PROBLEMS IN SERVICE OPERATIONS WITH HETEROGENEOUS CUSTOMERS

by

Mojtaba Araghi

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Rotman School of Management
University of Toronto

Copyright © 2014 by Mojtaba Araghi
We study three problems in service operations with heterogeneous customers. In Chapter 2, we consider the problem of locating facilities of two types at nodes of a tree network, where each facility incurs a fixed cost. Customers may need just one type of service, or both types. It is required to minimize the sum of the total transportation and fixed costs. For the problem with a single facility of each type, we present an $O(n)$ exact algorithm, and for the problem with multiple facilities of each type, we present an $O(n^5)$ algorithm.

In Chapter 3, we study the problem of maximizing profits for an inbound call center with abandonment by controlling customer acquisition, retention, and service quality via promotions, priorities, and staffing. This chapter makes four contributions. First, we develop a novel marketing-operations model of a call center that captures the evolution of the customer base as a function of past demand and queueing-related service quality. Second, we characterize the optimal controls analytically based on a deterministic fluid approximation and show via simulation that these prescriptions yield near-optimal performance for the underlying stochastic model. Third, we derive three metrics which play a key role in call center decisions and link customer and financial parameters with operational service quality, reflecting the system load and the priority policy. Fourth, we generate novel guidelines on managing a call center based on these metrics, the cost of promotions, and the capacity cost per call.

In Chapter 4, we extend our call center model by considering a firm that follows a
periodic advertising policy and receives time-varying demand response. The objective function is to maximize profit by controlling the advertisement policy and service levels for new and base customers. We show that the optimal service level allocation to new and base customers takes the form of a bang-bang policy. That is, at each moment during the advertisement period, the call center should serve all or none of the calls from each type of customers, and the policy transitions between these two extremes at most once during the advertisement period.
Dedication

To Leili

for all we shared through these years
Acknowledgements

I wish to express my profound gratitude to my supervisors, Professor Philipp Afęche and Professor Opher Baron, for their great help, support and guidance throughout these years. They have generously given their time in helping me to write the dissertation and to guide me through my professional life. I am also honored to have had the opportunity of working under the supervision of Professor Oded Berman and Professor Igor Averbakh in the second chapter of this thesis. Their irreplaceable guidance has improved this work tremendously. I truly appreciate all their help and support.

I am thankful to Professor Achal Bassamboo for accepting to act as the external examiner of this dissertation and Professor Joseph Milner and Professor Azarakhsh Malekian for serving on my committee. I would also like to thank everyone in the Rotman operations management group, particularly Professor Dmitry Krass, for his constructive comments and feedback.

Many of my friends and fellow graduate students at Rotman have contributed to this work through various discussions and helpful suggestions. In particular, I would like to thank Hossein Abouee-Mehrizi, Jianfu Wang, Adam Diamant, and Vahid Sarhangian. My Ph.D. journey would not have been as enjoyable if it wasn’t for the company of my great friends in Toronto. Aidin, Babak, Barzin, Hasti, Kosar, Pooyan, Sahar, Saharnaz, Sara, and Sara, thank you all for the happy moments we have shared together. Aidin, Hamid, Pantea, Reza, Reza, and Sina, thank you all for the delightful time we have virtually spent together.

I am forever indebted to my parents, my brother and my in-laws for all their love and moral support. I also would like to thank my best friend in life, Leili, for her endless support and encouragement. This work would have not been possible without her love and patience.
# Contents

1 Introduction

2 On the Minisum Multipurpose Trip Location Problem on Trees
   2.1 Introduction ......................................................... 4
   2.2 Notation and Problem Statement .................................. 6
   2.3 The Case of a Single Facility of Each Type .................... 8
   2.4 The Case of Multiple Facilities of Each Type ................. 10

3 Customer Acquisition, Retention, and Service Quality for a Call Center
   3.1 Introduction .......................................................... 16
   3.2 Literature Review ................................................... 19
   3.3 Model and Problem Formulation ................................... 21
      3.3.1 The Stochastic Queueing Model ............................. 23
      3.3.2 The Approximating Fluid Model ............................ 26
      3.3.3 Customer Value Metrics and Service Quality .............. 28
      3.3.4 The Profit-Maximization Problem in Terms of Customer Value
            Metrics ............................................................. 29
   3.4 Optimal Priority Policy, Staffing, and Promotion Level ........ 31
      3.4.1 Optimal Priority Policy for Fixed Staffing and Promotion Level . 31
      3.4.2 Jointly Optimal Priority and Staffing Policy for a Fixed Promotion
            Level ................................................................. 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.3 Jointly Optimal Priority, Staffing, and Promotion Policy</td>
<td>37</td>
</tr>
<tr>
<td>3.4.4 Ignoring the Effect of Service Probability on Customer Lifetime Value</td>
<td>38</td>
</tr>
<tr>
<td>3.5 Extensions</td>
<td>41</td>
</tr>
<tr>
<td>3.5.1 Multiple Types of Heterogeneous Base Customers</td>
<td>41</td>
</tr>
<tr>
<td>3.5.2 Quality-Driven Word-of-Mouth Effects on New Customer Acquisition</td>
<td>48</td>
</tr>
<tr>
<td>3.6 Fluid Model Validation: Simulation Results</td>
<td>53</td>
</tr>
<tr>
<td>3.6.1 Representative Case 1: Higher OTV for New Customers</td>
<td>54</td>
</tr>
<tr>
<td>3.6.2 Representative Case 2: Higher OTV For Base Customers</td>
<td>55</td>
</tr>
<tr>
<td>3.6.3 Robustness of Fluid Model Performance</td>
<td>56</td>
</tr>
<tr>
<td>3.7 Concluding Remarks</td>
<td>58</td>
</tr>
<tr>
<td>3.8 Appendix: Proofs</td>
<td>61</td>
</tr>
<tr>
<td>4 Customer Acquisition and Service Quality for a Call Center with Time-Varying Demand Response</td>
<td>65</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>65</td>
</tr>
<tr>
<td>4.2 Model Formulation</td>
<td>67</td>
</tr>
<tr>
<td>4.2.1 Fluid Model and System Evolution</td>
<td>70</td>
</tr>
<tr>
<td>4.2.2 The Optimization Problem</td>
<td>72</td>
</tr>
<tr>
<td>4.2.3 Service Value of Base or New Customers’ Calls</td>
<td>73</td>
</tr>
<tr>
<td>4.3 The Single Period Problem</td>
<td>75</td>
</tr>
<tr>
<td>4.3.1 Stationary Service Policy</td>
<td>76</td>
</tr>
<tr>
<td>4.3.2 Optimal Service Policy for a Fixed Promotion Level</td>
<td>77</td>
</tr>
<tr>
<td>4.3.3 Jointly Optimal Advertisement and Service Policy</td>
<td>83</td>
</tr>
<tr>
<td>4.4 Multi-period Problem</td>
<td>85</td>
</tr>
<tr>
<td>4.4.1 The Infinite Horizon Problem</td>
<td>86</td>
</tr>
<tr>
<td>4.5 Fluid Model Validation: Simulation Results</td>
<td>88</td>
</tr>
<tr>
<td>4.5.1 Parameter Values</td>
<td>88</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Summary of notation. ........................................... 24
3.2 Parameter values for representative Case 1: higher OTV for new customers. 54
3.3 Fluid model vs. simulation: minimum, maximum, and average percentage profit loss. .................................................. 58
4.1 Summary of notation. ........................................... 69
4.2 Parameter values for simulation. ................................ 88
4.3 Fluid model maximum approximation error, given service policy $q_n = q_b = 1$. 91
4.4 Fluid model maximum approximation error, given $\lambda_n(0) = 10,000$. .......... 91
List of Figures

2.1 An illustration for the problem. ................................. 7

3.1 Flow of new and base customers through the system. ........... 23
3.2 Optimal Priority policy as function of system load and new vs. base customer OTVs. ........................................ 33
3.3 Operating profit as a function of the service capacity and priority policy. .... 37
3.4 Ignoring the effect of service probability on customer lifetime value. ....... 40
3.5 Comparison of feasible capacity allocation regions: basic model and model with word-of-mouth effects on new customer acquisition. ............... 51
3.6 Considering word-of-mouth effects on new customer acquisition. ........... 52
3.7 Fluid model vs. simulation: percentage errors in optimal number of servers and new customer arrival rate, and percentage profit loss, as functions of the capacity cost (priority to new customers). ......................... 56
3.8 Fluid model vs. simulation: percentage errors in optimal number of servers and new customer arrival rate, and percentage profit loss, as functions of the capacity cost (priority to base customers). ......................... 57

4.1 Flow of new and base customers through the system. ............ 68
4.2 Value of serving a base customer call under optimal service policy. .... 80
4.3 Number of base customers and number of servers, given service policy $q_n = q_b = 1$. 90
4.4 Service probability realized by new and base customers, given service policy

\[ q_n = q_b = 1 \]

4.5 Number of base customers and number of servers, given \( \lambda_n(0) = 10,000 \)

4.6 Service probability realized by new and base customers, given \( \lambda_n(0) = 10,000 \)
Chapter 1

Introduction

In this thesis, we study three problems in service operations with heterogeneous customers.

In Chapter 2 titled “On the Minisum Multipurpose Trip Location Problem on Trees”, we focus on a facility location model with heterogeneous customers. Most classic facility location models assume that while customers differ in their demand intensities, they are homogeneous in terms of the nature of their demand, i.e., the product or the service they are asking for. Therefore, these models assume a single type of facility to fulfill customer demand. However, in practice, we usually encounter systems that consist of multiple types of facilities, and heterogeneous customers who have different types of demand.

In Chapter 2, we consider a facility location problem with two different types of facilities to be located at nodes of a tree network. Customers, also residing at nodes of the tree, are heterogeneous and can be categorized into three different groups: Customers in two of the groups each need only a single type of service provided by one facility type. Customers in the third group need services provided by both facility types, and they prefer to visit facilities of both types on a single trip to minimize their own transportation costs. We also assume that there is a fixed maintenance (or setup) cost associated with each facility type. The objective is to minimize the sum of the total weighted travel distance of all trips and the total fixed cost.

For the minisum multipurpose trip location problem where we are looking to locate only a single facility of each type, we establish an exact algorithm with linear complexity.
This improves upon the $O(n^3)$ algorithm presented in Berman and Huang (2007). For the problem with multiple facilities of each type, we determine the structural properties of the problem, and then use bottom-up dynamic programming to obtain the first polynomial exact algorithm for this problem, with complexity $O(n^5)$.

In Chapter 3, titled “Customer Acquisition, Retention, and Service Quality for a Call Center”, we consider heterogenous customers in a call center setting. Call centers are an integral part of many businesses, and the quality of the service they provide for customers can have a dramatic impact on customer loyalty. For example, Anton et al. (2004) reported that 92% of customers base their opinion of a company on their call center service experiences.

There is a vast literature on call center optimal staffing and allocation policies, and also on marketing and operational controls in call centers. Although several researchers have considered heterogenous customers, these call center and queueing research streams ignore the impact of service quality on customer retention, i.e., a firm’s customer base is independent of past interactions.

We consider an inbound call center that serves two types of customers. Base customers are part of the firm’s customer base and repeatedly interact with the call center. New customers are first-time callers who may turn into base customers. The firm’s manager aims to maximize the steady-state profit rate under three stationary controls: by controlling the staffing policy, the manager sets the call center service capacity; by controlling the promotion policy, the manager controls the new customer call arrival rate; and by controlling the priority policy, the manager prioritizes new or base customer calls.

To tackle this steady-state profit maximization problem, we propose a novel call center model in which the customer base depends on past demand and queueing-related service quality. We approximate this stochastic queueing network by a deterministic fluid model. We derive customer value metrics that are the base for the call center decisions, and depend not only on customer behavior and financial parameters, but also on operations through queueing related service quality. Based on these customer value metrics, we analyze the fluid model and characterize the optimal controls analytically. Finally, we show via simulation study that the optimal controls prescribed by the fluid model generate near-optimal profit for the underlying stochastic model, and that the insights from our analysis carry through to the stochastic model.
In Chapter 4, titled “Customer Acquisition and Service Quality for a Call Center with Time-Varying Demand Response”, we continue to investigate optimal acquisition and retention policies for an inbound call center with heterogenous customers. However, instead of focusing on steady-state system and stationary policies, in this chapter, we consider a firm that follows a periodic advertisement policy, where the demand response to the advertisement is changing over time during each period. We also assume that the firm may determine the service levels for new and base customers during each period by continuously adjusting the staffing level.

To obtain a tractable model, we approximate the resulting queueing network by a pointwise stationary fluid model (Bassamboo et al. 2009) that assumes our system reaches steady-state at each instant in time. This yields a dynamic model that can be solved as an optimal control problem. We extend the customer value metrics of Chapter 3, to consider non-stationary policies and time-varying demand response. We establish that a bang-bang service policy is optimal for a single- or multi-period problem, and derive the optimal promotion policy over different advertisement periods. Preliminary numerical results demonstrate that the optimal controls and insights that are generated by our approximations are valid for the stochastic model.
Chapter 2

On the Minisum Multipurpose Trip Location Problem on Trees

2.1 Introduction

Suzuki and Hodgson (2005) and Berman and Huang (2007) studied multipurpose trip location problems with two different types of facilities which provide two different types of service (see also Huang (2005)). Customers residing at nodes of a network may need one type of service, or both types. In the latter case, to minimize transportation times/costs, the customers visit facilities of both types on a single trip. The problem considered in Berman and Huang (2007) and Suzuki and Hodgson (2005), is to find locations for facilities of both types to minimize the total weighted travel distance of all trips. Integer linear programming formulations for the problem on general networks and computational results were presented both in Suzuki and Hodgson (2005) and Berman and Huang (2007). Berman and Huang (2007) also presented an analysis of the problem with a single facility of each type on a tree, and an $O(n^3)$ exact algorithm for solving it, where $n$ is the number of nodes.

Multipurpose trip location problems are related to round-trip location problems with only one type of facility to locate, where facilities of another type already exist in the network. A round-trip starts from a new facility (e.g., a distribution center), then goes to an existing facility (e.g., a collection depot where a package is loaded), continues to the customer (e.g., to deliver a package), and then returns back to the new facility (see Chan...
and Hearn (1977), Drezner and Wesolowsky (1982) and Wang et al. (2012) for further details). Round-trip location problems can be modeled as multipurpose trip location problems where all trips are multipurpose. One type of facility already exists, there is only one type of customer and only one type of new facility needs to be located.

The multipurpose trip location problem studied in this chapter has numerous applications discussed in Suzuki and Hodgson (2005) and Berman and Huang (2007). One example discussed in Berman and Huang (2007) is planning locations of book stores and coffee shops. By coordinating their locational decisions, book store companies and coffee shop companies can increase customer satisfaction (and, therefore, profits) since there are many customers who want to use both types of service on the same trip. This is illustrated by the popular colocation of coffee shops and book stores. Other examples include locating apparel and hardware stores by companies like Sears Canada (Berman and Huang (2007)), or supporting vertical integration for two companies that provide two partially complementary products (Berman and Huang (2007)); details can be found in (Berman and Huang (2007)).

Extensions of the problem were discussed in the recent papers Najafi et al. (2011) and Tong et al. (2012). Najafi et al. (2011) extend the problem by allowing competition where new facilities compete with existing facilities and the objective is to maximize the market share for the new facilities. Customers choose a facility to patronize probabilistically according to the distance to it. A hybrid genetic algorithm is used to solve the problem. Tong et al. (2012) study the flow interception version of the problem where travel flows are used instead of nodes' weights.

The purpose of this chapter is to continue the investigation of the problem on trees, initialized in Berman and Huang (2007). We consider a somewhat more general setting where facilities of both types incur fixed costs (that can be interpreted as set-up or maintenance costs) that depend on the nodes where the facilities are located, and the objective is to minimize the sum of the total weighted travel distance of all trips and the total fixed cost. We present two results in this note. First, for the case of a single facility of each type, in Section 3 we present an $O(n)$ exact algorithm, which improves upon the $O(n^3)$ algorithm presented in Berman and Huang (2007) for this case. Second, for the case of multiple facilities of each type on a tree, in Section 4 we present the first polynomial exact algorithm with complexity $O(n^5)$. The latter algorithm is based on
a combination of bottom-up dynamic programming that is typically used for minisum multifacility location problems such as the uncapacitated facility location problem or \( p \)-median on trees (see, e.g., Billionnet and Costa (1994), Kariv and Hakimi (1979), or Mirchandani (1990) Section 3.8.3) with use of structural properties of the problem.

### 2.2 Notation and Problem Statement

Let \( T \) be a tree network with \( V \) the set of nodes, \(|V|=n\). For any \( v_1, v_2 \in V \), let \( P(v_1, v_2) \) be the set of nodes of the only path between \( v_1 \) and \( v_2 \) (including \( v_1, v_2 \)), and let \( d(v_1, v_2) \) be the length of this path, i.e., the tree distance between \( v_1 \) and \( v_2 \). For any \( v, x, y \in V \), let \( d'(v, x, y) = d(v, x) + d(x, y) + d(y, v) \), and for any \( X \subseteq V \), \( Y \subseteq V \), let \( d(v, X) = \min_{x \in X} d(v, x), \ d'(v, X, Y) = \min_{x \in X, \ y \in Y} d'(v, x, y) \).

Facilities of two types, Type 1 and Type 2, need to be located at nodes of \( T \). Keeping a facility of Type \( i \) \((i \in \{1, 2\})\) at a node \( v \) incurs a fixed cost \( s_i(v) \) (a set-up cost or maintenance cost or their combination). The facilities will serve customers located at nodes, who will travel from their home nodes to the facilities to consume service and then back to the home nodes. There are customers of three types: A, B, and C. A-customers (B-customers) need only service from Type 1 (Type 2) facilities, and C-customers need both kinds of service. For any \( v \in V \), let \( w_A(v), w_B(v), w_C(v) \) be the average numbers of customers of types A, B, C, respectively, at node \( v \) on a typical day. We interpret the weights \( w_A(v), w_B(v), w_C(v) \) as average numbers of customers because the demand for each type of service can fluctuate from day to day and we are interested in long-term, or average, costs. \( X \) and \( Y \) will denote the sets of nodes where facilities of Type 1 and Type 2, respectively, are located (the location sets). Thus, an A-customer (B-customer) travels to an \( x \in X \) (a \( y \in Y \)) and returns home; a C-customer travels to an \( x \in X \) and to an \( y \in Y \) on a single trip and then returns home. The order of visiting \( x \) and \( y \) for a C-customer does not affect the distance traveled and thus is unimportant. It is assumed that distances are scaled so that for each customer the transportation cost is equal to the distance traveled, and that customers choose facilities to patronize so as to minimize their transportation costs.

**Problem P.** Find location sets \( X \subseteq V \) and \( Y \subseteq V \) for facilities of Type 1 and Type
2, respectively, so as to minimize the total fixed and transportation cost

\[
Z(X, Y) = \sum_{x \in X} s_1(x) + \sum_{y \in Y} s_2(y) + \sum_{v \in V} \left( 2w_A(v)d(v, X) + 2w_B(v)d(v, Y) + w_C(v)d'(v, X, Y) \right)
\]  

(2.1)

Let \( Z^* \) denote the optimum objective value for Problem P.

To illustrate the objective function, consider the tree in Figure 2.1, and suppose that for all nodes \( v \) of the tree \( w_A(v) = w_B(v) = w_C(v) = 1 \) and \( s_1(v) = s_2(v) = 1 \), and all edges have length 1. Consider \( X = \{c\} \) and \( Y = \{f\} \); then, the corresponding objective function value is 68 (the fixed cost is 2, the transportation costs for customers of types A, B, C are 14, 22, 30, respectively).

Figure 2.1: An illustration for the problem.

Let \( q \) be an arbitrarily chosen node of \( T \) which is not a leaf; this node is called the root. For any node \( v \neq q \), let \( f_v \) denote its father; that is, the node of \( P(v, q) \) adjacent to \( v \); \( v \) is called a son of \( f_v \). A node has at most one father (exactly one if it is not the root) but may have several sons. For any \( v \in V \), let \( D(v) \) denote the set of all descendants of \( v \), that is, the set of all nodes \( a \) such that \( v \in P(a, q) \). Observe that any node is a descendant of itself. For any node \( v \), we define its level as \( l(v) = \hat{l} - |P(v, q)| + 1 \), where \( \hat{l} = \max_{a \in V} |P(a, q)| \). Thus, the root \( q \) has the maximal level \( \hat{l} \). For any adjacent nodes \( a, b \), let \( B(a, b) = \{c \in V \mid b \in P(a, c)\} \).

For any integers \( k, p, k \leq p \), let \( [k : p] = \{k, k+1, ..., p\} \). We use the usual convention \( \min_{x \in \emptyset} F(x) = +\infty \) for any function \( F(x) \).
2.3 The Case of a Single Facility of Each Type

In this section, we consider Problem P with the additional restriction \(|X| = |Y| = 1\); this version will be called \textbf{Problem P}(1,1). We do not assume the tree distances to be pre-computed in advance; the necessary distances are computed during the course of the algorithm. Let \(x \in V\) \((y \in V)\) be the location of the only facility of Type 1 (Type 2). We denote \(Z(x, y) = Z(X, Y)\) where \(X = \{x\}, Y = \{y\}\). \(Z^*(1,1)\) will denote the optimal objective value for Problem P(1,1). Let

\[
F_1(x) = \sum_{v \in V} [(2w_A(v) + w_C(v)) d(v, x)];
\]

\[
F_2(y) = \sum_{v \in V} [(w_C(v) + 2w_B(v)) d(v, y)];
\]

\[
F_3(x, y) = \left(\sum_{v \in V} w_C(v)\right) d(x, y) = W_C d(x, y),
\]

where \(W_C = \sum_{v \in V} w_C(v)\). Observe that

\[
Z(x, y) = F_1(x) + F_2(y) + F_3(x, y) + s_1(x) + s_2(y). \tag{2.2}
\]

**Observation 2.1.** All values \(F_1(v), F_2(v)\) for all \(v \in V\) can be obtained in \(O(n)\) total time.

**Proof.** It is straightforward to obtain value \(F_1(v)\) for some \(v \in V\) in \(O(n)\) time, and to obtain values \(W_A(B(a, b)) = \sum_{v \in B(a, b)} (2w_A(v) + w_C(v))\) for all pairs of adjacent nodes \(a, b\) in \(O(n)\) total time. Now, observe that for adjacent \(a, b \in V\),

\[
F_1(a) - F_1(b) = [W_A(B(a, b)) - W_A(B(b, a))] d(a, b) \tag{2.3}
\]

which can be computed in \(O(1)\) time if the values \(W_A(B(a, b))\) and \(W_A(B(b, a))\) are already known. Thus, all values \(F_1(v)\) can be computed in \(O(n)\) total time, by computing \(F_1(v)\) for a specific node \(v\) first, then for the nodes adjacent to \(v\) using (2.3), etc. The same about values \(F_2(v), \ v \in V\). ■

Observation 2.1 implies that Problem P(1,1) can be solved in \(O(n^2)\) time by straight-
forward enumeration. Below, we present an $O(n)$ algorithm.

For any $v \in V$ and $i \in \{1, 2\}$, define

$$\beta_i(v) = \min_{a \in D(v)} (s_i(a) + F_i(a) + W_C d(a, v)),$$  \hspace{1cm} (2.4)

and let $a_i(v)$ be a minimizer in (2.4).

**Theorem 2.1.** (a) $Z^*(1, 1) = \min_{v \in V} (\beta_1(v) + \beta_2(v))$;

(b) If $v^* \in \arg \min \{\beta_1(v) + \beta_2(v) \mid v \in V\}$, then $(x = a_1(v^*), y = a_2(v^*))$ is an optimal solution for Problem P(1,1).

**Proof.** First, observe that

$$\beta_1(v) + \beta_2(v) \geq Z^*(1, 1) \text{ for any } v \in V,$$  \hspace{1cm} (2.5)

because $\beta_1(v) + \beta_2(v) \geq Z(a_1(v), a_2(v))$ since $d(a_1(v), a_2(v)) \leq d(a_1(v), v) + d(v, a_2(v))$. Consider now an optimal solution $(x^*, y^*)$ to Problem P(1,1), and let $v'$ be the closest to $q$ node of $P(x^*, y^*)$. Then,

$$Z^*(1, 1) = [s_1(x^*) + F_1(x^*) + W_C d(x^*, v')] + [s_2(y^*) + F_2(y^*) + W_C d(y^*, v')] \geq \beta_1(v') + \beta_2(v'),$$

according to the definition of $\beta_i(v)$. On the other hand, as observed above, $\beta_1(v') + \beta_2(v') \geq Z^*(1, 1)$. Therefore, $\beta_1(v') + \beta_2(v') = Z^*(1, 1)$. Along with (2.5), this implies statement (a) of the theorem. Statement (b) is now straightforward. ■

Theorem 2.1 implies that if values $\beta_1(v)$, $\beta_2(v)$ and the corresponding $a_1(v)$, $a_2(v)$ have been computed for all $v \in V$, then $Z^*(1, 1)$ and an optimal solution to Problem P(1,1) can be found in $O(n)$ time. To obtain an $O(n)$ algorithm for Problem P(1,1), it remains to find all values $\beta_1(v)$, $\beta_2(v)$ and the corresponding $a_1(v)$, $a_2(v)$ in $O(n)$ time.

The following recursive relation expresses $\beta_i(v)$, $i \in \{1, 2\}$, via the corresponding values for the sons of $v$.

**Lemma 2.1.** For any $v \in V$

$$\beta_i(v) = \min \left\{ s_i(v) + F_i(v), \min_{b \in \text{SON}(v)} (\beta_i(b) + W_C d(b, v)) \right\}, \quad i \in \{1, 2\},$$  \hspace{1cm} (2.6)
where SON(v) is the set of the sons of v.

**Proof.** The lemma follows from (2.4) and the observation that for any \( a \in D(v) \) such that \( a \neq v \), the node adjacent to \( v \) in the path \( P(a, v) \) is a son of \( v \). □

Lemma 2.1 implies that if \( s_i(v) + F_i(v) < \min_{b \in SON(v)} (\beta_i(b) + WCd(b, v)) \), then \( a_i(v) = v \), and if \( s_i(v) + F_i(v) \geq \min_{b \in SON(v)} (\beta_i(b) + WCd(b, v)) \), then we can set \( a_i(v) = a_i(b^*) \), where \( b^* \in \arg \min_{b \in SON(v)} (\beta_i(b) + WCd(b, v)) \). Thus, \( \beta_i(v), a_i(v), i \in \{1, 2\} \) can be obtained in \( O(|SON(v)|) \) time from the corresponding values for the sons of \( v \). Therefore, all \( \beta_i(v), a_i(v), i \in \{1, 2\} \) for all \( v \in V \) can be obtained in \( O(n) \) total time by a standard bottom-up dynamic programming, starting from the nodes \( v \) at level 1 and ending at the root node \( q \). (Note that for a leaf \( v \), \( SON(v) = \emptyset \) and therefore \( \beta_i(v) = s_i(v) + F_i(v) \) and \( a_i(v) = v \).) Then, using Theorem 2.1, the optimal objective value \( Z^*(1, 1) \) and an optimal solution are obtained in \( O(n) \) time. We obtain the following result.

