Constructing Learning Models from Data: The Dynamic Catalog Mailing Problem

by

Peng Sun

B.Eng., Tsinghua University (1998)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Author .....
Sloan School of Management
May 16, 2003

Certified by .....
Duncan Simester
Associate Professor of Management Science
Thesis Supervisor

Accepted by.....
John Tsitsiklis
Professor of Electrical Engineering and Computer Science
Co-director, Operations Research Center

Constructing Learning Models from Data: The Dynamic Catalog Mailing Problem

by

Peng Sun

Submitted to the Sloan School of Management
on May 16, 2003, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY
in Operations Research

Abstract

The catalog industry is a large and important industry in the US economy. One of the most important and challenging business decisions in the industry is to decide who should receive catalogs, due to the significant mailing cost and the low response rate. The problem is a dynamic one — when a customer is ready to purchase, s/he may order from a previous catalog if s/he does not have the most recent one. In this sense, customers' purchasing behavior depends not only on the firm's most recent mailing decision, but also on prior mailing decisions. From the firm's perspective, in order to maximize its long-term profit it should make a series of optimal mailing decisions to each customer over time.

Contrary to the traditional myopic catalog mailing decision process that is generally implemented in the catalog industry, we propose a model that allows firms to design optimal dynamic mailing policies using their own business data. We constructed the model from a large data set provided by a catalog mailing company. The computational results from the historical data show great potential profit improvement.

This application differs from many other applications of (approximate) dynamic programming in that an underlying Markov model is not *a priori* available, nor can it be derived in a principled manner. Instead, it has to be estimated or "learned" from available data. The thesis furthers the discussion on issues related to constructing learning models from data. More specifically, we discuss the so called "endogeneity problem" and the effects of inaccuracy in model parameter estimation.

The fact that the model parameter estimation depends on data collected according to a specific policy introduces an endogeneity problem. As a result, the derived optimal policy depends on the original policy used to collect the data. In the thesis we discuss a specific endogeneity problem, "attribution error." We also investigate whether online learning can solve this problem. More specifically, we discuss the existence of fixed point policies for potential on-line learning algorithms.

Imprecision in model parameter estimation also creates the potential for bias. We illustrate this problem and offer a method for detecting it.

Finally, we report preliminary results from a large scale field test that tests the effectiveness of the proposed approach in a real business decision setting.

Thesis Supervisor: Duncan Simester
Title: Associate Professor of Management Science

To the memory of my mother

Acknowledgments

This work benefited greatly from the superb advice and constant encouragement from my thesis supervisor, Professor Duncan Simester. I learned so much from Duncan and the thesis is a result of our cooperation.

I also want to thank Professor John Tsitsiklis for many discussions and his insightful advice on the directions of this research over the past couple of years.

It is very fortunate for me to have Professor Rob Freund bringing me to the ORC and serving as my research advisor, co-author and thesis committee member in the past few years. Rob's influence on me goes beyond research.

I want to thank Stephen Windsor for his work on maintaining the data sets and coding up many of the programs for this research project. And I want to thank the company who provides us with the data. This research has been partially supported by the Singapore-MIT Alliance and the Center for Innovative Product Development, both at MIT.

The ORC is great. And life would not have been quite the same without friends here, Melvyn, Adam, Fernando... especially Chen Xin, whose influence on me has been fundamental.

My friends at 440 Mass. Ave. #3, CSSA, at and beyond MIT have been such an important part of my life in the past five years. I thank you all.

It was my great honor to have my first Thanksgiving dinner in this country at Professor John Little's house and have him in my thesis defense.

Many ORC affiliated faculty members, co-directors and administrative staff have been very helpful.

Finally and most importantly, I want to thank my family. I feel so important to see my parents happy for me, considering their selfishless contribution over the years. My wife, Huina, provides the ultimate love and support that deserves everything in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objectives and Results

This thesis proposes a dynamic programming approach for designing a near optimal solution to the catalog mailing problem and addresses issues that arise when constructing reinforcement learning models from off policy sample trajectories.

