci and Ray, 2006; Xiao and Qi, 2016; Jalili Marand et al., 2017b). It has been shown that the trade-off that a firm has to make between offering a shorter delivery time with a lower delivery-reliability level and a longer delivery time with a higher delivery-reliability level is crucial in gaining competitive advantage in time-based competitions (Shang and Liu, 2011).

In addition to delivery performance, price is another important factor that influences customers’ purchase behavior, though some marketing studies show that less than 10% of the B2C customers and less than 30% of the B2B customers base their purchase decisions only on the price of the service or product (Baker et al., 2001). In a congested system where the capacity is limited, the price and time performance of the system are closely related. In fact, the pricing decision is a lever by which a firm can regulate the congestion level: increasing the price reduces the congestion level and lets the firm provide a service with a shorter delivery time and/or a higher delivery-reliability level, which in return increases the congestion level. The interactions between price, delivery time, and delivery-reliability level make it necessary to study them together. The importance of studying the interactions between price, delivery time, and delivery-reliability level is also acknowledged in the literature (So, 2000; Shang and Liu, 2011).

These interactions become even more complicated when firms serve boundedly rational customers in a competitive market. In this study, we address these issues. The bounded rationality assumption captures the human beings’ limited ability to process the information, and the competition assumption captures the more realistic situations
where multiple firms compete in terms of price and delivery performance. The problem is modeled as the competition among an arbitrary number of profit-maximizing firms facing boundedly rational customers who are sensitive to price, delivery time, and delivery-reliability level.

We make the following main contributions. (i) We show the existence of a unique Nash equilibrium to the price, delivery time, and delivery reliability competition problem. (ii) We propose a simple iterative algorithm that converges to this equilibrium. Furthermore, (iii) we conduct a sensitivity analysis on a monopolistic firm’s problem to answer the following question: How do the firm-specific parameters and the market-specific parameters affect the trade-offs between price, delivery time, and delivery reliability decisions for different bounded rationality levels? (iv) We also compare our price, delivery time, and delivery reliability competition model with the price and delivery time competition model in So (2000) and the delivery time and delivery reliability competition model in Ho and Zheng (2004) and answer the following question: When the customers are boundedly rational, how do the price, delivery time, and delivery-reliability level interact in a competitive environment? We show that a firm with a higher capacity level is not always better off in a more reliability-sensitive market, and even a firm with a lower capacity level may benefit from a more time-sensitive market.

The remainder of the paper is organized as follows. In Section 5.2, we review the most related literature. In Section 5.3, we formulate the problem. We analyze the best response problem for each firm in Subsection 5.3.2. Then, we investigate the competition problem in Subsection 5.3.3. In Section 5.4, the numerical results and main managerial insights are presented. We conclude the study and present future research directions in Section 5.5.

5.2 Literature Review

Empirical evidence supports the importance of delivery reliability as a measure of delivery performance for both customers and service providers/manufacturers. Considering the demand-oriented capabilities, Morash et al. (1996) show that the delivery reliability is rated highest in CEO-perceived importance. Rao et al. (2014) state that there is statistically significant evidence to suggest that the delivery reliability does, in fact, mitigate the risk of product returns. Rosenzweig et al. (2003) discuss the impact of reliability on customer satisfaction, and show that reliability directly impacts customer satisfaction. Rao et al. (2011) empirically examine the relationship between order fulfilment delays and subsequent shopping behavior for previously loyal customers in an on-line retailing environment. They emphasize that failure to keep these delivery time promises can be
detrimental for the retailer’s business. Using the data from a moderate-sized on-line retailer of printed material, they show that the delivery failures negatively affect the order frequency, order size, and customer anxiety.

In an empirical study, Peng and Lu (2017) analyze the transaction data collected from a heating, ventilation, and air conditioning control product supply chain, to examine the effect of the delivery performance on future customer transaction quantities and unit prices. They suggest that the delivery performance should not be treated as a single measure. They consider different dimensions of delivery performance, i.e., delivery speed (delivery time), on-time delivery rate (delivery-reliability level), and delivery date inaccuracy (late or early delivery), and find that the price is affected by on-time delivery rate and delivery speed, but not by delivery inaccuracy. They conclude that early deliveries do not appear to negatively affect the transaction quantity. According to them, the delivery-reliability level contributes to profitability by allowing the firms to charge premium prices. Although the empirical evidence emphasizes the existing interactions among price, delivery time, and delivery-reliability level, there is no integrated framework to analytically study these interactions in a competitive setting.

The competition problem in service delivery has been extensively studied (e.g., Hall and Porteus, 2000; Christ and Avi-Itzhak, 2002; Armony and Haviv, 2003; Allon and Federgruen, 2009; Li et al., 2012; Hu and Qiang, 2013; Saberi et al., 2014; Behzad and Jacobson, 2016; Chen et al., 2016a; Li et al., 2016a; Chen et al., 2016b; Jayaswal and Jewkes, 2016). Our study is closely related to four papers: So (2000), Ho and Zheng (2004), Allon and Federgruen (2007), and Shang and Liu (2011). We focus on reviewing these studies in details below.

So (2000) considers price and delivery time competition among N profit-maximizing service providers. Each service provider’s market share is determined using a multiplicative competitive interaction (MCI) model. Customers are assumed to be sensitive to the price and delivery time. Each firm is characterized by its capacity level (service rate) and unit operating cost. So (2000) proves the existence of a unique Nash equilibrium. While he assumes that the delivery-reliability level is exogenously set and does not consider its impact on the customers’ choice behavior, we assume that the customers are also sensitive to the delivery-reliability level. So (2000) assumes that no customer leaves the market without being served (no balking). This assumption is also relaxed in our study. Moreover, he implicitly assumes the customers to be boundedly rational.

Allon and Federgruen (2007) consider a model that is generally similar to So (2000), but they include the capacity level decision. They define the service level as the difference between the benchmark upper bound for the waiting time and the expected waiting time (or a fractile of the waiting time distribution). They consider three sequences through
which the firms make their choices. They show that in all three sequences an equilibrium pair of prices and service levels (waiting times) exists under some specified conditions. Allon and Federgruen (2009) extend Allon and Federgruen (2007) to a segmented market. Neither Allon and Federgruen (2007) nor Allon and Federgruen (2009) consider the impact of the delivery-reliability level on the customer behavior.

Ho and Zheng (2004) explicitly consider the delivery-reliability level along with the delivery time as the two main factors that influence customers’ utility and firms’ market shares. They employ a multinomial logit (MNL) model to determine the demand allocation of two competing firms. They assume that the price is exogenous, and hence, each firm aims at maximizing its demand rate. The firms are considered in both uncongested and congested settings. In addition, customers are implicitly assumed to be boundedly rational, but they cannot balk. Under these assumptions, the existence of a unique Nash equilibrium is proven for the delivery time and delivery reliability competition. They additionally study the impact of the capacity level on the optimal market share of a firm. They prove the existence of Nash equilibria in a duopolistic game and show that this game is similar to a prisoner’s dilemma when the cost of adding capacity is small. They also briefly discuss a more general oligopoly game.

Shang and Liu (2011) extend Ho and Zheng (2004) to a competition with \( N \) competing firms. They also employ the MNL model. The price is assumed to be fixed, and similar to Ho and Zheng (2004), the firms aim at maximizing their demand rates. They consider a lower bound for the delivery-reliability level. It is also assumed that customers are boundedly rational but cannot balk. They prove the existence of a unique Nash equilibrium and show that at the equilibrium, the firms are partitioned into two sets: a group of firms will offer the lower bound of delivery-reliability level and the rest will offer higher reliabilities. They also consider the capacity competition in a duopolistic game and find conditions for the existence of Nash equilibria.

In our study, we also consider an \( N \)-firm competition (similar to Shang and Liu, 2011; So, 2000). Assuming that the capacity cannot be adjusted in the short run, we fix the capacity level (similar to So, 2000). However, we relax the fixed price assumption in Ho and Zheng (2004) and Shang and Liu (2011), and the fixed delivery-reliability level assumption in So (2000) and Allon and Federgruen (2007). Moreover, we let the customers choose to balk or buy from one of the firms (unlike So, 2000; Ho and Zheng, 2004; Shang and Liu, 2011). Although this assumption does not cause extra difficulties in dealing with the problem, it yields some interesting results compared to the literature. Furthermore, we explicitly assume that the customers are boundedly rational. The bounded rationality assumption is represented by a noise term in customers’ utility function. It enables us to investigate the impact of the bounded rationality level on firms’ optimal decisions. The
bounded rationality assumption is of importance, because human beings are limited in their ability to process information (Simon, 1997). It has been shown that people rely on a limited number of heuristic principles that are generally useful but sometimes lead to severe and systematic errors (Kahneman et al., 1982).

Huang et al. (2013b) are the first to consider the customer bounded rationality in service systems using the MNL model. In a monopolistic setting, they show that ignoring bounded rationality can result in significant revenue and welfare losses even when the bounded rationality level is low. For an unobservable queue, they conclude that the revenue is strictly increasing in the bounded rationality level when it is sufficiently large. Li et al. (2016b) and Li et al. (2017) extend Huang et al. (2013b) to customer-intensive service competitions. Ren and Huang (2017) provide a review on the modeling of bounded rationality in operations management. We contribute to this stream of literature by assuming that the boundedly rational customers’ utility is a function of price, delivery time, and also delivery-reliability level, unlike the above mentioned studies.

It is worth noting that the interactions between price, delivery time, and delivery-reliability level have scarcely been addressed in the literature, e.g., Boyaci and Ray (2006), Xiao and Qi (2016), and Jalili Marand et al. (2017b). Jalili Marand et al. (2017b) consider a monopolistic revenue-maximizing firm facing price-, delivery-time-, and delivery-reliability-sensitive customers. Customers are assumed to be self-interested and rational. They investigate the customers’ equilibrium behavior and the firm’s optimal decisions under two assumptions: homogeneous and heterogeneous customers. Our work differs from theirs in two main assumptions: the bounded rationality and competition.