**Theorem 2.2.** Problem \( P(1, 1) \) can be solved in \( O(n) \) time.

### 2.4 The Case of Multiple Facilities of Each Type

In this section, we consider the general version of Problem \( P \) on a tree, without restrictions on the numbers of facilities to be located. For any \( v \in V \), let \( x_A(v) (y_B(v)) \) be the facility patronized by A-customers (B-customers) from \( v \), and let \( (x_C(v), y_C(v)) \) be the pair of facilities patronized by C-customers from \( v \). These facilities are chosen by customers to minimize the corresponding travel costs. For simplicity of presentation, we assume that there are no ties and the customer allocations \( x_A(v), y_B(v), x_C(v), y_C(v) \) are defined uniquely if \( X, Y \subset V \) are given. This is not a restrictive assumption, since a small perturbation of distances or a lexicographic rule can be used to break ties. The algorithm developed in this section has complexity \( O(n^5) \), so in this section for simplicity we assume that the tree distances for all pairs of nodes are pre-computed; it takes \( O(n^2) \) time using the depth-first search (Ahuja et al. 1993). Below we present a number of structural properties that will be used later to develop a polynomial algorithm.

**Lemma 2.2.** For any node \( a \neq q \), the following properties hold (\( v = f_a \) is the father of \( a \):
Property B1. If $x_A(v) \in D(a)$, then $x_A(a) = x_A(v)$; otherwise, $x_A(a) \in D(a) \cup \{x_A(v)\}$.

Property B2. If $y_B(v) \in D(a)$, then $y_B(a) = y_B(v)$; otherwise, $y_B(a) \in D(a) \cup \{y_B(v)\}$.

Property B3. If $[x_C(v) \in D(a)$ and $y_C(v) \in D(a)]$, then $x_C(a) = x_C(v)$ and $y_C(a) = y_C(v)$.

Property B4. If $[x_C(v) \notin D(a)$ and $y_C(v) \notin D(a)]$, then either
$$[x_C(a) \in D(a)$ and $y_C(a) \in D(a)],$$
$$[x_C(a) = x_C(v) \text{ and } y_C(a) = y_C(v)],$$
$$[x_C(a) \in D(a)$ and $y_C(a) = y_B(v)],$$
$$[y_C(a) \in D(a)$ and $x_C(a) = x_A(v)].$$

Property B5. If $x_C(v) \in D(a)$ and $y_C(v) \notin D(a)$, then either
$$[x_C(a) \in D(a)$ and $y_C(a) \in D(a)],$$
$$[x_C(a) = x_C(v) \text{ and } y_C(a) = y_C(v)].$$

Property B6. If $y_C(v) \in D(a)$ and $x_C(v) \notin D(a)$, then either
$$[y_C(a) \in D(a)$ and $x_C(a) \in D(a)],$$
$$[y_C(a) = y_C(v)$ and $x_C(a) = x_C(v)].$$

**Proof.** For proving the lemma, we use the assumption above that there are no ties and therefore customer allocations are defined uniquely. Properties B1-B2 follow from the observation that if $x$ is the closest node from an $X \subset V$ to a node $v_1$, then $x$ is the closest node from $X$ to any node $v_2 \in P(v_1, x)$. To see Property B3, observe that a better allocation for C-customers of node $a$ than $x_C(v), y_C(v)$ would result in the possibility of a better allocation for C-customers of node $v$. To see Property B4, observe that if the condition of the property holds, then:

a) If $x_C(a) \notin D(a)$ and $y_C(a) \notin D(a)$, then C-customers from $a$ start their trip by going from $a$ to $v$, and end it by going from $v$ to $a$, therefore their allocation is the same as for C-customers from $v$, i.e. $x_C(a) = x_C(v)$ and $y_C(a) = y_C(v)$.

b) If $x_C(a) \in D(a)$ and $y_C(a) \notin D(a)$, then C-customers from $a$ go outside of $D(a)$ only to consume Type B service, and go via $v$, so then $y_C(a) = y_B(v)$, because otherwise the allocation could be improved.

c) If $y_C(a) \in D(a)$ and $x_C(a) \notin D(a)$, then $x_C(a) = x_A(v)$ using an argument identical to that in the previous case.
To see Property B5, observe that if the condition of the property holds, then $x_C(v) = x_A(v)$ and $y_C(v) = y_B(v)$, because otherwise the allocation of C-customers from $v$ could be improved by setting $x_C(v) = x_A(v)$, $y_C(v) = y_B(v)$. Observe now that:

a) If $x_C(a) \in D(a)$ and $y_C(a) \not\in D(a)$, then $x_C(a) = x_C(v)$ and $y_C(a) = y_C(v)$.

b) The situation $x_C(a) \not\in D(a)$ is impossible, because then either the allocation of C-customers from $a$ could be improved, or the allocation of C-customers from $v$ could be improved.

Property B6 is verified analogously.

In the formulation of Problem P, the decision variables are the location sets $X, Y \subset V$, which define the allocations $x_A(v), y_B(v), x_C(v), y_C(v)$ for all $v \in V$. Now we give an equivalent formulation of the problem, where the decision variables are the allocations. A set $A = \{(x_A(v), y_B(v), x_C(v), y_C(v)) , v \in V\}$ which represents allocations for all nodes is called an allocation variant; the corresponding objective value is

$$Z'(A) = \sum_{v \in V} [2w_A(v)d(v, x_A(v)) + 2w_B(v)d(v, y_B(v)) + w_C(v)d'(v, x_C(v), y_C(v))] +$$

$$\sum_{v' \in \cup_{v \in V} \{x_A(v), x_C(v)\}} s_1(v') + \sum_{v' \in \cup_{v \in V} \{y_B(v), y_C(v)\}} s_2(v'),$$

and the corresponding location sets are $X = \cup_{v \in V} \{x_A(v), x_C(v)\}$, $Y = \cup_{v \in V} \{y_B(v), y_C(v)\}$.

Since location sets are defined here through the allocation variant, the allocations do not necessarily minimize the travel distances incurred by customers, e.g. $x_A(v)$ is not necessarily the closest to $v$ element of $X$.

**Problem P'.** Find an allocation variant that minimizes $Z'(A)$.

An allocation variant that satisfies properties B1-B6 for all $a \in V \setminus \{q\}$ will be called a feasible allocation variant (FAV). Consider

**Problem P''.** Find a FAV that minimizes $Z'(A)$.

Let $Z^*$ and $Z''^*$ be the optimal objective values for Problems P' and P''. Observe that Problem P' is equivalent to Problem P and $Z^* = Z^*$. Then, Lemma 2.2 implies that Problem P'' is equivalent to Problems P' and P and $Z''^* = Z^*$. Now we develop a dynamic programming procedure to find an optimal FAV.

For any $V' \subset V$, the cost incurred by the nodes from $V'$ is the total fixed cost of
the facilities located in \( V' \) plus the total transportation cost of all customers located at the nodes of \( V' \). For any \( v \in V \) and \((x_A, y_B, x_C, y_C) \in V^4\), where \( V^4 \) is the set of all 4-tuples of vertices, define value \( \alpha_v(x_A, y_B, x_C, y_C) \) as the minimum cost incurred by the nodes from \( D(v) \), where the minimum is taken over all \( \text{FAV} \) such that \( x_A(v) = x_A, y_B(v) = y_B, x_C(v) = x_C, y_C(v) = y_C \). For any node \( a \neq q \) and its father \( v = f_a \), and any \((x_A, y_B, x_C, y_C) \in V^4\), define \( Q_a(x_A, y_B, x_C, y_C) \) as the set of all \((x'_A, y'_B, x'_C, y'_C) \in V^4\) that satisfy properties B1-B6 where \( x_A(v), y_B(v), x_C(v), y_C(v) \) are replaced with \( x'_A, y'_B, x'_C, y'_C \), respectively. For any \( V' \subset V \) and \( v \in V \), define

\[
\delta(v, V') = \begin{cases} 
1 & \text{if } v \in V', \\
0 & \text{otherwise.}
\end{cases}
\]

The following result follows from Lemma 2.2 and the definitions of \( Q_a(x_A, y_B, x_C, y_C) \) and \( \alpha_v(x_a, y_B, x_C, y_C) \).

**Theorem 2.3.** Values \( \alpha_v(x_A, y_B, x_C, y_C) \) satisfy the following recursive relations:

\[
\alpha_v(x_A, y_B, x_C, y_C) = \delta(v, \{x_A, x_C\})s_1(v) + \delta(v, \{y_B, y_C\})s_2(v) + 2w_A(v)d(v, x_A) + 2w_B(v)d(v, y_B) + \frac{w_C}{2}d'(v, x_C, y_C) + \sum_{a \in \text{SON}(v)} \gamma_a(x_A, y_B, x_C, y_C),
\]

where

\[
\gamma_a(x_A, y_B, x_C, y_C) = \min \{ \alpha_v(x'_A, y'_B, x'_C, y'_C) \mid (x'_A, y'_B, x'_C, y'_C) \in Q_a(x_A, y_B, x_C, y_C) \}.
\]

If \( v \) is a leaf, then \( \text{SON}(v) = \emptyset \) and \( \sum_{a \in \text{SON}(v)} \gamma_a(x_A, y_B, x_C, y_C) = 0 \) in (2.7).

Using recursive relations (2.7), all values \( \alpha_v(x_A, y_B, x_C, y_C), v \in V \), \((x_A, y_B, x_C, y_C) \in V^4\) can be computed using the standard bottom-up dynamic programming (e.g., Kariv and Hakimi (1979)), starting with nodes \( v \) of level 1 and ending at \( v = q \). Then, the optimal objective value is

\[
Z^* = Z''^* = \min \{ \alpha_q(x_A, y_B, x_C, y_C) \mid (x_A, y_B, x_C, y_C) \in V^4 \},
\]

and an optimal solution is obtained using backtracking. Let us estimate the complexity
of this approach.

A straightforward implementation would take $O(n^9)$ time, since $|Q_a(x_A, y_B, x_C, y_C)| = O(n^4)$, and $O(n^5)$ values $\gamma_a(x_A, y_B, x_C, y_C)$ need to be computed using (2.8). A significant improvement is obtained by looking closer at sets $Q_a(x_A, y_B, x_C, y_C)$ and observing that most computations are repetitive and can be avoided.

For an $a \in V$, a $k$-dimensional $a$-slice, $k = 0, 1, 2, 3, 4$, is the set of all quadruples $(x_A, y_B, x_C, y_C)$ where some $4 - k$ components are fixed to be some specific nodes from $V$, and the remaining $k$ components vary everywhere in $D(a)$. For example, $(D(a))^4$ is the only 4-dimensional $a$-slice; $\{(x_A, y_B, x_C, y_C) \mid y_B = b, \ n_A, x_C, y_C \in D(a)\}$, where $b$ is some node from $V$, is a 3-dimensional $a$-slice. There is one 4-dimensional $a$-slice, $4n$ 3-dimensional $a$-slices, $6n^2$ 2-dimensional $a$-slices, $4n^3$ 1-dimensional $a$-slices, and $n^4$ 0-dimensional slices. Let $S_a$ be the set of all $a$-slices, $|S_a| = n^4 + 4n^3 + 6n^2 + 4n + 1$. For any $a$-slice $s \in S_a$, let $\phi(s) = \min\{\alpha_a(x'_A, y'_B, x'_C, y'_C) \mid (x'_A, y'_B, x'_C, y'_C) \in s\}$. Observe that the set $Q_a(x_A, y_B, x_C, y_C)$ is the union of a constant number of $a$-slices. Therefore, if all values $\phi(s)$, $s \in S_a$, are known, then the value $\gamma_a(x_A, y_B, x_C, y_C)$ can be obtained using (2.8) in a constant time. Observe also that all values $\phi(s)$, $s \in S_a$, can be obtained in $O(n^4)$ total time if all values $\alpha_a(x'_A, y'_B, x'_C, y'_C)$, $(x'_A, y'_B, x'_C, y'_C) \in V^4$ are already known. This implies that for any $a \in V$, if all values $\alpha_a(x'_A, y'_B, x'_C, y'_C)$, $(x'_A, y'_B, x'_C, y'_C) \in V^4$ are already known, then all values $\gamma_a(x_A, y_B, x_C, y_C)$, $(x_A, y_B, x_C, y_C) \in V^4$ can be obtained in $O(n^4)$ time. Thus, the overall dynamic programming procedure that computes all values $\alpha_v(x_A, y_B, x_C, y_C)$, $(x_A, y_B, x_C, y_C) \in V^4$, $v \in V$ can be implemented in $O(n^5)$ time. Then finding the optimal objective value using (2.9) takes $O(n^4)$ time, and the backtracking takes $O(n)$ time. We obtain the following result.

**Theorem 2.4.** Problem P can be solved in $O(n^5)$ time.

**Remark 2.1.** If there are only facilities of one type, then our problem becomes the classical uncapacitated facility location problem (UFLP). In this case, although the algorithm developed in this section is still applicable (just set the weights $w_B(v)$ and $w_C(v)$ equal to 0 for all $v \in V$), the problem is structurally simpler, and more efficient algorithms are available in the literature (see, e.g., Billionnet and Costa (1994), or Mirchandani (1990) Section 3.8.3). If there are $k$ types of facilities with $k > 2$, the problem becomes significantly more complicated, for the following reasons:

a) The order of visiting facilities of different types may become important, as some
customers may need to consume the services in a particular order. This increases the number of relevant customer classes, as different orders of visiting the same facilities generally result in different transportation costs if more than two facilities need to be visited.

b) Even with the assumption that each customer chooses the order of visiting facilities only based on transportation costs, there will still be $2^k - 1$ classes of customers, each class being defined by a specific combination of types of services needed.

Thus, the case of more than two types of facilities is beyond the scope of this chapter. However, we hope that this chapter will trigger additional research in this direction.
Chapter 3

Customer Acquisition, Retention, and Service Quality for a Call Center

3.1 Introduction

Call centers are an integral part of many businesses. By some estimates 70-80% of firms’ interactions with customers occur through call centers (Feinberg et al. 2002, Anton et al. 2004), and 92% of customers base their opinion of a company on their call center service experiences (Anton et al. 2004). More importantly, the call center service experience can have a dramatic impact on customer satisfaction and retention. Poor service is cited by 50% of customers as the reason for terminating their relationship with a business (Genesys Global Consumer Survey 2007). These findings underscore the key premise of customer relationship management (CRM), which is to view a firm’s interactions with its customers as part of ongoing relationships, rather than in isolation. As Akşin et al. (2007, p. 682) point out, “firms would benefit from a better understanding of the relationship between customers’ service experiences and their repeat purchase behavior, loyalty to the firm, and overall demand growth in order to make better decisions about call center operations.”

This paper provides a starting point for building such understanding. To our knowledge, this is the first paper to consider the impact of queueing-related service quality
on customer retention and long-term customer value. The standard approach in the call center literature has been to model a firm’s customer base as independent of past interactions. We model new and base (repeat) customers and study the problem of maximizing profits by controlling customer acquisition, retention, and service quality via promotions, priorities to new or base customers, and staffing.

This paper makes four contributions. First, we propose what seems to be the first call center model in which the customer base depends on past demand and queueing-related service quality. Second, we show via simulation that a deterministic fluid analysis yields near-optimal performance for the underlying stochastic model. This modeling framework can be extended to further problems of joint CRM and call center management. Third, we derive metrics which play a key role in call center decisions, and which link customer and financial parameters with operational service quality. Fourth, we generate novel results on how to manage a call center based on these metrics, the cost of promotions, and the capacity cost per call. We elaborate on these contributions in turn.

First, we develop a novel marketing-operations model of an inbound call center, which links elements of CRM with priority and staffing decisions. The model captures new customer arrivals in response to promotions, their conversion to base customers, the evolution and calls of the customer base, and call-related and call-independent profits and costs. Customers contribute to queueing and are impatient, which leads to abandonment and adversely affects customer retention. A notable feature of this model is that it can be tailored to a range of businesses that rely on a call center, such as credit card companies, phone service providers, or catalog marketing companies. In contrast to the marketing literature on CRM, the key novelty of our model is that customer flows and the customer base depend on the service quality, i.e., the probability of getting served, and in turn on customer acquisition, priorities, and capacity. In contrast to the call center literature, the key novelty of our model is that the customer base depends on past demand and service.

Second, we characterize the optimal controls analytically based on a deterministic fluid model approximation of the underlying stochastic queueing model, which is difficult to analyze directly. We validate these analytical prescriptions through a simulation study, which shows that they yield near-optimal performance for the stochastic system it approximates, with maximum profit losses below 1%. These results suggest that the
main insights and guidelines based on the fluid model apply to the stochastic model as well. More generally, these results suggest that our modeling and analytical approach may prove quite effective in tackling further problems in this important area.

Third, we derive three metrics which are the basis for call center decisions, the expected customer lifetime value (CLV) of a base customer and the expected one-time serving value (OTV) of a new and base customer. A key feature of these metrics is that they depend not only on customer behavior and financial parameters, but also on operations through the service quality, which reflects the system load and the priority policy. In particular, unlike standard CLV metrics in the marketing literature, the CLV in our model reflects the impact of abandonment on retention. The OTV metrics also capture interaction effects between customers’ service-related propensity to join and leave the customer base, and their call frequency while in the customer base.

Finally, we generate novel insights on how to manage a call center. We show that it is optimal to prioritize the customers with the higher OTV. A notable feature of this policy is that it accounts for the financial impact of customers’ future calls, in contrast to standard priority policies such as the $c\mu$ rule. Next, for situations where the capacity is fixed, e.g., due to lags in hiring or training, we characterize the jointly optimal promotion and priority policy as a function of the promotion cost, the CLV and OTVs, and the capacity level. We further show how the jointly optimal promotion, priority and staffing policy depend on the CLV, the OTVs, and the capacity cost. Under the optimal policy, the most striking operating regime arises if new customers have the higher OTV, e.g., due to prohibitive switching costs for base customers, and capacity is relatively expensive. Under these conditions it is optimal to prioritize new customers and to overload the system. In this regime the primary goal of the call center is to serve and acquire new customers to grow the customer base, whereas base customers receive deliberately poor service. This result lends some theoretical support for the anecdotal evidence that locked-in customers of firms such as mobile phone service providers commonly experience long waiting times when contacting the call center.

The plan of this paper is as follows. In §3.2 we review the related literature. In §3.3 we specify the stochastic queueing model and the approximating deterministic fluid model, and we formulate the firm’s profit maximization problem. In §3.4, we characterize the fluid model prescriptions on the optimal priority policy, promotion spending, and capacity
level. In §3.5 we extend the base model studied in §3.4 by consider heterogeneous base customers, and presence of service-quality-driven negative word-of-mouth effects on new customer acquisition. In §3.6 we present simulation results that evaluate the performance of the fluid model prescriptions of §3.4 against simulation-based optimization results for the stochastic system described in §3.3. Our concluding remarks are in §3.7. All proofs are in the Online Supplement.

### 3.2 Literature Review

This paper is at the intersection of research streams on advertising, CRM, and call center management. We relate our work first to these literatures, and then to operations papers outside the call center context which also consider demand as a function of past service, as we do in this paper.

There is a vast literature on advertising. We refer to Feichtinger et al. (1994), Hanssens et al. (2001), and Bagwell (2007) for surveys. In contrast to our study, the overwhelming majority of these papers ignore the firms’ supply constraints in fulfilling the demand generated by advertising. A number of papers consider advertising under supply constraints in different settings. Focusing on physical goods, Sethi and Zhang (1995) study joint advertising and production control, and Olsen and Parker (2008) study joint advertising and inventory control. Focusing on services, Horstmann and Moorthy (2003) study the relationship between advertising, capacity, and quality in a competitive market; in their model, unlike in ours, the quality attribute is independent of utilization.

CRM and models of CLV and related customer metrics are of growing importance in marketing. We refer to Rust and Chung (2006), Gupta and Lehmann (2008), and Reinartz and Venkatesan (2008) for surveys. Blattberg and Deighton (1996) develop a tool to optimize the (static) mix of acquisition and retention spending. Ho et al. (2006) derive static optimal spending policies in customer satisfaction in a model where customers’ purchase rates, spending amounts and retention depend on their satisfaction from their last purchase. Several papers study the design of dynamic policies, focusing on marketing instruments such as direct mail (cf. Bitran and Mondschein 1996), pricing (Lewis 2005), cross-selling (Günes et al. 2010), and service effort (Aflaki and Popescu 2013).
In contrast to our setup, the CRM literature ignores supply constraints and the interaction between capacity, demand, and service quality. To our knowledge, Pfeifer and Ovchinnikov (2011) and Ovchinnikov et. al. (2014) are the only papers that consider a capacity constraint. They study its impact on the value of an incremental customer and on the optimal spending policy for acquisition and retention. In contrast to our model, theirs do not consider the effect of queueing and service quality on customer acquisition, the CLV, and retention.

The call center literature is extensive and growing. We refer to Gans et al. (2003), Akşin et al. (2007), and Green et al. (2007) for surveys. The bulk of these papers focus on operational controls, i.e., staffing and allocation policies to serve an exogenous arrival process of call center requests to the system, and they often consider endogenous abandonment. Some papers consider exogenous arrivals of initial requests but model policies to manage endogenous retrials before service due to congestion (e.g., Armony and Maglaras 2004), or after service due to poor service quality on earlier calls (e.g., de Véricourt and Zhou 2005). A growing number of papers study marketing and operational controls; they consider exogenous arrivals of potential requests but model some aspects of the actual requests as endogenous. This framework characterizes the study of cross-selling which increases service times to boost revenue. (Akşin and Harker 1999 started this stream; Akşin et al. 2007 review it; Gurvich et al. 2009 jointly consider staffing and cross-selling; Debo et al. 2008 study a service time-revenue tradeoff outside the call center setting.) This framework is also standard in the stream on pricing, scheduling, and delay information policies for queueing systems in general, rather than call centers in particular (cf. Hassin and Haviv 2003). Randhawa and Kumar (2008) model both initial requests and retrials as endogenous, based on the price and service quality.

In contrast to this paper, these call center and queueing research streams ignore the impact of service quality on customer retention, i.e., a firm’s customer base is independent of past interactions. In a parallel effort, Farzan et al. (2012) do consider repeat purchases that depend on past service quality; however, in contrast to our model, they model service quality by a parameter that is independent of queueing. In the marketing literature, Sun and Li (2011) empirically estimate how the retention of customers depends on their allocation to onshore vs. offshore call centers, including on waiting and service time. In contrast to our paper, theirs does not model capacity constraints and the link to waiting
time. However, their numerical results underscore the value of considering customer retention and CLV in call center policies. Specifically, they numerically solve a stochastic dynamic program that matches customers to service centers to maximize long term profit, and show via simulation that considering customer retention and CLV can significantly improve performance.

Schwartz (1966) seems to be the first to consider how past service levels affect demand, focusing on inventory availability. The operations literature has seen a growing interest over the last decade in studying how past service levels affect demand and how to manage operations in such settings. These papers consider operations outside the call center context. As such their models are fundamentally different from ours. Gans (2002) and Bitran et al. (2008) consider a general notion of service quality and do not model capacity constraints. Gans (2002) considers oligopoly suppliers that compete on static service quality levels and models customers who switch among them in Bayesian fashion based on their service history. Bitran et al. (2008) model a price- and quality-setting monopoly and the evolution of its customer base depending on satisfaction levels and the number of past interactions. Hall and Porteus (2000), Liu et al. (2007), Gaur and Park (2007), and Olsen and Parker (2008) study equilibrium capacity/inventory control strategies and market shares of competing firms with customers that switch among them in reaction to poor service. The work of Olsen and Parker (2008) is distinct in that it considers nonperishable inventory, consumer backlogs, and firms that control not only inventory, but also advertising to attract new/reacquire dissatisfied customers. The authors study the optimality of base-stock policies for the monopoly and the duopoly case. Adelman and Mersereau (2013) study how a supplier of a physical good should dynamically allocate its fixed capacity among a fixed portfolio of heterogeneous customers who never defect, but whose stochastic demands depend on goodwill derived from past fill rates.

3.3 Model and Problem Formulation

The credit card company that motivated this research uses call centers as the main contact channel with its customers. Although the firm uses other contact channels to communicate with customers, such as SMS and e-mail alerts, these channels are typically used for requests that are of lower importance/interaction such as balance inquiries.
The focus of our model, one of the firm’s call centers, receives calls from both card holders and potential new customers. This call center operates two different queues, a sales queue and a service queue. The sales queue is mainly dedicated to calls from potential new customers in response to promotions. The service queue handles calls from existing customers, e.g., to request a credit limit increase, report a lost or stolen credit card, redeem loyalty points, and so on.

Each month, the marketing department sets the promotion spending level for that month based on some measures of customers’ profit contributions. However, in its promotion decisions marketing does not account for operational factors, such as service level considerations and the staffing cost. Furthermore, the operations department is not involved in promotion decisions.

Once a monthly promotion level is determined, the marketing department mails out offers fairly evenly over the course of that month. Offers have expiration dates and potential customers may call during this period to apply for a credit card.

Following the promotion budget decisions, the operations department forecasts the number of arriving calls to each of the queues during the month. The forecast for the new customer arrival rate considers the number of offers to be mailed out by the marketing department and other promotional mix activities that increase the brand awareness such as advertising and public relations activities. The forecast for the base customer call arrival rate considers the number of card holders.

Based on the demand forecast, the operations department decides on staffing for that month to meet certain predetermined service level targets.

This procedure of setting a promotion level, forecasting the demand, and staffing is followed each month with a rolling horizon of six months. Moreover, as far as we know, within each month there was no specific periodicity within either of these activities other than the daily and weekly fluctuations common in the demand pattern for call centers.

This decision-making procedure raises the following issues:

1. The marketing and operations departments do not coordinate their decisions.

2. Both promotion and staffing decisions either ignore service level considerations or make arbitrary assumptions about service level targets.
3. It is not clear whether to prioritize certain customers over others, and if so, on what basis.

4. There is no clarity about the value of a call, the value of a customer, and how this value depends on the call center’s service quality.

These and similar issues are pertinent not only for this particular operation but also for similar call centres. To address some of these issues we develop and analyze a model that integrates these marketing and operations decisions. We develop the model in two steps. In §3.3.1 we specify a conventional exact stochastic queueing model of the call center that captures variability in inter-arrival, service, and abandonment times. Then, in §3.3.2, we describe the approximating fluid model. Our analytical results in §3.4 and §3.5 are based on this fluid model.

### 3.3.1 The Stochastic Queueing Model

Consider a firm that serves two types of customers through its inbound call center. Base (type 1) customers are part of the firm’s customer base and repeatedly interact with the call center. New (type 0) customers are first-time callers who may turn into base customers. Figure 3.1 depicts the customer flow through the system, showing the flow of new customers by dashed lines and the flow of base customers by solid lines.

![Figure 3.1: Flow of new and base customers through the system.](image-url)
We model the call center as an \(N\)-server system. Service times for new and base customers are i.i.d. with mean \(1/\mu_0\) and \(1/\mu_1\), respectively. Calls arrive as detailed below. Customers wait in queue if the system is busy upon arrival, but they are impatient. Abandonment times are i.i.d with mean \(1/\tau_0\), and \(1/\tau_1\) for new and base customers, respectively. Table 3.1 summarizes the notation.

