Tiered Pricing for Volume and Priority:
Three Problems at the Intersection of Marketing and Operational Policies

by

Michael Pavlin

Graduate Department of the Joseph L. Rotman School of Management
University of Toronto

Copyright © 2012 by Michael Pavlin
Abstract

Tiered Pricing for Volume and Priority:

Three Problems at the Intersection of Marketing and Operational Policies

Michael Pavlin

Doctor of Philosophy

Graduate Department of the Joseph L. Rotman School of Management

University of Toronto

2012

This thesis addresses three problems where a focal agent’s operational policies (inventory and capacity allocation) interact with marketing decisions.

The first chapter studies how wholesale all-unit discounts may lead to products being shifted from authorized retailers to discounted gray market channels. Such discounts lead to discontinuous order-costs which may induce buyers to order up to a threshold where they receive a greater discount. The buyer in this chapter is a reseller who makes purchasing decisions while taking into account inventory holding costs, how their resale price affects consumer demand and whether or not they divert inventory to the gray market. I analyze factors which determine how the reseller balances between lowering resale prices and diverting to the gray market, both of which lower costs by shortening the time inventory is held. Modelling the decisions as a Stackelberg game, the welfare of the authorized channel participants is analyzed. Of import, consumer welfare may decrease if a gray market emerges when holding costs are low.

In the latter two chapters, the supplier sells a congested service. For example, this supplier may be a courier facing stochastic buyer arrivals. Buyers vary in their value for the service and how patient they are, so the supplier may improve outcomes by providing a menu of delay levels and prices. The system is modelled as a priority queue where congestion constrains the arrival rates at each delay level.

In the first study, the supplier has aggregate market data. I model the problem as an
optimization subject to incentive and congestion constraints. The novel contributions include a precise description of the optimal menu as a function of the supplier’s capacity (the rate at which buyers can be served). Findings include existence of distinct capacity regions where the supplier utilizes service pooling and strategic delay.

In the final chapter the related welfare maximization problem is considered. Sufficient conditions for optimal pricing are derived which depend only on operational information: the current revenue must be equal to the best-case revenue subject to current prices and congestion constraints. An associated performance measure is shown to bound deviation from maximum welfare and is used as a heuristic within an adaptive pricing protocol. This protocol is shown to converge to near welfare maximizing outcomes.
Acknowledgements

I would first like to acknowledge the support and guidance of my supervisor Joseph Milner as well as Philipp Afèche and Ming Hu throughout the conception, analysis and writing of this document. I would also like to thank Mengze Shi who helped me bridge the divide between marketing and operations research. Dmitry Krass provided many invaluable comments and a huge amount of support throughout the research and writing of this thesis. I would also like to thank my external examiner, Professor Edieal Pinker of the University of Rochester, for his inciteful comments. Throughout my studies, the entire Operations Management faculty has provided an environment full of collegiality, humour and exceptional scholarship for which I am very grateful. This work was supported by scholarships from NSERC and the PhD program at the Rotman School of Management.

Of course this document could not be in its present form without the support of my family who put up with many years of me being huddled in front of a computer. My Mom, Dad and Jessica have been a fount of support and wisdom throughout. I dedicate this thesis to my wife Joanne and my daughter Emily who showed up a little under two years ago to cheer me on!
## Contents

1 Thesis Introduction ........................................... 1

2 All-Unit Discounts & Gray Markets .......................... 4
   2.1 Introduction ........................................... 4
   2.2 Model and Preliminary Analysis .......................... 10
   2.3 Model Analysis: Exogenous Resale Price .............. 20
   2.4 Model Analysis: Endogenous Resale Price .......... 32
   2.5 Model Extension: Endogenous Gray Market Price .. 51
   2.6 Conclusions ........................................... 55

3 Revenue Maximizing Lead-Time Pricing ..................... 58
   3.1 Introduction ........................................... 58
   3.2 Model, Formulation and Analysis Roadmap .......... 64
   3.3 Feasible Customer Segmentation and Lead Times, Optimal Prices .. 71
   3.4 Optimal Menus: Strictly Increasing Virtual Delay Cost Functions .. 81
   3.5 Numerical Examples: Strictly Increasing Virtual Delay Cost Functions .. 105
   3.6 Discussion and Extensions .............................. 111
   3.7 Conclusion ............................................. 118
   3.8 Proof of Proposition 7 .................................. 120

4 Multiple Service Level Assignment Problem ............... 136
   4.1 Introduction ........................................... 136
Chapter 1

Thesis Introduction

Quantity and priority are generally considered to have one thing in common: more is better. How much better often depends on the customer and this presents a supplier with an opportunity to target a range of products differentiated by either quantity or priority to different customer segments. The simplest mechanism to translate this range of products into revenue is to present a menu of priced products and allow the customer to decide amongst the options. This thesis studies the optimal design of such a menu in light of operational considerations of customer or supplier. The common thread throughout this work is that taking into account operational factors can lead to improvements in a range of firm objectives including system wide welfare maximization, management of sales channels, and firm cash-flows. In the first chapter, we focus on the latter two of these objectives. We introduce the reader to a supply chain model where an intermediate reseller may partition sales between an authorized channel and a less profitable secondary (gray) market with better availability. We investigate how a discontinuous “all-unit discount” interacts with the reseller’s inventory, pricing and gray-market diversion decisions. The second and third chapters both deal with the following problem: how should a service provider facing customers heterogenous in their value for the service and their sensitivity to delay design a price-lead time menu. Chapter 3 considers the revenue maximization objective in a model. Results include identifying operational and market conditions under which strategies such as pooling and strategic delay are
optimal. The final chapter revisits this problem from the point of view of a welfare maximizing provider. We present analytical results characterizing welfare maximizing prices and develop a bound on deviation from welfare maximization. Finally, we implement an iterative price adjustment mechanism based on minimizing this bound which is implementable by a provider with limited market information.

To be more precise, the first chapter studies a distribution channel that consists of a supplier who offers all-unit quantity discounts for batch orders to enjoy cost savings, and a reseller who finds it costly to hold inventory and may divert some goods to the gray market. Gray markets are unauthorized channels of distribution for a supplier’s authentic products. Our analysis shows the impact of gray markets depends on the reseller’s batch inventory holding cost. When the reseller’s batch inventory holding cost is high, diversion to the gray markets improves the channel performance by enabling the reseller to make batch orders. Since the reseller’s order costs decrease through quantity discounts, diversion to the gray markets reduces the resale price and expands sales to the authorized channel. On the other hand, when the reseller’s batch inventory holding cost is low, the reseller would make the batch orders even without the gray markets. In this case the diversion to the gray markets may improve the reseller’s performance by shortening the order cycles and reducing the inventory holding costs. Since diversion to the gray markets decreases the reseller’s peak inventory volume, the reseller has the reduced incentive to push its inventory and, consequently, the resale price rises and sales volume decreases in the authorized channel. These results underscore the importance of firms integrating inventory and pricing decisions when managing distribution channels affected by gray markets.

In the third and fourth Chapters we turn our attention to the design of price-lead time menus when faced with heterogeneous time-sensitive customers with private information about their price and time-sensitivity. In Chapter 3 we study this problem when the provider is attempting to maximize revenues. Our findings are as follows. (i) The optimal lead time menu may deviate in two fundamentally different ways from the optimal menu when restricted to a work conserving strict priority policy. (i) Pooling of multiple different customer
types into a single service class. (ii) Strategic delay to artificially inflate the lead times of certain service classes. We identify conditions for such pooling and strategic delay to be optimal depending on the capacity level and market parameters. (ii) We find that two sets of virtual delay costs play an important role in determining the optimal customer segmentation and lead time menu. The functional form of these delay costs differ depending on which subset of customers are admitted. For monotone virtual delay cost functions pooling may be optimal for a single service class with medium lead time. For non-monotone virtual delay cost functions such pooling may be optimal for multiple service classes with short or long lead times. Whether this occurs depends on the level of customer valuation, the capacity level and the relative mix of high and low delay cost customers. (iii) Under the conditionally optimal menu given a work conserving strict priority policy, it may be optimal at high capacity levels to strategically exclude low value customers which would otherwise be served with strategic delay under the optimal menu.

Chapter 4 focuses on the implementation of a welfare maximizing menu. The welfare maximizing provider’s objective is not a function of price. This allows the underlying assignment problem to be studied without considering customer incentives. We formulate this problem when the provider is restricted to offering a finite menu of fixed service levels. Results of our analysis include existence of prices which are incentive compatible with the optimal assignment. Equally importantly, we show that these prices can be identified without requiring knowledge of the distribution of customer preferences. To identify these prices it is sufficient to nullify a measure called the apparent-loss which is the difference between actual revenue and the best-case feasible revenue. We also show that for any prices and incentive compatible assignment the apparent loss bounds the deviation from maximum welfare. These results are leveraged to develop a heuristic pricing decision criteria that lead to welfare maximizing menus but require little market intelligence or information elicitation from customers. In numerical examples this adaptive pricing protocol is shown to be an effective means of discovering optimal assignments when market information is not available.
Chapter 2

All-Unit Discounts & Gray Markets

2.1 Introduction

The diversion of branded goods to unauthorized channels, also known as gray markets, is of substantial strategic interest to manufacturers. Industry reports show that gray market channels account for a significant portion of markets in a broad set of industries ranging from pharmaceuticals to consumer electronics. According to a KPMG study, the gray market accounted for as much as $58 billion in the 2008 global information technology (IT) hardware industry (KPMG LLP, 2008). Gray market activity persisted in the IT hardware industry mainly because resellers exploited incentive programs offered by IT original equipment manufacturers (OEMs). The study shows that 90 percent of the OEM respondents offered incentives to their channel partners and customers. The two most popular incentive programs identified are promotional deals and quantity discounts, respectively adopted by 72 percent and 59 percent of the surveyed OEMs. The study indicates that channel partners, in order to obtain deeply discounted products, may deceive the OEMs into offering incentives on products for non-existent customers and then divert those products to the gray market. While a precise breakdown of the volume diverted through these means is not available, the survey identifies 24 percent of resellers as admitting to selling to unauthorized channels and gray market brokers.
The focus of this chapter is on the operational and marketing issues associated with the gray market diversions induced by quantity discounts. Suppliers often implement quantity discounts with the intent of promoting operational goals (Munson and Rosenblatt, 1998). Quantity discounts, in particular all-unit discounts, provide an incentive for the reseller to order in a manner consistent with efficient shipping and manufacturing. From this operational perspective, gray market and quantity discounts can be beneficial by allowing the supply chain to achieve economies of scale. The reseller can order up to an efficient threshold and then use the gray market to divert excess inventory. Our research goals are to understand how operational factors (e.g., inventory costs for the reseller and scale economies for the supplier) interact with the marketing problems (e.g., reseller pricing and supplier all-unit discounts) in markets where the resellers can divert to the gray market. We also investigate how the presence of the gray market affects the profits of the resellers and sales in the authorized channel.

The practice of offering all-unit discounts is consistent with both the operational perspective taken in this chapter and the “situation on the ground”. Munson and Rosenblatt (1998) reported that economies of scale in production and transportation are the predominant drivers of quantity discounts. They also identified all-unit discounts as the most prevalent quantity discount offered by suppliers (utilized by 94 percent versus 34 percent for incremental discounts). An incremental discount applies only to the units above the volume-threshold. In contrast, an all-unit discount applies to each and every unit purchased as soon as the volume-threshold is achieved. Like Lal and Staelin (1984), our model builds on the assumption that the supplier enjoys cost benefits when the resellers order in batches. Such cost savings can result from operational efficiency gained in batch handling and transportation. Our analysis indeed shows that only a schedule of all-unit discounts will lead resellers to order beyond the market demand to enjoy the quantity discount and divert the excess to the unauthorized gray market channel. In addition to the KPMG LLP (2008) survey, a wide variety of industrial literature has indicated reseller over-ordering in response to supplier’s discount pricing practices as a key driver for a persistent gray market (Lowe
and McCrohan, 1988; Jorgenson, 1999; Gilroy, 2004; Eban, 2005). An early example which
is commonly cited took place in the personal computer market when resellers’ responses to
quantity discount became a concern for IBM (Ramirez, 1985). The following quotation from
Gilroy (2004) on the subject of car audio equipment suppliers illustrates that this behavior
is not limited to the IT sector.

Suppliers place an enormous pressure on resellers to make product quotas in
order to receive volume discounts. Sometimes resellers dump overstock on the
Internet to make their quotas, admitted suppliers. “A lot of times [dealers] say
that the pressure placed on them by the supplier is very great”.

Despite its prevalence in practice, to the best of our knowledge, the quantity-discount
induced gray market has not been rigorously studied in the literature. This chapter ad-
dresses this gap by performing an economic analysis of a rich operational model with a
reseller responding to a supplier’s all-unit quantity discount offerings. We take the premise
that a supplier manages the gray market through tolerance of violation (Dutta et al., 1994;
Bergen et al., 1998; Antia et al., 2004). In other words, the supplier chooses not to pur-
sue enforcement through monitoring and legal action. Instead, the supplier anticipates the
reseller’s access to the gray market and formulates the pricing strategy accordingly. Both
the supplier and the reseller anticipate that the gray market, unlike other salvage channels,
can cannibalize their sales in the authorized channels. The tolerance of violation is consis-
tent with the estimate by KPMG LLP (2003, 2008) that, from 2003 to 2008, the IT gray
market experienced a 45 percent increase from $40 billion to $58 billion. Even in ostensibly
well-regulated industries such as pharmaceuticals, abuse of promotions has proved difficult
to counteract. Eban (2005) reports that up to four fifths of American nursing homes and
similar healthcare institutions take advantage of wholesaler discounting practices to profit
off sales of prescription medications to gray market channels. With this in mind, it is not
surprising that in the IT sector, 42 percent of OEMs still have no process to identify or
monitor gray market activity (KPMG LLP, 2003).
Gray market supply may also result from excess stock due to demand uncertainty (Ahmadi et al., 2010; Altug and van Ryzin, 2009; Xia and Bassok, 2005). When realized demand is below expectations, the gray market provides a channel for resellers to dispose of the overstock. This driver is not incompatible with diversion resulting from all-unit discounts discussed in this chapter. In practice we expect both drivers to contribute to facilitating the gray market. The setup leading to gray market diversion driven by demand uncertainty can be modeled by a newsvendor model and is particularly appropriate when goods are perishable and demand is periodic. This setup lends itself to seasonal and fashion goods. Both drivers may contribute to domestic gray market supplies as well as parallel imports. In the latter case, the drivers will act to reinforce price differences between countries leading to arbitrage opportunities resulting in parallel imports (Duhan and Sheffet, 1988; Ahmadi and Yang, 2000). Authorized resellers operating in the same channel and/or location as the gray market will compete with the gray market for customers. The extent of cannibalization will depend on the price sensitivity of customers, the degree of differentiation between the gray market and authorized market good (e.g. the gray market often lacks a warranty), and the trust in the gray market (e.g. the customer may worry that the product is counterfeit). Cannibalization with authorized market demand is what differentiates the gray market from a typical model of a salvage channel where the primary and secondary markets are often treated independently. Finally, while both the domestic gray market and parallel imports deal with authentic products sold through unauthorized channels, counterfeit products do not originate from the trademark owner and therefore are not authentic (Duhan and Sheffet, 1988).

2.1.1 Literature and Positioning

Our model and analysis integrate the operational and marketing decisions. First, our model considers a reseller optimizing over lot-sizing and resale price decisions when facing a gray market and all-unit quantity discounts. The closed-form analysis of the reseller’s dynamic lot-sizing problem yields a novel solution that links the cost of holding inventory to the
supply of goods to the gray market. Specifically, we find gray market diversion occurs only in a middle range of the batch inventory holding cost. Within this range, the reseller finds it beneficial to use the gray markets to reduce its inventory holding costs. The gray market may allow the reseller to improve operational efficiency while enjoying the batch quantity discounts. Second, we examine the impact of gray market diversion on the resale price and sales in the authorized channel. Interestingly, the effect depends on the reseller’s batch inventory holding costs. When the batch inventory holding cost is sufficiently high such that the reseller would not order in batches without the gray market, diversion allows the reseller to enjoy the quantity discount and in turn reduce the resale price. As a result, the presence of gray markets expands sales in the authorized channel. However, when the batch inventory holding cost is low enough that the reseller would order in batches even without the gray market, the diversion reduces the reseller’s peak cycle inventory and expedites the ordering cycles. As the reseller faces reduced pressure to push out the inventory, the resale price increases and sales decrease in the authorized channel. Third, we study the effect of gray market diversion on the performance of the distribution channel. When gray market diversion enables the reseller to take advantage of the quantity discount, the supplier enjoys increased operational efficiency and total channel performance improves. Due to the Stackelberg structure of our model, the supplier captures all or most of the efficiency gains.

This chapter contributes to both the marketing and operations management (OM) literature. The marketing literature typically neglects inventory costs and limits attention to single-period models (Howell et al., 1986; Wilcox et al., 1987; Banerji, 1990). These restrictions on the analysis do not permit operational characteristics, in particular inventory holding costs, to influence the reseller’s strategies or the diversion to a gray market. Though Lal and Staelin (1984) indeed consider the effect of the reseller’s holding cost on the supplier’s quantity discount design in an economic-order-quantity (EOQ) setting, the authors assume that all products are sold to the authorized channel at an exogenous resale price. In this chapter we incorporate the reseller’s operational decisions and derive the resale price as a function of the operating cost in the presence of a gray market. As a result, the supplier can
take a reseller’s operating environment into account when making the channel management decisions.

Closely related, are salvage channels which are directly controlled by the supplier. The outlet mall is the primary example. These channels generally serve to liquidate supplier overstock resulting from demand uncertainty and also serve a secondary purpose providing a market for high price sensitivity customers (see Coughlan and Soberman, 2005). From the perspective of the supplier these serve a very different purpose since they are vertically integrated. A Nike factory store is rather different than a Nike shoe sold over eBay since the differentiation (from inconvenient location and lack of service) and prices are directly under Nike’s control rather than indirectly controlled by levers such as discounting. Understanding these levers when the reseller network is not vertically integrated is the goal of this chapter.

The prior paragraph begs the question of why a supplier would not either go directly to the gray market reseller or integrate their own salvage channel. First and foremost ‘fairness’ should not be neglected selling at highly discounted price points to gray market resellers is often not feasible and may also be illegal. KPMG LLP (2008) cites such motivation as a primary reason for gray market abatement. Implementing a vertically integrated salvage channel which would accomplish a similar goal is a significant undertaking and also competes directly with the reseller network. Finally, such sales may not accomplish the economies of scale if those are associated with logistical issues such as transportation.

The OM literature tends to emphasize the algorithmic issues determining optimal lot sizes (reseller ordering policy) in response to offered discounts due to the lack of tractability of the general multi-period problem. An overview of the area is covered by a pair of surveys from Benton and Park (1996) and Munson and Rosenblatt (1998). Despite an impressive breadth of work including many extensions of the lot-sizing problem under all-unit discounts, gray markets and, more generally, the ability to salvage surplus inventory has been discussed in only a couple of instances. Sethi (1984) and Arcelus and Rowcroft (1992) have examined the optimal lot-sizing problem with an all-unit discount and a fixed value for salvaged inventory. They both develop algorithms to numerically solve the reseller’s lot-sizing problem. This
chapter contributes to this literature by developing an explicit solution to a representative
dynamic lot-sizing problem where the salvage channel is replaced by a gray market which
may cannibalize some portion of authorized channel demand. The stylized elements in our
model include no fixed order costs and deterministic demand. Without fixed order costs, the
reseller is able to order as demand arrives and eliminate holding costs. Adding a fixed order
cost would replace this strategy with a less profitable EOQ style policy and lead resellers to be
more likely to use batch strategies, possibly including gray market diversion. Deterministic
demand allows for better elucidation of the interaction of the gray market with the suppliers
discount policy. More importantly, the closed-form solution allows optimal analysis of the
reseller’s pricing and inventory decisions, and the resulting profit implications to the supplier
and welfare consequences to the consumers.

We organize the rest of the chapter as follows. In §2.2 we describe the main model, fol-
lowed by a preliminary analysis to demonstrate the quantity discount-induced gray market.
In §2.3 we follow the conventional operational approach and assume an exogenously deter-
mined resale price. In this section we focus on the effect of a gray market on the reseller’s
operational decisions and the subsequent effect on the supplier’s profit. In §2.4 we incorpo-
rate the reseller’s pricing decision and investigate the interaction between operational and
marketing decisions in managing a channel with a gray market. We then extend the model in
§2.5 to allow for an endogenous gray market wholesale price which responds to the reseller’s
supply. Finally, we conclude with a summary of results and managerial implications in §2.6.
All proofs can be found in the Online Appendix.

\section{Model and Preliminary Analysis}

We consider a market where a monopoly supplier sells its products to end consumers through
a single reseller. The reseller can sell the goods through an authorized channel and a gray
market. We consider a Stackelberg game: the supplier first sets an all-unit discount schedule.
Given the discount policy and gray market condition, the reseller responds with decisions
on ordering, inventory holding, gray market diversion, and possibly resale pricing based on deterministic demand and a constant lead time (see Lal and Staelin 1984 for a similar setting). For simplicity, we assume there is no fixed order costs because the all-unit discount has already incorporated an incentive for batch ordering to enjoy economies of scale. We will verify later that with diversion, EOQ-type cyclic policies can be the optimal inventory policies even without fixed order costs. The obtained insights from the main model on gray market diversion remain when there is a positive fixed order cost.

Following the convention in the OM literature, we represent the supplier’s lot size-based all-unit quantity discount by the reseller’s order cost function, denoted by $C(q)$ where $q$ is the order size, as follows:

$$C(q) = \begin{cases} 
  w_o q & \text{if } 0 \leq q < \eta, \\
  w_\eta q & \text{if } q \geq \eta,
\end{cases}$$

where $w_o$ is the list price before the discount, $w_\eta$ is the discounted price and $\eta$ is the threshold order quantity defining the change in the unit cost. For simplicity, we assume there is only one price breakpoint as it is sufficient to demonstrate the incentive for overbuying and gray-market diversion. If $w_o = w_\eta$, the discount schedule reduces to the trivial case of no quantity discount. We assume the list price $w_o$ is determined before the supplier optimizes the quantity discount by setting $w_\eta$. The list price may be determined either exogenously\(^1\) or by an optimization on top of the Stackelberg game which is discussed in greater depth in Section 2.4.2. As we have mentioned, all-unit discounts are the most popular type of quantity discounts and provide a unique incentive to order up to the threshold (Benton and Park, 1996). In practice, all-unit discounts shift the batch-breaking decision from the supplier to the reseller, who shares the cost savings through the quantity discount.\(^2\) These discounts are most commonly motivated by operational advantages (Munson and Rosenblatt, 1998) and are typically associated with an exogenous batch size. For instance, this batch size may

---

\(^1\)Fixing the regular wholesale price is not uncommon in studies of quantity discount practices, e.g., Lal and Staelin (1984).

\(^2\)In addition, this interpretation is consistent with the Robinson-Patman act which forbids discriminatory pricing in the US through a quantity discount unless justified by underlying costs (Coughlan et al., 2001).
correspond to a pallet, production lot size or the capacity of a truck. In this vein, we assume that the batch size \( \eta \) is exogenously determined.

We assume there exists economies of scale in production and distribution when the reseller orders in batches. Specifically, we let \( c_o \) represent the per unit supply cost when the reseller orders one unit of the product each time, and \( c_\eta (< c_o) \) be the per unit supply cost when the reseller orders in batches of \( \eta \). As a result, the discount threshold is assumed to be fixed.

After the supplier announces the wholesale pricing policy, the reseller makes lot-sizing and, possibly, resale pricing decision \( (p) \). The reseller in our model may be viewed as either a retailer selling directly to consumers or an intermediate distributor selling to authorized retailers. At any time \( t \), the reseller's lot-sizing decisions include ordering \( q(t) \) units from the supplier at a cost of \( C(q(t)) \) and diverting \( g(t) \) units into the gray market. We assume away demand uncertainty and allow replenishment to be instantaneous which is equivalent to assuming a deterministic lead time. We denote the reseller's inventory level at time \( t \) by \( I(t) \), which is the sum of all orders minus all sales through the authorized channel and the gray market up to time \( t \). For each unit of goods in inventory we let the holding cost be \( h \) per unit of goods per unit of time.

The reseller can sell through the authorized channel and gray market. We focus on all-unit discount induced gray market diversion and assume that demand is deterministic in both markets. Specifically, in the authorized channel, the market demand (or order) arrives continuously with a deterministic rate determined by a modified iso-elastic demand structure that takes into account the cannibalization between the authorized channel and gray market:

\[
\lambda(p, p_s) = m/(p - \gamma p_s)^\alpha.
\]

The parameter \( m \) denotes the size of the market, \( \gamma \) parameterizes the sensitivity of the authorized channel's demand to the gray market resale price and \(-\alpha\) measures the demand elasticity to the adjusted market price difference. The higher the value of \( \gamma \), the more sensitive the authorized channel's demand to changes in the gray market resale price. In addition to measuring demand sensitivity to changes in \( p_s \), \( \gamma \) has a secondary and counterintuitive
effect that an increase in this parameter increases demand. Thus, to study changes in
cannibalization comparative statics should be taken on $p_s$ rather than on $\gamma$. In Section
2.4, an unlimited profit opportunity might present itself when $p - \gamma p_s$. This opportunity is
eliminated by ensuring that $w_\eta > s > \gamma p_s$ which is not too restrictive when the markets are
relatively differentiated. The gray market resale price is treated as exogenously determined.
This is likely the case when quantity discount driven diversion is one of many contributors
to the total gray market. In §2.5, we will examine the robustness of our results by extending
the model to scenarios where the gray market prices are negatively affected by the focal
reseller’s diversion volume. In the numerical examples shown in sections 2.3 and 2.4 it is
assumed that the gray market uses a fixed markup $u$ to price to their consumers such that
$p_s = (1 + u)s$.

The degree of cannibalization depends on the structure of the gray market, for instance,
whether these goods are sold in the same geography as the reseller. If the quantity discount
is providing impetus for parallel importation into a separate marketplace $\gamma$ will be small
due to minimal cannibalization. Also, $\gamma$ will be small when goods are well differentiated
from their gray market counterparts. For instance, ancillary services such as warranties are
lacking in the gray market channel. In the markets where counterfeiting is prevalent, the
consumer is concerned that the gray market product is not genuine.

Profits for both reseller and supplier will be considered over an infinite horizon. In total,
the assumptions noted above lead to an EOQ style policy where inventory is ordered up to
a threshold and deterministic demand reduces holdings over the course of a constant length
cycle. This type of policy remains important for products with low demand seasonality and
moderate life cycles. These are characteristics of the prescription drug and audio electronics
markets cited in the introduction. The IT products cited also fall under this umbrella. The
subjects of the KPMG study are mainly commoditized IT products such as hard drives
and network routers. Such products are sold over very long timelines through periodic
incremental specification improvements (e.g. a 1GB hard drive is replaced by a 1.5GB hard
drive at the same price point).
Furthermore, the gray market wholesale price is assumed to be below the batch supply cost, i.e. $s < c_\eta$, to eliminate the suppliers indirect arbitrage opportunity. This is a common assumption in the OM literature, for example, similar assumptions are made in the newsvendor problem. The conventional OM literature often considers an exogenous resale price and focuses on cost-minimizing lot-sizing decisions. We will follow this approach in §2.3 and assume that the resale price $p$ is fixed and that the reseller adjusts only his ordering and diversion behavior to minimize costs. We then incorporate the reseller’s pricing decisions in §2.4 but assume $\alpha = 2$ to obtain closed form solutions. We will check the robustness of obtained insights for general values of $\alpha$ by numerical experiments. We summarize all notation in Table 2.1.

Finally, we summarize the sequence of events in Figure 2.1: First, the supplier sets the discount price $w_\eta$. After observing the all-unit discount, the reseller chooses the order size, inventory level and diversion volume. In the endogenous resale price model (§2.4), the reseller also selects the resale price. We assume that the supplier has perfect knowledge of the reseller’s holding costs as well as the demand structure of the authorized and gray markets. Therefore, the supplier can anticipate the reseller’s order, diversion and pricing decisions, and manage the authorized channel and possible diversion through the offering of the wholesale price schedule.

### 2.2.1 Preliminary Analysis: Quantity-Discount Induced Gray Market Diversion

We first analyze a simple one-period model to investigate the possible incentive for overbuying and diverting to the gray market under the all-unit quantity discount. Such a static model can be considered as a snapshot of our dynamic problem. Suppose the reseller orders $q$ units to fulfill demand $d$ in the authorized channel and diverts the remaining $g = q - d$ units to the gray market. Expressed in terms of $d$ and $q$ the reseller’s total effective cost associated with such a policy is $c(q, d) = C(q) - s(q - d)$. 
Table 2.1: Description of Principal Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta$</td>
<td>Order quantity discount threshold</td>
</tr>
<tr>
<td>$w_o$</td>
<td>List wholesale price per unit</td>
</tr>
<tr>
<td>$w_\eta$</td>
<td>Discounted wholesale price per unit</td>
</tr>
<tr>
<td>$c_o$</td>
<td>Supply cost per unit without economies of scale</td>
</tr>
<tr>
<td>$c_\eta$</td>
<td>Supply cost per unit with economies of scale</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Demand rate of the authorized channel</td>
</tr>
<tr>
<td>$p$</td>
<td>Resale price</td>
</tr>
<tr>
<td>$s$</td>
<td>Gray market wholesale price</td>
</tr>
<tr>
<td>$p_s$</td>
<td>Gray market resale price</td>
</tr>
<tr>
<td>$m$</td>
<td>Market size of the authorized channel</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Price elasticity</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Demand sensitivity of the authorized channel to the gray market resale price</td>
</tr>
<tr>
<td>$h$</td>
<td>Reseller’s inventory holding cost per unit per unit time</td>
</tr>
<tr>
<td>$I$</td>
<td>Reseller’s cycle inventory level</td>
</tr>
<tr>
<td>$I^o$</td>
<td>Reseller’s cycle inventory level in the diversion strategy</td>
</tr>
<tr>
<td>$H$</td>
<td>Reseller’s average inventory holding cost per unit per cycle when $I = \eta$ and $\lambda = m$</td>
</tr>
<tr>
<td>$G$</td>
<td>Gray market diversion per cycle</td>
</tr>
<tr>
<td>$g$</td>
<td>Gray market diversion per unit of time</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>Supplier’s profits</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Reseller’s profits</td>
</tr>
</tbody>
</table>
Supplier sets wholesale price schedule: selects a regular wholesale price $w_0$ and the all-unit quantity discount $w_\eta$ with batch size requirement $\eta$.

Reseller reacts: decides on the inventory policy composed of strategies for ordering and gray market diversion and the resale price.

End consumers make purchase decisions dependent on the resale and gray market price.

Figure 2.1: Model Structure
If the resale price is exogenous, then maximizing the reseller’s profit becomes equivalent to minimizing the reseller’s cost. We calculate the optimal order quantity as a function of replenishment need \(d\), \(q^*(d) := \max\{\arg\min_{q \geq d} c(q, d)\}\). We let \(c^*(d) := c(q^*(d), d)\) and \(g^*(d) := q^*(d) - d\) denote the reseller’s optimal order cost and the optimal size of diversion to the gray market, respectively. We summarize the reseller’s optimal order and diversion strategy in Proposition 1.

**Proposition 1. (Optimal One-Shot Order and Diversion Strategy).** Under the all-unit discount \(C(q)\), the optimal ordering and diversion strategy of the reseller with a one-shot demand size \(d \geq 0\) is

\[
q^*(d) = \begin{cases} 
\eta & \text{if } \hat{q} \leq d < \eta, \\
d & \text{otherwise}, 
\end{cases}
\]

where \(\hat{q} = \frac{(w_\eta - s)\eta - s}{w_\eta - s} \leq \eta\) is the threshold demand size above which the reseller orders up to the discount break point \(\eta\). The optimal effective order cost is a continuous function

\[
c^*(d) = \begin{cases} 
w_\eta d & \text{if } 0 \leq d < \hat{q}, \\
(w_\eta - s)\eta + sd & \text{if } \hat{q} \leq d < \eta, \\
w_\eta d & \text{otherwise.} 
\end{cases}
\]

The diversion to the gray market is

\[
G^*(d) = \begin{cases} 
\eta - d & \text{if } \hat{q} \leq d < \eta, \\
0 & \text{otherwise.} 
\end{cases}
\]

**Proof.** Proof of Proposition 1. To satisfy the demand \(d \geq 0\), the order size is \(q \geq d\) and the diversion size is \(q - d \geq 0\). If \(d \geq \eta\), then \(q \geq d \geq \eta\); the order cost \(c(q, d) = w_\eta q - s(q - d)\) that is linear in \(q\) with a slope \(w_\eta - s > 0\), thus it is minimized at \(q^*(d) = d\). If \(d < \eta\), the order cost is

\[
c(q, d) = \begin{cases} 
w_\eta q - s(q - d) & \text{if } d \leq q < \eta, \\
w_\eta q - s(q - d) & \text{if } q \geq \eta. 
\end{cases}
\]
On the above two regions the cost-minimizing solutions are respectively \( q = d \) and \( q = \eta \).

To find the optimum \( c^*(d) \) when \( d < \eta \), it suffices to compare the cost at the two solutions, i.e., \( c^*(d) = \min\{w_\eta d, (w_\eta - s)\eta + sd\} \): if \( 0 \leq d \leq \hat{q} \), \( c^*(d) = w_\eta d \) and if \( \hat{q} < d < \eta \), \( c^*(d) = (w_\eta - s)\eta + sd \).

\[ \square \]

**Remark 1.** Incremental discounts and two-part tariffs do not generate incentive for over-buying and diversion to the gray market.

**Proof.** Proof of Remark 1. First, consider an incremental discount:

\[
C(q) = \begin{cases} 
  w_\eta q & \text{if } 0 \leq q < \eta, \\
  w_\eta(q - \eta) + w_\eta \eta & \text{if } q \geq \eta,
\end{cases}
\]

To satisfy the demand \( d \geq 0 \), the order size is \( q \geq d \) and the diversion size is \( q - d \geq 0 \). If \( d \geq \eta \), then \( q \geq d \geq \eta \); the order cost \( c(q, d) = w_\eta(q - \eta) + w_\eta \eta - s(q - d) \) that is linear in \( q \) with a slope \( w_\eta - s > 0 \), thus it is minimized at \( q^*(d) = d \). If \( d < \eta \), the order cost is

\[
c(q, d) = \begin{cases} 
  w_\eta q - s(q - d) & \text{if } d \leq q < \eta, \\
  w_\eta(q - \eta) + w_\eta \eta - s(q - d) & \text{if } q \geq \eta.
\end{cases}
\]
On the above two regions the cost-minimizing solutions are respectively \( q = d \) and \( q = \eta \). To find the optimum \( c^*(d) \) when \( d < \eta \), it suffices to compare the cost at the two solutions, i.e.,
\[
c^*(d) = \min\{w_o d, w_o \eta - s(\eta - d)\} = w_o d.
\]
For both cases of \( d \geq \eta \) and \( d < \eta \), the optimal order size is \( q^*(d) = d \).

Second, consider a two-part tariff: \( C(q) = F + w_o q, \ q > 0 \). To satisfy the demand \( d \geq 0 \), the order size is \( q \geq d \) and the diversion size is \( q - d \geq 0 \). The order cost is
\[
c(q, d) = F + w_o q - s(q - d) \text{ for } q \geq d \text{ and is linear in } q \text{ with a slope } w_o - s > 0.
\]
Thus it is minimized at \( q^*(d) = d \).

The diversion to the gray market occurs only in the middle range of demand defined by \( \hat{q} \leq d < \eta \). In this range, the benefit from receiving the quantity discount outweighs the loss from diverting the excess purchases to the gray market. If \( d = \hat{q} \), the reseller is indifferent between 1) making an order \( \hat{q} \) at the list price \( w_o \) and 2) ordering up to \( \eta \) at the discounted unit price \( w_\eta \) and then diverting \( \eta - \hat{q} \) units to the gray market; for consistency, we designate the reseller’s behavior when \( d = \hat{q} \) to be the latter case of ordering and diversion.

We illustrate this optimal order cost \( c^*(d) \) in Figure 2.2. The optimal order cost function is a continuous piecewise linear function with three segments with respective slopes of \( w_o, s \), and \( w_\eta \). The reseller contributes to the gray market only over the middle segment where \( \hat{q} \leq d < \eta \). In this case, the reseller orders \( \eta \) units regardless of the size of demand; an incremental demand means a unit reduction of the reseller’s supply to the gray market, and hence an increase in the reseller’s total cost by \( s \).

The above analysis has clearly demonstrated the cause and consequence of a quantity discount-induced gray market at the reseller’s level. Next we analyze the full model where both the reseller and supplier consider their decisions over an infinite horizon.
2.3 Model Analysis: Exogenous Resale Price

In this section we study the case where the resale price $p$ is exogenous. As a result, the reseller faces an exogenous and deterministic demand $\lambda = m/(p - \gamma p_s)^\alpha$ per unit of time, where the gray market wholesale and resale prices $(s, p_s)$ are exogenously determined. We will solve the Stackelberg game backward by first analyzing the reseller’s optimal inventory decisions in response to the supplier’s quantity discount schedule, and then examining the supplier’s optimal quantity discount.

2.3.1 Reseller’s Inventory Policy

The reseller’s objective is to maximize long run average profits. With an exogenous resale price, the reseller’s profit-maximization problem can be solved by minimizing the total of the reseller’s order cost and inventory holding cost. In the following lemma, we prove that the optimal policy is of stationary type among all dynamic policies. The stationary inventory policy includes a target inventory level $S$, an order size $q$, a diversion amount $g$ to the gray market, and the timing of orders and diversion. Our analysis shows that it is optimal for the reseller to follow the zero-inventory policy described in Lemma 1.

Lemma 1. (Reseller’s Zero-Inventory Policy). The reseller’s optimal inventory policy consists of a target cycle inventory level $I$. The reseller orders $q^*(I)$ and diverts $G^*(I)$ to the gray market at precisely the times when the inventory drops to zero.

Proof. Proof of Lemma 1. At any time $t$ the reseller can change his inventory position from the current position $I(t)$ to a new position $I(t) + \Delta I(t)$ by a combination of order $q(t)$ from the supplier and gray market diversion $g(t)$. Fix any time $t$. We first argue that since replenishment is instantaneous, it is suboptimal for $\Delta I(t) = i > 0$ if $I(t) > 0$. The action of changing the current inventory position $I(t)$ to $I(t) + i$ could be profitably delayed to the time $t_0 := \inf\{x > t : I(x) = 0\}$. Letting $\Delta I(t) = 0$ and $\Delta I(t_0) = \Delta I(t_0) + i$ has an improvement $h(t_0 - t)i > 0$ in the holding costs up to time $t_0$. Recursively applying this
process results in an improved set of orders where $\Delta I(t) > 0$ only when $I(t) = 0$. We now argue that it is suboptimal for disposal of goods $\Delta I(t) = j < 0$ if $I(t) > 0$ since this action could have been profitably performed at an earlier time $t_{-1} := \sup\{x < t : I(x) = 0\}$, which has a holding cost improvement $h(t - t_{-1}) \mid j \mid > 0$. Therefore, in an efficient inventory policy, any ordering or gray market diversion occurs only at times when $I(t) = 0$. The set of times when $I(t) = 0$ represents a set of renewal points. Since the demand rate is stationary, the optimal action is identical at each of these times. To complete the proof let $I$ be equal to the optimal inventory adjustment when $I(t) = 0$ and then $q^*(I)$ and $g^*(I)$ correspond to the optimal order and gray market diversion quantities respectively.

The zero-inventory property in Lemma 1 is typically associated with EOQ models with constant demand and fixed order cost (see Zipkin 2000; Simchi-Levi et al. 2005). Under such an inventory policy, a reseller who diverts to the gray market will repeatedly place orders of volume $\eta$ whenever the inventory level hits zero, immediately divert a quantity of $G$ to the gray market, and serves the authorized channel from inventory afterwards until the inventory level reaches zero again. In the optimal inventory policy, the inventory level reaches a peak of $I = \eta - G$ at the beginning of each order cycle. Since the demand arrives at a constant rate $\lambda$, the length of each cycle is $I/\lambda = (\eta - G)(p - \gamma p_s)^\alpha / m$. For simplicity of notation, we define the average unit holding cost per cycle for a full batch order $I = \eta$ without diversion when the demand rate $\lambda = m$ as follows:

$$H := \frac{h\eta}{2m},$$

where $h$ is the unit holding cost for each unit of time. As an aggregate measurement, the holding cost $H$ contains information not only about $h$ but also about the cycle length, and is a more relevant measure of inventory costs in this chapter.

Given the optimal zero-inventory policy characterized by Lemma 1, the total costs for each order cycle of length $I/\lambda$ consist of order cost $c^*(I)$ given by Proposition 1 and holding cost $hI^2/(2\lambda)$. We can then calculate the long-run average cost per unit time as $G(I) =$
Chapter 2. All-Unit Discounts & Gray Markets

\[ c^*(I)\lambda/I + hI/2. \] The reseller’s optimal inventory policy minimizes \( G(I) \)

**Proposition 2. (Reseller’s Optimal Inventory Policy).** The optimal inventory policy of the reseller is a zero-inventory policy with the cycle inventory defined as follows:

If \( s < \omega < (\omega_0 + s)/2, \)

\[
I^* = \begin{cases} 
\eta & \text{if } H \leq \frac{\omega_0 - s}{(p - \gamma p_s)^{\alpha}}, \\
\frac{\omega_0 - s}{(p - \gamma p_s)^{\alpha}} & \text{if } \frac{\omega_0 - s}{(p - \gamma p_s)^{\alpha}} < H \leq \frac{(\omega_0 - s)^2}{4(\omega_0 - s)(p - \gamma p_s)^{\alpha}}, \\
0 & \text{otherwise};
\end{cases}
\]

If \( (\omega_0 + s)/2 \leq \omega \leq \omega_0, \)

\[
I^* = \begin{cases} 
\eta & \text{if } H \leq \frac{\omega_0 - \omega}{(p - \gamma p_s)^{\alpha}}, \\
0 & \text{otherwise}.
\end{cases}
\]

**Proof.** Proof of Proposition 2. To solve for the optimal inventory policy, the reseller selects the cycle inventory level \( I \) that minimizes the total costs. Given the optimal zero-inventory policy characterized by Lemma 1, the total costs for each order cycle of length \( I/\lambda \) consist of order cost \( c^*(I) \) given by Proposition 1 and holding cost \( hI^2/(2\lambda) \). By substituting the reseller’s optimal cost function given by Proposition 1 (where \( d = I \)) into \( G(I) = c^*(I)\lambda/I + hI/2 \) we obtain the expression:

\[
G(I) = \begin{cases} 
\frac{\omega_0 m}{(p - \gamma p_s)^{\alpha}} + hI/2 & \text{if } 0 \leq I < \hat{q}, \\
(\omega_0 - s)\eta m/[I(p - \gamma p_s)^{\alpha}] + sm/(p - \gamma p_s)^{\alpha} + hI/2 & \text{if } \hat{q} \leq I < \eta, \\
\frac{\omega_0 m}{(p - \gamma p_s)^{\alpha}} + hI/2 & \text{otherwise}.
\end{cases}
\]

Recall that in the first and third cases, the reseller orders up to the desired cycle inventory level \( I \) and sells the entire order through the authorized channel over time; no goods are diverted to the gray market in these two cases. However, in the second case, the reseller orders up to the quantity of \( \eta \) to enjoy the quantity discount and sells the excess amount \( \eta - I \) to the gray market. Within this range \( (\hat{q} \leq I < \eta) \), the reseller will choose a locally optimal cycle inventory level \( I^0 \) that minimizes the cost \( G(I) \), where \( I^0 := \sqrt{\frac{2(\omega - s)\eta m}{H(p - \gamma p_s)^{\alpha}}} = \eta \sqrt{\frac{w_0 - s}{H(p - \gamma p_s)^{\alpha}}}. \)

The reseller selects the optimal cycle inventory and gray market diversion by comparing the minimum cost \( G(I) \) in each of the three regions.
Since the demand and resale price is fixed for the reseller, the reseller’s revenue is fixed. The reseller is aiming at minimizing cost $G(I)$. To find the minimum of $G(I)$ we compare the optimal solutions for each region of $I \in [0, \hat{q}]$, $I \in (\hat{q}, \eta)$ and $I \in [\eta, \infty)$. Over the first and third regions, $G(I)$ is a linearly increasing function and is minimized at $I = 0$ and $I = \eta$ respectively. Over the second region $I \in (\hat{q}, \eta)$, $G(I)$ is convex and minimized at an interior point $I^o$ if it is indeed in $(\hat{q}, \eta)$. Otherwise, $G(I)$ is minimized at one of the boundary points $\hat{q}$ or $\eta$.

The necessary and sufficient condition for $I^* = I^o$ is $I^o \in (\hat{q}, \eta)$, $G(I^o) < G(0)$ and $G(I^o) < G(\eta)$. The feasibility condition $I^o > \hat{q}$ holds if and only if $H(p - \gamma p_s)^\alpha < (w_o - s)^2/(w_o - s)$; The other feasibility condition $I^o < \eta$ holds if and only if $H(p - \gamma p_s)^\alpha > (w_o - s)$. The optimality condition $G(I^o) \leq G(\eta)$ always holds since $G(I)$ is continuous at $I = \eta$. Finally, the other optimality condition $G(I^o) < G(0)$ holds if and only if $H(p - \gamma p_s)^\alpha < (w_o - s)^2/[4(w_o - s)]$. Taking the intersection of regions defined by the feasibility and optimality conditions yields that $I^* = I^o$ if and only if $(w_o - s) < H(p - \gamma p_s)^\alpha < (w_o - s)^2/[4(w_o - s)]$. Such holding costs exist only if $w_o \in (s, (w_o + s)/2]$.

$I^* = \eta$ if and only if $I^* \neq I^o$ and $G(\eta) \leq G(0)$. $G(\eta) \leq G(0)$ holds if and only if $H(p - \gamma p_s)^\alpha \leq w_o - w_o$ with equality holding at $H = w_o - w_o$. Intersecting $H(p - \gamma p_s)^\alpha \leq w_o - w_o$ with the region where $I^* \neq I^o$, i.e., $H(p - \gamma p_s)^\alpha \notin (w_o - s, (w_o - s)^2/[4(w_o - s)])$, results in $H(p - \gamma p_s)^\alpha \leq \min \{w_o - s, w_o - w_o\}$. Note that $w_o - s < w_o - w_o$ if and only if $w_o < (w_o + s)/2$. The necessary and sufficient condition for $I^* = \eta$ follows immediately.

The remaining possible holding cost regions are $H(p - \gamma p_s)^\alpha > (w_o - s)^2/[4(w_o - s)]$ if $w_o < (w_o + s)/2$, $H(p - \gamma p_s)^\alpha > w_o - w_o$ otherwise, which correspond to $I^* = 0$.

We define formally three inventory policies obtained in Proposition 2 and later may refer to them in their short names.

**Definition 1** (Order-as-You-Go Strategy). The reseller does not carry any inventory ($I = 0$ in short).
**Definition 2** (Batch Strategy). *At the beginning of each cycle the reseller orders \( \eta \) and does not divert any quantity to the gray market (\( I = \eta \) in short).

**Definition 3** (Diversion Strategy). *At the beginning of each cycle the reseller orders \( \eta \), diverts \( \eta - I^o \) into the gray market immediately and does not divert any order to the gray market within the cycle (\( I = I^o \) in short).*

Proposition 2 shows that if the holding cost is sufficiently high, the reseller follows the order-as-you-go strategy; and if the holding cost is sufficiently low, the reseller follows the batch strategy. As the reseller increases the order size to \( \eta \) to enjoy the discount, it may incur an additional unit holding cost or a negative profit margin for those units diverted to the gray market. Only when the quantity discount is attractive enough and the holding cost is in an intermediate range, does the reseller adopt the diversion strategy. The first necessary condition for an optimal diversion strategy is \( w_\eta - s < w_o - w_\eta \), namely, the benefit from the quantity discount must outweigh the margin loss on diverting a unit. This is essential to ensure a window of profit opportunity for the reseller. Moreover, in order for the reseller to be better off diverting goods to the gray market rather than keeping all products in inventory, the immediate diversion loss \( w_\eta - s \) needs to be smaller than the average unit holding cost \( H(p - \gamma p_s)^\alpha \) the reseller would have to incur in a cycle without diversion. Finally, in order for the reseller to be better off adopting the diversion strategy than the order-as-you-go strategy, costs associated with the optimal diversion strategy should be lower than those under the order-as-you-go strategy. This condition leads to an upper bound on the holding cost in order for the diversion strategy to be optimal.

Under the understanding that the gray market wholesale and resale prices are positively correlated, a higher gray market wholesale price will have two implications. First, as the per unit diversion loss becomes smaller, a diversion effect increases the reseller’s incentive to resort to the gray market. Second, a cannibalization effect makes the gray market purchase less attractive and increases demand in the authorized channel which reduces the reseller’s incentive to divert at a loss. These two effects clearly work in opposition. Depending on the
Figure 2.3: Reseller’s Best Responses with Exogenous Resale Price
The solid (resp. dashed) line indicates policy boundaries, $s = 2$ (resp. $s = 3$). $m = 10, \eta = 20, w_o = 10, \gamma = 0.5, p = 12$, gray market resale markup= 40%.
relative magnitude of each effect, an increase in the gray market prices may either reinforce or reduce the reseller’s incentive for gray market diversion. If the cannibalization parameter $\gamma$ is larger, the effect of an equivalent change in the gray market resale price will be magnified.