Boyaci and Ray (2006) consider the optimal differentiation strategy in terms of prices, delivery times, and delivery-reliability levels of a profit-maximizing firm that sells two variants of a product in a capacitated environment. They assume that each product’s demand rate linearly depends on both products’ prices, delivery times, and delivery-reliability levels. In a two-stage supply chain with one supplier and one manufacturer, Xiao and Qi (2016) study the equilibrium decisions in the supply chain with an all-unit quantity discount contract. Their demand function is linearly dependent on the price, delivery time, and delivery-reliability level. They conclude that an all-unit quantity discount scheme can coordinate the supply chain for most cases. Our work also differs from those of Boyaci and Ray (2006) and Xiao and Qi (2016), since those studies do not capture the horizontal competition among service providers.

To the best of our knowledge, none of the existing studies have considered the joint price, delivery time, and delivery reliability competition among \( N \) horizontally competing firms in a market with boundedly rational customers. Table 5.1 summarizes the key literature and shows the positioning of our study.
<table>
<thead>
<tr>
<th>Literature</th>
<th>Decisions</th>
<th>Comp</th>
<th>BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>So (2000)</td>
<td>✓ ✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Allon and Federgruen (2007)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>(implicitly)</td>
</tr>
<tr>
<td>Shang and Liu (2011)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Huang et al. (2013b)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Jalili Marand et al. (2017b)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Our study</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
</tr>
</tbody>
</table>

Table 5.1: Positioning of the current study in the literature (Comp and BR stand for Competition and Bounded Rationality, respectively)

5.3 The Competition Game and Analysis

5.3.1 Problem Description

Consider a market consisting of a homogeneous population of customers and $N$ service providers (henceforth, firms). Upon their arrivals, customers have $N + 1$ choices: they can choose between services provided by one of the $N$ firms, or they can balk (walk away without being served). Let $U_i$ denote a customer’s actual utility from using the service provided by Firm $i$, $i = 1, \ldots, N$. The reward that customers gain on the completion of service, $v_i$, the price, $p_i$, the delivery time, $t_i$, and the delivery-reliability level, $q_i$, together characterize Firm $i$’s service. It is assumed that there is no moral hazard problem involved.

A major part of the classical queue-pricing literature assumes that a customer’s utility is linearly dependent on the price and delivery time (see Hassin and Haviv, 2003, for a comprehensive literature review). Moreover, some recent studies assume that the customer’s utility is linearly dependent on the delivery time and delivery-reliability level (see, e.g., Ho and Zheng, 2004; Shang and Liu, 2011). In this study, we extend the literature by assuming that the utility is linearly dependent on the price, delivery time, and delivery-reliability level, and define

$$ U_i(p_i, t_i, q_i) = v_i - c_p p_i - c_t t_i + c_q q_i, \quad i = 1, \ldots, N, \quad (5.1) $$

in which $c_p$ denotes the customer price sensitivity, $c_t$ denotes the customer waiting cost rate (or simply the time sensitivity), and $c_q$ denotes the reliability sensitivity of the customers.

If we ignore the effect of the price, i.e., $c_p = 0$, our model is reduced to that of Ho and Zheng (2004) and Shang and Liu (2011), and disregarding the impact of the delivery reliability, i.e., $c_q = 0$, results in a model similar to So (2000). To complement their
results, we let both \( c_p \) and \( c_q \) be positive constants, i.e., \( c_p, c_q > 0 \). Furthermore, we assume that a customer’s actual utility of choosing to balk is normalized to zero, i.e., \( U_0 = 0 \). Note that subscript 0 denotes the balking option in the remainder of the paper.

The existing queue-pricing literature typically assumes that the customers are fully rational (see Ren and Huang, 2017, for a review on the exceptions). However, we assume that customers are boundedly rational in the sense that there may be a gap between their perceived utility and their actual utility at the time of making the joining/balking decisions. In order to characterize the bounded rationality, we associate a random noise with the customers’ perception of the service utility. Thus, the customers have a perceived utility from the service provided by Firm \( i \) as

\[
U_i(p_i, t_i, q_i; \epsilon_i) = U_i(p_i, t_i, q_i) + \epsilon_i, \; i = 1, \ldots, N,
\]

(similar to, e.g., Huang et al., 2013b; Li et al., 2016b). Similarly, the customers’ perceived utility from balking equals \( U_0 = \epsilon_0 \). As a result of the bounded rationality, customers will base their decisions on the perceived utilities, i.e., \( U_i, \; i = 0, 1, \ldots, N \), that depend on the realized value of the random noise. Therefore, they cannot guarantee that the best choice is always chosen, and they may even choose a service that yields a negative actual utility. Moreover, if it turns out to be the case that the perceived utility from balking is higher than the perceived utilities from the services provided by one of the firms, the customer will balk.

It is assumed that \( \epsilon_i \) are independent and identically Gumbel-distributed random variables with the mean \( E(\epsilon_i) = 0 \) and variance \( Var(\epsilon_i) = \frac{2\pi^2}{6} \), where \( \eta > 0 \) is the scale parameter of the corresponding Gumbel distribution (see Talluri and Van Ryzin, 2006, pages 301-308, for detailed discussions). By using a single parameter \( \eta \), we reflect the bounded rationality level of the customers. As \( \eta \) approaches positive infinity, the customers become completely irrational, and conversely, as \( \eta \) approaches zero, the customers become fully rational. When customers are completely irrational, they will join one of the firms or balk with equal fractions.

We employ the MNL model to capture the customer’s choice behavior. The MNL model is one of the most commonly used attraction models in the literature. See Huang et al. (2013a) for a review of the demand functions in decision modeling and Strauss et al. (2018) for a review of recent developments on choice-based revenue management. Based on the MNL model, the probability that a customer chooses Firm \( i \)’s service equals

\[
\theta_i = \frac{e^{U_i(p_i, t_i, q_i)/\eta}}{\sum_{j=0}^{N} e^{U_j(p_j, t_j, q_j)/\eta}}, \quad (5.2)
\]
and the probability that a customer chooses to balk is

\[ \theta_0 = \frac{1}{\sum_{j=0}^{N} e^{U_j(p_j,t_j,q_j)/\eta}}. \]

We assume that the customers arrive at the market following a Poisson process with the rate \( \Lambda \) (the potential market size). Let \( \lambda_i \) be the demand rate captured by Firm \( i \). From Eq. (5.2), the expected market share of Firm \( i \) equals \( \frac{\lambda_i}{\Lambda} = \theta_i \). Therefore, arrivals to Firm \( i \) will also follow a Poisson process with the rate \( \lambda_i = \Lambda \theta_i \) or equivalently

\[ \lambda_i(p_i,t_i,q_i|\beta_i) = \frac{\Lambda e^{U_i(p_i,t_i,q_i)/\eta}}{e^{U_i(p_i,t_i,q_i)/\eta} + \beta_i}, \tag{5.3} \]

where \( \beta_i = \sum_{j=0}^{N} e^{U_j(p_j,t_j,q_j)/\eta} \geq 1 \) is the aggregate impact of the other firms’ decisions on Firm \( i \)'s demand rate (arrival rate).

Furthermore, it is assumed that Firm \( i \)'s service times are exponentially distributed with the parameter \( \mu_i \). Thus, we model the operations of each firm as an \( M/M/1 \) queueing system similar to most of the related literature. The delivery time is measured by the sojourn time. The sojourn time in an \( M/M/1 \) queue is exponentially distributed (Asmussen, 2008). In fact, even for a \( G/G/1 \) queue, the tail distribution of the sojourn time can be accurately approximated by an exponential distribution for high delivery-reliability levels (Abate et al., 1996). Thus, setting a reasonably high lower bound, i.e., \( q_i \), for the delivery-reliability level, our results approximately hold for more general queueing systems. The lower bound \( q_i \) can be interpreted as the market entrance requirement (see Shang and Liu, 2011, for a similar assumption).

Let \( s \) be the random variable denoting the delivery time. According to our definition, Firm \( i \)'s delivery-reliability level equals

\[ q_i = Pr(s \leq t_i) = 1 - e^{-(\mu_i-\lambda_i)t_i}, \tag{5.4} \]

for \( 0 \leq \lambda_i < \mu_i \). From Eq. (5.4), the arrival rate of Firm \( i \) can be expressed as

\[ \lambda_i = \mu_i - \frac{1}{t_i} \ln\left(\frac{1}{1-q_i}\right), \tag{5.5} \]