<table>
<thead>
<tr>
<th>System Parameters</th>
<th>Economic Parameters</th>
<th>Steady-State Performance Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N) Number of servers</td>
<td>(p_0, p_1) Profit per served call of new, base customers</td>
<td>(x_1) Average number of base customers</td>
</tr>
<tr>
<td>(\mu_0, \mu_1) Service rate of new, base customers</td>
<td>(c_0, c_1) Cost per abandoned call of new, base customers</td>
<td>(q_0, q_1) Service probability of new, base customer calls</td>
</tr>
<tr>
<td>(\lambda_0) New customers arrival rate</td>
<td>(R) Profit rate per base customer (call-independent)</td>
<td>(\Pi) Call center profit rate</td>
</tr>
<tr>
<td>(r_1) Call arrival rate per base customer</td>
<td>(C) Cost rate per server</td>
<td></td>
</tr>
<tr>
<td>(\tau_0, \tau_1) Call abandonment rate of new, base customers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\theta_{01}) P(new customer joins customer base after service)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\theta_1) P(base customer remains in customer base after abandoning)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1) Attrition rate per base customer (call-independent)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Summary of notation.

We consider the system in steady-state under three stationary controls: the staffing policy sets the number of servers at a cost of \(C\) per server per unit time, the promotion policy controls the new customer call arrival rate, and the priority policy prioritizes new or base customer calls. Let \(q_1\) and \(q_0\) denote the steady-state probability that base and new customer calls are served, respectively. We subsequently refer to each of these measures simply as a “service probability.” These service probabilities depend on the system parameters and controls as discussed below.

New customer calls arrive to the system following a stationary Poisson process with rate \(\lambda_0\). The new customer call arrival rate depends on the firm’s advertising spending. To simplify the modeling, we do not differentiate solicited and ambient demand and use the terms advertising and promotion interchangeably. Let \(S(\lambda_0)\) denote the advertising spending rate per unit time as a function of the new customer arrival rate it gener-
ates. We assume that the response of new customers to advertising spending follows the law of diminishing returns (Simon and Arndt 1980), so $S(\lambda_0)$ is strictly increasing and strictly convex in $\lambda_0$. For analytical convenience we assume that $S$ is twice continuously differentiable and $S'(0) = 0$.

The following flows determine the evolution of the customer base. A new customer who receives service joins the customer base with probability $\theta_{01} > 0$, so new customers turn into base customers at an average rate of $\lambda_0 q_0 \theta_{01}$ per unit time. The times between successive calls of a base customer are independent and exponentially distributed with mean $1/r_1$, so $r_1 \geq 0$ is the average call rate per base customer per unit time. We assume that $1/r_1 >> 1/\mu_0, 1/\mu_1, 1/\tau_0, 1/\tau_1$, i.e., the mean time between calls from any given customer is much larger than the mean service and abandonment times.

It is prevalent in the service literature to assume that base customers think about terminating their relationship with service providers only when some critical incident happens (Gremler 2004; Keaveney 1995). We consider the call abandonment as the critical incident. Similar to Hall and Porteus (2000), we capture the “customer loyalty coefficient” (Reichheld 2000), by defining $\theta_1$ such that a base customer who abandons the queue remains in the customer base with probability $\theta_1$, but immediately terminates her relation with the company and leaves the customer base with probability $1 - \theta_1$. The customer base is also subject to attrition due to call-independent reasons as well as involuntarily switching incidents (e.g., relocation and death). The lifetimes of base customers in the absence of abandonment are independent and exponentially distributed with mean $1/\gamma_1$, so $\gamma_1 > 0$ is the average call-independent attrition rate of a base customer.

Let $x_1$ denote the long-run average number of base customers in steady-state; we also call $x_1$ simply the average customer base. In steady-state the average arrival rate of base customer calls is $x_1 r_1$. The system is stable since customer impatience ensures stable queues at the call center, and the average customer base is finite since $\lambda_0 < \infty$ and $\gamma_1 > 0$.

The profit of operating the firm is as follows. On average the firm generates a profit of $p_0$ per new customer call it serves and incurs a cost of $c_0 \geq 0$ per abandonment of any call from new customers. We model two potential profit streams from base customers. First, the firm may generate a call-independent profit at an average rate of $R \geq 0$ per unit time per base customer, which captures monetary flows that are independent of call
center interactions. Second, base customers may also call with a purchase or a service request. On average the firm generates a profit of $p_1$ per base customer call it serves and incurs a cost of $c_1 \geq 0$ per abandonment of any call from base customers.

To summarize, we model the system as a two-station queueing network with two types of impatient customers and state-dependent routing. The call center itself is a $G/G/N + G$. Between successive call center visits, base customers enter an orbit that operates like a $G/M/\infty$ system with service rate $r_1 + \gamma_1$. The new customer arrival process is Markovian and state-independent. The base customer arrival processes to the call center and to the orbit depend on the new customer arrival process, the capacity, and the priority policy.

Let $\Pi$ denote the firm’s average profit rate in steady-state. It is given by

$$\Pi := \lambda_0 (p_0 q_0 - c_0 (1 - q_0)) + x_1 (R + r_1 [p_1 q_1 - c_1 (1 - q_1)]) - CN - S (\lambda_0),$$

where the first product is the profit rate from new customer calls, the second product is the profit rate from base customers (both call-independent and call-dependent), the third term is the advertising cost rate, and the last term is the staffing cost rate. The firm aims to maximize its profit rate by choosing the number of servers $N$, the new customer arrival rate $\lambda_0$ and the priority policy which affects the service probabilities $q_1$ and $q_0$.

The profit rate (3.1) depends on three stationary performance measures, the average customer base $x_1$ and the service probabilities $q_1$ and $q_0$. The state-dependent nature of customer flows and feedbacks through the system make it difficult to analyze these measures for the stochastic model, even under Markovian assumptions. We therefore approximate the stochastic model by a corresponding deterministic fluid model that we describe in §3.3.2.

### 3.3.2 The Approximating Fluid Model

In this section we formulate the fluid approximation and the resulting profit-maximization problem. In steady-state, the size of the customer base must be constant in time:

$$x'_1 (t) = \lambda_0 q_0 \theta_{01} - x_1 (t) (\gamma_1 + r_1 (1 - q_1) (1 - \theta_1)) = 0,$$
where $\lambda_0 q_0 \theta_0$ is the rate at which new customers join the customer base, and the second term in (3.2) is the customer base decay rate which is proportional to the size of the customer base. As discussed in §3.3.1, the departure rate of any base customer has two components, the service-independent attrition rate $\gamma_1$ and the call-dependent term $r_1 (1 - q_1) (1 - \theta_1)$, which is the product of a base customer’s calling rate $r_1$, abandonment probability $1 - q_1$, and probability of leaving the customer base after abandonment $1 - \theta_1$. Solving (3.2) yields the steady-state average number of base customers

$$x_1 := \frac{\lambda_0 q_0 \theta_0}{\gamma_1 + r_1 (1 - q_1) (1 - \theta_1)}.$$  

(3.3)

The numerator in (3.3) is the inflow rate of new customers, the denominator the departure rate from the customer base, and its inverse is the mean sojourn time in the customer base. Note that, for the system to be profitable, the service probability of new customers must be positive, i.e., $q_0 > 0$.

From (3.1) and (3.3) the profit-maximization problem simplifies to the following non-linear program:

$$\max_{q \geq 0, N \geq 0, \lambda_0 \geq 0} \Pi = \lambda_0 (p_0 q_0 - c_0 (1 - q_0)) + x_1 (R + r_1 [p_1 q_1 - c_1 (1 - q_1)]) - CN - S \lambda \mu (3.4)$$

s.t.

$$x_1 = \frac{\lambda_0 q_0 \theta_0}{\gamma_1 + r_1 (1 - q_1) (1 - \theta_1)},$$

(3.5)

$$q_i \leq 1, \quad i = 0, 1,$$

(3.6)

$$\frac{\lambda_0 q_0}{\mu_0} + \frac{x_1 r_1 q_1}{\mu_1} \leq N,$$

(3.7)

where $q = (q_0, q_1)$, (3.6) captures the service probability constraints and (3.7) the capacity constraint. The left-hand side of (3.7) expresses the sum of processing time allocated (and consumed) by new and base customer calls.

One obvious caveat of the deterministic fluid model is that it does not account for queueing effects and customer impatience in evaluating the steady-state service probabilities: all customers are served if there is enough capacity. In contrast, in the stochastic system, customers may abandon even if there is enough capacity to serve them eventually. However, the great advantage of the fluid model is its analytical tractability. It yields clear results and insights on the optimal decisions, as shown in §3.4 and §3.5. More-
over, our simulation results in §3.6 show that the optimal decisions that the fluid model prescribes yield near-optimal performance for the stochastic system it approximates.

### 3.3.3 Customer Value Metrics and Service Quality

Before proceeding with the analysis, in this section we derive three customer value metrics that drive the optimal decisions. These metrics are novel in that they depend on the call center service quality. In §3.3.4 we use these metrics to reformulate the profit-maximization problem (3.4)-(3.7).

Let \( L(q_1) \) denote the mean base customer lifetime value (CLV), i.e., the total profit that she generates during her sojourn time in the customer base, as a function of the service probability \( q_1 \):

\[
L(q_1) := \frac{R + r_1 (p_1 q_1 - c_1 (1 - q_1))}{\gamma_1 + r_1 (1 - q_1) (1 - \theta_1)}. \tag{3.8}
\]

The CLV of a base customer is the product of her profit rate per unit time, the numerator in (3.8), by her average sojourn time in the customer base, reciprocal of the denominator in (3.8).

Let \( V_0 \) denote the mean one-time service value (OTV) of a new customer, i.e., the value of serving a new customer’s current call, but not any of her future calls:

\[
V_0 := p_0 + c_0 + \theta_{01} L(0). \tag{3.9}
\]

Serving a new customer yields instant profit \( p_0 \), and with probability \( \theta_{01} \) turns that customer into a base customer with lifetime value \( L(0) \). Not serving a new customer results in a penalty \( c_0 \).

Similarly, let \( V_1 \) denote the mean OTV of a base customer, which measures the value of serving her current call but not any of her future calls. The base customer OTV and the CLV satisfy the following intuitive relationship:

\[
V_1 := \frac{L(1) - L(0)}{r_1/\gamma_1}, \tag{3.10}
\]

where \( r_1/\gamma_1 \) is the mean number of calls during a base customer’s lifetime if all her calls
are served. Equivalently, the mean OTV of a base customer satisfies

\[ V_1 = p_1 + c_1 + (1 - \theta_1) L(0). \]  

Serving a base customer’s current call, but not any of her future calls, yields a profit \( p_1 + L(0) \), where \( p_1 \) is the immediate profit, and \( L(0) \) is the CLV given a zero service probability for this base customer. Not serving a base customer’s call yields \( -c_1 + \theta_1 L(0) \), where the first term captures the immediate cost and the second is her CLV given a zero service probability for this base customer. The difference between these two profits yields (3.11).

### 3.3.4 The Profit-Maximization Problem in Terms of Customer Value Metrics

In this section we transform the profit-maximization problem (3.4)-(3.7) by using the customer value metrics (3.8)-(3.10) and by formalizing the scheduling policy in terms of the capacity allocated to new versus base customers, rather than in terms of their service probabilities \( q_0 \) and \( q_1 \).

Let \( N_0 \) and \( N_1 \) denote the capacity allocated to (and consumed by) new and base customers, respectively, and \( \mathbf{N} = (N_0, N_1) \). We have

\[ N_0 : = \frac{\lambda_0 q_0}{\mu_0}, \]

\[ N_1 : = \frac{x_1 r_1 q_1}{\mu_1}. \]

The numerators of (3.12) and (3.13) represent, respectively, the throughput rates of new and base customer calls as functions of their service probabilities, and the denominators are their service rates. It follows from (3.3) and (3.12)-(3.13), after some algebra, that

\[ x_1 = \frac{N_0 \mu_0 \theta \theta_1 + N_1 \mu_1 (1 - \theta_1)}{\gamma_1 + r_1 (1 - \theta_1)}. \]  

Write \( s_0 \) for the mean service time of a new customer call. Let \( s_1 \) denote the total expected time required to serve all base customer calls that may be generated as a result
of serving a new customer.

\[ s_0 := \frac{1}{\mu_0} \quad \text{and} \quad s_1 := \theta_0 r_1 \frac{1}{\gamma_1 \mu_1}, \quad (3.15) \]

where \( \theta_0 \) is the probability that a new customer that is served joins the customer base, \( r_1/\gamma_1 \) is the mean number of calls during a base customer’s lifetime if all her calls are served, and \( 1/\mu_1 \) is the mean service time of a base customer call. In other words, \( s_1 \) is the offered load for base customer calls that a new customer generates. Then \( s_0 + s_1 \) is the total offered load per new customer to the system. Furthermore, \( \lambda_0 s_0 \) and \( \lambda_0 s_1 \) are the system’s offered loads of new and base customer calls, respectively, and \( \lambda_0 (s_0 + s_1) \) is the system’s total offered load. We call a system underloaded if \( \lambda_0 (s_0 + s_1) \leq N \), balanced if \( \lambda_0 (s_0 + s_1) = N \), and overloaded if \( \lambda_0 (s_0 + s_1) > N \).

Let \( \Pi (N, N, \lambda_0) \) denote the profit rate as a function of the capacity allocation vector \( N = (N_0, N_1) \), the total capacity \( N \), and the new customer arrival rate \( \lambda_0 \). From (3.4)-(3.7) we obtain the following equivalent profit-maximization problem for the fluid model, by substituting the customer value metrics \( L(q_1), V_0, \) and \( V_1 \) from (3.8)-(3.10); the capacity allocation vector \( N \) from (3.12)-(3.13); the customer base \( x_1 \) from (3.14); and the mean service times \( s_0 \) and \( s_1 \) from (3.15):

\[
\max_{N \geq 0, N \geq 0, \lambda_0 \geq 0} \quad \Pi (N, N, \lambda_0) = N_0 V_0 \mu_0 + N_1 V_1 \mu_1 - \lambda_0 c_0 - CN - S (\lambda_0) \quad (3.16)
\]

\[ \text{s.t.} \]

\[ N_0 \leq \lambda_0 s_0, \quad (3.17) \]

\[ N_1 \leq N_0 \mu_0 s_1, \quad (3.18) \]

\[ N_0 + N_1 \leq N. \quad (3.19) \]

For fixed new customer arrival rate \( \lambda_0 \) the problem is a linear in \( N \) and \( \lambda_0 \). In the profit rate (3.16), the term \( N_0 V_0 \mu_0 \) captures the value generated by serving new customer calls, where \( N_0 \mu_0 \) is their throughput and \( V_0 \) is the OTV of a new customer; similarly, \( N_1 V_1 \mu_1 \) captures the value generated by serving base customer calls. The capacity allocation constraints (3.17)-(3.18) correspond to the service probability constraints in (3.6). By (3.17) the capacity allocated to new customer calls must not exceed their offered load (mathematically, (3.17) follows from (3.12), (3.15) and \( q_0 \leq 1 \)). By (3.18) the capacity
allocated to base customer calls must not exceed the product of new customer throughput by the offered load for base customer calls that a new customer generates (mathematically, (3.18) follows from (3.3), (3.12)-(3.13), (3.15) and $q_1 \leq 1$). Finally, (3.19) captures the capacity constraint.

### 3.4 Optimal Priority Policy, Staffing, and Promotion Level

In this section we solve the profit-maximization problem (3.16)-(3.19) for the fluid model in three steps. The solution at each step serves as a building block for the next step. In §3.4.1 we consider the case in which the manager only controls the priority policy through the capacity allocation $N$, whereas the call center capacity $N$ and the new customer arrival rate $\lambda_0$ are fixed. In §3.4.2 we characterize the jointly optimal priority and staffing policies, taking the promotion level as fixed. In §3.4.3 we solve the optimization problem over all three controls.

#### 3.4.1 Optimal Priority Policy for Fixed Staffing and Promotion Level

Consider the case where the number of servers $N$ and the new customer arrival rate $\lambda_0$ are fixed. This captures situations where staffing and/or advertising may not be at their optimal levels, e.g., due to hiring lead times, time lags between advertising and demand response, or poor coordination between marketing and operations. The remaining decision is to control the priority policy through the capacity allocation $N$.

**Proposition 3.1.** Fix the new customer arrival rate $\lambda_0$ and the number of servers $N$. If $V_{0\mu_0} > V_{1\mu_1}$ then it is optimal to prioritize new customers, the optimal capacity allocation is

$$
N_0^* = \min (\lambda_0 s_0, N), \quad (3.20)
$$

$$
N_1^* = \min \left(\lambda_0 s_1, (N - \lambda_0 s_0)^+\right), \quad (3.21)
$$
and the optimal profit rate is

\[ \Pi^* (N, \lambda_0) = \min \left( \lambda_0 s_0, N \right) V_0 \mu_0 + \min \left( \lambda_0 s_1, (N - \lambda_0 s_0)^+ \right) V_1 \mu_1 - \lambda_0 c_0 - S(\lambda_0) - CN. \]  

(3.22)

If \( V_0 \mu_0 \leq V_1 \mu_1 \) then it is optimal to prioritize base customers, the optimal capacity allocation is

\[ N_0^* = \min \left( \lambda_0 (s_0 + s_1), N \right) \frac{s_0}{s_0 + s_1}, \]  

(3.23)

\[ N_1^* = \min \left( \lambda_0 (s_0 + s_1), N \right) \frac{s_1}{s_0 + s_1}, \]  

(3.24)

and the optimal profit rate is

\[ \Pi^* (N, \lambda_0) = \min \left( \lambda_0 (s_0 + s_1), N \right) \frac{s_0 V_0 \mu_0 + s_1 V_1 \mu_1}{s_0 + s_1} - \lambda_0 c_0 - S(\lambda_0) - CN. \]  

(3.25)

Remark 3.1. The profit depends on the priority policy only in conditions that result in throughput loss. In the fluid model, the system loses throughput if and only if it is overloaded. As confirmed by (3.22) and (3.25), the profit is independent of the priority policy in an underloaded system. An underloaded stochastic system, however, may experience throughput loss, due to queueing and abandonment. Therefore, prioritizing customers with the higher OTV, in line with Proposition 3.1, improves profits in stochastic systems even if \( \lambda_0 (s_0 + s_1) \leq N \), as our simulation results in §3.6 show.

Figure 3.2 illustrates how the optimal priority policy depends on the system’s offered load, the capacity, and the difference between the weighted OTVs of new and base customers, \( V_0 \mu_0 - V_1 \mu_1 \).

If \( V_0 \mu_0 \geq V_1 \mu_1 \), it is optimal to prioritize new customers, serving no base customers if \( N \leq \lambda_0 s_0 \), and only a fraction of them if \( \lambda_0 s_0 < N < \lambda_0 (s_0 + s_1) \). The condition \( V_0 \mu_0 \geq V_1 \mu_1 \) may hold in the example of a mobile phone service provider. Recall from (3.9) that \( V_0 = p_0 + c_0 + \theta_{01} L(0) \) and from (3.11) that \( V_1 = p_1 + c_1 + (1 - \theta_1) L(0) \). If potential new customers are likely to join the customer base upon being served (so \( \theta_{01} \) is significant), base customers do not easily leave the customer base, because of significant switching costs (so \( \theta_1 \) is high), and their per-call profit \( p_1 \) and abandonment cost \( c_1 \) are small in relation to their call-independent profit rate \( R \), then \( V_0 > V_1 \).
Conversely, in the region “Overloaded, Prioritize base”, the weighted OTV of a base customer exceeds that of a new customer, i.e., $V_1 \mu_1 > V_0 \mu_0$. The system serves all base customers and a fraction of new customers as the customer base must equilibrate at a level that leaves some (though insufficient) residual capacity for new customers to get served and join the customer base. The condition $V_1 \mu_1 > V_0 \mu_0$ may hold in the example of a catalog marketing company, where new and base customers purchase similar products and generate a significant profit per call (so $p_1 \approx p_0 > 0$), and base customers are prone to leave the customer base if not served (so $\theta_1$ is low).

*Effect of Priority Policy on Customer Lifetime Value and Profitability.* The optimal priority policy specified in Proposition 3.1 bears some resemblance to standard priority policies in the literature, such as the $c\mu$ rule. However, a novel feature of our model is that the priority policy considers the effect of the service level on customers’ future calls and their financial impact.

Specifically, if $V_0 \mu_0 > V_1 \mu_1$, so it is optimal to prioritize new over base customers, the value of serving a new customer call is limited to its OTV, $V_0$. In particular, it is independent of the OTV of base customer calls, $V_1$, that this new customer may generate subsequently. As a result, in the optimal profit rate (3.22), each unit of processing time allocated to new customer calls generates $V_0 \mu_0$, and each residual unit of processing time
allocated to base customer calls generates $V_1 \mu_1$.

In contrast, if $V_0 \mu_0 < V_1 \mu_1$, so it is optimal to prioritize base over new customers, the value of serving a new customer call must include not only its OTV, $V_0$, but also the OTVs of subsequent base customer calls. In particular, the expected total value that a new customer generates if the system serves her initial call and all her subsequent calls as a base customer equals

$$V_0 + \theta_{01} \frac{r_1}{\gamma_1} V_1 = s_0 V_0 \mu_0 + s_1 V_1 \mu_1 = p_0 + c_0 + \theta_{01} L(1),$$

(3.26)

where $\theta_{01}$ is the probability that a new customer that is served joins the customer base and $r_1/\gamma_1$ is the mean number of calls during a base customer’s lifetime if all her calls are served. The first equality follows from the definitions of $s_0$ and $s_1$ in (3.15), and the second equality from the definitions of $V_0$ and $V_1$ in (3.9)-(3.10). Serving a new customer and all of her future calls as base customer yields instant profit $p_0$ and saves the abandonment penalty $c_0$, and with probability $\theta_{01}$ turns that customer into a base customer with lifetime value $L(1)$. Therefore, as reflected in the optimal profit rate (3.25), the maximum value that the system generates on average per unit of processing time equals total value of a new customer from (3.26), divided by the total offered load per new customer:

$$\frac{s_0 V_0 \mu_0 + s_1 V_1 \mu_1}{s_0 + s_1}.$$  

(3.27)

As we show in §§3.4.2-3.4.3 for the base model and in §3.5.1 for settings with heterogeneous base customers, the optimal staffing and promotion levels crucially hinge on considering these optimal service-level-dependent customer value metrics. In §3.4.4, we demonstrate the suboptimal performance that may result if one takes customer value as independent of service levels (as is common in the marketing literature) and assumes that all calls must be served.

### 3.4.2 Jointly Optimal Priority and Staffing Policy for a Fixed Promotion Level

Consider situations where the promotion policy, and therefore the new customer arrival rate, is fixed, but the manager controls the priority policy and the call center capacity.
through the staffing budget. This setting is applicable as the promotion policy is often a strategic decision that affects the more tactical capacity planning decisions.

We henceforth refer to the operating profit as the profit rate before promotion costs. By Proposition 3.1, under the optimal priority policy the operating profit is piecewise linear and concave in the capacity $N$. Therefore, the (largest) optimal number of servers $N^*$ is the largest capacity level such that the marginal operating profit is nonnegative. The marginal operating profit of an extra server equals the maximum value of the additional throughput it generates, minus the server cost $C$. By Proposition 3.1, the maximum value of additional throughput depends as follows on the customer value metrics and the system load.

If $V_0\mu_0 \geq V_1\mu_1$, it is optimal to prioritize new customers. By (3.22) the marginal operating profit is\(^1\)

\[
\frac{\partial \Pi^* (N, \lambda_0)}{\partial N} = \begin{cases} 
V_0\mu_0 - C, & N < \lambda_0 s_0, \\
V_1\mu_1 - C, & \lambda_0 s_0 < N < \lambda_0 (s_0 + s_1), \\
-C, & \lambda_0 (s_0 + s_1) < N.
\end{cases}
\]

That is, server capacity is first assigned to new customers, up to the point where it matches their offered load $\lambda_0 s_0$. Additional capacity is then assigned to base customer calls, up to the point where it matches the total offered load $\lambda_0 (s_0 + s_1)$. Beyond this point, additional capacity has no value in the fluid model, so the operating profit decreases at the rate of the capacity cost $C$.

If $V_0\mu_0 \leq V_1\mu_1$, it is optimal to prioritize base customers. Here, additional capacity has the effect of increasing the throughput of new customers in a controlled fashion, ensuring that all base customer calls are served regardless of the load. By (3.25) the marginal operating profit equals

\[
\frac{\partial \Pi^* (N, \lambda_0)}{\partial N} = \begin{cases} 
\frac{s_0 V_0 \mu_0 + s_1 V_1 \mu_1}{s_0 + s_1} - C, & N < \lambda_0 (s_0 + s_1), \\
-C, & \lambda_0 (s_0 + s_1) < N.
\end{cases}
\]

Proposition 3.2 summarizes the jointly optimal staffing and priority policies.

---

\(^1\)For analytical convenience, in the fluid model analysis we ignore integrality constraints and treat the staffing variable $N$ as continuous.
Proposition 3.2. Fix the new customer arrival rate $\lambda_0$.

1. If $V_0\mu_0 > C > V_1\mu_1$ then the system is overloaded, and it is strictly optimal to prioritize and serve only new customers: $N^* = N_0^* = \lambda_0 s_0$. The optimal profit is

$$\Pi^* (\lambda_0) = \lambda_0 (V_0 - c_0 - C s_0) - S (\lambda_0),$$

where $V_0 - c_0 = p_0 + \theta_{01} L(0)$.

2. Otherwise, it is profitable to operate if, and only if, $C < (s_0 V_0 \mu_0 + s_1 V_1 \mu_1) / (s_0 + s_1)$. In this case the system is balanced: $N^* = \lambda_0 (s_0 + s_1)$. The optimal profit is independent of the priority policy:

$$\Pi^* (\lambda_0) = \lambda_0 (s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0 - C (s_0 + s_1)) - S (\lambda_0),$$

where $s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0 = p_0 + \theta_{01} L(1)$.

The jointly optimal staffing and priority policies give rise to one of two operating regimes.

By Part 1 of Proposition 3.2, overloading the system and prioritizing new customers is the unique optimal policy, if new customers have the higher weighted OTV and the server cost is between the weighted OTV of new and base customers: $V_0\mu_0 > C > V_1\mu_1$. Under these conditions, the primary goal of the call center is to increase the number of base customers, not to serve them, resulting in deliberately poor service to base customers. It is optimal to expand capacity so long as the offered load of new customer calls exceeds the capacity. This overloading boosts the new customer throughput and the customer base without raising base customer throughput. At optimum all new customers are served but the system’s total offered load may be well above capacity. As discussed in §3.4.1, the condition $V_0\mu_0 > V_1\mu_1$ may hold in the example of a mobile phone service provider. Indeed, existing mobile phone service customers commonly experience long waiting times when contacting the call center with a service request.

By Part 2 of Proposition 3.2, otherwise it is profitable to operate if, and only if, the total value of all calls generated by a new customer exceeds the capacity cost of matching her offered load: $s_0 V_0 \mu_0 + s_1 V_1 \mu_1 > C (s_0 + s_1)$. In this case it is optimal to increase the
call center capacity to serve all calls, so that the system is balanced, and to prioritize either new or base customers. This regime arises if base customers have the higher weighted OTV (i.e., $V_1\mu_1 \geq V_0\mu_0$), or if new customers have the higher weighted OTV and the server cost is sufficiently small (i.e., $C \leq V_1\mu_1 < V_0\mu_0$). Figure 3.3 illustrates Part 2 of Proposition 3.2 for the latter case $V_0\mu_0 > V_1\mu_1 > C$. It depicts the operating profit as a function of the number of servers $N$, both under the optimal priority policy to new customers, and under the suboptimal priority policy to base customers.