The diversion effect directly affects the reseller’s decision making, however, the cannibalization effect acts exogenously through the authorized market demand. This is relaxed in Section 2.4 where the gray market price is endogenously related to the rate of diversion. The effect is nonetheless of interest to the supplier who often has other levers to affect cannibalization parameters such as increasing or decreasing the degree of differentiation between authorized and gray market goods.

We illustrate the key results of Proposition 2 with Figure 2.3. The figures show the parameter spaces under which each of these three inventory policies are optimal. We assume that the gray market resale price is marked up by 40% so that $p_s = 1.4s$. Note that, while the scale for $w_\eta$ is the same in three figures, the scale for $H$ is very different, with much smaller units in Figure 2.3(b) and 2.3(c) owing to reduced demand from increased price elasticity. Overall, as the value of $\alpha$ increases from 0 to 2, the parametric spaces for both batch and diversion strategies become much smaller. This is consistent with Proposition 2 where an increase in $\alpha$ increases the lower bound on holding costs $H$ associated with diversion and batch strategies.

In order to illustrate the diversion effect, in Figure 2.3(a) we eliminate cannibalization by setting $\alpha = 0$. When $s$ increases from 2 to 3, the boundary lines for the diversion strategy versus batch and order-as-you-go strategies both shift towards the right as the diversion strategy becomes more attractive. The boundary between diversion and batch strategies shifts in parallel from $H = w_\eta - s_1$ to $w_\eta - s_2$. In this case, the gray market allows the reseller to reduce cycle inventory in addition to enjoying the discounted price. Such a benefit is independent of the holding cost. On other hand, the boundary between diversion and order-as-you-go strategies, present at higher values of $H$, shifts disproportionally further to the right as $H$ increases. In this case, the gray market offers the reseller a chance to order additional units to enjoy the discount which benefits the reseller more at higher holding
costs.

The cannibalization effect is illustrated in Figure 2.3(b) where $\alpha = 2$. Clearly the diversion effect still exists: when $s$ is increased from 2 to 3, the boundary lines shift rightwards and a diversion strategy is more likely to be optimal. However, positive values for parameters $\alpha$ and $\gamma$ lead to cannibalization of the authorized channel demand. As a result, in order to be optimal, the batch strategy requires a much smaller $H$. The interaction between the diversion and cannibalization effects on the reseller’s strategies can also be observed in Figure 2.3(b). In contrast to Figure 2.3(a), rather than shifting in parallel, the slope of the boundary line between diversion and batch strategies increases when $s$ is increased from 2 to 3. While a larger $s$ makes the diversion strategy more efficient, the effect is smaller with relatively larger values of $w_n$ due to the gray market resale price also increasing. The increased $p_s$ reduces the cannibalization effect, increasing market demand in the authorized channel and making the batch strategy more profitable. Figure 2.3(c) illustrates that the reduction in the cannibalization effect is stronger at a larger value of $\alpha$. The cannibalization effect eclipses the diversion effect when $w_n$ is close to 6. The boundary between diversion and order-as-you-go strategies is affected differently. As $s$ is increased, the diversion and reduced cannibalization effects are complementary make the diversion strategy even more attractive relative to the order-as-you-go policy and push this boundary further upwards.

2.3.2 Supplier’s Discount Policy

When deciding on the discounted price, the supplier needs to anticipate the best response from the reseller. Based on Proposition 2, we can derive the supplier’s profit function and optimize its all-unit discount policy. We summarize the results as follows.

**Lemma 2. (Supplier’s Profit Function under Exogenous Resale Price).** Given that the reseller employs the optimal inventory policy in response to a discounted wholesale price $w_n$, the supplier receives the following profit per unit of time:
if \( H < (w_o - s)/[2(p - \gamma p_s)^\alpha] \),

\[
\Pi(w_\eta) = \begin{cases} 
  m(w_\eta - c_\eta) \sqrt{H / (w_\eta - s)(p - \gamma p_s)^\alpha} & \text{if } s < w_\eta \leq s + H(p - \gamma p_s)^\alpha, \\
  m(w_o - c_\eta) / (p - \gamma p_s)^\alpha & \text{if } s + H(p - \gamma p_s)^\alpha < w_\eta \leq w_o, \\
  m(w_o - c_\eta) & \text{if } w_o - H(p - \gamma p_s)^\alpha < w_\eta \leq w_o;
\end{cases}
\]

if \( H \geq (w_o - s)/[2(p - \gamma p_s)^\alpha] \),

\[
\Pi(w_\eta) = \begin{cases} 
  m(w_\eta - c_\eta) \sqrt{H / (w_\eta - s)(p - \gamma p_s)^\alpha} & \text{if } s < w_\eta \leq s + \frac{(w_o - s)^2}{4H(p - \gamma p_s)^\alpha}, \\
  m(w_o - c_\eta) / (p - \gamma p_s)^\alpha & \text{if } s + \frac{(w_o - s)^2}{4H(p - \gamma p_s)^\alpha} < w_\eta \leq w_o.
\end{cases}
\]

**Proof.** Proof of Lemma 2. The supplier’s profit depends on the inventory strategy the reseller optimally selects according to conditions shown in Proposition 2. The supplier’s profit per unit of time is equal to the rate the supplier supplies the reseller multiplied by the profit margin per unit. The margin is \( w_o - c_o \) per unit if \( I^* = 0 \) and \( w_\eta - c_\eta \) per unit otherwise.

If no goods are sold to the gray market (i.e. \( I = 0 \) or \( I = \eta \)), the supplier’s demand rate is equivalent to the reseller’s demand rate \( \lambda \). However, if \( I = I^o \) then the supplier’s demand rate is \( \eta \lambda / I^o = \lambda \sqrt{H(p - \gamma p_s)^\alpha / (w_\eta - s)} \). The supplier’s profit function is therefore

\[
\Pi(w_\eta) = \begin{cases} 
  \lambda(w_o - c_o) & \text{if } I^* = 0, \\
  \lambda(w_\eta - c_\eta) \sqrt{H(p - \gamma p_s)^\alpha / (w_\eta - s)} & \text{if } I^* = I^o, \\
  \lambda(w_\eta - c_\eta) & \text{if } I^* = \eta.
\end{cases}
\]

It remains to show that the conditions denoted in the proposition are sufficient to entail the appropriate reseller’s best response.

We begin by considering the conditions on the discount which imply \( I^* = \eta \). The conditions where \( I^* = \eta \) denoted in Proposition 2 are equivalent to

\[
w_\eta \in (s + H(p - \gamma p_s)^\alpha, (w_o + s) / 2] \cup [(w_o + s) / 2, w_o - H(p - \gamma p_s)^\alpha).
\]

Hence \( I^* = \eta \) if and only if

\[
w_\eta \in (s + H(p - \gamma p_s)^\alpha, w_o - H(p - \gamma p_s)^\alpha)
\]
and such a $w_{\eta}$ exists if and only if $w_{o} - s > 2H(p - \gamma p_{s})^{\alpha}$. Therefore, as stated in the proposition, $\Pi(w_{\eta}) = \lambda(w_{\eta} - c_{\eta})$ if and only if

$$w_{\eta} \in (s + H(p - \gamma p_{s})^{\alpha}, w_{o} - H(p - \gamma p_{s})^{\alpha})$$

which is non-empty only if $w_{o} - s > 2H(p - \gamma p_{s})^{\alpha}$.

Similarly the conditions from Proposition 2 which imply $I^{*} = I^{o}$ can be summarized as $w_{\eta} \in (s, s + H(p - \gamma p_{s})^{\alpha}) \cap (s, s + (w_{o} - s)^{2}/(4H(p - \gamma p_{s})^{\alpha})) \cap (s, (w_{o} + s)/2)$, which can be simplified by considering whether $w_{o} - s \leq h_{\eta}/\lambda$. If $w_{o} - s \leq 2H(p - \gamma p_{s})^{\alpha}$ then $I^{*} = I^{o} \iff w_{\eta} \in (s, s + (w_{o} - s)^{2}/(4H(p - \gamma p_{s})^{\alpha}))$, and if $w_{o} - s > h_{\eta}/\lambda$ then $I^{*} = I^{o} \iff w_{\eta} \in (s, s - H(p - \gamma p_{s})^{\alpha})$.

The remaining scenarios of $w_{\eta}$ are attributed to when the reseller selects $I^{*} = 0$. Instantiating the regions corresponding to each of the reseller’s inventory policy into $\Pi(w_{\eta})$ is sufficient to verify that Lemma 2 holds.

To derive the supplier’s optimal discounted price $w_{\eta}$, it suffices to compare the profits at the local maximum points. We summarize the results in Proposition 3.

**Proposition 3. (Supplier’s Optimal Discount with Exogenous Resale Price).**

*Given that the reseller employs the optimal inventory policy, the profit-maximizing supplier’s optimal discounted wholesale price is:*

if $c_{o} - c_{\eta} > (w_{o} - s)/2$,

$$w_{\eta}^{*} = \begin{cases} w_{o} - H(p - \gamma p_{s})^{\alpha} & \text{if } H \leq \frac{w_{o} - s}{2(p - \gamma p_{s})^{\alpha}}, \\ s + \frac{(w_{o} - s)^{2}}{4H(p - \gamma p_{s})^{\alpha}} & \text{if } \frac{w_{o} - s}{2(p - \gamma p_{s})^{\alpha}} < H \leq \frac{(2c_{o} - w_{o} - s)(w_{o} - s)}{4(c_{\eta} - s)(p - \gamma p_{s})^{\alpha}}, \\ w_{o} & \text{otherwise}; \end{cases}$$

if $c_{o} - c_{\eta} \leq (w_{o} - s)/2$,

$$w_{\eta}^{*} = \begin{cases} w_{o} - H(p - \gamma p_{s})^{\alpha} & \text{if } H \leq \frac{c_{o} - c_{\eta}}{(p - \gamma p_{s})^{\alpha}}, \\ w_{o} & \text{otherwise}. \end{cases}$$
Proof. Proof of Proposition 3. Given the profit function in Lemma 2, we solve the problem of optimizing \( \Pi(w_\eta) \) over \( w_\eta \in (s, w_o] \). For those discounted prices \( w_\eta \) which are close enough to \( w_o \) and elicit \( I^* = 0 \), the profit function \( \Pi(w_\eta) = \lambda(w_o - c_\eta) \) remains a constant. It is sufficient to set \( w_\eta = w_o \) to generate this profit. Consider \( H \geq (w_o - s)/[2(p - \gamma p_s)^\alpha] \). By Lemma 2, if \( w_\eta < s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \), then \( I^* > 0 \). Under the assumption that \( s < c_\eta \), \( \Pi(w_\eta) \) is increasing over \( w_\eta \in (s, s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha)] \), and is maximized at \( w_\eta = s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \) generating profit \( \lambda[(w_o - s)/2 - 2H(p - \gamma p_s)^\alpha(c_\eta - s)/(w_o - s)] \) for the supplier. Consider \( H < (w_o - s)/[2(p - \gamma p_s)^\alpha] \). If \( w_\eta \leq w_o - H(p - \gamma p_s)^\alpha \), then \( I^* > 0 \). Again since \( s < c_\eta \), it is easy to see that \( \Pi(w_\eta) \) is continuous at \( w_\eta = s + H(p - \gamma p_s)^\alpha \) and increasing over \( w_\eta \in (s, w_o - H(p - \gamma p_s)^\alpha] \), and is therefore maximized at \( w_\eta = w_o - H(p - \gamma p_s)^\alpha \) generating profit \( \lambda(w_o - H(p - \gamma p_s)^\alpha - c_\eta) \) for the supplier. We can now compare supplier’s profits given the reseller’s best response of \( I^* > 0 \) or \( I^* = 0 \) depending on whether \( (w_o - s)/2 > H(p - \gamma p_s)^\alpha \). When \( H(p - \gamma p_s)^\alpha \geq (w_o - s)/2 \), setting \( w_\eta = w_o \) generates greater profit for the supplier than \( w_\eta = s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \) if and only if \( H(p - \gamma p_s)^\alpha > (2c_o - w_o - s)(w_o - s)/[4(c_\eta - s)] \). When \( H(p - \gamma p_s)^\alpha < (w_o - s)/2 \), setting \( w_\eta = w_o \) generates greater profit for the supplier than \( w_\eta = w_o - H(p - \gamma p_s)^\alpha \) if and only if \( H(p - \gamma p_s)^\alpha > c_o - c_\eta \). 

Proposition 3 shows that the supplier’s decision to offer a quantity discount largely depends on the supplier’s benefit from economies of scale \( c_o - c_\eta \) and the reseller’s holding cost \( H \). When the benefit from economies of scale is sufficiently small and the holding cost is sufficiently large, the supplier does not offer a discount, i.e., \( w^*_\eta = w_o \). Inadequate economies of scale allow only a small window of profitable discounts, which is not enough to make the reseller hold inventory. On the other hand, a significant benefit from economies of scale may lead the supplier to provide a quantity discount. The size of the discount also depends on the inventory holding cost.

When the holding cost is low, the supplier offers a discount of \( H(p - \gamma p_s)^\alpha \) off the list wholesale price \( w_o \). The discount exactly accounts for the reseller’s incremental holding cost.
from ordering in batches without diversion. In this case, \( w_\eta^* = w_o - H(p - \gamma p_s)^\alpha > (w_o + s)/2 \), which is equivalent to \( H(p - \gamma p_s)^\alpha < w_\eta^* - s \). Thus, the supplier can induce batch orders with a discounted price \( w_\eta^* \) sufficiently high in comparison to the gray market wholesale price to discourage any diversion. The supplier enjoys all the net benefits for the entire channel, \( c_o - c_\eta - H(p - \gamma p_s)^\alpha \) per unit, resulting from economies of scale.

Only when the holding cost is in an intermediate range does gray market diversion occur. Proposition 3 clearly indicates the importance of considering the reseller’s operational parameters when investigating gray market diversion. Such an intermediate range exists when the condition \( c_o - c_\eta > (w_o - s)/2 \) holds. In this diversion range, the optimal discounted price \( w_\eta^* = s + (w_o - s)^2/[4H(p - \gamma p_s)^\alpha] \), which implies a discount size less than \( H(p - \gamma p_s)^\alpha \). This discount size perfectly offsets the reseller’s holding cost for a less-than-full-cycle-inventory as well as the loss incurred in gray market diversion. The supplier enjoys economies of scale by making it just incentive-compatible for the reseller to order in batches followed by an immediate diversion.

**Corollary 1. (Benefit Allocation under Exogenous Resale Price).** In the case of an exogenous resale price, when it is optimal for the supplier to offer an all-unit quantity discount to the reseller to enjoy economies of scale, the supplier takes all the net benefits.

**Proof.** Proof of Corollary 1. By Proposition 3, if \( H(p - \gamma p_s)^\alpha < \min\{(w_o - s)/2, c_o - c_\eta\} \), then \( w_\eta^* = w_o - H(p - \gamma p_s)^\alpha \) and the reseller’s best response is \( I^* = \eta \). The quantity discount \( H(p - \gamma p_s)^\alpha \) per unit off \( w_o \) is just to offset the increased holding cost in induced strategy \( I^* = \eta \) compared to the inventory strategy \( I^o = 0 \) when no quantity discount is offered. Again by Proposition 3, if \( (w_o - s)/2 \leq H(p - \gamma p_s)^\alpha \leq (2c_o - w_o - s)(w_o - s)/[4(c_\eta - s)] \), then \( w_\eta^* = s + (w_o - s)^2/[4H(p - \gamma p_s)^\alpha] \) and the best response is \( I^* = I^o = \sqrt{(w_\eta^* - s)/H(p - \gamma p_s)^\alpha\eta} = (w_o - s)\eta/(2H(p - \gamma p_s)^\alpha) \). For the reseller, the increased holding cost per cycle with length \( I^o/\lambda \) is \( (I^o/\lambda)h = (w_o - s)^2\eta/(4H(p - \gamma p_s)^\alpha) \) as compared to the inventory strategy \( I^* = 0 \); the loss in gray market diversion within the same cycle is \( (w_\eta^* - s)(\eta - I^o) = (w_o - s)^2\eta[1 - (w_o - s)/(2H(p - \gamma p_s)^\alpha)]/(4H(p - \gamma p_s)^\alpha) \). The reseller’s gain from the quantity discount for the
same cycle is $(w_o - w^*)I^o = [w_o - s - (w_o - s)^2/(4H(p - \gamma p_s)^\alpha)](w_o - s)\eta/(2H(p - \gamma p_s)^\alpha)$, which is exactly equal to the sum of the increased holding cost and loss in gray market diversion for the same cycle. Therefore, we can conclude that the supplier’s optimal all-unit quantity discount leaves the reseller with zero profits.

In Lal and Staelin (1984), the motivation behind the supplier’s quantity discount is increased channel efficiency resulting from economies of scale. The size of the discount is just to offset the reseller’s extra inventory holding costs. However, in this chapter the possible channel efficiency from economies of scale is also affected by the gray market. When the gray market prices (both wholesale and resale prices) increase, the diversion and cannibalization effects are complementary in helping the supplier to achieve economies of scale. First, the diversion loss becomes smaller. Second, the higher gray market price leads to less cannibalization and hence increases the authorized channel demand. The diversion option allows the supply chain to achieve the economies of scale when

$$\frac{c_o - c_s}{(p - \gamma p_s)^\alpha} < H \leq \frac{(2c_o - w_o - s)(w_o - s)}{4(c_s - s)(p - \gamma p_s)^\alpha}$$

### 2.4 Model Analysis: Endogenous Resale Price

In this section we extend the previous analysis to allow the resale price to be an endogenous decision variable of the reseller. For tractability, we fix the elasticity by setting $\alpha = 2$, which leads to a demand rate function $\lambda(p) = m/(p - \gamma p_s)^2$, where the gray market resale price $p_s$ is exogenously given. Recall that we also assume $\gamma < 1$. As in the previous section, we will solve the Stackelberg game backward by first examining the reseller’s optimal pricing and inventory decisions in response to the supplier’s quantity discount schedule, and then solving the supplier’s optimal quantity discount.
2.4.1 Reseller’s Pricing and Inventory Policies

The reseller jointly determines the optimal resale price and inventory policy by balancing revenues with ordering, diversion and inventory costs. We solve for the optimal resale price and inventory policy given the supplier’s discount schedule and summarize the results in the following proposition.

Proposition 4. (Reseller’s Optimal Pricing and Inventory Policy). The optimal pricing and inventory policy for the reseller with the demand function \( \lambda(p) = m/(p - \gamma p_s)^2 \) is:

- if \( s < w_\eta < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \),
  \[
  (p^*, I^*) = \begin{cases} 
  (2w_\eta - \gamma p_s, \eta) & \text{if } H \leq \frac{w_\eta - s}{4(w_\eta - \gamma p_s)}, \\
  \left( \frac{2(s - \gamma p_s)}{1 - 2\sqrt{(w_\eta - s)H}} + \gamma p_s, \left[ \sqrt{\frac{w_\eta - s}{4H}} - (w_\eta - s) \right] \frac{\eta}{(s - \gamma p_s)} \right) & \text{if } \frac{w_\eta - s}{4(w_\eta - \gamma p_s)} < H \leq \frac{1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s))^2}}{4(w_\eta - s)}, \\
  (2w_o - \gamma p_s, 0) & \text{otherwise}; 
  \end{cases}
  \]

- if \( \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \leq w_\eta \leq w_o \),
  \[
  (p^*, I^*) = \begin{cases} 
  (2w_\eta - \gamma p_s, \eta) & \text{if } H \leq \frac{1}{4(w_\eta - \gamma p_s)} - \frac{1}{4(w_o - \gamma p_s)}, \\
  (2w_o - \gamma p_s, 0) & \text{otherwise}. 
  \end{cases}
  \]

Proof. Proof of Proposition 4. The reseller selects his resale price \( p \) and inventory policy \( I \) to maximize the expected profit per unit of time. We write the reseller’s profit in terms of the endogenously determined demand rate \( \lambda \): \( \pi(I, \lambda) = \sqrt{m\lambda} + \gamma p_s \lambda - G(I, \lambda) \), where \( \sqrt{m\lambda} + \gamma p_s \lambda \) is the reseller’s revenue per unit of time and \( G(I, \lambda) \) is the total costs per unit of time among all zero-inventory policies characterized by the initial cycle inventory level \( I \). By Proposition 2, for any given resale price \( p \) and its corresponding demand rate \( \lambda \), the reseller will choose the unique inventory policy that minimizes \( G(I, \lambda) \). Using the reseller’s optimal inventory response to an arrival rate, we derive the minimum inventory cost function \( G^*(\lambda) = G(I^*(\lambda)) \) as follows:
if \( s < w_\eta < (w_o + s)/2 \),

\[
G^*(\lambda) = \begin{cases} 
    w_o \lambda & \text{if } 0 < \lambda < \frac{4(w_o - s)mH}{(w_o - s)^2}, \\
    \sqrt{4(w_o - s)mH/\lambda} + s\lambda & \text{if } \frac{4(w_o - s)mH}{(w_o - s)^2} \leq \lambda < \frac{mH}{(w_\eta - s)}, \\
    w_\eta \lambda + mH & \text{otherwise};
\end{cases}
\]

if \((w_o + s)/2 \leq w_\eta \leq w_o\),

\[
G^*(\lambda) = \begin{cases} 
    w_o \lambda & \text{if } 0 < \lambda < \limsup_{c \to w_\eta -} - \frac{mH}{(w_o - c)}, \\
    w_\eta \lambda + mH & \text{otherwise}.
\end{cases}
\]

The above cost function includes the cost of ordering, diversion and holding inventory. When the order size is \( \eta \) and there is no gray market diversion, the reseller enjoys the low unit cost \( w_\eta \) but suffers an average holding cost of \( mH \) per unit. In the case where the reseller diverts to the gray market, the reseller optimizes the diversion quantity \( \eta - I \) by comparing the holding cost \( hI/2 \) with the diversion cost \( w_\eta - s \). The reseller profit function with demand function \( \lambda(p) = m/(p - \gamma p_o)^2 \) is

\[
\pi(\lambda) = \begin{cases} 
    \pi_0(\lambda) := \sqrt{m\lambda} + (\gamma p_s - w_o)\lambda & \text{if } 0 < \lambda \leq \frac{4(w_\eta - s)mH}{(w_o - s)^2}, \\
    \pi^o(\lambda) := \sqrt{m\lambda} + (\gamma p_s - s)\lambda - \sqrt{4(w_\eta - s)mH\lambda} & \text{if } \frac{4(w_\eta - s)mH}{(w_o - s)^2} < \lambda < \frac{mH}{(w_\eta - s)}, \\
    \pi_\eta(\lambda) := \sqrt{m\lambda} + (\gamma p_s - w_\eta)\lambda - mH & \text{otherwise},
\end{cases}
\]

where \( \pi_0, \pi^o \) and \( \pi_\eta \) correspond to when the reseller adopts the inventory policy \( I^* = 0 \), \( I^* = I^o \) and \( I^* = \eta \) respectively. Note that \( \pi_0(\lambda) \) and \( \pi_\eta(\lambda) \) are concave since \( m > 0 \) and \( \pi^o(\lambda) \) is concave if \( \sqrt{m} - \sqrt{4(w_\eta - s)mH} > 0 \). We take the derivative of \( \pi_0(\lambda), \pi^o(\lambda) \) and \( \pi_\eta(\lambda) \) with respect to \( \lambda \) as \( \partial \pi_0(\lambda)/\partial \lambda = \sqrt{m}/(2\sqrt{\lambda}) + \gamma p_s - w_o \), \( \partial \pi^o(\lambda)/\partial \lambda = \sqrt{m}/(2\sqrt{\lambda}) + (\gamma p_s - s) - \sqrt{(w_\eta - s)mH/(2\lambda)} \) and \( \partial \pi_\eta(\lambda)/\partial \lambda = \sqrt{m}/(2\sqrt{\lambda}) + \gamma p_s - w_\eta \). The local optima satisfying the first-order conditions are \( \lambda_1^* = m/4(w_o - \gamma p_s)^2 \), \( \lambda_2^* = \left( \sqrt{m} - \sqrt{4(w_\eta - s)mH} \right)^2/(4(s - \gamma p_s)^2) \) and \( \lambda_3^* = m/(4(w_\eta - \gamma p_s)^2) \) respectively, and the corresponding profits are

\[
\pi_0(\lambda_1^*) = m/(4(w_o - \gamma p_s)), \\
\pi^o(\lambda_2^*) = \left( \sqrt{m} - \sqrt{4(w_\eta - s)mH} \right)^2/(4(s - \gamma p_s)), \\
\pi_\eta(\lambda_3^*) = m/(4(w_\eta - \gamma p_s)) - mH.
\]
Chapter 2. All-Unit Discounts & Gray Markets

The continuity of $\pi(\lambda)$ is easily verified by checking at the two breakpoints

$\lambda_A = 4(w_\eta - s) mH/[(w_o - s)^2]$ and $\lambda_B = mH/[(w_\eta - s)]$. Since

$$\lim_{\lambda \to \lambda_A^-} \frac{\partial \pi_0(\lambda)}{\partial \lambda} < \lim_{\lambda \to \lambda_A^+} \frac{\partial \pi(\lambda)}{\partial \lambda}$$

, we eliminate the breakpoint $\lambda_A$ as a global optimum. Since

$$\lim_{\lambda \to \lambda_B^-} \frac{\partial \pi(\lambda)}{\partial \lambda} = \lim_{\lambda \to \lambda_B^+} \frac{\partial \pi(\lambda)}{\partial \lambda}$$

, the global optimum $\lambda^* = \lambda_B$ only if $\lambda_2^* = \lambda_3^* = \lambda_B$. Hence, we conclude that the global optimum $\lambda^*$ must be one of the local optima $\lambda_1^*$, $\lambda_2^*$ and $\lambda_3^*$.

It remains to check under what conditions each local optimum dominates. First, note that $\lim_{\lambda \to \lambda_B^-} \frac{\partial \pi(\lambda)}{\partial \lambda} = \lim_{\lambda \to \lambda_B^+} \frac{\partial \pi(\lambda)}{\partial \lambda} \leq 0$ if and only if $H \geq (w_\eta - s)/(4(w_\eta - \gamma p_s)^2]$. Hence, a necessary condition for $\lambda_2^*$ to be a global optimum is $H \geq (w_\eta - s)/(4(w_\eta - \gamma p_s)^2]$ and a necessary condition for $\lambda_3^*$ to be a global optimum is $H \leq (w_\eta - s)/(4(w_\eta - \gamma p_s)^2]$. Second, we compare the profit of each of the batch order policies $I^* = I^o$ or $I^* = \eta$ to the profit of the order-as-you-go policy $I^* = 0$:

$$\pi_0(\lambda_1^*) > \pi_0(\lambda_2^*) \leftrightarrow H > (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2/[4(w_\eta - s)]$$

and

$$\pi_0(\lambda_1^*) > \pi_0(\lambda_3^*) \leftrightarrow H > 1/[4(w_\eta - \gamma p_s)] - 1/[4(w_o - \gamma p_s)] \]$$

Lastly, conditioned on whether $w_\eta - \gamma p_s < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}$, the break points on $H$ can be ordered as follows: if

$$w_\eta - \gamma p_s < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}$$

, then

$$(1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2/[4(w_\eta - s)] > (w_\eta - s)/(4(w_\eta - \gamma p_s)^2]$$

and if

$$w_\eta - \gamma p_s \geq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}$$
\[
\frac{1}{4(w - \gamma p_s)} - 1/[4(w_o - \gamma p_s)] \geq (w - s)/[4(w - \gamma p_s)^2].
\]

Therefore, it is not hard to conclude that when \( w - \gamma p_s \geq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} \), the optimal demand rate is \( \lambda^* = \lambda^*_1 \) if \( H > (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2/[4(w - s)] \), \( \lambda^* = \lambda^*_2 \) if \((w - s)/[4(w - \gamma p_s)^2] \leq H \leq (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2/[4(w - s)] \) and \( \lambda^* = \lambda^*_3 \) otherwise; when \( w \geq \sqrt{w_o s} \) the optimal demand rate is \( \lambda^* = \lambda^*_1 \) if \( H > 1/[4(w - \gamma p_s)] - 1/[4(w_o - \gamma p_s)] \) and \( \lambda^* = \lambda^*_3 \) otherwise. By the relationship between price and demand rate \( p(\lambda) = m/\sqrt{\lambda} + \gamma p_s \) and Proposition 2, the corresponding reseller’s optimal pricing and inventory policy follows immediately.

The above proposition describes the impact of market and operational parameters on the resale price in the authorized market. The interesting case occurs when the reseller diverts a portion of its orders to the gray market. As in the case when the resale price is exogenous, only when the discounted wholesale price is sufficiently low and the inventory holding cost is moderate does diversion to the gray market occur. We illustrate this result with Figure 2.4, which depicts the parameter spaces under which diversion strategy becomes optimal. Overall the patterns look very similar to the case under exogenous resale price illustrated by Figure 3. Specifically, Figure 2.4(a) shows the diversion effect in the case without demand cannibalization (\( \gamma = 0 \)). When the gray market wholesale price increases from \( s = 2 \) to \( s = 4 \), the diversion strategy becomes more attractive. All else equal, when the holding cost \( H \) is small, there is a range of \( w \) where the reseller’s optimal response switches from a batch to a diversion strategy in order to reduce the inventory cost. When the holding cost \( H \) is large, there is a range of \( w \) where the reseller’s optimal response changes from order-as-you-go to a diversion strategy in order to take advantage of the quantity discount. Figure 2.4(b) and (c) show the interaction of diversion and cannibalization effects. With positive \( \gamma \) the diversion effect remains, such that when \( s \) is increased from 2 to 4 the diversion strategy is more likely to outperform the order-as-you-go strategy at large \( H \) and outperform the batch strategy at small \( H \). In our examples, as \( s \) increases the gray market resale price \( p_s \) increases
Chapter 2. All-Unit Discounts & Gray Markets

according to a 20% markup over the wholesale gray market wholesale price $s$. As a result, market demand in the authorized channel increases as cannibalization is reduced. This lowers inventory costs associated with higher inventory strategies. Consequently, at high $H$, the benefits to the diversion strategy are greater than to the order-as-you-go strategy. However, at low $H$, the cannibalization effect promotes the batch strategy at the expense of the diversion strategy.

It is useful to note that lowering the resale price and diverting to the gray market are two alternatives for the reseller to order enough to enjoy the quantity discount. Lowering the resale price generates additional demand in the authorized channel. This yields a shorter reorder cycle which reduces cycle holding costs. Diversion to the gray market, by reducing the cycle inventory, decreases cycle length and holding costs of the remaining inventory. In solving its optimal resale price and inventory policy, the reseller balances between lowering the price to increase demand and diverting to the gray market to lower the cycle inventory. When a batch strategy is the optimal inventory policy ($I^* = \eta$), the reseller can sufficiently stimulate demand from the authorized channel to enjoy the quantity discount without resorting to the gray market. Adoption of the diversion strategy occurs when the difference between the discounted wholesale price and the gray market wholesale price is small resulting in correspondingly small diversion loss and an effective wholesale price sufficiently close to the discounted wholesale price. Alternatively, at higher $H$, the diversion strategy may be used at more substantial diversion losses because attempting to achieve the same goal by lowering the price would result in even greater profit loss.

We use Figure 2.5 to illustrate the effect of a gray market on the optimal resale price. In Figure 2.5(a), without a gray market ($s=0$ and $\gamma = 0$), the optimal resale price (solid line) is low when the firm enjoys the discounted wholesale price $w_\eta$ due to low holding cost ($H$), and is high when the firm incurs high holding cost ($H$) and pays regular wholesale price $w_0$. With a positive gray market wholesale price ($s = 2$) but still no demand cannibalization $\gamma = 0$, the resale price increases at medium-low range of $H$ but decreases at medium-high range of $H$. First, when the holding cost is relatively low, the reseller would order in batches
Figure 2.4: Reseller’s Best Responses with Endogenous Resale Price

The solid (resp. dashed) line indicates policy boundaries, $s = 2$ (resp. $s = 4$). $m = 10, \eta = 20, w_o = 10, \alpha = 2$, gray market resale markup = 20%.
Figure 2.5: Reseller’s Optimal Endogenous Resale Price

The solid (resp. dashed) line indicates policy boundaries, $s = 0$ (resp. $s = 2$). $m = 10, \eta = 20, w_o = 10, w_{\eta} = 2.5, \alpha = 2$, gray market resale markup= 20\%.
regardless of the gray market wholesale price. An increase in the gray market wholesale price makes diversion more attractive which reduces the reseller’s inventory. Now that the reseller has a reduced incentive to use a lower price to attract consumers the resale price increases. As a consequence, sales in the authorized channel would decrease. Second, when the holding cost is relatively high, without the gray market the reseller would follow the order-as-you-go strategy. With an increased gray market wholesale price, the reseller would overbuy and benefit from the discount while diverting part of the order to the gray market. The reduction in order costs decreases the resale price which expands the market coverage in the authorized channel.

Figures 2.5(b) and 2.5(c) show the interaction of diversion effect in Figure 2.5(a) and demand cannibalization effect. With positive $\gamma$, a higher gray market resale price leads to a weaker cannibalization effect and higher demand in authorized channel. Given the assumed demand function, the price elasticity increases with the cannibalization parameter $\gamma$. Thus, as shown in Proposition 4, given the reseller’s inventory strategy, the optimal resale price decreases as $\gamma$ and/or $p_s$ increases. In our numerical example featuring a larger value of $\gamma$ and increased price elasticity, the optimal resale price in the authorized channel is always lower when the gray market resale price is increased from $s = 0$ to $s = 2$. The reduced resale price is expected to further expand the market coverage in the authorized channel.

**Corollary 2. (Optimal Bulk Inventory Strategy).** The optimal reseller bulk strategy is:

$$I^n(w_\eta) = \begin{cases} 
\eta & \text{if } H < \frac{w_\eta - s}{4(w_\eta - \gamma p_s)^2} \\
I^o & \text{otherwise}
\end{cases}$$

where $I^o$ is the optimal reseller bulk inventory strategy for all $w_\eta$ if $H \geq 1/[16(s - \gamma p_s)]$.

**Proof.** Proof of Corollary 2. This corollary follows from the profit functions described in the proof of Proposition 4. Namely, a diversion strategy will be preferred to a batch strategy if $\pi^o(\lambda_2^*) \geq \pi_\eta(\lambda_2^*)$. Solving for $H$ shows the transition point occurs at $H_\eta = \frac{w_\eta - s}{4(w_\eta - \gamma p_s)^2}$ confirming the first part of the Corollary. The transition point is a unimodal function of
Given $w_o \eta > s$ with a unique maximizer at $w^*_\eta = 2s - \gamma p_s$. This is confirmed by observing that $\partial H_\eta(s + \epsilon)/\partial w_o \eta > 0$ for small $\epsilon$ and then determining that there is a unique critical point. The proof of the Corollary is completed by evaluating $H_\eta(w^*_\eta)$.

The above corollary highlights the tradeoff between the batch and diversion strategies. It is consistent with Figures 2.5(a)-(c) which show the optimal bulk order strategy transitions from batch to diversion as holding costs increase. The corollary highlights that this tradeoff is independent of the wholesale price but depends strongly on the gray market prices. There are two contributing factors in this tradeoff. (i) As competition from the gray market increases ($p_s$ decreases), the diversion strategy becomes more valuable owing to the increased holding costs from a reduction in authorized channel sales. (ii) The diversion strategy is more valuable when the per unit cost from diversion decreases ($s$ increases). From the supplier’s point of view, this corollary implies that irrespective of $w_0$, if the holding costs are high, the supplier can only achieve economies of scale with prices which lead to diversion.

### 2.4.2 Supplier’s Discount Policy

In this section we examine the supplier’s optimal discount price, taking into consideration the reseller’s best response to the discount schedule. The supplier’s profit depends on the effective wholesale price with or without discount, the reseller’s best-response resale pricing and inventory policies, and the effective supply cost with or without economies of scale. By Proposition 4, it is easy to derive the supplier’s profit function given the reseller’s best response. Then we can proceed to solve the supplier’s profit maximization problem. In addition to the previous assumptions $s < c_\eta$ and $\gamma < 1$, we make an additional assumption $w_o < 4s - 3\gamma p_s$ for tractability. This is not restrictive when the two markets are relatively differentiated (i.e. $\gamma$ is not too close to 1).

**Proposition 5.** (Supplier’s Optimal Discount with Endogenous Resale Price). We assume $w_o < 4s - 3\gamma p_s$. Given that the reseller employs the optimal inventory policy,
the profit-maximizing supplier’s optimal discounted price is:

case (i). if \( 0 \leq 4H < \max\{0,1/(2(c_\eta - \gamma p_s)) - 1/(w_o - \gamma p_s)\} \),

\[
w^*_\eta = \begin{cases} 
2c_\eta - \gamma p_s & \text{if } (w_o - \gamma p_s)^2 \geq 4(c_\eta - \gamma p_s)(w_o - c_o), \\
      w_o & \text{otherwise}; 
\end{cases}
\]

case (ii). if \( \max\{0,1/(2(c_\eta - \gamma p_s)) - 1/(w_o - \gamma p_s)\} \leq 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \),

\[
w^*_\eta = \begin{cases} 
\tilde{w} & \text{if } 4H \leq \frac{1}{2(c_\eta - \gamma p_s)} - \frac{1}{(w_o - \gamma p_s)} + \sqrt{\frac{1}{4(c_\eta - \gamma p_s)^2} - \frac{w_o - c_o}{(c_\eta - \gamma p_s)(w_o - \gamma p_s)^2}}, \\
      w_o & \text{otherwise}; 
\end{cases}
\]

case (iii). if \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \),

\[
w^*_\eta = \begin{cases} 
\tilde{w} & \text{if } 4H \leq \frac{1}{c_\eta - s} \left( \sqrt{\frac{s - \gamma p_s}{w_o - \gamma p_s}} - \frac{(s - \gamma p_s)}{w_o - \gamma p_s} \right) \left( \sqrt{\frac{w_o - \gamma p_s}{s - \gamma p_s}} - \frac{w_o - c_o}{w_o - \gamma p_s} - 1 \right), \\
      w_o & \text{otherwise}; 
\end{cases}
\]

where \( \tilde{w} := 1/\left[ 4H + 1/(w_o - \gamma p_s) \right] + \gamma p_s \) and \( \tilde{w} := s + \left( 1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)} \right)^2 / (4H) \).

Proof. Proof of Proposition 5. It is readily apparent that when the reseller’s best response is to order in batches without any gray market diversion, the supplier enjoys economies of scale from batch processing at the same rate as the demand rate \( \lambda(p^* = 2w_\eta - \gamma p_s) = m/[4(w_\eta - \gamma p_s)^2] \) in the authorized channel. As a result, the supplier’s profit per unit of time is \( \Pi_\eta(w_\eta) := (w_\eta - c_\eta)m/[4(w_\eta - \gamma p_s)^2] \). When the reseller’s best response is to order on demand and not to hold inventory at all, the supplier delivers the product at the list price and the same rate as the demand rate \( \lambda(p^* = 2w_o - \gamma p_s) = m/[4(w_o - \gamma p_s)^2] \) in the authorized channel, and hence the supplier’s profit per unit of time is \( \Pi_0 := (w_o - c_o)m/[4(w_o - \gamma p_s)^2] \) which is independent of the size of the all-unit discount. Finally, when the reseller’s best response is to order in batches with part of the order diverted to the gray market, the supplier enjoys economies of scale from orders of size \( \eta \) every \( I^*/\lambda(p^*) \) time units, and hence the supplier’s profit per unit of time can be shown to be

\[
\Pi^\eta(w_\eta) := \frac{m(w_\eta - c_\eta)}{2(s - \gamma p_s)} \left( \sqrt{\frac{H}{w_\eta - s}} - 2H \right).
\]
As a precursor to establishing this proposition we derive the supplier’s profit function. By Proposition 4, the supplier’s profit function given that the reseller employs the optimal pricing and inventory decisions can be described as

\[
\Pi(w) = \begin{cases} 
\Pi_\eta(w) & \text{if } w \in R^L_\eta \cup R^H_\eta, \\
\Pi_0(w) & \text{if } w \in R^o,
\end{cases}
\]

We let

\[
\bar{w} := \frac{1}{16H + 1/(w_o - \gamma p_s)} + \gamma p_s,
\]

\[
\hat{w} := s + \left(1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)}\right)^2 / 4H,
\]

\[
\bar{w} := \left(1 - \sqrt{1 - 16(s - \gamma p_s)H}/(8H)\right) + \gamma p_s,
\]

\[
\bar{w} := \left(1 + \sqrt{1 - 16(s - \gamma p_s)H}/(8H)\right) + \gamma p_s,
\]

where \(w\) and \(\bar{w}\) are the two real roots, if they exist, of the quadratic equation \(f(w) := 4H(w - \gamma p_s)^2 - w + s = 0\). Then, the regions of quantity discount that induce different reseller pricing and inventory decisions are

\[
R^H_0 := \left\{ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \leq w \leq w_o \mid w \leq \bar{w} \right\},
\]

\[
R^H_\eta := \left\{ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \leq w_o \mid w \leq \bar{w} \right\},
\]

\[
R^L_\eta := \left\{ s < w \leq \frac{4H(w - \gamma p_s)^2 + w}{4H} \mid s \leq 0 \right\},
\]

\[
R^o := \left\{ s < w \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \mid w \leq \hat{w}, 4H(w - \gamma p_s)^2 - w + s > 0 \right\},
\]

\[
R^L_0 := \left\{ s < w \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \mid w > \hat{w} \right\}.
\]

Regions \(R^H_0\) and \(R^H_\eta\) are mutually exclusive, with one of them possibly being an empty set. Regions \(R^L_0\), \(R^o\) and \(R^H_0\) are mutually exclusive, with no more than two of them possibly being an empty set.

To simplify the profit function, we condition on the magnitude of \(H\) according to the following three cases.
case (i). Consider $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$, which is equivalent to $\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s < \bar{w}$, hence $R^H_\eta = \left[\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s, \bar{w}\right]$ and $R^H_0 = (\bar{w}, w_o)$. It can be easily verified that $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$ is also equivalent to $\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s < \bar{w}$, hence $R^L_\eta = \emptyset$. Note that $f(w = s) = 4H(w - \gamma p_s)^2 \geq 0$ and $f(w = \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s) < 0$ when $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$, hence the smaller root $\bar{w}$ must be real-valued and exist between $s$ and $\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s$, i.e., $\bar{w} \in \left[s, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s\right]$. Hence $R^L_\eta = \left[\bar{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s\right]$ and $R^o = (s, w)$. In summary, $R^L_0 \cup R^H_0 = (\bar{w}, w_o)$, $R^L_\eta \cup R^H_\eta = [w, \bar{w}]$ and $R^o = (s, w)$.

case (ii). Consider $1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/[4(s - \gamma p_s)]$. Such an interval of $H$ indeed exists since

$$1/[4(s - \gamma p_s)] + 1/(w_o - \gamma p_s) \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)},$$

with equality holding if and only if $w_o - \gamma p_s = 4(s - \gamma p_s)$. Note that

$$4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$$

is equivalent to

$$\bar{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s,$$

hence $R^H_\eta = \emptyset$ and $R^H_0 = \left[\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s, w_o\right]$. Also note that

$$4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$$

is equivalent to

$$\bar{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s,$$

hence

$$R^L_0 = \left(\bar{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s\right).$$

Furthermore, when $4H \leq 1/[4(s - \gamma p_s)]$, the discriminant of the quadratic equation $f(w) = 0$ is non-negative, hence the roots $\bar{w}$ and $\bar{w}$ of equation $f(w) = 0$ must be real-valued. It is easy
to check that $4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$ is equivalent to $w \leq \hat{w}$. Since $f(w = s) \geq 0$, $f(w = \hat{w}) \geq 0$ and $f(w) = \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \geq 0$ with the last two inequalities ensured by $4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$, the roots $w$ and $\bar{w}$ must exist between $s$ and $\hat{w}$, i.e., $[w, \bar{w}] \subseteq [s, \hat{w}]$. Hence $R^L_\eta = [w, \bar{w}]$ and $R^o = (s, w) \cup (\bar{w}, \hat{w}]$.

In summary, $R^H_0 \cup R^L_0 = (\hat{w}, w_o], R^o = (s, w) \cup (\bar{w}, \hat{w}]$ and $R^L_\eta \cup R^H_\eta = [w, \bar{w}]$.

Case (iii). Consider $4H \leq 1/[4(s - \gamma p_s)]$. According to the case (ii), we know that

$$4H \leq 1/[4(s - \gamma p_s)] \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$$

and

$$4H > 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$$

which leads to

$$R^H_\eta = \emptyset,$$

$$R^H_0 = [\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s}, w_o],$$

$$R^L_0 = (\hat{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s}).$$

Moreover, $4H > 1/[4(s - \gamma p_s)]$ guarantees that $f(w) > 0$ for any $w$, hence $R^L_\eta = \emptyset$ and $R^o = (s, \hat{w}]$. In summary, $R^H_0 \cup R^L_0 = (\hat{w}, w_o], R^o = (s, \hat{w}]$ and $R^L_\eta \cup R^H_\eta = \emptyset$.

Thus, the profit function can be expressed as follows which completes the derivation:

Case (i). if $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$,

$$\Pi(w_\eta) = \begin{cases} 
\Pi^o(w_\eta) & \text{if } s < w_\eta < w, \quad (I^* = I^o) \\
\Pi^*_\eta(w_\eta) & \text{if } w \leq w_\eta \leq \hat{w}, \quad (I^* = \eta) \\
\Pi^*_0 & \text{if } \hat{w} < w_\eta \leq w_o, \quad (I^* = 0)
\end{cases}$$

Case (ii). if $1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/[4(s - \gamma p_s)]$,

$$\Pi(w_\eta) = \begin{cases} 
\Pi^*_\eta(w_\eta) & \text{if } w \leq w_\eta \leq \bar{w}, \quad (I^* = \eta) \\
\Pi^o(w_\eta) & \text{if } s < w_\eta < w \text{ and } \bar{w} < w_\eta \leq \hat{w}, \quad (I^* = I^o) \\
\Pi^*_0 & \text{if } \hat{w} < w_\eta \leq w_o, \quad (I^* = 0)
\end{cases}$$
profit function, \( w \). If \( H < 4 \), the form of the supplier's profit function corresponding to case (i) of the supplier profit function, \( \Pi(w) = \left\{ \begin{array}{ll}
\Pi^o(w) & \text{if } s < w \leq \hat{w}, \quad (I^* = I^o) \\
\Pi_0 & \text{if } \hat{w} < w \leq w_o. \quad (I^* = 0) 
\end{array} \right. \)

Given that the reseller employs the optimal pricing and inventory policy in response to a discounted wholesale price \( w^*_o \), the supplier earns the following profit per unit of time:

Taking the first-order derivative of \( \Pi_\eta(w_\eta) \) with respect to \( w_\eta \), we have \( \partial \Pi_\eta(w_\eta)/\partial w_\eta = m/(4(w_\eta - \gamma p_s)^2)(1 - 2(w_\eta - c_\eta)/(w_\eta - \gamma p_s)) \), hence the function \( \Pi_\eta(w_\eta) \) is increasing on \((0, 2c_\eta - \gamma p_s)\) and decreasing on \([2c_\eta - \gamma p_s, \infty)\). Note that under the assumption that \((w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)\), we have \( \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c_\eta - \gamma p_s \). Taking the first-order derivative of \( \Pi^o(w_\eta) \) with respect to \( w_\eta \), we have

\[
\frac{\partial \Pi^o(w_\eta)}{\partial w_\eta} = \frac{m\sqrt{H}}{4(s - \gamma p_s)(w_\eta - s)^{\frac{3}{2}}}
\left[-4\sqrt{H}(w_\eta - s)^{\frac{3}{2}} + w_\eta + c_\eta - 2s\right].
\]

Taking the second-order derivative of \( \Pi^o(w_\eta) \) with respect to \( w_\eta \), we have

\[
\frac{\partial^2 \Pi^o(w_\eta)}{\partial w_\eta^2} = \frac{m\sqrt{H}(4s - 3c_\eta - w_\eta)}{8(s - \gamma p_s)(w_\eta - s)^{\frac{5}{2}}} < 0.
\]

Under the assumption that \( s < c_\eta \), then \( \partial^2 \Pi^o(w_\eta)/(\partial w_\eta^2) < 0 \) for \( w > s \), namely, \( \Pi^o(w_\eta) \) is strictly concave on \((s, \infty)\). Furthermore, since \( s < c_\eta \), \( \lim_{w_\eta \to s^+} \partial \Pi^o(w_\eta)/\partial w_\eta = \infty \). Under the assumption that \( w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)\),

\[
\frac{\partial \Pi^o(w_\eta)}{\partial w_\eta} \bigg|_{w_\eta = \hat{w}} = \frac{m\sqrt{H}}{4(s - \gamma p_s)(\hat{w} - s)^{\frac{3}{2}}}
\left[(1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2 \right. \\
\left. \frac{4H}{2\sqrt{(s - \gamma p_s)/(w_o - \gamma p_s) - 1}} + c_\eta - s\right] > 0.
\]

Therefore, \( \Pi^o(w_\eta) \) is strictly increasing on \((s, \hat{w})\).