Catalog firms in the USA mailed almost 17 billion catalogs in 2000 [2], generating over $88 billion in consumer and business to business spending. Printing and mailing these pieces was the second largest contribution to these firms' expenses, representing almost 25% of total costs. As a result, direct mail managers view improvements to their mailing policies as one of their highest priorities [14], [16].

Traditionally the catalog mailing process, similar to many other direct marketing decisions, is considered myopic – firms mail to customers who they judge are most likely to respond immediately, overlooking the subsequent implications of this mailing decision. This research explores the dynamic nature of catalog mailing decision strategies and develops dynamic programming models using state-of-the-art advances in approximate dynamic programming and reinforcement learning. Rather than focusing on improving the response to each mailing, our solution approach emphasizes learning near optimal mailing policies that maximize long-term profits directly from historical data. We formulate the model as

an infinite time horizon discounted Markov Decision Process, with a look up table state space. The state space is constructed based on the encoded historical information of each customer's past purchases and received catalogs. Computational results using data provided by a retail catalog mailer show that the models have the potential to greatly improve long-run profits.

Since historical data was generated from a sub-optimal mailing policy and there is no perfect simulation model for customers' behavior under a different mailing policy, we test the effectiveness of our approach through a field test. A large-scale field test of the proposed model is currently underway with a catalog retailer. The field test is expected to run for six months and involves a total of $60,000$ customers, randomly assigned to Treatment and Control groups. Mailing decisions for the customers in the Control group are made by the company using their current mailing policy, while mailing decisions for customers in the Treatment group are made using the new model. We not only observe the profits earned from different groups over the six months but also compare the trajectory of those profits and the final state distribution of the customers.

As will be explained in detail in the thesis, the catalog mailing problem has the following characteristics:

(1) We do not have an *a priori* model describing the dynamic of the system — all the model parameters have to be learned directly from data.

(2) There is no online data collection scheme — we only have access to historical data generated from a historical policy.

These two restrictions imply that we are limited to constructing a dynamic programming model from off policy sample trajectories. In this thesis, we address several problems related to constructing and solving infinite time horizon dynamic programming models with state space and model parameters constructed from historical data.

More specifically, we address the following issues in this thesis:

1. *Hidden Information*:

In Chapter 2 we construct a discrete state space S to summarize the available information for decision making. Because the state space is not a perfect encoding of the available information, there could be hidden information not considered in the state space construction. This means that a state s in the current state space S could be an aggregation of many (tiny) "true" states. The distribution of data on these "true" states affects the estimation of model parameters associated with state s (for example, the immediate reward and the transition probabilities out of state s). Under different policies the distribution of data on the "true" states varies. This implies that the model parameter estimation from data and the optimal policy based on the model are affected by the original policy used to generate the data.

When historical actions are different on the "true" states, an "attribution error" may arise — the policy improvement procedure of the dynamic programming algorithm attributes the effects of different "true" states (distinguished by hidden information) to different actions. The dynamic programming algorithm based on such a model generates a sub-optimal policy and a misleadingly upward biased value function. In this research we provide theoretical justifications and empirical evidence of this bias.

2. *Batch Online Learning*:

   One approach to resolving the problems that result from hidden information is to use "batch online learning" — we cumulate some data under a near optimal policy (derived from the model estimated from historical data) and re-evaluate the model parameters and optimize it. Ideally, we hope that by repeating such a procedure, we are able to learn a close to optimal policy for the aggregated state space. In this thesis, we investigate the merits of this sequential approach. More specifically, we show that there exist self-enforcing policies such that data collected under such a policy reinforces the optimality of the policy. This implies that potentially we could design batch online learning algorithms to reach some fixed points. However, we also provide examples that there could be multiple self-enforcing polices with

very different profit-to-go performance. This result implies that an algorithm merely reaching a fixed point is not sufficient to guarantee "optimality."