where \( 0 < q_i \leq q_i < 1 \). According to Eq. (5.5), we have \( \lambda_i < \mu_i \) and the stability condition is always met. To be able to derive our results we need the condition \( q > 0.64 \) (or equivalently \( k = -\ln(1-q) > 1 \)) to hold, which does not affect the practical generality of our results (specifically, in today’s markets). In fact, we set \( q \) at much higher levels in the numerical analysis part.
Table 5.2: Notation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_p$</td>
<td>price sensitivity of customers</td>
<td>$\Pi$</td>
<td>profit of Firm $i$</td>
</tr>
<tr>
<td>$c_t$</td>
<td>time sensitivity of customers</td>
<td>$U_i$</td>
<td>actual expected utility from Firm $i$’s service</td>
</tr>
<tr>
<td>$c_q$</td>
<td>reliability sensitivity of customers</td>
<td>$U_i$</td>
<td>perceived utility from Firm $i$’s service</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>potential market size</td>
<td>$p_i$</td>
<td>price of Firm $i$</td>
</tr>
<tr>
<td>$q_i$</td>
<td>market entrance reliability requirement</td>
<td>$t_i$</td>
<td>delivery time of Firm $i$</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>unit operating cost of Firm $i$</td>
<td>$q_i$</td>
<td>delivery reliability of Firm $i$</td>
</tr>
<tr>
<td>$\nu_i$</td>
<td>customer reward from Firm $i$’s service</td>
<td>$\lambda_i$</td>
<td>arrival rate at Firm $i$</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>capacity of Firm $i$</td>
<td>$x_i$</td>
<td>strategy profile of Firm $i$</td>
</tr>
<tr>
<td>$\epsilon_i$</td>
<td>Gumbel-distributed noise random variable</td>
<td>$\theta_0$</td>
<td>probability of choosing to balk</td>
</tr>
<tr>
<td>$\eta_i$</td>
<td>bounded rationality measure</td>
<td>$\theta_i$</td>
<td>probability of choosing Firm $i$’s service</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>competition impact on Firm $i$’s problem</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the described setting, we consider the competition problem between $N$ profit-maximizing firms in a Nash game, where all firms simultaneously decide on their prices, delivery times, and delivery-reliability levels. Let $x_i = (p_i, t_i, q_i)$ denote the strategy profile of Firm $i$, $i = 1, ..., N$. A strategy profile $\phi^* = (x_1^*, ..., x_N^*)$ is a Nash equilibrium if $\Pi_i(p_i, t_i, q_i|\beta_i) \geq \Pi_i(p_i, t_i, q_i|\beta_i^*)$ for any feasible $(p_i, t_i, q_i)$ and for all $i = 1, ..., N$. In words, a Nash equilibrium solution to the competition problem is a vector consisting of the individual firms’ decisions such that no firm has an incentive to unilaterally deviate from the equilibrium solution. In order to solve the competition problem, we first analyze a tagged firm’s best response problem given the decisions of other firms. The notation is summarized in Table 5.2.

5.3.2 The Best Response Problem

In this subsection, we formulate and solve the best response problem of a tagged firm, i.e., Firm $i$, $i = 1, ..., N$, which is equivalent to finding Firm $i$’s optimal decisions given the decisions of its competitors. Given $\beta_i$ as the combined impact of other firms’ decisions on Firm $i$’s problem, Firm $i$’s best response problem is formulated as

$$\max_{p_i, t_i, q_i} \Pi_i(p_i, t_i, q_i|\beta_i) = \lambda_i(p_i, t_i, q_i|\beta_i)(p_i - \gamma_i)$$  \hspace{1cm} (5.6)

s.t. $\gamma_i \leq p_i$, \hspace{1cm} (5.7)

$0 \leq t_i$, \hspace{1cm} (5.8)

$q_i \leq q_i < 1$. \hspace{1cm} (5.9)

in which $\lambda_i$ and $q_i$ are defined by Eq. (5.3) and Eq. (5.4), respectively. Equating right-hand sides of Eq. (5.3) and Eq. (5.5) gives

$$U_i(p_i, t_i, q_i|\beta_i) = \eta \ln \frac{\beta \lambda_i}{\Lambda - \lambda_i}.$$  \hspace{1cm} (5.10)
From Eq. (5.10), when $\eta \to 0$, we have $U_i \to 0$. In words, when customers become fully rational, the firm extracts all customer surplus at the equilibrium (Hassin and Haviv, 2003).

Substituting Eq. (5.1) in Eq. (5.10) results in

$$p_i(t_i, q_i | \beta_i) = \frac{1}{c_p} \left( v_i - c_t t_i + c_q q_i - \eta \ln \frac{\beta_i (\mu_i - \frac{1}{t_i} \ln \frac{1}{1-q_i})}{\Lambda - \mu_i + \frac{1}{t_i} \ln \frac{1}{1-q_i}} \right).$$

(5.11)

The price $p_i(t_i, q_i | \beta_i)$ is defined for

$$1 - \frac{1}{\mu_i} \ln(\frac{1}{1-q_i}) < t_i < \frac{1}{\mu_i - \Lambda} \ln(\frac{1}{1-q_i})$$

when $\mu_i > \Lambda$, and for

$$\frac{1}{\mu_i} \ln(\frac{1}{1-q_i}) < t_i$$

when $\mu_i \leq \Lambda$. Eq. (5.11) expresses the price in terms of the delivery time and delivery-reliability level. Substituting Eq. (5.5) and Eq. (5.11) into Eq. (5.6) results in the following optimization problem:

$$\max_{t_i, q_i} \Pi_i(t_i, q_i | \beta_i) = (\mu_i - \frac{1}{t_i} \ln(\frac{1}{1-q_i}))(p_i(t_i, q_i | \beta_i) - \gamma_i)$$  

(5.12)

s.t. Eq. (5.7) - Eq. (5.9).

Proposition 5.1 characterizes the optimal solution to the best response problem. Refer to Appendix 5.A for the proofs.

**Proposition 5.1.** Given the other firms’ decisions, $\beta_i$,

(a) there exists a unique optimal price, $p_i^*$, delivery time, $t_i^*$, and delivery-reliability level, $q_i^*$, maximizing Firm $i$’s profit function, $\Pi_i(p_i, t_i, q_i | \beta_i)$.

(b) If

(i) $\Lambda > \mu_i$, $q < 1 - \frac{\mu_i}{c_q \Lambda}$ and $H(q | \beta_i) > 0$, or

(ii) $\Lambda < \mu_i$, $1 - \frac{\mu_i}{c_q (\mu_i - \Lambda)} < q < 1 - \frac{\mu_i}{c_q \Lambda}$ and $H(q | \beta_i) > 0$, or

(iii) $\Lambda < \mu_i$, $q < 1 - \frac{\mu_i}{c_q (\mu_i - \Lambda)}$,

then $p_i^* = p_i(t_i^*, q_i^* | \beta_i)$, $t_i^* = \frac{c_q (1-q_i^*)}{c_t} \ln \frac{1}{1-q_i^*}$, and $q_i^*$ is the unique root of $H(q | \beta_i) = 0$, where

$$H(q | \beta_i) = \frac{c_q^2 \mu_i (1-q)}{c_t} \ln \frac{1}{1-q} + \frac{\eta \Lambda}{\Lambda - \mu_i + \frac{c_t}{c_q (1-q)}}$$

$$+ \eta \ln \frac{\mu_i - \frac{c_q (1-q)}{c_q (1-q)}}{\Lambda - \mu_i + \frac{c_t}{c_q (1-q)}} - v_i - c_q q + \eta \ln \beta_i + c_p \gamma_i,$$
otherwise, \( p_i^* = p_i(t_i^*, q_i^*|\beta_i), q_i^* = q, \) and \( t_i^* \) is the unique root of \( G(t|\beta_i) = 0, \) where
\[
G(t|\beta_i) = \frac{c_t \mu_i t^2}{k} + \frac{\eta \Lambda}{\Lambda - \mu_i + \frac{k}{t}} + \eta \ln \left( \frac{\mu_i - \frac{k}{t}}{\Lambda - \mu_i + \frac{k}{t}} \right) - v_i - c_q q + \eta \ln \beta_i + c_p \gamma_i,
\]
in which \( k = \ln(\frac{1}{1-q}). \)

Proposition 5.1 states that, having the decisions of other firms, the optimal decisions of Firm \( i \) is uniquely determined. Moreover, although no closed-form solution exists, finding the optimal solution is not challenging, because \( H(q|\beta_i) \) and \( G(t|\beta_i) \) are monotone and the roots of \( H(q|\beta_i) = 0 \) and \( G(t|\beta_i) = 0 \) can easily be found using efficient numerical methods such as the bisection method.

The next proposition states how Firm \( i \)'s decisions are influenced by other firms’ decisions and plays a key role in the iterative solution method, proposed in the next subsection, for solving the competition problem.

**Proposition 5.2.** For Firm \( i \), the optimal price and delivery time are strictly decreasing in \( \beta_i \), and the optimal delivery-reliability level is increasing in \( \beta_i \).

Proposition 5.2 implies that under increased competition, a firm not only offers lower price and shorter delivery time, but it also increases the delivery-reliability level. For a higher level of competition, i.e., a higher \( \beta_i \), So (2000) also shows that the price and delivery time decrease. Correspondingly, Ho and Zheng (2004) show through numerical studies that the delivery time would be tighter for a higher attraction level of competitors, \( \beta_i \). Proposition 5.2 also implies that the competition decreases the firm’s market share (and also the arrival rate), since from Eq. (5.5) we have \( \frac{\partial \lambda_i}{\partial \beta_i} = \frac{1}{t_i} \ln(\frac{1}{1-q_i}) \frac{\partial t_i}{\partial \beta_i} - \frac{1}{t_i(1-q_i)} \frac{\partial q_i}{\partial \beta_i} < 0. \) For the same reason, the firm’s profit also decreases with the competition.

### 5.3.3 The Equilibrium Analysis

In this subsection, we consider the Nash game between \( N \) competing firms, where all the firms simultaneously decide on their prices, delivery times, and delivery-reliability levels. As mentioned in Subsection 5.3.1, a Nash equilibrium solution to the competition problem is a vector consisting of individual firms’ decisions such that no firm has an incentive to unilaterally deviate from the equilibrium solution. We employ the following iterative procedure, adopted from So (2000) and modified to fit our problem, to solve the competition problem.

1. **Initialization:** For Firm \( i \), choose \( p_i = \gamma_i, t_i = \frac{1}{\mu_i} \ln(\frac{1}{1-q}) \), and \( q_i = 1. \)
(2) Iterative step: Start with $i = 1$. Apply the results in Proposition 5.1 to find Firm $i$'s optimal decisions given the current decisions of other firms. Repeat this for all $i = 2, ..., N$.

(3) Termination condition: Repeat step (2) until the difference between the decisions of two successive steps falls below a pre-specified tolerance level $\epsilon$, i.e., $\Delta \beta_1 < \epsilon$.

Now consider Proposition 5.3 that ensures both the existence and uniqueness of the Nash equilibrium.

**Proposition 5.3.** There exists a unique Nash equilibrium to the N-firm competition problem, and the above-mentioned iterative procedure converges to this equilibrium solution.

In the next section, we utilize the described iterative procedure to conduct our numerical study. Before proceeding to the numerical results in Section 5.4, we present two corollaries.