Cases (i) and (ii). In both cases, \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \) with the form of the supplier's profit function corresponding to case (i) of the supplier profit function. If \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \), by the derivation of the supplier profit function, \( w < \hat{w} \). Hence, \( \Pi^o(w_\eta) \) is increasing on \((s, w)\). Note that \( \Pi_\eta(w_\eta) \) and \( \Pi^o(w_\eta) \) are continuous at \( w \). Furthermore, by the derivation of the supplier profit function,
\( \omega < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} \) and we know under the assumption \((w_o - \gamma p_s)/4 < (s - \gamma p_s) < \omega < (e - \gamma p_s)\), \(\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c - \gamma p_s\), hence \(\omega < 2c - \gamma p_s\). Recall that \(\Pi_\eta(w_\eta)\) is increasing on \((0, 2c - \gamma p_s]\) and decreasing on \([2c - \gamma p_s, \infty)\). Therefore, if \(2c - \gamma p_s < \omega\), \(\Pi_\eta(w_\eta)\) is increasing on \([\omega, 2c - \gamma p_s]\) and decreasing on \([2c - \gamma p_s, \omega]\); otherwise, \(\Pi_\eta(w_\eta)\) is increasing on \([\omega, \omega]\). Finally, it remains to compare the supplier’s profit at the list price \(w_o\) and the one at the discounted price \(\min\{2c - \gamma p_s, \omega\}\) that maximizes the profit when a discount is offered.

Case (iii). We consider \(4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)\). First, we consider \(\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/(4(s - \gamma p_s))\) with the form of the supplier’s profit function corresponding to case (ii) of the supplier profit function. Note that if \(4H \leq 1/(4(s - \gamma p_s))\), \(\omega \leq \bar{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c - \gamma p_s\), where the second inequality is shown in the derivation of the supplier profit function and the last is due to the assumption \((w_o - \gamma p_s)/4 < (s - \gamma p_s) < (e - \gamma p_s)\). Hence \(\Pi_\eta(w_\eta)\) is increasing on \([\omega, \bar{w}]\) and maximized at \(\bar{w}\). We also know that \(\Pi^o(w_\eta)\) is increasing on \((s, \bar{w})\) and \([\bar{w}, \bar{w}]\), and obtains its maximum at \(\bar{w}\). Note that \(\Pi_\eta(w_\eta)\) and \(\Pi^o(w_\eta)\) are continuous at \(w\) and \(\bar{w}\). Therefore, if \(1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/(4(s - \gamma p_s))\), the discounted price \(\hat{w}\) maximizes the supplier’s profit when a discount is offered. Second, we consider \(4H > 1/(4(s - \gamma p_s))\) with the form of the supplier’s profit function corresponding to case (iii) of the supplier profit function. Recall that \(\Pi^o(w_\eta)\) is strictly increasing on \((s, \hat{w})\). Therefore, if \(4H > 1/(4(s - \gamma p_s))\), the discounted price \(\hat{w}\) maximizes the supplier’s profit when a discount is offered. Finally, combining the two subcases, it remains to compare the supplier’s profit at the list price \(w_o\) and the one at the optimal discounted price \(\hat{w}\).

\[
\text{If the holding cost } H \text{ is relatively low, the supplier sets the optimal all-unit discount such that, in response, the reseller sets a low resale price to drive up demand and achieve the minimum batch threshold without diversion (see the first sub-cases in cases (i) and (ii)). However, if the holding cost } H \text{ is high enough, the high inventory cost may lead to an optimal discount policy under which the reseller simultaneously resorts to gray market }
\]
diversion and a lower resale price (see the first sub-case in case (iii)). Consistent with its exogenous resale price counterpart, Proposition 5 shows that gray market diversion occurs if and only if the economies of scale are sufficiently large and the reseller’s inventory holding cost falls in an intermediate range of values. Unlike the case with exogenous resale prices, here the reseller can also use a lower resale price to increase demand and reduce inventory costs. If the supplier jointly optimizes the wholesale price \( w_0 \) as well as \( w_\eta \) the results will have a similar structure: if reseller holding costs are moderate or greater, the supplier will select between a diversion and zero inventory policy depending on their scale economies.  

**Corollary 3.** *(Benefit Allocation under Endogenous Resale Price).* In the case of an endogenous resale price, when it is optimal for the supplier to offer an all-unit quantity discount to the reseller to enjoy economies of scale, the reseller shares either none or part of the net benefits of economies of scale. Consumers in the authorized channel are always better off with economies of scale than without.

**Proof.** Proof of Corollary 3. First we check case by case how in Proposition 5 the benefits from economies of scale are allocated between the supplier and reseller. When it is optimal for the supplier to offer a discounted price \( w_\eta^* < w_\eta \) the supplier enjoys economies of scale. By Proposition 4, the reseller’s best response \( (p^*(w_\eta^*), I^*(w_\eta^*)) \) of pricing and inventory is either to take the \( I^* = I^0 \) strategy, namely, \( (p^*(w_\eta), I^*(w_\eta^*)) = \left( \frac{2((s-\gamma p_s)}{1-2\sqrt{(w_\eta-s)H} \right) + \gamma p_s, \left[ \frac{w_\eta-s}{s-\gamma p_s} \right]) \) or the \( I^* = \eta \) strategy, namely, \( (p^*(w_\eta), I^*(w_\eta^*)) = (2w_\eta^* - \gamma p_s, \eta) \).

If the best response of the reseller \( (p^*(w_\eta^*), I^*(w_\eta^*)) = (2w_\eta^* - \gamma p_s, \eta) \), then the reseller’s profit per unit of time is \( \pi(\lambda(p^*(w_\eta^*))) = m/(4(w_\eta - \gamma p_s)) - mH \). When the supplier sets \( w_\eta = w_0 \) and does not enjoy economies of scale, the reseller’s profit per unit of time in the

---

3 The joint optimization problem can be solved easily if resellers are homogenous because only one of the wholesale or discount prices are active at any time and the profits from each price can be computed via a piecewise convex optimization. When the supplier is facing a reseller network made up of heterogenous firms the problem is analytically not tractable. The two supplier prices may be targeted to different resellers and there is a complex interaction resulting from incentive constraints. We have conducted numerical analyses to ensure that there exists a holding cost interval where sufficient economies of scale lead the supplier to induce gray market diversion continue to hold.
best response is \( \pi(\lambda(p^*(w_o))) = m/(4(w_o - \gamma p_s)^2) \). It is easy to see that if \( w^*_\eta \leq \) (resp. <) \( \tilde{w} \), the reseller is (resp. strictly) better off when the supplier offers a discounted price \( w^*_\eta = w^* \) as compared to when the supplier does not, i.e., \( \pi(\lambda(p^*(w^*_\eta))) \geq (\text{resp. } >) \pi(\lambda(p^*(w_o))) \). By the proof of Proposition 5, we can verify that to elicit \( I^* = \eta \), in cases (i) and (ii) of Proposition 5, it is optimal for the supplier to set \( w^*_\eta = \min\{2c_\eta - \gamma p_s, \tilde{w}\} \leq \tilde{w} \). Hence, in case (i), \( w^*_\eta = 2c_\eta - \gamma p_s < \tilde{w} \) and the reseller shares part of the benefits from economies of scale; in case (ii), \( w^*_\eta = \tilde{w} \) and the reseller shares no benefits.

By Proposition 5, the other scenario is that the supplier sets \( w^*_\eta = \hat{w} \) to induce the reseller to take the \( I^* = I^o \) strategy. By Proposition 4, the corresponding best response of the reseller is

\[
(p^*(\hat{w}), I^*(\hat{w})) = \left(2\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - \gamma p_s, \left(\sqrt{(s - \gamma p_s) \over w_o - \gamma p_s} - (s - \gamma p_s) \over w_o - \gamma p_s\right) \eta/(s - \gamma p_s)4H\right)
\]

and the reseller’s profit per unit of time is

\[
\pi(\lambda(p^*(\hat{w}))) = \left(\sqrt{m} - \sqrt{4(\hat{w} - s)mH}\right)^2/(4(s - \gamma p_s), = m/(4(w_o - \gamma p_s)) = \pi(\lambda(p^*(w_o))).
\]

Hence in this scenario, the reseller earns the same profit as without the quantity discount and shares no benefits from economies of scale.

From the analysis of all three cases, we can see that the resale price in equilibrium when the supplier enjoys economies of scale is always strictly smaller than the resale price \( 2w_o - \gamma p_s \) that the reseller will charge if the supplier does not enjoy economies of scale and offers the wholesale price at \( w_o \). Hence, consumers will always enjoy a lower resale price with the supplier’s economies of scale than without.

Unlike the case of exogenous resale prices, when the resale price is endogenous the reseller may share part of the benefits from batch orders induced by the supplier’s optimal quantity discount. When resale prices are exogenous, the supplier’s optimal wholesale price discount just offsets the reseller’s increased inventory holding cost in the induced batch strategy. In the case of an endogenous resale price, the supplier can benefit from additional authorized
market demand when the reseller uses the discounted wholesale price and lowers the resale price. Note that in a typical Stackelberg game with the supplier as the leader and the reseller as the follower facing the downward-sloping demand function \( \lambda(p) = m/(p - \gamma p_s)^2 \), the supplier’s optimal wholesale price with an effective supply cost \( c \) is \( 2c - \gamma p_s \). When \( 2c - \gamma p_s < w_o \), the supplier may enjoy a reduced supply cost \( c \eta \) through batch orders and charge a discounted wholesale price \( 2c - \gamma p_s \). When \( 2c - \gamma p_s < w_o \) and the holding cost is sufficiently low as in case (i), the net benefit of charging the discounted price at \( 2c - \gamma p_s \) is greater than the benefit of charging a wholesale price just offsetting the increased inventory holding cost. In this case, the benefit of increasing sales in the authorized channel dominates the additional subsidy sufficiently to offset the inventory holding cost. As a result, the benefits from the supplier’s economies of scale can trickle down the supply chain from the supplier to the reseller and to the end users.

The above reseller benefits do not carry through to case (ii) because \( 2c - \gamma p_s \) is no longer a feasible discounted price. When the holding cost is at the low end of the holding cost interval corresponding to this case, the supplier provides an optimal quantity discount which is just sufficient to compensate for the reseller’s additional inventory holding costs. The supplier anticipates that the reseller diverts no inventory to the gray market.

The most interesting situation is case (iii) where the reseller follows the diversion strategy if their holding costs are at the low end of the corresponding interval. In Figure 2.4 these moderate levels of \( H \) are where the diversion strategy region borders the zero inventory strategy region. In this case, the reseller again shares none of the economic benefits resulting from batch orders. When the holding cost falls into the bottom end of the range corresponding to this case, the supplier provides the precise quantity discount sufficient to make up for the reseller’s higher inventory and diversion costs. Compared with the case where the supplier does not offer batch discounts, all else equal, when the channel members are able to order in batches, the resale price is always lower and consumers in the authorized channel are always better off.
2.5 Model Extension: Endogenous Gray Market Price

So far we have assumed the gray market prices are exogenous. This is reasonable when a large number of resellers contribute to the gray market such that no one reseller’s diversion decision will change gray market wholesale price. In this section we extend the model by allowing the gray market wholesale price to decrease with the amount of diversion from the focal reseller. This section serves to examine the extent to which the main insights generated in previous sections are robust to endogenous gray market prices. Due to analytical complexity, we limit this investigation to a computational exercise.

To allow the gray market wholesale price to change with the reseller’s diversion, we assume the gray market wholesale price obeys a decreasing linear function, \( S(g) = b - ag \). Parameter \( b \) represents the gray market wholesale price when the focal reseller does not divert. Parameter \( a \) measures sensitivity of the gray market wholesale price to the rate of diversion from the focal reseller. If \( a = 0 \), we revert to cases already analyzed in previous sections. The parameter \( a \) would be expected to be small if diversion from quantity discounts is only one of several sources of supply to the gray market. In this exercise, we further assume that the gray market employs a fixed percentage markup \( u \) in its sale to the end users. The markup \( u \) should be smaller for competitive markets with commodity-type products. After incorporating the above assumptions, the primary market demand in the authorized channel is equal to \( \lambda(p, s) = m / (p - (1 + u)\gamma S(g))^\alpha \). Finally, we assume that the cannibalization parameter \( \gamma \) is sufficiently small so that peak inventory decreases with gray market wholesale price \( s \). That is, \( I_s^o < 0 \). Then, with a lower gray market wholesale price, the reseller has less incentive to divert and consequently holds a larger peak inventory in equilibrium.

An endogenous gray market wholesale price leads to substantial additional complexity in the analysis. When the reseller diverts to the gray market, in equilibrium the reseller needs to optimize the amount of diversion to the gray market conditional on the gray market wholesale price \( s \). We have already solved for the optimal rate \( G_r(s, p) = (\eta - I_s^o(p, s))\lambda(p, s) / I_s^o(p, s) \). With an endogenous gray market wholesale price, gray market demand is also a function of...
the market price \( s \), \( G_g(s) = (b - s)/a \). For the equilibrium to be consistent, we equate the demand with supply and solve for the gray market wholesale price, \( S(p) = \{ s : G_r(s, p) = G_g(s) \} \). Such a \( S(p) \) exists and is unique as long as \( I^*_s < 0 \) holds. The gray market policy selects the profit-maximizing price \( p(w^s) = \arg\max_p (\Pi(p, I^o(p, s), S(p))) \). We use standard numerical optimization techniques (e.g., bisection) to perform root finding and maximization.

Our numerical analysis covers a wide range of parameter values. Overall, as illustrated in Figures 2.6(a) and 2.6(b), our results and qualitative insights uncovered in previous sections remain intact when gray market prices are endogenous and dependent on the reseller’s diversion volume. Figure 2.6(a) shows the reseller’s optimal inventory policy under a low elasticity parameter \( \alpha = 0.5 \) and Figure 2.6(b) shows the case of a moderate elasticity parameter \( \alpha = 2 \). In both cases, once again the reseller follows the diversion strategy when \( w^s \) is sufficiently small and holding cost \( H \) falls in an intermediate range of values. Alternatively, the reseller will follow the batch strategy if \( H \) is too small, and follow the order-as-you-go strategy if \( H \) is too large. When \( a \) is increased from 0 to 2, the gray market prices become more sensitive to the reseller’s diversion rate. Since diversion becomes a less profitable option, there is a range in the parameter space where an increased \( a \) leads the reseller to switch from diversion to order-as-you-go strategies.

Lower gray market price can dampen the reseller’s incentive because, first, more diversion leads to a lower gray market wholesale price and hence a higher unit diversion loss. Second, a lower gray market resale price will lead to increased cannibalization and hence lower profit in the authorized channel. Thus, all else being equal, with a larger value of \( a \), it is less likely for the reseller to divert to the gray market. Figure 2.6(c) shows the boundary defined by \( \alpha \) and \( \gamma \), beyond which the cannibalization effect is too strong for the reseller to divert to gray market.

At lower demand elasticities such as \( \alpha = 0.5 \) shown in Figure 2.6(a), gray market diversion has a smaller effect on the demand in the authorized channel. As a result, the diversion decision is determined primarily by diversion loss, the difference between \( w^s \) and \( s \). As shown in Figure 2.6(a), if \( H \) is low the reseller decides between the diversion and batch strategies. As shown in detail in Figure 2.7(a), when \( H = 3 \) the reseller prefers the diversion
Figure 2.6: Reseller’s Response to Endogenous Gray Market Price

$b = 2, m = 10, \eta = 20, w_o = 10$. Panels (a) and (b) the solid (resp. dashed) refers to $a = 0$ (resp. $a = 2$), $\gamma = 0.5$ and $u = 20\%$. In panel (c) $a = 1$ (resp. $a = 2$) and $u = 0\%$
strategy as long as the discounted price $w_{\eta}$ is less than 5. In this case, when sensitivity parameter $a$ increases from 0 to 2, the amount of diversion $g$ decreases. Interestingly, the decision on the adoption of a diversion strategy remains the same. In contrast, if $H$ is high, the reseller decides between diversion and order-as-you-go strategies. As shown in Figure 2.7(b), when $H = 6$ the reseller selects the diversion strategy as long as the discounted price $w_{\eta}$ is less than 4.7. At this higher $H$, when the gray market wholesale price sensitivity parameter $a$ increases from 0 to 2, not only does the rate of diversion decrease, but, the reseller requires a greater discount in order to adopt the diversion strategy. Overall, when the gray market wholesale price is more sensitive to the reseller’s diversion, we expect the optimal diversion rate to be lower.

Under a supply sensitive gray market price the supplier would receive benefits from eliciting a gray market under similar qualitative requirements: higher holding costs and scale economies. The basic structure does not change except at very high elasticity and cannibalization parameters. In particular, at sufficient holding costs the only bulk policy
which the supplier can elicit is a diversion policy.

2.6 Conclusions

This chapter examines the impact of a gray market on authorized channel members and their decision making, specifically, the reseller’s inventory and pricing decisions, the supplier’s all-unit quantity discount policy, and the welfare of consumers in the authorized channel. We study a Stackelberg game and develop closed-form solutions for the reseller’s and the supplier’s subgame perfect equilibrium decisions. When the gray market wholesale price is sufficiently high, the resellers can use gray market diversion to reduce inventory costs as efficiently as resale price reduction. From an operational perspective, the presence of a gray market, by reducing the reseller’s inventory cost, can help the supplier to take advantage of economies of scale in batch processing. Overall, gray market diversion occurs only in markets where the inventory holding costs fall in an intermediate range of values. In the lower part of this range, the reseller would order in batches even without the option of gray market diversion. The gray market diversion reduces effective inventory size and shortens order cycles. Since the reseller faces reduced pressure to push out the inventory, the resale price goes up and sales volume goes down in the authorized channel. On the other hand, in the higher part of the intermediate range of holding cost, the reseller would follow the order-as-you-go strategy without the option of using the gray market. The gray market diversion reduces inventory holding costs and induces the reseller to order in batches. Now that the reseller enjoys a discounted price through batch orders, the resale price goes down and sales increase in the authorized channel. Given the monopoly nature of the model, the supplier is able to extract all the benefits resulting from gray market diversion. Finally, we extend the model and allow the gray market prices to depend on the amount of diversion from the reseller. We find our results regarding the effect of a gray market to be robust.

These results yield several useful implications on how firms may manage their distribution channels with potential gray market leakage. First, in industries with sufficiently attractive
gray market prices, only those resellers in the intermediate range of inventory holding costs may divert goods to the gray market. This may help managers to identify the resellers prone to gray market diversion. Managers can easily monitor the gray market wholesale price and examine its gap from the discounted price \( w_\eta \) to assess the attractiveness of the gray market. Although our results cannot prescribe specific parameter ranges to predict a reseller’s gray market activities, the managers could use the values of parameters \( \eta \), \( h \), and \( m \) to estimate the gray market activities. Consider a given industry with a constant batch size \( \eta \). One may expect a reseller’s unit holding cost and market demand to depend on the reseller’s geographical location. For instance, a downtown location may have higher unit holding cost than a suburban location, a young and educated city may have a higher consumption rate for IT products than a rural and less tech-savvy market.

Second, managers should be aware of the complex effect of gray market diversion on the authorized channel. Among resellers engaging in gray market activities, those with relatively smaller holding costs use gray market diversion to reduce inventory costs and increase the prices in the authorized channel. This would reduce sales in the authorized channel which typically serve more valuable customers. The managers may consider treating these resellers differently from those with relatively large inventory holding costs. For the latter, the gray market diversion enables the resellers to enjoy batch discounts and pass a part of these discounts on to consumers in the authorized channel. Such incidences of gray market diversions could be treated with more tolerance because of the positive effect on the sales in the authorized channel.

This chapter also provides guidance when control of the gray market is the paramount concern of the supplier. The batch cycle holding cost largely determines whether the reseller will elect to utilize a batch or diversion strategy. If economies of scale are not substantial, the supplier may want to alter the break point \( \eta \). By lowering \( \eta \) and as a result \( H \), resellers will divert a lower volume to the gray market.

Finally, this chapter shifts from the usual focus on legal issues in gray markets and studies the strategic impact to the channel members. Gray markets have existed for a very
long time and have gained momentum, growing in recent years through the online channel. The emergence of online markets makes it easier to establish independent resale operations and to reach a much larger geographic market. Since the gray market will remain for the foreseeable future, firms should fully consider its operational and marketing impacts on their authorized resellers. This chapter demonstrates that simply looking at one aspect (e.g., the pricing decisions) but ignoring another (e.g., the inventory decisions) could lead to erroneous conclusions on gray markets.
Chapter 3

Revenue Maximizing Lead-Time Pricing

3.1 Introduction

Motivation. A range of service and manufacturing firms, such as Amazon, Dell or Federal Express, operate in markets with time-sensitive customers whose willingness to buy and pay for a product or service also depends on the lead time between order placement and delivery. To take advantage of heterogeneous customer preferences – some value speedy service more than others – such firms often offer a menu of differentiated price-lead time options (same-day, two-day, etc.), giving impatient customers the option to pay more for faster delivery while charging less for longer lead times. Such lead time-based price and service differentiation can serve as a valuable revenue management tool. However, the joint problem of designing the revenue-maximizing price-lead time menu and the corresponding capacity scheduling policy is significantly complicated if providers cannot tell apart individual customers but only have aggregate information about their attributes, e.g., based on market research. In this common scenario, all customers choose among all menu options based on their own self-interest, and firms must in turn consider these customer choice constraints in their menu design decisions. While the joint problem of revenue-maximizing price-lead
time design under customer choice has recently attracted considerable interest, as detailed below significant, gaps remain in understanding the properties of the optimal menu and its relationship to the system capacity and to market attributes. This paper contributes to closing these gaps.

**Research questions.** Our analysis addresses the following sets of questions.

1. *Properties of the revenue-maximizing price-lead time menu and scheduling policy.* Should different customer types be offered different lead times or be “pooled” into a common service class with a single lead time? Should the scheduling policy be work conserving or involve strategic delay?

2. *Properties of customer targeting and segmentation.* Which customer types (most, least or moderately time-sensitive?) should be served and which, if any, should be pooled into a common class?

3. *Impact of operations and market attributes.* How do the optimal menu and customer targeting and segmentation strategy depend on the capacity, the market size, and the value-delay cost distribution?

**Positioning and contribution.** Three problem characteristics – revenue maximization, customer choice, and optimization over scheduling policies – jointly distinguish this paper from most of the literature on pricing and lead time decisions: (i) the provider’s objective is to maximize her own revenue or profits, not the total system benefit (see Mendelson and Whang (1990), and Van Mieghem (2000)); (ii) customer types have private information about their own delay cost and can therefore choose among all the offered service classes, unlike settings where a single class can be targeted to each customer type (see Maglaras and Zeevi (2005)); (iii) the provider optimizes over scheduling policies as opposed to optimizing over prices for a given policy (see Rao and Petersen (1998)).

Only a few papers share these three attributes. Like this paper, they study static price-lead time menus for customer types that are heterogeneous in their valuations for instant delivery and in their linear delay costs: the net value of a type with valuation $v$ and delay cost $c$ for expected lead time $w$ equals $v - c \cdot w$. Afeche (2004, 2010) analyze strategic delay in
a model with *binary delay costs*.\(^1\) Katta and J. Sethuraman (2005) provides a partial analysis of pooling in a model with *multiple delay costs* but an exogenous *lead time-invariant ranking of types*.

We make the following contributions to the modeling and solution of the multiple delay cost case.

1. *Generalized customer type model*. The model we propose is distinctive in that it *jointly* considers *multiple delay costs*, unlike Afêche (2004, 2010), and an *endogenous lead time-dependent ranking of types*, unlike Katta and J. Sethuraman (2005). It yields novel insights and offers an unifying framework for explaining the presence or absence of these phenomena in related studies. We return to these connections in Section 3.6.

2. *Insights on the optimal menu and segmentation*. We provide necessary and sufficient conditions, in terms of the capacity, market size and properties of the \((v,c)\)-distribution, for three striking features of the optimal menu and segmentation. *(i)* *Pooling* types with different delay costs; our results are more general and more informative than those in Katta and J. Sethuraman (2005). *(ii)* *Pricing out the middle of the delay cost spectrum* while serving both ends; this feature arises neither in Afêche (2004, 2010) nor in Katta and J. Sethuraman (2005). *(iii)* *Strategic delay* to artificially inflate lead times; our results complement those of Afêche (2004, 2010).

We briefly contrast our study with Katta and J. Sethuraman (2005) and Afêche (2004, 2010) – see Section 3.6 for details. We consider i.i.d. delay costs with a continuous distribution. Valuations and delay costs are perfectly correlated: the valuation of a type with delay cost \(c\) is given by a monotone function \(V(c)\). The *valuation-delay cost ratio* \(V(c)/c\) may be increasing or decreasing. The difference between these cases is important. An increasing \(V(c)/c\) ratio, as in Katta and J. Sethuraman (2005), implies an *exogenous lead time-invariant ranking of types*: for every profitable lead time, net values are increasing in

\(^1\)Yahalom et al. (2006) consider a version of this model with convex increasing delay cost functions and fixed demand rates.
the delay cost \( c \). By contrast, if \( V(c)/c \) is decreasing then the ranking of types is endogenous and lead time-dependent: patient types have a higher net value for long lead times than impatient ones, and vice versa for speedier service.

Katta and J. Sethuraman (2005) is relevant for our results on pooling. In their model, neither pricing out the middle of the delay cost spectrum, nor strategic delay is optimal. To our knowledge, Katta and J. Sethuraman (2005) is the first and only other paper in the queueing literature that considers the pooling phenomenon studied here. Their model is in essence a special case of ours: it assumes an increasing \( V(c)/c \) ratio, implying an exogenous lead time-invariant ranking of types. In their model pooling can be optimal only if the hazard rate of the delay cost distribution is not increasing, and they provide no sufficient optimality conditions in terms of the system parameters. Our optimal pooling results are more general and more informative. First, we show that the necessary conditions for optimal pooling are less restrictive when \( V(c)/c \) is decreasing, i.e., the ranking of customer types is lead time-dependent: pooling may be optimal for every delay cost distribution at some threshold lead time. Second, we provide sufficient conditions for optimal pooling in terms of the capacity, market size and properties of the \((v,c)\)-distribution.

Afèche (2004, 2010) is relevant for our results on strategic delay. Pooling two-dimensional types with binary delay costs is not optimal\(^2\), and pricing out the middle of a binary delay cost spectrum is meaningless. To our knowledge, Afèche (2004, 2010) is the first in the queueing literature to identify strategic delay and characterize necessary and sufficient conditions for its optimality. His model, unlike ours, allows heterogenous valuations for each delay cost level. Our results on optimal strategic delay complement his. In our multiple delay cost case, optimal strategic delay targets a range of types at the low end of the delay cost spectrum, and it may occur jointly with pooling. With binary delay costs, strategic delay is optimal only for a single delay cost level. In our model strategic delay is optimal only if capacity is relatively abundant. His also yields cases with optimal strategic delay only at relatively

\(^2\)Pooling types with the same delay cost but different valuations is known to be without loss of optimality. Hence it is optimal in general to offer at most one, in the binary case exactly one, class per delay cost.
scarce capacity; they are the result of valuation heterogeneity at each delay cost level.

Following the literature review, the paper proceeds as follows. Section 3.2 specifies the model and problem formulation and outlines the analysis. Section 3.3 characterizes the properties of feasible mechanisms. Section 3.4 characterizes the structure of the optimal price-lead time menu in cases where the delay cost distribution yields monotone "virtual delay cost functions" which in turn restrict the possible structures of the optimal menu; Section 3.5 presents numerical examples to illustrate these results. Section 6 discusses some generalizations of our results, and Section 3.7 concludes.

### 3.1.1 Literature Review

This paper bridges streams of research on queueing system control and pricing, and on mechanism design. A vast literature considers the analysis, design and control of queueing systems in settings where, unlike in this paper, the system manager is omniscient and omnipotent, able to fully observe all information and determine all job flows. See Stidham Jr (2002) for a recent survey. As noted, minimizing the delay cost or related measures such as average waiting time or system inventory is a prevalent optimality criterion in these settings. We use the achievable-region approach, pioneered by Coffman Jr and Mitrani (1980) and extended by Federgruen and Groenevelt (1988) and others, to characterize the feasible set of expected delays. As a result, customers are restricted to valuations over expected delays or equivalently are risk neutral over delay. As a result, customers must have linear delay costs. This is a standard assumption in almost all priority pricing literature. Exceptions are Afêche and Mendelson (2004) who discusses multiplicative delay costs and Van Mieghem (2000, 1995) which introduces the generalized $c\mu$ scheduling rule which is asymptotically optimal under the presence of convex delay costs. These papers consider only a welfare maximizing service provider.

Our analysis also draws on standard approaches from the theory of mechanism design, which studies optimal resource allocation problems under private information. Myerson (1981) provides a seminal analysis of optimal auction design. Our work is also related to
papers on mechanism design and the design of price-quality menus, see Mussa and Rosen (1978), Maskin and Riley (1984), Armstrong (1996), Rochet and Chone (1998). However, these papers focus on a more generic notion of quality and, unlike ours, do not consider the relationship between capacity and lead times. Jehiel et al. (1996) consider a problem with externalities. Unlike in their setting, the structure of the externality is endogenous in our model since it depends on the scheduling policy.

Numerous papers study pricing and scheduling for queueing systems with strategic agents. See Hassin and Haviv (2003) for an excellent survey. Naor (1969) is widely credited with the first published analysis in this area. He shows for a FIFO $M/M/1$ queue that individual customers’ decisions are not socially optimal and that this problem can be remedied with a static admission price. There is no question of incentive-compatibility in his nor in other papers that consider customers with identical time-sensitivities and mean service times (e.g., Yechiali (1971), Balachandran (1972), Lippman and Stidham Jr (1977) and Dewan and Mendelson (1990)). Among papers that do consider heterogeneous customers, Kleinrock (1967) first studied priority pricing by ignoring customer incentives, whereas Marchand (1974), Ghanem (1975), Dolan (1978), Mendelson and Whang (1990), and Van Mieghem (2000) focus on and jointly provide a thorough understanding of the incentive-compatible and socially optimal mechanism. Ha (2001) derives incentive-compatible and socially optimal prices in systems where each customer chooses a service rate. Among recent studies on revenue-maximizing price and scheduling policies in the absence of customer choice, Maglaras and Zeevi (2005) consider a system without queueing that offers a Guaranteed Service class and one with Best-effort that shares capacity among customers. By contrast, only partial insights appear to be available on incentive-compatible revenue-maximizing mechanisms.

Considering two customer types with equal mean service times but different linear delay cost rates, Plambeck (2004) uses diffusion approximations to study the joint problem of dynamic lead time quotation, static pricing and capacity sizing for an $M/M/1$ queue in heavy traffic. A key assumption to justify the use of heavy traffic theory is that the patient customers tolerate long lead times. She proposes a policy that is approximately incentive-compatible at
utilizations below 100% and is asymptotically optimal in the limit as the patient customers become very delay-tolerant and the utilization is near 100%. Rao and Petersen (1998) focus on incentive-compatible pricing to maximize revenues but assume (as opposed to derive) certain expected delay functions. They consider a generic congestion model without specific queueing structure: the expected delay for each of a fixed number of priority classes is given by an exogenous function of flow rates. Lederer and Li (1997) study price-delay equilibria under perfect competition; in their setting the $c\mu$ scheduling rule is optimal due to its delay cost-minimizing property.

A couple of papers, including Allon and Federgruen (2009), consider the design of revenue-maximizing price-lead time menus and scheduling policies under competition.

### 3.2 Model, Formulation and Analysis Roadmap

This section introduces the model, formulates the provider’s revenue management problem in a mechanism design framework and describes the analysis roadmap for the problem.

#### 3.2.1 Model

We consider a make-to-order service or manufacturing operation which we model as an $M/M/1$ queueing system. The market consists of a heterogeneous pool of small customers with unit demand who arrive to the facility according to an exogenous Poisson process with rate or market size $\Lambda$ per unit time. The facility has i.i.d. exponential service times with mean $1/\mu$, where $\mu$ is the system’s capacity or service rate. The capacity is not a decision variable, but our results specify how the optimal menu varies with $\mu$. We assume that the marginal cost of service is zero. We study this queueing system in a game-theoretic framework, considering it in the steady-state corresponding to customers’ and the provider’s equilibrium actions.

**Customer preferences.** The preferences of each customer are summarized by two attributes, a value or willingness to pay for immediate delivery and a delay cost parameter
which specifies how delay reduces her willingness to pay. We use the terms “delay”, “lead time”,

and “response time” interchangeably to refer to the entire time interval between order
placement and delivery. We consider a continuum of customer types indexed by $c$, which
denotes the customer’s linear delay cost parameter. It measures her disutility per unit
of lead time. Customer types $c$ are i.i.d. draws from a continuous distribution $F$ with
p.d.f. $f$, which is assumed strictly positive and continuously differentiable on the interval
$C := [c_{\text{min}}, c_{\text{max}}] \subset [0, \infty)$. Let $F = 1 - F$. The service time and delay cost rate distributions
are mutually independent and independent of the arrival process.

A type $c$ customer has a positive valuation $V(c)$ for immediate delivery of the product
or service, where $V : C \to \mathbb{R}_+$ is a strictly increasing and continuous function. Thus the
delay cost rate and valuation of customers are perfectly positively correlated. This model is
well suited for settings where delays deflate values. A variety of important phenomena lead
to delay-driven value losses (see Af`eche and Mendelson (2004)), including physical decay
of perishable goods during transportation delays, technological or market obsolescence of
short life-cycle products such as computer chips or fashion items, and delayed information
in industrial and financial markets.

The analysis focuses on the case where $V(c)$ is affine and given by $V(c) = v + c \cdot d$, where
$v > 0$ and $d > 0$ are constants. Section 3.6.2 briefly discusses how the main properties of the
optimal menu generalize in cases where $V(c)$ is not affine. A type $c$ customer who experiences
a lead time $t$ and pays a price $p$ has net value or willingness to pay $V(c) - c \cdot t = v + c \cdot (d - t)$
and her utility is $v + c \cdot (d - t) - p$. Observe that the ranking of customer types’ net values
depend on the available lead times: net values for long lead times $t > d$ decrease while
those for short lead times $t < d$ increase in the time-sensitivity $c$. This plausibly describes
situations where customers with high time-sensitivity get very little utility from a product or
service if their delay is long while more patient customers with a smaller budget have lower
willingness-to-pay for speedy service but are willing to pay more than impatient customers
for somewhat delayed delivery. For example, a customer may be willing to pay a lot for
overnight delivery of a time-critical shipment but only very little if it takes two or three
days, while a more patient customer may be willing to pay more than impatient customers
for delivery in several business days.

Customers are risk neutral. They decide which service class to purchase, if any, based
on the prices and expected steady-state lead times posted on the menu as discussed below.
To rule out the case where it is unprofitable to operate we assume that the most patient
type has positive expected net value in the absence of waiting: $v + c_{\text{min}} \left( d - \frac{1}{\mu} \right) > 0$ or
$\mu > \frac{c_{\text{min}}}{v + c_{\text{min}} \cdot d}$.

Information structure. The arrival process, the delay cost rate distribution $f$, the
function $V(c)$ and the service time distribution are known to the provider. A customer’s
type $c$ is her private information. However, a customer does not know her exact service time
when making her purchase decision; it only becomes known – to her and to the provider – once
her job is processed to completion. Only the provider observes the system state; customers
lack this information when making their decisions and base their lead time forecasts on the
posted mean lead times.

Provider and customer decisions. The provider’s objective is to choose a price-lead
time menu and a scheduling policy to maximize her long run average revenue rate. We
study the system in steady-state under static menus of service classes, each characterized by
a per job price and an expected steady-state delay. We henceforth simply say “lead time”,
“mean delay” or “delay” when referring to the expected steady-state delay of a service class.
Since customers’ types are their private information the provider must make all service classes
available to them and consider the resulting customer choice behavior in designing her menu.
Customers are strategic in deciding whether to place an order and in choosing their service
class upon arrival at the facility. However, they do not choose their arrival time, do not
affect the subsequent arrival process if they do not place an order, and cannot later renege
if they place an order.

To formalize this revenue management problem, the provider selects a menu of lead
time-price options $\{(w, P(w)) : w \in W\}$ where $W$ denotes the set of offered lead-times and
$P : \mathcal{W} \rightarrow \mathbb{R}$ the function which assigns a price for service at a given lead time. Upon arrival at the facility a customer of type $c$ determines the lead time-price pair $(w, P(w))$ which maximizes her expected utility from service $U(w, P(w); c) := v + c \cdot (d - w) - P(w)$. Formally, denote by

\[ w(c) := \arg \max_{w \in \mathcal{W}} \{U(w, P(w); c)\}, \tag{3.1} \]

\[ p(c) := P(w(c)) \tag{3.2} \]

the preferred lead time-price pair for type $c$ and let $U(c) := U(w(c), p(c); c) = v + c \cdot (d - w(c)) - p(c)$ denote her corresponding expected utility from service. Customers who do not purchase service balk and receive zero utility, so they only purchase service if their expected utility from service is non-negative. Our model keeps track of customers’ purchase decisions with the acceptance function $a : \mathcal{C} \rightarrow \{0, 1\}$ where $a(c) = 1$ if type $c$ customers buy service, choosing the lead time-price pair $(w(c), p(c))$, and $a(c) = 0$ otherwise. Let $\mathcal{C}_a := \{c \in \mathcal{C} : a(c) = 1\}$ denote the set of customer types that buy service and $\overline{\mathcal{C}}_a = \mathcal{C} \backslash \mathcal{C}_a$ be its complement. Customer types’ best responses to a menu must satisfy $c \in \mathcal{C}_a$ if $U(c) > 0$ and $c \not\in \mathcal{C}_a$ if $U(c) < 0$. However, customer types with zero expected utility may or may not purchase service (if $U(c) = 0$ then $c \in \mathcal{C}_a$ or $c \not\in \mathcal{C}_a$) as further discussed below.

**Lead times and scheduling policy.** We assume that customers are risk neutral with respect to lead time uncertainty and that they base their decisions on the *announced* mean delays, based on the notion that they generally possess neither the required information (about queue lengths, scheduling policy, arrival rates, etc.) nor the analytical sophistication to reliably forecast their actual delays at the time of their order decision. However, the provider is committed to ensure that the announced lead times equal the realized mean delays given the capacity $\mu$, the scheduling policy, and customers’ equilibrium purchase decisions. Specifically, given the capacity of $\mu$ jobs per unit time, the provider must determine a scheduling policy which in expectation will provide delay of $w(c)$ for each customer $c$ where $a(c) = 1$.

We do not a priori assume a specific scheduling policy but allow the provider to select
any policy from the set of admissible policies, which has the following properties:

1. We restrict attention to nonanticipative and regenerative policies. This appears to be the most general, easily describable restriction under which the existence of long-run averages of waiting times may be verified, see Federgruen and Groenevelt (1988). In particular, it excludes policies where scheduling decisions depend on prior knowledge of customers’ actual (remaining) service times since this information is unavailable, or where decisions in one busy period may depend on information concerning prior busy periods.

2. We do not restrict attention to work conserving policies. In particular, we allow for the insertion of strategic delay whereby the provider artificially increases the mean lead times for a subset of service classes above the levels that are operationally achievable given the system utilization. See Afeche and Mendelson (2004); Afeche (2010) for a detailed discussion of strategic delay and its implementation through the control of server idleness, server speed and/or delivery delays of completed jobs.

3. We allow preemption. This assumption does not affect our results but simplifies the analysis: it implies that the lead time of a given service class depends only on the arrival rates to classes with higher priority but not on the total arrival rate.

We build on the achievable region approach, see Coffman Jr and Mitrani (1980), which allows us to circumvent the problem of determining an optimal scheduling policy by considering the equivalent problem of finding the optimal mean lead times in the set of achievable mean lead times corresponding to the set of admissible scheduling policies. The achievable region for the class of scheduling policies described by 1-3 is defined as follows for an $M/M/1$ system. If the system is stable, i.e.,

$$\mu > \Lambda \int_{x \in C_u} f(x) \, dx,$$

(3.3)
then the mean delays $\{w(c) : c \in C_a\}$ are achievable if and only if

$$\frac{\Lambda}{\mu} \int_{x \in s} f(x) w(x) dx \geq \frac{\Lambda}{\mu - \Lambda} \int_{x \in s} f(x) dx, \quad \forall \text{ subsets } s \subset C_a. \tag{3.4}$$

$$w(c) \geq \frac{1}{\mu}, \quad \forall c \in C_a. \tag{3.5}$$

The constraints (3.4) require that the long run average work in the system for every subset of admitted customers $s \subset C_a$ (the LHS) exceeds the average work under a work conserving policy that gives all admitted customers in the set $s$ strict priority over all admitted customers in $C_a \setminus s$, where the RHS of (3.4) equals the average work in a FIFO $M/M/1$ system with arrival rate $\Lambda \int_{x \in s} f(x) dx$ and service rate $\mu$. A scheduling policy is work conserving if (3.4) is binding for $s = C_a$.

### 3.2.2 Mechanism Design Formulation

A lead time-price menu $\{(w, P(w)) : w \in W\}$ induces customers to make purchase decisions and service class choices that maximize their expected utility and are characterized by the triple of functions $(a, w, p)$ as described above. It is analytically convenient to view the provider’s problem of choosing a menu to maximize her expected revenue rate subject to these customer decisions as a mechanism design problem. Based on the revelation principle, mechanism design problems restrict attention w.l.o.g. to direct revelation mechanisms in which each customer directly reveals her type to the provider who then allocates products and charges customers following previously announced rules. The procedure described above is strictly speaking not a direct revelation mechanism – customers reveal their types only indirectly, but it is more descriptive of how services are sold. It is also de facto equivalent to a direct revelation mechanism whereby the provider selects the functions $(a, w, p)$ such that all customers truthfully reveal their types. This requires that $(a, w, p)$ satisfy the individual rationality (IR) and incentive-compatibility (IC) constraints. The IR constraints require that the expected utility from service is non-negative for types who are targeted for service and non-positive for all others: $U(c) \geq 0$ for $c \in C_a$ and $U(c) \leq 0$ for $c \notin C_a$. The IC constraints
require that for each type \( c \) truthfully revealing her type maximizes her expected utility from service: \( U(c) \geq U(w(c'), p(c'); c) \) for \( c \neq c' \). The provider solves the following mechanism design problem.

**Problem 1.**

\[
\max_{a: \mathcal{C} \to \{0, 1\}, \ w: \mathcal{C} \to \mathbb{R}, \ p: \mathcal{C} \to \mathbb{R}} \quad \Lambda \int_{c_{\min}}^{c_{\max}} a(x) f(x) p(x) dx \quad (3.6)
\]

\[
s.t.
\]

\[
\mu > \Lambda \int_{x \in \mathcal{C}_a} f(x) dx, \quad (3.7)
\]

\[
\frac{\Lambda}{\mu} \int_{x \in s} f(x) w(x) dx \geq \frac{\Lambda}{\mu} \int_{x \in s} f(x) dx, \quad \forall s \subset \mathcal{C}_a, \quad (3.8)
\]

\[
w(c) \geq \frac{1}{\mu}, \quad \forall c \in \mathcal{C}_a, \quad (3.9)
\]

\[
U(c) = v + c(d - w(c)) - p(c) \geq 0, \quad \forall c \in \mathcal{C}_a, \quad (3.10)
\]

\[
U(c) = v + c(d - w(c)) - p(c) \leq 0, \quad \forall c \in \overline{\mathcal{C}}_a, \quad (3.11)
\]

\[
w(c) \cdot c + p(c) \leq w(c') \cdot c + p(c'), \quad \forall c \neq c'. \quad (3.12)
\]

The constraints (3.7)-(3.9) restate (3.3)-(3.5), where (3.7) ensures that the system is stable and (3.8)-(3.9) that the lead times \( \{w(c) : c \in \mathcal{C}_a\} \) are *operationally achievable*; (3.10)-(3.11) capture the IR constraints and (3.12) the IC constraints. We call the triple \((a, p, w)\) feasible if it satisfies (3.7)-(3.12). Given a feasible \((a, p, w)\) the corresponding menu satisfies \( \mathcal{W} = \{w(c) : c \in \mathcal{C}\} \) and \( P(w(c)) = p(c) \) for \( w(c) \in \mathcal{W} \).

### 3.2.3 Analysis Roadmap

We develop the solution of Problem 1 using the following 3-step approach.

**STEP 1. Feasible customer segmentation and lead times, optimal prices.** We translate the IR and IC constraints (3.10)-(3.12) into equivalent properties that any feasible and revenue-maximizing triple \((a, p, w)\) must satisfy. These properties yield a segmentation of customer types into those receiving “low” \((w(c) > d)\), “medium” \((w(c) = d)\) and “high” \((w(c) < d)\)
lead time quality. They also characterize the optimal prices for given segmentation and lead times, which reduces Problem 1 into one of choosing the arrival rates for these segments and the lead times to offer them.

STEP 2. Optimal customer segmentation and lead times for fixed arrival rate. In this step we solve the reduced problem for a fixed arrival rate. The solution characterizes the optimal segmentation structure and lead time menu depending on $\lambda, \Lambda, \mu$, the parameter $d$ and the function $f$.

STEP 3. Optimal arrival rate, customer segmentation and lead times. This final step consists of optimizing over arrival rates. The solution yields insights on how the optimal segmentation structure and lead time menu (at the optimal arrival rate) depends (i) on the demand parameter $v$ for given capacity $\mu$ and (ii) on the capacity for given demand parameters.

We discuss STEP 1 in Section 3.3, STEP 2 in Section 3.4.1 and STEP 3 in Section 3.4.2.

### 3.3 Feasible Customer Segmentation and Lead Times, Optimal Prices

Given a triple $(a, w, p)$ we segment the admitted customer types by the following partition of $\mathcal{C}_a$:

$$
C_l \triangleq \{ c \in \mathcal{C}_a : w(c) > d \}, \quad C_m \triangleq \{ c \in \mathcal{C}_a : w(c) = d \}, \quad \text{and} \quad C_h \triangleq \{ c \in \mathcal{C}_a : w(c) < d \},
$$

(3.13)

where $\mathcal{C}_a = C_l \cup C_m \cup C_h$. For simplicity we suppress the dependence of $C_l, C_m$ and $C_h$ on $a$. We refer to service classes with lead time $> d$ as having, and to $C_l$ as the segment of types buying, low lead time quality. Similarly, we refer to classes with lead time $< d$ ($= d$) as having, and to $C_h$ ($C_m$) as the segment of types buying, high (medium) lead time quality. The subscripts $l, m$ and $h$ are mnemonic for low, medium and high lead time quality, respectively, keeping in mind that lower quality refers to a higher lead time and vice versa. As
shown below the optimal segmentation of types into $C_l$, $C_m$ and $C_h$ depends on the demand parameters and the capacity level; e.g., while low lead time qualities, if purchased at all, are purchased by the type with lowest time-sensitivity ($c_{\min}$), types with significant time-sensitivity (large $c$) may also buy them. The following Proposition translates the IR and IC constraints (3.10)-(3.12) into equivalent properties that any feasible and revenue-maximizing triple $(a, p, w)$ must satisfy.

**Proposition 6. [Properties of Incentive-Compatible Menus]**

Fix a triple $(a, w, p)$. Define the boundary types $c_l$ and $c_h$ as follows:

$$
c_l := \begin{cases} 
\sup C_l & \text{if } C_l \neq \emptyset \\
\min c & \text{otherwise}
\end{cases}, \quad \text{and} \quad 
c_h := \begin{cases} 
\inf C_h & \text{if } C_h \neq \emptyset \\
\max c & \text{otherwise}
\end{cases}.
$$

Suppose that $(a, w, p)$ maximizes the revenue rate. Then $(a, w, p)$ satisfies the IR and IC constraints (3.10)-(3.12) if and only if the following properties hold.

1. Lead times $w(c)$ are non-increasing, prices $p(c)$ are non-decreasing, and $c_l \leq c_h$.

2. If there is a segment of types who buy low lead time qualities ($C_l \neq \emptyset$) then: (i) it is an interval that includes $c_{\min}$, i.e., $c < c_l \Rightarrow c \in C_l$; (ii) prices and expected utilities from service satisfy:

$$
p(c) = v + c \cdot (d - w(c)) - \int_c^{c_l} (w(x) - d) \, dx, \quad \forall c \in [c_{\min}, c_l],
$$

where $p(c) < v$ for $c < c_l$, \hspace{1cm} (3.14)

$$
U(c) = \int_c^{c_l} (w(x) - d) \, dx, \quad \forall c \in [c_{\min}, c_l],
$$

where $U(c) > 0 = U(c_l)$ for $c < c_l$. \hspace{1cm} (3.15)

3. If there is a segment of types who buy high lead time qualities ($C_h \neq \emptyset$) then:

   (i) it is an interval that includes $c_{\max}$, i.e., $c > c_h \Rightarrow c \in C_h$;
(ii) prices and expected utilities from service satisfy:

\[ p(c) = v + c \cdot (d - w(c)) - \int_{c_h}^{c} (d - w(x)) \, dx, \quad \forall c \in [c_h, c_{\text{max}}], \]

where \( p(c) > v \) for \( c > c_h \), \quad (3.16)\]

\[ U(c) = \int_{c_h}^{c} (d - w(x)) \, dx, \quad \forall c \in [c_h, c_{\text{max}}], \]

where \( U(c) > 0 = U(c_h) \) for \( c > c_h \). \quad (3.17)\]

4. If there is a segment of types who buy the medium lead time (\( C_m \neq \emptyset \)) then: (i) \( C_m \subset [c_l, c_h] \); (ii) the prices and expected utilities from service satisfy \( p(c) = v \) and \( U(c) = 0 \) for \( c \in C_m \).