3. *Imprecision in the Observed Model*:

   Because the transition probabilities and expected rewards are estimated from data, inevitably randomness in the data cause estimation errors in the model parameters. The dynamic programming procedure tends to direct the policy towards actions taking advantage of the random errors and therefore generates an upward biased value function and a sub-optimal policy. Theoretically we show the conditions under which the bias occurs and the effects of different model parameters on this bias. We also provide a test for identifying the extent of the problem in the computation of the catalog mailing problem. Empirical evidence shows that the problem exists even in a seemingly "massive data" environment. This demonstrates the managerial importance of the problem when we use dynamic programming model constructed from data.

1.2 Structure of Thesis

The structure of thesis is as follows: In Chapter 2, we present the solution framework for the catalog mailing problem.

Motivated by the computational experiences for solving this problem, we further the discussion on hidden information in Chapter 3 and the model imprecision problem in Chapter 4.

In Chapter 3, we present theoretical justifications, empirical evidence and ways of mitigating the attribution error. Then we extend the discussion to batch online learning schemes. We show the existence of fixed point policies and provide an example of multiple fixed points.

In Chapter 4, we discuss the effect of model parameter estimation error through theoretical justifications and computation. Chapter 5 describes the on-going field test and presents

some preliminary empirical results. The last chapter, Chapter 6, concludes the thesis.

Chapter 2

Dynamic Catalog Mailing Problem

2.1 Introduction

Identifying an optimal mailing policy is a difficult task. Customer response functions are highly stochastic, reflecting in part the relative paucity of information that firms have about each customer. Moreover, the problem is a dynamic one. Although it may appear that receiving the current catalog is a necessary condition for making a purchase, this is not true in practice. A customer who is ready to make a purchase will often purchase from an earlier catalog if they did not receive the most recent catalog. More generally, customers often have an extensive stock of experience with a catalog company, stretching over many prior catalogs and purchase experiences. This prior experience will often play a more important role in determining customers' purchase probabilities than receipt of the most recent catalog.

As a result, customers' purchasing decisions are influenced not just by the firm's most recent mailing decision, but also by prior mailing decisions. From the firm's perspective, the probability that a customer will respond to the current catalog may be less important than the impact on the likelihood of a future purchase. This leads to a very difficult optimization problem for the firm; the optimal mailing decision depends not just upon the customer's response to the current catalog, but also upon the firm's past and future mailing

decisions.

Current mailing policies are almost invariably myopic. Firms mail catalogs to customers who they judge are most likely to respond to that catalog, overlooking the subsequent implications. We develop a model that allows firms to address the dynamic implications of mailing decisions. We test the proposed model using a large data set provided by a mail order catalog firm. The findings confirm that the distinction between a myopic and a dynamic strategy is an important one. A myopic policy may argue against mailing to some customers because the likelihood of an immediate high value order is low. In contrast, the dynamic policy will sometimes mail to these customers because doing so may increase the probability of future purchases.

2.1.1 Current Catalog Mailing Policies

Catalogs are mailed on specific mailing dates, which are pre-determined up to a year prior to the mailing date. This long lead-time reflects the period required to design the catalogs and coordinate product-purchasing (inventory) decisions. The frequency of these mailing dates varies across firms and seasons. However, most firms mail between 15 and 50 catalogs a year. We will treat the catalog mailing dates as exogenous.

Catalogs are mailed to a combination of past customers and prospective customers. Past customers are often described as "house" customers, while prospective customers are commonly referred to as "prospects". The procedures used to identify a mailing policy for house customers differ considerably from the procedures used for prospective customers. This reflects both the difference in the likelihood of a response and the difference in the amount of information that firms have about each type of customer.