Figure 3.3: Operating profit as a function of the service capacity and priority policy ($V_0\mu_0 > V_1\mu_1 > C$).

### 3.4.3 Jointly Optimal Priority, Staffing, and Promotion Policy

Finally, we characterize the optimal promotion policy, taking into account the jointly optimal staffing and priority policies discussed in §3.4.2. Proposition 3.3 summarizes the solution to the profit-maximization problem (3.16)-(3.19).

**Proposition 3.3.** The jointly optimal promotion, priority and staffing policies depend as follows on the customer OTVs and the capacity cost:

1. If $(V_0 - c_0)\mu_0 > C > V_1\mu_1$ then the system is overloaded, it is strictly optimal to
prioritize and serve only new customers, \( N^* = N_0^* = \lambda_0^* s_0 \), and

\[
\lambda_0^* = \arg \{ \lambda_0 \geq 0 : V_0 - c_0 - C s_0 = S'(\lambda_0) > 0 \}, \tag{3.32}
\]

where \( V_0 - c_0 = p_0 + \theta_01 L(0) \).

2. Otherwise, it is profitable to operate if, and only if, \( C < (s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0) / (s_0 + s_1) \).

In this case the system is balanced, \( N^* = \lambda_0^* (s_0 + s_1) \), profits are independent of the priority policy, and

\[
\lambda_0^* = \arg \{ \lambda_0 \geq 0 : s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0 - C (s_0 + s_1) = S'(\lambda_0) > 0 \}, \tag{3.33}
\]

where \( s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0 = p_0 + \theta_01 L(1) \).

The conditions for the two operating regimes described in Proposition 3.3 parallel those for Proposition 3.2, with one minor adjustment: When the new customer arrival rate is a decision variable, the value of a new customer call must be reduced by the abandonment cost \( c_0 \), because it is suboptimal for the firm to turn away new customers it pays to attract.

In each operating regime of Proposition 3.3 the optimal promotion spending balances the marginal cost of attracting a new customer with the marginal operating profit she generates under the jointly optimal staffing and priority policies that is specified in Proposition 3.2. In Part 1 of Proposition 3.3, it is optimal to only serve new customer calls. In this case, by (3.32) the marginal operating profit of a new customer equals \( p_0 + \theta_01 L(0) - C s_0 \), which is the value minus the server cost when serving only her initial call. In Part 2 of Proposition 3.3, it is optimal to serve all customer calls. In this case, by (3.33) the marginal operating profit of a new customer equals \( p_0 + \theta_01 L(1) - C (s_0 + s_1) \), which is the value minus the server cost when serving her initial and all her future calls.

### 3.4.4 Ignoring the Effect of Service Probability on Customer Lifetime Value

Standard measures of customer lifetime value in the marketing literature ignore the service probability. In this section we demonstrate the suboptimal performance that may
result if decision makers take customer value as independent of service levels (as is common in the marketing literature) and assume that all calls must be served. We focus on the case \((V_0 - c_0) \mu_0 > V_1 \mu_1\), so

\[
\frac{p_0 + \theta_{01} L(1)}{s_0 + s_1} = \frac{s_0 (V_0 - c_0) \mu_0 + s_1 V_1 \mu_1}{s_0 + s_1} < \frac{(V_0 - c_0) \mu_0}{s_0}.
\] (3.34)

That is, the lifetime value of a new customer is lower if all her calls as base customer are served. By Proposition 3.3 it is optimal to only serve new customers if \((V_0 - c_0) \mu_0 > C > V_1 \mu_1\) and to serve all customers if \(C \leq V_1 \mu_1\). We contrast this optimal policy with two alternatives.

In the marketing-driven policy the marketing department controls both promotions and staffing, but assumes that all calls must be served in making these decisions. Let \(\lambda_0^M\) and \(N_0^M\) denote, respectively, the optimal new customer arrival rate and staffing level under this policy. The new customer arrival rate \(\lambda_0^M\) is chosen to balance the marginal cost of attracting a new customer with the marginal operating profit she generates when serving all her calls:

\[
\lambda_0^M = \arg \{ \lambda_0 \geq 0 : p_0 + \theta_{01} L(1) - C (s_0 + s_1) = S' (\lambda_0) \} > 0 \text{ if } C < \frac{p_0 + \theta_{01} L(1)}{s_0 + s_1},
\]

and \(\lambda_0^M = 0\) otherwise. The staffing level satisfies \(N_0^M = \lambda_0^M (s_0 + s_1)\), because all calls are served.

In the uncoordinated policy the marketing department optimizes the promotion level, assuming that all calls will be served, whereas the operations department optimizes the priority policy and staffing level (in line with Proposition 3.2), given the new customer arrival rate set by marketing. Let \(\lambda_0^U\) and \(N_0^U\) denote, respectively, the optimal new customer arrival rate and staffing level under this policy. Then \(\lambda_0^U = \lambda_0^M\), because the new customer arrival rate in both policies is determined by same decision rule. Because \(V_0 \mu_0 > V_1 \mu_1\), by Proposition 3.2, the optimal staffing level corresponding to \(\lambda_0^U\) satisfies \(N_0^U = \lambda_0^U s_0\) if \(V_0 \mu_0 > C > V_1 \mu_1\), and \(N_0^U = \lambda_0^U (s_0 + s_1)\) if \(V_1 \mu_1 \geq C\).

Figure 3.4 illustrates how the optimal staffing level \(N^*\) and the new customer arrival rate \(\lambda_0^*\) compare to their counterparts under the marketing-driven and coordinated policies, depending on the server cost \(C\). As shown in the right panel, both the marketing-driven and the uncoordinated policies adversely affect the promotion level.
for $C > V_1\mu_1$, and may even cause the system to shut down (for $C > (s_0(V_0 - c_0)\mu_0 + s_1V_1\mu_1)/(s_0 + s_1)$). This is the direct consequence of imposing a suboptimal service level: As reflected in (3.34), compared to serving only new customer calls, serving all base customer calls reduces the marginal operating profit per new customer. Considering the diminishing return of promotion spending, the firm chooses a lower promotion level as a result. As shown in the left panel of Figure 3.4, the uncoordinated policy also yields a lower than optimal staffing level for $C > V_1\mu_1$. However, in this cost range, the marketing-driven policy may yield lower or higher than optimal staffing, which is due to two countervailing effects: The new customer arrival rate is lower than optimal (i.e., $\lambda_0^M < \lambda_0^*$) but the marketing-driven policy serves all calls, rather than only those of new customers (as is optimal), i.e., $N_0^M = \lambda_0^M(s_0 + s_1)$ whereas $N_0^* = \lambda_0^*s_0$.

The profit loss as a result of these suboptimal decisions can be very considerable. Numerical examples suggest that the profit loss relative to the optimal profit can easily amount to tens of percentage points. This discussion underscores the importance of accounting for the optimal service level in evaluating the customer lifetime value, particularly when this measure serves as the basis of substantial resource allocation decisions on promotion and staffing levels.

Figure 3.4: Ignoring the effect of service probability on customer lifetime value: $(V_0 - c_0)\mu_0 > V_1\mu_1$. (The rate $\lambda_0^*$ satisfies (3.32) and (3.33) for $C = V_1\mu_1$. The rate $\lambda_0^*$ satisfies (3.33) for $C = 0$.)
3.5 Extensions

In this section we extend the model studied so far in two directions. In §3.5.1 we consider heterogeneous base customers. In §3.5.2 we reexamine our results in the presence of service-quality-driven negative word-of-mouth effects on new customer acquisition.

3.5.1 Multiple Types of Heterogeneous Base Customers

In the basic model studied so far all base customers are inherently homogeneous. In this section, we generalize the model by considering multiple types of heterogeneous base customers that differ in their main characteristics. Let the firm have \( m \) types of base customers. We index all variables and quantities pertaining to base customers by subscript \( i \in \{1, 2, ..., m\} \). As before, \( S(\lambda_0) \) denotes the advertising spending rate that generates the new customer arrival rate \( \lambda_0 \). A new customer who receives service joins type \( i \) base customers with probability \( \theta_0i \). We assume that base customers do not switch among types. The call arrival rate per base customer of type \( i \) is \( r_i \), and if she abandons her call, she remains in the customer base with probability \( \theta_i \). The service rate of a type-\( i \) base customer is \( \mu_i \), and her call-independent attrition rate is \( \gamma_i \). Let \( q_i \) denote the steady-state service probability for calls of type \( i \) base customers, \( q := (q_0, q_1, ..., q_m) \), and \( x_i \) denote the mean number of type-\( i \) base customers in steady state. Following (3.3) we have

\[
x_i = \frac{\lambda_0 q_0 \theta_{0i}}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)}.
\]

(3.35)

The base customer types may also differ in their financial parameters, that is, their abandonment penalty costs \( c_i \), and their call-dependent and call-independent profits, \( p_i \) and \( R_i \), respectively.
The profit-maximization problem (3.4)-(3.7) generalizes as follows for the multi-type model:

$$\max_{q \geq 0, n \geq 0, \lambda \geq 0} \Pi = \lambda_0 (p_0 q_0 - c_0 (1 - q_0)) + \sum_{i=1}^{m} x_i (R_i + r_i [p_i q_i - c_i (1 - q_i)]) - CN - S (\lambda)$$

subject to

$$x_i = \frac{\lambda_0 q_0 \theta_i}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)}, \quad i = 1, 2, ..., m,$$

$$q_i \leq 1, \quad i = 0, 1, 2, ..., m,$$

$$\lambda_0 q_0 + \sum_{i=1}^{m} \frac{x_i r_i q_i}{\mu_i} \leq N.$$

We transform (3.36)-(3.39) by generalizing the development of §§3.3.3-3.3.4 in the obvious way. Let

$$L_i(q_i) := \frac{R_i + r_i (p_i q_i - c_i (1 - q_i))}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)} \gamma_i + r_i (1 - q_i) (1 - \theta_i)$$
denote the mean CLV of a type-\(i\) base customer,

$$V_0 := p_0 + c_0 + \sum_{i=1}^{m} \theta_0 L_i(0)$$
denote the mean OTV of a new customer, and

$$V_i := \frac{L_i(1) - L_i(0)}{r_i \gamma_i}$$
denote the mean OTV of a type-\(i\) base customer. Without loss of generality we assume that \(V_i \mu_i \geq V_{i+1} \mu_{i+1}\) for \(i = 1, 2, ..., m - 1\). For simplicity we further assume that \(V_i \mu_i > V_{i+1} \mu_{i+1}\) for \(i = 1, 2, ..., m - 1\). (Settings where \(V_i \mu_i = V_{i+1} \mu_{i+1}\) for some pair(s) of base customer types add somewhat cumbersome detail to the analysis without generating additional insights.)

Denote the capacity allocated to (and consumed by) type-\(i\) base customers by

$$N_i := \frac{x_i r_i q_i}{\mu_i},$$

and let \(N := (N_0, N_1, N_2, ..., N_m)\). It follows from (3.12), (3.35) and (3.42) that the mean
size of the type-\(i\) customer base satisfies
\[
x_i = \frac{N_0 \mu_0 \theta_0 + N_i \mu_i (1 - \theta_i)}{\gamma_i + r_i (1 - \theta_i)}.
\]

Finally, let \(s_i\) denote the total expected time required to serve all base customer calls that may be generated as a result of serving a new customer:
\[
s_i := \theta_0 \frac{r_i}{\gamma_i \mu_i}.
\] (3.43)

Then we obtain from (3.36)-(3.39) the following equivalent profit-maximization problem:
\[
\max_{N \geq 0, N_i \geq 0, \lambda_0 \geq 0} \Pi (N, N_i, \lambda_0) = \sum_{i=0}^{m} N_i \mu_i - \lambda_0 c_0 - CN - S (\lambda_0)
\] (3.44)
\[
\text{s.t.}
\]
\[
N_0 \leq \lambda_0 s_0,
\] (3.45)
\[
N_i \leq N_0 \theta_0 s_i, \quad i = 1, 2, \ldots, m,
\] (3.46)
\[
\sum_{i=0}^{m} N_i \leq N.
\] (3.47)

We solve the problem (3.44)-(3.47) following the same three steps as in §3.4.

### 3.5.1.1 Optimal Priority Policy for Fixed Staffing and Promotion Level

Denote by \(\bar{s}_i\) the offered load of a new customer due to her initial call and all subsequent calls placed as a base customer of type \(j \in \{1, 2, \ldots, i\}\). Then
\[
\bar{s}_i := \sum_{j=0}^{i} s_j, \quad i = 0, 1, 2, \ldots, m.
\] (3.48)

Let \(\bar{V}_i\) denote the expected total value that a new customer generates if the system serves her initial call and all her subsequent calls as a base customer of type \(j \in \{1, 2, \ldots, i\}\), divided by the corresponding offered load \(\bar{s}_i\):
\[
\bar{V}_i = \frac{\sum_{j=0}^{i} s_j \mu_j j}{\bar{s}_i}, \quad i = 0, 1, 2, \ldots, m.
\] (3.49)
Note that $\overline{V}_0 = V_0 \mu_0$. The parameters $\overline{V}_i$ are measures of service-level-dependent customer lifetime value, as introduced in Section 3.4.1. In particular, it follows from (3.40), (3.41), (3.43), and (3.49) that

$$\overline{V}_i = \frac{p_0 + c_0 + \sum_{j=0}^{i} \theta_{0j} L_j(1) + \sum_{j=i+1}^{m} \theta_{0j} L_j(0)}{\bar{s}_i}, \quad i = 0, 1, 2, \ldots m. \quad (3.50)$$

The numerator of $\overline{V}_i$ captures the total value of a new customer as a function of the service levels for the base customer types: The calls of type $j \in \{1, 2, \ldots, i\}$ are served, so their CLV is evaluated for $q_j = 1$. However the calls of type $j > i$ are not served, so their CLV is evaluated for $q_j = 0$. The denominator of $\overline{V}_i$ captures the total capacity consumed by a new customer under this service policy. Lemma 3.1 summarizes key properties of $\overline{V}_i$.

**Lemma 3.1.** Define

$$k = \begin{cases} 0, & \text{if } \overline{V}_0 > \overline{V}_1, \\ \max \{1 \leq i \leq m : \overline{V}_{i-1} \leq \overline{V}_i\}, & \text{if } \overline{V}_0 \leq \overline{V}_1. \end{cases} \quad (3.51)$$

Then $\overline{V}_k$ is the maximum value of a new customer per unit of processing time:

$$\overline{V}_k = \max_{i \in \{0, 1, \ldots, m\}} \overline{V}_i. \quad (3.52)$$

Moreover

$$\overline{V}_0 < \ldots < \overline{V}_{k-1} \leq \overline{V}_k > \overline{V}_{k+1} > \ldots > \overline{V}_m, \quad (3.53)$$

$$\overline{V}_k \mu_k \geq \overline{V}_k > \overline{V}_{k+1} \mu_{k+1}, \quad \text{if } k < m. \quad (3.54)$$

Proposition 3.4 generalizes Proposition 3.1.

**Proposition 3.4.** Fix the new customer arrival rate $\lambda_0$ and the number of servers $N$. It is optimal to prioritize base customers of type $i \in \{1, \ldots, k\}$ over new customers, new customers over base customers of type $i > k$, and base customers in decreasing order of their $V_i \mu_i$-index.
The optimal capacity allocation satisfies

\[ N_0^* = \min \left( \lambda_0, \frac{N}{\overline{s}_k} \right) s_0, \quad (3.55) \]

\[ N_i^* = \min \left( \lambda_0, \frac{N}{\overline{s}_k} \right) s_i, \quad i \in \{1, \ldots, k\}, \quad (3.56) \]

\[ N_i^* = \min \left( \lambda_0 s_i, (N - \lambda_0 \overline{s}_{i-1})^+ \right), \quad i > k, \quad (3.57) \]

and the optimal profit rate is

\[ \Pi^* (N, \lambda_0) = \min (\lambda_0 \overline{s}_k, N) \overline{V}_k + \sum_{i=k+1}^{m} \min (\lambda_0 s_i, (N - \lambda_0 \overline{s}_{i-1})^+) \overline{V}_i \mu_i - \lambda_0 c_0 - S (\lambda_0) - CN. \quad (3.58) \]

The optimal priority policy and the corresponding capacity allocation are quite intuitive: The capacity should be allocated in a manner that maximizes the customer value generated per unit of processing time. By Lemma 3.1, this is achieved if all base customer calls of type \( i \in \{1, \ldots, k\} \) are served. Therefore, if \( N < \lambda_0 \overline{s}_k \), it is optimal to turn away enough new customers to ensure that all base customer calls of type \( i \in \{1, \ldots, k\} \) can be served. No base customers of type \( i > k \) are served as a result. However, if \( N \geq \lambda_0 \overline{s}_k \), then it is optimal to serve all new customers and base customer calls of type \( i \in \{1, \ldots, k\} \), and base customers of type \( i > k \) in decreasing order of their type.

### 3.5.1.2 Jointly Optimal Priority and Staffing Policy for a Fixed Promotion Level

Proposition 3.5 generalizes Proposition 3.2. The optimal capacity hinges on the structure of the optimal profit rate in (3.58), which is piecewise linear and concave in the staffing level \( N \).

**Proposition 3.5.** Fix the new customer arrival rate \( \lambda_0 \). Under the optimal priority and staffing policies it is profitable to operate if, and only if, \( C < \overline{V}_k \). In this case:

1. It is optimal to serve new customers, base customers of type \( i \in \{1, \ldots, k\} \), and base customers of type \( j > k \) if and only if \( V_j \mu_j \geq C \).
2. The optimal staffing is

\[ N^* = \lambda_0 \left( \bar{s}_k + \sum_{i=k+1}^{m} s_i 1_{\{V_i \mu_i \geq C\}} \right), \]  

and the system is overloaded if and only if \( C > V_m \mu_m \).

3. The optimal profit rate is

\[ \Pi^* (\lambda_0) = \lambda_0 \left( \bar{s}_k (V_k - C) - c_0 + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C) 1_{\{V_i \mu_i \geq C\}} \right) - S(\lambda_0). \] 

An important result of Proposition 3.5 is that the optimality conditions for operating an overloaded system are considerably more general and more plausible in the presence of heterogeneous base customer types, compared to the basic model with homogeneous base customers. For one, it is only necessary that the weighted OTV of a new customer call exceed that of some (not all) base customer calls. Second, even when the optimal system is overloaded, it is may well be optimal to serve some base customers, provided their OTV is sufficiently high. More generally, Proposition 3.5 underscores the importance of using the priority policy, not only for differentiating the service level across base customers depending on their value, but also to influence the composition of the customer base. Specifically, it follows from (3.35) that the optimal size of the customer bases of two types \( l < j \) that have the same calling behavior (\( \theta_{0l} = \theta_{0j}, \theta_l = \theta_j, r_l = r_j, \) and \( \gamma_l = \gamma_j \)) but receive different service levels, \( q_l = 1 \) and \( q_j = 0 \), satisfy

\[ \frac{x_l}{x_j} = 1 + \frac{r_j}{\gamma_l} (1 - \theta_j). \]

3.5.1.3 Jointly Optimal Priority, Staffing, and Promotion Policy

As noted in §3.4.3, when the new customer arrival rate is a decision variable, the value of a new customer call must be reduced by the abandonment cost \( c_0 \), because it is suboptimal for the firm to turn away new customers it pays to attract. To make this adjustment, define

\[ \tilde{V}_i = V_i - \frac{c_0}{s_i}, \quad i = 0, 1, 2, ... m, \]
and note that $\tilde{V}_0 = (V_0 - c_0) \mu_0$. Lemma 3.2 parallels Lemma 3.1.

**Lemma 3.2.** Define

$$k^* = \begin{cases} 0, & \text{if } \tilde{V}_0 > \tilde{V}_1, \\ \max \{1 \leq i \leq m : \tilde{V}_{i-1} \leq \tilde{V}_i\}, & \text{if } \tilde{V}_0 \leq \tilde{V}_1. \end{cases} \quad (3.62)$$

Then $\tilde{V}_k$ is the maximum value of a new customer per unit time, net of abandonment cost:

$$\tilde{V}_{k^*} = \max_{i \in \{0, 1, \ldots, m\}} \tilde{V}_i. \quad (3.63)$$

Moreover

$$\tilde{V}_0 < \ldots < \tilde{V}_{k^*-1} \leq \tilde{V}_{k^*} > \tilde{V}_{k^*+1} > \ldots > \tilde{V}_m, \quad (3.64)$$

$$V_{k^*} \mu_{k^*} \geq \tilde{V}_{k^*} > V_{k^*+1} \mu_{k^*+1}, \quad \text{if } k^* < m, \quad (3.65)$$

$$k \leq k^*. \quad (3.66)$$

Proposition 3.6 generalizes Proposition 3.3.

**Proposition 3.6.** Under the jointly optimal promotion, priority and staffing policies it is profitable to operate if, and only if, $C < \tilde{V}_{k^*}$. In this case:

1. It is optimal to serve new customers, base customers of type $i \in \{1, \ldots, k^*\}$, and base customers of type $j > k^*$ if and only if $V_j \mu_j \geq C$.

2. The optimal staffing is

$$N^* = \lambda_0^* \left( \bar{s}_{k^*} + \sum_{i=k^*+1}^{m} s_i 1_{\{V_i \mu_i \geq C\}} \right), \quad (3.67)$$

and the system is overloaded if and only if $C > V_m \mu_m$.

3. The optimal new customer arrival rate satisfies

$$\lambda_0^* = \arg \left\{ \lambda_0 \geq 0 : \bar{s}_{k^*} \left( \tilde{V}_{k^*} - C \right) + \sum_{i=k^*+1}^{m} s_i (V_i \mu_i - C) 1_{\{V_i \mu_i \geq C\}} = S' (\lambda_0) \right\} > 0. \quad (3.68)$$
Like the optimal staffing level, the optimal promotion level is highly dependent on the optimal service policy and the resulting CLV.

### 3.5.2 Quality-Driven Word-of-Mouth Effects on New Customer Acquisition

In this section, we generalize the basic model by considering quality-driven word-of-mouth effects on new customer acquisition. For simplicity, we assume that base customers are homogeneous, as in §3.4. However, the results of this section can also be extended to heterogeneous base customers.

We assume that for a fixed advertisement spending, the new customer arrival rate decreases proportionally to the abandonment rate of base customers. That is, spending $S(\lambda_0)$ in advertisement yields the *effective* new customer arrival rate

$$\lambda_0 - \delta x_1 r_1 (1 - q_1),$$

where $\delta \geq 0$ is a scaling factor. Note that this model only considers negative word-of-mouth. In the absence of base customer abandonment, the new customer arrival rate is $\lambda_0$. Substituting the new customer arrival rate with the effective new customer arrival rate and following (3.3) we have

$$x_1 = \frac{\lambda_0 q_0 \theta_{01}}{\gamma_1 + r_1 (1 - q_1) (1 - \theta_1 + \delta q_0 \theta_{01})}.$$  

(3.70)

The profit-maximization problem (3.4)-(3.7) generalizes as follows in the presence of the negative word-of-mouth:

$$\max_{q \geq 0, N \geq 0, \lambda_0 \geq 0} \Pi = (\lambda_0 - \delta x_1 r_1 (1 - q_1)) (p_0 q_0 - c_0 (1 - q_0))$$

$$+ x_1 (R_1 + r_1 (p_1 q_1 - c_1 (1 - q_1))) - CN - S(\lambda_0)$$  

(3.71)

subject to

$$x_1 = \frac{\lambda_0 q_0 \theta_{01}}{\gamma_1 + r_1 (1 - q_1) (1 - \theta_1 + \delta q_0 \theta_{01})},$$  

(3.72)

$$q_i \leq 1, \quad i = 0, 1,$$

(3.73)

$$\frac{(\lambda_0 - \delta x_1 r_1 (1 - q_1)) q_0}{\mu_0} + \frac{x_1 r_1 q_1}{\mu_1} \leq N.$$  

(3.74)
The expected sojourn time of a base customer whose calls are not served is \( \frac{1}{\gamma_1 + r_1 (1 - \theta_1)} \).

During this time she calls and abandons with rate \( r_1 \). Considering the effect of these abandonments on the new customer arrival rate, she contributes to reducing the new customer arrival rate by \( \delta r_1 / (\gamma_1 + r_1 (1 - \theta_1)) \). Depending on the congestion of the system, this reduction reduces the profit (if all new customer calls are served) or decreases the abandonment cost of new customers (otherwise).

Let \( V_0^w \) denote the mean OTV of a new customer in a call center where some new customers abandon their call, in the presence of word-of-mouth:

\[
V_0^w := p_0 + c_0 + \theta_{01} L(0) + \theta_{01} c_0 \frac{\delta r_1}{\gamma_1 + r_1 (1 - \theta_1)}. \tag{3.75}
\]

Serving a new customer yields instant profit \( p_0 \), and with probability \( \theta_{01} \) turns that customer into a base customer with lifetime value \( L(0) \) and a contribution of \( c_0 \delta r_1 / (\gamma_1 + r_1 (1 - \theta_1)) \) to decreasing the new customer abandonment cost. Not serving a new customer results in a penalty \( c_0 \). The difference between the profits of serving and not serving a new customer’s call yields (3.75).

Similarly, let \( V_1^w \) denote the mean OTV of a base customer in a call center where some new customers abandon their call, in the presence of word-of-mouth:

\[
V_1^w = p_1 + L(0) - c_0 \frac{\delta \gamma_1}{\gamma_1 + r_1 (1 - \theta_1)}. \tag{3.76}
\]

Serving a base customer’s current call, but not any of her future calls, yields a profit \( p_1 + L(0) + c_0 \delta r_1 / (\gamma_1 + r_1 (1 - \theta_1)) \), where \( p_1 \) is the immediate profit, \( L(0) \) is her lifetime value, and the third term is her contribution to decreasing the new customer abandonment cost. Not serving a base customer’s call yields \(-c_1 + \delta c_0 + \theta_1 (L(0) + c_0 \delta r_1 / (\gamma_1 + r_1 (1 - \theta_1))) \), where the first term captures the immediate cost of not serving her, the second term captures the immediate contribution in decreasing the new customer abandonment cost, and the third term is her expected CLV plus her contribution to decreasing the new customer abandonment cost, if she remains in the system after abandoning her current call. The difference between the profits of serving and not serving a base customer’s call yields (3.76).

Note that in the absence of new customer abandonment cost (i.e., if \( c_0 = 0 \), which is common in practice), the OTVs of new and base customers become independent of
word-of-mouth, and we have $V_1^w = V_1, V_0^w = V_0$.