5. Types in \((c_l, c_h)\) buy the medium lead time or do not buy at all, i.e., \((c_l, c_h) \subset C_m \cup \overline{C_a}\), and

\[ U(c) = U(c_l) - \int_{c_l}^{c} (w(x) - d) \, dx = U(c_h) - \int_{c}^{c_h} (d - w(x)) \, dx \leq 0, \quad \forall c \in [c_l, c_h], \]

where \( U(c) = 0 \ \forall c \in [c_l, c_h] \) if some types buy the medium lead time (\( C_m \neq \emptyset \)). \quad (3.18)\]

**Proof of Proposition 6**

See Problem 1. *The constraints (3.10)-(3.12) \Rightarrow Parts 1.-5.*

Write \( U(c'; c) \) for the expected utility of a type \( c \) who reports type \( c' \), and \( U(c) \triangleq U(c; c) \), where

\[ U(c'; c) \triangleq U(w(c'), p(c') \mid c) = v + c (d - w(c')) - p(c') = U(c') + (c - c') (d - w(c')) . \quad (3.19)\]

The IC constraints (3.12) require that the expected utilities from service satisfy for any pair of types:

\[ U(c) = v + c (d - w(c)) - p(c) \geq U(c'; c) \Leftrightarrow c \cdot w(c) + p(c) \leq c \cdot w(c') + p(c') \text{ for } \forall c \neq c'. \quad (3.20)\]
Similarly, we must have $U(c') \geq U(c; c')$, so the IC constraints (3.12) are equivalent to

\[(c - c') (d - w(c)) \geq U(c) - U(c') \geq (c - c') (d - w(c')) \quad \text{for } \forall c \neq c'. \tag{3.21}\]

It follows from (3.21) that the expected utility from service $U(c)$ is continuous in the type.

**Part 1.** If $c < c'$ then (3.21) implies $w(c') \leq w(c)$, and (3.20) implies $p(c') - p(c) \geq c (w(c) - w(c')) \geq 0$. It follows from (3.20)-(3.21) that $w(c)$ is nonincreasing and $p(c)$ is nondecreasing in $c$. The fact $c_l \leq c_h$ follows since $w$ is nonincreasing in $c$. Since $w$ is nonincreasing it is Riemann integrable, and (3.21) implies

\[U(c'') - U(c') = \int_{c'}^{c''} (d - w(x))dx \quad \text{for all } c' < c''. \tag{3.22}\]

**Part 2.** We first show (i). The case $C_l = \{c_{\min}\}$ is trivial. Suppose that $C_l \neq \emptyset$ and $c_l > c_{\min}$. Fix $c \in [c_{\min}, c_l)$. We show that $c \in C_l$. Apply (3.22) with $c = c'$ and $c_l = c''$ to get

\[U(c) = U(c_l) + \int_{c}^{c_l} (w(x) - d)dx > U(c_l) \geq 0. \tag{3.23}\]

The first (strict) inequality follows since $w(x) > d$ for $x < c_l$; otherwise, if $w(x) \leq d$ for some $x < c_l$ then $w(x') \leq w(x) \leq d$ for $x' > x$ since $w(c)$ is nonincreasing, contradicting that $c_l = \sup C_l$. That $U(c_l) \geq 0$ follows since $U(c)$ is continuous in $c$: if $U(c_l) < 0$ then by continuity it must be that $U(c) < 0$ for all $c$ in some interval $[c_s, c_l]$ and the IR constraint (3.10) for $c \in [c_s, c_l]$ can only hold if $c \notin C_l$. But this contradicts that $c_l = \sup C_l$, so $U(c_l) \geq 0$. Therefore, $U(c) > 0$ for $c < c_l$; the IR constraint (3.11) for $c$ only holds if $c \in C_a$, and since $w(c) > d$ for $c < c_l$ it follows that $c \in C_l$. To prove (ii) it remains to show that $C_l \neq \emptyset \implies U(c_l) = 0$ (which we prove with Part 5) for then (3.23) reduces to (3.15) and the price equation (3.14) follows since $U(c) = v + c (d - w(c)) - p(c)$.

**Part 3.** Follows from the same line of argument as in the proof of Part 2, by applying (3.22) with $c_h = c' \leq c = c''$ and showing that $C_h \neq \emptyset$ implies $U(c_h) = 0$, which we prove with Part 5.

**Part 4.** Suppose that $C_m \neq \emptyset$. Parts 2-3 imply $C_m \subset [c_l, c_h]$. Fix $c \in C_m$. Then
\[ U(c') \geq U(c; c') = U(c) = v - p(c) \geq 0 \text{ for } \forall c' \neq c. \] (3.24)

The first inequality follows from the IC constraints (3.12), the equalities hold since \( w(c) = d \), and the IR constraint (3.10) implies the second inequality. Let \( U_m \triangleq U(c) = v - p(c) \). It follows from (3.24) that \( U(c') = U_m \) for all \( c' \in C_m \). It remains to show that \( U_m = 0 \), which we prove with Part 5.

**Part 5.** That \( (c_l, c_h) \subset C_m \cup \overline{C}_a \) is immediate from the definitions of \( c_l \) and \( c_h \). The expression (3.18) for \( U(c) \) follows from (3.22). We prove that \( U(c) \leq 0 \) for \( c \in [c_l, c_h] \) for three exhaustive cases.

(i) Not all types are served \( (\overline{C}_a \neq \emptyset) \) but some types buy the medium lead time \( (C_m \neq \emptyset) \). This implies that \( c_l < c_h \). Then (3.11) and (3.24) imply \( 0 \geq U(c') \geq U(c) = U_m \geq 0 \) for any types \( c \in C_m \) and \( c' \in \overline{C}_a \); therefore we have \( U(c) = 0 \) for all \( c \in (c_l, c_h) \). Since \( U(c) \) is continuous it follows that \( U(c_l) = 0 = U(c_h) \). Since \( C_m \subset [c_l, c_h] \) it follows that \( U(c) = U_m = 0 \) for \( c \in C_m \).

(ii) Not all types are served \( (\overline{C}_a \neq \emptyset) \) and no types buy the medium lead time \( (C_m = \emptyset) \). It follows that \( (c_l, c_h) \subset \overline{C}_a \). The IR constraints (3.11) and the continuity of \( U(c) \) imply \( U(c) \leq 0 \) for \( c \in [c_l, c_h] \). If \( C_l \neq \emptyset \) then (3.23) implies \( U(c_l) \geq 0 \) and so \( U(c_l) = 0 \). Similarly, \( U(c_h) = 0 \) if \( C_h \neq \emptyset \).

(iii) All types are served \( (\overline{C}_a = \emptyset) \). Let \( U_{\text{min}} = \min_c U(c) \). The IR constraints (3.10) require \( U_{\text{min}} \geq 0 \), and revenue-maximization requires \( U_{\text{min}} = 0 \). The proof is complete if \( U(c) = U_{\text{min}} \) for \( c \in [c_l, c_h] \). First note that \( U(c) = U(c_l) = U(c_h) \) for \( c \in [c_l, c_h] \). If \( c_l = c_h \) this is trivial. If \( c_l < c_h \) this holds since then \( (c_l, c_h) \subset C_m \), Part 4 implies \( U(c) = U_m \) for \( c \in C_m \), and by continuity \( U_m = U(c_l) = U(c_h) \). By Parts 2-3 we have \( U(c) > U(c_l) = U(c_h) \) for \( c \notin [c_l, c_h] \) which proves that \( U(c) = U_{\text{min}} \) for \( c \in [c_l, c_h] \).

**Parts 1.-5. ⇒ the constraints (3.10)-(3.12).** Parts 2-5. imply the IR constraints (3.10)-(3.11). The IC constraints (3.12) are equivalent to (3.21). Substituting for \( U(c) \) from Parts
Chapter 3. Revenue Maximizing Lead-Time Pricing

2-5, (3.21) is equivalent to

\[(c'' - c')(d - w(c'')) \geq U(c'') - U(c') = \int_{c'}^{c''} (d - w(x))dx \geq (c'' - c')(d - w(c')) \text{ for } \forall c' < c''.\]

By Part 1, \(w(c') \geq w(x) \geq w(c'')\) for \(x \in [c', c'']\), which establishes both inequalities.

By Proposition 6 any lead time price-menu that satisfies the IR and IC constraints segments the types who purchase service into up to three sets. The sets \(C_l\) and \(C_h\) are intervals that, if non-empty, include the least \((c_{\text{min}})\) and the most \((c_{\text{max}})\) time-sensitive type, respectively. In particular, if \(C_l \neq \emptyset\) and \(w(c) > d\) for some type \(c \in C_l\) then every less time-sensitive type with \(c' < c\) will purchase service with a low lead time quality \((c' \in C_l)\) and get strictly more utility than \(c\). Similarly, if \(C_h \neq \emptyset\) and \(w(c) < d\) for some type \(c \in C_h\) then every more time-sensitive type with \(c' > c\) will purchase service with a high lead time quality \((c' \in C_h)\) and get strictly more utility than \(c\).

The set of customers \(C_m\) who purchase the medium lead time quality \(d\) is a subset of \([c_l, c_h]\) but need not itself be an interval. If \(C_m \neq \emptyset\), then lead time \(d\) is offered at a price of \(v\); every type has zero expected utility from this lead-time price option, but only types in \([c_l, c_h]\) have no better service option available and are indifferent between purchasing service and not doing so. If \(\lambda_m > 0\) then different types buy the same lead time; we call this service pooling and further discuss it below.

Based on Proposition 6 choosing the segments \(C_l, C_m\) and \(C_h\) can be reduced to choosing the corresponding demand rates. Let \(\lambda_l, \lambda_m\) and \(\lambda_h\) denote the demand rates for low, medium and high lead time quality, respectively, and let \(\lambda \triangleq \lambda_l + \lambda_m + \lambda_h\). A feasible function \(a\) determines \(c_l, c_h\) and

\[\lambda_l = \Lambda F(c_l), \lambda_h = \Lambda F(c_h), \lambda_m = \Lambda \int_{c_l}^{c_h} a(x)f(x)dx = \lambda - \lambda_l - \lambda_h.\]  

(3.26)

Conversely, the demand rates \((\lambda_l, \lambda_m, \lambda_h)\) determine the boundary types \(c_l = F^{-1}(\lambda_l/\Lambda)\) and \(c_h = F^{-1}(\lambda_h/\Lambda),\) where \(a(c) = 1\) for \(c < c_l\) and \(c > c_h\). However, the function \(a\) is
not uniquely determined for \( c \in [c_l, c_h] \). If \( C_m \neq \emptyset \) and \( \lambda_m < \Lambda - \lambda_l - \lambda_h \) then only a fraction \( \lambda_m / [\Lambda - \lambda_l - \lambda_h] < 1 \) of types in \([c_l, c_h]\) buy service with medium lead time. Any two feasible triples \((a, w, p)\) and \((a', w, p)\) with the same mass of types in the interval \([c_l, c_h]\) buying the medium lead time are revenue equivalent.

The mechanics underlying the admission of zero utility customers deserves additional discussion. There are several ways in which such admission can be supported. Most prominent, is a mixed strategy equilibrium where types with \( c \in [c_l, c_h]\) buy service with probability \( \lambda_m / [\Lambda - \lambda_l - \lambda_h] < 1 \). Deviation from this strategy would lead the system to deliver a delay not equal to \( d \) leading to an appropriate correction. If for instance demand were low, delay would be less than \( d \) incentivizing an increase in the probability of participating. Two remarks are in order on this equilibrium structure with a positive mass of zero-utility customers who buy the service. First, this structure is quite standard in the literature whenever a positive mass of types have identical preferences with respect to a service option, see Hassin and Haviv (2003). In the simplest case customers are identical – they have the same valuation and delay cost. Therefore the provider charges a single price that leaves all customers with zero expected utility from service. Unless the entire market is served, any equilibrium has the feature that some, but not all, zero-utility customers buy the service. Second, while this equilibrium structure follows in our model because the valuation-delay cost relationship \( V(c) \) is affine, our main insights on the structure of the optimal customer segmentation and menu remain valid for a broader class of \( V(c) \) functions; we discuss this point in 3.6.2.

A second option is akin to a direct revelation mechanism and relies on the precise indifference for these customers. The provider may simply ask whether the customer is indifferent between balking and service level \( d \) and then select between these two options via an independent coin flip. This places greater computational burden on the provider.

Finally, while less prominent in the literature, it may be appropriate to require strictly positive utility for participation (i.e. \( a(c) = 1 \) if and only if \( U(c) > 0 \)). In which case, the option of receiving delay \( d \) may be replaced by an option of receiving delay \( d + \epsilon \). Then by appropriately adjusting prices for this and lower quality options an approximately equivalent
3.3.1 Simplified Problem Formulation

Based on Proposition 6 we drop the pair \((a, p)\) from the problem and write the revenue rate (3.6) as a function of the demand rates \(\lambda_l, \lambda_m, \lambda_h\), and the lead times \(w\). Substituting the price functions (3.14) and (3.16) yields the following revenue rates from low, medium and high lead time qualities:

\[
\Pi_l(\lambda_l, w) := \Lambda \int_{c_{\min}}^{c_l} p(x) f(x) dx = \Lambda \int_{c_{\min}}^{c_l} f(x) \cdot \{v - f_t(x) (w(x) - d)\} dx, \quad (3.27)
\]

\[
\Pi_m(\lambda_m) := \Lambda \int_{c_h}^{c_{\max}} a(x) p(x) f(x) dx = \lambda_m \cdot v = (\lambda - \lambda_l - \lambda_h) \cdot v, \quad (3.28)
\]

\[
\Pi_h(\lambda_h, w) := \Lambda \int_{c_h}^{c_{\max}} p(x) f(x) dx = \Lambda \int_{c_h}^{c_{\max}} f(x) \cdot \{v + f_h(x) (d - w(x))\} dx, \quad (3.29)
\]

where the functions \(f_t\) in (3.27) and \(f_h\) in (3.29) are defined in terms of \(f\) as follows:

\[
f_t(c) := c + \frac{F(c)}{f(c)} \text{ for } c \in [c_{\min}, c_l], \quad (3.30)
\]

\[
f_h(c) := c - \frac{F(c)}{f(c)} \text{ for } c \in [c_h, c_{\max}]. \quad (3.31)
\]

We refer to \(f_t\) and \(f_h\) as the virtual delay cost functions: they measure the sensitivity of the overall revenue to changes in the lead times of individual types. (In this respect they loosely play a similar role to the concept of virtual valuations in auction theory.) Since in our model the ranking of types’ willingness to pay for a given lead time \(w\) depends on whether \(w < d\) or \(w > d\), which of the virtual delay costs applies to a given type \(c\) is endogenous to how that type is served under a given menu: \(f_t(c)\) applies if the lead time of type \(c\) is of low quality \((w(c) > d)\) and \(f_h(c)\) if it is of high quality \((w(c) < d)\). The first summand in (3.30) and (3.31), the type’s delay cost parameter \(c\), simply measures how an increase in that type’s lead time \(w(c)\) reduces its own price \(p(c)\). The second terms capture the price reduction for types with lower (if \(c \in (c_{\min}, c_l]\)) or higher (if \(c \in [c_h, c_{\max})\)) time-sensitivity in response to an increase in \(w(c)\). While \(f_t(c)\) is strictly positive, \(f_h(c)\) may be negative.

Let \(\Pi(\lambda_l, \lambda_h, \lambda, w) \triangleq \Pi_l(\lambda_l, w) + \Pi_m(\lambda - \lambda_l - \lambda_h, w) + \Pi_h(\lambda_h, w)\) denote the expected revenue rate as a function of the total demand rate \(\lambda\), the segmentation characterized by \(\lambda_l\),
\( \lambda_h \) and \( \lambda_m = \lambda - \lambda_l - \lambda_h \geq 0 \), and the purchased high and low lead time qualities \( w \). It satisfies

\[
\Pi (\lambda_l, \lambda_h, \lambda, w) = \lambda v - \Lambda \int_{c_{\min}}^{c_l(\lambda_l)} f(x)f_l(x)(w(x) - d) \, dx - \Lambda \int_{c_{\min}}^{c_{\max}} f(x)f_h(x)(w(x) - d) \, dx,
\]

where \( c_l(\lambda_l) = F^{-1}(\lambda_l/\Lambda) \) and \( c_h(\lambda_h) = \overline{F}^{-1}(\lambda_h/\Lambda) \). For given segmentation \((\lambda_l, \lambda_m, \lambda_h)\) the revenue only depends on the system’s aggregate virtual delay cost. Problem 1 simplifies as follows.

**Problem 2.**

\[
\max_{\lambda_l, \lambda_h, \lambda, w} \Pi (\lambda_l, \lambda_h, \lambda, w) = \lambda v - \Lambda \int_{c_{\min}}^{c_l(\lambda_l)} f(x)f_l(x)(w(x) - d) \, dx - \Lambda \int_{c_{\min}}^{c_{\max}} f(x)f_h(x)(w(x) - d) \, dx
\]

\( s.t. \)

\[
\lambda_l, \lambda_h \geq 0, \quad (3.34)
\]

\[
\lambda_m = \lambda - (\lambda_l + \lambda_h) \geq 0, \quad (3.35)
\]

\[
\lambda \leq \Lambda, \quad (3.36)
\]

\[
\lambda < m u, \quad (3.37)
\]

\[
\Lambda \int_{c}^{c_{\max}} f(x)w(x)dx \geq \frac{\Lambda F(c)}{\mu - \lambda \Lambda F(c)}, \quad \forall c \in [c_h(\lambda_h), c_{\max}],
\]

\[
\Lambda \int_{x \in [c_l, c_h \cup c_{\max}]} f(x)w(x)dx + \lambda_m \cdot d \geq \frac{\lambda - \Lambda F(c)}{\mu - [\lambda - \Lambda F(c)]}, \quad \forall c \in [c_{\min}, c_l(\lambda_l)],
\]

\[
w(c) \geq \frac{1}{\mu}, \quad \forall c, \quad (3.40)
\]

\[
w(c) > d > w(c'), \quad \forall c \in [c_{\min}, c_l(\lambda_l)], c' \in (c_h(\lambda_h), c_{\max}],
\]

\[
w(c) \geq w(c'), \quad \forall c < c' \in [c_{\min}, c_l(\lambda_l)] \cup [c_h(\lambda_h), c_{\max}].
\]

The constraints (3.34)-(3.35) ensure a feasible segmentation; (3.37) system stability; (3.38)-(3.40) that the lead times are operationally achievable, and (3.41)-(3.42) that Property 1. of Proposition 6 holds.
3.3.2 Solution Preview and Contrast to Case With Observable Types

A simple but key fact to keep in mind: if the provider does know customer types, unlike in our setting, then the optimal lead times are obtained by a standard work conserving policy which strictly prioritizes customers in the order of their delay costs. The optimal lead times and scheduling policy in our setting may differ in two fundamental ways from this standard policy.

1. “Service Pooling” of multiple types into a single class. Instead of targeting a distinct service class to each type, it may be optimal to target a common class to multiple types with different delay costs – we call this service pooling. It may be optimal under either of the following conditions. If $f_l(c)$ is nonmonotone on a subset of $C_l$, the set of types targeted for low lead time quality, it may be optimal to serve some of these types with a common low lead time quality. Similarly, if $f_h(c)$ is nonmonotone on a subset of $C_h$, it may be optimal to serve multiple types with a common high lead time quality. Finally, if $f_l(c_1) > f_h(c_2)$ for some types $c_1 < c_2$ with $w(c_1) > d > w(c_2)$, it may be optimal to serve multiple types with the medium lead time quality.

2. Strategic delay. It may be optimal to deviate from a work conserving policy by artificially inflating some lead times, see Afèche (2004, 2010). Doing so is optimal if the best work conserving policy yields a negative virtual delay cost $f_h(c) < 0$ for some types $c$ with high lead time quality.

Sections 3.4 and 3.5 focus on the case of strictly increasing virtual delay cost functions $f_l$ and $f_h$. Section 3.6.1 discusses how the analysis and results generalize when this restriction is relaxed.
Chapter 3. Revenue Maximizing Lead-Time Pricing

3.4 Optimal Menus: Strictly Increasing Virtual Delay Cost Functions

The IC constraints require the delays $w(c)$ to be non-increasing in customer types $c$, whereas it is evident from (3.27) and (3.29) that maximizing revenues calls for prioritizing customers in the order of their virtual delay costs $f_l(c)$ or $f_h(c)$. If the virtual delays costs are not monotone in $c$ on the subset of types targeted for service, then designing the lead time menu is subject to a conflict between maintaining incentive-compatibility and increasing revenues. This conflict can be resolved by pooling multiple adjacent types into a single service class. In this Section we restrict attention to the case where the virtual delay cost functions $f_l$ and $f_h$ are each strictly increasing in customer type $c$. Under this assumption this conflict between revenue-maximization and incentive-compatibility occurs at a given arrival rate only if the following two conditions hold. (1) It is not optimal to offer only low ($C_l \neq \emptyset$) or only high ($C_h \neq \emptyset$) lead time qualities. (2) It is impossible to offer only low and high lead time qualities (but not the medium lead time, i.e., $C_m = \emptyset$) such that $f_l(c_l) \leq f_h(c_h)$; in other words, some type $c \in C_l$ targeted for a low lead time quality has a larger virtual delay cost than some type $c' \in C_h$ targeted for a high lead time quality: $f_l(c) > f_h(c')$ and $c < c'$. In this situation the revenue-maximizing lead time menu is designed such that multiple types are pooled into a single class with medium lead time quality; the result is to offer the medium lead time class ($C_m \neq \emptyset$), together with low lead time qualities ($C_l \neq \emptyset$), high lead time qualities ($C_h \neq \emptyset$) or both.

Pooling is optimal only if these two conditions hold at the optimal arrival rate. Section 3.4.1 characterizes the optimal customer segmentation and lead times, including the conditions under which pooling is optimal, for a fixed arrival rate. Section 3.4.2 characterizes how the optimal segmentation structure and lead time menu at the optimal arrival rate depend on the (fixed) capacity $\mu$. 
3.4.1 Optimal Customer Segmentation and Lead Times for Fixed Arrival Rate

We discuss STEP 2 of the solution approach outlined in Section 3.2.3. Let $\lambda^*_l(\lambda)$, $\lambda^*_m(\lambda)$ and $\lambda^*_h(\lambda)$ denote the optimal customer segmentation, and the function $w^*(c;\lambda)$ the optimal lead times as a function of the total arrival rate $\lambda$. Write $c^*_l(\lambda) \triangleq F^{-1}(\lambda^*_l(\lambda)/\Lambda)$ and $c^*_h(\lambda) \triangleq F^{-1}(\lambda^*_h(\lambda)/\Lambda)$ for the corresponding boundary types, where $\lambda^*_l(\lambda) > 0 \iff c^*_l(\lambda) > c_{\text{min}}$ and $\lambda^*_h(\lambda) > 0 \iff c^*_h(\lambda) < c_{\text{max}}$, and $C^*_l(\lambda)$, $C^*_m(\lambda)$ and $C^*_h(\lambda)$ for the corresponding sets of types buying low, medium and high lead time qualities, respectively. Lemma 3 establishes necessary conditions for the optimal lead times, scheduling policy and segmentation for fixed $\lambda$.

**Lemma 3.** [Optimal Customer Lead Times] Fix $\lambda \in (0, \Lambda] \cap (0, \mu)$. Assume strictly increasing virtual delay cost functions $f_l, f_h$.

1. The optimal lead time menu and corresponding scheduling policy have the following properties:

   (a) The lead times satisfy:

   $$w^*(c;\lambda) = \begin{cases} \frac{\mu}{(\mu - \Lambda F(c))^{2}} < d & \text{if } c \in C^*_h(\lambda) \neq \emptyset \\ d & \text{if } c \in C^*_m(\lambda) \neq \emptyset \\ \frac{\mu}{(\mu - [\lambda - \Lambda F(c))]^{2}} > d & \text{if } c \in C^*_l(\lambda) \neq \emptyset \end{cases} \quad (3.43)$$

   (b) Types with high lead time quality in $C^*_h(\lambda)$ receive strict priority over types in $C^*_m(\lambda)$ which receive strict priority over types in $C^*_l(\lambda)$. Different types within $C^*_l(\lambda)$ and $C^*_h(\lambda)$ buy different lead times and are strictly prioritized in the order of their delay cost. Different types within $C^*_m(\lambda)$ buy the same lead time and are pooled into a single FIFO service class.

   (c) If low lead time qualities are sold ($C_l \neq \emptyset$) then the optimal policy is work conserving.
2. Under the optimal customer segmentation, the virtual delay costs are positive and increasing over types with low or high lead time quality: (a) \( f_h(c_h^*(\lambda)) \geq 0 \). (b) If \( \lambda_t^*(\lambda) > 0 \) and \( \lambda_h^*(\lambda) > 0 \) then \( f_l(c_l^*(\lambda)) \leq f_h(c_h^*(\lambda)) \); if in addition \( \lambda_m^*(\lambda) > 0 \) then \( f_l(c_l^*(\lambda)) = f_h(c_h^*(\lambda)) \).

**Proof of Lemma 3**

Refer to Problem 2 and the revenue rate (3.32). Fix \( \lambda < \mu \). Let \( D(\lambda_t, \lambda_h, w) \) be the virtual delay cost rate as a function of \( \lambda_t, \lambda_h \) and the lead time function \( w \):

\[
D(\lambda_t, \lambda_h, w) \triangleq \Lambda \int_{c_{\min}}^{c_{\max}} f_l(x) f_l(x)(w(x) - d) \, dx + \Lambda \int_{c_h(\lambda_h)}^{c_{\max}} f_h(x) f_h(x)(w(x) - d) \, dx.
\]

For fixed \( \lambda \), maximizing revenue is equivalent to minimizing \( D(\lambda_t, \lambda_h, w) \) over \( \lambda_t, \lambda_h \) and \( w : C \to \mathbb{R} \), subject to the constraints (3.35)-(3.42). Recall that \( f_l, f_h \) satisfy (3.30)-(3.31), \( f_l' > 0, f_h' > 0 \), and that \( c_l(\lambda_t) = F^{-1}(\lambda_t/\Lambda) \) and \( c_h(\lambda_h) = \bar{F}^{-1}(\lambda_h/\Lambda) \) are the marginal types corresponding to \( \lambda_t \) and \( \lambda_h \), respectively. Suppose the scalars \( \lambda_t^*, \lambda_h^* \) and the function \( w^* \) are a solution of this problem and write \( c_l^* = c_l(\lambda_t^*) \) and \( c_h^* = c_h(\lambda_h^*) \). The proof hinges on three necessary optimality conditions.

(i) If \( c \in (c_h^*, c_{\max}] \) then \( f_h(c) > 0 \). We argue by contradiction. If \( f_h(c_0) = 0 \) for some \( c_0 \in (c_h^*, c_{\max}] \), where \( f_h(c_{\max}) = c_{\max} > 0 \), then \( f_h(c) < 0 \) for \( c \in [c_h^*, c_0] \) since \( f_h \) is strictly increasing, and \( w^*(c) < d \) for \( c \in (c_h^*, c_0] \). By inspection it is clear that we can reduce the virtual delay cost rate by perturbing the lead time function from \( w^* \) to \( w^o \), where \( w^o \) agrees with \( w^* \) except that \( w^o(c) = d \) for \( c \in [c_h^*, c_0] \). Then \( w^o \) is feasible and \( D(\lambda_t^*, \lambda_h^*, w^o) < D(\lambda_t^*, \lambda_h^*, w^*) \). Under the menu \( w^o \) the marginal type \( c_h^* \) moves to \( c_h^* = c_0 > c_h^* \) and \( f_h(c) > 0 \) holds for \( c \in (c_h^*, c_{\max}] \).

(ii) If the set of types with high lead time qualities is nonempty \( (C_h^* \neq \emptyset) \) then the constraints (3.38) are binding for \( c \in [c_h^*, c_{\max}] \), and \( w^*(c_{\max}) = 1/\mu \). This is trivial if \( c_h^* = c_{\max} \in C_h^* \). Suppose that \( c_h^* < c_{\max} \). If the property is not satisfied, there exists a feasible perturbation \( w^o \) of \( w^* \) which reduces the virtual delay cost rate \( D(\lambda_t, \lambda_h, w) \) by lowering the lead times for types \( (c_2, c_2 + \epsilon_2] \subset [c_h^*, c_{\max}] \) and by increasing the lead times
for lower types \((c_1, c_1 + \epsilon_1) \subset [c_h^*, c_{\max}]\), where \(\epsilon_1, \epsilon_2 > 0\) and \(c_1 + \epsilon_1 \leq c_2\). This holds since \(f_h(c) > 0\) for \(c > c_h^*\) by (i), and because \(f_h' > 0\).

\((iii)\) If the set \(C_l^*\) is nonempty then the constraints (3.39) are binding for \(c \in [c_{\min}, c_l^*]\), and \(w^*(c_{\min}) = \mu / (\mu - \lambda)^2\). This follows from a similar argument as for \((ii)\), because \(f_l > 0\) and \(f_l' > 0\).

Part 1(a). Properties \((ii)-(iii)\) imply (3.43). This is immediate for \(c_h^* = c_{\max} \in C_h^*\) and/or \(c_{\min} = c_l^* \in C_l^*\). If \(c_h^* < c_{\max}\) then by \((ii)\) the constraints (3.38) are binding for \(c \in [c_h^*, c_{\max}]\).

Solving the resulting integral equation in \(w^*\) yields \(w^*(c) = \mu / (\mu - \Lambda F(c))^2\). If \(c_l > c_{\min}\) the RHS of (3.39) satisfies

\[
\frac{\lambda - \Lambda F(c)}{\mu - [\lambda - \Lambda F(c)]} = \frac{[\lambda_l - \Lambda F(c)] \mu}{(\mu - [\lambda - \Lambda F(c)])(\mu - \lambda_m - \lambda_h)} + \frac{\lambda_m \mu}{(\mu - \lambda_m - \lambda_h)(\mu - \lambda_h)} + \frac{\lambda_h}{\mu - \lambda_h}, c \in [c_{\min}, c_l].
\]

By \((ii)-(iii)\) for an optimal solution the constraints (3.39) therefore simplify to

\[
\Lambda \int_c f(x) w^*(x) dx + \lambda^*_m d = \frac{\Lambda [F(c_l^*) - F(c)] \mu}{(\mu - [\lambda - \Lambda F(c)])(\mu - \lambda_m^* - \lambda_h^*)} + \frac{\lambda_m^* \mu}{(\mu - \lambda_m^* - \lambda_h^*)(\mu - \lambda_h^*)}, c \in [c_{\min}, c_l^*].
\]

Solving this integral equation in \(w^*\) yields \(w^*(c) = \mu / (\mu - [\lambda - \Lambda F(c)])^2\).

Parts 1(b)-(c). These claims follow directly from \((ii)-(iii)\). Since (3.38) is binding for \(c \in [c_h^*, c_{\max}]\) these types are strictly prioritized in order of their delay costs and receive strict priorities over all lower types. Since (3.39) is binding for \(c \in [c_{\min}, c_l^*]\) these types are strictly prioritized in order of their delay costs and receive strictly lower priority than all higher types. If \(C_l^* \neq \emptyset\) then (3.39) is binding for \(c = c_{\min}\), which implies that the policy is work conserving.

Part 2(a). Follows from Property (i) above since \(f_h\) is continuous.

Parts 2(b)-(c). Follow by substituting \(w^*(c)\) from (3.43) in the revenue function (3.32) and analyzing its partial derivatives with respect to \(\lambda_l\) and \(\lambda_h\). The details are in the proof of Proposition 7.

\[\blacksquare\]
Substituting the lead time function (3.43) in (3.32) yields the revenue rate as a function of $(\lambda_l, \lambda_h, \lambda)$:

\[
\Pi(\lambda_l, \lambda_h, \lambda) \triangleq \lambda v - \Lambda \int_{c_{\min}}^{c_{l}(\lambda_l)} f(x)f_l(x) \left( \frac{\mu}{(\mu - [\lambda - \Lambda F(x)])^2} - d \right) dx \\
+ \Lambda \int_{c_{h}(\lambda_h)}^{c_{\max}} f(x)f_h(x) \left( d - \frac{\mu}{(\mu - \Lambda F(x))^2} \right) dx. \tag{3.45}
\]

**Problem 2** reduces to the following program, where the constraints (3.49)-(3.50) correspond to (3.40)-(3.41).

**Problem 3.**

\[
\max_{\lambda_l, \lambda_h, \Lambda} \Pi(\lambda_l, \lambda_h, \lambda) \tag{3.46}
\]

\text{s.t.}

\[
\lambda_l, \lambda_h \geq 0, \tag{3.47}
\]

\[
\lambda_l + \lambda_h \leq \lambda \leq \Lambda, \tag{3.48}
\]

\[
\lambda < \mu, \tag{3.49}
\]

\[
\frac{\mu}{(\mu - [\lambda - \Lambda_l])(\mu - \lambda_h)} \leq d \text{ if } \lambda - \lambda_l > 0, \tag{3.49}
\]

\[
\frac{\mu}{(\mu - [\lambda - \Lambda_l])^2} \geq d \text{ if } \lambda_l > 0. \tag{3.50}
\]

**Proposition 7.** \textbf{[Optimal Allocation For Fixed Admission Rate]} Fix a capacity $\mu > 0$ and assume that $f'_l > 0$ and $f'_h > 0$. Let $h$ be shorthand for $\lambda_h^* (\lambda) > 0$, $m$ for $\lambda_m^* (\lambda) > 0$, and $l$ for $\lambda_l^* (\lambda) > 0$. The optimal customer segmentation and lead time menu depend as follows on the market size $\Lambda > 0$ and the arrival rate $\lambda \in (0, \Lambda] \cap (0, \mu)$.

1. For $d \leq \mu^{-1}$ the segmentation $(l)$ is optimal for all $\lambda$ and $\Lambda$: $\lambda_l^* (\lambda) > 0 = \lambda_m^* (\lambda) = \lambda_h^* (\lambda)$.

2. For $d > \mu^{-1}$ denote by $\lambda_F \triangleq \mu - \sqrt{\mu/d}$ and $\lambda_F \triangleq \mu - 1/d$ the arrival rates at which the maximum lead time equals $d$ under workconserving priority and FIFO service, respectively.
(a) If \( f_h(c_{\text{min}}) \geq 0 \) and \( \mu^{-1} \leq F(f_f^{-1}(c_{\text{max}})) \cdot d \) then there are unique thresholds \( \Lambda_1 < \Lambda_2 < \Lambda_3 < \Lambda_4 \) where \( \Lambda_1 = \lambda_P < \Lambda_2 < \lambda_F < \mu < \Lambda_4 \), which yield the following segmentation structure:

\[
\begin{array}{|c|c|}
\hline
\text{Market Size} & \text{Classes with positive rate as } \lambda \text{ increases on } [0, \Lambda] \cap (0, \mu) \\
\hline
\Lambda \in (0, \Lambda_1] & (h) \\
\hline
\Lambda \in (\Lambda_1, \Lambda_2] & (h) \rightarrow (h, m) \\
\hline
\Lambda \in (\Lambda_2, \Lambda_3) & (h) \rightarrow (h, m) \rightarrow (h, m, l) \\
\hline
\Lambda \in (\Lambda_3, \Lambda_4) & (h) \rightarrow (h, l) \rightarrow (h, m, l) \\
\hline
\Lambda \in [\Lambda_4, \infty) & (h) \rightarrow (h, l) \\
\hline
\end{array}
\]

(3.51)

Segmentations \((m, l)\) and \((h, m_{sd})\) are never optimal. The optimal policy is work conserving.

(b) It is optimal for some \((\lambda, \Lambda)\) to sell only medium and low quality classes \((m, l)\), if and only if \( F(f_f^{-1}(c_{\text{min}})) \cdot d < \mu^{-1} < d \). If \( f_h(c_{\text{min}}) \geq 0 \) and \( F(f_f^{-1}(c_{\text{max}})) \cdot d < \mu^{-1} < d \) then (3.51) is modified by additional thresholds \( \underline{\Lambda}_{ml} < \overline{\Lambda}_{ml} \) where \( \Lambda_1 < \Lambda_2 < \lambda_F < \underline{\Lambda}_{ml} < \mu < \overline{\Lambda}_{ml} < \Lambda_4 \).

\[
\begin{array}{|c|c|}
\hline
\text{Market Size} & \text{Classes with positive rate as } \lambda \text{ increases on } [0, \Lambda] \cap (0, \mu) \\
\hline
\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml}) & (h) \rightarrow (h, m) \rightarrow (h, m, l) \rightarrow (m, l), \text{ if } \Lambda < \Lambda_3 \\
& (h) \rightarrow (h, l) \rightarrow (h, m, l) \rightarrow (m, l), \text{ if } \Lambda > \Lambda_3 \\
\hline
\end{array}
\]

(3.52)

For \( \Lambda \notin [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml}) \) the structure of (3.51) applies.

(c) It is optimal for some \((\lambda, \Lambda)\) to use strategic delay if and only if \( f_h(c_{\text{min}}) < 0 \) and \( d > \mu^{-1} \). If \( f_h(c_{\text{min}}) < 0 \) and \( \mu^{-1} \leq F(f_f^{-1}(c_{\text{max}})) \cdot d \) then (3.51) changes in that two thresholds \( \underline{\Lambda}_{sd} < \overline{\Lambda}_{sd} \) replace \( \Lambda_1 \) and yield the following segmentation structure, where \( \lambda_P < \underline{\Lambda}_{sd} \leq \Lambda_2 \) and \( \underline{\Lambda}_{sd} < \overline{\Lambda}_{sd} \leq \Lambda_3 < \Lambda_4 \).
Figure 3.1: Illustration of Case 2(a) of Proposition 7. Optimal Customer Segmentation as a Function of $\lambda$ and $\Lambda$. Capacity $\mu = 3$, $d = 1$, uniform delay cost rate distribution with $c_{\min} = 1$ and $c_{\max} = 2$.

<table>
<thead>
<tr>
<th>Market Size</th>
<th>Classes with positive rate as $\lambda$ increases on $[0, \Lambda] \cap (0, \mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda \in (0, \overline{\Lambda}_{sd})$</td>
<td>$(h) \rightarrow (h, m_{sd})$</td>
</tr>
</tbody>
</table>
| $\Lambda \in [\Lambda_{sd}, \overline{\Lambda}_{sd})$ | $(h) \rightarrow (h, m_{sd}) \rightarrow (h, m)$, if $\Lambda \leq \Lambda_2$  
| | $(h) \rightarrow (h, m_{sd}) \rightarrow (h, m) \rightarrow (h, m, l)$, if $\Lambda > \Lambda_2$ |

(3.53) where $m_{sd}$ indicates that the lead time $d$ involves strategic delay.

For $\Lambda \geq \overline{\Lambda}_{sd}$ the optimal scheduling policy is work conserving and (3.51) applies.

Due to the space required to prove this proposition, we relegate the proof of Proposition 7 to Appendix 3.A. This proposition introduces several points worthy of discussion.

1. Magnitude of capacity $\mu$ relative to $d$. If capacity is relatively tight in the sense that $\mu^{-1} \geq d$, as in Part 1, then the provider can only sell low lead time qualities ($w > d$). Since $f_l$ is strictly increasing and no high lead time qualities can be offered, the optimal segmentation structure is invariant to $\lambda$ and $\Lambda$. Pooling is not optimal; it is optimal to differentiate the low lead time qualities through a strict priority discipline. If capacity
Chapter 3. Revenue Maximizing Lead-Time Pricing

is relatively ample in the sense that \( \mu^{-1} < d \), as in Part 2, then the provider can sell high lead time qualities \( (w < d) \) to some fraction of the market. Since it is optimal to differentiate these lead times through a strict priority discipline, the maximum rate of types to which these classes can be sold is \( \lambda_P \), the rate at which the largest such lead time equals \( d \). If \( \mu^{-1} < d \) then it is always optimal to offer some high lead time qualities for some \( \lambda \) and \( \Lambda \); the remaining features of the optimal segmentation are, however, highly sensitive to \( \lambda \) and \( \Lambda \).

2. Main structure of optimal segmentation if \( \mu^{-1} < d \): Part 2(a). Refer to Figure 3.1 which illustrates the optimal segmentation specified in Table (3.51). For this example with \( \mu = 2 \) the market size thresholds are \( \Lambda_1 = \lambda_P = 1.27 \), \( \Lambda_2 = 1.63 \), \( \Lambda_3 = 2.54 \) and \( \Lambda_4 = 6 \). For \( \lambda \leq \min(\lambda_P, \Lambda) \) it is optimal to only offer high lead time qualities so \( \lambda_h^* (\lambda) = \lambda > 0 = \lambda_m^* (\lambda) = \lambda_m^* (\lambda) \); this segmentation is denoted by \( (h) \). For \( \lambda \in (\lambda_P, \Lambda] \) it is optimal to sell, in addition, either the medium lead time, low lead time qualities, or both, corresponding to \( (h, m) \), \( (h, l) \) and \( (h, m, l) \), respectively. Under \( (h, m) \) and \( (h, m, l) \), different types are pooled into the same medium lead time class, with some rate \( \lambda_m^* (\lambda) > 0 \), and the rate of types buying high lead time qualities is strictly lower than in the absence of such pooling, i.e., \( \lambda_h^* (\lambda) < \lambda_P \). The market size \( \Lambda \) and the arrival rate \( \lambda \) play the following role in determining which of the segmentations, \( (h, m) \), \( (h, l) \) or \( (h, m, l) \), is optimal for \( \Lambda > \lambda > \lambda_P \). First consider the impact of \( \Lambda \) for fixed \( \lambda \). An increase in \( \Lambda \) increases the mass \( \Lambda f(c) \) of every type. Therefore, for given \( \lambda \) the maximum range of types that can be offered high lead time qualities becomes smaller as \( \Lambda \) increases: mathematically, if \( c_h \) is the (lowest) boundary type of segment \( C_h \) when its rate \( \lambda_h = \lambda_P \) then \( \Lambda F(c_h) = \lambda_P \), so \( c_h \) increases and \( c_{\text{max}} - d c_h \) decreases in \( \Lambda \) for given \( \mu \). Similarly, for given \( \lambda \) the maximum range of types that can be offered low lead time qualities also becomes smaller as \( \Lambda \) increases: if \( c_l \) is the (largest) boundary type of segment \( C_l \) when its rate \( \lambda_l = \lambda - \lambda_P \) then \( \Lambda F(c_l) = \lambda - \lambda_P \), so \( c_l - c_{\text{min}} \) decreases in \( \Lambda \) for given \( \mu \) and \( \lambda \geq \lambda_P \). In
summary as $\Lambda$ increases for fixed $\lambda > \lambda_P$, the segments $C_l$ and $C_m$ each become more homogeneous while the minimum difference between their respective types, $\bar{c}_l - \bar{c}_t$, increases. As a result the virtual delay cost difference $f_h(c_h) - f_l(c_l)$ increases in $\Lambda$ for given $\mu$ and $\lambda > \lambda_P$, and pooling is not optimal ($\lambda_m^*(\lambda) = 0$) for large values of $\Lambda$.

In Figure 3.1 fix $\lambda > \lambda_P = 1.27$ and observe that for sufficiently large $\Lambda$ the optimal segmentation is $(h, l)$. For $\Lambda > \Lambda_4 = 6 > \mu$, the optimal segmentation is $(h, l)$ for all $\lambda \in (\lambda_P, \mu)$. For $\Lambda < \Lambda_4$ pooling is always optimal for large enough $\lambda$. For $\Lambda < \Lambda_3$, pooling is optimal for all $\lambda > \lambda_P$, with $(h, m)$ optimal for smaller $\lambda$, as long as capacity is sufficient to sell all customers medium or high lead time quality, and $(h, m, l)$ optimal for larger $\lambda$.

3. **Optimality of strategic delay if $\mu^{-1} < d$: Part 2(b).** If $f_h(c_{\text{min}}) < 0$ then an interval of types $[c_{\text{min}}, c_0)$, where $c_0 < c_{\text{max}}$, have a negative virtual delay cost $f_h(c)$, if targeted for high lead time quality $(w < d)$; reducing their lead time therefore lowers the overall revenue. These types should always be targeted for a lead time $\geq d$, even if serving them with lead time $< d$ is operationally feasible. If $\lambda$ is small relative to the capacity $\mu$ but large relative to the market size $\Lambda$, then segmentation $(h, m_{\text{sd}})$ is optimal: the highest types in $[c_0, c_{\text{max}}]$ are served with high lead time quality, while types in a subset of $[c_{\text{min}}, c_0]$ are pooled into the medium quality class and the lead time $d$ is attained by inserting strategic delay. This segmentation is optimal for all sufficiently large $\lambda$ if $\Lambda \in (0, \Lambda_{sd})$ but only for intermediate $\lambda$ if $\Lambda \in [\Lambda_{sd}, \Lambda_{sd})$. If $\Lambda \geq \Lambda_{sd}$, strategic delay is not optimal for any $\lambda$: the maximum possible arrival rate $\lambda_P$ to the high quality segment $C_h$ can be ‘sold out’ to types in $[c_0, c_{\text{max}}]$ with non-negative virtual delay cost $f_h(c) \geq 0$.

4. **Selling only medium and low lead time qualities if $\mu^{-1} < d$: Part 2(c).** If $F\left(f_l^{-1}(c_{\text{max}})\right) \cdot d < \mu^{-1} < d$ then the capacity is large enough such that high lead time qualities can be offered, but only to a small set of types. For $\Lambda \in (\Lambda_{ml}, \Lambda_{ml})$, the maximum type range of segment $C_h$, measured by $c_{\text{max}} - c_h$, is small relative to the maximum type
range of the low lead time quality segment $C_l$. For sufficiently large $\lambda$ it is therefore optimal \textit{not} to offer high lead time qualities; instead offer the medium lead time to more customers.

The following lemma determines concavity properties of the optimal profit function. These properties are useful in showing that the optimal arrival rate of admitted customers increases as capacity increases. With the segmentation properties described in the previous proposition this lemma leads to the characterization of the optimal menu as the provider’s capacity increases described in the following section.

\textbf{Lemma 4. [Properties of Optimal Revenue Function]} Fix the market size $\Lambda$. Let $\lambda_l^*(\lambda, \mu)$ and $\lambda_h^*(\lambda, \mu)$ denote the optimal arrival rates to $l$ and $h$ classes, respectively, as a function of $\lambda$ and $\mu$, as specified in Proposition 7; the argument $\mu$ makes the possible dependence on the service rate explicit. Write $\Pi(\lambda, \mu) \triangleq \Pi(\lambda_l^*(\lambda, \mu), \lambda_h^*(\lambda, \mu), \lambda, \mu)$ for the optimal revenue as a function of $\lambda$ and $\mu$.

1. $\Pi(\lambda, \mu)$ satisfies: $\Pi_\lambda(\lambda, \mu) \geq v > 0$ for every $(\lambda, \mu)$ where the optimal segmentation is (h) or (h, msd); $\Pi_{\lambda\lambda}(\lambda, \mu) \leq 0 \leq \Pi_{\lambda\mu}(\lambda, \mu)$ for every $(\lambda, \mu)$; and $\Pi_{\lambda\lambda}(\lambda, \mu) < 0 < \Pi_{\lambda\mu}(\lambda, \mu)$ for every $(\lambda, \mu)$ where the optimal segmentation is (l), (h, m), (h, l), (h, m, l) or (m, l).