2.1.2 Prospective Customers

Prospective customers are generally identified by renting a mailing list from a third-party vendor at a cost of between $\$60$ and $\$120$ per thousand names. When choosing which lists

to rent, firms try to match the demographic characteristics of their existing customers with the characteristics of prospects on rental lists. Firms receive the name and address of each prospect and acquire the right to mail to them only once. The rental agreements require that the company delete from its database all information about households that do not respond to this mailing. Mailing lists are seeded with disguised names that allow the third party vendor to detect violations (Anderson and Simester 2002 [1]).

Mailing policies with prospective customers focus on selecting which rental list to use. Firms generally have too little information about each prospective household to make informed mailing decisions for each households. List vendors also generally require that a firm rent the entire list (less some allowance for overlap with the house list). Interestingly, the level of overlap with the house list is often positively correlated with the likelihood that other prospects on the rental list will respond.

The average response rate when mailing to prospective customers is low, often less than 0.5%. Catalogs typically lose money in the short-term when mailing to prospects as the mailing costs generally exceed the profits from the resulting orders. They only mail to prospective customers in order to increase their pool of house customers. Although, the value of acquiring a customer depends on the subsequent mailing policy, these dynamic considerations have little influence on mailing policies for prospective customers. The low average response rate and the cost of re-mailing prospects who do not respond to a first mailing reduce the objective to choosing a rental list for which the response to a single mailing will be high. Because we focus in this paper on the dynamic characteristics of the optimal mailing policy we will restrict attention to house customers.

2.1.3 House Customers

The names, addresses and purchase histories of house customers are generally considered to be amongst a catalog's most valuable assets. The procedures used to select mailing policies for house customers vary across firms, but they all share three common components: (a) firms collect data describing different customer characteristics; (b) this data is used to

group customers into discrete segments; and finally (c) firms evaluate the probability that customers in each segment will respond.

The primary variables used to segment house customers are the so-called RFM measures, which describe the recency, frequency and monetary value of customers' prior purchases. The Direct Marketing Association (DMA 2001) [2] reports that amongst catalogs selling consumer products, 84% use the Recency measure in their mailing policy, 80% use the Monetary Value measure and 78% use the Frequency measure. In addition, 28% of companies use information about customers' purchases from competing catalogs. This competitive information is pooled by third party firms and supplied to cooperating catalogs. Just 4% of firms use other sources of data, such as the history of catalogs mailed to each customer.

To group customers into distinct segments, many firms simply discretize the (continuous) RFM measures. For example, managers from a clothing catalog reported that they identify whether the period since the last purchase is: less than 6 months, between 6 months and 1 year, between 1 and 2 years, between 2 and 3 years, between 3 and 4 years, or more than 4 years. The segments are then defined by the intersection of the discretized RFM measures. Some firms use a more sophisticated approach, in which the RFM variables and other purchase history measures are used to develop customized models that predict how likely customers are to respond to a catalog. The predictions from these models are then discretized to identify separate segments.

Having segmented the customers, the third component focuses on determining which customer segments to mail to. A common approach is to use a simple breakeven analysis, in which the firm estimates the average response rate required to breakeven. It then mails to the customers in a segment if and only if the historical response rate for that segment is sufficiently high. This standard policy may be enhanced by a series of exceptions, such as arbitrary policies to mail to all customers who have made a recent purchase. Notably current mailing policies invariably focus on the probability of a response to the next catalog. They do not consider the dynamic implications of the current mailing decision.

The catalog mailing problem context raises two types of issues that have prevented the application of standard Dynamic Programming techniques:

(a) The dimensionality of the problem is large. For example, if we take the "state" of a customer to be a representation of the customer's past history, then the large number of possible histories translates to a large number of possible states. This leads to the need for approximations, either of the model or in the solution algorithm.

(b) An underlying Markov model is not a priori available, nor can it be derived in a principled manner. Instead, it has to be estimated or "learned" from available data.