The capacity allocated to (and consumed by) new customers should consider the effective new customer arrival rate. Thus, we redefine $N_0$ as

$$N_0 := \frac{(\lambda_0 - \delta x_1 r_1 (1 - q_1)) q_0}{\mu_0}. \tag{3.77}$$

It follows from (3.70), (3.77) and (3.13) that the mean size of the customer base satisfies

$$x_1 = \frac{N_0 \mu_0 \theta_{01} + N_1 \mu_1 (1 - \theta_1)}{\gamma_1 + r_1 (1 - \theta_1)}. \tag{3.78}$$

Then we obtain from (3.71)-(3.74) the following equivalent profit-maximization problem:

$$\max_{N \geq 0, N_0 \geq 0, \lambda_0 \geq 0} \Pi (N, N, \lambda_0) = N_0 V_0^w \mu_0 + N_1 V_1^w \mu_1 - \lambda_0 c_0 - C N - S(\lambda_0) \tag{3.79}$$

s.t.

\begin{align*}
N_0 &\leq \frac{\lambda_0 s_0}{\gamma_1 + r_1 (1 - \theta_1)} + N_1 \mu_1 s_0 \frac{\delta \gamma_1}{\gamma_1 + r_1 (1 - \theta_1 + \delta \theta_{01})} \tag{3.80} \\
N_1 &\leq N_0 \mu_0 s_1, \tag{3.81} \\
N_0 + N_1 &\leq N. \tag{3.82}
\end{align*}

We can solve the problem (3.79)-(3.82) following the same three steps as in §3.4. However, the structure of the optimal decisions can be obtained by comparing this problem with its counterpart (3.16)-(3.19) for the basic model without word-of-mouth effects.

Comparing the objective functions (3.16) and (3.79), $V_0$ and $V_1$ are substituted by their corresponding values in the presence of word-of-mouth, i.e., $V_0^w$ and $V_1^w$, respectively. The feasible regions defined by (3.17)-(3.19) and (3.80)-(3.82) are depicted in Figure 3.5 in the left and right panels, respectively. Note that both regions are based on arbitrary values for $N$ and $\lambda_0$ and would be different for other values of $N$ and $\lambda_0$, e.g., in the basic model increasing $N$ moves (3.19) to the right until it becomes redundant.

To compare the optimal policies of these two models, first consider the problem of finding the optimal priority policy for fixed staffing and promotion level. In the basic model the optimal capacity allocation $N^*$ (and the resulting priority policy) can readily be determined by considering $V_0 \mu_0, V_1 \mu_1$, and the feasible region (3.17)-(3.19). Specifically, if
Figure 3.5: Comparison of feasible capacity allocation regions: basic model and model with word-of-mouth effects on new customer acquisition.

$V_0 \mu_0 > V_1 \mu_1$, then if $N$ is large enough to make (3.19) redundant, $N^*$ is at the intersection of (3.17) and (3.18). Otherwise, $N^*$ is at the intersection of (3.17) and (3.19) or at the intersection of (3.19) with the $N_0$-axis, whichever is a feasible point. Similarly, if $V_0 \mu_0 < V_1 \mu_1$, then $N^*$ is at the intersection of (3.18) and (3.19) if this point is feasible, or at the intersection of (3.17) and (3.18) if (3.19) is redundant. The same analysis applies in the presence of word-of-mouth effects to $V_0^w \mu_0$, $V_1^w \mu_1$ and the feasible region (3.80)-(3.82).

Next, consider the problem of determining the jointly optimal priority and staffing policy for a fixed promotion level. In the basic model, if it is profitable to operate, based on the capacity cost, then the number of servers $N$ should be increased (i.e., (3.19) should be moved to the right) until (3.19) reaches either the intersection of (3.17) with the $N_0$-axis (where only new customers are served), or the intersection of (3.17) and (3.18) (where all new and base customers are served). Similarly, when we consider the word-of-
mouth effect, based on the capacity cost, the number of servers \( N \) should be increased (i.e., (3.82) should be moved to the right) until (3.82) reaches either the intersection of (3.80) with the \( N_0 \)-axis or the intersection of (3.80) and (3.81). Therefore, similar to the basic model, based on the capacity cost the firm either serves only new customers or all new and base customers.

Note that if \( V_0 \mu_0 > V_1 \mu_1 \) and \( V_0^w \mu_0 > V_1^w \mu_1 \), in both models at small capacity cost, the firm serves all new and base customers, and at higher capacity cost, it only serves new customers. However, in the presence of word-of-mouth, the firm switches to not serving base customers at a higher capacity cost since by not serving a base customer, it also partially loses the profit of serving new customers (see the left panel in Figure 3.6 for illustration).

![Figure 3.6](image_url)

**Figure 3.6:** Considering word-of-mouth effects on new customer acquisition: \((V_0 - c_0) \mu_0 > V_1 \mu_1 \) and \((V_0^w - c_0) \mu_0 > V_1^w \mu_1 \). (The rate \( \lambda_0^* \) satisfies (3.32) and (3.33) for \( C = V_1 \mu_1 \). The rate \( \lambda_0^* \) satisfies (3.33) for \( C = 0 \).)

Finally, consider the problem of jointly optimizing the priority, staffing, and promotion policy. From the above discussion, we know that for any \( \lambda_0 \) in the basic model, \( N^* \) is either at the intersection of (3.17) with the \( N_0 \)-axis or at the intersection of (3.17) and (3.18). Increasing \( \lambda_0 \) moves (3.17) (and thus \( N^* \)) to the right. Therefore, the objective function (3.16) increases with rate \((V_0 - c_0 - Cs_0) \) in the former and \((s_0 V_0 \mu_0 + s_1 V_1 \mu_1 - c_0 - C (s_0 + s_1)) \) in the latter case (refer to Proposition 3.3 for further discussion on
optimizing $\lambda_0$). By similar discussion, in the presence of word-of-mouth, $\lambda^*_0$ depends on how moving $N^*$ affects the profit function (3.79). In particular, if for any $\lambda_0$, the optimal policy is to serve all new and base customers, then the optimal new customer arrival rate in the presence of the word-of-mouth would be the same as the one in the basic model. Otherwise, $N^*$ is at the intersection of (3.80) and the $N_0$-axis, and hence, moves to the right by $\frac{\gamma_1 + r_1 (1 - \theta_1)}{\gamma_1 + r_1 (1 - \theta_1 + 4 \theta_0)}$ units for each unit increase in $\lambda_0$. Therefore, in the presence of word-of-mouth, the marginal profit increases in $\lambda_0$ more slowly than the basic model. Therefore, if a firm chooses not to serve base customers, it should also decrease its advertisement spending and its optimal new customer arrival rate would be lower as a result (see the right panel of Figure 3.6 for illustration).

3.6 Fluid Model Validation: Simulation Results

In this section, we evaluate the fluid model results developed in §3.4 by comparing their performance with simulation results for the stochastic system described in §3.3.1. In §§3.6.1-3.6.2 we show how the fluid model error in approximating the optimal decisions, and the resulting profit loss, depend on the server cost $C$. We consider representative parameter settings for two cases: In §3.6.1 the OTV of new customers exceeds that of base customers, and vice versa in §3.6.2. In §3.6.3 we evaluate the robustness of these results by reporting the profit loss for a range of additional parameter settings.

We assume a 24x7 operation. One time unit equals one day. For each simulation we set the initial size of the customer base to the equilibrium value suggested by the fluid model, run the simulation for 1,100,000 new customer arrivals, and discard results from the first 100,000 arrivals in computing performance measures. (Starting from an empty system, a much longer warm-up period is required to reach steady state). Throughout, we assume exponentially distributed service times.

To represent the advertising cost function, we use the commonly assumed power model

$$S(\lambda_0) = \alpha \lambda_0^\beta, \quad \alpha > 0, \beta > 1,$$

where $\alpha$ is a scale factor and $\beta$ is the inverse of the customers’ response elasticity to the advertising level (cf. Hanssens et al. 2001). We let $\alpha = 0.5$, $\beta = 1.5$. 

53
3.6.1 Representative Case 1: Higher OTV for New Customers

\( (V_0 - c_0 > V_1) \)

In this section, we report the results for a representative case in which the OTV of new customers exceeds that of base customers. Table 3.2 summarizes the parameter values for this case, which may be representative of a call center of a business with significant call-independent revenue, such as a mobile phone service provider.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service rate of new and base customers (per day)</td>
<td>( \mu_0, \mu_1 )</td>
</tr>
<tr>
<td>Call abandonment rate of new and base customers (per day)</td>
<td>( \tau_0, \tau_1 )</td>
</tr>
<tr>
<td>Call arrival rate per base customer (per day)</td>
<td>( r_1 )</td>
</tr>
<tr>
<td>P(new customer joins customer base after service)</td>
<td>( \theta_{01} )</td>
</tr>
<tr>
<td>P(base customer remains in customer base after abandoning)</td>
<td>( \theta_1 )</td>
</tr>
<tr>
<td>Attrition rate per base customer (per day)</td>
<td>( \gamma_1 )</td>
</tr>
<tr>
<td>Profit rate per base customer (per day)</td>
<td>( R )</td>
</tr>
<tr>
<td>Profit per served call</td>
<td>( p_{01} )</td>
</tr>
<tr>
<td>Cost per abandoned call</td>
<td>( c_0, c_1 )</td>
</tr>
<tr>
<td>Advertising cost function parameters (power model)</td>
<td>( \alpha, \beta )</td>
</tr>
</tbody>
</table>

Table 3.2: Parameter values for representative Case 1: higher OTV for new customers \((V_n - c_n > V_b)\).

From (3.8) the CLV ranges from \( L(0) = 331.67 \) to \( L(1) = 450.00 \) depending on the base customer service probability. From (3.9) and (3.10) the OTVs of new and base customers are \( V_0 = 109.75 \) and \( V_1 = 23.67 \), respectively.

We evaluate the fluid model performance through two comparisons. First, we compare the fluid model prescriptions (optimal priority policy, number of servers, and new customer arrival rate) with those from simulation-based optimization results for the stochastic system. Second, we compare the simulation-based optimal profit with the simulated system profit under the fluid model prescriptions. We perform these comparisons for a range of capacity costs, varying the server cost \( C \) from 0 to 7200, in increments of 100. (For \( C > 7200 \) the fluid model prescribes a system with fewer than 25 servers. The fluid model is not accurate for such small systems.)

Both in the fluid model and in the simulation, prioritizing new customers is optimal for all capacity costs. Figure 3.7 reports the percentage errors in the fluid model prescriptions, relative to their counterparts from simulation-based optimization, and the
corresponding percentage profit loss. Figure 3.7(a) shows that the error in the new customer arrival rate (i.e., $\lambda^*_0$) prescribed by the fluid model is lower than 5%, except for $C > 5300$. The error in the number of servers (i.e., $N^*$) prescribed by the fluid model is also lower than 5%, except for $C \in [2300, 2700]$ and $C > 5300$. The larger errors for $C > 5300$ simply arise because such a high cost implies a relatively small system – the fluid model prescribes fewer than 57 servers – too small for a fluid approximation to be accurate. The larger errors in the server cost range $[2300, 2700]$ are more significant – up to around 60%. They arise because by Proposition 3.3, the fluid-optimal number of servers is discontinuous in the server cost at $C = V_1 \mu_1 = 2,367$ (here, the optimal server number $N^*$ drops from 325 to 130 as $C$ crosses 2,367 from below), whereas the simulation-optimal server number decreases gradually in the neighbourhood of $C = V_1 \mu_1$.

Figure 3.7(b) shows that the corresponding percentage profit loss peaks in the neighbourhood of this cost (for $C = 2400$). However, at 5.9% this profit loss is significantly smaller than the error in the number of servers prescribed by the fluid model. Whereas the fluid-optimal number of servers is discontinuous in the server cost at $C = V_1 \mu_1$ the profit rate is continuous at this cost: deploying 325 servers and serving all customers is as profitable as reducing the number of servers to 130 and serving only new customers. Outside the cost range $[2300, 2700]$, the profit loss remains smaller than 1%, again so long as the capacity cost is not too large.

### 3.6.2 Representative Case 2: Higher OTV For Base Customers

$(V_0 - c_0 < V_1)$

In this section, we report the results for a representative case in which the OTV of base customers is higher than the new customer OTV. Specifically, we consider the same parameters as in Table 3.2, except that we set $\theta_1 = 0.3$, so that $V_0 = 43.42 < V_1 = 67.89$.

We perform the comparisons outlined in §3.6.1 for a range of capacity costs, varying the server cost $C$ from 0 to 4800, in increments of 100. (For $C > 4800$ the fluid model prescribes a system with fewer than 25 servers. As noted above, the fluid model is not accurate for such small systems.)

Both in the fluid model and in the simulation, prioritizing base customers is optimal for all capacity costs. Figure 3.8 reports the same performance measures as Figure 3.7 for
Figure 3.7: Fluid model vs. simulation: percentage errors in optimal number of servers and new customer arrival rate, and percentage profit loss, as functions of the capacity cost (priority to new customers).

Case 1. The results are qualitatively similar to those for Case 1 in that (i) the percentage profit loss is typically smaller than the error in the fluid-optimal number of servers and new customer arrival rate, and (ii) the fluid model performance deteriorates in all of these measures as the capacity cost gets large. However, unlike in Case 1, in this case the error in the fluid-optimal prescriptions and the corresponding profit loss do not peak in some intermediate range. This follows because when base customers are prioritized, it is never optimal to overload the system in the fluid model. Therefore, the fluid-optimal number of servers is continuous in the server cost.

3.6.3 Robustness of Fluid Model Performance

In this section, we evaluate the robustness of the results for the two representative cases in §§3.6.1-3.6.2, by reporting the profit performance of the fluid model for a variety of additional parameter settings. Specifically, we consider twelve additional scenarios, by combining $\theta_{01} \in \{0.1, 0.9\}$, $\theta_1 \in \{0.1, 0.9\}$, and $R \in \{0, 0.5, 1\}$. For all of these 12 scenarios, we set $p_0 = p_1 = 20$. All other parameters remain as specified in Table 3.2. For each of the twelve scenarios, we compare the simulation-based optimal profit with the simulated system profit under the fluid model prescriptions for a server cost ranging
from $C = 0$ to $C = C_{\text{max}}$ where $C_{\text{max}}$ is the largest cost for which the fluid-optimal number of servers remains acceptably large (typically $N^* > 25$). For each scenario we report the average, minimum, and maximum percentage profit errors across the server cost range.

Table 3.3 summarizes the results. As discussed above, the deterministic fluid model does not account for queueing effects and customer impatience in evaluating steady-state measures. However, when base customers make a major contribution to the profit rate, the fluid model error in estimating the abandonment rates does not affect the optimal profit significantly. All else equal, an increase in $R$ increases the base customers’ contribution to the profit rate, which reduces the percentage profit loss, as can be seen in each of the parts (a)-(d) of Table 3.3). Similarly, all else equal, an increase in $\theta_{01}$ increases the base customers’ contribution to the profit rate, which in turn reduces the percentage profit loss. (In Table 3.3 compare parts (a) and (c), and parts (b) with (d)). Finally, all else equal, an increase in $\theta_1$ increases the base customers’ contribution to the profit rate. However, the fluid model performance in our experiments seems to be fairly insensitive to $\theta_1$ (Compare sections (a) with (b), and sections (c) with (d)) of Table 3.3). The reason is that when we give priority to base customers (or when we give priority to new customers and the capacity cost is small), the fluid model prescribes to increase the capacity to serve all base customers: Therefore $\theta_1$ does not significantly affect the flow.
of base and new customers.

<table>
<thead>
<tr>
<th>R</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>24.75</td>
<td>61.00</td>
<td>0.89</td>
<td>8.73</td>
<td>3.54</td>
</tr>
<tr>
<td>1</td>
<td>29.30</td>
<td>101.9</td>
<td>0.60</td>
<td>4.42</td>
<td>1.01</td>
</tr>
</tbody>
</table>

(a) $\theta_{01} = 0.1$, $\theta_1 = 0.1$

<table>
<thead>
<tr>
<th>R</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>36.75</td>
<td>37.00</td>
<td>1.17</td>
<td>8.29</td>
<td>3.38</td>
</tr>
<tr>
<td>1</td>
<td>53.45</td>
<td>53.70</td>
<td>0.30</td>
<td>3.89</td>
<td>1.36</td>
</tr>
</tbody>
</table>

(b) $\theta_{01} = 0.1$, $\theta_1 = 0.9$

<table>
<thead>
<tr>
<th>R</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19.85</td>
<td>20.10</td>
<td>1.32</td>
<td>4.22</td>
<td>2.85</td>
</tr>
<tr>
<td>0.5</td>
<td>60.75</td>
<td>61.00</td>
<td>0.06</td>
<td>0.74</td>
<td>0.24</td>
</tr>
<tr>
<td>1</td>
<td>101.7</td>
<td>101.9</td>
<td>0.04</td>
<td>0.27</td>
<td>0.07</td>
</tr>
</tbody>
</table>

(c) $\theta_{01} = 0.9$, $\theta_1 = 0.1$

<table>
<thead>
<tr>
<th>R</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>18.50</td>
<td>20.33</td>
<td>2.02</td>
<td>4.20</td>
<td>2.81</td>
</tr>
<tr>
<td>0.5</td>
<td>168.5</td>
<td>37.00</td>
<td>0.05</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>1</td>
<td>318.5</td>
<td>53.67</td>
<td>0.12</td>
<td>0.50</td>
<td>0.29</td>
</tr>
</tbody>
</table>

(d) $\theta_{01} = 0.9$, $\theta_1 = 0.9$

Table 3.3: Fluid model vs. simulation: minimum, maximum, and average percentage profit loss.

### 3.7 Concluding Remarks

This paper proposes and analyzes a novel call center model that considers the impact of past demand and service quality on customer retention. We study the problem of maximizing profits by controlling customer acquisition, retention, and service quality via
promotions, priorities, and staffing. The key feature of our model is that the customer base depends on the abandonment rates of new and base customers, reflecting their priority and the system load. We specify a stochastic queueing model, characterize the optimal controls analytically based on a deterministic fluid model, and show via simulation that these prescriptions yield near-optimal performance for the underlying stochastic model. The following findings and implications emerge from our analysis.

First, we provide novel insights and guidelines on call center management. We derive three metrics which form the basis for call center decisions, the CLV of a base customer and the OTVs of new and base customers. These metrics, unlike standard ones in marketing, depend on operations through the service quality, i.e., the probability of getting served. We show that it is optimal to prioritize the customers with the higher OTV. In contrast to standard priority policies such as the $c\mu$ rule, this policy accounts for the financial impact of customers’ future calls. We further show how the jointly optimal promotion, priority and staffing policy depends on the CLV, the OTVs, and the capacity cost. The results on the optimal promotion and staffing levels underscore the importance of considering the interaction between customer metrics and operations; i.e., the contribution of an additional new customer to net revenues depends on the system load and the priority policy. These results also provide insights on the optimal service quality. Specifically, offering deliberately poor service to base customers is optimal only if new customers have the higher OTV, e.g., due to prohibitive switching costs for base customers, and the capacity cost per call exceeds the OTV of a base customer. Under these conditions it is optimal to prioritize new customers and to overload the system, possibly significantly. In this regime the call center’s primary goal is to serve and acquire new customers to grow the customer base, not to serve base customers. This result lends some theoretical support for the anecdotal evidence that locked-in customers of firms such as mobile phone service providers commonly experience long waiting times when contacting the call center.

Second, from a modeling and methodological perspective, we conclude that the benefits of our solution approach via the analysis of the deterministic fluid model outweigh its disadvantages. The fluid model is particularly appealing because it is both analytically tractable, as is obvious from §§3.4-3.5, and its prescriptions yield near-optimal (gross) profit performance for the approximated stochastic system, as shown in §3.6. Further-
more, the fluid model analysis provides valuable intuition on the problem structure. The tractability and accuracy of the fluid model demonstrated in this paper suggest that a similar approach may prove effective on further problems of joint CRM and call center management.

We close by outlining three future research directions. First, in terms of customer modeling, we model base customers that defect based only on their last call center interaction. It would be interesting to model customer behaviour that responds to longer service histories. Second, in terms of system modeling and solution methodology, one potentially fruitful avenue is to consider refinements to the fluid approximation we use in this paper, and to establish formal limit results. Third, in terms of data, as discussed in §3.3.1, our model can be tailored to a range of call center characteristics. Many of our model inputs are reasonably well measurable based on data that call centers track. It would be quite interesting to estimate our model parameters and also to refine our model, based on such data. The results could be of value to measure and compare CLV and OTV metrics within and across call centers, and more importantly, to study the impact of service quality attributes such as waiting time on these metrics.
3.8 Appendix: Proofs

We omit the proofs for the results in Section 3.4, as they are subsumed by the results in Section 3.5.1.

Proof of Lemma 3.1. Clearly (3.53) implies (3.52). To prove (3.53), because $\overline{V}_{i-1} > \overline{V}_i$ for $i > k$ by (3.51), it suffices to show that $\overline{V}_{i-1} \geq \overline{V}_i \Rightarrow \overline{V}_i > \overline{V}_{i+1}$ for $i \in \{1, ..., m-1\}$. From (3.49) note that

$$\overline{V}_i = \frac{s_{i-1} \overline{V}_{i-1} + s_i V_i \mu_i}{\bar{s}_i}, \quad i = 1, 2, ..., m, \quad (3.83)$$

so that $\overline{V}_{i-1} \geq \overline{V}_i \Rightarrow \overline{V}_i \geq V_i \mu_i$. Since $V_i \mu_i > V_{i+1} \mu_{i+1}$ we have $\overline{V}_i > V_{i+1} \mu_{i+1}$ so that $\overline{V}_i > \overline{V}_{i+1}$ by (3.83). Next, (3.54) follows because $\overline{V}_{k-1} \leq \overline{V}_k > \overline{V}_{k+1}$ by (3.53) and from (3.83). 

Proof of Proposition 3.4. Prioritizing base customers in decreasing order of their $V_i \mu_i$-index is equivalent to the capacity allocation constraints

$$N_i < N_0 \mu_0 s_i \Longrightarrow N_{i+1} = 0, \quad i \in \{1, ..., m-1\}, \quad (3.84)$$

i.e., customers of type $i + 1$ are only served if all of type $i$ are served (constraint (3.46) is binding for $i$). This is optimal because $V_i \mu_i > V_{i+1} \mu_{i+1}$ for $i \geq 1$ and by inspection of (3.44) and (3.46)-(3.47).

We next establish the optimal priority of new customers. Let $\Pi^l$ denote the maximum profit under the policy that prioritizes base customers of type $i \in \{1, ..., l\}$ over new customers, new customers over base customers of type $i > l$, and base customers in decreasing order of their $V_i \mu_i$-index. This policy is equivalent to (3.84) and the following capacity allocation constraints. All base customer calls of type $i \in \{1, ..., l\}$ are served, i.e., (3.46) is binding:

$$N_i = N_0 \mu_0 s_i, \quad i \in \{1, ..., l\}, \quad (3.85)$$

and if some new customers are not served (so (3.45) is slack) then no calls of type $i > l$ are served:

$$N_0 < \lambda_0 s_0 \Longrightarrow N_i = 0, \quad i > l. \quad (3.86)$$
To prove the claim we show that \( \Pi^k \geq \Pi^l \) for \( l \neq k \). The profit \( \Pi^l \) is the maximum value of the objective function in (3.44) under a capacity allocation vector \( N \) that satisfies (3.45)-(3.47) and (3.84)-(3.86). Using (3.48), (3.49), and (3.85) we get

\[
\sum_{i=0}^{l} N_i V_i \mu_i = N_0 \mu_0 \bar{s}_l \bar{V}_l \quad \text{and} \quad \sum_{i=0}^{l} N_i = N_0 \mu_0 \bar{s}_l,
\]

and substituting into (3.44) and (3.47), respectively, it is straightforward to verify that

\[
\Pi^l(\lambda_0, N) = \min (\lambda_0 \bar{s}_l, N) \bar{V}_l + \sum_{i=l+1}^{m} \min (\lambda_0 s_i, (N - \lambda_0 \bar{s}_{l-1})^+) V_i \mu_i - \lambda_0 c_0 - S(\lambda_0) - CN. \tag{3.87}
\]

It remains to show \( \Pi^k \geq \Pi^l \) for \( l \neq k \), which also proves (3.58) in light of (3.87). Furthermore, the capacity allocation specified in (3.55)-(3.57) is implied by (3.84)-(3.86) for \( l = k \).

We have

\[
\Pi^l - \Pi^{l-1} = \min (\lambda_0 \bar{s}_l, N) \bar{V}_l - \min (\lambda_0 \bar{s}_{l-1}, N) \bar{V}_{l-1} - \min (\lambda_0 s_l, (N - \lambda_0 \bar{s}_{l-1})^+) V_l \mu_l. \]

Using (3.83) to substitute for \( \bar{V}_l \), and noting that

\[
\min (\lambda_0 s_l, (N - \lambda_0 \bar{s}_{l-1})^+) = \min (\lambda_0 \bar{s}_l, N) - \min (\lambda_0 \bar{s}_{l-1}, N),
\]

we have

\[
\bar{s}_l (\Pi^l - \Pi^{l-1}) = (V_l \mu_l - \bar{V}_{l-1}) (\bar{s}_l \min (\lambda_0 \bar{s}_{l-1}, N) - \bar{s}_{l-1} \min (\lambda_0 \bar{s}_l, N)).
\]

Note that \( V_l \mu_l \geq \bar{V}_{l-1} \iff \bar{V}_l \geq \bar{V}_{l-1} \) by (3.83) and \( \bar{s}_l \min (\lambda_0 \bar{s}_{l-1}, N) - \bar{s}_{l-1} \min (\lambda_0 \bar{s}_l, N) \geq 0 \), so

\[
\Pi^l \geq \Pi^{l-1} \iff \bar{V}_l \geq \bar{V}_{l-1}.
\]

It follows from (3.53) in Lemma (3.1) that \( \Pi^k \geq \Pi^l \) for \( l \neq k \).

**Proof of Proposition 3.5.** Parts 1 and 2 follow from the profit function (3.58) and from (3.53)-(3.54) in Lemma (3.1): (i) if \( C \geq \bar{V}_k \) then \( \Pi \leq 0 \) for \( N \geq 0 \), and (ii) if \( C < \bar{V}_k \) then \( \Pi \) is strictly increasing for \( N \in [0, \lambda_0 \bar{s}_k) \), nondecreasing for \( N \in [\lambda_0 \bar{s}_{l-1}, \lambda_0 \bar{s}_l] \) and
\(i \in \{k + 1, \ldots, m\}\) if and only if \(V_i \mu_i \geq C\), and \(\Pi\) is strictly decreasing for \(N > \lambda_0 s_m\). This establishes (3.59). It follows by the optimal priority rule of Proposition 3.4 that it is optimal to serve the customer types as claimed. Furthermore, because \(V_i \mu_i > V_{i+1} \mu_{i+1}\) for \(i \geq 1\), it follows that the system is overloaded, i.e., \(N^* < \lambda_0 s_m\) if and only if \(C > V_m \mu_m\).

Part 3. Using (3.59) to substitute for \(N^*\) into (3.58) yields (3.60).

**Proof of Lemma 3.2.** The proof of (3.64)-(3.65) is exactly the same as that of (3.53)-(3.54). By the definition of \(k^*\) in (3.62), the inequality (3.66) is equivalent to \(\tilde{V}_{k-1} \leq \tilde{V}_k\) if \(k \geq 1\). Using (3.61) we have

\[
\tilde{V}_{k-1} \leq \tilde{V}_k \iff \tilde{V}_{k-1} - \tilde{V}_k \leq c_0 \left( \frac{1}{\tilde{s}_{k-1}} - \frac{1}{\tilde{s}_k} \right),
\]

which is satisfied because \(\tilde{V}_{k-1} - \tilde{V}_k \leq 0\) by (3.53), \(c_0 \geq 0\), and \(\tilde{s}_k > \tilde{s}_{k-1}\).