2. $\Pi_\lambda(\lambda, \mu)$ is continuous in $(\lambda, \mu)$. The partial derivatives $\Pi_{\lambda\lambda}(\lambda, \mu)$, $\Pi_{\lambda\mu}(\lambda, \mu)$ are continuous in $(\lambda, \mu)$ under each optimal segmentation; they are also continuous at every $(\lambda, \mu)$ where a transition takes place between two of the optimal segmentations (l), (h, m), (h, l), (h, m, l), (m, l).
Proof of Lemma 4

The maximum revenue satisfies \( \Pi^\ast(\lambda, \mu) = \Pi(\lambda, \lambda^\ast_{mh}, \lambda^\ast_h, \mu) \) where

\[
\Pi(\lambda, \lambda^\ast_{mh}, \lambda^\ast_h, \mu) \triangleq \lambda v - \Lambda \int_{c_{\text{min}}}^{c_{\text{max}}} f(x) f_l(x) \left( \frac{\mu}{(\mu - [\lambda - \Lambda F(x)])^2} - d \right) dx + \Lambda \int_{c_h(\lambda^\ast_h)}^\infty f(x) f_h(x) \left( d - \frac{\mu}{(\mu - \Lambda F(x))^2} \right) dx,
\]

and \( \lambda^\ast_{mh}, \lambda^\ast_h \) depend on \( (\lambda, \mu) \) as tabulated in (3.54). Its entries for \( \lambda, \lambda^\ast_{mh}, \lambda^\ast_h \) and \( \mu \) are from the proof of Proposition 7; refer in particular to (3.84) and (3.88) and their discussion. They directly imply the partial derivatives of \( \lambda^\ast_{mh} \) and \( \lambda^\ast_h \), except for \( (h, m, l) \) where we derive them below. The proof refers to these properties of \( \lambda^\ast_{mh} \) and \( \lambda^\ast_h \) and derives those of \( \Pi \) below. We first review some important facts. Recall from Proposition 7 and its proof:

- \( \lambda_P \triangleq \mu - \sqrt{\mu/d} \) and \( \lambda_F \triangleq \mu - 1/d \) are well defined if \( \mu^{-1} < d \), so \( \lambda_P < \lambda_F \); \( \lambda^\ast_{mh} \) is defined in (3.77); and \( \lambda_0 \) is defined in (3.82).
Chapter 3. Revenue Maximizing Lead-Time Pricing

92

\[ \Pi_\lambda = \frac{\partial \Pi}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda_{mh}} \frac{\partial \lambda_{mh}}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda_{h}} \frac{\partial \lambda_{h}}{\partial \lambda}, \]

where

\[ \frac{\partial \Pi}{\partial \lambda} = v - \Lambda \int_{c_{\text{min}}}^{c_1} \frac{f(x)f_i(x)2\mu}{(\mu + \lambda + \Lambda F(x))^3} dx - \frac{\partial \Pi}{\partial \lambda_{mh}} \frac{\partial \lambda_{mh}}{\partial \lambda} \leq 0, \text{ and } \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} > 0. \]

(3.56)

\[ \frac{\partial \Pi}{\partial \lambda_{mh}} = f_i(c_i(\lambda - \lambda_{mh}))(\frac{\mu}{\mu - \lambda_{mh}} - d) \geq 0, \quad \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}} \geq 0, \text{ and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda_{mh}} < 0. \]

(3.57)

\[ \frac{\partial \Pi}{\partial \lambda_{h}} = f_i(c_h(\lambda))\left(d - \frac{\mu}{\mu - \lambda_{h}} \right) > 0, \quad \frac{\partial^2 \Pi}{\partial \lambda_{h}} \leq 0, \text{ and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda_{h}} \geq 0. \]

(3.58)

Part 1(i). It follows from (3.54) and (3.55)-(3.58) that \( \Pi_\lambda = v \) for \((h, m_{sd})\) and \( \Pi_\lambda \geq v \) for \((h)\).

Part 1(ii). We first show \( \Pi_{\lambda\lambda} < 0 < \Pi_{\lambda\mu} \) for \((h, m, l)\). Table (3.54) claims \( 0 < \partial \lambda_{mh}/\partial \lambda, \partial \lambda_{mh}/\partial \mu < 1 \), which is implied by equations \((i) - (ii)\) in the table and the fact that \( f_i^\prime c_i^\prime > 0 > f_i^\prime c_h^\prime \):

\[ \text{sign}\left(1 - \frac{\partial \lambda_{mh}}{\partial \lambda}\right) = -\text{sign}\left(\frac{\partial \lambda_{mh}}{\partial \lambda}\right) = \text{sign}\left(\frac{\partial \lambda_{h}}{\partial \mu}\right), \text{ so } 0 < \frac{\partial \lambda_{mh}}{\partial \lambda} < 1. \]

(3.59)

\[ \text{sign}\left(\frac{\partial \lambda_{mh}}{\partial \mu}\right) = \text{sign}\left(\frac{\partial \lambda_{h}}{\partial \mu}\right) \text{ and } 1 - \frac{\partial \lambda_{mh}}{\partial \mu} + \frac{\partial \lambda_{h}}{\partial \mu} = \frac{1}{\mu}, \text{ so } 0 < \frac{\partial \lambda_{mh}}{\partial \mu} < 1. \]

(3.60)

The first equations in (3.59)-(3.60) each follow from \((ii)\) and \( f_i^\prime c_i^\prime > 0 > f_i^\prime c_h^\prime \), the second equations each follow from \((i)\). For fixed \( \lambda \) the total derivative \( d\Pi/d\lambda_{mh} = 0 \) at \( \lambda_{mh} = \lambda_{mh}^* \); see (3.6) in proof of Proposition 7 where \( g(\lambda, \lambda_{mh}^*) = 0 \) for \((h, m, l)\). It follows that

\[ \Pi_\lambda = \frac{\partial \Pi}{\partial \lambda} + \frac{\partial \lambda_{mh}}{\partial \lambda} \left[ \frac{\partial \Pi}{\partial \lambda_{mh}} + \frac{\partial \Pi}{\partial \lambda_{h}} \frac{\partial \lambda_{h}}{\partial \lambda} \right] = \frac{\partial \Pi}{\partial \lambda} \text{ for all } (\lambda, \mu) \text{ with } (h, m, l). \]

(3.61)

We show that the following holds:

\[ \Pi_{\lambda\lambda} = \frac{\partial^2 \Pi}{\partial \lambda^2} + \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}} - \frac{f_i(c_i(\lambda - \lambda_{mh}))2\mu}{(\mu - \lambda_{mh})^3} - \frac{\partial^2 \Pi}{\partial \lambda_{mh} \partial \lambda} \left(1 - \frac{\partial \lambda_{mh}}{\partial \lambda}\right) < 0, \]

(3.62)

\[ \Pi_{\lambda\mu} = \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} + \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}} \frac{\partial \lambda_{mh}}{\partial \mu} > 0. \]

(3.63)

The equations for \( \Pi_{\lambda\lambda}, \Pi_{\lambda\mu} \) hold by (3.61) and \( \partial^2 \Pi/\partial \lambda \partial \lambda_{h} \) = 0 by (3.58). The first inequality in (3.62) holds by (3.56)-(3.57); the second by (3.57) and (3.59). The inequality in (3.63) holds by (3.56)-(3.57) and (3.60).

We next show \( \Pi_{\lambda\lambda} \leq 0 \leq \Pi_{\lambda\mu} \) for any segmentation other than \((h, m, l)\).

First consider \( \Pi_{\lambda\lambda}^* \). Since \( \partial^2 \Pi/\partial \lambda \partial \lambda_{h} \) = 0 by (3.58) and \( \partial^2 \lambda_{mh}/\partial \lambda^2 = 0 \) by (3.54) we have
\[
\Pi^*_{\lambda \lambda} = \frac{2}{\partial \lambda^2} \frac{\partial^2 \Pi}{\partial \lambda^2} + \frac{2}{\partial \lambda \partial \lambda^*_{\text{mh}}} \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} + \frac{2}{\partial \lambda^2} \left( \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right)^2 + \frac{2}{\partial \lambda^2} \left( \frac{\partial \lambda^*_{\text{hl}}}{\partial \lambda} \right)^2 + \frac{\partial \Pi}{\partial \lambda^2} \left( \frac{\partial \lambda^*_{\text{hl}}}{\partial \lambda} \right)^2.
\]

The terms in brackets are nonpositive by (3.58) and \(\partial^2 \lambda^*_h / \partial \lambda^2 \leq 0\) from (3.54). The other terms satisfy
\[
\frac{\partial^2 \Pi}{\partial \lambda^2} + \frac{2}{\partial \lambda \partial \lambda^*_{\text{mh}}} \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} + \frac{2}{\partial \lambda^2} \left( \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right)^2 \leq \left( \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right)^2 - \left( 1 - \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right)^2 \frac{\partial^2 \Pi}{\partial \lambda^2}
\]
by (3.56)-(3.57). The RHS is nonpositive since \(\partial^2 \Pi / \partial \lambda^2 \partial \lambda \geq 0\) by (3.57) and \(\partial \lambda^*_{\text{mh}} / \partial \lambda \leq 1\) by (3.54).

Next consider \(\Pi^*_{\lambda \mu}\). By (3.54) we have \((\partial \lambda^*_{\text{mh}} / \partial \lambda)(\partial \lambda^*_{\text{mh}} / \partial \mu) = \partial^2 \lambda^*_{\text{mh}} / (\partial \lambda \partial \mu) = 0\) which implies
\[
\Pi^*_{\lambda \mu} = \left( \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} + \frac{\partial^2 \Pi}{\partial \lambda^*_{\text{mh}} \partial \mu} \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right) + \left( \frac{\partial \Pi}{\partial \lambda \partial \lambda^*_{\text{mh}}} \frac{\partial \lambda^*_{\text{mh}}}{\partial \mu} \right) + \left( \frac{d}{\partial \mu} \left[ \frac{\partial \Pi}{\partial \lambda^*_{\text{hl}}} \frac{\partial \lambda^*_{\text{hl}}}{\partial \lambda} \right] \right).
\]
(3.64)

We show that each bracket is nonnegative. For the first, (3.54) and (3.56)-(3.57) imply
\[
\frac{\partial^2 \Pi}{\partial \lambda \partial \mu} + \frac{\partial^2 \Pi}{\partial \lambda^*_{\text{mh}} \partial \mu} \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \geq \frac{\partial^2 \Pi}{\partial \lambda^*_{\text{mh}} \partial \mu} \left( 1 - \frac{\partial \lambda^*_{\text{mh}}}{\partial \lambda} \right) \geq 0.
\]
The second bracket of (3.64) is nonnegative by (3.54) and (3.57). The third bracket of (3.64) is also nonnegative. For segmentations other than \((h, m)\) and \((h, m, l)\) we have \(\partial \lambda^*_h / \partial \lambda = 0\) or \(= 1\) by (3.54) and \(\partial^2 \Pi / (\partial \lambda^*_h \partial \mu) \geq 0\) by (3.58). For \((h, m)\) substitute for \(\lambda^*_h = \mu - \mu / d (\mu - \lambda)^{-1}\) from (3.54) to get:
\[
\frac{\partial \Pi}{\partial \lambda^*_{\text{hl}}} \frac{\partial \lambda^*_{\text{hl}}}{\partial \lambda} = f_h(c_h(\lambda^*_h)) \left( \frac{d - \frac{\mu}{(\mu - \lambda)^2}}{d - \frac{\mu}{(\mu - \lambda)^2}} \right) \frac{\partial \lambda^*_h}{\partial \lambda} = v + f_h(c_h(\lambda^*_h)) \left( \frac{d - \frac{\mu}{(\mu - \lambda)^2}} {d - \frac{\mu}{(\mu - \lambda)^2}} \right).
\]
(3.65)

This expression increases in \(\mu\): the virtual delay cost \(f_h(c_h(\lambda^*_h)) \geq 0\) decreases in \(\mu\) since \(f_h'(c_h(\lambda^*_h)) < 0\) and \(\partial \lambda^*_h / \partial \mu > 0\) by (3.54), and its multiplier is negative \((\lambda > \lambda_p)\) and increases in \(\mu\).

We omit the remaining straightforward checks that \(\Pi^*_{\lambda \lambda} < \Pi^*_{\lambda \mu}\) for \((l), (h, m), (h, l)\), and \((m, l)\).

**Part 2.** The function \(\Pi^* (\lambda, \mu) = \Pi(\lambda, \lambda^*_{\text{mh}}, \lambda^*_h, \mu)\) is defined piecewise: the functions \(\lambda^*_{\text{mh}} (\lambda, \mu)\) and \(\lambda^*_h (\lambda, \mu)\) depend on the optimal segmentations which vary with \((\lambda, \mu)\) as
specified by Proposition 7. By the definition of $\Pi$ and table (3.54) all first and second order derivates of $\Pi$, $\lambda^*_{mh}$ and $\lambda^*_h$ with respect to $(\lambda, \mu)$ are continuous for each segmentation. Therefore so are the functions $\Pi^*_\lambda(\lambda, \mu)$, $\Pi^*_{\lambda\lambda}(\lambda, \mu)$ and $\Pi^*_{\lambda\mu}(\lambda, \mu)$. It remains to show the stated properties at each $(\lambda, \mu)$ with a transition between two optimal segmentations. By Proposition 7, the possible transitions are as follows.

<table>
<thead>
<tr>
<th>Transitions in optimal segmentation involving $(h)$ or $(h, m_{sd})$</th>
<th>from</th>
<th>to</th>
<th>at $(\lambda, \mu)$</th>
<th>$\lambda^*_{mh}$</th>
<th>virtual delay cost</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(h)$</td>
<td>$(h, m_{sd})$</td>
<td>$(h, m_{sd})$</td>
<td>$\lambda = \lambda_0, \mu &gt; d^{-1}$</td>
<td>$\lambda$</td>
<td>$f_h(c_h(\lambda_0)) = 0$</td>
<td>9.1</td>
</tr>
<tr>
<td>$(h, m_{sd})$</td>
<td>$(h, m)$</td>
<td>$(h, m)$</td>
<td>$\lambda = \lambda_{sd}, \mu &gt; d^{-1}$</td>
<td>$\lambda$</td>
<td>$f_h(c_h(\lambda_{sd})) = 0$</td>
<td>9.2</td>
</tr>
<tr>
<td>$(h)$</td>
<td>$(h, l)$</td>
<td>$(h, l)$</td>
<td>$\lambda = \lambda_P, \mu &gt; d^{-1}$</td>
<td>$\lambda$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(h)$</td>
<td>$(h, m)$</td>
<td>$(h, m)$</td>
<td>$\lambda = \lambda_P, \mu &gt; d^{-1}$</td>
<td>$\lambda$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transitions in optimal segmentation involving neither $(h)$ nor $(h, m_{sd})$</th>
<th>from</th>
<th>to</th>
<th>at $(\lambda, \mu)$</th>
<th>$\lambda^*_{mh}$</th>
<th>virtual delay cost</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(l)$</td>
<td>$(l)$</td>
<td>$(l)$</td>
<td>$\lambda &lt; \mu = d^{-1}$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(l)$</td>
<td>$(m, l)$</td>
<td>$(m, l)$</td>
<td>$\lambda &lt; \mu = d^{-1}$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(h, m)$</td>
<td>$(h, m, l)$</td>
<td>$(h, m, l)$</td>
<td>$\lambda = \lambda_1, \mu &gt; d^{-1}$</td>
<td>$\lambda$</td>
<td>$f_l(c_l(\lambda_1)) = f_l(c_l(\lambda - \lambda^*<em>{mh})) = c</em>{\min}$</td>
<td>6.1(b)</td>
</tr>
<tr>
<td>$(h, m, l)$</td>
<td>$(m, l)$</td>
<td>$(m, l)$</td>
<td>$\lambda = \lambda_3, \mu &gt; d^{-1}$</td>
<td>$\lambda_F$</td>
<td>$f_l(c_l(\lambda_3 - \lambda_F)) = c_{\max}$</td>
<td>6.2(a)</td>
</tr>
<tr>
<td>$(h, l)$</td>
<td>$(h, m, l)$</td>
<td>$(h, m, l)$</td>
<td>$\lambda = \lambda_2, \mu &gt; d^{-1}$</td>
<td>$\lambda_P$</td>
<td>$f_h(c_h(\lambda_P)) = f_l(c_l(\lambda_2 - \lambda_P))$</td>
<td>7.2</td>
</tr>
</tbody>
</table>

The following facts establish $(i)$ and $(ii)$. At $(\lambda, \mu)$ where the segmentation transitions from $(h)$ or $(h, m_{sd})$ we have $\Pi^*_{\lambda} = v$. At $(\lambda, \mu)$ with a transition involving neither $(h)$ nor $(h, m_{sd})$ the two expressions for $\Pi^*_{\lambda\lambda}$ (one for each segmentations) agree, and ditto for $\Pi^*_{\lambda\mu}$. We omit these checks; they are straightforward using table (3.54) and the formulae for $\Pi^*_\lambda$, $\Pi^*_{\lambda\lambda}$ and $\Pi^*_{\lambda\mu}$ derived above. ■
3.4.2 Optimal Arrival Rate, Customer Segmentation and Lead Times

We turn to STEP 3 of the solution approach outlined in Section 3.2.3 and discuss how the optimal segmentation and lead time menu at the optimal arrival rate depends (i) on the demand parameter $v$ for given capacity $\mu$, and (ii) on the capacity for given demand parameters.

**Sensitivity to demand parameter $v$.** The market size thresholds of Proposition 7 are analytically defined, some in closed form. (See the proof; we omit these details from the statement of the Proposition in order to highlight the structural properties of the solution.) These thresholds only depend on $d$, $\mu$ and $f$; they do not depend on $v$ which can be seen as a scale parameter of customer valuations since $V(c) = v + c \cdot d$ for $c \in [c_{\min}, c_{\max}]$. Based on this fact and the observation that the revenue rate (3.45) depends on $v$ only through the linear term $\lambda v$, it is straightforward to infer from Proposition 7 how the optimal segmentation at the optimal arrival rate depends on the capacity and the demand parameters. For example, suppose that $\mu$, $d$, and $f$ satisfy the conditions of Part 2(a) and that $\Lambda \in (\Lambda_2, \Lambda_3)$. For $v = 0$ the optimal arrival rate, call it $\lambda^*$, equals $\lambda_P$ and the segmentation is $(h)$. Since $\lambda^*(\mu)$ increases in $v$ with $\lambda^*(\mu) \to \min(\mu, \Lambda)$ as $v \to \infty$, there is a threshold $v_1 > 0$ such that the optimal segmentation is $(h, m)$ for $v \in (0, v_1]$ and $(h, m, l)$ for $v > v_1$.

Proposition 7 implies certain additional properties of the customer segmentation. In the example just discussed, the rate of types who buy high quality lead times equals $\lambda_P(\mu)$ for $v = 0$ and decreases in $v$. Therefore the larger customer valuations ($v$) the smaller the optimal size of the high lead time quality segment and the larger the time-sensitivity (types) of customers in that segment.

**Sensitivity to capacity $\mu$.** We have the following Proposition.

**Proposition 8.** [Optimal Allocation and Provider Capacity] The optimal cus-
tomer segmentation and lead times depend as follows on the capacity. Define the thresholds

\[ \mu_{\text{min}} := \frac{1}{d + v/c_{\text{min}}} < \frac{1}{d} < \mu_H := \Lambda + \frac{1 + \sqrt{1 + 4d \Lambda}}{2d}. \]  

(3.66)

If \( f_h(c_{\text{min}}) < 0 \) then let \( \mu_{SD} \) be the unique solution in \( \mu \in (\Lambda + d^{-1}, \mu_H) \) of the equation

\[ \mu - \frac{\mu/d}{\mu - \Lambda} = \Lambda f^{-1}_h(0). \]  

(3.67)

Let \( \lambda^*(\mu) \) be the optimal arrival rate as a function of the capacity \( \mu \).

1. Market coverage threshold. For \( \mu \leq \mu_{\text{min}} \) it is not profitable to operate: \( \lambda^*(\mu) = 0 \).

The optimal arrival rate \( \lambda^*(\mu) \) is strictly increasing on \( [\mu_{\text{min}}, \mu_A] \) where \( \mu_A > \mu_{\text{min}} \) is a unique threshold such that \( \lambda^*(\mu) = \Lambda \Leftrightarrow \mu \geq \mu_A \). If \( f_h(c_{\text{min}}) \geq 0 \) then \( \mu_A < \mu_H \), and if \( f_h(c_{\text{min}}) < 0 \) then \( \mu_A < \mu_{SD} < \mu_H \).

2. Pooling threshold. There is a unique \( \mu_P \) such that it is optimal to pool different types into the medium lead time class as follows: (a) if \( f_h(c_{\text{min}}) \geq 0 \) then \( \mu_P \in [d^{-1}, \mu_H] \) and pooling is optimal if and only if \( \mu \in (\mu_P, \mu_H) \). (b) if \( f_h(c_{\text{min}}) < 0 \) then \( \mu_P \in [d^{-1}, \mu_{SD}] \) and pooling is optimal if and only if \( \mu > \mu_P \). (c) The capacity thresholds for market coverage and pooling are ranked as follows:

\[ \mu_A \leq d^{-1} = \mu_P, \quad \text{if} \quad \int_{c_{\text{min}}}^{c_{\text{max}}} f(x) f_l(x) \frac{2d^2 \Lambda}{(1 - \Lambda d F(x))^3} dx. \]  

(3.68)

\[ d^{-1} \leq \mu_P < \mu_A, \quad \text{otherwise.} \]  

(3.69)

3. Maximum differentiation through strict priorities. It is optimal to sell classes (l) iff \( \mu \in (\mu_{\text{min}}, 1/d) \), classes (h,l) iff \( \mu \in (1/d, \mu_P) \), and classes (h) iff \( \mu \geq \mu_H \) and \( f_h(c_{\text{min}}) \geq 0 \). This involves neither pooling nor strategic delay.

4. Pooling without strategic delay. It is optimal to sell classes (h,m), (h,m,l) or (m,l) iff: \( \mu \in (\mu_P, \mu_H) \) and \( f_h(c_{\text{min}}) \geq 0 \), or \( \mu \in (\mu_P, \mu_{SD}) \) and \( f_h(c_{\text{min}}) < 0 \). This involves pooling without strategic delay for the medium class and strict priorities for all others.

5. Pooling with strategic delay. It is optimal to sell classes (h,m_{sd}) iff \( \mu > \mu_{SD} \) and \( f_h(c_{\text{min}}) < 0 \). The medium lead time pools multiple types and requires strategic delay.
Proposition 8 establishes two differences between cases with \( f_h(c_{\text{min}}) \geq 0 \) and \( f_h(c_{\text{min}}) < 0 \).

1. If \( f_h(c_{\text{min}}) \geq 0 \) then pooling types (for the medium lead time) is optimal only at intermediate capacity levels, if and only if \( \mu \in (\mu_P, \mu_H) \). By contrast, if \( f_h(c_{\text{min}}) < 0 \) then pooling types is optimal for all capacity levels above a threshold, if and only if \( \mu > \mu_P \).

2. If \( f_h(c_{\text{min}}) \geq 0 \) then all lead times are achieved through a work conserving scheduling policy. By contrast, if \( f_h(c_{\text{min}}) < 0 \) then the policy is work conserving if and only if \( \mu \leq \mu_{SD} \).

In Section 3.5 we discuss numerical examples that illustrate these differences. We now prove the proposition

**Proof of Proposition 8**

Let \( \lambda^*(\mu) = \arg \{ \max_{\lambda} \Pi^*(\lambda, \mu) \text{ s.t. } \lambda \in [0, \Lambda] \cap [0, \mu] \} \). We first prove key properties of the thresholds \( \mu_H \), defined by (3.66), and \( \mu_{SD} \), defined by (3.67).

**p1.** Suppose that \( f_h(c_{\text{min}}) \geq 0 \). Then (i) the optimal rate \( \lambda^*(\mu) = \Lambda \) for \( \mu \geq \mu_H \), and (ii) the optimal segmentation is \((h)\) if and only if \( \mu \geq \mu_H \). Recall that \( \lambda_P = \mu - \sqrt{\mu/d} \) for \( \mu > d^{-1} \) as defined in Proposition 7. We write \( \lambda_P(\mu) \) to make its dependence on \( \mu \) explicit. For fixed \( \mu \) the following facts imply that **p1** holds if the condition “\( \mu \geq \mu_H \)” is replaced by “\( \Lambda \leq \lambda_P(\mu) \)”.

First, segmentation \((h)\) is optimal for fixed \( \lambda \) if and only if \( \mu > d^{-1} \) and \( \lambda \leq \lambda_P(\mu) \); this holds by Proposition 7.2 and its proof. Second, \( \Pi^*_\lambda(\lambda, \mu) \geq v > 0 \) for all \((\lambda, \mu)\) where segmentation \((h)\) is optimal, and \( \Pi^*_\lambda(\lambda, \mu) \) is continuous in \((\lambda, \mu)\); see Lemma 4. The proof of **p1** is complete if \( \mu \geq \mu_H \iff \Lambda \leq \lambda_P(\mu) \). This holds since \( \Lambda = \lambda_P(\mu_H) \) by the definition (3.66), \( \lambda_P(d^{-1}) = 0 \), and \( \lambda_P'(\mu) = 1 - 1/(2\sqrt{\mu d}) > 0 \) for \( \mu \geq d^{-1} \).

**p2.** Suppose that \( f_h(c_{\text{min}}) < 0 \). Then (i) \( \lambda^*(\mu) = \Lambda \) for \( \mu \geq \mu_{SD} \), and (ii) the optimal segmentation is \((h, m_{sd})\) if and only if \( \mu > \mu_{SD} \). Recall the threshold \( \Lambda_{sd} \) from Proposition 7.2(c) and its proof; it is defined in (3.104) of Lemma 9. We write \( \Lambda_{sd}(\mu) \) to make its dependence on \( \mu \) explicit. For fixed \( \mu \) the following facts imply that **p2** holds if the conditions “\( \mu \geq \mu_{SD} \)” and “\( \mu > \mu_{SD} \)” are replaced by “\( \Lambda \leq \Lambda_{sd}(\mu) \)” and “\( \Lambda < \Lambda_{sd}(\mu) \)”, respectively.

First, the optimal segmentation is \((h, m_{sd})\) at the largest feasible \( \lambda \) if and only if \( \Lambda < \Lambda_{sd}(\mu) \);
this holds by Proposition 7.2(c); for details see Lemma 9. Second, the revenue $\Pi^*$ satisfies $\Pi^*_\lambda(\lambda, \mu) \geq v > 0$ for all $(\lambda, \mu)$ where segmentation $(h, m_{sd})$ is optimal, $\Pi^*_\lambda(\lambda, \mu) \geq 0$ for all $(\lambda, \mu)$, and $\Pi^*_\lambda(\lambda, \mu)$ is continuous in $(\lambda, \mu)$; see Lemma 4. We complete the proof of p2 by showing that $\mu = \mu_{SD}(\Lambda) \Leftrightarrow \Lambda = \Lambda_{sd}(\mu)$ and $\mu > \mu_{SD}(\Lambda) \Leftrightarrow \Lambda < \Lambda_{sd}(\mu)$. We write $\mu_{SD}(\Lambda)$ to emphasize that $\mu_{SD}$ depends on $\Lambda$ through its defining equation (3.67). Fix $\mu$ and solve (3.67) for $\Lambda$ to get $\Lambda = \Lambda^*$. By Lemma 4, $\Pi^*_{\lambda}(\lambda, \mu) < \Pi^*_{\lambda'}(\lambda', \mu')$ if $\lambda < \lambda'$ for any $\mu$. It remains to show that there exists an unique threshold $\lambda_m(\mu) > 0$ since the LHS of (3.67) increases in $\mu$ and decreases in $\Lambda$. For fixed $\lambda$ the fact that $\Lambda + d^{-1} < \mu_{SD} < \mu_H$ follows since the LHS of (3.67) is 0 for $\mu = \Lambda + d^{-1}$, and 1 for $\mu = \mu_H$ since $\lambda_p(\mu_H) = \mu_H - \sqrt{\mu_H/d} = \Lambda$.

**Part 1.** For $\mu \leq \mu_{\min}$, $\lambda^*(\mu) = 0$ since no type buys at a positive price: $w(c) \geq 1/\mu_{\min} = d + v/c_{\min}$ implies $v + c \cdot (d - w(c)) \leq v (1 - c/c_{\min}) \leq 0$. For $\mu = \mu_{\min}$ we have $\Pi_{\lambda}(0, \mu) = 0$. For $\mu \in (\mu_{\min}, d^{-1})$ segmentation $(l)$ is optimal for all $\lambda$ by Proposition 7.1. By Lemma 4, $\Pi^*_{\lambda\mu}(\lambda, \mu) > 0 > \Pi^*_{\lambda\lambda}(\lambda, \mu)$ under segmentation $(l)$; therefore $\lambda^*(\mu) > 0$ for $\mu > \mu_{\min}$. Since $\Pi^*_{\lambda}(\lambda, \mu)$ is continuous in $(\lambda, \mu)$ we have $\lambda^*(\mu) < \Lambda$ for $\mu \in (\mu_{\min}, \mu_{\min} + \varepsilon)$ and small $\varepsilon > 0$.

We next show that there exists an unique threshold $\mu_A > \mu_{\min}$ such that $\lambda^*(\mu) = \Lambda$ if and only if $\mu \geq \mu_A$, where $\mu_A < \mu_H$ if $f_h(c_{\min}) \geq 0$ and $\mu_A < \mu_{SD}$ if $f_h(c_{\min}) < 0$. By Lemma 4, $\Pi^*(\lambda, \mu)$ is concave in $\lambda$ for fixed $\mu$, and $\Pi^*_{\lambda\mu}(\lambda, \mu) \geq 0$ for all $(\lambda, \mu)$. It follows that $\lambda^*(\mu) = \Lambda \Leftrightarrow \Pi^*_{\lambda}(\Lambda, \mu) \geq 0$ for any $\mu$, and if $\lambda^*(\mu) = \Lambda$ for some $\mu$ then $\lambda^*(\mu') = \Lambda$ for all $\mu' > \mu$. It remains to show that there exists $\mu$ that satisfies $\lambda^*(\mu) = \Lambda$ and either $\mu < \mu_H$ if $f_h(c_{\min}) \geq 0$, or $\mu_A < \mu_{SD}$ if $f_h(c_{\min}) < 0$. This holds since p1 implies that $\Pi^*_{\lambda}(\Lambda, \mu_H) = v > 0$, p2 implies that $\Pi^*_{\lambda}(\Lambda, \mu_{SD}) = v > 0$ if $f_h(c_{\min}) < 0$, and because $\Pi^*_{\lambda\mu}(\lambda, \mu)$ is continuous in $(\lambda, \mu)$ by Lemma 4.2.

It remains to show that $\lambda^*(\mu)$ is strictly increasing on $[\mu_{\min}, \mu_A]$. Lemma 4.1 implies that for fixed $\mu$, $\Pi^*(\lambda, \mu)$ has a unique maximizer $\lambda^*(\mu)$, and that if $\lambda^*(\mu) < \Lambda$ then the optimal segmentation is $(l), (h, l), (h, m), (m, l)$ or $(h, m, l)$. Under each of these segmentations, $\Pi^*_{\lambda\lambda}(\lambda, \mu) < 0 < \Pi^*_{\lambda\mu}(\lambda, \mu)$ (Lemma 4.1) and $\Pi^*_{\lambda\lambda}(\lambda, \mu)$ and $\Pi^*_{\lambda\mu}(\lambda, \mu)$ are continuous in $(\lambda, \mu)$ (Lemma 4.2). We have $\Pi_{\lambda}(\lambda^*(\mu), \mu) = 0$ for $\mu \in [\mu_{\min}, \mu_A]$. By the implicit function theorem
\( \lambda^*(\mu) \) is differentiable and \( \lambda''(\mu) = -\Pi^*_\mu(\lambda^*(\mu), \mu) / \Pi^*_\Lambda(\lambda^*(\mu), \mu) > 0 \) for \( \mu \in [\mu_{\min}, \mu_\Lambda] \).

**Part 2.** We first prove two key properties.

**p3.** If \( \mu \in (d^{-1}, \mu_H) \) and \( \lambda^*(\mu) = \Lambda \), then pooling must be optimal. For \( \mu \in (d^{-1}, \mu_H) \), by **p1-p2** and Proposition 7, only one of \((h, l), (h, m), (m, l), (h, m, l), (h, m_{sd})\) can be optimal. Only \((h, l)\) has no pooling, but it cannot be optimal if \( \lambda^*(\mu) = \Lambda \), for this rules out the necessary optimality condition \( f_l(c_l^*(\Lambda)) \leq f_h(c_h^*(\Lambda)) \) (Lemma 3.2). Recall that \( c_l^*(\Lambda) = F^{-1}(\lambda_l^*(\Lambda) / \Lambda) \) and \( c_h^*(\Lambda) = F^{-1}(\lambda_h^*(\Lambda) / \Lambda) \) where \( \lambda_l^*(\Lambda) \) and \( \lambda_h^*(\Lambda) \) are, respectively the optimal arrival rates to \( l \) and \( h \) classes for fixed \( \Lambda \). Under \((h, l)\) with \( \lambda^*(\mu) = \Lambda \), by (3.54) they satisfy \( \lambda_h^*(\Lambda) = \mu - \sqrt{\mu/d} \) and \( \lambda_l^*(\Lambda) = \Lambda - \lambda_h^*(\Lambda) \). It follows that \( c_l^*(\Lambda) = c_h^*(\Lambda) \). The proof is complete since by definition \( f_l(c) > f_h(c) \) for all \( c \).

**p4.** Suppose that pooling is not optimal for some fixed \( \mu < \mu_H \). Then pooling is also not optimal for all \( \mu' < \mu \). If \( \mu \leq d^{-1} \) this follows since segmentation \( (l) \) without pooling is optimal for all \( \mu' < \mu \) by Proposition 7.1. Suppose that \( \mu \in (d^{-1}, \mu_H) \). By **p3** segmentation \((h, l)\) must be optimal for \( \mu \), and \( \lambda^*(\mu) < \Lambda \). Let \( \lambda_h^*(\mu) = \mu - \sqrt{\mu/d} \) and \( \lambda_l^*(\mu) = \lambda^*(\mu) - \lambda_h^*(\mu) \) be the corresponding optimal rates. Optimality requires \( f_l(F^{-1}(\lambda_l^*(\mu)/\Lambda)) \leq f_h(F^{-1}(\lambda_h^*(\mu)/\Lambda)) \) (Lemma 3.2) and

\[
0 = \Pi_\Lambda(\lambda^*(\mu), \mu) \Leftrightarrow v = \Lambda \int_{c_{\min}}^{F^{-1}(\lambda_l^*(\mu)/\Lambda)} \frac{f(x) f_l(x) 2\mu}{(\mu - \lambda^*(\mu) + \Lambda F(x))^3} dx \\
= \Lambda \int_{c_{\min}}^{F^{-1}(\lambda_l^*(\mu)/\Lambda)} \frac{f(x) f_l(x) 2\mu}{(\sqrt{\mu/d} - \lambda_h^*(\mu) + \Lambda F(x))^3} dx,
\]

where the second equation holds since \( \mu - \lambda^*(\mu) = \sqrt{\mu/d} - \lambda_h^*(\mu) \). We have \( \lambda_l''(\mu) > 0 \), since the RHS of the equation strictly decreases in \( \mu \) for fixed \( \lambda_l^*(\mu) \) and strictly increases in \( \lambda_l^*(\mu) \) for fixed \( \mu \). Noting that \( \lambda_l''(\mu) = 1 - 1/(2\sqrt{\mu d}) > 0 \) implies that \( f_l(F^{-1}(\lambda_l^*(\mu)/\Lambda)) - f_h(F^{-1}(\lambda_h^*(\mu)/\Lambda)) \) strictly increases in \( \mu \). Therefore the optimality conditions for \((h, l)\) hold for every \( \mu' \in (d^{-1}, \mu) \).

Part 1 and **p3-p4** imply that there is an unique threshold \( \mu_P \in [d^{-1}, \mu_H] \) such that pooling is not optimal for \( \mu \leq \mu_P \) and optimal for \( \mu \in (\mu_P, \mu_H) \). These properties combined
with properties \(p_1-p_2\) shown above, establish Parts 2(a)-(b). Part 2(c) is based on two facts. If \(\mu_A > d^{-1}\), then \(\mu_P < \mu_A\) since pooling is optimal for \(\mu = \mu_A\) by \(p_3\) (since \(\lambda^*(\mu_A) = \Lambda\) and \(\mu_A < \mu_H\) by Part 1) and since the solution is continuous in \(\mu\). Second, if \(\mu_A \leq d^{-1}\), then \(\mu_P = d^{-1}\) by \(p_3\). By Part 1, the necessary and sufficient condition for \(\mu_A \leq d^{-1}\) is that \(\Pi_\lambda(\Lambda, \mu) \geq 0\) for \(\mu = d^{-1}\). Segmentation \((l)\) is optimal for \(\mu = d^{-1}\); substituting in (3.70) with \(\lambda_l^*(\mu) = \Lambda\) yields

\[
\Pi_\lambda (\Lambda, \mu)|_{\mu = d^{-1}} = v - \Lambda \int_{c_{\min}}^{c_{\max}} \frac{f(x) f_l(x) 2d^2}{(1 - d\Lambda F(x))^2} dx. \tag{3.71}
\]

**Part 3.** Follows by Proposition 7.1 for \(\mu \in (\mu_{\min}, d^{-1}]\), Part 2 for \(\mu \in (d^{-1}, \mu_P]\), and \(p_1\) for \(\mu \geq \mu_H\).

**Part 4.** Follows by Part 2 and property \(p_2\).

**Part 5.** Follows by property \(p_2\).

3.4.3 Optimal Menu under Work Conserving Strict Priority Discipline

As noted in Section 3.3.2, when types are observable, unlike in our setting, then the optimal lead times are obtained by a standard work conserving policy which strictly prioritizes customers in the order of their delay costs. The results of Sections 3.4.1 and 3.4.2 show that when customer types are private information the optimal lead time menu may involve two deviations from this standard policy: 1. Pooling of types with different delay costs into a common service class, and 2. Strategic delay to artificially inflate the mean delays of certain classes above the operationally feasible levels.

For comparison, in this Section we characterize the customer segmentation and lead times that are optimal given the provider operates a work conserving scheduling policy where customers are assigned strict priorities based on their delay cost (type). In particular, no customer types with different delay costs can be pooled and no strategic delay may be added.
to the system. From the perspective of the lead time menu, this restriction implies that the set of offered – and purchased – lead times must form a single interval. Mathematically the provider modifies Problem 3 by setting $\lambda_l + \lambda_h = \lambda$, which forces $\lambda_m = 0$.

Proposition 9 establishes that given a work conserving strict priority scheduling policy, it may be optimal at higher capacity to exclude low delay cost types that are served at lower capacity levels. We call this strategic exclusion and discuss it in Section 3.5 with a numerical example.

**Proposition 9.** [Segmentation and Lead Times Under Strict Priorities]

Let $\lambda^p(\mu)$ be the optimal arrival rate as a function of $\mu$ if the offered lead times are restricted to those attainable by a work conserving scheduling policy which strictly prioritizes customers in the order of their delay costs. Suppose that $f_h(c_{\text{min}}) < 0$.

1. Define the valuation threshold $v_H := -df_h(c_{\text{min}})$. If $v \geq v_H$ then there exists a capacity threshold $\mu_A < \mu_H$ such that $\lambda^p(\mu)$ increases to $\Lambda$ on $[\mu_{\text{min}}, \mu_A]$ and $\lambda^p(\mu) = \Lambda$ for $\mu \geq \mu_A$.

2. There exists a valuation threshold $v_M \in (0, v_H)$ where for $v \in [v_M, v_H)$ there are two capacity thresholds $\mu_A < \mu_H < \mu_E$ such that $\lambda^p(\mu)$ increases to $\Lambda$ on $[\mu_{\text{min}}, \mu_A]$, stays constant at $\lambda^p(\mu) = \Lambda$ on $[\mu_A, \mu_E]$ and decreases for $\mu > \mu_E$.

3. There exists a valuation threshold $v_L \in (0, v_M)$ such that for $v < v_L$ there exists no capacity level at which it is optimal to serve the entire market: $\lambda^p(\mu) < \Lambda$ for all $\mu$.

**Proof of Proposition 9**

Let $\Pi^P(\lambda, \mu)$ denote the revenue function under the work conserving priority (WCP) policy which strictly prioritizes types in the order of their delay costs; it rules out strategic delay and pooling. Recall that $\Pi(\lambda, \mu)$ is the revenue under optimal segmentation and scheduling, without restriction to the WCP policy. Let $\lambda(\mu)$ denote a maximizer of $\Pi^P(\lambda, \mu)$ as a function of $\mu$. By Proposition 7.1 if $\mu^{-1} \geq d$ the WCP policy and the segmentation $(l)$ are
optimal for all \( \lambda \), so \( \Pi^P(\lambda, \mu) \) agrees with \( \Pi(\lambda, \mu) \) for all \( \lambda \). This is not the case for \( \mu^{-1} < d \).

We first characterize \( \Pi^P(\lambda, \mu) \) for \( \mu^{-1} < d \), then the valuation thresholds, and then prove Parts 1-3.

**Revenue function** \( \Pi^P(\lambda, \mu) \) for \( \mu^{-1} < d \). Note that \( \lambda_P(\mu) = \mu - \sqrt{\mu/d} \), defined in Proposition 7, is the arrival rate at which the maximum lead time equals \( d \) under the WCP policy. It satisfies \( \lambda_P(d^{-1}) = 0 \) and \( \lambda_P(\mu) > 0, \lambda_P'(\mu) > 0 \) for \( \mu^{-1} < d \). Mathematically, the WCP policy amounts to forcing \( \lambda_t + \lambda_h = \lambda \) in Problem 3. With this restriction constraints (3.49)-(3.50) require: if \( \lambda \leq \lambda_P(\mu) \) then \( \lambda_h = \lambda \), and if \( \lambda > \lambda_P(\mu) \) then \( \lambda_h = \lambda - \lambda_P(\mu) \). For \( \mu^{-1} < d \), we have

\[
\Pi^P(\lambda, \mu) = \begin{cases}
\Pi^h(\lambda, \mu) \triangleq (\lambda - \lambda_P(\mu)) v - \Lambda \int_{c_{\min}}^{c_{\max}} f(x) f_h(x) \left( d - \frac{\mu}{(\mu - \Lambda F(x))} \right) dx, & \lambda \leq \lambda_P(\mu), \\
\Pi^{hl}(\lambda, \mu) \triangleq \Pi^h(\lambda, \mu) v - \Lambda \int_{c_{\min}}^{c_{\max}} f(x) f_1(x) \left( d - \frac{\mu}{(\mu - \Lambda F(x))} \right) dx + \Pi^h(\lambda_P(\mu), \mu), & \lambda > \lambda_P(\mu),
\end{cases}
\]

where \( c_l(\lambda - \lambda_P(\mu)) = F^{-1}(\lambda - \lambda_P(\mu) / \Lambda) \) and \( c_h(\lambda) = F^{-1}(\lambda / \Lambda) \). Since \( f_h(c_{\min}) < 0 \) and \( f_h' > 0 \) we have \( f_h^{-1}(0) \in (c_{\min}, c_{\max}) \), and \( \lambda_0 = \Lambda F^{-1}(f_h^{-1}(0)) \) satisfies \( \lambda_0 \in (0, \Lambda) \). The revenue \( \Pi^h(\lambda, \mu) \) agrees with \( \Pi(\lambda, \mu) \) if and only if segmentation \( (h) \) is optimal for \( (\lambda, \mu) \); this holds if \( \lambda \leq \min(\lambda_P(\mu), \lambda_0) \) but not if \( \lambda_0 < \lambda \leq \lambda_P(\mu) \). Similarly, \( \Pi^{hl}(\lambda, \mu) \) agrees with \( \Pi(\lambda, \mu) \) if and only if segmentation \( (h, l) \) is optimal for \( (\lambda, \mu) \). We have

\[
\Pi^P(\lambda, \mu) = \begin{cases}
\Pi^h(\lambda, \mu) = v + f_h(c_h(\lambda)) \left( d - \frac{\mu}{(\mu - \lambda)^2} \right), & \lambda \leq \lambda_P(\mu), \\
\Pi^{hl}(\lambda, \mu) = v - \Lambda \int_{c_{\min}}^{c_{\max}} f(x) f_1(x) \left( d - \frac{\mu}{(\mu - \Lambda F(x))} \right) dx, & \lambda > \lambda_P(\mu).
\end{cases}
\]

The following properties of \( \Pi^P(\lambda, \mu) \) are immediate from (3.72), noting that \( f_h'(c_h(\lambda)) c_h'(\lambda) < 0 \), \( f_h(c_h(\lambda_0)) = 0 \) and the multiplier of \( f_h(c_h(\lambda)) \) is positive for \( \lambda < \lambda_P(\mu) \) and zero for \( \lambda = \lambda_P(\mu) \).

**p1.** If \( \lambda \leq \min(\lambda_P(\mu), \lambda_0) \) then \( \Pi^P(\lambda, \mu) = \Pi^h(\lambda, \mu) \geq v \) and \( \Pi^{hl}(\lambda, \mu) = \Pi^{hl}(\lambda, \mu) \geq 0 \).

**p2.** If \( \lambda_0 < \lambda_P(\mu) \) then \( \Pi^P(\lambda, \mu) = \Pi^h(\lambda, \mu) = v \) for \( \lambda = \lambda_0 \) or \( \lambda = \lambda_P(\mu) \), and \( \Pi^P(\lambda, \mu) = \Pi^{hl}(\lambda, \mu) < v \) for \( \lambda_0 < \lambda < \lambda_P(\mu) \).


p3. If \( \lambda_0 < \lambda_P(\mu) \) then \( \Pi^P_{\lambda\mu}(\lambda_0, \mu) = \Pi^h_{\lambda\mu}(\lambda_0, \mu) = 0 \) and \( \Pi^P_{\lambda\mu}(\lambda, \mu) = \Pi^h_{\lambda\mu}(\lambda, \mu) < 0 \) for \( \lambda \in (\lambda_0, \lambda_P(\mu)) \).

p4. If \( \lambda > \lambda_P(\mu) \) then \( \Pi^P_{\lambda\lambda}(\lambda, \mu) = \Pi^h_{\lambda\lambda}(\lambda, \mu) < 0 < \Pi^P_{\lambda\mu}(\lambda, \mu) = \Pi^h_{\lambda\mu}(\lambda, \mu) \), see Lemma 4.

Note that \( \Pi^P(\lambda, \mu) \) is linear in the parameter \( v \). We write \( \Pi^P(\lambda, \mu; v) \) and \( \lambda(\mu; v) \) when making the dependence of these quantities on \( v \) explicit.

Valuation threshold \( v_M \). Let

\[
v_M \triangleq - \min_{\lambda_0 \leq \lambda \leq \lambda_P(\mu_H)} \left\{ \Pi^P_{\lambda}(\lambda, \mu_H; 0) = f_h(c_h(\lambda)) \left( d - \frac{\mu_H}{(\mu_H - \lambda)^2} \right) \right\},
\]

where (3.66) in Proposition 8 implies that \( \lambda_P(\mu_H) = \Lambda \). We have \( v_M > 0 \), since \( f_h(c_h(\lambda)) < 0 \) for \( \lambda > \lambda_0 \) and its multiplier is strictly positive for \( \lambda < \lambda_P(\mu_H) \).

Valuation threshold \( v_L \). For all \( \mu \) it is optimal to serve less than \( \Lambda \) if \( v = 0 \), i.e., \( \lambda(\mu; 0) < \Lambda \). For fixed \( \mu \), p1-p2 and (3.72) imply: \( \Pi^P_{\lambda}(\lambda, \mu; 0) > 0 \) if and only if \( \lambda < \min(\lambda_P(\mu), \lambda_0) \); \( \Pi^P_{\lambda}(\lambda, \mu; 0) = 0 \) for \( \lambda = \min(\lambda_P(\mu), \lambda_0) \); for \( \lambda > \min(\lambda_P(\mu), \lambda_0) \) we have \( \Pi^P_{\lambda}(\lambda, \mu; 0) \leq 0 \) with equality if and only if \( \lambda = \lambda_P(\mu) > \lambda_0 \). Hence for \( v = 0 \) and fixed \( \mu \) the maximizer satisfies \( \lambda(\mu; 0) = 0 \) if \( \mu \leq d^{-1} \) and \( \lambda(\mu; 0) = \min(\lambda_P(\mu), \lambda_0) \) for \( \mu > d^{-1} \). It follows that \( \Lambda - \lambda(\mu; 0) \geq \Lambda - \lambda_0 = F(f^{-1}_h(0)) > 0 \).

Based on this one can construct \( v_L > 0 \) such that if \( v < v_L \) then \( \lambda(\mu; v) < \Lambda \) for all \( \mu \). We use several properties that follow since \( \lambda_P(\mu_H) = \Lambda \) and \( \lambda_P(\mu) > 0 \): p2 implies that \( \Pi^P_{\lambda}(\lambda, \mu_H; 0) = 0 \), p3 implies that \( \Pi^P_{\lambda\mu}(\lambda, \mu) < 0 \) for \( \mu > \mu_H \), and p4 implies that \( \Pi^P_{\lambda\mu}(\lambda, \mu; 0) > 0 \) for \( \mu \in (\Lambda, \mu_H) \).

Let \( v_L = \min(v_1, v_2) \), where \( v_1 \) and \( v_2 \) are defined as follows.

Fix \( \mu_1 \in (\Lambda, \mu_H) \) and let \( v_1 = -\Pi^P_{\lambda}(\Lambda, \mu_1; 0) \), where \( v_1 > 0 \) since \( \Lambda > \min(\lambda_P(\mu_1), \lambda_0) \).

For \( v < v_1 \) we have \( \lambda(\mu; v) < \Lambda \) for all \( \mu \leq \mu_1 \). For \( \mu \leq \Lambda \) this is obvious. For \( \mu \in (\Lambda, \mu_1) \) this follows since \( \Pi^P_{\lambda\mu}(\Lambda, \mu) > 0 \), so \( \Pi^P_{\lambda}(\Lambda, \mu; v) \leq \Pi^P_{\lambda}(\Lambda, \mu_1; v) = v + \Pi^P_{\lambda}(\Lambda, \mu_1; 0) = v - v_1 < 0 \).