**Corollary 5.1.** Suppose $\mu_N \leq ... \leq \mu_1$ and $i, j \in \{1, ..., N\}$ such that $i \leq j$. If $\frac{c_t}{c_q} > (1 - q)\mu_i$, then Firm $j$ cannot differentiate in the delivery-reliability level and $q_j = q$.

Corollary 5.1 implies that if $\frac{c_t}{c_q} > (1 - q)\mu_1$, then our model would be similar to So (2000) with no delivery-reliability level competition. It also shows how the availability of capacity and market parameters together influence a firm’s offered delivery-reliability level; when the ratio of the time sensitivity to the reliability sensitivity, $\frac{c_t}{c_q}$, is too high, no firm has an incentive to offer a reliability greater than the market entrance requirement. As this ratio decreases, however, more firms may benefit from differentiating in reliability. The ratio $\frac{c_t}{c_q}$ plays an important role in our analysis in the next section. Corollary 5.2 is an immediate result of Proposition 5.2 and the fact that in a monopolistic market $\beta_i = 1$.

**Corollary 5.2.** Everything else being equal, a firm offers a lower price, a shorter delivery time, and a higher delivery-reliability level and gains a lower market share and consequently a lower profit in a competitive market compared to a monopolistic market.

## 5.4 Numerical Study and Managerial Insights

In this section, we first analyze the sensitivity of a monopolistic firm’s decisions to the problem parameters. Then, we investigate the competition problem. We use the iterative procedure introduced in Subsection 5.3.3 to analyze the competition model. In order to keep the results interpretable, we consider two-player games. We compare our results with those in the literature. We refer to our model, the fixed price model (Ho and Zheng, 2004), and the fixed delivery-reliability level model (So, 2000) as the *Price, Time, and...*
Reliability (PTR), Time and Reliability (TR), and Price and Time (PT) competition models, respectively.

In our numerical studies, we are interested to see how the firm-specific parameters, i.e., capacity level and unit operating cost, affect the competition. It is obvious that if a firm has both a capacity advantage (a higher capacity level) and a cost advantage (a lower unit operating cost), then the firm will have the dominating power in the competition. As a result, it gains a higher market share and also makes more profit compared to its competitors. Thus, we focus on the more interesting competition situations where one firm has the capacity advantage and another firm has the cost advantage. Furthermore, in the PTR and TR competitions, we are more interested in the cases where the optimal delivery-reliability level is an interior point, because such cases enable us to study the impact of the delivery-reliability decision on firms’ profitability.

5.4.1 Monopoly Market

The model structure for the monopoly market is similar to the best response problem, where $\beta_i = 1$. Proposition 5.4 shows the sensitivity of the interior and boundary solutions to variations of both market-specific (i.e., the market size, price sensitivity, time sensitivity, and reliability sensitivity) and firm-specific (i.e., the basic reward, unit operating cost, and capacity level) parameters. The results are elaborated afterwards.

**Proposition 5.4.** For a monopolistic firm, the optimal price, delivery time, delivery-reliability level, and arrival rate are sensitive to problem parameters as shown in Table 5.3.

<table>
<thead>
<tr>
<th>Sensitivity to</th>
<th>Interior solution</th>
<th>Boundary solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda$</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>$c_p$</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>$c_t$</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>$c_q$</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$v_i$</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>↑</td>
<td>↓</td>
</tr>
</tbody>
</table>

Table 5.3: Sensitivity analysis of a monopolist

**Sensitivity to $\Lambda$:** In a larger market, the firm’s arrival rate is also higher. However, as the capacity level remains unchanged, the congestion level increases. As a result of the increased congestion level, the firm not only offers a longer delivery time (a result that is consistent with the results in Ho and Zheng, 2004), but it also provides its service
with a lower delivery-reliability level. So (2000) also shows that as the size of the overall
market shrinks, the firm offers a shorter delivery time to increase its arrival rate in order
to fill its available capacity.

**Sensitivity to** $c_p$: As the customers become more price sensitive, the firm both
decreases its price and enhances its delivery performance. Surprisingly, despite a lower
price and a superior delivery performance (a lower delivery time and a higher delivery-
reliability level), the firm’s optimal arrival rate decreases as the price sensitivity increases.
Note that when the unit operating cost, $\gamma_i$, is equal to zero, the variations in the price
sensitivity do not influence the optimal delivery time and delivery-reliability level.

**Sensitivity to** $c_t$: As the customers become more time sensitive, the firm decreases its
delivery-reliability level. Figures 5.1a-5.1d show the variations of the optimal decisions
with respect to the time sensitivity. According to these figures, as the time sensitivity
increases, the firm delivers its service in a shorter time. Moreover, it reduces its price
and delivery-reliability level. The market share of the firm also decreases. Consequently,
the profit of the firm decreases in a more time-sensitive market.

**Sensitivity to** $c_q$: When the optimal delivery-reliability level is on the boundary, the
optimal delivery time, price, and arrival rate increase as the customers become more
reliability sensitive. However, when the optimal delivery-reliability level is an interior
point, it is not straightforward to analytically show the impact of the reliability sensitivity
on the optimal decisions. Figures 5.1e-5.1h show these results for a number of sample
problems. According to these figures, as the reliability sensitivity increases, the firm
increases its delivery-reliability level at the expense of a longer delivery time. It also
increases its price. The market share increases. From Figures 5.1e and 5.1h, we can
conclude that the firm gains a higher profit in a more reliability sensitive market.

**Sensitivity to** $v_i$: As the service reward increases, even though the firm increases its
price and reduces its delivery performance, it gains a higher market share and makes more
profit. So (2000) also concludes that both the optimal delivery time and the optimal price
increase as the service reward increases.

**Sensitivity to** $\gamma_i$: A firm with a higher unit operating cost charges a higher price and
offers a superior delivery performance. This can be explained as follows. At a higher
unit operating cost, the firm has to increase its price to avoid a higher marginal profit
loss. Increasing the price has a negative impact on customers which is compensated for
with a superior delivery performance. Employing the MCI model, So (2000) also shows
Figure 5.1: The impact of the bounded rationality level on a monopolistic firm’s decisions ($\Lambda = 1, v = 2, \mu = 5, \gamma = 0.8, c_p = 1, c_t = 0.5, c_q = 8, q = 0.90$).
that the firm increases its price and decreases its delivery time as the unit operating cost increases.

**Sensitivity to $\mu_i$:** A firm with a higher capacity level offers a faster and more reliable service. It is also shown by So (2000) that the firm shortens its delivery time when its capacity level increases because such a move allows for an increase in its market share to fill the additional capacity. It is also intuitive that when the capacity level increases, the optimal profit also increases as shown in Figure 5.1, because the feasible space broadens. Figures 5.1i-5.1k show the impact of the capacity level on the optimal price, delivery-reliability level, and arrival rate. Figures 5.1i and 5.1j together reveal that the optimal price first increases in the capacity level and then decreases as long as the optimal delivery-reliability level is at the boundary. However, when the delivery-reliability level is an interior point, the optimal price first decreases and then increases in the capacity level. According to Figure 5.1k, the optimal market share is increasing in the capacity level.

![Figure 5.2: The impact of the bounded rationality level on a monopolistic firm’s decisions ($v = 2$, $\mu = 5$, $\gamma = 0.8$, $c_p = 1$, $c_t = 0.5$, $c_q = 8$, $q = 0.90$)](image)

In addition, we look into the impact of the bounded rationality level on the firm’s optimal decisions through the numerical examples presented in Figure 5.2. In this figure, we note that the optimal market share of the firm decreases as the customers become less rational. Moreover, when the bounded rationality level is sufficiently high, the optimal
profit is increasing in the bounded rationality level. It can be shown that the optimal profit is increasing (decreasing) in the bounded rationality level as long as the optimal market share is less (greater) than $\frac{1}{1+\beta} = \frac{1}{2}$ (or equivalently 50%). For instance, in the Figure 5.2a for $\Lambda = 4$, the market share is less (greater) than $\frac{1}{2}$ for $\eta > 4 \ (\eta < 4)$, and as shown in Figure 5.2b, the optimal profit is increasing (decreasing) in $\eta$ in this range, i.e., $\eta > 4 \ (\eta < 4)$. A market share less than $\frac{1}{2}$ induces each customer to receive a strictly negative expected utility in equilibrium. Our results are consistent with Huang et al. (2013b) who found that a higher bounded rationality level can lead to a lower optimal price and a lower profit, but it also leads to a higher optimal price and a higher profit when it is sufficiently large. For this observation, the following intuition can be offered: From Figure 5.2a, as the bounded rationality level increases, the firm’s market share (the joining fraction of the customers) decreases$^1$. Moreover, as $\eta$ increases the customers become indifferent to different options and their joining fraction almost stabilizes. As shown by Figure 5.2c, however, as $\eta$ increases, the firm sharply increases the optimal price, when $\eta$ is sufficiently large, in order to compensate for the market share loss and also to benefit from the customers’ bounded rationality. The interaction between the sharply increasing price and the almost stabilizing market share results in an increasing optimal profit for sufficiently high bounded rationality levels. Furthermore, Figure 5.2 shows that the optimal delivery time and the optimal delivery-reliability level are monotone in the bounded rationality level.

5.4.2 Comparison: The PTR and TR Competitions

In this subsection, we compare the PTR and TR competitions. The TR competition is studied by Ho and Zheng (2004). They also employ the MNL model. However, they do not mention the fact that the customers are boundedly rational, which is an inherent property of the MNL model. In other words, the bounded rationality measure, $\eta$, is set equal to one in their study. In addition, they assume that customers do not balk. Although the balking assumption does not impose extra difficulty on the equilibrium analysis, it significantly influences the outcomes. When balking is an option, a major part of the market may choose to walk away without being served. Based on numerical examples, we show how the customers’ bounded rationality and the balking option could influence the firms’ optimal decisions. Since firms’ objective functions are their market shares in the TR competition, we mainly do the comparison in terms of the firms’ market shares in this subsections.