**Proof of Proposition 3.6.** We first prove the condition for profitable operation. First note that (3.53), (3.61), (3.64), and (3.66) imply that \(e V_{k^*} \leq V_{k^*}\). If \(C \geq V_{k^*}\) then it is not optimal to operate by Proposition 3.5. If \(C < V_{k^*}\) then

\[
\Pi (\lambda_0) = \lambda_0 \left( \tilde{s}_k (\tilde{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C) \right) 1\{V_i \mu_i \geq C\} - S(\lambda_0)
\]

by (3.60)-(3.61). Noting that \(\Pi'' < 0 < S''\), it follows that \(\lambda_0^*\) is uniquely defined as \(\lambda_0^* = 0\) if \(\Pi'(0) \leq 0\) and \(\lambda_0^* > 0\) if \(\Pi'(0) > 0\). Because \(S'(0) = 0\) we have

\[
\Pi'(0) = \tilde{s}_k (\tilde{V}_k - C) + \sum_{i=k+1}^{k^*} s_i (V_i \mu_i - C) 1\{V_i \mu_i \geq C\}.
\] (3.88)

Note that \(\Pi'(0)\) strictly decreases in \(C\). Therefore, \(\Pi'(0) > 0 \iff C < \tilde{V}_{k^*}\) holds if and only if \(\Pi'(0) = 0\) for \(C = \tilde{V}_{k^*}\), which we show next. For \(C = \tilde{V}_{k^*}\) we have

\[
\Pi'(0) = \tilde{s}_k (\tilde{V}_k - C) + \sum_{i=k+1}^{k^*} s_i (V_i \mu_i - C) = \tilde{s}_{k^*} (\tilde{V}_{k^*} - C) = 0,
\]

where the first equality follows from (3.88) and from (3.65)-(3.66) in Lemma 3.2, and the second equality follows by the definitions (3.49) and (3.61).
If \( C < \tilde{V}_{k^*} \) then Parts 1-2 follow from Parts 1-2 of Proposition 3.5 because \( \tilde{V}_{k^*} \leq \tilde{V}_{k^*} \) by (3.61), whereas (3.68) in Part 3 follows from (3.88): from (3.65)-(3.66) in Lemma 3.2, and by (3.49) and (3.61),

\[
\Pi'(0) = \bar{s}_{k^*} \left( \tilde{V}_{k^*} - C \right) + \sum_{i=k^*+1}^{m} s_i \left( V_i \mu_i - C \right) 1_{\{V_i \mu_i \geq C\}} \text{ for } C < \tilde{V}_{k^*},
\]

and \( \Pi'(\lambda_0) = 0 \iff \Pi'(0) = S'('\lambda_0). \]
Chapter 4

Customer Acquisition and Service Quality for a Call Center with Time-Varying Demand Response

4.1 Introduction

As discussed in Chapter 3, call centers are an integral part of many businesses and call center service experience significantly affects customers’ opinions about the firm, their repeat purchase behavior, and overall demand growth. However, the relationship between queueing-related service quality and customer retention and long-term customer value is mainly overlooked in the literature. The standard approach in the call center literature has been to model a firm’s customer base as independent of past interactions and the CRM literature usually ignores supply constraints. (see §3.2 for a detail literature review).

In Chapter 3 we discussed issues raised in the call center of a credit cart company (as explained in §3.3), assuming that the optimal advertisement can be kept at a continues and fixed expenditure level to maintain the sales rate at its optimal rate. In this chapter, in contrast, we consider a firm that follows a periodic advertising policy and advertises at the start of each period. The demand response to the advertisement diminishes over time during the period. We also assume the firm may determine the service levels for new and base customers during each period by continuously adjusting the staffing level. Dynamic
staffing is common for call centers with time-varying demand; see, e.g., Jennings et. al (1996), Green et. al (2007), Hampshire et. al (2008), and Feldman et al. (2008).

These assumptions yield a non-stationary queueing system. To obtain a tractable model, we approximate the non-stationary system by a pointwise stationary fluid model (Bassamboo et al. 2009) that assumes the queueing system reaches steady-state at each instant in time. For detailed discussion on pointwise stationary model we refer to Green and Kolesar (1991) and Massey and Whitt (1998).

The pointwise stationary fluid model framework, enables us to capture the system evolution over time and model the interactions of promotion and service policy during and between different periods. Because the resulting fluid model is dynamic, we analyze it using optimal control theory (see Sethi and Thompson 2000) and derive the optimal promotion and service policies over different advertisement period.

The contributions of this chapter are threefold: First, we extend the model developed in Chapter 3 to consider time-varying demand response and the customer base evolution over different advertising periods. This model keeps track of new and “base” (repeat) customers to provide guidelines on how to maximize the profit of an inbound call center by using three levers: Periodic promotions to control customer acquisition under a time-varying demand response; and staffing and priority policies to control service levels to new and base customers. The specific service level we focus on is the percentage of calls that are served before customers abandon, but the model can be used to consider other service levels such as the expected waiting time.

Second, we derive the expected customer lifetime value (CLV) of a base customer by considering how she will be served during the advertisement period under any service policy. Unlike standard CLV metrics, the expected customer lifetime value in our model is time-dependent. It may increase or decrease when approaching the end of the advertisement period. Furthermore, we derive the expected value of serving a call of new or base customers based on the time of the call during the advertisement period. These metrics relate customer characteristics with service quality and play an important role in characterizing the optimal call center decisions.

Finally, we approximate the stochastic model by a pointwise stationary fluid model and generate novel results on how to coordinate service and promotion policies in a call center over different advertisement periods. For a single period problem, we determine
the effect of the terminal value of a base customer on the optimal service policy. We established that the optimal service level allocation to new and base customers takes the form of a bang-bang policy. That is, at each moment during the advertisement period, the call center should serve all or none of the calls from each type of customers, and the policy transitions between these two extremes at most once during the advertisement period. We then consider the multi-period (finite or infinite) advertisement problem, and showed that a bang-bang policy is still optimal. Furthermore, we established, based on a simulation study, that investigating the pointwise stationary fluid version of the stochastic model leads to good approximations of main performance measures such as evolution of the customer base, the abandonment rate, and the profit rate. The accuracy of the approximation suggests that the optimal controls, i.e., the bang-bang service policy, and insights derived based on this model would hold for the stochastic model as well.

The rest of the chapter is organized as follows: In §4.2 we specify the stochastic queueing model, approximate the stochastic model with a point-wise stationary fluid model, and formulate the firm’s profit maximization problem as a dynamic programming problem over a finite horizon. In §4.3, we characterize the optimal promotion and service policies for a single period problem. Then, in §4.4, we derive the optimal prescriptions on service and promotion policy over multi periods of advertisement. In §5 we verify the accuracy of the fluid approximation by comparing the performance measures under the fluid assumption with measures derived from simulating the stochastic model. Our concluding remarks are presented in §4.6. All proofs are in the appendix.

4.2 Model Formulation

Consider a firm that serves two types of customers. New customers (type $n$) are first-time callers attracted by the firm advertisement that may join the customer base after receiving service. Base customers (type $b$) are repeating customers. Figure 4.1 depicts the customer flow through the system, showing the flow of new customers by dashed lines and the flow of base customers by solid lines.

We model the call center as a multiple server system, where the call center manager can adjust the number of servers over short intervals. Service times are i.i.d. with mean $1/\mu_n$ and $1/\mu_b$, for new and base customers, respectively. Calls arrive as detailed below.
Customers wait in queue if the system is busy upon arrival, but they are impatient. Abandonment times have mean, $1/\tau_n$ and $1/\tau_b$ for new and base customers, respectively. Table 4.1 summarizes the notation.

Assume that the firm follows a periodical advertisement policy over a finite horizon with $N$ periods of advertisement (the infinite horizon problem will be discussed as a limiting case in §4.4.1). At the start of each period, the firm sends out the promotions and advertisements designed for that period. New customers are attracted by these offers/promotions, however, as time passes, customers’ response decays until a new advertisement emerges at the start of the next period.

For simplicity of the model, we assume that all advertisement periods have identical and exogenously given length $T$. This assumption can be generalized to settings with different advertisement period lengths. However, it is common in practice where firms advertisement cycles are often independent of the level of advertisement and coincide with fixed time intervals such as a month or a quarter.

Consider the $i^{th}$ advertisement period. We assume that new customer calls arrive to the system following a non-stationary Poisson process with rate

$$\lambda_n^i(t) = \lambda_n^i(0)e^{-\gamma_n t},$$

at time $t \in (0, T]$, where $\lambda_n^i(0)$ is the new customer arrival rate at the start of the $i^{th}$
period, and $\gamma_n$ is the decay rate of new customers’ demand response to the advertisement.

To keep the model tractable, we assume $\lambda_n^i(0)$ depends only on the firm’s advertisement spendings at the $i^{th}$ period and is independent of previous advertisements and service levels, e.g. a time limited promotion ($\lambda_n^i(t) = 0$ for each $t > T$).

Let $S(\lambda_n^i(0))$ denote the advertising spending needed to generate arrival rate pulse of the size $\lambda_n^i(0)$. We assume that the initial response rate of new customers to advertising spending follows the law of diminishing returns (Simon and Arndt 1980), so $S(\lambda_n^i(0))$ is strictly increasing and strictly convex in $\lambda_n^i(0)$. For analytical convenience, we assume that $S$ is twice continuously differentiable and $S'(0) = 0$.

Let $x_n^i(t)$ denote the number of base customers at time $t \in [0,T]$ during the $i^{th}$ advertisement period. We require that the size of the customer base at the start of each period equals its size at the end of the previous period: $x_n^i(0) = x_n^{i-1}(T)$. During the advertisement period, new and base customer calls affect the evolution of the customer base over time. If a new customer receives service, she joins the customer base with probability $\theta_n > 0$. The times between successive calls of a base customer are independent
and exponentially distributed with mean $1/r_b$, so $r_b \geq 0$ is the average call rate per base customer per unit time. We assume that $1/r_b >> 1/\mu_b, 1/\tau_b$, i.e., the mean time between calls from any given customer is much larger than the mean service and abandonment times.

We capture the “customer loyalty coefficient” (Reichheld 2000, Hall and Porteus 2000) by defining $\theta_b$ such that a base customer who abandons her call stays in the customer base with probability $\theta_b$, but departs from the customer base and does not call back again with probability $1 - \theta_b$. There is also an attrition from the customer base due to involuntarily and/or call-independent reasons (e.g., relocation or death). We assume the call-independent attrition of a base customer happens with rate $\gamma_b > 0$.

The profit of operating the firm is as follows. On average the firm generates a profit of $p_n$ per new customer call it serves and incurs a cost of $c_n \geq 0$ per abandonment of any call from new customers. We model two potential profit streams from base customers. First, the firm may generate a call-independent profit at an average rate of $R \geq 0$ per unit time per base customer, which captures monetary flows that are independent of call center interactions. Second, base customers may also call with a purchase or a service request. On average the firm generates a profit of $p_b$ per base customer call it serves and incurs a cost of $c_b \geq 0$ per abandonment of any call from base customers.

Finally, let $A$ represent the terminal value of a base customer who remains in the customer base after the last advertisement period. This value measures the expected profit a base customer generates outside the advertisement period, while the firm operates on a regular basis.

### 4.2.1 Fluid Model and System Evolution

The exact model that we described is not tractable analytically. To study the problem under a tractable modeling framework, we approximate it by a pointwise stationary fluid model (PSFM) (Green and Kolesar, 1991, Bassamboo, et. al, 2009). As a fluid model, we substitute mean flow rate for the stochastic new and base customer arrival flows, which is justifiable by considering the relatively high arrival rate and service capacity of a large call center, as discussed in Chapter 3. As a pointwise stationary model, at each point of time, we “freeze” the arrival rate of new and base customers, use steady state analysis to compute the performance measures and decide on service and staffing levels.
In our model, we consider abandonment, and we assume a relatively large number of servers. Bassamboo et al (2009) discussed that such models are well approximated by a pointwise stationary fluid model if the service and reneging rates are large relative to the rate of changes in arrival rate. We, indeed, assume that arrival rates are changing over time periods of days (as the advertisement period length, $T$, may be a quarter), so these rates do not change significantly during the time an arbitrary customer spends in the call center.

Define $q^i_n(t)$ and $q^i_b(t)$ as the probabilities that new and base customers arriving at time $t \in (0,T]$ receive service, respectively. Under the PSFM assumption, at any time $t \in (0,T]$, serving $q^i_n(t)$ percent of new customer and $q^i_b(t)$ percent of base customer calls requires at least \((\lambda^i_n(t)/\mu_n) q^i_n(t) + (x^i_b(t) r_b/\mu_b) q^i_b(t)\) servers. Positive capacity cost and infinitely flexible staffing guarantee that the optimal system should never be underloaded (though, it can be overloaded if the firm decides not to serve all new and/or base customer calls). Thus, at optimality, the required number of servers is

$$K^i(t) = \frac{\lambda^i_n(t)}{\mu_n} q^i_n(t) + \frac{x^i_b(t) r_b}{\mu_b} q^i_b(t),$$

and controlling the service probabilities becomes equivalent to controlling the staffing and allocation policies.

Let $q^i_n$ and $q^i_b$ represent service policies for new and base customers, respectively. These policies are functions from $(0,T]$ to $[0,1]$ indicating $q^i_n(t)$ and $q^i_b(t)$ for any time $t \in (0,T]$ during the $i^{th}$ period.

Considering $q^i_n$ and $q^i_b$, under PSFM assumption, we next determine the evolution of the customer base over time. At any time $t \in (0,T]$, new customers are served with rate $\lambda^i_n(t) q^i_n(t)$ and therefore they join the customer base with rate $\lambda^i_n(t) q^i_n(t) \theta_n$. Base customers arrive to the call center with rate $x^i_b(t) r_b$ at time $t$. The fraction of base customers who abandon their calls is $1 - q^i_b(t)$, from which $(1 - \theta_b)$ percent terminate their relation with the firm and leave the customer base. Also, a call-independent attrition rate of $x^i_b(t) \gamma_b$ happens at time $t$. Therefore, at each moment $t$, the size of the customer base changes with time as follows:

$$\frac{d x^i_b(t)}{dt} = \lambda^i_n(t) q^i_n(t) \theta_n - x^i_b(t) \left( \gamma_b r_b \left( 1 - q^i_b(t) \right) (1 - \theta_b) \right).$$

71
Note that unlike the stationary case in Chapter 3, the size of the customer base cannot be derived as a closed form function of new and base customers’ throughputs.

### 4.2.2 The Optimization Problem

We consider the dynamic of our system under three main levers: by controlling the advertisement policy the firm controls the new customer arrival pulse at the start of each period, \( \lambda^i_n(0) \), and by controlling the servers’ staffing and allocation policy, the firm sets the service policies for both new and base customers, i.e., \( q^i_n \) and \( q^i_b \), respectively.

Focusing on a finite horizon problem, we maximize the total profit the firm may generate during the advertisement period, while considering the terminal effect of our decisions.

Define \( P(\lambda^i_n(0), q^i_n, q^i_b; x^i_b(0)) \) as the firm’s total profit during the \( i^{th} \) advertisement period, given the customer base of size \( x^i_b(0) \) at the start of period \( i \):

\[
P(\lambda^i_n(0), q^i_n, q^i_b; x^i_b(0)) = \int_0^T \left( \lambda^i_n(t) \left( p_n q^i_n(t) - c_n \left(1 - q^i_n(t)\right)\right) + x^i_b(t) \left( R + r_b \left( p_b q^i_b(t) - c_b \left(1 - q^i_b(t)\right)\right)\right) - CK^i(t) \right) dt - S(\lambda^i_n(0)),
\]

where the first product in the integral is the profit from new customer calls, the second is the profit from base customers (call-independent and call-dependent profit), and the third is the staffing cost. Finally, the term outside the integral captures the advertisement spending for the period. Substituting (4.2) in (4.4) yields:

\[
P(\lambda^i_n(0), q^i_n, q^i_b; x^i_b(0)) = \int_0^T \left( \lambda^i_n(t) \left( p_n q^i_n(t) - c_n \left(1 - q^i_n(t)\right)\right) + x^i_b(t) \left( R + r_b \left( p_b q^i_b(t) - c_b \left(1 - q^i_b(t)\right)\right)\right) - CK^i(t) \right) dt - S(\lambda^i_n(0)),
\]

in which the staffing cost is incorporated in the profit of serving new and base customers.

The firm’s profit optimization problem may be written as a dynamic programming problem, in which stages are the advertisement periods and the state of the system at each period is the size of the customer base at the start of that period. Firm’s decisions in each period affect the profit in the following periods through the ending level of the
The customer base in that period as

\[ x_{i+1}^b (0) = x_i^b (T), \quad i = 1, ..., N. \] (4.6)

Let \( \Pi^i (x_i^b (0)) \) denote the maximum profit the firm can generate from stage \( i \) to the end of the problem horizon, given that at the start of period \( i \), there are \( x_i^b (0) \) base customers present in the system. Dynamic programming recursion can be written as:

\[
\Pi^i (x_i^b (0)) = \max_{\lambda_n^i (0), q_i^1, q_i^b} \left\{ P \left( \lambda_n^i (0), q_i^1, q_i^b; x_i^b (0) \right) + \Pi^{i+1} (x_{i+1}^b (T)) \right\} \quad i = 1, ..., N \quad (4.7)
\]

subject to (4.3), (4.6), given the initial value \( x_1^b (0) \), and the boundary condition \( \Pi^{N+1} (x_N^{N+1} (0)) = A x_N^{N+1} (0) \) where \( A \) is the terminal value of a base customer.

### 4.2.3 Service Value of Base or New Customers’ Calls

In this section we derive the value of serving new or base customers’ calls. These metrics depend on the service policy and govern the structure of the optimal decisions.

Consider a base customer who is in the customer base at time \( k \in (0, T] \) during the \( i^{th} \) advertisement period. At this instant, the customer base decays with a rate \( (\gamma_b + r_b (1 - q_i^b (k)) (1 - \theta_b)) \) per base customer (see the discussion above (4.3)). In other words, given that a base customer has remained in the customer base until time \( k \), \( (\gamma_b + r_b (1 - q_i^b (k)) (1 - \theta_b)) \) represents the conditional probability density function that she leaves the customer base at time \( k \).

Given the service policy \( q_i^b \) for the \( i^{th} \) advertisement period, let \( U^i (t, s, q_i^b) \) denote the probability that a base customer who is in the customer base at time \( t \in (0, T] \) remains in the customer base until time \( s \in [t, T] \):

\[
U^i (t, s, q_i^b) = 1 - \int_t^s U^i (t, k, q_i^b) \left( \gamma_b + r_b (1 - q_i^b (k)) (1 - \theta_b) \right) dk. \quad (4.8)
\]

We call \( U^i (t, s, q_i^b) \) the base customers’ survival probability. In particular, \( U^i (t, T, q_i^b) \) represents the probability that a base customer who is in the system at time \( t \), remains at the customer base until the end of the current advertisement period.

Different service policies for base customers during the advertisement period affect
base customers lifetime value by changing both the profit generated by a base customer and her survival probability. Let \( \overrightarrow{q}^b = (q^b_1, q^b_{i+1}, \ldots, q^b_N) \) represent the vector of service policies over periods \( i \) to \( N \). Define \( L^i(t, \overrightarrow{q}^b) \) as the expected lifetime value of a base customer who is in the customer base at time \( t \) during the \( i^{th} \) advertisement period, given \( \overrightarrow{q}^b \):

\[
L^i(t, \overrightarrow{q}^b) = \int_t^T U^i(t, s, q^b_i) \left( R + r_b \left( \left( \frac{p_b - \frac{C}{\mu_b}}{\mu_b} \right) q^b_i(s) - c_b \left( 1 - q^b_i(s) \right) \right) \right) ds
+ U^i(t, T, q^b_i) L^{i+1} \left( 0, \overrightarrow{q}^{i+1}_b \right), \quad (4.9)
\]

with the boundary condition

\[
L^{N+1} \left( 0, \overrightarrow{q}^{N+1}_b \right) = A. \quad (4.10)
\]

The expected lifetime value of a base customer consists of the sum of the expected profit she generates during the advertisement period and her expected value after time \( T \). The expected profit of a base customer at time \( s \in [t, T] \) is derived by multiplying her survival probability \( U^i(t, s, q^b_i) \) by the rate of the profit she generates (as in (4.5)). The value of a base customer after advertisement period \( i \) is her survival probability until time \( T \), i.e., \( U^i(t, T, q^b_i) \), multiplied by her expected lifetime value from period \( i+1 \) to the end. Finally, (4.10) follows from that the terminal value of a base customer at the end of the last advertisement period is \( A \).

Let \( V^i_b(t, \overrightarrow{q}^b) \) denote the value of serving a base customer call at time \( t \) during the \( i^{th} \) advertisement period under service policy \( \overrightarrow{q}^b \):

\[
V^i_b(t, \overrightarrow{q}^b) = p_b + c_b + (1 - \theta_b) L^i \left( t, \overrightarrow{q}^b \right). \quad (4.11)
\]

Serving a base customer call at time \( t \) yields a profit \( p_b + L^i \left( t, \overrightarrow{q}^b \right) \), where \( p_b \) is the immediate profit of serving her, and \( L^i \left( t, \overrightarrow{q}^b \right) \) is her expected value to the firm at time \( t \). Not serving the base customer call yields a profit \( -c_b + \theta_b L^i \left( t, \overrightarrow{q}^b \right) \), where \( -c_b \) is the immediate abandonment penalty, and the second term is the expected value of a base customer who abandons her call at time \( t \). The difference between these two gives (4.11).

Similarly, let \( V^i_n(t, \overrightarrow{q}^b) \) denote the value of serving a new customer call at time \( t \)
during the $i^{th}$ advertisement period, under base customers’ service policy $\overrightarrow{q_b}$:

$$V^i_n(t, \overrightarrow{q_b}) = p_n + c_n + \theta_n L^i(t, \overrightarrow{q_b}).$$  \hfill (4.12)

Serving a new customer call at time $t$ yields an instant profit $p_n$, and expected profit $\theta_n L^i(t, \overrightarrow{q_b})$ over her lifetime after joining the customer base. Not serving a new customer results in an instant abandonment penalty $-c_n$. The difference between these two yields (4.12). Note that the value in (4.12) depends on base customers service policy, $\overrightarrow{q_b}$, i.e., the value of serving a new customer depends on the quality of service she receives later if she joins the customer base.

### 4.3 The Single Period Problem

In this section we focus on the single period problem. This problem may be seen as the last period of a multi-period problem discussed in §4.2, and its solution is a building block to determine the optimal controls for the multi-period problem. The single period problem may also be used to model an ad-hoc advertisement over a limited period of time, or to model a shift in customers’ demand due to exceptional market conditions over a limited time interval. For simplicity of the notation, we drop the advertisement period index.

From (4.9) we have that at the end of the advertisement period, the lifetime value of a base customer equals her terminal value, independent of base customers’ service policy $\overrightarrow{q_b}$:

$$L(T, \overrightarrow{q_b}) = A.$$  \hfill (4.13)

Therefore, from (4.5), (4.7), and (4.13) the optimization problem becomes:

$$\Pi(x_b(0)) = \max_{\lambda_n(0), \delta_n, q_b} \int_0^T \left( \lambda_n(t) \left( \left(p_n - \frac{c_n}{\mu_n}\right) q_n(t) - c_n \left(1 - q_n(t)\right) \right) + x_b(t) \left( R + \delta_b \left( \left(p_b - \frac{c_b}{\mu_b}\right) q_b(t) - c_b \left(1 - q_b(t)\right) \right) \right) \right) dt$$

$$- S(\lambda_n(0)) + Ax_b(T)$$ \hfill (4.14)

subject to (4.3).
In §4.3.1 we discuss the customer value metrics under fixed service probability for base customers. Then in §4.3.2 we determine the optimal service policy for new and base customers for any fixed promotion level and finally in §4.3.3 we jointly optimize the advertisement and service policy.

### 4.3.1 Stationary Service Policy

In this section, we characterize customer value metrics under a stationary service policy for base customers $q_b = q_b$, i.e., $q_b(t) = q_b$ for any $t \in (0, T]$ where $q_b$ is fixed.

**Proposition 4.1.** Under any base customers’ stationary service policy, $q_b = q_b$, the lifetime value of a base customer at time $t$ during the advertisement period is

$$L(t, q_b) = \frac{R + r_b ((p_b - C/\mu_b) q_b - c_b (1 - q_b))}{\gamma_b + r_b (1 - q_b) (1 - \theta_b)} \left( 1 - e^{-(\gamma_b + r_b (1 - q_b) (1 - \theta_b))(T-t)} \right) + Ae^{-(\gamma_b + r_b (1 - q_b) (1 - \theta_b))(T-t)}, \quad (4.15)$$

which is decreasing in time if

$$A < \frac{R + r_b ((p_b - C/\mu_b) q_b - c_b (1 - q_b))}{\gamma_b + r_b (1 - q_b) (1 - \theta_b)}, \quad (4.16)$$

and is increasing in time, otherwise.

The right hand side in (4.16) equals $\lim_{T \to \infty} L(t, q_b)$ that represents the expected profit a base customer may generate during a very long advertisement period. In other words, if (4.16) holds, a base customer’s terminal value is relatively small in comparison to the value she is generating during the advertisement period. Therefore, a base customer lifetime value is decreasing when the end of the advertisement period is approaching. Conversely, if (4.16) does not hold, the terminal value of a base customer is relatively large, and her lifetime value increases with time.

Next, we consider the value of serving a base customer call under the two extreme stationary policies, $q_b = 0$, and $q_b = 1$. From (4.11) and (4.15), the value of serving a call of a base customer at time $t$ but not any of her future calls during the advertisement
period (i.e., \( q_b = 0 \)) is
\[
V_b(t, 0) = p_b + c_b + (1 - \theta_b) \left( \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \left( 1 - e^{-(\gamma_b + r_b (1 - \theta_b))(T-t))} + A e^{-(\gamma_b + r_b (1 - \theta_b))(T-t)} \right) \right).
\]
(4.17)

Similarly, the value of serving a call of base customer at time \( t \) and all of her future calls during the advertisement period (\( q_b = 1 \)) is
\[
V_b(t, 1) = p_b + c_b + (1 - \theta_b) \left( \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \left( 1 - e^{-\gamma_b(T-t)} \right) + A e^{-\gamma_b(T-t)} \right).
\]
(4.18)

Note that value of serving a base customer call is monotone in time for both (4.17) and (4.18). In particular, it follows from (4.17) that if
\[
A < \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)},
\]
(4.19)
\( V_b(t, 0) \) is decreasing in time and the firm becomes less willing to serve a base customer call if it plans not to serve any of her future calls. Similarly, it follows from (4.18) that if
\[
A \geq \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b},
\]
(4.20)
the firm becomes more willing to serve a base customer’s current and all future calls near the end of the period.

### 4.3.2 Optimal Service Policy for a Fixed Promotion Level

Consider the case where the promotion level and thus the new customer arrival rate is fixed but the manager still controls the service policy for new and base customers. This setting is applicable when the promotion policy is taken by the marketing department independently of the capacity and service planning done by the operations department. It is also possible that the new customer arrival rate is predictable but not controllable by the firm, e.g., the burst in new customer arrival rate to a phone service provider company after introducing a new long-awaited cell phone.

Base customers are repeating customers whose calls are independent of new customer arrival stream, therefore the manager may set the optimal staffing policy (which corresponds to optimal service policy) for base customers independent of new customer
arrival and service rate. But, to determine the optimal service policy for a new customer, the manager should consider the service she may receive if she joins the customer base. Therefore, we first determine the optimal service policy for base customers in Proposition 4.2, and then in Proposition 4.3 we characterize the optimal service policy for new customers as a function of the optimal service policy for base customers.