Let \( v_2 = \Lambda^{-1}\min\{\mu \in [\mu_1, \mu_H]: \Pi^P(\lambda(\mu; 0), \mu; 0) - \Pi^P(\lambda, \mu; 0)\} \), where \( v_2 > 0 \) because the difference \( \Pi^P(\lambda(\mu; 0), \mu; 0) - \Pi^P(\lambda, \mu; 0) \) is strictly positive for every \( \mu \) and continuous on \([\mu_1, \mu_H]\). For \( v \in (0, v_2) \) we have \( \lambda(\mu; v) < \Lambda \) for all \( \mu \in [\mu_1, \mu_H] \). This follows since for each \( \mu \in [\mu_1, \mu_H] \) the revenue rate at \( \lambda(\mu; 0) \) is strictly larger than at \( \Lambda: \Pi^P(\lambda(\mu; 0), \mu; v) - \Pi^P(\lambda(\mu; 0), \mu; 0) \).
\( \Pi^P (\lambda, \mu; v) = \Pi^P (\lambda (\mu; 0), \mu; 0) - \Pi^P (\Lambda, \mu; 0) + v (\lambda (\mu; 0) - \Lambda) \geq v_2 \Lambda + v (\lambda (\mu; 0) - \Lambda) > v \lambda (\mu; 0) \geq 0. \)

**Part 1.** Fix \( v \geq v_H = -df_h (c_{\min}) > 0. \) First note that for \( \mu > d^{-1} \) the maximizer satisfies
\[ \lambda (\mu; v) > \lambda_P (\mu) \text{ if } \lambda_P (\mu) < \Lambda \text{ and } \lambda (\mu; v) = \Lambda \text{ if } \Lambda \leq \lambda_P (\mu). \]
If \( \lambda \leq \min (\lambda_P (\mu), \lambda_0) \) then \( \Pi^P (\lambda, \mu; v) \geq v > 0 \) by p1. If \( \lambda \in (\lambda_0, \lambda_P (\mu)] \), then p3 and the fact that \( f_h (c_h (\lambda)) \) decreases in \( \lambda \) with \( f_h (c_h (\lambda)) < 0 \) imply \( \Pi^P (\lambda, \mu; v) > \lim_{\mu \to \infty} \Pi^P (\lambda, \mu; v) = v + df_h (c_h (\lambda)) \geq v + df_h (c_{\min}) \geq 0. \)
Therefore the optimal segmentation is \( (l) \) for \( \mu \leq d^{-1} \) and \( (h, l) \) for \( \mu > d^{-1} \). Since \( \Pi^P (\lambda, \mu) < 0 < \Pi^P (\lambda, \mu) \) under either segmentation, it follows from a similar argument as in the proof of Proposition 8.1 that there exists \( \mu_A \in (\mu_{\min}, \mu_H) \) such that \( \lambda (\mu; v) \) strictly increases to \( \Lambda \) on \([\mu_{\min}, \mu_A]\) and \( \lambda (\mu; v) = \Lambda \) for \( \mu \geq \mu_A. \)

**Part 2.** That \( v_M < v_H \) follows since \( \Pi^P (\lambda, \mu_H; 0) > df_h (c_{\min}) \) for \( \lambda \in [\lambda_0, \lambda_P (\mu)]. \) We have \( \Pi^P (\lambda_0, \mu_H; 0) = 0 > df_h (c_{\min}) \) and \( \Pi^P (\lambda, \mu_H; 0) > \lim_{\mu \to \infty} \Pi^P (\lambda, \mu; 0) = df_h (c_h (\lambda)) \geq df_h (c_{\min}) \) for \( \lambda \in (\lambda_0, \lambda_P (\mu)] \); the first inequality follows by p3, the second since \( f_h (c_h (\lambda)) \) decreases in \( \lambda \) and \( f_h (c_h (\lambda)) = f_h (c_{\min}) \). The definition of \( v_M \) and continuity of \( \Pi^P (\lambda, \mu_H; 0) \) imply that \( v_M < v_H. \)

Fix \( v \in [v_M, v_H) \). First note that for \( \mu \in (d^{-1}, \mu_H) \) the maximizer satisfies \( \lambda (\mu; v) > \lambda_P (\mu) \) and \( \lambda (\mu_H; v) = \Lambda = \lambda_P (\mu_H) \). If \( \lambda \leq \min (\lambda_P (\mu), \lambda_0) \) then \( \Pi^P (\lambda, \mu; v) \geq v > 0 \) by p1. If \( \lambda \in (\lambda_0, \lambda_P (\mu)] \), then \( \Pi^P (\lambda, \mu; v) = v + \Pi^P (\lambda, \mu; 0) \geq v + \Pi^P (\lambda, \mu_H; 0) \geq v - v_M \geq 0 \) where the first inequality follows since \( \Pi^P (\lambda, \mu) < 0 \) by p3, and the second by definition of \( \Pi^P (\lambda, \mu) = v_M. \) The same argument as in Part 1 implies that there exists \( \mu_A \in (\mu_{\min}, \mu_H) \) such that \( \lambda (\mu; v) \) strictly increases to \( \Lambda \) on \([\mu_{\min}, \mu_A]\) and \( \lambda (\mu; v) = \Lambda \) for \( \mu \in [\mu_A, \mu_H] \). It remains to show that there exists \( \mu_E \in (\mu_H, \infty) \) such that \( \lambda (\mu; v) = \Lambda \) for \( \mu \in [\mu_H, \mu_E] \) and \( \lambda (\mu; v) \) strictly decreases on \([\mu_E, \infty) \). That \( \lambda (\mu; v) = \Lambda \) for all \( \mu \) in some interval \([\mu_H, \mu_H + \varepsilon]\) follows since \( \Pi^P (\lambda, \mu; v) \) is continuous in \( \mu \), and for \( v \geq v_M \) we have \( \lambda (\mu_H; v) = \Lambda \) as the unique maximizer of \( \Pi^P (\lambda, \mu_H; v) \). The existence of \( \mu > \mu_H \) such that \( \lambda (\mu; v) < \Lambda \) follows since \( \lim_{\mu \to \infty} \Pi^P (\lambda, \mu; v) = v + df_h (c_{\min}) = v - v_H < 0 \). That the threshold \( \mu_E > \mu_H \) is unique and \( \lambda (\mu; v) \) strictly decreases on \([\mu_E, \infty) \) follows because \( \lambda (\mu; v) > \lambda_0 \) for all \( \mu > \mu_H, \)
\( \Pi^P (\lambda (\mu; v), \mu; v) = 0 > \Pi^P (\lambda (\mu; v), \mu; v) \) if \( \lambda (\mu; v) < \Lambda \), and \( \Pi^P (\lambda (\mu; v), \mu; v) < 0 \) for all \( \mu > \mu_H \).
and $\lambda \in (\lambda_0, \Lambda]$.

**Part 3.** For $v < v_L$ we have $\lambda(\mu; v) < \Lambda$ for all $\mu$. For $\mu \leq \mu_H$ this is immediate from the definition of $v_L = \min(v_1, v_2)$. For $\mu > \mu_H$ this follows since the maximizer for $v = 0$ is constant in $\mu$ and satisfies $\lambda(\mu; 0) = \lambda_0$, and p3 implies $\Pi^P_{\lambda, \mu}(\lambda, \mu) < 0$ for $\lambda \in [\lambda_0, \Lambda]$. As a result we have $\Pi^P (\lambda(\mu_H; 0), \mu_H; v) < \Pi^P (\Lambda, \mu_H; v) \leq \Pi^P (\lambda(\mu; 0), \mu; v) - \Pi^P (\Lambda, \mu; v)$ for $\mu > \mu_H$.

We show that $v_L < v_M$ by contradiction, i.e., if $v_L \geq v_M$ then there exists $v < v_L$ with maximizer $\lambda(\mu_H; v) = \Lambda$. By definition, $v_L \geq v_M$ implies that $v_L \geq -\Pi^P(\lambda, \mu_H; 0)$ for $\lambda \in [\lambda_0, \lambda_P(\mu_H)]$, and so $\Pi^P_{\lambda}(\lambda, \mu_H; v_L) = \Pi^P(\lambda, \mu_H; 0) + v_L \geq 0$ for $\lambda \in [\lambda_0, \lambda_P(\mu_H)]$. Since $\Lambda = \lambda_P(\mu_H)$ we have $\Pi^P_{\lambda}(\Lambda, \mu_H; v_L) = v_L > 0$ by p2; therefore $\Lambda$ is the unique maximizer of $\Pi^P(\lambda, \mu_H; v)$ for $v = v_L$ and, since $\Pi^P_{\lambda}(\lambda, \mu_H; v)$ is continuous in $\lambda$, also for some $v < v_L$.

3.5 Numerical Examples: Strictly Increasing Virtual Delay Cost Functions

To illustrate the results, we consider the case where the delay cost rate distribution $f$ is uniformly distributed on the unit length interval $[c_{\min}, c_{\max}]$ and the market size $\Lambda = 1$. This distribution has strictly increasing virtual delay cost functions: $f_l(c) = 2c - c_{\min}$ and $f_h(c) = 2c - c_{\min} - 1$. Let $c_0 := \inf\{c \geq c_{\min} : f_h(c) > 0\}$ be the boundary type between those with positive and negative virtual delay costs under high lead time quality. It satisfies $c_0 = \max(c_{\min}, (c_{\min} + 1)/2)$.

3.5.1 Optimal Segmentation and Lead Times: Impact of Capacity

This Section illustrates the results of Propositions 8 and 9 on how the segmentation and lead time menu depend on the capacity $\mu$. In particular, we highlight the role of the delay cost distribution $f$ in the optimality of pooling, strategic delay and strategic exclusion.
Non-negative virtual delay cost function: \( f_h(c_{\text{min}}) \geq 0 \). Figure 3.2 shows the segmentation of types into low \((w(c) > d)\), medium \((w(c) = d)\) and high \((w(c) < d)\) lead time qualities as a function of \(\mu\); in Panel (a) for the optimal lead time menu and in Panel (b) for the menu that is optimal if restricted to a work conserving strict priority policy. The type domain satisfies \(c_{\text{min}} = 1\) and \(c_{\text{max}} = 2\), so \(c_0 = c_{\text{min}}\) and \(f_h(c) \geq 0\) on \([c_{\text{min}}, c_{\text{max}}]\). Therefore strategic delay is not optimal at any capacity level. In each panel, the capacity is recorded on the horizontal, and the type domain \([c_{\text{min}}, c_{\text{max}}]\) on the vertical axis. The boundaries indicated by the dashed, dotted, and solid lines denote, respectively, the lowest customer type buying \(w(c) < d\), the lowest type buying \(w(c) = d\), and the highest type buying \(w(c) > d\). At any capacity level where (i) pooling is optimal (the lead time \(d\) is sold) and (ii) only a fraction of the market is served, we show the set of types buying \(w(c) = d\) as an interval contiguous to the interval of types buying \(w(c) < d\), bounded above by the dashed line and below by the dotted line – we follow this convention in all such Figures. E.g., in Figure 3.2(a) at \(\mu = 1.5\) the set of types buying \(w(c) = d\) is shown as the type interval \([1.2, 1.6]\). Recall from the discussion of Proposition 6 that while this interval represents only one of many such equilibrium type sets, all such sets have the same arrival rate, where in this example \(\lambda^*_m(\mu) = 0.4\).

The optimal segmentation and lead time menu shown in Figure 3.2(a) has the following features. The capacity thresholds satisfy \(\mu_{\text{min}} = 1/4 < \mu_L := 1/d = 1/2 < \mu_P \approx 1 < \mu_A \approx 1.65 < \mu_H = 2\). The provider sells only low lead time qualities at low capacity \(\mu \in (1/4, 1/2]\); low and high lead time qualities at larger capacity \(\mu \in (1/2, 1]\); low, medium and high lead time qualities at \(\mu \in (1, 2)\) with multiple types pooled into the medium lead time class; and only high lead time qualities for \(\mu > 2\). Notice that as \(\mu\) increases from 1 to 2, the average delay cost rate of the set of types that are pooled decreases. When a work conserving strict priority policy is enforced, see Figure 3.2(b), the pooling of types is precluded. The two segmentations differ for \(\mu \in (1, 2)\) in that pooling is part of the optimal menu and the menu under strict priorities yields larger demand rates for low and high lead time qualities.

Negative virtual delay costs \(f_h(c)\) for low types: \(f_h(c_{\text{min}}) < 0\). Figure 3.3 shows
the same comparison as Figure 3.2, for a customer population that only differs in the domain of the delay cost distribution: $c_{\text{min}} = 0$, $c_{\text{max}} = 1$, so $c_0 = 1/2$ and $f_h(c) < 0$ on $[0, 1/2)$. As a result the optimal menu involves the insertion of strategic delay at larger capacity levels. The capacity thresholds satisfy $\mu_{\text{min}} = 0 < \mu_L := 1/d = 1/2 < \mu_P \approx 0.8 < \mu_A \approx 1.4 < \mu_{sd} \approx 1.7 < \mu_H = 2$. Under the optimal scheduling policy this case yields two main differences compared to the case where all types have positive virtual delay costs $f_h(c)$: (i) pooling types (for the lead time $d$) is optimal for all $\mu > \mu_P$, whereas if $f_h(c) \geq 0$ doing so is optimal only at intermediate capacity levels; (ii) for $\mu > \mu_{sd}$ the lead time $d$ involves the pooling of all low types with $f_h(c) < 0$ (in this case equal to half the market) and requires strategic delay, i.e., the optimal policy is not work conserving. Under a work conserving strict priority policy the optimal segmentation in Figure 3.3(b) for $f_h(c_{\text{min}}) < 0$ exhibits the same structure as in the case where $f_h \geq 0$ (Figure 3.2(b)), but notice in Figure 3.3(b) that the optimal revenue decreases in the capacity for $\mu \geq 1.7$. The reason is that at large capacity levels, selling short lead times $w(c) < d$ to low types $c$ with $f_h(c) < 0$ without the ability to artificially

Figure 3.2: Optimal Customer Segmentation and Lead Times: Positive Virtual Delay Costs $f_h(c)$
delay them, reduces the price the provider can charge for more time-sensitive types. This is also the reason why it may be optimal at higher capacity to strategically exclude some fraction of these low delay cost types that are served at lower capacity levels. The underlying intuition is that at larger capacity levels, the lead times attained under a work conserving strict priority discipline differentiate less between the highest and lowest delay cost types. In the extreme, as \( \mu \to \infty \), all customers would receive the same delay. Incentive constraints would require that the offered prices also converge and approach that paid by the lowest type served. Figure 3.4 illustrates this result of Proposition 9. The parameters are the same as in Figure 3.3, except that \( v \) is lowered to \( v = 1 \) in Figure 3.4(a) and \( v = 0.1 \) in Figure 3.4(b). In both panels the rate of admitted customers increases for \( \mu < \mu_E \) and decreases for \( \mu > \mu_E \). For \( v = 1 \), there is a capacity range \( \mu \in [\mu_H, \mu_E] \) for which it is profitable to serve the entire market, whereas for very low \( v = 0.1 \) it is not profitable to serve the entire market at any capacity level. These cases correspond to Parts 2 and 3 of Proposition 9, respectively.

Figure 3.5 shows the percentage gain in revenue from using optimal scheduling over the
Figure 3.4: Work Conserving Strict Priority Scheduling: Strategic Exclusion when \( v \) is low and \( f_h(c_{\text{min}}) < 0 \).

Figure 3.5: Revenue improvement of Optimal over Work Conserving Strict Priority menu. Dashed: \( f_h(c_{\text{min}}) < 0 \), \( c_{\text{min}} = 0 \) and \( c_{\text{max}} = 1 \). Solid: \( f_h(c) > 0 \), \( c_{\text{min}} = 0 \) and \( c_{\text{max}} = 1 \). Uniform \( f() \), \( \Lambda = 1 \), \( d = 2 \), \( v = 1 \).
conditionally optimal menu for customer distributions with strictly positive and negative virtual delay costs. When virtual delay costs are strictly positive the optimal menu differs from a strictly prioritized menu only in the intermediate segment where customers are pooled into a medium lead time segment. This provides a modest revenue improvement visible between capacities 1.25 to 2. When negative virtual costs exist, there are revenue benefits for all capacities greater than 1.25. The benefits are substantially greater, in particular, when strategic delay is being used.

3.5.2 Optimal Segmentation and Lead Times: Impact of Parameters $d$ and $v$

The parameters $d$ and $v$ define the value-delay cost function $V(c) = v + d \cdot c$ for $c \in [c_{\text{min}}, c_{\text{max}}]$.

Increasing $v$ increases the types’ valuations while keeping the variance constant. As a result, higher prices can be charged to all customers so that the optimal menu serves more customers and yields a higher utilization. Figure 3.6(a) shows for fixed $\mu$ that the optimal rate of customers served increases in $v$ and how the optimal segmentation changes with $v$. The parameters correspond to Part 2(c) of Proposition 7: the optimal segmentation follows $(h) \rightarrow (h, l) \rightarrow (h, m, l) \rightarrow (m, l)$ as $v$ increases from 0 to 30.

Increasing $d$ increases the valuations of all types and the variance of valuations. For one the optimal rate of customers served increases in $d$, as in the case of increasing $v$, see Figure 3.6(b). An increase in $d$ also alters the relative ranking of types’ willingness to pay for a given lead time, making high types with high delay cost $c$ more profitable than low types. In particular, the willingness to pay for a given lead time $w$ satisfies $v + c(d - w)$; for $d < w$ it is below $v$ and decreasing in $c$, but for $d > w$ it is larger than $v$ and increasing in $c$. This explains the pattern shown in Figure 3.6(b). For small $d < \mu^{-1} \approx 0.8$ all feasible lead times are of “low” quality, i.e., $w(c) > d$, and the segment of types buying these lead times always includes the lowest type; more impatient types are less profitable. For sufficiently large $d \geq 14.5$ all customers can be served under strict priorities with a lead time less than
Chapter 3. Revenue Maximizing Lead-Time Pricing

Figure 3.6: Optimal Segmentation and Leadtimes: Sensitivity to Parameters $v$ and $d$ ($c_{\text{min}} = 1$, $c_{\text{max}} = 2$).

$d$; in the example all types are served with “high” lead time quality. For intermediate values of $d$ some fraction of types are pooled in the medium lead time class; the size of this segment first increases and then decreases in $d$.

3.6 Discussion and Extensions

We discuss how our results generalize under (i) nonmonotone virtual delay cost functions and (ii) a nonlinear value-delay cost relationship.

3.6.1 Nonmonotone Virtual Delay Cost Functions

The analysis and results have so far focused on cases where the virtual delay cost functions $f_l$ and $f_h$ are strictly increasing. Under this assumption, the pooling of multiple types into a common service class can only be optimal for the medium lead time $d$. In particular, types $c \in C_l$ and/or $c \in C_h$ who buy low or high lead time qualities always purchase different lead
times – it is not optimal to pool multiple types into a common class with lead time strictly smaller or larger than \( d \).

In general the function \( f_l \) and \( f_h \) need not be increasing. We briefly discuss (i) market properties that yield nonmonotone virtual delay cost functions, (ii) how our results on the optimal segmentation and lead time menu generalize to such settings, and (iii) a numerical example that illustrates the properties of the optimal segmentation and menu that may arise in such settings.

**Market properties that yield nonmonotone virtual delay cost functions.** Nonmonotone virtual delay cost functions arise quite naturally in settings where the market is composed of two or more segments where delay costs within each segment have a unimodal distribution and the differences in the mean delay costs of these segments are large relative to the within-segment delay cost variances. The simplest example is that of a bimodal distribution, which we illustrate in a numerical example below. In general the definitions of \( f_l \) and \( f_h \) in (3.30) and (3.31) imply the following necessary and sufficient conditions for the virtual delay cost functions to be decreasing at type \( c \) (for simplicity we assume that \( f \) is differentiable):

\[
f_l'(c) < 0 \iff 2 < \frac{F(c)f'(c)}{(f(c))^2} \quad \text{and} \quad f_h'(c) < 0 \iff 2 < \frac{-F(c)f'(c)}{(f(c))^2}.
\]

Specifically, \( f_l(c) \) is decreasing if the density \( f \) is increasing at \( c \), there is a substantial mass of lower delay cost types and the local density at \( c \) is small. Also note that the subsets of \([c_{\text{min}}, c_{\text{max}}]\) on which \( f_l \) and \( f_h \) are decreasing are mutually exclusive.

**Properties of optimal segmentation and lead time menus.** The presence of nonmonotone virtual delay cost functions has the following impact on our results.

1. The basic property of the optimal structure is that types are segmented into low, medium and high lead time qualities (\( C_l \), \( C_m \) and \( C_h \)). This result remains valid; it is established by Proposition 6 which does not rely on any assumptions regarding the virtual delay cost functions.
2. The main structural change is that the pooling of multiple types may be optimal not only for the medium lead time $d$ but also for service classes with lead time strictly smaller or larger than $d$. The necessary conditions of Lemma 3 for the optimal segmentation, lead times and scheduling policy remain partially valid. In particular, Part 2 generalizes if applied to the notion of types’ mean virtual delay costs, as stated in Proposition 10 below. The pooling of multiple types into low and/or high lead time qualities has a couple of implications for the properties of the corresponding optimal scheduling policy. First, only a fraction of types buying low and high lead time qualities are strictly prioritized in the order of their delay costs; second, it need not be optimal to give strict priority to types who buy high lead time qualities ($w(c) < d$) over those buying the medium lead time as explained in the numerical example below; the statement of Lemma 3.1 needs to be appropriately modified.

3. The optimal number of high and low lead time qualities with pooling is of course highly dependent on the specific properties of the delay cost distribution $f$, the parameters of the value-delay cost function $d$ and $v$, the market size $\Lambda$ and capacity $\mu$. The results of Section 3.4.2 do generate some insights into the solution properties under nonmonotone $f_l$ and $f_h$. For example, Proposition 8 implies that the optimal pooling structure depends only on the properties of $f_l$ if capacity $\mu \leq 1/d$ and only on the properties of $f_h$ if capacity $\mu \geq \mu_H$. In either case there is no interaction between low and high lead time qualities and the optimal pooling structure can be readily computed (see Katta and Sethuraman, 2005, for such a pooling algorithm.) For a given set of types served, the optimal number of service classes with pooling within each segment is closely related to the number of modes/valleys of the appropriate virtual delay cost function.

**Proposition 10. [Properties of Mean Virtual Delay Costs]**

Suppose that the lead times are specified by the function $w(c)$. Define the mean virtual
delay cost functions:

\[
\overline{f}_l(c) = \frac{\int_{c_{\min}}^{c_{l}} f_l(x) f(x) \cdot 1 \{w(x) = c\} \, dx}{\int_{c_{\min}}^{c_{l}} f(x) \cdot 1 \{w(x) = c\} \, dx}, \quad \forall c \in C_l, \text{ if } c_l > c_{\min} \tag{3.73}
\]

\[
\overline{f}_h(c) = \frac{\int_{c_{h}}^{c_{\max}} f_h(x) f(x) \cdot 1 \{w(x) = c\} \, dx}{\int_{c_{h}}^{c_{\max}} f(x) \cdot 1 \{w(x) = c\} \, dx}, \quad \forall c \in C_h, \text{ if } c_h < c_{\max}. \tag{3.74}
\]

If the lead times \(w(c)\) are optimal then \(\overline{f}_l\) and \(\overline{f}_h\) satisfy the following properties:

1. They are positive, increasing and continuous on their respective domains.

2. They are increasing over all types that buy a low or high lead time quality:

\[
\text{if } c_l > c_{\min} \text{ and } c_h < c_{\max} \text{ then } \overline{f}_l(c_l) \leq \overline{f}_h(c_h), \tag{3.75}
\]

where \(\overline{f}_l(c_l) = \overline{f}_h(c_h)\) if there is a positive arrival rate to the medium lead time class \(\lambda_m > 0\).

Proof of Proposition 10

Follows by perturbation arguments similar to those used to prove Lemma 3.

Unlike a types’ individual virtual delay cost, its mean virtual delay cost depends not only on whether it receives low or high lead time quality, but also on whether it is pooled together with other types who buy the same lead time, and if so, how.

**Numerical example.** Figure 3.7 shows how the optimal segmentation and lead time menu vary with capacity, for the following bimodal delay cost distribution. It is the mixture of two truncated Normal distributions on the type domain \([c_{\min}, c_{\max}] = [1, 2]\) with mean and standard deviation \((\eta_1, \sigma_1) = (1.5, 0.1)\) and \((\eta_2, \sigma_2) = (1.9, 0.1)\). The weights \(w_1 \approx 0.54\) and \(w_2 \approx 0.46\) are chosen such that the p.d.f. has equal mass at each of the modes. The corresponding virtual delay cost functions \(f_l\) and \(f_h\) are each decreasing only on a single
Figure 3.7: Optimal Segmentation and Lead Times as a Function of $\mu$: Nonmonotone Virtual Delay Cost Functions

$\Lambda = 0.5$, $d = 1$, $v = 3$. Truncated normal delay cost distribution $f(c)$ with $(\eta_1, \sigma_1) = (1.5, 0.1)$ and $(\eta_2, \sigma_2) = (1.9, 0.1)$.

interval located, respectively, to the right and to the left of $c = 1.7$. which is where $f$ has a local minimum. For each capacity level the menu is computed with an algorithm that builds on the conditions of Proposition 10. As the Figure shows, some structural properties established by Lemma 3 and Proposition 8 continue to hold. The most notable difference is that pooling is optimal not only for the medium lead time, but at certain capacity levels also for the low or high lead time qualities. At most two classes pool customers at any capacity level, which follows from the fact that the delay cost distribution is bimodal. Another notable difference is that it need not be optimal to give strict priority to types who buy high lead time qualities ($w(c) < d$) over those buying the medium lead time. In particular, at $\mu = 1.67$ there are two neighboring pools of types that buy different lead times, one equal to and the other smaller than $d$; types buying the latter do not receive strict priority over those buying $d$. The alternative of giving types with $w < d$ strict priority over those with $w = d$ would require reducing the arrival rate of types buying lead time $< d$ and yield less revenue.
3.6.2 Nonlinear Value-Delay Cost Relationship $V(c)$

The value-delay cost relationship $V(c) = v + c \cdot d$ in our model has the distinctive feature that the ranking of types’ net willingness to pay for a given lead time $t$, given by $V(c) - ct = v + c(d - t)$, depends on the lead time: patient-low valuation types have higher willingness-to-pay for long lead times $t > d$ than impatient high-valuation types, and vice versa for short lead times $t < d$. This model feature plausibly describes situations where customers with high time-sensitivity get very little utility from a product or service if their delay exceeds a certain threshold. For example, a customer may be willing to pay a lot for overnight delivery of a time-critical shipment but only very little if it takes two or three days. By contrast a more patient customer with a small budget has much lower willingness-to-pay for such speedy delivery but is willing to pay more than impatient customers for delivery in several business days. In markets for fashion products customers that are early-adopters can be viewed as willing to pay a lot so long as they receive the product quickly and before others, but because they value the uniqueness of the product their willingness to pay for long delivery times may be below that of more patient customers.

Our model has the restriction that all customers value the medium lead time $d$ equally; in other words the lead time threshold at which early-adopters cease to be higher value is fixed. However, several properties of the optimal segmentation and lead times under this model generalize quite naturally if one relaxes the assumption that $V(c)$ is affine. The following facts are helpful in visualizing how our results generalize to such cases: first, given any menu the minimum full price is concave and increasing in the type; second, types only purchase if their full price is (weakly) lower than their valuation. Therefore, if $V(c)$ is convex increasing, then the optimal segmentation structure does not change, but when pooling is optimal the corresponding medium lead time is to be determined endogenously. If $V(c)$ is concave increasing in the type $c$, meaning that more impatient types have a disproportionately lower valuation, then the optimal segmentation structure depends on the relative curvatures of $V(c)$ and the (concave) full price function $p(c) + cw(c)$. If the curvature of $V(c)$ is slight
then structure is similar to that for convex $V(c)$, except that there can be no intermediate pooled segment unless all customers are admitted. If the curvature of $V(c)$ is significant, then only one contiguous segment of types can be served in equilibrium. Notably, in this case, an intermediate segment of customers may be admitted while both low and high types are excluded. Finally, if the curvature of $V(c)$ is sufficiently variable, it may be optimal to exclude types in several disjoint intervals.

### 3.6.3 Finite length menu

The optimal menu presented above will consist of an infinite number of distinct service classes most of which correspond to a single customer type. This is the case both for welfare and revenue maximizing menus when the delay sensitivities are continuously distributed. In practice, menus are generally of discrete length (e.g. 1 day, 2 day and 1 week delivery) and it is worth asking what implications the results presented above have on such a menu.

First, Proposition 6 continues to hold for a discrete menu. This implies the presence of at most three distinct customer regions of customer admission, delay decreasing with type, prices increasing with type and two sets of virtual delay cost functions. Since delay increases with type, any positive utility service class in a finite length menu will admit a contiguous set of customers.

Clearly, in a finite length menu, neighbouring customers must be pooled into common service classes. Our results indicating which customers should be pooled together also carries over. Proposition 7 indicates where, for particular customer admissions, customers should be strictly prioritized over lower sensitivity customers and where they should be pooled with neighbouring lower sensitivity customers. These results carry over to the finite menu. (i) A service class $C_w$ should be fully prioritized over lower priority service classes unless Proposition 7 indicates that some customers in $C_w$ would be pooled. (ii) Adding a strategically delayed or intermediate service pool to a finite length menu is revenue improving if that service class was predicted for the same admission profile by Proposition 7.

The globally optimal menu with a particular number of entries remains a difficult com-
binatorial problem. However, the pair of virtual delay cost functions will continue to play a large role in an optimal menu under such constraints.

3.7 Conclusion

This paper studies the optimal design of price-lead time menus to maximize revenues from heterogeneous time-sensitive customers with private information about their price and time-sensitivity. We believe the results provide insight into a broad range of situations a capacitated service provider may encounter. In order to gain these insights there are certain concessions that were made which include the requirement of an infinite number of service classes, affine relationship between delay sensitivity and base-value, and linear delay costs. The first two of these points are discussed in this chapter as extensions in Section 3.6 and many results carry over. Linear delay costs places limitations on where this model is applicable. For instance, a customer waiting for a video game to arrive via courier is likely to experience fairly constant delay costs. The same customer using the same courier to forward a credit card payment with a deadline will have a very different delay cost function. In welfare maximizing contexts, Van Mieghem (1995, 2000) show that many basic principles carry over though the optimal ordering of types may change as time in the queue progresses. Similar results may be expected for a revenue maximizing provider resulting in lower types exerting externalities on higher types.

To summarize our major findings and put them in context we remind the reader of which strategies raise revenue over a welfare maximizing strictly prioritized menu and how these strategies depend on capacity and market characteristics:

Service pooling may be optimal either within the high or low quality segments or in the intermediate segment. Within segment pooling occurs as a result of non-monotonicities within $f_l$ and $f_h$ which is often associated with multimodal customer distributions. Analogous phenomena have been documented in the economics literature (c.f. Guesnerie and Laffont, 1984). With the exception of low capacities where admission is low and the segment
of customers with non-monotone virtual costs may be excluded, within segment pooling is not associated with a particular capacity.

Between segment service pooling is optimal when non-monotonicity of virtual costs would otherwise occur between at the boundaries of the high and low quality segments. Since a customer type’s high quality virtual costs are smaller than their low quality counterparts, the distribution of customers does not play a large role in this type of service pooling. The strategy however relies on admission of high and low quality customers which is associated with intermediate capacities. The strategy is often accompanied by exclusion of an intermediate segment of customers.

Finally, strategic delay occurs when customers with negative high quality virtual costs could be admitted into the high quality segment. Strategic delay is associated with the existence of sufficiently patient customers for the negative external component of their high quality virtual costs to dominate resulting in negative virtual costs. These customers can only be admitted into the high quality segment at very high capacities.

We also note that when the provider is restricted to work conserving strictly prioritized schedules, it may be optimal at high capacity levels to strategically exclude low value customers which would otherwise be served and strategically delayed under the unrestricted optimal menu.

Finally, in addition to these central results, this chapter discusses how it may be beneficial to exclude an intermediate segment customers at intermediate capacities which to the best of our knowledge has not been previously documented. This exclusion would occur for both welfare maximizing, revenue maximizing and conditionally optimal menus. The exclusion requires both the capacity and congestion featured in our model which leads to insufficient capacity to profitably serve all customers while leaving residual capacity to serve low quality customers.
3.A Proof of Proposition 7

The optimal customer segmentation and lead time menu are obtained by solving Problem 3 for fixed \( \lambda \). Refer to (3.45)-(3.50) in Section 3.4.1.

**Part 1.** If \( \mu^{-1} \geq d \) then (3.49) only holds if \( \lambda_l = \lambda \); hence \( \lambda^*_l (\lambda) = \lambda \) and \( \lambda^*_m (\lambda) = \lambda^*_h (\lambda) = 0 \).

**Part 2.** Suppose that \( \mu^{-1} < d \). We first reformulate Problem 3 for analytical convenience.

**Definitions and problem formulation.** Let \( \lambda_{mh} \triangleq \lambda_m + \lambda_h \) be the aggregate rate for the \( m \) class (lead time \( d \)) and \( h \) classes (lead times \( < d \)). For fixed \( \lambda \) write the revenue (3.45) as

\[
\Pi(\lambda_{mh}, \lambda_h) \triangleq \lambda v - \Lambda \int_{\epsilon_{\text{min}}}^{c_t(\lambda-\lambda_{mh})} f(x)f_l(x) \left( \frac{\mu}{\mu - [\lambda - \Lambda F(x)]^2} - d \right) dx
\]

\[
+ \Lambda \int_{\epsilon_{\text{max}}}^{c_h(\lambda_h)} f(x)f_h(x) \left( d - \frac{\mu}{\mu - \Lambda F(x)} \right) dx,
\]

where \( \lambda_l = \lambda - \lambda_{mh} \), \( c_t(x) \triangleq F^{-1}(x/\Lambda) \) with \( c_t' > 0 \), \( c_h(x) \triangleq F^{-1}(x/\Lambda) \) with \( c_h' < 0 \); and \( f_l, f_l', f_h > 0 \).

By Lemma 3 types in \( h \) or \( l \) classes are prioritized by their delay costs, where \( h \) classes have lead times \( < d \) and get priority over the \( m \) class (lead time \( = d \)) which gets priority over \( l \) classes (lead times \( > d \)). Let \( \lambda_P \triangleq \mu - \sqrt{\mu/d} \): it is the maximum feasible rate for \( h \) classes; at this rate the maximum lead time is \( d \) under work conserving priority service: \( \mu/ (\mu - \lambda_P) = d \). Let \( \lambda_F \triangleq \mu - 1/d \): it is the maximum feasible aggregate rate for \( m \) and \( h \) classes; at this rate the lead time is \( d \) under work conserving FIFO service: \( (\mu - \lambda_F)^{-1} = d \), where \( \mu d > 1 \) implies \( 0 < \lambda_P < \lambda_F \).

Let \( \lambda_h (\lambda_{mh}) \) be the maximum feasible rate of \( h \) classes as a function of the total rate \( \lambda_{mh} \) to \( m \) and \( h \) classes, such that the medium lead time \( d \) is achievable i.e., (3.49) holds.

\[
\lambda_h (\lambda_{mh}) \triangleq \min \left( \lambda_{mh}, \mu - \frac{\mu/d}{(\mu - \lambda_{mh})} \right) = \begin{cases} 
\lambda_{mh}, & \text{if } \lambda_{mh} \in [0, \lambda_P] \\
\mu - \frac{\mu/d}{\mu - \lambda_{mh}} < \lambda_{mh}, & \text{if } \lambda_{mh} \in [\lambda_P, \lambda_F] 
\end{cases}
\]

(3.77)
For $\lambda_{mh} \leq \lambda_P$ all classes can have lead times $< d$, so $\lambda_h(\lambda_{mh}) = \lambda_{mh}$ increases on $[0, \lambda_P]$ with $\lambda_h(\lambda_P) = \lambda_P$. For $\lambda_{mh} > \lambda_P$ only a portion $\lambda_h(\lambda_{mh}) < \lambda_{mh}$ can be allocated to $h$ classes, and $\lambda_h(\lambda_{mh})$ decreases on $[\lambda_P, \lambda_F]$, with $\lambda_h(\lambda_F) = 0$. Having $\lambda_{mh} > \lambda_F$ violates (3.49).

Constraint (3.49) in Problem 3 holds if and only if $\lambda_{mh} \leq \lambda_F$ and $\lambda_h \leq \lambda_h(\lambda_{mh})$; constraint (3.50) holds if and only if $\lambda_P \leq \lambda_{mh} < \lambda$ or $\lambda_{mh} = \lambda \leq \lambda_P$. Problem 3 for fixed $\lambda$ is thus equivalent to

$$\max_{\lambda_{mh}, \lambda_h} = \Pi(\lambda_{mh}, \lambda_h)$$

subject to

$$\min(\lambda, \lambda_P) \leq \lambda_{mh} \leq \min(\lambda, \lambda_F),$$

$$0 \leq \lambda_h \leq \lambda_h(\lambda_{mh}).$$

The lower bound of (3.79) ensures a work conserving policy if $\lambda_l > 0$: $\lambda_{mh} < \min(\lambda, \lambda_P)$ is not feasible$^3$. It implies strategic delay for $l$ classes ($d > \mu / (\mu - \lambda_{mh})^2$) which is suboptimal by Lemma 3.1(c).

**Overview of proof.** We prove Parts 2(a)-(c) by solving (3.78)-(3.80) in the following three steps.

1. **Step 1.** For fixed $\lambda_{mh}$, we characterize the optimal rate $\lambda_h(\lambda_{mh})$.

2. **Step 2.** For fixed $\lambda$, we characterize the optimal rates $\lambda^*_mh(\lambda)$ and $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_mh(\lambda))$, which determine the optimal segmentation. The conditions for $\lambda^*_mh(\lambda)$ and $\lambda^*_h(\lambda)$ are stated in terms of the virtual delay costs of the corresponding marginal types as summarized in (3.84) and (3.88).

3. **Step 3.** We translate the optimality conditions of Step 2 for $\lambda^*_mh(\lambda)$ and $\lambda^*_h(\lambda)$ into conditions on $\lambda$ and $\Lambda$. These conditions imply the optimal segmentation structure specified in Parts 2(a)-(c) of the Proposition. We organize this analysis into Lemmas 5-9 as further detailed below.

**Step 1. Optimal $\lambda_h$ for fixed $\lambda_{mh}$.** Fix $\lambda_{mh} \in (0, \min(\lambda, \lambda_F)]$ and let $\lambda_h(\lambda_{mh})$ define

---

$^3$We say ‘feasible’ with respect to (3.79)-(3.80), which are implied by the *optimal* lead times in Lemma 3.1. Choosing $\lambda_{mh}$ and/or $\lambda_h$ that violate these constraints may be feasible for a *suboptimal* lead time function.
arg \{\max_{\lambda_h} \Pi(\lambda_{mh}, \lambda_h) \text{ s.t. } 0 \leq \lambda_h \leq \overline{x}_h (\lambda_{mh})\} be the corresponding optimal \lambda_h. From (3.76) we have

\[
\frac{\partial \Pi(\lambda_{mh}, \lambda_h)}{\partial \lambda_h} = f_h(c_h(\lambda_h)) \left( d - \frac{\mu}{(\mu - \lambda_h)^2} \right), \text{ for } \lambda_h \leq \overline{x}_h (\lambda_{mh}),
\]

where \(\Lambda f(c_h(x))c_h'(x) = -1\). The multiplier of \(f_h(c_h(\lambda_h))\) is nonnegative since \(\overline{x}_h (\lambda_{mh}) \leq \lambda_P\) by (3.77). Since \(f_h(c_h(0)) = c_{\text{max}} > 0\) and \(f_h c_h' < 0\), the maximizer \(\lambda_h (\lambda_{mh})\) is unique and satisfies

\[
\lambda_h (\lambda_{mh}) = \begin{cases} 
\overline{x}_h (\lambda_{mh}), & \text{if } f_h(c_h(\overline{x}_h (\lambda_{mh}))) \geq 0 \\
\lambda_0 \triangleq \Lambda F (f_h^{-1}(0)) < \overline{x}_h (\lambda_{mh}), & \text{if } f_h(c_h(\overline{x}_h (\lambda_{mh}))) < 0
\end{cases}.
\]

If \(f_h(c_h(\overline{x}_h (\lambda_{mh}))) \geq 0\) then it is optimal to sell the maximum possible rate \(\overline{x}_h (\lambda_{mh})\) to \(h\) classes and \(\lambda_{mh} - \overline{x}_h (\lambda_{mh}) \geq 0\) to the medium lead time class. This policy is work conserving.

If \(f_h(c_h(\overline{x}_h (\lambda_{mh}))) < 0\) then it is optimal to sell \(\lambda_0\) to \(h\) classes, less than the maximum possible rate \(\overline{x}_h (\lambda_{mh})\), and \(\lambda_{mh} - \lambda_0 > 0\) to the medium lead time \(d\). At \(\lambda_0\) defined in (3.82) the virtual delay cost of the corresponding marginal type is zero: \(f_h(c_h(\lambda_0)) = f_h(F^{-1}(\lambda_0/\Lambda)) = 0\). This policy is not work conserving: the lead time \(d\) involves strategic delay since \(\lambda_0 < \overline{x}_h (\lambda_{mh})\):

\[
\frac{\mu}{(\mu - \lambda_{mh})(\mu - \lambda_0)} < \frac{\mu}{(\mu - \lambda_{mh})(\mu - \overline{x}_h (\lambda_{mh}))} \leq d.
\]

**Step 2. Optimal segmentation for fixed \(\lambda\): optimal \(\lambda_{mh}^*(\lambda)\) and \(\lambda^*_l (\lambda)\).** Let \(\lambda_{mh}^*(\lambda) \triangleq \arg\{\max_{\lambda_{mh}} \Pi(\lambda_{mh}, \lambda_h (\lambda_{mh})) \text{ s.t. } \min(\lambda, \lambda_P) \leq \lambda_{mh} \leq \min(\lambda, \lambda_F)\}\) be the optimal \(\lambda_{mh}\) for fixed \(\lambda\). Write \(\lambda_h^*(\lambda), \lambda_m^*(\lambda)\) and \(\lambda_l^*(\lambda)\) for the optimal rates of \(h, m\) and \(l\) classes for fixed \(\lambda\), where \(\lambda_h^*(\lambda) = \lambda_h (\lambda_{mh}^*(\lambda))\) by (3.82), whereas \(\lambda_m^*(\lambda) = \lambda_{mh}^*(\lambda) - \lambda_h^*(\lambda)\) and \(\lambda_l^*(\lambda) = \lambda - \lambda_{mh}^*(\lambda)\).

**Optimal segmentation for \(\lambda \leq \lambda_P\).** For \(\lambda \leq \lambda_P\) it is not optimal to sell \(l\) classes: (3.79) requires \(\lambda_{mh} = \lambda\), so the maximizer satisfies \(\lambda_{mh}^*(\lambda) = \lambda\). By (3.77) the entire \(\lambda\) can be sold to \(h\) classes, so \(\overline{x}_h (\lambda_{mh}^*(\lambda)) = \overline{x}_h (\lambda) = \lambda\). Therefore (3.82) yields the following optimal segmentations, where \(m_{sd}\) denotes that the medium lead time involves strategic delay.
Chapter 3. Revenue Maximizing Lead-Time Pricing

<table>
<thead>
<tr>
<th>segments</th>
<th>virtual delay cost condition</th>
<th>( \lambda_{mh}^* (\lambda) )</th>
<th>( \lambda_h^* (\lambda) )</th>
<th>( \lambda_m^* (\lambda) )</th>
<th>( \lambda_l^* (\lambda) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((h))</td>
<td>( f_h (c_h (\lambda_h (\lambda))) = f_h (c_h (\lambda)) \geq 0 )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>((h, m_{sd}))</td>
<td>( f_h (c_h (\lambda_h (\lambda))) = f_h (c_h (\lambda)) &lt; 0 )</td>
<td>( \lambda )</td>
<td>( \lambda_0 )</td>
<td>( \lambda - \lambda_0 &gt; 0 )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>

(3.84)

**Optimal segmentation for** \( \lambda > \lambda_p \). (3.79) requires \( \lambda_{mh} \in [\lambda_p, \min (\lambda, \lambda_F)] \). From (3.76):

\[
\frac{d\Pi(\lambda_{mh}, \lambda_h (\lambda_{mh}))}{d\lambda_{mh}} = \frac{\partial\Pi(\lambda_{mh}, \lambda_h (\lambda_{mh}))}{\partial\lambda_{mh}} + \frac{\partial\Pi(\lambda_{mh}, \lambda_h (\lambda_{mh}))}{\partial\lambda_h} \cdot \lambda_h' (\lambda_{mh}) =
\]

\[
= f_l (c_l (\lambda - \lambda_{mh})) \left( \frac{\mu}{(\mu - \lambda_{mh})^2} - d \right) + f_h (c_h (\lambda_h (\lambda_{mh}))) \left( d - \frac{\mu}{(\mu - \lambda_h (\lambda_{mh}))^2} \right) \lambda_h' (\lambda_{mh}),
\]

(3.85)

where \( \Lambda f(c_l (x))c_l' (x) = 1 \) and \( \lambda_h (\lambda_{mh}) \) satisfies (3.82). If \( f_h (c_h (\lambda_h (\lambda_{mh}))) < 0 \) then \( \lambda_h' (\lambda_{mh}) = 0 \). If \( f_h (c_h (\lambda_h (\lambda_{mh}))) \geq 0 \), then \( \lambda_h (\lambda_{mh}) = \lambda_h (\lambda_{mh}) = \mu - \mu/d (\mu - \lambda_{mh}) \) by (3.77) and (3.82); in this case \( \lambda_h' (\lambda_{mh}) = -\mu/d (\mu - \lambda_{mh}) \) and \( \mu/ (\mu - \lambda_h (\lambda_{mh}))^2 = d^2 (\mu - \lambda_{mh})^2/\mu \). Substituting into (3.85) yields

\[
\frac{d\Pi(\lambda_{mh}, \lambda_h (\lambda_{mh}))}{d\lambda_{mh}} = \begin{cases} 
  f_l (c_l (\lambda - \lambda_{mh})) \left( \frac{\mu}{(\mu - \lambda_{mh})^2} - d \right), & \text{if } f_h (c_h (\lambda_h (\lambda_{mh}))) < 0 \\
  g (\lambda, \lambda_{mh}) \left( \frac{\mu}{(\mu - \lambda_{mh})^2} - d \right), & \text{if } f_h (c_h (\lambda_h (\lambda_{mh}))) \geq 0
\end{cases},
\]

(3.86)

for \( \lambda_{mh} \in [\lambda_p, \min (\lambda, \lambda_F)] \), where

\[
g (\lambda, \lambda_{mh}) \triangleq f_l (c_l (\lambda - \lambda_{mh})) - f_h (c_h (\lambda_h (\lambda_{mh}))).
\]

(3.87)

The function \( g (\lambda, \lambda_{mh}) \) is important in determining the solution. It measures the difference between the virtual delay costs of the marginal types \( c_l \) and \( c_h \) as a function of \( \lambda \) and \( \lambda_{mh} \), when allocating the corresponding maximum feasible rate \( \lambda_h = \lambda_h (\lambda_{mh}) \) to \( h \) classes and \( \lambda_l = \lambda - \lambda_{mh} \) to \( l \) classes.

The sign of the revenue derivative in (3.86) only depends on \( f_h (c_h (\lambda_h (\lambda_{mh}))) \) and \( g (\lambda, \lambda_{mh}) \), because \( f_l > 0 \) and the common factor in both cases of (3.86) is zero at \( \lambda_{mh} = \lambda_p \) and positive for \( \lambda_{mh} > \lambda_p \). The maximizer \( \lambda_{mh}^* (\lambda) \) is unique since the following holds for \( \lambda_{mh} \in [\lambda_p, \min (\lambda, \lambda_F)] \).
(i) \( f_h (c_h (\bar{\lambda}_h (\lambda_{mh}))) \) increases in \( \lambda_{mh} \) since \( \bar{\lambda}_h' (\lambda_{mh}) < 0 \) by (3.77) and \( f_h' c_h' < 0 \).

(ii) \( g (\lambda, \lambda_{mh}) \) decreases in \( \lambda_{mh} \) since \( f_l' c_l' > 0 \), and by (i).

By (3.79)-(3.80) the threshold \( \lambda_P \) is the smallest feasible \( \lambda_{mh} \) and the largest feasible \( \lambda_h \).

The choice of \( \lambda_{mh} \) determines the maximum feasible \( h \) rate \( \bar{\lambda}_h (\lambda_{mh}) \), and the \( l \) rate \( \lambda_l = \lambda - \lambda_{mh} \). The rates of both segments decrease in \( \lambda_{mh} \), and their marginal types and virtual delay costs move apart: \( c_h (\bar{\lambda}_h (\lambda_{mh})) \) and \( f_h (c_h (\bar{\lambda}_h (\lambda_{mh}))) \) increase while \( c_l (\lambda - \lambda_{mh}) \) and \( f_l (c_l (\lambda - \lambda_{mh})) \) decrease in \( \lambda_{mh} \). Since \( \bar{\lambda}_h (\lambda_{mh}) = 0 \) if \( \lambda_{mh} = \lambda_F \) and \( \lambda_l = 0 \) if \( \lambda = \lambda_{mh} \), \( \min (\lambda_F, \lambda) \) is the largest feasible \( \lambda_{mh} \).

Properties (i)–(ii) and (3.86) determine the maximizer \( \lambda^*_{mh} (\lambda) \), and (3.82) determines \( \lambda^*_h (\lambda) = \lambda_h (\lambda^*_{mh} (\lambda)) \). The optimal segmentations and the corresponding conditions are summarized in (3.88). The optimal segmentation sets \( \lambda_{mh} \) and \( \lambda_h \) to minimize the virtual delay cost of types served, subject to the constraints (3.79)-(3.80). Since \( f_l \) and \( f_h \) are increasing, a segmentation is optimal if and only if it satisfies the following criteria. It minimizes \( \lambda_{mh} \in [\lambda_P, \min (\lambda_F, \lambda)] \), and maximizes \( \lambda_h \leq \bar{\lambda}_h (\lambda_{mh}) \), subject to two conditions.