$^1$In this numerical example, the reason is that the service provided by the firm results in a sufficiently high utility at low bounded rationality levels.
Balking Option: The Hidden Competitor (I)

Consider two firms as $(v_1, \mu_1, \gamma_1) = (2, 2.5, 0.2)$ and $(v_2, \mu_2, \gamma_2) = (2, 5.0, 0.8)$. We see that customers gain the same reward on the service completion from both firms, i.e., $v_1 = v_2$, so the firms would have captured equal market shares in the TR competition if the customers were not time and reliability sensitive, i.e., $c_t = c_q = 0$. Moreover, Firm 1 (Firm 2) has the cost advantage (the capacity advantage) over Firm 2 (Firm 1).

Figure 5.3a demonstrates the market shares of the two firms with respect to the reliability sensitivity, $c_q$, with and without the balking option for the same level of the customer bounded rationality, i.e., $\eta = 1$. The figure shows that when the reliability sensitivity is low, Firm 1 gains a higher market share. Note that in the TR competition a firm with the capacity advantage over its competitors always gains a higher market share, while this fact does not hold in the PTR competition. The reason is that when the reliability sensitivity is low, Firm 2 cannot exploit its capacity advantage over Firm 1, and Firm 1 captures a higher market share due to its cost advantage. It is for the case where the customers cannot balk. However, when the customers can balk the results are different, because the customers may choose not to buy from the firms that offer a service with a low utility. In this case, when the reliability sensitivity is low, the firms can hardly use their capacities to attract the customers by offering a better delivery performance. Therefore, as shown in Figure 5.3a, a considerable part of the market chooses to balk, and the market shares of both firms decrease.

When the customers do not have the balking option, as the reliability sensitivity increases, Firm 2’s market share increases, because now it can exploit its capacity advantage to gain a higher market share than Firm 1. Thus, without the balking option, Firm 2’s (Firm 1’s) market share is strictly increasing (decreasing) in the delivery-reliability level (see Figure 5.3a). In contrast to this observation, when boundedly rational customers have the balking option, as the reliability sensitivity increases, both firms’ market shares may initially increase, as shown in Figure 5.3a. In this case, as the reliability sensitivity increases up to a threshold (here, $c_q = 6$), both firms can exploit their capacities to offer a higher delivery-reliability level and attract some of the balking customers. Beyond the threshold, however, Firm 2’s (Firm 1’s) market share increases (decreases) in the reliability sensitivity, because the impact of the balking option is diminished.

Figure 5.3a also shows that even a firm that is in at a capacity disadvantage may benefit from an increase in the reliability sensitivity when the balking option is considered, whereas the same firm is always worse off in a more reliability sensitive market when the balking option is not considered. This observation is in contrast to the results observed in the literature (see Ho and Zheng, 2004; Shang and Liu, 2011).
Bounded Rationality: The Hard Fact

Figure 5.3 shows the profits and market shares of the two firms with respect to the reliability sensitivity for different bounded rationality levels. Moving from Figure 5.3b to Figure 5.3a and Figure 5.3c, \( \eta \) increases, i.e., the customers become less rational. While Ho and Zheng (2004) marginalize the impact of bounded rationality by setting \( \eta \) equal to one, these figures show the important impact of the customer bounded rationality on the equilibrium market shares.

Figure 5.3 also shows that as the bounded rationality level increases, the gap between two firms’ market shares decreases. When the bounded rationality level is high, i.e., the variance of the noise parameters, \( \epsilon_i \), is large, the firms’ advantages over each other diminish. Additionally, when the customers are more rational, the cost advantage of Firm 1 (in low reliability sensitivity levels) and the capacity advantage of Firm 2 (in high reliability sensitivity levels) sharply distinguish the two firms’ positioning in the market as shown in Figure 5.3b. Figures 5.3d-5.3f show a similar behavior for the firms’ profits for different combinations of customer bounded rationality levels and reliability sensitivity levels.
5.4.3 Comparison: The PTR and PT Competitions

In this subsection, we compare the PTR and PT competitions. So (2000) studies the PT competition. He assumes that customers choose the firms according to the MCI model without considering the balking option. He shows that as the time sensitivity increases, the profit and market share become increasingly beneficial to the firm with the capacity advantage. However, our results show that even a firm with a capacity advantage may suffer from an increase in the time sensitivity level when boundedly rational customers have the balking option. In order to have a meaningful comparison, we employ the MNL model to the PT competition. Later, we discuss how the MCI model influences the results.

First, we investigate the impact of the delivery reliability decision on the firms’ profitability. We compare the optimal profits of two firms under PTR and PT competitions and show that ignoring the delivery reliability decision may result in significant profit losses.

The Importance of the Delivery Reliability

In this subsection, we compare the optimal decisions, profits, and market shares of two firms in the PT and PTR competitions. The results are presented in Table 5.4 for different bounded rationality levels and market sizes and for the two firms described in Subsection 5.4.2, i.e., \((v_1, \mu_1, \gamma_1) = (2, 2.5, 0.2)\) and \((v_2, \mu_2, \gamma_2) = (2, 5.0, 0.8)\). For the PT competition, we fix the reliability level at \(q = 0.9\).

Table 5.4 shows that at each bounded rationality level, both firms offer a relatively higher price in the PTR competition compared to the PT competition. For instance, when \(\eta = 0.5\) and \(\Lambda = 1\), Firm 1 (Firm 2) charges the customers a 18.2% (19.2%) higher price in the PTR competition compared to the PT competition. For Firm 1 (Firm 2), this value increases to 324.6% (305.6%) when the market size increases to \(\Lambda = 7\). The delivery time and delivery reliability also demonstrate a similar behavior: they are higher in the PTR competition compared to the PT competition in this set of numerical examples.

When \(\eta = 0.5\) and \(\Lambda = 1\), Firm 1 (Firm 2) captures a market share in the PTR competition that is 16.5% (50.2%) higher compared to the PT competition. This shows how the two firms in the PTR competition can exploit their capacities to offer more reliable services (even with higher prices and longer delivery times) in order to attract the balking customers. As a result, both of the firms make relatively higher profits in the PTR competition compared to the PT competition. Table 5.4 reveals the importance of the delivery reliability as a strategic tool for the firms to position themselves in the market. This table also reveals that even a firm that is at a capacity disadvantage may
be better off in the PTR competition.

<table>
<thead>
<tr>
<th>η</th>
<th>A</th>
<th>Firm 1</th>
<th>Firm 2</th>
<th>Firm 1</th>
<th>Firm 2</th>
<th>Firm 1</th>
<th>Firm 2</th>
<th>Firm 1</th>
<th>Firm 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.5</td>
<td>50.2</td>
<td>42.1</td>
<td>108.8</td>
<td>19.2</td>
<td>56.0</td>
<td>92.7</td>
<td>7.7</td>
<td>9.6</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>35.5</td>
<td>81.5</td>
<td>166.9</td>
<td>351.7</td>
<td>83.3</td>
<td>79.2</td>
<td>66.0</td>
<td>135.8</td>
</tr>
<tr>
<td>7</td>
<td>56.7</td>
<td>83.2</td>
<td>635.5</td>
<td>1072.7</td>
<td>324.6</td>
<td>305.6</td>
<td>115.0</td>
<td>238.6</td>
<td>0.0</td>
</tr>
<tr>
<td>1.5</td>
<td>4</td>
<td>66.1</td>
<td>138.0</td>
<td>218.5</td>
<td>389.7</td>
<td>84.5</td>
<td>76.2</td>
<td>92.5</td>
<td>146.0</td>
</tr>
<tr>
<td>7</td>
<td>49.5</td>
<td>115.9</td>
<td>286.3</td>
<td>531.9</td>
<td>147.4</td>
<td>140.7</td>
<td>96.6</td>
<td>229.5</td>
<td>0.6</td>
</tr>
<tr>
<td>2.5</td>
<td>4</td>
<td>89.1</td>
<td>107.4</td>
<td>169.0</td>
<td>199.3</td>
<td>39.8</td>
<td>35.5</td>
<td>67.3</td>
<td>95.8</td>
</tr>
<tr>
<td>7</td>
<td>69.1</td>
<td>123.2</td>
<td>173.8</td>
<td>265.3</td>
<td>58.6</td>
<td>51.2</td>
<td>91.3</td>
<td>129.1</td>
<td>4.5</td>
</tr>
<tr>
<td>1</td>
<td>43.7</td>
<td>112.4</td>
<td>176.9</td>
<td>316.5</td>
<td>88.2</td>
<td>77.6</td>
<td>86.2</td>
<td>190.3</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table 5.4: Comparison of the PTR and PT competitions (c_p = 1, c_t = 0.5, c_{q} = 8, q = 0.9, (v_1, \mu_1, \gamma_1) = (2, 2.5, 0.2), and (v_2, \mu_2, \gamma_2) = (2, 5.0, 0.8))

**Balking Option: The Hidden Competitor (II)**

Figure 5.4 shows the profits and market shares of the two firms in the PTR competition with and without the balking option. When the customers cannot balk and the time sensitivity is relatively low, Firm 1 wins a higher market share and makes more profit than Firm 2. In this case, as the time sensitivity increases, Firm 2 can exploit its capacity advantage to increase its market share and profit (see Figures 5.4a and 5.4b). When the customers have the balking option, however, the results are different. As shown in Figure 5.4b, Firm 2’s market share first increases and then decreases as the time sensitivity increases. It first increases because Firm 2 attracts some of Firm 1’s customers. But when the time sensitivity is too high, more customers choose to balk. Thus, in contrast to So (2000)’s conclusion, a firm with a capacity advantage is not necessarily better off when customers are more time sensitive.