4.3.2.1 Optimal Service Policy for Base Customers

To determine the optimal service policy for base customers, the firm needs to compare the value of serving a base customer who arrives at any time instant \( t \in (0, T] \) (i.e. \( V_b(t, q_b) \)) with the cost of serving her call (i.e., \( C/\mu_b \)). We derive the optimal service policy for a base customer backwards in time, starting by deciding on the optimal service probability for a customer who arrives at the end of the advertisement period.

It follows from (4.11) and (4.13) that the value of serving a base customer call at the end of the advertisement period, i.e., at time \( T \), is independent of the service policy:

\[
V_b(T, q_b) = p_b + c_b + (1 - \theta_b) A.
\]

Comparing the cost and value of serving a base customer call at time \( T \), a base customer call at time \( T \) should be served if \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \), and should not be served, otherwise.

To determine the optimal service policy for base customers during the entire period, in addition to their terminal value, the firm should also consider the profit a base customer may generate during the advertisement period. For any fixed advertisement level, solving (4.14) suggests that by considering the lifetime value of a base customer, the firm should decide to serve or not to serve her call. Following the homogeneity assumption for base customers and infinite flexibility assumption for staffing, we expect that optimal service level for base customers to take the form of a bang-bang policy, and at each moment during the advertisement period the call center should serve all or none of the calls from base customers. This intuition is formalized in Proposition 4.2.

Proposition 4.2 also shows that the optimal policy transitions between these two extremes at most once. In other words, there is at most one time-threshold after which the service policy changes from serve all to serve none of the calls (or vice versa). We let

78
Proposition 4.2. Optimal service policy for base customers. Fix the advertisement level, then:

(a) If \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \), let

\[
T^\uparrow_b = \inf \{0 < t \leq T : C/\mu_b \leq V_b(t, \ 1)\},
\]

the optimal service policy for base customers is:

\[
q^*_b = \begin{cases} 
0 & \text{if } 0 < t < T^\uparrow_b \\
1 & \text{if } T^\uparrow_b \leq t \leq T 
\end{cases},
\]

(b) If \( C/\mu_b > p_b + c_b + (1 - \theta_b) A \), let

\[
T^\downarrow_b = \inf \{0 < t \leq T : C/\mu_b > V_b(t, \ 0)\},
\]

the optimal service policy for base customers is:

\[
q^*_b = \begin{cases} 
1 & \text{if } 0 < t < T^\downarrow_b \\
0 & \text{if } T^\downarrow_b \leq t \leq T 
\end{cases}.
\]

Part (a) of Proposition 4.2 describes the case where serving base customers at the end of the advertisement period is profitable. For this case, consider the stationary policy of serving all base customer calls, \( q_b = 1 \). If under this policy the value of serving a base customer decreases with time (i.e., (4.20) does not hold), we have

\[
C/\mu_b \leq V_b(T, \ 1) \leq V_b(0, \ 1),
\]

which means that the value of serving base customer calls is higher than the capacity cost during the entire period, thus the optimal service policy for base customers is \( q^*_b = 1 \) (Figure 4.2.(i)).
However, if (4.20) holds, the value of serving base customers, $V_b(t, 1)$, increases as we approach the end of the period. Therefore, either

$$C/\mu_b \leq V_b(0, 1) \leq V_b(T, 1),$$

which means it is optimal to serve all base customers during the entire period (Figure 4.2.(ii)), or there is time threshold $T_b^+$ such that

$$V_b(0, 1) < C/\mu_b = V_b\left(T_b^+, 1\right) \leq V_b(T, 1)$$

which means that it is profitable to serve all base customers at the end, but this is not the case earlier in the period (Figure 4.2.(iii)). In other words, since the profit generated during the advertisement period is relatively small (in comparison to the terminal value and the capacity cost), it is not worthwhile to serve and keep base customers who call early in the period, but it is worth to serve and keep them later in the period.

Figure 4.2: Value of serving a base customer call under optimal service policy.

Similarly, part (b) of Proposition 4.2 indicates that if it is not worth to serve a base
customer who calls at time $T$, either it is not worth to serve any base customer call, or it is profitable to only serve base customers who call early. The latter case corresponds to systems in which the profit generated during the advertisement period, from base customer calls or from call-independent revenue of base customers, is relatively large (in comparison to the terminal value and the capacity cost). Thus the firm does not want to lose a customer early in the period.

Proposition 4.2 indicates that the optimal service policy for base customers, $q^*_b$, is stationary before and after the corresponding time threshold, thus the advertisement period may be divided to two intervals with stationary service probability to base customers in each interval. An important implication of this result is that it simplifies the determination of the lifetime value of a base customer.

**Corollary 4.1.** Under the optimal service policy during a single period, the lifetime value of a base customer is monotone in time. Specifically,

(a) If $C/\mu_b \leq p_b + c_b + (1 - \theta_b) A$, then

\[
L(t, q^*_b) = \begin{cases}
\frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} + \left(\frac{C/\mu_b - p_b - c_b}{1 - \theta_b} - \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}\right) e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^t - t)} & 0 \leq t \leq T_b^t \\
\frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} + \left(A - \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b}\right) e^{-\gamma_b (T - t)} & T_b^t < t \leq T
\end{cases}
\]  

which is increasing in time if

\[
A \geq \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b},
\]

and decreasing otherwise.

(b) If $C/\mu_b > p_b + c_b + (1 - \theta_b) A$, then

\[
L(t, q^*_b) = \begin{cases}
\frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} + \left(\frac{C/\mu_b - p_b - c_b}{1 - \theta_b} - \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b}\right) e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^t - t)} & 0 \leq t \leq T_b^t \\
\frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} + \left(A - \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}\right) e^{-\gamma_b (T - t)} & T_b^t < t \leq T
\end{cases}
\]

which is increasing in time if

\[
A \geq \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)},
\]

and decreasing otherwise.
4.3.2.2 Optimal Service Policy for New Customers

We determine the value of serving a new customer, considering how she will be served if she joins the customer base. It follows from Corollary 4.1 and the definition of $V_n(t, q_b^*)$ in (4.12) (ignoring the period index), that under the optimal service policy for base customers, the value of serving a new customer call is changing monotonically with time. Therefore, it is possible that while new customers’ early calls receives good service, the firm prefers to serve no new customer call who arrives later. This change may be caused by relatively low termination value of base customers that also lowers the value of serving new customers closer to the end of the period. In contrast, if the terminal value is relatively large, it may be profitable to serve new customer calls later in the period, even if it is not the case for customers who arrive early. Since the value of serving a new customer call $V_n(t, q_b^*)\mu_n$ monotonically changes with time, it equals the capacity cost $C$ at most once. The following proposition prescribes the optimal service policy for new customer calls:

**Proposition 4.3.** Fix the advertisement level, then under the optimal service policy $q_b^*$, we have:

(a) If $C/\mu_n \leq p_n + c_n + \theta_n A$, let

$$T_n^+ = \inf \{0 < t \leq T : C/\mu_n \leq V_n(t, q_b^*)\},$$

then, the optimal service policy for new customers is:

$$q_n^* = \begin{cases} 
0 & \text{if } 0 < t < T_n^+ \\
1 & \text{if } T_n^+ \leq t \leq T 
\end{cases}.$$  

(b) If $C/\mu_n > p_n + c_n + \theta_n A$, let

$$T_n^- = \inf \{0 < t \leq T : C/\mu_n > V_n(t, q_b^*)\},$$

then, the optimal service policy for new customers is:

$$q_n^* = \begin{cases} 
1 & \text{if } 0 < t < T_n^- \\
0 & \text{if } T_n^- \leq t \leq T 
\end{cases}.$$
Proposition 4.3 indicates that the optimal service policy for new customer calls is also a bang-bang policy.

It follows from Propositions 4.2 and 4.3 that the optimal service policy is independent of the new customer arrival rate $\lambda_n(t)$ and the size of the customer base $x_b(t)$, and it only depends on the capacity cost and the value of serving a base/new customer call. This property plays an important role in optimizing the new customer arrival rate in the next section by determining the marginal profit contribution of a new or a base customer independent of the size of the customer base.

4.3.3 Jointly Optimal Advertisement and Service Policy

In §4.3.2, we determined the optimal service policy for any given advertisement level. Applying the optimal service policy, in this section we establish the evolution of the customer base over the advertisement period, assess the marginal profit of acquiring a new customer, and characterize the jointly optimal advertisement and service policy.

Customer base evolution over any time interval during the advertisement period depends on the arrival rate of new customers, the size of the customer base at the start of the interval, and the service policy during the interval. Following Propositions 4.2 and 4.3, the advertisement period $(0, T]$ may be divided into shorter time intervals, during each, service policies for new and base customers are stationary. Specifically, as service policies are changing at most once during the advertisement period, we have a finite number (at most 3) of time intervals.

Given (4.3), for any time interval $(t_j, t_{j+1}]$, given $x_b(t_j)$, $\lambda_n(t_j)$ and stationary service policy $q_n$ and $q_b$ during the interval, the size of the customer base for any $t \in (t_j, t_{j+1}]$ is:

$$x_b(t) = x_b(t_j) e^{-(\gamma_n + r_b(1-\theta_b)(1-q_b))(t-t_j)} + \lambda_n(t_j) q_n \theta_n \frac{e^{-\gamma_n(t-t_j)} - e^{-(\gamma_n + r_b(1-\theta_b)(1-q_b))(t-t_j)}}{\gamma_b + r_b (1-\theta_b) (1-q_b) - \gamma_n}$$

(4.31)

where the first term represents the number of base customers who are in the customer base at time $t_j$ and remain until time $t$, and the second term counts base customers who make their first call as new customers during $(t_j, t]$, join the customer base, and remain as base customers until time $t$.

We discussed in §4.3.2 that $q_n^*$ and $q_b^*$ are independent of the size of the customer base. Therefore, it follows from (4.31) that to determine $x_b(t)$ during any time interval
\((t_j, t_{j+1}]\), it is sufficient to know the arrival rate of new and base customers at the start of that time interval (i.e., time \(t_j\)). From (4.1), the new customers arrival rate at time \(t_j\) is \(\lambda_n(t_j) = \lambda_n(0) e^{-\gamma_n t_j}\). The size of the customer base at time \(t_j\) is a function of previous time intervals. In particular, \(x_b(t_j)\) equals to the size of the customer base at the end of the previous time interval \((t_{j-1}, t_j]\). Therefore, by applying (4.31) for time interval \((t_{j-1}, t_j]\), we can determine \(x_b(t_j)\) as a function of \(x_b(t_{j-1})\) and \(\lambda_n(0)\) (under service policies \(q_n^*\) and \(q_b^*\)).

By successively applying (4.31), we can recursively find the size of the customer base at any time \(t \in (0, T]\) as an additively separable function of the initial values \(x_b(0)\) and \(\lambda_n(0)\). Therefore, we can break the total profit into two parts: profit due to initial base customers, and profit due to customers acquired by advertisement during \((0, T]\). We rewrite the optimization problem (4.14) as an additively separable function of \(x_b(0)\) and \(\lambda_n(0)\):

**Lemma 4.1.** For any fix advertisement level, under the optimal service policy, there exist time independent functions \(f(A, q_n^*, q_b^*)\) and \(g(A, q_n^*, q_b^*)\), such that:

\[
P(\lambda_n(0), q_n^*, q_b^*) + Ax_b(T) = \lambda_n(0) f(A, q_n^*, q_b^*) + x_b(0) g(A, q_n^*, q_b^*) - S(\lambda_n(0)),
\]

where the functions \(f(A, q_n^*, q_b^*)\) and \(g(A, q_n^*, q_b^*)\) are independent of \(\lambda_n(0)\) and \(x_b(0)\).

In fact, \(\lambda_n(0) f(A, q_n^*, q_b^*)\) captures the expected profit generated by customers acquired by advertisement during \((0, T]\), which includes the profit of serving their first call, the profit they generate as base customers during \((0, T]\) if they join the customer base, and their terminal value if they stay in the customer base until the end of the advertisement period. The term \(x_b(0) g(A, q_n^*, q_b^*)\) captures the expected profit generated by base customers who are in the system at the start of advertisement period, which includes the profit they generate during \((0, T]\), and their terminal value if they remain in the customer base until time \(T\).

It follows from Lemma 4.1 that under the optimal service policy:

\[
\Pi(x_b(0)) = \max_{\lambda_n(0)} \left\{ P(\lambda_n(0), q_n^*, q_b^*; x_b(0)) + Ax_b(T) \right\}
\]

\[= \max_{\lambda_n(0)} \left\{ \lambda_n(0) f(A, q_n^*, q_b^*) - S(\lambda_n(0)) \right\} + x_b(0) g(A, q_n^*, q_b^*). \quad (4.32)\]
Although the value of $f(A, q_n^*, q_b^*)$ and $g(A, q_n^*, q_b^*)$ functions change under different parameter values, their interpretations remain unchanged: the marginal profit contribution of acquisition efforts and initial customer base, respectively. Therefore, to optimize the advertisement spendings, we just need to compare the marginal acquisition cost with $f(A, q_n^*, q_b^*)$:

**Proposition 4.4.** Given the optimal service policy, the optimal advertisement level is

$$\lambda_n^*(0) = \arg \{ \lambda_n(0) \geq 0 : S'(\lambda_n(0)) = f((A, q_n^*, q_b^*)) \}, \quad (4.33)$$

where $f(A, q_n^*, q_b^*)$ is independent of system initial state $x_b(0)$.

Note that since $S(\lambda_n(0))$ is strictly increasing and strictly convex in $\lambda_n(0)$, (4.33) has a unique solution if $f((A, q_n^*, q_b^*)) \geq 0$, otherwise $\lambda_n^*(0) = 0$, and it is not profitable to advertise.

### 4.4 Multi-period Problem

In this section we convert the multi-period problem to a series of single period problems, and show that how the termination value of each period depends on the optimal service policies of new and base customers in future periods.

The optimal advertisement and service policies over the single period problem illustrate two main characteristics: (i) By propositions 4.2 and 4.3, the optimal new and base customers’ service policies are independent of the arrival rate of new and base customers, and (ii) by Lemma 4.1 the contribution of initial base customers to the total profit can be decoupled from the contribution of the customers who are acquired during the advertisement period.

Consider the last period of a multi-period problem (i.e., period $N$) as a single period problem with initial state value $x_b^N(0)$. It follows from these characteristics that optimal decisions in period $N$ are independent of the state of the system at the start of that period. In other words, the last advertisement period may be optimized independent of earlier periods.

Indicating the last period’s optimal decisions by $\lambda_n^{*N}(0), q_n^{*N},$ and $q_b^{*N}$, it follows from
(4.7) and (4.32) that:

$$\Pi^N (x_b^N (0)) = \lambda_n^N (0) f(A, q_n^N, q_b^N; x_b^N (0)) - S (\lambda_n^N (0)) + x_b^N (0) g(A, q_n^N, q_b^N).$$

(4.34)

For any given $x_b^{N-1} (0)$ at the start of period $N - 1$, (4.7) yields:

$$\Pi^{N-1} (x_b^{N-1} (0)) = \max_{\lambda_n^{N-1}(0), q_n^{N-1}, q_b^{N-1}} \{ P (\lambda_n^{N-1} (0), q_n^{N-1}, q_b^{N-1}; x_b^{N-1} (0)) + \Pi^N (x_b^{N-1} (T)) \}.$$

(4.35)

By substituting (4.34) in (4.35), and using $x_b^N (0) = x_b^{N-1} (T)$ we get:

$$\Pi^{N-1} (x_b^{N-1} (0)) = \max_{\lambda_n^{N-1}(0), q_n^{N-1}, q_b^{N-1}} \{ P (\lambda_n^{N-1} (0), q_n^{N-1}, q_b^{N-1}; x_b^{N-1} (0)) + x_b^{N-1} (T) g(A, q_n^N, q_b^N) \}$$

$$+ \lambda_n^N (0) f(A, q_n^N, q_b^N; x_b^N (0)) - S (\lambda_n^N (0)),$$

(4.36)

for which the optimization part is a single period problem with initial state $x_b^{N-1} (0)$ and termination value $g(A, q_n^N, q_b^N)$. Therefore, the optimal controls over this period are independent of the history of decisions in previous periods (e.g., they are independent of the initial state $x_b^{N-1} (0)$). Similarly, the generated profit due to new customers acquired in the last period does not affect optimal decisions in the current period, either. In fact, optimal decisions only depend on the parameters of the current period, and on the future profit the firm can generate from the base customers who remain until the end of this period (i.e. the terminal value of base customers at the end of this period).

By successively applying this argument, each period may be considered as a single period problem with a termination value which only depends on how ending base customers will be treated during next periods Therefore, by starting from the last advertisement period, we may recursively determine the optimal decisions over the entire time horizon.

### 4.4.1 The Infinite Horizon Problem

As the limiting case of the multi-period problem, consider the infinite horizon problem with stationary model parameter values. As discussed in §4.4 each period may be considered as a single period problem with a termination value which depends on future period decisions. Considering the service-independent attrition rate, none of the base customers at the end of a period remain in the customer base for infinite time even if she receives
prefect service. Therefore, the termination value for each period is also finite as it equals
the profit the firm can generate over the infinite future due to base customers that exist
at the end of this period. Although, the sum of the profit over infinite horizon goes to
infinity,

Proposition 4.5. By applying the optimal decisions over an infinite horizon, the terminal
value of a base customer during different advertising periods converges. In particular,
(a) if \( \frac{C}{\mu_b} \leq p_b + c_b + (1 - \theta_b) \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \), the terminal value of a base customer
converges to \( \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \).
(b) Otherwise, this value converges to \( \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \).

The main implication of Proposition 4.5 is that under optimal service policy, lifetime
value of base customers converges and is independent of the time of the call. In particular,
if

\[
\frac{C}{\mu_b} \leq p_b + c_b + (1 - \theta_b) \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)},
\]

(4.37)
it follows from part (a) of Proposition (4.5) and (4.18), that the value of serving a base
customer current and all future calls converges to

\[
V_{b}^i (t, 1) = p_b + c_b + (1 - \theta_b) \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b},
\]

(4.38)
independent of the time of the call. Therefore, by Proposition (4.2), all calls from base
customers over all advertisement periods should be served, i.e. \( q_{b}^{*i} = 1 \) for any period
\( i \geq 1 \).

With similar discussion, for part (b) in which (4.37) does not hold, the value of serving
only the current call of a base customer converges to

\[
V_{b}^i (t, 0) = p_b + c_b + (1 - \theta_b) \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}.
\]

(4.39)

Thus, it follows from (4.37) that \( C/\mu_b > V_{b}^i (t, 0) \) for any time instant \( t \) during any period
\( i \), i.e. \( q_{b}^{*i} = 0 \) for any period \( i \geq 1 \).

The value of serving a new customer call, \( V_{n}^i (t, q_{b}^{*}) \) also converges and becomes inde-
dependent of the time instant \( t \) during any period \( i \). Following Propositions 4.3 and 4.4, if
\( C/\mu_n \leq V_{n}^i (t, q_{b}^{*}) \), it is profitable for the firm to advertise to acquire new customers and
all new customer calls should be served, i.e. $q_i^{ri} = 1$. Otherwise, it is not profitable for the firm to advertise.

Note that under the infinite horizon assumption, the fluid model prescribes identical policies to the optimal ones in the stationary problem discussed in Chapter 3.

4.5 Fluid Model Validation: Simulation Results

In this section, we study the accuracy of the pointwise stationary fluid model approximation by comparing its performance with simulation results for the stochastic system described in §4.2 over a single period advertisement. We assume exponentially distributed service times. §4.5.1 specifies the parameter values for this simulation study. In §4.5.2 we report the accuracy of the pointwise stationary fluid model in approximating key performance measures under different levels of promotion and service policies.

4.5.1 Parameter Values

Table 4.2 summarizes the parameter values for the simulation study. These values may be representative for a call center of a business with significant call-independent revenue, such as a mobile phone service provider company.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertisement period length (in days)</td>
<td>$T$</td>
</tr>
<tr>
<td>Service rate of new and base customers (per day)</td>
<td>$\mu_n, \mu_b$</td>
</tr>
<tr>
<td>Call abandonment rate of new and base customers (per day)</td>
<td>$\tau_n, \tau_b$</td>
</tr>
<tr>
<td>Call arrival rate per base customer (per day)</td>
<td>$r_b$</td>
</tr>
<tr>
<td>P(new customer joins customer base after service)</td>
<td>$\theta_n$</td>
</tr>
<tr>
<td>P(base customer remains in customer base after abandoning)</td>
<td>$\theta_b$</td>
</tr>
<tr>
<td>Attrition rate per base customer (per day)</td>
<td>$\gamma_b$</td>
</tr>
<tr>
<td>Decay rate of new customers arrival rate (per day)</td>
<td>$\gamma_n$</td>
</tr>
<tr>
<td>Profit rate per base customer (per day)</td>
<td>$R$</td>
</tr>
<tr>
<td>Profit per served call</td>
<td>$p_n, p_b$</td>
</tr>
<tr>
<td>Cost per abandoned call</td>
<td>$c_n, c_b$</td>
</tr>
<tr>
<td>Cost of serving a new and a base customer call</td>
<td>$C/\mu_n, C/\mu_b$</td>
</tr>
<tr>
<td>Advertising cost function parameters (power model)</td>
<td>$\alpha, \beta$</td>
</tr>
</tbody>
</table>

Table 4.2: Parameter values for simulation.

We consider each time unit as one day. We assume a 24x7 operation. We consider quarterly advertisement and set the length of the period to 90 days. We assume $\gamma_n = 0.02,$
i.e., at the end of 90 days, new customer arrival rate decreases to 15% of the original arrival rate. Values of the rest of the model parameters are given in Table 4.2 and are identical to the ones in Chapter 3, thus their justifications are omitted.

4.5.2 Accuracy of Steady-State Performance Measures under Fixed Policies

In this section, we report the accuracy of the fluid model in approximating four key performance measures at any time instant $t$: the size of the customer base $x_b(t)$, the number of servers $k(t)$, and the probability of receiving service by new and base customers $q_n(t)$ and $q_b(t)$, respectively. We also discuss the accuracy of the fluid model in approximating the total profit for specific policies.

We simulated the stochastic system under two sets of service policies: (i) stationary policies, where both customer types are served during the entire period, and (ii) bang-bang policies where the service probability of one customer type (new or base) drops to zero in the middle of the advertisement period, while the other type calls are served during the entire period. For the first policy, we considered three levels of advertising spendings to generate pulses of size 2500, 10000, and 25000 in new customers arrival rate per day. For the non-stationary policies, we considered a pulse of size $\lambda_n(0) = 10,000$. These rates diminish over time during the 90-day advertising period. Following (4.2), we adjusted staffing at each new or base customer arrival, service completion and abandonment.

In every case, results are averaged out over 1000 iterations, and in each iteration 10 days warm-up period is considered before the start of the advertisement period.

(i) Stationary service policies:

Figure 4.3 compares the evolution of the size of the customer base, and the number of servers for the simulation (solid line) vs. the fluid (dotted line) model, under the stationary service policy $q_n = q_b = 1$. When the arrival rate of new customers increases (moving left to right in Figure 4.3), the size of the call center grows to keep the service level fixed. Figure 4.3 shows that the approximation errors decrease in the size of the system, as expected for a fluid model (See Table 4.3 for comparing the maximum approximation error for each case).
Figure 4.3: Number of base customers $x_b(t)$ and number of servers $K(t)$, given service policy $q_n = q_b = 1$.

Next, we measure the realized service probability of new and base customers in the simulation model, given the staffing recommended by the fluid approximation, and compare it with the predicted service policy in the fluid model. Figure 4.4 shows $q_n(t)$ and $q_b(t)$ (averaged over daily intervals) for a simulation model under the stationary service policy $q_n = q_b = 1$. The fluid model error in approximating the service probability (or equivalently in approximating the abandonment rate) is the difference between the $q_n(t)$ and $q_b(t)$ graphs in Figure 4.4 and the $q_n(t) = q_b(t) = 1$ line. Again, the fluid model error decreases with the size of the system, from a maximum error of 11.3% ($\lambda_n(0) = 2,500$) to 2.9% ($\lambda_n(0) = 25,000$).

Note that the relatively large fluctuation in the staffing level over the advertisement period, as depicted in Figure 4.3, is due to assuming a single advertisement. Whereas, in practice, usually there are multiple advertisement campaigns with some overlap in their
Table 4.3: Fluid model maximum approximation error, given service policy $q_n = q_b = 1$.

<table>
<thead>
<tr>
<th>$\lambda_n(0)$</th>
<th>Max error of $x_b(t)$</th>
<th>Max error of $K(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,500</td>
<td>2.3 %</td>
<td>5.9 %</td>
</tr>
<tr>
<td>10,000</td>
<td>1.0 %</td>
<td>1.8 %</td>
</tr>
<tr>
<td>25,000</td>
<td>0.6 %</td>
<td>0.9 %</td>
</tr>
</tbody>
</table>

(ii) Non-stationary service policies

Figure 4.5 compares the size of the customer base and the number of servers for the simulation (solid line) vs. the fluid (dotted line) model, for the case where the service probability of new customers (left) or base customers (right) drops to zero at $t = 45$, i.e., at the middle of the advertisement period, while the other type calls are served during the entire period (See Table 4.4 for the maximum approximation error for each case).

Figure 4.6 shows the realized service probabilities (averaged over daily intervals) by new and base customers. Note that in both cases, before time $t = 45$ the policy is to

Table 4.4: Fluid model maximum approximation error, given $\lambda_n(0) = 10,000$

<table>
<thead>
<tr>
<th>$q_n$</th>
<th>$q_b$</th>
<th>Max error of $x_b(t)$</th>
<th>Max error of $k(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 if $t \leq 45$; 0 otherwise</td>
<td>1.1 %</td>
<td>5.4 %</td>
</tr>
<tr>
<td>1 if $t \leq 45$; 0 otherwise</td>
<td>1</td>
<td>0.9 %</td>
<td>2.1 %</td>
</tr>
</tbody>
</table>
serve all new and base customers, and thus the approximation errors are similar to the ones in the case of the stationary service policy with $\lambda_n(0) = 10,000$. However, after time $t = 45$, one customer type is not served, thus the size of the system drops, and the fluid model accuracy in approximating the service probabilities decrease. This effect is more significant when the policy is not to serve any base customer after $t = 45$, because the size of the system is much smaller in this case (see Figure 4.5).

The firm’s total profit over the advertisement period is a function of the arrival rates of new and base customers, their realized service probability, and the staffing cost. The above figures show that the fluid model accuracy in approximating these factors is reason-
able, especially if the size of the call center is large. In particular, for the cases reported in tables 4.3 and 4.4, the error of the fluid model in approximating the total profit of the firm remains smaller than 1%.

4.6 Concluding Remarks

In this paper we develop a stochastic queueing model that captures the characteristics of a call center with both new and base customers under time-varying demand response to different combinations of promotions and service policies. We approximate this stochastic model with a dynamic pointwise stationary fluid model and study the problem of maximizing the total profit over the advertisement horizon. The key feature of this model is that, similar to Chapter 3, the customer base depends on customers’ past service experiences. We establish optimal controls base on the pointwise stationary fluid model and show the robustness of our approximations by comparing them with simulation results.