1. The marginal \( h \) type’s virtual delay cost is nonnegative: \( f_h (c_h (\lambda_h)) \geq 0 \).
2. If both \( h \) and \( l \) are offered \( (\lambda_h, \lambda - \lambda_{mh} > 0) \) then the virtual delay cost of the marginal \( h \) type (with the shorter lead time) is higher: \( f_h (c_h (\lambda_h)) \geq f_l (c_l (\lambda - \lambda_{mh})) \).

We review these criteria for the cases of (3.88) in the order \((h, l) - (h, m, l) - (m, l) - (h, m) - (h, m_{sd})\).

<table>
<thead>
<tr>
<th>segments</th>
<th>conditions (other than ( \lambda &gt; \lambda_P ))</th>
<th>demand rates for each segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (h, m_{sd}) )</td>
<td>( \lambda &lt; \lambda_F ) ( f_h (c_h (\bar{\lambda}_h (\lambda))) &lt; 0 )</td>
<td>( \lambda^*_{mh} (\lambda) )</td>
</tr>
<tr>
<td>( (h, m) )</td>
<td>( \lambda &lt; \lambda_F ) ( f_h (c_h (\bar{\lambda}_h (\lambda))) \geq 0 ), ( g (\lambda, \lambda) \geq 0 )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>( (h, l) )</td>
<td>( \lambda &gt; \lambda_F ) ( g (\lambda, \lambda_P) \leq 0 )</td>
<td>( \lambda_P )</td>
</tr>
<tr>
<td>( (h, m, l) )</td>
<td>( \lambda &gt; \lambda_F ) ( g (\lambda, \lambda_P) &gt; 0 ), ( g (\lambda, \min (\lambda, \lambda_F)) )</td>
<td>( \lambda^*_{mh} (\lambda) )</td>
</tr>
<tr>
<td>( (m, l) )</td>
<td>( \lambda &gt; \lambda_F ) ( g (\lambda, \lambda_F) \geq 0 )</td>
<td>( \lambda_F )</td>
</tr>
</tbody>
</table>

(3.88)

Segmentation \( (h, l) \): if \( g (\lambda, \lambda_P) \leq 0 \), conditions (1) – (2) hold at \( \lambda_{mh} = \lambda_P \), which yields
maximum allocations to $h$ and $l$, using the entire $\lambda$. Segmentation $(h, m, l)$: if $g(\lambda, \lambda_P) > 0 > g(\lambda, \min(\lambda, \lambda_F))$, then $\lambda_{mh} = \lambda_P$ violates (2) whereas $\lambda_{mh} = \min(\lambda, \lambda_F)$ satisfies (2) but does not minimize $\lambda_{mh}$. In this case increase $\lambda_{mh}$ to the point where the marginal types’ virtual delay costs are equal: $g(\lambda, \lambda_{mh}) = 0$ for $\lambda_{mh} = \lambda_{mh}^*$. This reduces the $l$ and $h$ rates and increases the rate $\lambda_m$.

Segmentations $(m, l), (h, m)$ and $(h, m_{sd})$: in these cases $g(\lambda, \lambda_{mh}) > 0$ for every $\lambda_{mh} < \min(\lambda, \lambda_F)$, so condition (2) is violated for every segmentation that sells both $h$ and $l$ classes; note that $f_h(c_h(\lambda_h)) < 0$, the condition for $(h, m_{sd})$, implies $g(\lambda, \lambda) > 0$. In each case condition (2) is met by not offering both $l$ and $h$ classes: set $\lambda_{mh} = \lambda_{mh}^*$.

If $\lambda > \lambda_F$, no $h$ classes are sold: allocate $\lambda_F$ to the $m$ class and the rest to $l$ classes. If $\lambda < \lambda_F$, no $l$ classes are sold: allocate $\lambda$ to $m$ and $h$ classes, then maximize $\lambda_h \leq \lambda_h^*(\lambda)$ subject to (1). If all of $\lambda_h(\lambda)$ can be sold to types with nonnegative $f_h$ this yields $(h, m)$. If not, then $(h, m_{sd})$, with strategic delay is optimal. In this case sell $h$ classes to $\lambda_0 < \lambda_h(\lambda)$, where $f_h(c_h(\lambda_0)) = 0$, and the $m$ class (lead time = $d$) to the remaining $\lambda - \lambda_0$ consisting of types with negative $f_h$. Since $\lambda_0 < \lambda_h(\lambda)$ the lead time $d$ involves strategic delay: it exceeds the minimum achievable lead time given $\lambda_0$ have higher priority – see also (3.82).

Step 3. Optimal segmentations depending on $\lambda$ and $\Lambda$. The optimality conditions of Step 2, (3.84) for $\lambda \leq \lambda_F$ and (3.88) for $\lambda > \lambda_P$, are stated in terms of the signs of the virtual delay cost $f_h(c_h(\lambda_h(\lambda)))$ and of the virtual delay cost difference $g(\lambda, \lambda_{mh})$ for $\lambda_{mh} = \lambda_P$ and $\lambda_{mh} = \min(\lambda, \lambda_F)$. Next we translate these conditions into conditions on $\lambda$ and $\Lambda$. We proceed as follows. Lemmas 5-7 characterize the signs of $g(\lambda, \lambda_P)$ and $g(\lambda, \min(\lambda, \lambda_F))$ depending on $\lambda$ and $\Lambda$. Together with (3.84) and (3.88) they imply Parts 2(a)-(b), in which $f_h(c_{\min}) \geq 0$. Lemmas 8-9 characterize the sign of $f_h(c_h(\lambda_h(\lambda)))$; these additional results imply Part 2(c) in which $f_h(c_{\min}) < 0$.

Lemma 5. Fix $\mu > d^{-1}$ and $\Lambda > \lambda_F$. Consider $g(\lambda, \lambda_{mh})$ for $\lambda_{mh} = \lambda_P$ and $\lambda_{mh} = \min(\lambda, \lambda_F)$.

1. The virtual delay cost difference $g(\lambda, \lambda_P)$, where $\lambda_h = \lambda_P$, $\lambda_m = 0$, increases in $\lambda \geq \lambda_P$. 
where

\[ g(\lambda, \lambda_P) = f_l(c_l(\lambda - \lambda_P)) - f_h(c_h(\lambda_P)). \]  
(3.89)

2. The virtual delay cost difference \( g(\lambda, \min(\lambda, \lambda_F)) \) varies as follows with \( \lambda \).

(a) It decreases in \( \lambda \in [\lambda_P, \min(\lambda, \Lambda)] \), where \( \lambda_h = \lambda_h(\lambda) \), \( \lambda_m = \lambda - \lambda_h(\lambda) > 0 \), and

\[ g(\lambda, \min(\lambda_F, \lambda)) = g(\lambda, \lambda) = c_{\min} - f_h(c_h(\lambda_h(\lambda))). \]  
(3.90)

(b) It increases in \( \lambda \in [\lambda_F, \min(\Lambda, \mu)] \), where \( \lambda_h = 0 \), \( \lambda_m = \lambda_F \), \( g(\lambda_F, \lambda) = c_{\min} - c_{\max} \) and

\[ g(\lambda, \min(\lambda_F, \lambda)) = g(\lambda, \lambda_F) = f_l(c_l(\lambda - \lambda_F)) - c_{\max}, \]  
(3.91)

**Proof.** By (3.87) we have \( g(\lambda, \lambda_{mh}) = f_l(c_l(\lambda - \lambda_{mh})) - f_h(c_h(\lambda_h(\lambda_{mh}))). \) For \( \lambda_{mh} = \lambda_P \) in (3.89) note that \( \lambda_h(\lambda_P) = \lambda_P \). For \( \lambda_{mh} = \lambda \) in (3.90) note that \( c_l(0) = c_{\min} \). For \( \lambda_{mh} = \lambda_F \) in (3.91) note that \( \lambda_h(\lambda_F) = 0 \) and \( c_h(0) = c_{\max} \).

Parts 1. and 2(b) follow since \( f'_l c'_l > 0 \) and the rate of \( l \) classes \( \lambda - \lambda_{mh} \) increases in \( \lambda \), while \( \lambda_h \) is fixed. Part 2(a) follows since \( f'_h c'_h < 0 \) and \( \lambda_h'(\lambda) < 0 \) for \( \lambda \in [\lambda_P, \lambda_F] \) by (3.77), while \( \lambda_l = 0 \).

\[ \square \]

Figure 3.8 illustrates Lemma 5. In both panels \( g(\lambda, \lambda_P) \) increases in \( \lambda \) (Part 1). At left \( \Lambda < \lambda_F \) and \( g(\lambda, \min(\lambda, \lambda_F)) \) decreases in \( \lambda \). At right \( \Lambda > \lambda_F \) and \( g(\lambda, \min(\lambda, \lambda_F)) \) is minimized and equal to \( c_{\min} - c_{\max} < 0 \) at \( \lambda = \lambda_F \), when all types are served FIFO (Part 2). In both panels \( g(\lambda, \lambda_P) > g(\lambda, \min(\lambda, \lambda_F)) \) for fixed \( \lambda > \lambda_P \): as explained in Step 2, \( g(\lambda, \lambda_{mh}) \) decreases in \( \lambda_{mh} \) for fixed \( \lambda \).

Increasing \( \Lambda \) has two effects on \( g(\lambda, \lambda_{mh}) \) for \( \lambda_m = \lambda_P \) and \( \lambda_m = \min(\lambda, \lambda_F) \). It increases the set of feasible \( \lambda \). It also decreases \( g(\lambda, \lambda_{mh}) \), see Figure 3.8: for fixed \( \lambda, \lambda_{mh} \) a larger market increases the virtual delay cost difference as it “pulls apart” the marginal types.
Figure 3.8: Virtual Delay Cost Differences $g(\lambda, \lambda_P)$ and $g(\lambda, \min(\lambda, \lambda_F))$ as Functions of $\lambda$

$c_l(\lambda - \lambda_{mh}) = F^{-1}((\lambda - \lambda_{mh})/\Lambda)$ and $c_h(\lambda_{mh}) = F^{-1}(\lambda_{mh}/\Lambda)$. To make this dependence explicit we henceforth write

$g(\lambda, \lambda_{mh}) = f_l(c_l(\lambda - \lambda_{mh}, \Lambda)) - f_h(c_h(\lambda_{mh}, \Lambda))$, for $\lambda \in [\lambda_P, \min(\mu, \Lambda)]$, $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$.

Lemma 6 and 7 characterize, respectively, the sign of $g(\lambda, \min(\lambda, \lambda_F), \Lambda)$ and $g(\lambda, \lambda_P, \Lambda)$.

\[\square\]

**Lemma 6.** Fix $\mu > d^{-1}$. Consider $g(\lambda, \min(\lambda, \lambda_F), \Lambda)$ as a function of $\lambda > \lambda_P$ and $\Lambda$.

where

$$g(\lambda, \min(\lambda, \lambda_F), \Lambda)=\begin{cases} g(\lambda, \lambda, \Lambda) = c_{\text{min}} - f_h \left( F^{-1} \left( \frac{\lambda}{\Lambda} \right) \right), & \lambda \in [\lambda_P, \min(\Lambda, \lambda_F)], \\ g(\lambda, \lambda_F, \Lambda) = f_l \left( F^{-1} \left( \frac{\lambda - \lambda_F}{\Lambda} \right) \right) - c_{\text{max}}, & \lambda \in [\lambda_F, \min(\Lambda, \mu)]. \end{cases}$$

(3.92)

1. For $\lambda \in [\lambda_P, \min(\Lambda, \lambda_F)]$ the sign of $g(\lambda, \min(\lambda, \lambda_F), \Lambda) = g(\lambda, \lambda, \Lambda)$ depends on the thresholds

$$\Lambda_2 \triangleq \frac{\mu}{2} \left( \frac{1}{z} + 1 \right) - \sqrt{\frac{\mu^2}{4} \left( \frac{1}{z} - 1 \right)^2 + \frac{\mu}{zd}}$$

and $\Lambda_3 \triangleq \frac{\lambda_P}{z}$, where $z = F \left( f_h^{-1}(c_{\text{min}}) \right) \in (0, 1)$

(3.93)

where $\Lambda_2 \in (\lambda_P, \lambda_F)$ satisfies $g(\Lambda_2, \Lambda_2, \Lambda_2) = 0$ and $\Lambda_3 > \Lambda_2$ satisfies $g(\lambda_P, \lambda_P, \Lambda_3) = 0.$
Chapter 3. Revenue Maximizing Lead-Time Pricing

128

(a) If $\Lambda \leq \Lambda_2$ then $g(\lambda, \lambda, \Lambda) \geq 0$ for $\lambda \in [\lambda_P, \Lambda]$, where $\Lambda_2 < \lambda_F$.

(b) If $\Lambda \in (\Lambda_2, \Lambda_3)$, then there is a demand rate threshold $\lambda_1 \in (\lambda_F, \lambda_F)$ such that

\[
g(\lambda, \lambda, \Lambda) = \begin{cases} 
0, & \text{if } \lambda = \lambda_1 \triangleq \mu - \frac{\mu/d}{\mu - \Lambda F(f_l^{-1}(c_{\min}))}, \\
< 0, & \text{if } \lambda \in (\lambda_1, \min(\lambda_F, \Lambda)].
\end{cases}
\]  

(c) If $\Lambda > \Lambda_3$, then $g(\lambda, \lambda, \Lambda) < 0$ for $\lambda \in [\lambda_P, \min(\lambda_F, \Lambda)]$.

2. For $\lambda \in [\lambda_F, \min(\Lambda, \mu)]$ the sign of $g(\lambda, \min(\lambda_F, \lambda), \Lambda) = g(\lambda, \lambda, \Lambda)$ depends on the thresholds

\[
\Lambda_{ml} \triangleq \frac{\lambda_F}{F(f_l^{-1}(c_{\max}))} > \lambda_F \text{ and } \Lambda_{m} \triangleq \frac{1}{dF(f_l^{-1}(c_{\max}))},
\]  

where $g(\Lambda_{ml}, \lambda_F, \Lambda_{ml}) = 0$ and $g(\mu, \lambda_F, \Lambda_{ml}) = 0$.

(a) If $F(f_l^{-1}(c_{\max})) \cdot d < \mu^{-1} < d$ then $\Lambda_{ml} < \mu < \Lambda_{ml}$. If $\Lambda \in [\Lambda_{ml}, \Lambda_{ml})$ then

\[
g(\lambda, \lambda_F, \Lambda) = \begin{cases} 
< 0, & \text{if } \lambda \in (\lambda_F, \lambda_3), \\
0, & \text{if } \lambda = \lambda_3 \triangleq \lambda_F + \Lambda F(f_l^{-1}(c_{\max})), \\
> 0, & \text{if } \lambda > \lambda_3.
\end{cases}
\]  

If $\Lambda \notin [\Lambda_{ml}, \Lambda_{ml})$ then $g(\lambda, \lambda_F, \Lambda) < 0$ for all feasible $\lambda > \lambda_F$.

(b) If $\mu^{-1} \leq F(f_l^{-1}(c_{\max})) \cdot d$ then $g(\lambda, \lambda_F, \Lambda) < 0$ for all feasible $\lambda > \lambda_F$.

**Proof.** By Lemma 5.2, for fixed $\Lambda$ the function $g(\lambda, \min(\lambda_F, \lambda), \Lambda)$ is decreasing in $\lambda \in [\lambda_P, \min(\Lambda, \lambda_F)]$, negative at $\lambda = \lambda_F$ and increasing in $\lambda \in [\lambda_F, \min(\Lambda, \mu)]$. Hence it has at most two roots in $\lambda$, one smaller than, the other larger than $\lambda_F$. The number of roots depends on the sign of $g(\lambda, \min(\lambda_F, \lambda), \Lambda)$ for $\lambda = \lambda_P$, for $\lambda = \min(\Lambda, \lambda_F)$ and for $\lambda = \min(\Lambda, \mu)$ when $\Lambda > \lambda_F$.

**Part 1.** The threshold $\Lambda_3$ determines the sign of $g(\lambda, \min(\lambda_F, \lambda), \Lambda)$ for $\lambda = \lambda_P$. It is the unique solution of $g(\lambda_P, \lambda_P, \Lambda) = c_{\min} - f_h(F^{-1}(\lambda_P/\Lambda)) = 0$ in $\Lambda \geq \lambda_P$. This follows because $g(\lambda_P, \lambda_P, \lambda) = c_{\min} - f_h(c_{\min}) > 0$ and since $g(\lambda_P, \lambda_P, \Lambda)$ decreases in $\Lambda$ with

\[
\lim_{\Lambda \to \infty} g(\lambda_P, \lambda_P, \Lambda) = c_{\min} - c_{\max} < 0.
\]

Therefore $g(\lambda_P, \lambda_P, \Lambda) > 0$ if $\Lambda < \Lambda_3$ and conversely for $\Lambda > \Lambda_3$. Solving for $\Lambda_3$ yields (3.93).

A
The threshold $\Lambda_2$ determines the sign of $g(\lambda, \min(\lambda_F, \lambda), \Lambda)$ for $\lambda = \min(\Lambda, \lambda_F)$. It is the unique solution of $g(\Lambda, \Lambda, \Lambda) = c_{\min} - f_h(\overline{F}^{-1}(\overline{X}_h(\Lambda)/\Lambda)) = 0$ in $\Lambda \in (\lambda_F, \Lambda_F)$. This follows since $g(\lambda_F, \lambda_F, \lambda_F) = c_{\min} - f_h(c_{\min}) > 0 > g(\lambda_F, \lambda_F, \lambda_F) = c_{\min} - c_{\max}$, and $g(\Lambda, \Lambda, \Lambda)$ decreases in $\Lambda \in [\lambda_F, \Lambda_F]$ as the fraction $\overline{X}_h(\Lambda)/\Lambda$ allocated to $h$ classes decreases in $\Lambda$. It follows that $g(\Lambda, \Lambda, \Lambda) \geq 0$ if $\Lambda \leq \Lambda_2$ and $g(\min(\Lambda, \lambda_F), \min(\Lambda, \lambda_F), \Lambda) < 0$ if $\Lambda > \Lambda_2$. Solving for $\Lambda_2$ yields (3.93).

To show that $\Lambda_2 < \Lambda_3$, first note that $g(\lambda_F, \lambda_F, \Lambda_2) > g(\Lambda_2, \Lambda_2, \Lambda_2) = 0$ by Lemma 5.2(a) since $\lambda_F < \Lambda_2 < \Lambda_F$. Since $g(\lambda_F, \lambda_F, \Lambda)$ decreases in $\Lambda$ it follows that $\Lambda_2 < \Lambda_3$.

Parts 1(a)-(c). The above analysis implies that for $\lambda \in [\lambda_F, \min(\Lambda, \lambda_F)]$ the function $g(\lambda, \lambda, \Lambda)$ is non-negative if $\Lambda \leq \Lambda_2$, as in 1(a), and strictly negative if $\Lambda > \Lambda_3$, as in 1(c). If $\Lambda \in (\Lambda_2, \Lambda_3)$ then $g(\lambda, \lambda, \Lambda)$ has an unique root $\lambda_1 \in [\lambda_F, \min(\Lambda, \lambda_F)]$ since $g(\lambda_F, \lambda_F, \Lambda) > 0 > g(\min(\Lambda, \lambda_F), \min(\Lambda, \lambda_F), \Lambda)$. Solving $g(\lambda_1, \lambda_1, \Lambda) = c_{\min} - f_h(c_h(\overline{X}_h(\lambda_1), \Lambda)) = 0$ yields (3.94).

Part 2. The thresholds $\underline{\Lambda}_{ml}$ and $\overline{\Lambda}_{ml}$ determine the sign of $g(\lambda, \min(\lambda_F, \lambda), \Lambda)$ for $\lambda = \min(\Lambda, \mu)$ when $\Lambda > \lambda_F$. When $\Lambda > \lambda_F$ and $\lambda = \min(\Lambda, \mu)$, the maximum feasible total rate of $h$ and $m$ classes is $\lambda_F$: $\lambda_{mh} = \min(\lambda_F, \lambda) = \min(\lambda_F, \min(\Lambda, \mu)) = \lambda_F$. By (3.92) the virtual delay cost difference is

$$g(\min(\Lambda, \mu), \lambda_F, \Lambda) = \begin{cases} 
  g(\Lambda, \lambda_F, \Lambda) = f_l\left(F^{-1}\left(\frac{\Lambda - \lambda_F}{\Lambda}\right)\right) - c_{\max}, & \Lambda \in [\lambda_F, \mu], \\
  g(\mu, \lambda_F, \Lambda) = f_l\left(F^{-1}\left(\frac{1}{\mu}\right)\right) - c_{\max}, & \Lambda \geq \mu.
\end{cases}$$

(3.97)

The function $g(\Lambda, \lambda_F, \Lambda)$ increases in $\Lambda \in [\lambda_F, \mu]$ as the fraction $1 - \lambda_F/\Lambda$ in $l$ classes increases, $g(\mu, \lambda_F, \Lambda)$ decreases in $\Lambda \geq \mu$ as the fraction $1/(d\Lambda)$ in $l$ classes decreases, and

$$\lim_{\Lambda \to \infty} g(\mu, \lambda_F, \Lambda) = c_{\min} - c_{\max} < 0.$$ 

Hence $g(\min(\Lambda, \mu), \lambda_F, \Lambda)$ has an unique maximum for $\Lambda \geq \lambda_F$ at $\Lambda = \mu$ where

$$g(\mu, \lambda_F, \mu) = f_l\left(F\left(\frac{1}{d\mu}\right)\right) - c_{\max} > 0 \iff F(f_l^{-1}(c_{\max})) \cdot d < \mu^{-1}.$$ 

(3.98)

Part 2(a). It follows that if $F(f_l^{-1}(c_{\max})) \cdot d < \mu^{-1} < d$ then $\underline{\Lambda}_{ml} \in (\lambda_F, \mu)$ is the unique solution of $g(\Lambda, \lambda_F, \Lambda) = 0$ and $\overline{\Lambda}_{ml} > \mu$ is the unique solution of $g(\mu, \lambda_F, \Lambda) = 0$, where $\underline{\Lambda}_{ml}$ and $\overline{\Lambda}_{ml}$ satisfy (3.95). We further have $g(\Lambda, \lambda_F, \Lambda) > 0$ for $\Lambda \in (\underline{\Lambda}_{ml}, \mu)$ and
Chapter 3. Revenue Maximizing Lead-Time Pricing

130

g(µ, λF, Λ) > 0 for Λ ∈ [µ, λF]. Recalling from Lemma 5.2(b) that for fixed Λ > λF, the function g(λ, λF, Λ) satisfies g(λF, λF, Λ) < 0 and increases in λ ≥ λF establishes the unique root λ3 as claimed. Solving g(λ3, λF, Λ) = 0 yields (3.96).

If Λ ∉ [λm, λF], the above discussion and Lemma 5.2(b) imply that g(λ, λF, Λ) < 0 for all λ ≥ λF.

Part 2(b). If µ − 1 ≤ λ = 1 − F(f − 1(cmax)) · d then the above analysis of g(min(Λ, µ), λF, Λ) and (3.98), together with Lemma 5.2(b) again imply that g(λ, λF, Λ) < 0 for all feasible λ ≥ λF.
□

Lemma 7. Fix µ > d−1. The sign of g(λ, λP, Λ) depends as follows on λ > λP and Λ. Let

Λ4 = \left\{ \Lambda \geq \lambda_P : \Lambda = \frac{\sqrt{\mu/d}}{F(f^{-1}(f_{h}(c_{h}(\lambda_P, \Lambda))))} \right\}, \quad (3.99)

where Λ4 satisfies g(µ, λP, Λ4) = 0. Moreover, Λ4 > max(Λ3, λF, µ).

1. If Λ < Λ3 then g(λ, λP, Λ) > 0 for λ > λP.

2. If Λ ∈ (Λ3, Λ4) then there is a demand rate threshold λ2 > λP such that

\begin{align}
g(λ, λP, Λ) &= \begin{cases} < 0, & \text{if } λ \in [λP, λ2), \\ = 0, & \text{if } λ = λ2 \triangleq λP + ΛF(f^{-1}(f_{h}(c_{h}(λP, Λ)))) \\ > 0, & \text{if } λ > λ2. \end{cases} \quad (3.100)\end{align}

3. If Λ ≥ Λ4 then g(λ, λP, Λ) ≤ 0 for λ > λP.

Proof. By Lemma 5.1, for fixed Λ > λP the virtual delay cost difference g(λ, λP, Λ) increases in λ. Hence for fixed Λ the function g(λ, λP, Λ) has at most one root in λ ∈ [λP, min(Λ, µ)].

The threshold Λ3 > λP is defined in Lemma 6 and determines the sign of g(λ, λP, Λ) for λ = λP, where g(λP, λP, Λ3) = 0 and g(λP, λP, Λ) > 0 (< 0) if Λ < (>) Λ3.

The threshold Λ4 defined in (3.99) determines the sign of g(λ, λP, Λ) for λ = min(Λ, µ)
where

\[
g(\min(\Lambda, \mu), \Lambda P, \Lambda) = \begin{cases} 
geq (\Lambda, \lambda P, \Lambda) = f_i \left( F^{-1} \left( \frac{\Lambda - \lambda P}{\Lambda} \right) \right) - f_h \left( F^{-1} (\frac{\mu}{\Lambda}) \right), & \Lambda \in [\lambda P, \mu], 
\geq (\mu, \lambda P, \Lambda) = f_i \left( F^{-1} (\frac{\mu - \lambda P}{\Lambda}) \right) - f_h \left( F^{-1} (\frac{\mu}{\Lambda}) \right), & \Lambda \geq \mu. 
\end{cases}
\]

(3.101)

For \( \Lambda \leq \mu \) and \( \lambda = \min(\Lambda, \mu) \) we have \( g(\Lambda, \lambda P, \Lambda) > 0 \): when all types are served the segments with high and low lead time qualities have the same marginal type

\[
(3.101)
\]

For \( \Lambda \leq \mu \) and \( \lambda = \min(\Lambda, \mu) \) we have \( g(\Lambda, \lambda P, \Lambda) > 0 \): when all types are served the segments with high and low lead time qualities have the same marginal type

\[
(3.101)
\]

For \( \Lambda \leq \mu \) and \( \lambda = \min(\Lambda, \mu) \) we have \( g(\Lambda, \lambda P, \Lambda) > 0 \): when all types are served the segments with high and low lead time qualities have the same marginal type

\[
(3.101)
\]

For \( \Lambda \leq \mu \) and \( \lambda = \min(\Lambda, \mu) \) we have \( g(\Lambda, \lambda P, \Lambda) > 0 \): when all types are served the segments with high and low lead time qualities have the same marginal type
hold. Since $\frac{1}{\mu} \leq F(f^{-1}_l(c_{\text{max}})) \cdot d$, segmentation $(m, l)$ cannot be optimal: the condition is $g(\lambda, \lambda_F, \Lambda) \geq 0$, see (3.88), and Lemma 6.2(b) rules it out.

The conditions (3.84) and (3.88), combined with Lemmas 6-7, imply the transitions among $(h)$, $(h, m)$, $(h, l)$ and $(h, m, l)$ listed in Table (3.51). We illustrate how for $\Lambda \in (\Lambda_2, \Lambda_3)$. By (3.51) the optimal segmentation transitions as $(h) \rightarrow (h, m) \rightarrow (h, m, l)$ as $\lambda$ increases. For $\lambda \leq \lambda_P$ segmentation $(h)$ is optimal by (3.84). For $\lambda > \lambda_P$ Lemma 6 specifies the sign of $g(\lambda, \min(\lambda, \lambda_F), \Lambda)$; Lemma 6.1(b) applies since $\Lambda \in (\Lambda_2, \Lambda_3)$, and Lemma 6.2(b) since $\frac{1}{\mu} \leq F(f^{-1}_l(c_{\text{max}})) \cdot d$. Together they specify that $g(\lambda, \min(\lambda, \lambda_F), \Lambda) \geq 0$ for $\lambda \leq \lambda_1$ and $g(\lambda, \min(\lambda, \lambda_F), \Lambda) < 0$ otherwise. Lemma 7 specifies the sign of $g(\lambda, \lambda_P, \Lambda)$, where Lemma 7.1 applies since $\Lambda < \Lambda_3$: it specifies that $g(\lambda, \lambda_P, \Lambda) > 0$ for $\lambda > \lambda_P$. It follows from (3.88) that $(h, m)$ is optimal for $\lambda \leq \lambda_1$, and $(h, m, l)$ is optimal for $\lambda > \lambda_1$.

\[ \square \]

**Proof of Proposition 7.2(b)**. Suppose that $f_h(c_{\text{min}}) \geq 0$ and $F(f^{-1}_l(c_{\text{max}})) \cdot d < \mu^{-1} < d$. The proof follows the same logic as explained for Part 2(a). However, by Lemma 6.2(a) segmentation $(m, l)$ is optimal if and only if $\Lambda \in [\Lambda_{ml}, \overline{\Lambda}_{ml})$ and $\lambda \geq \lambda_3$: in this case $g(\lambda, \lambda_F, \Lambda) \geq 0$, which is the optimality condition for $(m, l)$ by (3.88). The optimal segmentation for $\Lambda \in [\Lambda_{ml}, \overline{\Lambda}_{ml})$ and $\lambda < \lambda_3$ follows from Lemma 6-7 and (3.88) in exactly the same way as in Part 2(a). This yields Table (3.52).

\[ \square \]

**Proof of Proposition 7.2(c)**. Suppose that $f_h(c_{\text{min}}) < 0$ and $\frac{1}{\mu} \leq F(f^{-1}_l(c_{\text{max}})) \cdot d$. By the analysis of Step 2 the segmentation $(h, m_{sd})$ with strategic delay is optimal if and only if $f_h(c_h(\overline{\lambda}_h(\Lambda))) < 0$ where $\overline{\lambda}_h(\Lambda)$ is defined in (3.77). Refer to (3.84) for $\lambda \leq \lambda_P$ and (3.88) for $\lambda > \lambda_P$. We translate this condition into conditions on $\lambda$ and $\Lambda$. To this end, Lemma 8 characterizes the slope of $f_h(c_h(\overline{\lambda}_h(\Lambda)))$ for fixed $\Lambda$, and Lemma 9 specifies its sign depending on $\lambda$ and $\Lambda$.

\[ \square \]
Lemma 8. Fix $\mu > d^{-1}$ and $\Lambda$. Consider the virtual delay cost $f_h (c_h (\overline{\lambda}_h (\lambda)))$.

1. It decreases in $\lambda \in [0, \min (\lambda_P, \Lambda)]$, where $\overline{\lambda}_h (\lambda) = \lambda$, the maximum $f_h (c_h (0)) = c_{\max} > 0$, and

\[
 f_h (c_h (\lambda))|_{\lambda=\min(\lambda_P,\Lambda)} = \begin{cases} 
 f_h (c_{\min}) < 0, & \lambda \leq \lambda_P \\
 f_h \left( F^{-1} \left( \frac{\lambda_P}{\lambda} \right) \right), & \Lambda > \lambda_P.
\end{cases}
\] (3.102)

2. If $\Lambda > \lambda_P$, it increases in $\lambda \in [\lambda_P, \min (\lambda_F, \Lambda)]$, where $\overline{\lambda}_h (\lambda) = \mu - \mu/d (\mu - \lambda)$ and

\[
 f_h (c_h (\overline{\lambda}_h (\lambda)))|_{\lambda=\min(\lambda_F,\Lambda)} = \begin{cases} 
 f_h \left( F^{-1} \left( \frac{\overline{\lambda}_h (\lambda)}{\lambda} \right) \right), & \Lambda \in (\lambda_P, \lambda_F) \\
 c_{\max} > 0, & \lambda \geq \lambda_F.
\end{cases}
\] (3.103)

Proof. The claims on the slope of $f_h (c_h (\overline{\lambda}_h (\lambda)))$ hold since $f_h c'_h < 0$, and $\overline{\lambda}_h (\lambda)$ increases in $\lambda < \lambda_P$ and decreases in $\lambda \in (\lambda_P, \lambda_F)$ by (3.77). In Part 1. the values of $f_h (c_h (\lambda))$ for $\lambda = 0$ and $\lambda = \min (\lambda_P, \Lambda)$ follow because $c_h (\lambda) = F^{-1} (\lambda/\Lambda)$ and $f_h (c_{\max}) = c_{\max}$. In Part 2. the fact that $f_h (c_h (\overline{\lambda}_h (\lambda))) = c_{\max}$ for $\lambda = \lambda_F \leq \Lambda$ holds since $\overline{\lambda}_h (\lambda_F) = 0$ by (3.77) and $c_h (0) = c_{\max} = f_h (c_{\max})$.

Henceforth write $f_h (c_h (\overline{\lambda}_h (\lambda), \Lambda))$ to emphasize the dependence on $\Lambda$. Note: For fixed $\lambda$ the marginal type $c_h (\overline{\lambda}_h (\lambda)) = F^{-1} (\overline{\lambda}_h (\lambda)/\Lambda)$ increases in $\Lambda$, hence $f_h (c_h (\overline{\lambda}_h (\lambda), \Lambda))$ increases in $\Lambda$.

Lemma 9. Fix $\mu > d^{-1}$. The sign of $f_h (c_h (\overline{\lambda}_h (\lambda), \Lambda))$ depends as follows on $\lambda$ and $\Lambda$.

Let

\[
 \Lambda_{sd} \triangleq \frac{\mu}{2} \left( \frac{1}{x} + 1 \right) - \sqrt{\frac{\mu^2}{4} \left( \frac{1}{x} - 1 \right)^2 + \frac{\mu}{xd}} \quad \text{and} \quad \Lambda_{sd} \triangleq \frac{\lambda_P}{x}, \quad \text{where} \quad x = F (f_h^{-1} (0)) \in (0, 1),
\] (3.104)

$f_h (c_h (\overline{\lambda}_h (\Lambda_{sd}), \Lambda_{sd})) = 0$ and $f_h (c_h (\lambda_P, \Lambda_{sd})) = 0$. Moreover $\lambda_P < \Lambda_{sd} \leq \Lambda_2$ and $\Lambda_{sd} < \overline{\lambda}_{sd} \leq \Lambda_3$. Define the demand rate thresholds

\[
 \lambda_0 \triangleq \Lambda F (f_h^{-1} (0)) \quad \text{and} \quad \lambda_{sd} \triangleq \mu - \frac{\mu/d}{\mu - \lambda_0}.
\] (3.105)
1. If $\Lambda < \underline{\Lambda}_{sd}$ then $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) < 0$ if and only if $\lambda > \lambda_0$, where $\lambda_0 < \lambda_P$.

2. If $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd}]$ then $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) < 0$ if and only if $\lambda \in (\lambda_0, \lambda_{sd})$, where $\lambda_0 < \lambda_P < \lambda_{sd} < \lambda_F$.

3. If $\Lambda \geq \overline{\Lambda}_{sd}$ then $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) \geq 0$ for all feasible $\lambda$.

**Proof.** Parts 1-3 follow by Lemma 8 and since $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda))$ increases in $\Lambda$.

*Part 3.* For fixed $\Lambda$, if $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) \geq 0$ at $\lambda = \min (\lambda_P, \Lambda)$, then Lemma 8 implies that $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) \geq 0$ throughout. The condition holds if and only if $\Lambda \geq \overline{\Lambda}_{sd}$: by (3.102) in Lemma 8, for $\lambda = \min (\lambda_P, \Lambda)$ this virtual delay cost equals $f_h (c_{min}) < 0$ if $\Lambda \leq \lambda_P$, it equals $f_h (c_h (\lambda_P, \Lambda)) = f_h (\overline{\Lambda}_{sd} (\lambda_P / \Lambda))$ and increases in $\Lambda > \lambda_P$, with $\lim_{\Lambda \to \infty} f_h (c_h (\lambda_P, \Lambda)) = f_h (c_{max}) = c_{max} > 0$. Solving $f_h (c_h (\lambda_P, \overline{\Lambda}_{sd})) = 0$ yields $\overline{\Lambda}_{sd} > \lambda_P$ as in (3.104).

*Part 1.* For fixed $\Lambda$, if $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) < 0$ at $\lambda = \min (\lambda_F, \Lambda)$, then Lemma 8 implies that $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda))$ has an unique root $\lambda_0$, where $\lambda_0 < \lambda_P$ and $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda))$ is nonnegative iff $\lambda \leq \lambda_0$. The condition holds if and only if $\Lambda < \underline{\Lambda}_{sd}$: we have

$$f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda))_{\lambda=\min (\lambda_F, \Lambda)} = \begin{cases} f_h (c_h (\Lambda, \Lambda)) = f_h (c_{min}) < 0, & \Lambda \leq \lambda_P, \\
 f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) = f_h (\overline{\Lambda}_h (\lambda_P / \Lambda)), & \Lambda \in (\lambda_P, \lambda_F), \\
 f_h (c_h (\overline{\Lambda}_h (\lambda_F), \Lambda)) = c_{max} > 0, & \Lambda \geq \lambda_F,
\end{cases}$$

(3.106)

where $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) = f_h (\overline{\Lambda}_h (\lambda_P / \Lambda))$ increases in $\Lambda \in (\lambda_P, \lambda_F)$ since $\overline{\Lambda}_h (\Lambda) / \Lambda$ decreases in $\Lambda$. Hence $\underline{\Lambda}_{sd} \in (\lambda_P, \lambda_F)$ is the unique solution of $f_h (c_h (\overline{\Lambda}_h (\lambda), \Lambda)) = 0$ and satisfies (3.104). The fact that $\underline{\Lambda}_{sd} < \overline{\Lambda}_{sd}$ follows by the properties of $\overline{\Lambda}_{sd}$.

*Part 2.* If $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$ then Parts 1 and 3 together with Lemma 8 imply the stated properties.

It remains to rank $\underline{\Lambda}_{sd}$ and $\overline{\Lambda}_{sd}$ relative to $\Lambda_2$ and $\Lambda_3$, which are defined in Lemma 6. To see that $\underline{\Lambda}_{sd} \leq \Lambda_2$, recall that $g (\Lambda_2, \Lambda_2, \Lambda_2) = 0$ which is equivalent to $f_h (c_h (\overline{\Lambda}_h (\Lambda_2), \Lambda_2)) = c_{min}$. The ranking follows since $f_h (c_h (\overline{\Lambda}_h (\underline{\Lambda}_{sd}), \underline{\Lambda}_{sd})) = 0 \leq c_{min}$ and $f_h (c_h (\overline{\Lambda}_h (\Lambda), \Lambda))$ increases in $\Lambda \in [\lambda_P, \lambda_F]$. To see that $\overline{\Lambda}_{sd} \leq \Lambda_3$ recall that $g (\lambda_P, \lambda_P, \Lambda_3) = 0$ which is
equivalent to \( f_h (c_h (\lambda P, \Lambda_3)) = c_{\min} \). The ranking follows since \( f_h (c_h (\lambda P, \Lambda_{sd})) = 0 \leq c_{\min} \) and \( f_h (c_h (\lambda P, \Lambda)) \) increases in \( \Lambda \).

\[ \Box \]

Refer to Proposition 7.2(c). The case \( \Lambda < \Lambda_{sd} \) in Table (3.53) is immediate from (3.84), (3.88) and Lemma 9.1. For \( \Lambda \in [\Lambda_{sd}, \Lambda_{sd}] \) Lemma 9.2. implies the transition \((h) \rightarrow (h, m_{sd})\) for \( \lambda < \lambda_{sd} \). For \( \lambda \geq \lambda_{sd} \) Lemmas 6-7 and (3.88) imply \((h, m)\) if \( \Lambda \leq \Lambda_2 \) or \((h, m, l)\) if \( \Lambda > \Lambda_2 \).

\[ \blacksquare \]
Chapter 4

Multiple Service Level Assignment Problem

4.1 Introduction

This paper addresses a similar problem to Chapter 3. Service providing firms, from large logistics providers such as FedEx to neighbourhood dry-cleaners, market their services by posting a menu of price-service level options. An arriving customer’s selection results in a payment from the customer and a service level commitment from the provider. Delay is often the most important differentiator of service quality. Since customers typically vary in their sensitivity to delay, carefully crafting a menu of lead-time qualities and prices allows the provider to offer differentiated service and improve system wide outcomes.

Since Marchand (1974), there has been a substantial amount of work which has considered the question of how a capacitated firm should prioritize customers and charge appropriate prices. In particular, there is a broad literature examining characteristics of price-lead time menus for priority queueing systems which maximize social welfare when customers are free to select from menu options.

A welfare maximizing or efficient outcome is one which maximizes the total net-value received by customers for the service. As the following quote from Nisan (2009) makes clear,
while revenue may be the natural objective for a firm in the short term, when a longer timeframe and competitive factors are taken into account efficiency is equally important.

*I often hear cynical doubts about whether anyone optimizes efficiency rather than revenue, and specifically such disbelief regarding the big companies running ad auctions (such as my current employer, Google). As far as I can tell, reality seems to be quite the opposite: companies aim to optimize their long-term or middle-term revenue rather than the revenue of a single auction. In a competitive environment the only way of optimizing long term revenue is by gaining and maintaining market share which in turn requires providing high added-value i.e. optimizing efficiency.*

Further, under common auction settings there is substantial evidence that as the number of participants increases the welfare maximizing solution provides near optimal revenue (e.g. Bulow and Klemperer, 1994).

Important properties such as the ordering of customers in welfare maximizing menus are well understood, however, applying these results requires full knowledge of market characteristics such as the distribution of customer’s sensitivities to lead times. Acquisition of market information is an expensive and difficult problem for a firm. With this perspective in mind, the principle goal of this undertaking is to design an adaptive pricing protocol which (i) converges to a welfare maximizing solution and (ii) requires minimal market information. (iii) A related goal is to characterize the welfare of solutions of the underlying assignment problem.

To the best of our knowledge, this is the first attempt to design an adaptive pricing protocol for the assignment of multiple congested service levels to heterogenous customers. In order to accomplish this we characterize the underlying assignment problem and identify a performance measure deemed the *apparent-loss* which provides an absolute bound on the deviation of current customer welfare from the optimum and allows identification of a certain class of welfare maximizing solutions. This measure can be calculated entirely
from operational information: the current prices and the current service level utilizations. We find that the apparent-loss provides a valuable heuristic to guide the adaptive pricing protocol.

Following a discussion of related literature, Section 4.2 presents the model of the system. In Section 4.3 we define and analyze the underlying welfare maximizing assignment problem. Then, in Section 4.4, we propose a price adjustment mechanism and demonstrate effectiveness through empirical results. Finally we present conclusions in Section 4.5.

4.1.1 Related Literature and Positioning

This paper is related to several areas in the operations, economics and data networks literatures. We will begin by discussing the relation to the economics of quality differentiated menus which do not feature congestion. We will then discuss the operations literature assigning prices to service qualities in congested queueing situations. In both of these cases, the construction of the menu requires aggregate market information. We then discuss robust pricing mechanisms which attempt to find optimal allocations despite information asymmetries. We discuss both auctions and adaptive pricing mechanisms in this context. Finally we contrast adaptive pricing techniques with online learning algorithms.

Economics of Quality: The economic screening literature, pioneered in Mussa and Rosen (1978) and Maskin and Riley (1984), directly addresses the adverse selection problems posed by information asymmetry. In these models, customers are permitted to self select based on a quality-price menu. This literature is rich with respect to the description of the preference and information modellings. This includes cases where customer types are two dimensional. Our customer type model is a special case of such a two dimensional model where customers differ in their initial willingness to pay and in their sensitivity to waiting. A full survey of multidimensional screening is described in Rochet and Stole (2003). Despite the breadth of preference models, these papers feature a very consistent operational model. The supplier can improve quality of an individual service-level according to a marginal cost curve. This
model differs strongly from the capacitated supplier with stochastic arrivals which brings about congestion. In the model in this and the previous chapter, congestion leads to a set of nested capacity constraints. These constraints may lead to very different optimal pricing and service-level decisions as discussed in the previous chapter.

**Congestion Pricing:** In the operations literature, services subject to capacity constraints are studied. For more detailed surveys of the area we refer the reader to Hassin and Haviv (2003) and Afèche and Mendelson (2004). Emphasis is moved from the preference modelings to the operational structure of the service provider. Naor (1969) is widely recognized as the first attempt to use prices to moderate congestion. This seminal paper presents a model where tolls are used to moderate entry into a first-in-first-out (FIFO) queueing system from a set of customers with common delay sensitivity. In this simple environment, both welfare and revenue maximizing prices are considered.

The seminal papers of Kleinrock (1967) and Marchand (1974) are the first to consider welfare maximization where a queued server faces customers with heterogenous delay costs. Kleinrock (1967) and Marchand (1974) respectively consider the welfare maximization problem without and with incentive constraints. Since these papers, a wide range of related settings have been considered. An important result is the $c\mu$ rule which states that in the presence of customers with linear delay sensitivities $c$ and service requirements $\mu$ it is welfare maximizing to prioritize customers in order of this multiple (c.f. Kakalik, 1969). Pricing consistent with these heterogeneous service requirements are described in Ha (2001). Other important settings include competitive settings discussed in Lederer and Li (1997), Allon and Federgruen (2007) and Allon and Federgruen (2009). Of note, as in our paper, almost without exception delay costs in these models are assumed to be linear. Non-linear delay costs have been discussed in only a couple instances. Notably, Afèche and Mendelson (2004) discusses multiplicative delay costs and contrasts revenues associated with popular mechanisms under these new assumptions. Van Mieghem (2000, 1995) introduce the generalized $c\mu$ scheduling rule which is asymptotically optimal under the presence of convex delay costs.
Other important contributions to priority pricing under the welfare maximization objective include Ghanem (1975); Dolan (1978); Mendelson and Whang (1990).

In addition to the above work, there is related work which considers the revenue maximization objective. The previous Chapter 3 studies such a priority pricing problem and includes an indepth review of the related literature. Revenue has the technical disadvantage that the optimal outcome is not separable from prices. This chapter uses a two stage approach as used in this chapter which first establishes the optimal assignment (solution to a linear program) and then identifies incentive compatible prices (via the duality). This type of approach can not be used for a revenue maximizing objective.

Robust pricing mechanisms: The papers discussed in the previous paragraph consider optimal prices for a service provider but the constructions of such prices are premised on the existence of accurate aggregate descriptions of the customer preferences. For instance descriptions of the distribution of and correlation between delay sensitivity and initial value for the service must be known to find prices which lead customers with freedom to choose to welfare maximizing decisions. We have argued in the introduction that there are many cases where such information is not available. Here we discuss mechanisms which attempt to overcome this difficulty and lead to optimal allocations in a range of settings despite lacking information. While the previous paragraphs described mechanisms which dealt with information asymmetries, the following paragraphs deal with information deficits. We primarily discuss robust mechanisms which do not directly make use of nor maintain explicit models of the market. We discuss both one-shot auction mechanisms and ongoing adaptive pricing algorithms. We briefly contrast these approaches with Bayesian pricing protocols which maintain and update explicit beliefs over the course of the run.

We discuss here a small part of the auction literature which considers one-shot monopolistic mechanisms where customers have private values. We are focusing primarily on mechanisms which can be implemented under (myopically) dominant equilibria which do not rely on prior information. By one shot, we are indicating that all potential customers are
assumed to be present throughout the running of the mechanism. This includes mechanisms which may require more than one interaction between buyers and seller as in the case of an English auction with ascending price. In these cases the allocation and payments occur at the conclusion of the mechanisms. This is clearly a detriment as many of the classical results cannot be expected to carry over to the setting where arrivals are stochastic and valuations dependant on delay which may require a sequence of auctions over time.

The most well known robust auction mechanism is the Vickrey-Clarke-Groves (VCG) mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973). The mechanism can be summarized as the customer receives in utility their contribution to the welfare of the system. In a simple setting with a single item for sale and private customer values, the mechanism corresponds to the standard second price auction where customers bid their valuation and the high bidder is awarded the object at the second highest bid price. This payment scheme induces truth-telling as a dominant strategy. A disadvantage is that to calculate the welfare requires full revelation of the customers utility function. In addition to excessive information revelation, the mechanism requires a highly centralized computation and may not result in budget balanced payments (buyers may have to be paid to participate).

As a result of the above disadvantages many alternatives have been proposed. Most relevant for this chapter is the family of iterative first price auctions. These are extensions of the classic English auction which is widely used in practice by auction firms such as eBay and Sotheby’s. There is a close relation to adaptive pricing protocols since each iteration both schemes propose new prices in response to buyer actions.

Bertsekas (1988) describes the auction algorithm as an optimization method for the classical assignment problem. The assignment problem features a set of goods to be assigned to a set of buyers each of which has a value associated with each good. The problem is to assign at most one good to each buyer in order to maximize total value (social welfare). Bertsekas and Castanon (1989) and Bertsekas (1991) extends these techniques to transportation and shortest path problems.

This type of mechanism has also been applied to combinatorial auctions. The assignment
problem underlying combinatorial auctions generalizes the classical assignment problem by allowing assignment of sets of goods (also referred to as packages or bundles) to buyers who have general private value functions over sets of goods. Bikhchandani and Ostroy (2002) discuss analytical and economic results of this model which they call the package assignment model. Analysis of both the package assignment model and the classical assignment model benefit from a linear programming formulation whose dual formulation is sufficient to show the existence of incentive compatible discriminative prices which are incentive compatible with the optimal assignment. Discriminative prices imply that each customer may have to be presented with a separate set of prices. Our analysis benefits from this dual approach though we reach stronger results which do not rely on discriminative prices. Bikhchandani and Ostroy (2002) as well as Parkes and Ungar (2000) and Ausubel and Milgrom (2002) discuss the analog of the English auction in this setting. The equilibrium underlying optimal bidding strategies is typically stable only for myopic agents who consistently strategize as if the current bidding round is the final round.