**5.4.4 The MCI Model**

In Subsection 5.4.3, we applied the MNL model to the PTR and PT competitions. In this subsection, we employ the MCI model to the PTR competition. Based on the MCI model, the probability that a customer chooses Firm i’s service equals

\[ \theta_i = \frac{v_ip_i^{c_p}t_i^{c_t}q_i^{c_{q}}}{\sum_{j=1}^{N} v_jp_j^{c_p}t_j^{c_t}q_j^{c_{q}}}. \]
Figure 5.4: The impact of the time sensitivity on the PTR competition (Λ = 1, cp = 1, cq = 8, q = 0.90, η = 1, (v1, μ1, γ1) = (2, 2.5, 0.2) and (v2, μ2, γ2) = (2, 5.0, 0.8))

and the demand rate captured by Firm i equals

\[ \lambda_i = \frac{\Lambda v_i p_i^{-c_p t_i^{-c_t}} q_i^{-c_q}}{v_i p_i^{-c_p t_i^{-c_t}} q_i^{-c_q} + \beta_i}, \quad (5.13) \]

where \[ \beta_i = \sum_{j=1}^{N} v_j p_j^{-c_p t_j^{-c_t}} q_j^{-c_q}. \] Now consider Firm i’s best response problem.

**Proposition 5.5.** An interior delivery reliability solution to Firm i’s best response problem exists only if \[ \frac{c_t}{c_q} \leq \left( \frac{1-q}{q} \right) \ln(\frac{1}{1-q}) \], and it is the unique root of \[ F(q) = 0 \] where

\[ F(q) = \frac{c_t}{c_q} q + (1-q) \ln(1-q). \quad (5.14) \]

Proposition 5.5 provides a number of interesting insights. First, when the ratio of the time sensitivity to the reliability sensitivity, \[ \frac{c_t}{c_q} \], is higher than a threshold value, no firm has an incentive to offer a reliability greater than the market entrance requirement. This observation is consistent with Corollary 5.1 for the MNL model. However, this threshold depends on both the market entrance requirement and the firm’s capacity level in the MNL model, i.e., \( (1-q) \mu_i \), whereas it only depends on the market entrance requirement in the MCI model, i.e., \( (1-q) \ln(\frac{1}{1-q}) \).

Moreover, \[ F(q) \] is independent of Firm i’s operational properties. In other words, the interior delivery-reliability level only depends on the market-specific parameters, i.e., the time and reliability sensitivities, and not on the firm-specific parameter, i.e., the capacity level. Hence, when all of the firms offer an interior delivery-reliability level, they will all offer the same delivery-reliability level. As a result, no firm can differentiate in the delivery reliability, and the competition is reduced to the PT competition. We leave the equilibrium analysis of the PTR competition under the MCI model for future studies.
5.5 Conclusion and Future Research

Empirical research has shown that price and delivery performance of a firm have significant impact on customers’ perceived utility. Two dimensions of the delivery performance are generally acknowledged in the literature: promised delivery time and delivery-reliability level. Although the interactions between price and these two delivery performance measures and their impact on customers’ purchase behavior have been studied previously, no research considers them together in a competitive setting.

In this study, we model the problem as competition among an arbitrary number of profit-maximizing firms facing boundedly rational customers who are sensitive to price, promised delivery time, and delivery-reliability level. Customers can choose to buy the service from one of the firms or balk. We show the existence of a unique Nash equilibrium. A simple iterative method, that converges to this equilibrium, is proposed to solve the competition problem. We thoroughly compare our results with those in the existing literature. We show that, in contrast to the literature, a firm with a higher capacity level is not always better off in a more reliability-sensitive market, and even a firm with a lower capacity level may benefit from a more time-sensitive market. In addition, the impacts of bounded rationality level and the balking option are addressed in depth in the numerical study. Several interesting managerial insights are also found and presented.

There are several directions for further research. In this study, the utility function is assumed to be linear. Specifically, the separability of the delivery time and delivery-reliability level may not capture the tight relationship between them. It would be interesting to consider other utility functions. Moreover, we have assumed that the customers are homogeneous. Taking the customers’ heterogeneity into account is another possible direction for future studies. In this case, the market can also be segmented, and price and delivery performance differentiation can be included in the model. Additionally, we model the operations of each firm as an $M/M/1$ queueing system. It would be interesting to study the problem in a setting with more general queueing systems i.e., $G/G/k$. Another direction is to consider the capacity competition. Delivery time, delivery-reliability level, and capacity competition is studied by Shang and Liu (2011). However, they assume that the price is exogenous. An integrated model may provide interesting results.
Appendix

5.A Proofs

Proof of Proposition 5.1. For the sake of simplicity, we drop the subscript $i$ for the remainder of this proof as long as it does not make any confusion. We start by proving part (b). From the first-order conditions, we have

$$
\frac{\partial \Pi(t,q|\beta_i)}{\partial t} = \frac{k}{t^2} \left( \frac{t^2}{k} (\mu - \frac{k}{t}) \frac{\partial p(t,q|\beta_i)}{\partial t} + p(t,q|\beta_i) - \gamma \right),
$$

$$
\frac{\partial \Pi(t,q|\beta_i)}{\partial q} = \frac{-1}{t(1-q)} \left( -t(1-q)(\mu - \frac{k}{t}) \frac{\partial p(t,q|\beta_i)}{\partial q} + p(t,q|\beta_i) - \gamma \right)
$$

where

$$
\frac{\partial p(t,q|\beta_i)}{\partial t} = \frac{-1}{c_p} \left( c_t + \frac{k \Lambda}{t^2(\Lambda - \mu + k/t)(\mu - k/t)} \right) < 0,
$$

$$
\frac{\partial p(t,q|\beta_i)}{\partial q} = \frac{1}{c_p} \left( c_q + \frac{\Lambda}{t(1-q)(\Lambda - \mu + k/t)(\mu - k/t)} \right) > 0.
$$

In the system of equations \( \frac{\partial \Pi(t,q|\beta_i)}{\partial t} = 0, \frac{\partial \Pi(t,q|\beta_i)}{\partial q} = 0 \) equating \( \left( \frac{c^2}{k} (\mu - \frac{k}{t}) \frac{\partial p(t,q|\beta_i)}{\partial t} \right) + p(t,q|\beta_i) - \gamma \) and \( \left( -t(1-q)(\mu - \frac{k}{t}) \frac{\partial p(t,q|\beta_i)}{\partial q} + p(t,q|\beta_i) - \gamma \right) \) results in

$$
t(q) = \frac{c_q (1-q)}{c_t} \ln \frac{1}{1-q}.
$$

(5.15)

To have an interior delivery-reliability level solution there should exist a solution to the first-order conditions. Substituting Eq. (5.15) in \( \frac{\partial \Pi(t,q|\beta_i)}{\partial t} = 0 \) results in

$$
\frac{\partial \Pi(t(q),q|\beta_i)}{\partial t} = \frac{k}{(t(q))^2} H(q|\beta_i) = 0.
$$

136
\[ H(q|\beta_i) \text{ is well-defined for } 1 - \frac{c_i}{c_q(\mu_i - \Lambda)} < q < 1 - \frac{c_i}{c_q\mu_i} \text{ when } \mu_i > \Lambda, \text{ and for } q < 1 - \frac{c_i}{c_q\mu_i}, \text{ when } \mu_i < \Lambda. \]  
\[ H(q|\beta_i) \text{ is decreasing in } q, \text{ for } 0.64 < q < 1, \text{ since} \]
\[
\frac{\partial H(q|\beta_i)}{\partial q} = \frac{c_i^2\mu}{c_t}(2q \ln \frac{1}{1-q} + 1 - q) - \frac{c_i\eta \Lambda}{c_q(1-q)^2(\Lambda - \mu + \frac{c_i}{c_q(1-q)})^2} - \frac{c_i\eta \Lambda}{c_q(1-q)^2(\Lambda - \mu + \frac{c_i}{c_q(1-q)})(\mu - \frac{c_i}{c_q(1-q)})} - c_q < 0.
\]

For (i) \( \Lambda > \mu_i \), the acceptable region for \( q \) is \( q < q < 1 - \frac{c_i}{c_q\mu_i} \), and we know that \( \lim_{q \to (1 - \frac{c_i}{c_q\mu_i})^-} H(q|\beta_i) = -\infty \). If \( \lim_{q \to 2} H(q|\beta_i) > 0 \), then there is a unique solution to \( H(q|\beta_i) = 0 \). For \( \Lambda < \mu_i \), the acceptable region for \( q \) is \( \max\{1 - \frac{c_i}{c_q(\mu_i - \Lambda)}, q\} < q < 1 - \frac{c_i}{c_q\mu_i} \). We have \( \lim_{q \to (1 - \frac{c_i}{c_q\mu_i}^-)} H(q|\beta_i) = -\infty \). For (ii) \( 1 - \frac{c_i}{c_q(\mu_i - \Lambda)} < q \), if \( \lim_{q \to 2} H(q|\beta_i) > 0 \), then there is a unique solution to \( H(q|\beta_i) = 0 \). For (iii) \( 1 - \frac{c_i}{c_q(\mu_i - \Lambda)} > q \), there exists a unique solution to \( H(q|\beta_i) = 0 \), because \( \lim_{q \to (1 - \frac{c_i}{c_q(\mu_i - \Lambda)}^+)} H(q|\beta_i) = +\infty \). Firm \( i \)'s profit function is unimodal in \( q_i \) (first increasing and then decreasing). Due to this fact, when an interior solution for delivery-reliability level exists, the boundary delivery-reliability level, i.e., \( q \), cannot be optimal. This completes the proof to part (b).