In this chapter we describe a model that address these issues and allows for the dynamic optimization of mailing policies. In doing so, the proposed model provides modifications to all three components in firms' current mailing policies.

2.2 Literature review

2.2.1 Direct marketing literature

There is an extensive literature investigating topics relevant to the catalog industry. This includes a series of studies that use catalog data to investigate pricing cues or the impact of price promotions (see for example Anderson and Simester 2002 [1]). Other topics range from customer merchandise returns (Hess and Mayhew 1997 [17]), to customer privacy (Schönbachler and Gordon 2002 [21]) and catalog copy issues (Fiore and Yu 2001 [12]).

There have also been several prior studies investigating optimal catalog mailing strategies. Bult and Wansbeek (1995) [9] present a model for making mailing decisions that builds on work by Banslaben (1992) [4]. They develop a model to predict whether customers will respond to a catalog and link the model to the firm's profit function in order to derive a profit maximizing decision rule. This approach is more rigorous, but conceptually similar, to the final component of the procedure that many firms currently use (described

above). The authors evaluate their model using a sample of 13,828 customers from a direct marketing company selling books, periodicals and music in the Netherlands. They show that their methodology offers strong predictive accuracy and the potential to generate higher net returns than traditional approaches.

Bitran and Mondschien (1996) [8] focus on the role of cash flow constraints when making catalog mailing decisions. The cash flow constraint introduces a tradeoff between mailing to prospective customers and mailing to house customers. Mailing to prospective customers is an investment that yields negative cash flow in the short term but builds the company's house list, while mailing to the house list enables the firm to harvest value from its earlier investments. The model incorporates inventory decisions, so that the profitability of the mailing policy depends upon the availability of inventory. The authors present heuristics that approximate a solution to their model and test the model using a series of Monte Carlo simulations.

Gönül and Shi (1998) [14] propose a model of mailing policies that explicitly recognizes that a mailing policy may affect demand beyond the current period. The primary focus is on the customer response model. The model assumes that customers understand both the firm's mailing strategy and the stochasticity in their own purchasing decisions. When making purchasing decisions customers consider both the current and future impact of their decisions. In particular, customer utility is an increasing function of whether they receive catalogs and so customers contemplate how their purchasing decisions will affect the likelihood that they will receive catalogs in the future. The firm's mailing policy and the customers' purchasing decisions are jointly optimized using successive maximum likelihood approximations. The authors test their predictions using the purchase histories for 530 households selected from the house list of a retailer of durable household products. Their findings indicate that their proposed policy has the potential to increase the firm's profits by approximately 16%.

The catalog-mailing decision shares many similar features to the problem of deciding whom to offer price promotions to. Pednault, Abe and Zadrozny (2002) [19] recently pro-

posed a dynamic approach to address this question. They observe that promotion decisions are also typically made myopically and argue that maximizing profits on each promotion in isolation may not be as profitable as a strategy that seeks to maximize the dynamic sequence of promotion decisions. The authors use function approximation to estimate the value function directly without an underlying response model.

2.2.2 Reinforcement learning

The methodologies employed in this prior work fall under the general umbrella of "approximate dynamic programming" and "reinforcement learning" (Bertsekas and Tsitsiklis 1996 [7] and Sutton and Barto 1998 [22]). In particular, the Gönül and Shi (1998) [14] and Pednault, Abe and Zadrozny (2002) [19] papers are examples of standard approaches to applying approximate dynamic programming methods to social science data. In the Gönül and Shi (1998) [14] paper the proposed algorithm proceeds in two distinct steps: the authors first estimate a statistical model of the underlying response function and then apply standard dynamic programming methods to this model. Similar approaches have been used to address airline pricing (yield management) together with a range of applications in the finance industry. A limitation of this approach is that the dynamic programming results are potentially sensitive to errors in the statistical model of the response function. The function approximation approach used by Pednault, Abe and Zadrozny (2002) [19] does not rely on standard model-based dynamic programming methods and instead estimates the value function directly without specifying an underlying model. The major limitation of this approach is that it is not guaranteed to yield accurate solutions when using data obtained under an historical policy that differs from the evaluated policy (Baird 1995 [3]; Tsitsiklis and Van Roy 1997 [23]). Even convergence can be problematic and may require experimentation in order to set parameters, such as learning rates.