The main insight from our analysis is that, under flexible staffing assumption, the manager can set the service levels for new and base customers independent of the promotion level and the size of the customer base. Since changes in the arrival rate is negligible in comparison to the service rate, the flexible staffing is reasonable assumption and it is achievable by changing the capacity level every hour which is common in large call
centers.

Our model also suggests that the optimal promotion policy at the start of each period depends on the service policy over that period and the following periods, and it is independent of the current size of the customer base. This result holds under two assumptions: (i) the market size is large, i.e., there are virtually infinite potential customers to be targeted by advertisement, and (ii) the effect of word-of-mouth is negligible. Both assumptions are reasonable for call centers of companies such as a major credit card issuer or a nationwide phone service provider.

This work can be extended in multiple directions. Similar to Chapter 3, we can consider heterogeneous base customers with different service dependent and service independent attributes. Another viable research direction is to study systems with limited number of servers. In these systems, the manager still benefits from flexible staffing, but she should consider promotion policy and new and base customer arrival rates in determining the optimal service policy. One can also consider systems where word-of-mouth significantly affects demand response to advertisement, therefore the optimal promotion policy depends on the current size of the customer base and their experiences.
4.7 Appendix: Proofs

Proof of Proposition 4.1

Under fixed service probability for base customers, the survival probability (4.8) becomes:

\[ U(t, s, q_b) = e^{-\left(\gamma_b + r_b (1-q_b)(1-\theta_b)\right)(s-t)}, \tag{4.40} \]

i.e., the lifetime of a base customer follows an exponential distribution with a rate \( \gamma_b + r_b (1-\theta_b) (1-q_b) \).

Substituting \( q_b = q_b \) and (4.40) in (4.9), we determine the lifetime value of a base customer who is in the customer base at time \( t \):

\[
L(t, q_b) = \int_t^T U(t, s, q_b) \left( R + r_b \left( (p_b - C/\mu_b) q_b - c_b (1-q_b) \right) \right) ds + U(t, T, q_b) A
+ A e^{-\left(\gamma_b + r_b (1-q_b)(1-\theta_b)\right)(T-t)},
\]

which is identical to (4.15). Showing that \( L(t, q_b) \) is decreasing if (4.16) holds and is increasing otherwise is straightforward.

Proposition 4.1 follows from (4.41).

Proof of Proposition 4.2

Given any exogenous new customer arrival rate, \( \lambda_n(t) \), the optimization problem in (4.14) may be considered as an optimal control problem with control parameters \( 0 \leq q_n(t), q_b(t) \leq 1 \), corresponding dynamics

\[
\begin{aligned}
\frac{dx_b(t)}{dt} &= \lambda_n(t) q_n(t) \theta_n - x_b(t) \left( \gamma_b + r_b (1-q_b(t)) \right) (1-\theta_b) \\
x_b(0) &= x^0_b
\end{aligned}
\tag{4.42}
\]

and with the payoff function

\[
\Pi = \int_0^T \left( \lambda_n(t) \left( (p_n - C/\mu_n) q_n(t) - c_n (1-q_n(t)) \right) + x_b(t) \left( R + r_b \left( (p_b - C/\mu_b) q_b(t) - c_b (1-q_b(t)) \right) \right) \right) dt \\
+ A x_b(T) - S(\lambda_n(0)),
\tag{4.43}
\]

where \( x_b(t) \) solves (4.42) for controls \( q_n(t) \) and \( q_b(t) \). The integral term captures the running payoff and \( A x_b(T) \) is the terminal payoff for any given end-time \( T \).

We use the Pontryagin Maximum Principle (Sethi and Thompson 2000, pp 33-34) to
solve this optimal control problem. First, we construct the Hamiltonian as:

\[ H(x_b(t), \omega(t), (q_n(t), q_b(t))) := \omega(t) (\lambda_n (t) q_n(t) \theta_n - x_b (t) (\gamma_b + r_b (1 - q_b(t)) (1 - \theta_b))) + \lambda_n (t) ((p_n - C/\mu_n) q_n (t) - c_n (1 - q_n(t))) + x_b (t) (R + r_b ((p_b - C/\mu_b) q_b (t) - c_b (1 - q_b (t)))) , \]

which, by factoring controls \( q_n(t) \) and \( q_b(t) \), may be rewritten as

\[ H(x_b(t), \omega(t), (q_n(t), q_b(t))) := q_n(t) \lambda_n (t) (\omega(t) \theta_n + p_n - C/\mu_n + c_n) + q_b(t)x_b (t) r_b(\omega(t) (1 - \theta_b) + p_b - C/\mu_b + c_b) + x_b (t) (R - r_b c_b - \omega(t) (\gamma_b + r_b (1 - \theta_b))) - \lambda_n (t) \theta_n \theta(t) \]

Now, by the maximality principle we have:

\[ H(x_b^*(t), \omega^*(t), (q_n^*(t), q_b^*(t))) = \max_{0 \leq q_n(t), q_b(t) \leq 1} \{ H(x_b^*(t), \omega^*(t), (q_n(t), q_b(t))) \} . \]  

(4.45)

According to (4.45), at each time instant \( t \) the control value \( (q_n(t), q_b(t)) \) must be selected to maximize \( H(x_b^*(t), \omega^*(t), (q_n(t), q_b(t))) \). From (4.44), for any \( x_b(t) \) and \( \omega(t) \), we may optimize the controls \( q_n(t) \) and \( q_b(t) \) independent of each other. In particular, since \( \lambda_n (t) \geq 0 \), it follows from the first term of (4.44) that:

\[ q_n^*(t) = \begin{cases} 1 & \text{if } \omega^*(t) \geq (C/\mu_n - p_n - c_n) / \theta_n, \\ 0 & \text{otherwise} \end{cases} \]  

(4.46)

Similarly, since \( x_b(t) \geq 0 \), it follows from the second term of (4.44) that:

\[ q_b^*(t) = \begin{cases} 1 & \text{if } \omega^*(t) \geq (C/\mu_b - p_b - c_b) / (1 - \theta_b), \\ 0 & \text{otherwise} \end{cases} \]  

(4.47)

Therefore, the optimal service policies of new and base customers are both bang-bang policies.

The remaining step to determine the optimal control \( (q_n^*(t), q_b^*(t)) \) is to find the values of \( \omega^*(t) \). By Pontryagin Maximum Principle, we write the adjoint equation:

\[ \dot{\omega}^*(t) = - \frac{\partial}{\partial x_b} H = \omega^*(t) (\gamma_b + r_b (1 - q_b^*(t)) (1 - \theta_b)) - (R + r_b ((p_b - C/\mu_b) q_b^*(t) - c_b (1 - q_b^*(t)))) , \]  

(4.48)
with the terminal condition:

\[ \omega^* (T) = \frac{\partial}{\partial x_b} A x_b^* (T) = A. \] (4.49)

From (4.48) we see that \( \omega^* (t) \) is independent of \( q_n^* (t) \). Thus, we next solve the system of equations (4.47) and (4.48) to find \( \omega^* (t) \) and \( q_b^* (t) \) to complete the proof of Proposition 4.2. (Later, in the proof of Proposition 4.3, we use the value of \( \omega^* (t) \) to find \( q_n^* (t) \) from (4.46).)

Consider the following cases:

a) If \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \):

We first focus on the case where \( C/\mu_b < p_b + c_b + (1 - \theta_b) A \), and discuss the case \( C/\mu_b = p_b + c_b + (1 - \theta_b) A \) at the end of this part.

We derive the optimal service policy for a base customer backwards in time. Starting from the end of the advertisement period, it follows from (4.49) and \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \) that:

\[ \frac{(C/\mu_b - p_b - c_b)}{(1 - \theta_b)} < A = \omega^* (T). \]

Therefore by (4.47), the optimal service probability at time \( T \) is \( q_b^* (T) = 1 \).

We deduce from the continuity of \( \omega^* (t) \) that there is a neighborhood of \( T \), say \( [T_b^\uparrow, T] \), during which \( \omega^* (t) \geq \frac{(C/\mu_b - p_b - c_b)}{(1 - \theta_b)} \). It follows from (4.47) that for \( t \in [T_b^\uparrow, T] \), we have \( q_b^* (t) = 1 \). Substituting \( q_b^* (t) \) in (4.48) gives:

\[ \dot{\omega}^* (t) = \omega^* (t) \gamma_b - R + r_b p_b + Cr_b/\mu_b \] (4.50)

Solving 4.50 with boundary condition \( \omega^* (T) = A \), yields that for \( t \in [T_b^\uparrow, T] \) we have \( \omega^* (t) = \omega_1 (t) \) where

\[ \omega_1 (t) := \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} + \left( A - \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \right) e^{-\gamma_b (T-t)}. \] (4.51)

Since \( \omega_1 (t) \) is monotone in time, going backwards in time, \( q_b^* (t) = 1 \) and \( \omega^* (t) = \omega_1 (t) \) as long as \( \omega_1 (t) \geq \frac{(C/\mu_b - p_b - c_b)}{(1 - \theta_b)} \). Therefore

\[ T_b^\uparrow = \inf \{ 0 < t \leq T : \omega_1 (t) \geq \frac{(C/\mu_b - p_b - c_b)}{(1 - \theta_b)} \}. \] (4.52)

Note that if at time 0, i.e., at the start of the period, \( \omega_1 (0) \geq \frac{(C/\mu_b - p_b - c_b)}{(1 - \theta_b)} \), it follows from (4.52) and the monotonicity of \( \omega_1 (t) \) that \( T_b^\uparrow = 0 \) and the firm should always serve all base customer calls. Otherwise, there exists \( T_b^\uparrow > 0 \). The necessary
condition for $T_b^+ > 0$ is that $\omega_1(t)$ increases with time, which from (4.51) yields

$$\frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b} \leq \omega_1(0) \leq \frac{C/\mu_b - p_b - c_b}{1 - \theta_b} < \omega_1(T) = A,$$

(4.53)

Let $T_b^{\uparrow-}$ denote the time just before $T_b^+$. We have $q_b^*(T_b^{\uparrow-}) = 0$, since otherwise $\omega^*(T_b^{\uparrow-}) = \omega_1(T_b^{\uparrow-})$ which contradicts the definition of $T_b^+$ in (4.52). For the neighborhood of $T_b^+$, substituting $q_b^*(t) = 0$ in (4.48) gives:

$$\dot{\omega}^*(t) = \omega^*(t) \left( \gamma_b + r_b (1 - \theta_b) \right) - (R - r_b c_b),$$

(4.54)

Solving (4.54) with the boundary condition $\omega^*(T_b^+) = (C/\mu_b - p_b - c_b) / (1 - \theta_b)$, yields that in the neighborhood of $T_b^+$, $\omega^*(t) = \omega_2(t)$ where

$$\omega_2(t) = \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} + \left( \frac{C/\mu_b - p_b - c_b}{1 - \theta_b} - \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \right) e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^+ - t)}.$$

(4.55)

It follows from (4.53) that

$$\frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} < \frac{C/\mu_b - p_b - c_b}{1 - \theta_b},$$

(4.56)

which can be written as

$$\frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} < \frac{C/\mu_b - p_b - c_b}{1 - \theta_b}.$$

(4.57)

It follows from (4.55) and (4.57) that $\omega_2(t)$ is monotone increasing in time. Therefore, going backwards in time, $\omega_2(t) < (C/\mu_b - p_b - c_b) / (1 - \theta_b)$ for any $t \in \left(0, T_b^+\right)$, which implies that $T_b^+$ is unique, and for $t \in \left(0, T_b^+\right)$ we have $\omega^*(t) = \omega_2(t)$ and $q_b^*(t) = 0$.

The proof of Part (a) now follows from (4.52), the definition of $V_b(t, q_b)$ in (4.11), and $L(t, q_b^*) = \omega^*(t)$ which is proved as Lemma 4.2 below.

Finally, in the case where $(C/\mu_b - p_b - c_b) / (1 - \theta_b) = A$, the above discussion still holds. Specifically if (4.56) holds, then $T_b^+ = 0$, and otherwise, $T_b^+ = T$.

b) If $C/\mu_b > p_b + c_b + (1 - \theta_b) A$:

Starting from the end of the advertisement period ($\omega^*(T) = A < (C/\mu_b - p_b - c_b) / (1 - \theta_b)$, thus $q_b^*(T) = 0$) and following similar steps as in the proof of part a, we may solve the
system of equations (4.47) and (4.48) to obtain:

\[
\omega^*(t) = \begin{cases} 
\omega_3(t) = \frac{R + r_b p_b - r_b c_b}{\gamma_b} + \left(\frac{C}{\mu_b - p_b - c_b} - \frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b}\right) e^{-\gamma_b (1 - \theta_b)}(T_b^+ - t) & 0 < t < T_b^+ \\
\omega_4(t) = \frac{R - r_b C}{\gamma_b + r_b (1 - \theta_b)} + \left( A - \frac{R - r_b C}{\gamma_b + r_b (1 - \theta_b)} \right) e^{-\gamma_b (T - t)} & T_b^+ \leq t \leq T
\end{cases}
\]

and

\[
q_b^*(t) = \begin{cases} 
0 & 0 < t < T_b^+ \\
1 & T_b^+ \leq t \leq T
\end{cases}
\]

where

\[
T_b^+ = \inf \{ 0 < t \leq T : \omega_3(t) < (C/\mu_b - p_b - c_b) / (1 - \theta_b) \}.
\]

The proof of Part (b) now follows from (4.60), the definition of \( V_b(t, q_b) \) in (4.11), and \( L(t, q_b^*) = \omega^*(t) \) which is proved in Lemma 4.2.

**Lemma 4.2.** Under the optimal service policy for base customers, \( L(t, q_b^*) = \omega^*(t) \).

**Proof of Lemma 4.2**

a) If \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \):

For \( t \leq T_b^+ \), we have

\[
L(t, q_b^*) = \int_t^{T_b^+} \left( R - r_b c_b \right) e^{-(\gamma_b + r_b (1 - \theta_b))(s-t)} ds 
+ e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^+ - t)} \left( \int_{T_b^+}^T \left( R + r_b p_b - r_b C/\mu_b \right) e^{-\gamma_b (s-T_b^+)} ds + A e^{-\gamma_b (T-T_b^+)} \right)
= \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \left( 1 - e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^+ - t)} \right)
+ \left( \frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b} \left( 1 - e^{-\gamma_b (T-T_b^+)} \right) + A e^{-\gamma_b (T-T_b^+)} \right) e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^+ - t)}
= \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} + \left( \frac{C/\mu_b - p_b - c_b}{1 - \theta_b} - \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \right) e^{-(\gamma_b + r_b (1 - \theta_b))(T_b^+ - t)}
\]

(4.61)

where the last inequality follows from the definition of \( T_b^+ \):

\[
\omega^*(T_b^+) = \frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b} \left( 1 - e^{-\gamma_b (T-T_b^+)} \right) + A e^{-\gamma_b (T-T_b^+)} = \frac{C/\mu_b - p_b - c_b}{1 - \theta_b}.
\]
Similarly, for \( t > T_b^+ \)

\[
L(t, q_b^n) = \int_t^T (R + r_b p_b - r_b C/\mu_b) e^{-\gamma_b(s-t)} ds + A e^{-\gamma_b(T-t)}
\]

\[
= \frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b} + \left( A - \frac{R + r_b p_b - r_b C/\mu_b}{\gamma_b} \right) e^{-\gamma_b(T-t)}
\]

(4.62)

The proof follows by comparing (4.62) and (4.61) with (4.51) and (4.55), respectively.

b) If \( C/\mu_b > p_b + c_b + (1 - \theta_b) A \),

The proof is similar to this of Part (a). Thus it is omitted.

\[\text{Proof of Corollary 4.1}\]

From Lemma 4.2, we have \( L(t, q_b^n) = \omega^*(t) \). Thus the remaining step is to show the monotonicity of \( \omega^*(t) \).

a) If \( C/\mu_b \leq p_b + c_b + (1 - \theta_b) A \),

First consider the case where \( A \geq (R + r_b (p_b - C/\mu_b))/\gamma_b \), it follows from (4.51) that \( \omega_1(t) \) is increasing in time for \( [T_b^+, T] \), and we showed that if \( T_b^+ > 0 \), \( \omega_2(t) \) is increasing in time for \( (0, T_b^+) \). Thus, \( \omega^*(t) \) is increasing in time. Otherwise, if \( A < (R + r_b (p_b - C/\mu_b))/\gamma_b \), \( \omega_1(t) \) is decreasing in time for \( (0, T] \), i.e., \( T_b^+ = 0 \), and \( \omega^*(t) \) is decreasing in time.

b) If \( C/\mu_b > p_b + c_b + (1 - \theta_b) A \),

The proof is similar to this of Part (a). Thus it is omitted.

\[\text{Proof of Proposition 4.3}\]

It follows from (4.46) that to find \( q_n^*(t) \), we need to compare \( \omega^*(t) \) (obtained in Proposition 4.2) with \((C/\mu_n - p_n - c_n)/\theta_n\):

a) If \( C/\mu_n \leq p_n + c_n + \theta_n A \),

Since \( \omega^*(T) = A \geq (C/\mu_n - p_n - c_n)/\theta_n \), from (4.46) we have \( q_n^*(T) = 1 \). Next, if \( \omega^*(0) \geq (C/\mu_n - p_n - c_n)/\theta_n \), by monotonicity of \( \omega^*(t) \), all new customers should be served, i.e., \( q_n^*(t) = 1 \) for \( t \in (0, T] \). Otherwise, by monotonicity of \( \omega^*(t) \), there is a unique time \( T_n^+ > 0 \) such that \( \omega^*(T_n^+) = (C/\mu_n - p_n - c_n)/\theta_n \). From (4.46) we have \( q_n^*(t) = 0 \) for \( t \in (0, T_n^+) \) and \( q_n^*(t) = 1 \) for \( t \in [T_n^+, T] \).

The proof of Part (a) follows from the definition of \( V_n(t, q_b) \) in (4.12) and \( L(t, q_b^n) = \omega^*(t) \).

b) If \( C/\mu_n > p_n + c_n + \theta_n A \),

The proof is similar to this of Part (a). Thus it is omitted.

\[\text{Proof of Lemma 4.1}\]

Following Propositions 4.2 and 4.3, advertisement time period \((0, T]\) may be divided into 3 intervals (with length 0 or longer), during each, the service probabilities of new and base customers are stationary. Let \((t_1, t_2], (t_2, t_3], \) and \((t_3, t_4]\) represent these time
intervals, where \( t_1 = 0, t_4 = T \), and either \( t_2 \in \{ T_n^+, T_n^l \} \) and \( t_3 \in \{ T_b^+, T_b^l \} \), or \( t_2 \in \{ T_b^+, T_b^l \} \) and \( t_3 \in \{ T_n^+, T_n^l \} \), based on the optimal service policy for new and base customers.

We determine the functions \( f(A, q_{n}^*, q_{b}^*) \) and \( g(A, q_{n}^*, q_{b}^*) \) for a case where \( 0 \leq T_n^i \leq T_b^l \leq T \). Other cases can be determined similarly. Thus, they are omitted.

We serve both new and base customers at the first interval: \( q_n(t) = q_b(t) = 1 \) for \( t \in (0, T_n^i) \). It follows from (4.31):

\[
x_b(t) = x_b(0) e^{-\gamma_b t} + \lambda_n(0) \theta_n e^{-\gamma_n t} - e^{-\gamma_n t} \over \gamma_b - \gamma_n.
\]  

(4.63)

During the second interval, we serve none of the new but all of the base customers: \( q_n(t) = 0 \) and \( q_b(t) = 1 \) for \( t \in \left[ T_n^+, T_b^+ \right) \). Deriving \( x_b(T_n^i) \) from (4.63), it follows from (4.31) that:

\[
x_b(t) = \left( x_b(0) e^{-\gamma_T^i_n} + \lambda_n(0) \theta_n e^{-\gamma_n T_n^i - e^{-\gamma_n T_n^i} \over \gamma_b - \gamma_n} \right) e^{-\gamma_b (t - T_n^i)}
\]

\[
= x_b(0) e^{-\gamma_b t} + \lambda_n(0) \theta_n e^{-\gamma_n T_n^i} - e^{-\gamma_n T_n^i} \over \gamma_b - \gamma_n} e^{-\gamma_b (t - T_n^i)}.
\]  

(4.64)

Finally, during the third interval, we do not serve any customer: \( q_n(t) = q_b(t) = 0 \) for \( t \in \left[ T_b^l, T \right) \). Finding \( x_b(T_b^l) \) from (4.64), we have:

\[
x_b(t) = \left( x_b(0) e^{-\gamma_T^i_b} + \lambda_n(0) \theta_n e^{-\gamma_n T_b^i - e^{-\gamma_n T_b^i} \over \gamma_b - \gamma_n} e^{-\gamma_b (t - T_b^i)} \right) e^{-\gamma_b + \gamma_b (1 - \theta_b)} (t - T_b^i).
\]  

(4.65)

We expand the profit function as

\[
\Pi = \int_0^{T_n^+} (\lambda_n(t) (p_n - C/\mu_n) + x_b(t) (R + r_b p_b - r_b C/\mu_b)) dt + \int_{T_n^+}^{T_b^+} (-\lambda_n(t) c_n + x_b(t) (R + r_b p_b - r_b C/\mu_b)) dt + \int_{T_b^+}^{T} (-\lambda_n(t) c_n + x_b(t) (R - r_b c_b)) dt + A x(T) - S(\lambda_n(0)),
\]  

(4.66)

By substituting \( x_b(t) \) from (4.63), (4.64), and (4.65) in (4.66), we can determine
\[ f(A, q^*_n, q^*_b) = \int_0^{T^*_n} \left( e^{-\gamma t_n} (p_n - C/\mu_n) + \theta_n e^{-\gamma t_n} - e^{-\gamma t_n} \right) \left( R + r_b p_b - r_b C/\mu_b \right) dt \]
\[ + \int_{T^*_n}^{T^*_i} \left( e^{-\gamma t_n} (-c_n) + \theta_n e^{-\gamma t_n} - e^{-\gamma t_n} \right) e^{-(\gamma + r_n (1 - \theta_n)) (t - T^*_n)} \left( R - r_b c_b \right) dt \]
\[ + A \theta_n e^{-\gamma t_n} - e^{-\gamma t_n} \right) e^{-(\gamma + r_n (1 - \theta_n)) (T - T^*_n)} \]

and

\[ g(A, q^*_n, q^*_b) = \int_0^{T^*_n} e^{-\gamma t_n} \left( R + r_b (p_b - C/\mu_b) \right) dt + \int_{T^*_n}^{T^*_i} e^{-\gamma t_n} \left( R - r_b c_b \right) e^{-(\gamma + r_n (1 - \theta_n)) (t - T^*_n)} dt \]
\[ + A e^{-\gamma t_n} e^{-(\gamma + r_n (1 - \theta_n)) (T - T^*_n)} \]

Proof of Proposition 4.4

It follows from Lemma 4.1 and the fact that \( f(A, q^*_n, q^*_b) \) and \( g(A, q^*_n, q^*_b) \) are independent of \( \lambda_n(0) \).

Proof of Proposition 4.5

Since \( q^*_n \) and \( q^*_b \) are both functions of the terminal value \( A \), for the simplicity of notation, we denote \( g(A, q^*_n, q^*_b) \) by \( g(A) \). As we discussed, \( g(A) \) is the terminal value for the optimal control problem in period \( N - 1 \). Thus the terminal value in period \( N - 2 \) becomes \( g(g(A)) \). Same discussion applies to other periods.

Define

\[ g^{(i)} (A) = \underbrace{g \left( g \left( \ldots \left( g \left( A, q^*_n, q^*_b \right) \right) \right) \right)}_{i \text{ times}}, \tag{4.67} \]

and let \( g^{(0)} (A) = A \). The optimization problem in the \( i^{th} \) period becomes

\[ \Pi_i \left( x^i_0 \right) = \max_{\lambda^i_n(0), q^i_n(t), q^i_b(t)} \left\{ P \left( \lambda^i_n(0), q^i_n(t), q^i_b(t) \right) + x^i_0 (T) g^{(N-i)} (A) \right\} \]
\[ + \sum_{j=i+1}^{N} \lambda^*_{n,j} (0) f(g^{(N-j)} (A)) - S \left( \lambda^*_{n,j} (0) \right). \tag{4.68} \]

Next we show that the terminal value converges over time for different capacity costs:
a) If $C/\mu_b \leq p_b + c_b + (1 - \theta_b) \cdot \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}$:

First note that the following conditions are equivalent

$$\frac{C/\mu_b - p_b - c_b}{1 - \theta_b} \leq \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \Leftrightarrow \frac{C/\mu_b - p_b - c_b}{1 - \theta_b} \leq \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} \quad (4.69)$$

Now, consider following cases

a.i) If

$$\frac{C/\mu_b - p_b - c_b}{1 - \theta_b} \leq g^n(A), \quad (4.70)$$

where $g^n(A)$ is the terminal value of the advertisement period $N - n$. Therefore, we use Proposition 4.2 to determine optimal service policy over this period.

It follows from (4.51), (4.52), (4.69), and (4.70) that $T^*_{b} = 0$, and $q^*_b(t) = 1$ during the entire period. Thus,

$$g^{n+1}(A) = \int_0^T e^{-\gamma_b t} (R + r_b (p_b - C/\mu_b)) dt + g^n(A) e^{-\gamma_b T} = \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} (1 - e^{-\gamma_b T}) + g^n(A) e^{-\gamma_b T},$$

which is a convex combination of $\frac{R + r_b (p_b - C/\mu_b)}{\gamma_b}$ and $g^n(A)$. Therefore, from (4.69) and (4.70)

$$\frac{C/\mu_b - p_b - c_b}{1 - \theta_b} \leq g^{n+1}(A).$$

Repeating the above step

$$\lim_{n \to \infty} g^n(A) = \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b}.$$  

a.ii) If

$$g^n(A) \leq \frac{C/\mu_b - p_b - c_b}{1 - \theta_b}, \quad (4.71)$$

103
by Proposition 4.2, \( q^*_b(t) \) follows from (4.59). Thus,

\[
g^{(n+1)}(A) = \int_0^{T^*_b} e^{-\gamma t} \left( R + r_b \left( p_b - C/\mu_b \right) \right) dt + \int_{T^*_b}^{T} e^{-\gamma t} \left( R - r_b c_b \right) e^{-\gamma (t-T^*_b)} dt \\
+ g^{(n)}(A) e^{-\gamma T^*_b} e^{-\gamma (T-T^*_b)} \\
= \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \left( 1 - e^{-\gamma T^*_b} \right) + \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)} e^{-\gamma t} \left( 1 - e^{-\gamma (t-T^*_b)} \right) \\
+ g^{(n)}(A) e^{-\gamma T^*_b} e^{-\gamma (T-T^*_b)} \\
\geq g^{(n)}(A)
\]

where the last inequality follows from (4.69) and (4.71). Therefore, repeating the above step, \( g^{(n)}(A) \) increases in \( n \) until it reaches \( \frac{C/\mu_b - p_b - c_b}{1 - \theta_b} \). After that, as in the proof in part (a.i), it converges to \( \frac{R + r_b (p_b - C/\mu_b)}{\gamma_b} \).

\[ b) \text{ If } C > p_b + c_b + (1 - \theta_b) \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}, \text{ similar discussion shows} \]

\[
\lim_{n \to \infty} g^{(n)}(A) = \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}.
\]

\[ (4.72) \]
References