When there is no interior solution for the delivery-reliability level, the firm has to set the boundary value, \( q \). Substitute \( q = \bar{q} \) in \( \partial p(t, q|\beta_i) \) and \( \lambda_i \). Taking the first derivative of the objective function with respect to \( t \) results in
\[
\frac{\partial \Pi(t, q)}{\partial t} = \frac{k}{t^2} G(t|\beta_i) = 0.
\]
\( G(t|\beta_i) \) is increasing in \( t \), because
\[
\frac{\partial G(t|\beta_i)}{\partial t} = \frac{2c_i\mu t}{k} + \frac{\eta k}{t^2(\Lambda - \mu_i + \frac{c_i}{c_q})^2} + \frac{\eta k}{t^2(\mu_i - \frac{c_i}{c_q})^2} > 0.
\]
Moreover, \( \lim_{t \to (\frac{k}{\lambda})^+} G(t|\beta_i) = -\infty \) and \( \lim_{t \to +\infty} G(t|\beta_i) = +\infty \) when \( \Lambda > \mu_i \), and \( \lim_{t \to (\frac{k}{\lambda})^-} G(t|\beta_i) = +\infty \) when \( \Lambda < \mu_i \). Therefore, there is a unique solution to \( G(t|\beta_i) = 0 \). This completes the proof to part (c). It also shows that for any given delivery-reliability level, the optimal delivery time can be uniquely determined from the first-order condition with respect to \( t \).

Note that from \( \frac{\partial \Pi(t, q)}{\partial q} = 0 \) and Eq. (5.15), the price can be found as
\[
p(q) = \frac{1}{c_p} \left( \frac{c_i^2(1-q)^2}{c_t}(\mu - \frac{c_i}{c_q(1-q)}) \ln \frac{1}{1-q} + \frac{\eta \Lambda}{(\Lambda - \mu + \frac{c_i}{c_q(1-q)})} + c_p \gamma \right), \quad (5.16)
\]
137
and from \( \frac{\partial \Pi(t, q)}{\partial t} = 0 \) the price is equal to

\[
p(t) = \frac{1}{c_p} \left( \frac{c_t^2}{k} \left( \mu - \frac{k}{t} \right) + \frac{\eta \Lambda}{\left( \Lambda - \mu + \frac{k}{t} \right)} + c_p \gamma \right).
\] (5.17)

In either case of an interior or a boundary delivery-reliability level, according to Eq. (5.16) and Eq. (5.17), the optimal price is greater than \( \gamma \), and therefore constraint (5.7) is met.

Parts (b) and (c), together, uniquely determine the optimal solution. This completes the proof.

**Proof of Proposition 5.2.** The optimal delivery-reliability level is either an interior point, i.e., \( q \in (q, 1) \), or on the boundary, i.e., \( q = q \).

*Interior delivery-reliability level:* If the optimal delivery-reliability level is an interior point, according to Proposition 5.1, it is the unique root of \( H(q|\beta) = 0 \). We have

\[
\frac{\partial q}{\partial \beta} = -\frac{\partial H(q|\beta)}{\partial q} > 0,
\]

because \( \frac{\partial H(q|\beta)}{\partial \beta} = \frac{1}{\beta} > 0 \), and \( \frac{\partial H}{\partial q} < 0 \) as shown in Proof of Proposition 5.1. In addition, the optimal delivery time is found from Eq. (5.15). We have

\[
\frac{\partial t(q)}{\partial q} = \frac{c_q}{c_t} \left( 1 - \ln \frac{1}{1 - q} \right) < 0,
\]

because \( \ln \frac{1}{1 - q} > 1 \) for \( q > q > 0.64 \). Thus,

\[
\frac{\partial t(q)}{\partial \beta} = \frac{\partial t(q)}{\partial q} \frac{\partial q}{\partial \beta} < 0.
\]

According to Eq. (5.16), for all \( q > q > 0.64 \), we have

\[
\frac{\partial p(q)}{\partial q} = \frac{1}{c_p} \left( \frac{c_t^2}{c_t} \left( 1 - \ln \frac{1}{1 - q} \right) \left( 1 - q \right) \left( \mu - \frac{c_t}{c_q(1 - q)} \right) - \frac{c_q^2 \mu}{c_t} (1 - q) \ln \frac{1}{1 - q} - \frac{\eta \Lambda c_t}{c_q(1 - q)^2} \left( \Lambda - \mu + \frac{c_t}{c_q(1 - q)} \right)^2 \right) < 0.
\]

Therefore,

\[
\frac{\partial p(q)}{\partial \beta} = \frac{\partial p(q)}{\partial q} \frac{\partial q}{\partial \beta} < 0.
\]

*Boundary delivery-reliability level:* If the optimal delivery-reliability level is determined by \( q \), according to Proposition 5.1, the optimal delivery time is the unique root of
Therefore, according to Proposition 5.2, we have
\[ \frac{\partial t}{\partial \beta} = -\frac{\partial G(t|\beta_i)}{\partial \beta} < 0, \]

because \( \frac{\partial G(t|\beta_i)}{\partial \beta} = \frac{1}{\beta} > 0, \) and \( \frac{\partial G(t|\beta_i)}{\partial t} > 0. \) According to Eq. (5.17), we have
\[ \frac{\partial p(t)}{\partial t} = \frac{1}{c_p} \left( \frac{2c_i}{k} t - \frac{k}{\mu} + c_i + \frac{\Lambda k}{t^2(\Lambda - \mu + \frac{k}{\mu})} \right) > 0. \]

Therefore,
\[ \frac{\partial p(t)}{\partial \beta} = \frac{\partial p(t)}{\partial t} \frac{\partial t}{\partial \beta} < 0. \]

Thus, it can be concluded that the optimal price and delivery time are strictly decreasing in \( \beta. \) If the conditions stated in Proposition 5.1, part (a) are not satisfied, then the optimal delivery-reliability level would be the boundary point and invariant with respect to \( \beta. \) The optimal delivery-reliability level is therefore increasing in \( \beta. \) \( \square \)

**Proof of Proposition 5.3.** This proof follows from Proof of Proposition 4 in So (2000). We prove this proposition in two parts: proof of the convergence and proof of the uniqueness.

**Proof of the convergence:** We use inductive reasoning to prove the convergence of the procedure. Define \( A_i = e^{U_i/\eta} \) as the attraction of Firm \( i. \) Let \( p_i^{(j)}, t_i^{(j)} \), and \( q_i^{(j)} \) denote the price, delivery time, and delivery-reliability level of Firm \( i \) in the \( j \)-th iteration. Set \( p_i^{(0)} = \gamma_i, \ t_i^{(0)} = \frac{k}{\mu}, \) and \( q_i^{(0)} = 1. \) Obviously, \( p_i^{(j)} > \gamma_i, \ t_i^{(j)} > \frac{k}{\mu}, \) and \( q_i^{(j)} < 1 \) for all \( j. \) Particularly, \( p_i^{(1)} > p_i^{(0)}, \ t_i^{(1)} > t_i^{(0)} \) and \( q_i^{(1)} < q_i^{(0)}. \) Assume \( p_i^{(j)} \geq p_i^{(j-1)}, \ t_i^{(j)} \geq t_i^{(j-1)} \) and \( q_i^{(j)} \leq q_i^{(j-1)}. \) Consider \( j = n. \) \( \beta_i^{(j)} \) denotes the combined impact of the other firms’ actions on Firm \( i \)'s problem at the \( j \)-th iteration. For Firm 1 we have,
\[
\beta_1^{(n)} = \sum_{m=2}^{N} e^{U_m/\eta} + 1
= \sum_{m=2}^{N} e^{(\nu_m-a_{p_m}^{(n-1)}-b_{t_m}^{(n-1)}+c_{q_m}^{(n-1)})/\eta} + 1
\leq \sum_{m=2}^{N} e^{(\nu_m-a_{p_m}^{(n-2)}-b_{t_m}^{(n-2)}+c_{q_m}^{(n-2)})/\eta} + 1 = \beta_1^{(n-1)}.
\]

Therefore, according to Proposition 5.2, we have \( p_1^{(n)} \geq p_1^{(n-1)}, \ t_1^{(n)} \geq t_1^{(n-1)} \) and \( q_1^{(n)} \leq q_1^{(n-1)}, \) because \( \beta_1^{(n)} \leq \beta_1^{(n-1)} \).

Now, suppose that \( p_i^{(n)} \geq p_i^{(n-1)}, \ t_i^{(n)} \geq t_i^{(n-1)} \) and \( q_i^{(n)} \leq q_i^{(n-1)} \) for all \( i = 1, 2, \ldots, s-1. \)
For Firm \( s \) we have

\[
\beta_{s}^{(n)} = \sum_{m=1}^{s-1} e^{\frac{U_m}{\eta}} + \sum_{m=s+1}^{N} e^{\frac{U_m}{\eta}} + 1
\]

\[
= \sum_{m=1}^{s-1} e^{(v_m - a p_m^{(n)} - bt_m^{(n)} + c q_m^{(n)})/\eta} + \sum_{m=1}^{s-1} e^{(v_m - a p_m^{(n-1)} - bt_m^{(n-1)} + c q_m^{(n-1)})/\eta} + 1
\]

\[
\leq \sum_{m=1}^{s-1} e^{(v_m - a p_m^{(n-1)} - bt_m^{(n-1)} + c q_m^{(n-1)})/\eta} + \sum_{m=1}^{s-1} e^{(v_m - a p_m^{(n-2)} - bt_m^{(n-2)} + c q_m^{(n-2)})/\eta} + 1 = \beta_{s}^{(n-1)}.
\]

Similarly, according to Proposition 5.2, we have \( p_{s}^{(n)} \geq p_{s}^{(n-1)}, t_{s}^{(n)} \geq t_{s}^{(n-1)} \) and \( q_{s}^{(n)} \leq q_{s}^{(n-1)} \), because \( \beta_{s}^{(n)} \leq \beta_{s}^{(n-1)} \).

From Proposition 5.1, for a given \( \beta_{i} \), the decisions of Firm \( i \) are uniquely determined. We also know that through the above-mentioned iterative procedure, \( \beta_{i} \) decreases for each firm in each step. Moreover, the balking option implies that there is a lower bound for \( \beta_{i} > 1 \). Thus, the iterative procedure converges.