As opposed to an auction, an adaptive pricing (AP) protocol allows customers to arrive over time. Each time a customer arrives they are presented with a menu from which they select a good at the corresponding price. Based on their decision, the protocol may update the offered prices. In the early going, customers are likely to be presented with prices which are not fulfilling the seller’s objective and the system may be lose substantial welfare or utility in the early stages. As time progresses and prices are adjusted, the protocol should find better prices which provide some guarantee that the provider is near their objective. When the guarantee is sufficiently tight the protocol is terminated. AP protocols have been proposed for the allocation of manufactured goods as well as services. In keeping with the topic of this paper we generally refer to the sold good as a service level.

AP protocols have been applied to both uncongested and congested settings. With respect to non-congested settings both revenue and welfare maximization objectives have been considered. (Berg and Ehtamo, 2008, 2009; Ehtamo et al., 2010) consider revenue maximization for a menu of quality differentiated manufactured products. Low and Varaiya
Thomas et al. (2002) consider welfare maximization in a network with guaranteed service levels. AP protocols have also been used in network optimizations to achieve fairness objectives (c.f. Kelly et al., 1998).

There are several papers which feature congestion but where there is only a single service level per link. Stidham Jr (1992); Rump and Stidham Jr (1998) consider welfare maximization on a single link subject to best-effort service where arriving customers who pay the entry toll are scheduled in a first-in-first-out (FIFO) manner. These papers consider both the short term problem of achieving equilibrium prices and admission and the longer term problem of adjusting capacity. They show that in general there may be many equilibria corresponding to an entry price and propose an adaptive protocol where a price equilibrium is achieved in one stage and capacity is adjusted in a second stage. The protocol results in a relative (not necessarily global) maximum. Similar approaches have also been studied for pricing in a decentralized network. Gupta et al. (1997); Masuda and Whang (1999) provide relevant examples of such protocols. Again, optimality is difficult to guarantee.

There are two primary drawbacks of adaptive pricing (i) there is no performance guarantee over the entire length of the exercise and (ii) market information is not acquired leaving the provider with a limited picture at the conclusion of the exercise. Focused on (i), adaptive pricing protocols do not explicitly concern themselves with the quality of prices at iterations prior to termination despite the fact that revenue and welfare is being earned (and lost) at each of these stages. Relevant reinforcement learning approaches balance exploration (learning) and exploitation (earning) to improve outcomes over the course of the run. Bayesian reinforcement learning approaches to pricing are generally model based (the true distribution of customer preferences is assumed to belong to a family of parametric distributions) and attempt to rectify both (i) and (ii) by maintaining and updating beliefs over customer distributions while iteratively adjusting prices to promote their immediate revenue or welfare objective and to provoke informing customer decisions. The earliest analytical example in a pricing context is Rothschild (1974) which extends the two armed bandit problem to a pricing setting with a restricted set of possible prices. More recent Bayesian learning papers
such as Harrison et al. (2011) allow the price to be chosen more generally. Harrison et al. (2011) shows that a myopic Bayesian policy which considers only exploitation at each stage performs well relative to a clairvoyant policy.

For systems featuring congestion there are very few Bayesian reinforcement learning papers. To the best of our knowledge, Afèche and Ata (2005) is the unique paper using Bayesian updating to dynamically price lead-time. They consider a limited setting: the server is modelled by an M/M/1/2 queue which can hold at most two customers resulting in three possible states for the system and there are two customers classes differentiated by how patient they are. With this limited statespace, Afèche and Ata (2005) is able to characterize closed form results. Related papers which focus on learning traffic characteristics in a congested settings include Gibbens et al. (1995).

While in theory Bayesian approaches are clearly appealing, they suffer from both analytical and computational tractability which limits their applicability. The large unobservable state space in our model leads us to believe that an adaptive pricing protocol leveraging known characteristics of optimal pricing is a more promising approach.

**Positioning our paper:** This chapter has two sets of results, first a novel analysis of the welfare maximizing lead-time assignment problem and second an exercise in the design of an adaptive pricing protocol. The chapter is unlike other lead-time pricing analyses in that we consider a related linear programming version of the problem. Most importantly the analyses identifies sufficient conditions and related bounds on optimality which are to the best of our knowledge novel. The chapter is differentiated from the other adaptive pricing papers discussed above since (i) it features multiple service levels and (ii) there is congestion which implies interdependence between allowable arrival rates at different service levels. The most closely related papers are Stidham Jr (1992) and the Berg and Ehtamo (2008, 2009); Ehtamo et al. (2010) series of papers. Stidham Jr (1992) considers only a provider who offers a single service level whose delay is defined by best-effort service (work conserving FIFO) and the customer decision to join. In contrast our service provider offers multiple service
levels and firmly designates their associated delay at the outset. We discuss the benefits and
detrimentsof server specified service levels vs. best effort service in detail at the end of the
model. Berg and Ehtamo (2008, 2009); Ehtamo et al. (2010) consider a provider selling a
range of quality differentiated products not subject to congestion. Their protocol attempts
to find revenue maximizing prices.

4.2 Model

This chapter features a provider delivering a service (or a make-to-order product) with IID
service times drawn from a known distribution. As in the previous chapter, the customers (i)
value an identical (and the unique) service from this monopoly provider, (ii) will experience a
delay between ordering and completion of the service and (iii) differ in their linear sensitivity
to that delay and their base-value for the service. The server maintains an unlimited length
queue and presents a menu of prices for specific service-levels defined by the expected delay
to delivery. Customers arrivals are stochastic and regulated by a set of prices assigned to
service levels. The formal details are described in the following paragraphs.

Service time is IID with mean $1/\mu$. Customers are drawn independently from $N$ types:
$\Theta = \{1, 2, \ldots, N\}$. Customers arrive at the server according to a Poisson process with total
rate $\Lambda$ and from type $i$ with rate $\lambda_i$. Of course $\sum_{i \in \Theta} \lambda_i = \Lambda$ and the probability of an
arriving customer being of type $i$ is $\lambda_i/\Lambda$. Each customer of type $i$ has a delay sensitivity
$c_i$ and a base value $v_i$. For a customer of type $i$ the value for receipt of the service after a
delay of $w$ is

$$V(w : i) = v_i - c_iw. \quad (4.1)$$

Customers select from a menu of service levels $\mathcal{W} = \{w_1, \ldots, w_z\}$ and prices $p : \mathcal{W} \to \mathbb{R}$. A service level $w$ denotes the expected time between entering the system and completion of
the service. The origin of the components of this menu are discussed later with details of
the provider. Customer $i$ receives expected utility

$$ U(w, \rho : i) = V(w : i) - \rho $$

(4.2)

from service option $w$ and payment $\rho$.

As an alternative to joining the system, the customer may balk on inspection of the menu and leave the system without receiving the service. The decision to balk is denoted by $b$. We normalize the value of the outside option to zero such that the customers expected utility from balking is $U(b, \rho : i) = 0$. Customer $i$’s utility maximizing choices correspond to the set

$$ W^*(i : W, p) = \arg \max_{W \cup \{b\}} \{U(w, p(w))\}. $$

(4.3)

To describe the allocation of customers to service-levels (or balking) we introduce the assignment function $A : W \cup \{b\} \rightarrow \mathbb{R}$. $A(i, w)$ defines the arrival rate of customers who are of type $i$ which select service level $w$. The stream of customers of a particular type may be partitioned between two or more service levels (or balking). In such a case we assume that the choice of each customer is selected independently where the probability of selecting option $w$ is $A(i, w)/\lambda_i$.

We assume that customers are utility maximizing. An assignment $A(i, w)$ is consistent with this assumption if it is *incentive compatible* which holds if customers select only from their utility maximizing choices:

$$ \sum_{w \in W^*(i : W, p)} A(i, w) = \lambda_i, $$

(4.4)

$$ \sum_{w \in W \cup \{b\} \setminus W^*(i : W, p)} A(i, w) = 0. $$

(4.5)

For prices $p$, an assignment has revenue $\Pi(A, p) = \sum_i \sum_w A(i, w)p(w)$ and the contribution to revenue from each service level as $\Pi(w, A, p) = \sum_i A(i, w)p(w)$.

We say that an assignment $A$ is *supported* by pricing function $p$ if for all types $i$ and $w \in W \cup \{b\}$, if $A(i, w) > 0$ then $U(w, p(w) : i) \geq U(w', p(w) : i)$. In other words the prices and assignments are incentive compatible.
In addition to utility maximization, we assume that customers are completely indifferent between choices in their utility maximizing set. This is formalized in the following stronger assumption: customers’ actual decisions correspond to choosing with equal probability from $W^*(i : W, p)$. This assumption implies a unique assignment $A^{R\text{IC}}(\cdot : p)$ which we call the random incentive compatible (RIC) assignment. This assignment depends on $\Theta, W$ as well as $p$ though we will generally suppress the dependence on $\Theta$ and $W$. $A^{R\text{IC}}(\cdot : p)$ is of the following form:

$$A^{R\text{IC}}(w, i, p, \Theta, W) = \frac{\lambda_i 1_{w \in W^*(i : W, p)}}{|W^*(i : W, p)|},$$

(4.6)

where $1_{w_k \in W^*(i : W, p)}$ is the indicator taking value 1 if $w_k$ is a utility maximizer. The RIC assignment will be used as the standard in the computational experiments.

The service provider’s decisions variables are limited to the prices $p$. The set of service levels is assumed to be fixed a priori. This likely corresponds to reality where there is often a finite and discrete set of reasonable service levels such as one day delivery, two day delivery and so on in the case of a courier. When a customer selects a service level $w$, the provider is committed to delivery after this period of time in expectation. Beyond that, the provider has complete freedom in how they schedule the various service levels. These assumptions lead to following notion of operational feasibility:

**Definition 4.** An assignment $A$ is operationally feasible for the provider if there exists a (possibly preemptive) non-anticipative scheduling scheme which for all $w_k \in W$ gives average delay of $w_k$ for all customers assigned to this option.

By using the achievable region approach introduced in Coffman Jr and Mitrani (1980) and generalized in Federgruen and Groenevelt (1988), operational feasibility of a particular assignment can be assessed by assuring that the following set of constraints hold:

$$\sum_{w \in \Theta : w \leq x} w \sum_{i \in \Theta} A(i, w) \geq C(\sum_{w \in \Theta : w \leq x} \sum_{i \in \Theta} A(i, w)) \quad \forall x \in W.$$

(4.7)

Where $C(\lambda)$ is the total average delay ($\lambda$ times the average delay) associated with a FIFO (or any other work-conserving) scheduling protocol serving customers arriving according to
a Poisson process with rate $\lambda$.

The provider adjusts only prices. The service levels or number of priority classes is fixed a priori. Of course options may be priced such that they are not utilized, e.g. $p(w) = \max_{i \in \Theta} v_i + 0.01$ will ensure no customers select $w$. So, by picking a large $W$, this assumption is not too restrictive.

The provider’s objective is to maximize welfare which is the steady state expected value received by customers:

$$SW(A : \Theta, W) = \sum_{i \in \Theta} \sum_{w \in W} A(i, w)V(i, w)$$  \hspace{1cm} (4.8)

We also define the contribution to welfare from each service level as $SW(w, A : \Theta, W) = \sum_{i \in \Theta} A(i, w)V(i, w)$. Maximum welfare is a common objective objective in the literature and is particularly appropriate for a monopoly provider where regulation may play a role.

The basic model is very similar to Chapter 3 however it differs in the following ways. (i) The customer’s base value for the service may not be correlated with their delay sensitivity. (ii) Customer types are discretely distributed. (iii) The set of service levels is discrete. (iv) The objective is welfare maximization and not revenue maximization.

### 4.2.1 Revenue Upper Bound

As an upper bound on the revenue outcome from any incentive compatible assignment we define the best-case revenue. The best-case revenue is the maximum revenue achievable from any operationally feasible stream of demand to service levels. We define a demand vector $\bar{\lambda}$ with elements $\bar{\lambda}^w$ defining the arrival rate at service level $w$. Then the best case demand vector is:

$$\bar{\lambda}^* = \arg\max_{\lambda} \left\{ \sum_{w \in W} \bar{\lambda}^w p(w) \right\}$$  \hspace{1cm} (4.9)

Subject to:

$$\sum_{w \in W : w \leq x} w \bar{\lambda}^w \geq C \left( \sum_{w \in W : w \leq x} \bar{\lambda}^w \right) \quad \forall x \in W,$$  \hspace{1cm} (4.10)

and the best-case revenue is $\Pi(p) = \{\sum_{w \in W} \bar{\lambda}^* w p(w)\}$ with contribution from each service level $\Pi(w, p) = \bar{\lambda}^* w p(w)$. 

4.3 Analysis of the Assignment Problem

The goal of this chapter is the development of a protocol for improving prices subject to information limitations. The direction we take is to develop a heuristic, the apparent-loss, which is closely related to the welfare of the system. The theoretical properties of the apparent-loss are determined through the analysis of the underlying assignment problem. The objective is to implement the welfare maximizing, operationally feasible assignment. The objective does not depend on prices. A mathematical formulation of this assignment problem \((AP)\) is as follows:

\[
\max_{A(\cdot, \cdot)} SW(A) \quad (4.11)
\]

Subject to:

\[
\sum_{w \in W \cup \{b\}} A(i, w) = \lambda_i \quad \forall i \in \Theta \quad (4.12)
\]

\[
\sum_{w \in W : w \leq x} w \sum_{i} A(i, w) \geq C \left( \sum_{w \in W : w \leq x} \sum_{i} A(i, w) \right) \quad \forall x \in W \quad (4.13)
\]

Foremost, the intent of this section is to show existence and characteristics of prices which are incentive compatible with the assignment solving the above problem. In order to accomplish this it is useful to characterize an approximate linear version of the above problem which constrains the set of feasible assignments using a discrete set of operationally feasible demand vectors \(D = \{\bar{\lambda}_1 \ldots \bar{\lambda}_d\}\). Recall that a demand vector \(\bar{\lambda}\) defines usage at each service level and is feasible if \(\sum_{w \in W : w \leq x} w \bar{\lambda}^w \geq C \left( \sum_{w \in W : w \leq x} \bar{\lambda}^w \right)\) for all service levels \(x\). In the approximate formulation Constraint 4.13 is replaced by the requirement that the demand vector associated with a potential assignment must have less demand at each service level than a convex combination of elements of \(D\):

\[
\sum_{\lambda \in D} y_{\lambda} \leq 1 \quad (4.14)
\]

\[
\sum_{i \in \Theta} A(i, w) \leq \sum_{\lambda \in D} y_{\lambda} \bar{\lambda}^w \quad \forall w \in W \quad (4.15)
\]

\[
y_{\lambda} \geq 0 \quad (4.16)
\]
Let $AP$ be the program corresponding to replacing Constraint 4.13 in $AP$ with Constraints 4.14-4.16.

**Lemma 10. [Lower Bound of the Assignment Problem]** If $C$ is convex then an assignment $A$ solving $\bar{AP}$ is a feasible assignment in $AP$ and provides a lower bound on the optimal objective value of $AP$.

**Proof:** We show that since $C$ is convex any convex combination of feasible demand vectors is also a feasible demand vector. This is sufficient to complete the proof since then any operationally feasible assignment to $\bar{AP}$ also provides a feasible solution to $AP$ and hence a lower bound on the optimal solution to $AP$.

A fractional $y_{\bar{\lambda}}$ corresponds to a new demand vector $\tilde{\lambda}$ where $\tilde{\lambda}^w = \sum_{\bar{\lambda} \in D} y_{\bar{\lambda}} \bar{\lambda}^w$. For $\tilde{\lambda}$ to be feasible the following must hold for all $x \in W$: $\sum_{w: w \leq x} w \tilde{\lambda}^w \geq C\left( \sum_{w: w \leq x} \bar{\lambda}^w \right)$. Observing that for all $x \in W$,

$$\sum_{w: w \leq x} \bar{\lambda}^w = \sum_{w: w \leq x} \sum_{\bar{\lambda} \in D} y_{\bar{\lambda}} \bar{\lambda}^w w$$

$$= \sum_{\bar{\lambda} \in D} y_{\bar{\lambda}} \sum_{w: w \leq x} \bar{\lambda}^w w$$

$$\geq \sum_{\bar{\lambda} \in D} y_{\bar{\lambda}} C\left( \sum_{w: w \leq x} \bar{\lambda}^w \right)$$

$$\geq C\left( \sum_{w: w \leq x} \sum_{\bar{\lambda} \in D} y_{\bar{\lambda}} \bar{\lambda}^w \right)$$

$$= C\left( \sum_{w: w \leq x} \tilde{\lambda}^w \right)$$

Where the final inequality holds for convex $C$ and corresponds to the operational feasibility constraint.

$\blacksquare$

The $AP$ formulation is a linear program and has an associated dual program with dual variables $u_i$, $\pi$ and $p_w$. Respectively, $u_i$, $\pi$ and $p_w$ correspond to Constraints 4.12, 4.14 and
4.15. The dual program is as follows:

$$\begin{align*}
\min_{u_i, p_w, \pi} & \sum_{i \in \Theta} u_i \lambda_i + \pi \\
\text{Subject to:} & \\
& u_i \geq V(w : i) - p(w) \quad \forall i \in \Theta, w \in W \quad (4.18) \\
& \pi \geq \sum_{w \in W} p_w \bar{\lambda}^w \quad \forall \bar{\lambda} \in D \quad (4.19) \\
& u_i, p_w, \pi \geq 0 \quad (4.20)
\end{align*}$$

By observation, it is readily apparent that the dual variables admit a convenient interpretation: $u_i$ as the expected utility of customer type $i$; $\pi$ as the total revenue of the provider; $p_w$ the price of entry for service level $w$ ($p(w)$).

The complementary slackness conditions below provide necessary and sufficient conditions for variables $A(i, w)$, $y_{\bar{\lambda}}$ and $u_i$, $\pi$, $p_w$ to be optimal solutions to the primal and dual programs:

$$\begin{align*}
u_i = V(w : i) - p_w & \iff A(i, w) > 0 \quad (4.21) \\
u_i > V(w : i) - p_w & \Rightarrow A(i, w) = 0 \quad (4.22) \\
\pi = \sum_{w \in W} p_w \bar{\lambda}^w & \iff y_{\bar{\lambda}} > 0 \quad (4.23) \\
\pi > \sum_{w \in W} p_w \bar{\lambda}^w & \Rightarrow y_{\bar{\lambda}} = 0 \quad (4.24)
\end{align*}$$

Let $\tilde{A}(i, w)$ and $\tilde{y}_{\bar{\lambda}}$ be the optimal solution to the primal and, let $\tilde{u_i}$, $\tilde{\pi}$ and $\tilde{p_w}$ be the optimal solution to the dual program. It is also immediately of value to define $\Pi_{\tilde{A}\tilde{P}}(p)$ as the best-case
revenue contingent on Constraints 4.14-4.16:

\[ \bar{\Pi}_{AP}(p) = \max_{\bar{\lambda}} \{ \sum_{w \in W} \bar{\lambda}^w p(w) \} \]  

(4.25)

Subject to:

\[ \sum_{\ell \in D} y_{\ell} \leq 1 \]  

(4.26)

\[ \sum_{i \in \Theta} \bar{\lambda} \leq \sum_{\ell \in D} y_{\ell}^w \]  

\[ \forall w \in W \]  

(4.27)

\[ y_{\ell} \geq 0 \]  

(4.28)

Studying the complementary slackness conditions leads to the following result:

Lemma 11. [Existence of Incentive Compatible Prices] Any assignment \( \bar{A}(i, w) \) which maximizes \( \bar{AP} \) is supported by a pricing function \( \bar{p}(w) \) where \( \bar{p}(w) \) and \( \bar{A}(i, w) \) are incentive compatible and \( \Pi(A : p) = \bar{\Pi}_{AP}(p) \).

Proof: Let \( u_i, \pi \) and \( p_w \) be the corresponding solution of the dual. We will show that prices \( \bar{p}(w) = p_w \) are incentive compatible with \( \bar{A} \). Recall that utility is \( U(i, w : p) = V(w : i) - \bar{p}(w) \). Since complementary slackness must hold, Condition 4.21 shows if customers of type \( i \) are assigned to service level \( w \) (\( \bar{A}(i, w) > 0 \)) then they receive precisely utility \( u_i \). By Condition 4.22 if \( U(i, w : p) < u_i \) no customers of type \( i \) are assigned to this option. Since \( U(i, w : p) \leq u_i \), the utility maximizing set is

\[ W^*(i : \mathcal{W}, \bar{p}) = \{ w : V(w : i) - \bar{p}(w) = u_i \} = \{ w : \bar{A}(i, w) > 0 \}. \]

Then for each customer type, customers are assigned only to utility maximizing service levels and \( \bar{A} \) and \( \bar{p} \) satisfying the requirements for incentive compatibility in Equations 4.4 and 4.5.

We now show \( \Pi(A : p) = \bar{\Pi}_{AP}(p) \). Note, \( \Pi(A : p) = \sum_{w \in W} \sum_{i \in \Theta} A(i, w)p(w) = \sum_{w \in W} \sum_{\ell \in D} \bar{y}_{\ell}^w p(w) \) by Equation 4.15. Since Condition 4.23 holds, \( \Pi(A : p) = \sum_{w \in W} \bar{y}_{\ell} \pi = \pi \). Further, from Constraint 4.19, \( \pi \) is the maximum revenue from any \( \bar{\lambda} \in D \). Since \( \bar{\Pi}_{AP}(p) \) is a convex combination of \( \{ \sum_{w \in W} \bar{\lambda}^w p(w) \}_{\bar{\lambda} \in D} \), \( \bar{\Pi}_{AP}(p) \leq \pi \). Since \( \bar{\Pi}_{AP}(p) \geq \Pi(A : p) \) the
required statement must hold with equality.

Then, the following sufficient condition for optimality also follows from the complementary slackness.

**Lemma 12. [Sufficient Condition for Welfare Maximization]** Let $A$ be operationally feasible with respect to program $\bar{AP}$ and incentive compatible with pricing function $p$, then, $A$ is an optimal solution of $\bar{AP}$ if $\Pi(A : p) = \bar{\Pi}_{\bar{AP}}(p)$.

**Proof:** To show this holds we show that the requirements of the lemma are consistent with a set of dual variables which satisfy 4.21 to 4.24. These complementary slackness conditions are sufficient conditions for optimality and implying $A(i, w)$ is an optimal solution to $\bar{AP}$. The dual variables are $p_w = p(w)$, $u_i = \max_{w \in \mathcal{W} \cup \{b\}} U(i, w : p)$ and $\pi = \Pi(A : p)$.

First, we show that Condition 4.21 holds. By the above definition of $u_i$, $V(w : i) - p(w) = u_i$ implies $w \in W^*(p, i)$ then, from incentive compatibility (Equation 4.4) $A(i, w) > 0$ as required.

To show that Condition 4.22 holds, consider a customer type $i$ and service level $w$ for which $V(w : i) - p(w) < u_i$. Then clearly the service level is not utility maximizing $w \notin W^*(i : \mathcal{W}, p)$ and by incentive compatibility (Equation 4.5) $A(i, w) = 0$ as required.

Prior to showing that Conditions 4.23 and 4.24 hold note that the best-case revenue can be expressed as a convex combination of revenues from the basis demand vectors. So, $\bar{\Pi}_{\bar{AP}} = \sum_{\lambda \in \mathcal{D}} y_\lambda \sum_{w \in \mathcal{W}} \bar{\lambda}^w p(w)$ for optimal $y_\lambda$. It is well known that the maximum of convex combination of a basis of real numbers exists at a member of the basis so there is a $\bar{\lambda}^* \in \mathcal{D}$ such that $\bar{\Pi}_{\bar{AP}} = \sum_{w \in \mathcal{W}} \bar{\lambda}^* w p(w)$. As a result, for a focal member of the basis $\bar{\lambda}$ the following holds

$$\pi = \Pi(A : p) = \sum_{\ell \in \mathcal{D}} y_\ell \sum_{w \in \mathcal{W}} \ell^w p(w)$$

$$\leq y_\lambda y_\bar{\lambda} \sum_{w \in \mathcal{W}} \bar{\lambda}^w p(w) + (1 - y_\lambda) \pi$$

(4.29)

(4.30)
Then, to show Condition 4.23 holds it suffices to note that for a focal member of the basis where \( y_{\lambda} > 0 \) if \( \pi > \sum_{w \in W} \bar{\lambda}^w p(w) \) the inequality in Equation 4.30 would not hold.

Finally, Condition 4.24 holds from a similar argument. If for a focal member of the basis \( \pi > \sum_{w \in W} \bar{\lambda}^w p(w) \) then the inequality in Equation 4.30 would not hold if \( y_{\lambda} > 0 \).

The existence and sufficient condition results translate easily to the original assignment problem using Lemma 10. These results are summarized in the following proposition:

**Proposition 11.** [Properties of the Optimal Assignment Problem] Consider an instance of \( AP \) where \( C \) is convex,

1. any optimal solution \( A^* \) of \( AP \) is supported by incentive compatible price function \( p^* \) where \( \Pi(A^* : p^*) = \bar{\Pi}(p^*) \),

2. an operationally feasible assignment \( A \) supported by incentive compatible price function \( p \) is an optimal solution of \( AP \) if

\[
\Pi(A : p) = \bar{\Pi}(p), \quad (4.31)
\]

3. for any operationally feasible assignment \( A \) with supporting incentive compatible price function \( p \),

\[
\bar{\Pi}(p) - \Pi(A, p) \geq SW(A^*) - SW(A). \quad (4.32)
\]

**Proof of Proposition:** We prove the properties of Proposition 11 in turn.

**Property (1):** Solving \( AP \) reveals \( A^* \) with associated demand vector \( \bar{\lambda}^* \) where \( \bar{\lambda}^w = \sum_{i \in \Theta} A^*(i, w) \). Clearly, if \( \bar{\lambda}^* \in D \) then \( A^* \) is a feasible and optimal solution of \( AP \). Lemma 11 shows that the assignment is supported by incentive compatible prices \( p^*(w) = \bar{p}(w) \). Then, to show that \( \Pi(A^* : p^*) = \bar{\Pi}(p^*) \) note that there is a demand vector \( \bar{\lambda} \) which results in \( \bar{\Pi}(p^*) \). Assume \( D \) contains both \( \bar{\lambda} \) and \( \bar{\lambda}^* \), then \( \bar{\Pi}_{AP}(p^*) = \bar{\Pi}(p^*) \) so by Lemma 11 \( \Pi(A^* : p^*) = \bar{\Pi}_{AP}(p^*) = \bar{\Pi}(p^*) \) as required.
Property (2): Assume \( A \) and \( p \) are incentive compatible and \( \Pi(A : p) = \tilde{\Pi}(p) \). Consider \( \bar{\lambda}_A = \sum_{i \in \Theta} A(i, w) \) and \( \bar{\lambda} \) which results in \( \tilde{\Pi}(p) \). Now for the sake of contradiction assume there exists an assignment \( A^o \) where \( SW(A^o) > SW(A) \). Let \( \bar{\lambda}_{A^o} = \sum_{i \in \Theta} A(i, w) \) and \( \bar{\lambda}_A \) be contained in \( D \) so that both \( A \) and \( A^o \) are feasible solutions to \( AP \). Now note that \( \Pi(A : p) = \tilde{\Pi}(p) \geq \tilde{\Pi}_{AP}(p) \) and since \( \Pi(A : p) \) cannot be greater than \( \tilde{\Pi}_{AP}(p) \), from Lemma 12, \( A \) is an optimal solution to \( AP \) and thus \( SW(A) \geq SW(A^o) \) showing a contradiction and completing the proof.

Property (3): Note immediately that

\[
\tilde{\Pi}(p) - \Pi(A, p) \geq SW(A^*) - SW(A) \iff SW(A) - \Pi(A, p) \geq SW(A^*) - \tilde{\Pi}(p).
\]

Observe that \( SW(A^*) - \tilde{\Pi}(p) > SW(A^*) - \Pi(A^*) \) it suffices to show

\[
SW(A) - \Pi(A, p) \geq SW(A^*) - \Pi(A^*). \tag{4.33}
\]

To perform this comparison, we will consider a reallocation of the assignment of customers from \( A \) to generate assignment \( A^* \) and consider how each reassignment contributes to the left and right side of Equation 4.33. The reallocation is defined for each \( i, w_o \) (\( o \) denotes origin) and partitions the arrival rate \( A(i, w_o) \) and describes a set of alternative decisions \( r(i, w_o, w_d) \) (\( d \) denotes destination) consistent with \( A^* \). Formally, for a reallocation \( r \) to be valid, for any customer type \( i \), \( \sum_{w_d} r(i, w_o, w_d) = A_p(i, w_o) \) and \( \sum_{w_o} r(i, w_o, w_d) = A^*(i, w_d) \) must hold.

Now, the surplus for each customer and arrival rate where \( A(i, w_o) > 0 \) is \( V(i, w_o) - p(w_o) \). Since the customer is utility maximizing given \( p \), \( V(i, w_o) - p(w_o) \geq V(i, w_d) - p(w_d) \). From
above we have the following string of inequalities:

\[
SW(A) - \Pi(A, p) = \sum_{i \in \Theta} \sum_{w_o \in W} A(i, w_o)[V(i, w_o) - p(w_o)]
\]

\[
= \sum_{i \in \Theta} \sum_{w_o \in W} \sum_{w_d \in W} r(i, w_o, w_d)[V(i, w_o) - p(w_o)]
\]

\[
\geq \sum_{i \in \Theta} \sum_{w_o \in W} \sum_{w_d \in W} r(i, w_o, w_d)[V(i, w_d) - p(w_d)]
\]

\[
= \sum_{i \in \Theta} \sum_{w_d \in W} A^*(i, w_d)[V(i, w_d) - p(w_d)]
\]

\[
= SW(A^*) - \Pi(A^*)
\]

This shows that Equation 4.33 holds completing the proof.

\[\blacksquare\]

It is quite evident that \(\tilde{\Pi}(p) - \Pi(A, p)\) plays an important role in our results. We define this difference between the best-case and actual revenue as the apparent-loss:

\[AL(A, p) = \tilde{\Pi}(p) - \Pi(A, p)\]

with contributions from each service level \(AL(w, A, p) = \tilde{\Pi}(w, p) - \Pi(w, A, p)\). From the above proposition, the apparent-loss provides a bound on deviation from welfare maximization for any set of prices and a sufficient condition to identify a welfare maximizing outcome. The apparent-loss also has the important advantage that it can be calculated without direct knowledge of the distribution of customer types. The provider however has to have an accurate picture of its operational capabilities in order to calculate the best case revenue.

**Discussion of the apparent-loss** The apparent-loss is positive when high value balking customers can not be ruled out. In particular, for a particular set of prices and an assignment, a provider lacking information can not eliminate the possibility of a hypothetical set of customers who would produce revenue within a small \(\epsilon\) of the best-case revenue and improve welfare. To promote this interpretation of the apparent-loss consider a simple example with one customer type \(\Theta = \{(c_1, v_1, \lambda_1)\}\) and one service level \(W = \{w\}\). Assume that an arrival
rate of $\lambda_w > \lambda_1$ can be served at service-level $w$. The customer receives value for the service $v_1 - c_1 w$. If $p(w) > v_1 - c_1 w$ the customer balks and the apparent loss is $\lambda_w p(w)$. This is because from the point of view of a provider lacking information, a customer type with arrival rate $\lambda_w$ may be willing to pay $p(w) - \epsilon$. This hypothetical customer type would deliver welfare equal to the apparent-loss. If $p(w) < v_1 - c_1 w$ the customers of type 1 purchase the service. The apparent-loss is then $(\lambda_w - \lambda_1)p(w)$ because again a customer type may be willing to pay $p(w) - \epsilon$. Only when $p(w) = 0$ is the apparent-loss eliminated because only then can the possibility of the hypothetical customer be eliminated.

This example is also useful for observing the limitations of apparent-loss with respect to revenue. Revenue is maximized when $p(w) = v_1 - c_1 w$. In this case, the provider can extract the full surplus. However, the apparent-loss from such a policy is $(\lambda_w - \lambda_1)v_1 - c_1 w$. The apparent-loss is nullified only when the provider charges $p(w) = 0$ for the service and receives zero revenue. This is of course a pathological example and when demand for the service increases minimizing apparent-loss may deliver acceptable revenue. Using the same example but with a higher demand $\lambda_1 > \lambda_w$, apparent-loss is minimized when revenue is maximized.

**Approximation results** An appropriate question to ask is what is the welfare loss from the approximations and assumptions made in the model. First, with the same set of discrete service levels $\mathcal{W}$, how does the optimal RIC assignment compare to the optimal assignment. Second, how does the optimal assignment with discrete service levels compare to the welfare maximizing assignment with continuous service levels (i.e. $\mathcal{W} = \mathbb{R}^+$). Unfortunately, in the most general setting, the RIC assignment may result in an arbitrarily large welfare loss. This again may be illustrated with an example with one customer type $\Theta = \{(c_1, v_1, \lambda_1)\}$ and one service level $\mathcal{W} = \{w\}$. Assume that an arrival rate of $\lambda_w < \lambda_1/2$ customers can be served. The IC assignment would assign an arrival rate of $\lambda_w$ to $w$ and the remaining $\lambda_1 - \lambda_w$ customers would balk. It is IC to charge $v_1 - c_1 w$ such that no customers receive positive utility. However, the RIC assignment would assign zero customers to $w$ because at
least $\lambda_w/2 > \lambda_1$ customers would have to be assigned to $w$ leading to infeasibility.

This performance is driven by zero utility customers and we expect these circumstances will arise only rarely. By making a simple assumption resulting in a model similar to the customer model used in Chapter 3 these difficulties can be eliminated. In this result we consider a setting where base value and delay sensitivity are affinely related. Specifically, $v_i = v(c_i)$ where $v()$ is linear. In this case, the optimal assignment has at most two zero utility customer types (discussed at length in Chapter 3). Then with a large number of types we may show that the quality of the approximation is linear in the delay difference between neighbouring service levels. This shows that the system is robust and not overly sensitive to the precise selection of $W$. Let $A^{**}$ be the optimal assignment for $AP$ with $W = \mathbb{R}^+$. Let $A^*$ be the optimal assignment with the standard $W$ and $A^{RIC}$ be the optimal RIC assignment.

To simplify the exposition assume that $W = W_z$ where there are $z$ evenly spaced service-levels over the interval $[0, w_{max}]$. The service level $w_{max} = \max_{i \in \Theta} \{v_i/c_i\}$ where all customers would receive non-positive value. Also let $v_{max} = \max_{i \in \Theta} \{v_i\}$.

**Theorem 1.** Consider $\Theta$ where $v_i = v(c_i)$ for convex $v(\cdot)$ and $C$ is also convex. Then,

$$SW(A^{**}) \leq SW(A^*) + \frac{\Lambda c_{max} w_{max}}{z - 1} \leq SW(A^{RIC}) + \frac{3\Lambda c_{max} w_{max}}{2(z - 1)} + 2v_{max} \quad (4.39)$$

**Proof of Theorem:** The solution $A^{**}$ is well understood. In particular, it is well known that (i) it is welfare maximizing to strictly prioritize customers and schedule them in order of their delay sensitivity. In $A^{**}$ let customer $i$ be assigned to $w^{**}(i)$. From this allocation we can easily construct an assignment $\bar{A}$ which is a feasible solution of $AP$ with service levels $W_z$ by assigning $i$ to the 'next best' service level:

$$\bar{A}(i, w) = \begin{cases} 
\lambda_i & \text{if } w = \min \{w \in W_z : w \geq w^{**}(i)\} \\
0 & \text{otherwise}
\end{cases}$$

For each $i$, the welfare change from $A^{**}$ to $\bar{A}$ is less than $\frac{c_i w_{max}}{z - 1}$ since $\frac{c_i w_{max}}{z - 1}$ is the maximum increase in delay. Then

$$SW(A^*) \geq SW(\bar{A}) \geq SW(A^{**}) - \sum_{i \in \Theta} \frac{c_i w_{max}}{z - 1}$$
satisfying the first inequality in Equation 4.39.

To show the second inequality holds, prices \( \bar{p}_{RIC} \) must be shown to exist incentive compatible with a (possibly suboptimal) RIC assignment \( \bar{A}^{RIC} \) where \( SW(A^*) \leq SW(\bar{A}^{RIC}) + \frac{\Lambda c_{max} w_{max}}{2(z-1)} \). First note that prices \( p^* \) exist incentive compatible with \( A^* \). Let \( \bar{w}_i \) and \( w_i \) be the maximum and minimum service level \( i \) is allocated to in \( A^* \). If \( \bar{w}_i = w_i \) the assignment of this type is feasibly compatible with a RIC assignment. Otherwise, \( \bar{w}_i > w_i \) and the pair of service levels are neighbouring (if allocated to three or more service levels, an interchange argument suffices to show that welfare can be increased by removing a service level without affecting operational feasibility). Then, consider two cases depending on whether \( A^*(i, w) \geq A^*(i, \bar{w}) \):

(i) If \( A^*(i, w) \geq A^*(i, \bar{w}) \) then the RIC assignment where \( \bar{A}(i, w) \geq A^*(i, \bar{w}) = (A^*(i, w) + A^*(i, \bar{w}))/2 \) is operationally feasible and RIC compatible with \( p^* \). The welfare from \( i \) in \( \bar{A} \) is decreased by at most \( \frac{\Lambda c_{i} w_{max}}{2(z-1)} \) since less than one half of customers of type \( i \) are affected.

(ii) Otherwise, \( A^*(i, w) < A^*(i, \bar{w}) \) in which case prices can be adjusted such that only the lower quality option is utility maximizing again at a welfare cost of at most \( \frac{\Lambda c_{i} w_{max}}{2(z-1)} \).

Let \( W_0 = \{ w : \exists is.t.A(i, w) > 0, U(w, i) = 0 \} \) be the set of delays allocated to zero utility customers. Let \( W_2 = \{ \bar{w}_i : A^*(i, w) < A^*(i, \bar{w}) \} \) be the set of low quality options compatible with case (ii). Then consider the following prices

\[
\bar{p}(w) = p^*(w) + \epsilon |W_0| - \frac{\epsilon}{|\Theta|} |\{ w' : w' > w \} W_2|.
\]

For a small enough \( \epsilon \) the zero utility customers are eliminated, all case (ii) customers are relegated to their higher delay choice (\( \bar{w} \)) and the assignment of all other customers is not affected. Summing up the impact of these changes, \( SW(A^*) \leq SW(\bar{A}) + \frac{\Lambda c_{max} w_{max}}{2(z-1)} + 2v_{max} \) leading to the desired result.

\[ \square \]

The results in Theorem 1 show that under reasonable assumptions, the bound on the approximation error decreases linearly in the number of service levels. This provides evidence supported in the following empirical results that there is unlikely to be much loss from the
restrictions placed on the assignment. The pathological example discussed above is unlikely to be relevant in practice.

4.4 Adaptive Pricing Protocol

The goal of this section is to take advantage of the results developed above to construct an adaptive pricing protocol to discover prices which lead to a near welfare maximizing assignment. The idea developed here is to decrease the apparent-loss over a series of iterations. In addition to providing a bound on welfare loss, the apparent loss is easily calculated from operational information. The apparent-loss requires an accurate calculation of the best-case revenue which has as inputs the operational capabilities of the provider described by $C(\cdot)$ and the current prices. The second component of apparent-loss is the actual revenue at current prices. The protocol is designed to operate in an environment where neither the distribution of customer types nor the identity of arriving customers is known. As a result, we maintain that the provider can not direct a customer to particular options within their optimal choice set. Rather, the customer randomly and uniformly selects from this set and the revenue is the result of the random incentive compatible assignment $A^{RIC}(p)$ for a proposed set of prices $p$. The protocol is named Min-AL and we describe it in general terms while breaking it up into an initialization, a series of price adjustment iterations and a termination condition which halts and determines the success of the algorithm.

**Initialization:** The protocol requires a set of prices which lead to customers decisions consistent with an operationally feasible assignment. In practice, it may be that there is a readily available set of historic prices already in use which a manager is looking to improve. Alternatively, the minimal amount of information required to be assured of feasible prices is an upper bound on the maximum customer valuation. Then a trivial feasible assignment where no customers participate will be result from setting $p(w) = p_o \geq \max_{i \in \Theta} v_i$. 
Price Adjustment (Iterations): Each iteration the service level which contributes most to the apparent-loss is selected. The price of that service level is decremented. The new prices result in an assignment corresponding to the new set of decisions by customers. Operational feasibility is tested. If the assignment is not operationally feasible the prices are reset to the last prices resulting in a feasible assignment and the size of the next decrement is halved. If the assignment is feasible prices are accepted and the decrement is reset.

Termination: Each iteration is followed by a continuation test. The total apparent-loss is measured and if it is smaller than a predetermined $\epsilon$ the run halts and these prices are returned. At completion, Proposition 1 shows that the final assignment is within $\epsilon$ of the actual maximum welfare for the set of service levels.

Below we provide a pseudo-code implementation of the protocol. We assume customer decisions are consistent with the RIC assignment ($A^{RIC}(p)$) and this assumption is used in the notation below.

Algorithm Skeleton 1. $Min\text{-}AL(\delta_o, p_o)$

1: Set $p^* = p_o$, $\delta = \delta_o$
2: While $(\sum_{w \in W} AL(w, A^{RIC}(p^*), p^*) > \epsilon)$:
3:   Set $\bar{w} = \arg\max_{w \in W} \{AL(w, A^{RIC}, p^*)\}$
4:   Set $\rho$ s.t. $\rho(w) = \begin{cases} p^*(w) - \delta & \text{if } w = \bar{w} \\ p^*(w) & \text{otherwise} \end{cases}$
5:   If $\forall x \in W$, $\sum_{w \in W : w \leq x} \sum_{i \in \Theta} A^{RIC}(i, w) \geq C(\sum_{w \in W : w \leq x} \sum_{i \in \Theta} A^{RIC}(i, w))$
6:      Then: Set $p^* = \rho$, $\delta = \delta_o$
7:   Else: Set $\delta = \delta/2$
8: Return $p^*$

The termination parameter $\epsilon$ is a positive real parameter. The potential price function $\rho$ is the perturbation of the current price function $p$ such that the price of the service level which contributes most to apparent-loss $\bar{w}$ is changed to $p(\bar{w}) - \delta$. Line 6 ensures that the allocation resulting from the change is operationally feasible.
Of note there always exists a decrement which decreases apparent-loss. By focusing on the largest contributor to the apparent-loss, the algorithm is intuitively selecting the most overpriced/most underutilized service level. An increase in utilization at this service level is unlikely to result in infeasibility.

**Description of empirical methodology:** The operational structure described by an M/M/1 queue: $C(\lambda) = \lambda / (\mu - \lambda)$. The capacity $\mu$ was equal to 1.5 for all runs.

Two customer type distributions (examples 1 and 2) are studied. In both cases one hundred customer types are selected. In both cases, the arrival rates $\lambda_i$ are selected uniformly in $[0, 1]$ and normalized such that $\sum_{i \in \Theta} \lambda_i = 1.2$. The delay sensitivities $c_i$ are selected independently and uniformly from $[0, 1]$. In example 1 the base value and delay sensitivity are uncorrelated. The base value $v_i$ are selected independently and uniformly from $[1, 5]$. In example 2 the base value and delay sensitivity are perfectly affinely correlated such that $v_i = 1 + 4c_i$.

Prices are initialized to $\max_{i \in \Theta} \{v_i\}$ to ensure that initially no customers participate. At each iteration the customer behaviour is consistent with the RIC assignment.

The approximation of the feasible space of demand vectors is very important for accurately assessing the apparent-loss. In the approximate version of the problem analyzed in section 4.3, the space of feasible demand vectors was described by linear combinations of a basis of demand vectors $D$. In this computational exercise, we also approximate the feasible space however we now do so in a more direct fashion by approximating $C(\lambda)$ with a piecewise linear function. The approximation is below $C(\lambda)$ everywhere and is tangent at 101 evenly spaced points between 0 and $\Lambda$. In practice, $\Lambda$ does not need to be known, only a reasonable upper bound. This function is added to a linear programming formulation to determine the best-case revenue in the calculation of apparent-loss (Step 2 of the algorithm) and to determine operational feasibility (Step 5 of algorithm). The examples are implemented in JAVA and Gurobi is used to solve linear programming subproblems.
Discussion of empirical examples: The results of Example 1 are shown in Figure 4.1 (uncorrelated base value and delay sensitivity). In the first panel the set of 100 potential types are shown scattered on base value and sensitivity axes. Multiple types, denoted by diamonds in Figure 4.1a, balked in the final welfare maximizing allocation. Customer types denoted by squares correspond to customers where participation is incentive compatible. Consistent with known results in the screening literature (see Mussa and Rosen, 1978), the participation constraints draw an increasing concave line partitioning the type space into higher value more patient customers who are participating and lower value, impatient customers who balk.

The second panel shows the algorithms trajectory where the x-axis is the iteration and y-axis denotes the value of various measures at the prices selected for each iteration. Measures in Figure 4.1b include the apparent-loss (AL), current welfare (CW), the apparent-loss plus current welfare (AL+CW) and revenue (RV). As desired, AL decreases over the course of the run though, at certain points the AL appears to increase slightly because of the discrete decrement and discrete customer classes. AL+CW is an upper bound on maximum welfare and decreases monotonically throughout the run. At completion, \( AL(A_{RIC}^R(p^*)) \leq \epsilon = 0.01 \)
and the algorithm is approximately welfare maximizing. Initially, the revenue increases in tandem with the welfare. As welfare grows however, the two lines diverge. Revenue actually decreases as welfare peaks because the increased participation resulting from lower prices is not sufficient to account for decreased prices on the already incumbent customers.

Example 2 is illustrated in Figure 4.2 (correlated base value and delay sensitivity) Panel 4.2a shows the welfare maximizing menu admits high value customers and excludes low value high patience customers. Panel 4.2b shows revenue and welfare are more closely tied than the uncorrelated case. Poor initialization as about 150 iterations elapse before the provider elicits any participation. Though, in later stages the algorithm converges very quickly.

As discussed when the model was introduced, the provider is committed to realizing the service levels selected by customers. With such a model, there is a concern that a price adjustment may result in an infeasibility—the provider may not have sufficient capacity to realize the selected service levels. Figure 4.3 shows how much slack there is in the tightest operational constraint (Constraint 4.13) for each iteration of the runs of Examples 1 and 2.
Formally the quantity is

$$\text{MinSlack} = \min_{w \in W} \left\{ \sum_{w \in W : w \leq x} w \right\} \left( \sum_{i \in \Theta} A^{RIC}(i, w) - C\left( \sum_{w \in W : w \leq x} \sum_{i \in \Theta} A^{RIC}(i, w) \right) \right\}.$$

MinSlack is labelled in the figures and shown with the apparent-loss. In both examples, as expected the minimum slack tends to decrease along with the apparent-loss. In several areas of both plots the minimum slack oscillates rapidly. This appear to be caused by the price changes alternating between a pair of higher and lower quality service levels. The resulting movement of customers back and forth between these high and low service requirements leads to a corresponding increase then decrease in the minimum slack. Despite these misbehaviours, very rarely does a price adjustment result in an infeasibility (i.e. a negative minimum slack). In Example 1, shown in Figure 4.3a, only one of the 498 iterations results in an infeasibility. In the second last iteration utilization at a single service level exceeds available capacity by 0.071 units of population weighted delay. In Example 2, shown in Figure 4.3b, the system is infeasible 15 of 421 iterations but on average by only 0.011 units of population weighted delay. To put this in context, an infeasibility of this magnitude could be rectified by moving 5% of the total arrival rate $\Lambda$ from the highest quality service level (0.2 units of delay) or 0.5% of the total arrival rate from the lowest priority service level (2 units of delay). Whether this volume and magnitude of infeasibilities is manageable will depend on the individual circumstances for the provider. The algorithm eliminates infeasibilities by returning to the last feasible state and reducing the step size for the next price adjustment. The examples find that this is effectively redirects the protocol and leads to a very good solution. It is worth investigating whether a more proactive mechanism which adjusts the step size based on the minimum slack can be used to decrease the number of infeasibilities and simultaneously the number of iterations.

### 4.5 Conclusions

We identified an adaptive pricing scheme based on the dual prices of the full information version of this assignment problem. The protocol features include (i) requiring only local
Figure 4.3: Minimum slack of feasibility constraints.

\( N = 100, \Lambda = 1.2, \mu = 1.5 \) (negative number indicates infeasibility)

operational information on the part of the provider, (ii) halting at an approximately welfare maximizing outcome. We demonstrated that this protocol can be implemented easily and applied effectively to a range of market parameters.

The protocol is based on minimizing the apparent-loss. Apparent-loss has the useful characteristic that it can be calculated without market information and requires knowledge of current prices, utilization at each service level and the operational bounds. Analytical properties of the apparent-loss include forming a bound on welfare loss and nullification implying welfare maximization. In addition to its value as a heuristic in the adaptive pricing algorithm, we suggest that this measure may be more generally valuable in assessing the welfare quality of current prices in capacitated systems.


Kakalik, J.S. 1969. Optimal dynamic operating policies for a service facility.