**Proof of the uniqueness:** We prove the uniqueness by contradiction. Suppose there exist two equilibrium solutions denoted by \( \phi = (x_1, \ldots, x_N) \) and \( \phi' = (x'_1, \ldots, x'_N) \) with corresponding optimal solutions \( (p_i, t_i, q_i) \) and \( (p'_i, t'_i, q'_i) \). Let the attractions of each firm in the two equilibrium solutions be expressed as \( A'_i = (1 + r_i) A_i \). Without loss of generality, the firms can be numbered such that \( r_1 \geq r_2 \geq \ldots \geq r_N \) and \( r_1 > 0 \). Next, we show that \( r_2 \) must be positive, i.e., \( r_2 > 0 \).

Assume \( r_2 \leq 0 \). It implies \( A'_i \leq A_i \) for \( i = 2, \ldots, N \), and subsequently \( \beta'_1 \leq \beta_1 \). From Proposition 5.2, we have \( p'_1 \geq p_1, t'_1 \geq t_1 \) and \( q'_1 \leq q_1 \). This implies

\[
A_1 = e^{(v_1 - a p_1 - b t_1 + c q_1)/\eta} \geq e^{(v_1 - a p'_1 - b t'_1 + c q'_1)/\eta} = A'_1 = (1 + r_1) A_1,
\]

that contradicts \( r_1 > 0 \). Therefore, we must have \( r_2 > 0 \).

Consider Firm 1’s problem. Assume \( r_i = r_2 \) for \( i = 2, \ldots, N \). Denote the corresponding equilibrium solution by \( \phi'' = (x''_1, \ldots, x''_N) \) where \( A''_i = (1 + r_2) A_i \) for \( i = 2, \ldots, N \). Let \( (p''_1, t''_1, q''_1) \) represent the corresponding optimal solution. From Eq. (5.3) and Eq. (5.5), for \( \phi \) and \( \phi'' \) we have

\[
\mu_1 = \frac{\Lambda A_1}{A_1 + \sum_{i=2}^{N} A_i} = \frac{k_1}{t_1},
\]

140
and
\[ \mu_1 - \frac{\Lambda A''_1}{A''_1 + \sum_{i=2}^{N} A''_i} = \mu_1 - \frac{\Lambda A''_1}{A''_1 + (1 + r_2) \sum_{i=2}^{N} A_i} = \mu_1 - \frac{\Lambda A''_1}{A''_1 + (1 + r_2) + \sum_{i=2}^{N} A_i} = \frac{k''}{t''}. \]

Since \( r_2 > 0 \), \( A''_i > A_i \) for \( i = 2, \ldots, N \), and subsequently \( \beta''_1 > \beta_1 \). From Proposition 5.2, we have \( p''_i < p_i, t''_i < t_i \) and \( q''_i \geq q_i \) (equivalently \( k''_i \geq k_i \)), which gives \( \frac{k''}{t''} > \frac{k}{t} \). It implies
\[ \mu_1 - \frac{\Lambda A''_1}{A''_1 + (1 + r_2) + \sum_{i=2}^{N} A_i} > \mu_1 - \frac{\Lambda A_1}{A_1 + \sum_{i=2}^{N} A_i} \]
or equivalently
\[ A''_1 < (1 + r_2)A_1. \] (5.18)

Again, consider Firm 1’s problem. Under equilibrium solution \( \phi' \), we have \( A'_i \leq A''_i \) for \( i = 2, \ldots, N \), because \( r_i \leq r_2 \) for \( i = 3, \ldots, N \). This implies \( \beta'_1 \leq \beta''_1 \). From Proposition 5.2, we know \( p''_i \leq p'_i, t''_i \leq t'_i \) and \( q''_i \geq q'_i \). Thus, we have
\[ A''_1 = e^{(v_1 - ap''_1 + bt''_1 + cq''_1)/n} \geq e^{(v_1 - ap'_1 + bt'_1 + cq'_1)/n} = A'_1 = (1 + r_1)A_1. \] (5.19)

From Eq. (5.18) and Eq. (5.19), we have
\[ (1 + r_1)A_1 = A'_1 \leq A''_1 < (1 + r_2)A_1, \] (5.20)
that implies \( r_1 < r_2 \) and contradicts the assumption \( r_1 \geq r_2 \). Therefore, the equilibrium solution must be unique. \( \square \)

**Proof of Proposition 5.4.** The proof is immediate from Proposition 5.1 and thus omitted. \( \square \)

**Proof of Proposition 5.5.** Substituting Eq. (5.1) in Eq. (5.13) results in
\[ p_i(t_i, q_i|\beta_i) = \left( \frac{(M_i - \mu t_i + k_i) t_i^{-\alpha_i} q_i^{\alpha_i}}{\beta_i (\mu t_i - k_i)} \right)^{\frac{1}{\alpha_i}}. \] (5.21)

The objective function is similar to Eq. (5.12) where \( p_i(t_i, q_i|\beta_i) \) is defined by Eq. (5.21).
We drop the subscript $i$ for the rest of this proof. The first-order conditions are

\[
\frac{\partial \Pi(t, q|\beta_i)}{\partial q} = -\frac{p(t, q|\beta_i)}{c_p t} G(t, q|\beta_i) = 0,
\]

\[
\frac{\partial \Pi(t, q|\beta_i)}{\partial t} = \frac{kp(t, q|\beta_i)}{c_p t^2} H(t, q|\beta_i) = 0,
\]

in which

\[
G(t, q|\beta_i) = \frac{-\Lambda t}{\Lambda t - \mu t + k} + c_p - \frac{c_p \gamma}{p(t, q|\beta_i)} - \frac{c_q (1-q)(\mu t - k)}{q},
\]

\[
H(t, q|\beta_i) = \frac{-\Lambda t}{\Lambda t - \mu t + k} + c_p - \frac{c_p \gamma}{p(t, q|\beta_i)} - \frac{c_t (\mu t - k)}{k}.
\]

Obviously, the system of the first-order conditions \((\frac{\partial \Pi(t, q|\beta_i)}{\partial q} = 0, \frac{\partial \Pi(t, q|\beta_i)}{\partial t} = 0)\) is equivalent to the system \((G(t, q|\beta_i) = 0, H(t, q|\beta_i) = 0)\). Equating \(G(t, q|\beta_i)\) and \(H(t, q|\beta_i)\) results in \(F(q) = 0\).

\(F(q)\) is strictly convex, i.e., \(\frac{d^2 F(q)}{dq^2} = \frac{1}{1-q} > 0\). Furthermore, we have \(\frac{dF(q)}{dq} = \frac{c_t}{c_q} - 1 - \ln(1-q), F(0) = 0,\) and \(F(1) = \frac{c_t}{c_q}\). For \(\frac{c_t}{c_q} \geq 1, \frac{dF(q)}{dq} > 1,\) and \(F(q)\) is strictly increasing. Since \(F(0) = 0,\) there is no positive solution to \(F(q) = 0\). For \(\frac{c_t}{c_q} < 1, F(q)\) is first decreasing and then increasing, and since \(F(0) = 0,\) a unique positive solution to \(F(q) = 0\) must exist. It is obvious that if \(F(q) > 0\) or equivalently \(\frac{(1-q)}{q} \ln\left(\frac{1}{1-q}\right) < \frac{c_t}{c_q}\), there exists no solution to \(F(q) = 0\) within \([q, 1]\). It is easy to conclude that if \(\frac{c_t}{c_q} \leq \frac{(1-q)}{q} \ln\left(\frac{1}{1-q}\right)\), there exists a unique positive solution to \(F(q) = 0\) within \([q, 1]\). \(\square\)


Declaration of co-authorship

Full name of the PhD student: Ata Jalili Marand

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>Joint Inventory Control and Pricing in a Service-Inventory System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Ata Jalili Marand, Hongyan Li, Anders Thorstenson</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published ■ Accepted □ Submitted □ In preparation □


If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ■ Yes □ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>B</td>
</tr>
<tr>
<td>2. Planning of the experiments/methodology design and development</td>
<td>B</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>A</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>B</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>B</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>A</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.08.2018</td>
<td>Hongyan Li</td>
<td></td>
</tr>
<tr>
<td>21.08.2018</td>
<td>Anders Thorstenson</td>
<td></td>
</tr>
</tbody>
</table>

In case of further co-authors please attach appendix

Date: 30.08.2018

Ata Jalili Marand

Signature of the PhD student
Declaration of co-authorship

Full name of the PhD student: Ata Jalili Marand

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>Quandary of Service Logistics: Fast or Reliable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Ata Jalili Marand, Ou Tang, Hongyan Li</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: European Journal of Operational Research

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☐ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

| A. Has essentially done all the work |
| B. Major contribution                |
| C. Equal contribution                |
| D. Minor contribution                |
| E. Not relevant                      |

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>A</td>
</tr>
<tr>
<td>2. Planning of the experiments/methodology design and development</td>
<td>B</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>A</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>B</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>B</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>A</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>08.08.17</td>
<td>Ou Tang</td>
<td></td>
</tr>
<tr>
<td>18.08.18</td>
<td>Hongyan Li</td>
<td></td>
</tr>
</tbody>
</table>

In case of further co-authors please attach appendix

Date: 30.08.2018

Ata Jalili Marand

Signature of the PhD student
Declaration of co-authorship

Full name of the PhD student: Ata Jalili Marand

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>Price, Time, and Reliability Competition for Service Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Ata Jalili Marand, Hongyan Li, Anders Thorstenson</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published □ Accepted □ Submitted ■ In preparation □

If published, state full reference:

If accepted or submitted, state journal: Service Science

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No □ Yes ■ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>B</td>
</tr>
<tr>
<td>2. Planning of the experiments/methodology design and development</td>
<td>A</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>A</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>B</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>B</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>A</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.08.2018</td>
<td>Hongyan Li</td>
<td></td>
</tr>
<tr>
<td>21.08.2018</td>
<td>Anders Thorstenson</td>
<td></td>
</tr>
</tbody>
</table>

In case of further co-authors please attach appendix

Date: 30.08.2018

Signature of the PhD student

Ata Jalili Marand