The method proposed in this study addresses these limitations using a fundamentally different approach. The method, which we discuss in greater detail below, begins by designing a discrete state space to approximate customers' histories. We then calculate tran-

sition probabilities and one-step rewards directly from the data. This direct estimation of the customers' response function from the data provides an extremely flexible functional form and allows us to greatly expand the dimensionality of the problem. The method has its own limitations, which we identify and propose solutions for.

2.3 Overview of model and notations

In this section we provide an overview of our model and notations.

We interpret the company's sequence of mailing decisions as an infinite horizon task (there is no end point) and seek to maximize the discounted stream of future profits. Time is measured in discrete steps defined by the exogenously determined catalog mailing dates. The intervals between mailing dates typically vary and so we will allow time steps to have different lengths. We use the term "reward" to describe the profit earned in any time period. This reward is calculated as the net profits earned from a customer's order (if any) less mailing costs (if a catalog was mailed that period). We attribute the profits from a purchase to the time step in which the purchase occurred, rather than the date of the catalog that the customer orders from.

This approach offers two advantages. First, it is consistent with our claim that profits earned during a period are affected by factors other than the most recent mailing decision. For example, catalogs may cannibalize from each other, so that customers may be less likely to purchase from a specific catalog if they are mailed another catalog two weeks later (see later discussion). Second, it overcomes the practical problem that it is often difficult to link a purchase to a specific catalog. This problem arises for approximately 15% of the transactions in our data set.

Customers' histories (and their current status) will be described at each point in time by a set of n variables such that each customer at each time period is represented by a point in an n-dimensional space. These n variables summarizes the whole historical information about a customer. Details on the construction of the variables is in the next section.

Formally, we will define a vector space **X** to be Cartesian product of the range of the n variables. Each customer's historical movement in this space **X** provides a sample trajectory. At each time period, the collection of a customer's historical information before that time period formulates an "observation". In the rest of the chapter, we use "observations" to refer to the information we have from each customer at each time period. Each observation corresponds to a data point in space **X**.

We will segment the space into mutually exclusive and collectively exhaustive discrete states defined by linear demarcations in the **X** space, reasons discussed later in the chapter. This requires that we identify a mapping from **X** to the discrete state space S. Intuitively, the states group neighboring observations that have similar histories and are expected to respond in a similar way to future mailing decisions.

Because the historical information describing a customer at a time period is encoded in the vector space **X**, the **X** space preserves the Markovian property. However, the aggregation of the **X** space into a discrete state space S may not preserve this property. This introduces a trade-off between the computational complexity of the model and the accuracy of the Markovian assumption. We will resolve this trade-off by creating an approximate model and solving it exactly, rather than finding an approximate solution to an exact model. In particular, we will assume that the evolution of the aggregate state is Markov. Thus we assume the state a customer is in completely summarizes all of the information that we will use about that customer in that mailing period. Obviously, the design of the states is an important challenge, which we address in Section 4.

These assumptions define a Markov Decision Process (MDP) on the state space S for which out of each state the action space is binary $U := \{0, 1\}$ where $a = 0$ represents "not to mail" and $a = 1$ represents "to mail". A stationary mailing policy $\pi : S \to U$ determines that out of state $s \in S$, action $\pi(s) \in U$ is taken. The policy embedded in the historical data is a randomized policy because of the nature of business practice. In other words, out of each state s, we are able to observe both actions from the historical data. We then define the historical policy $\tilde{\pi}$ to be a stochastic stationary policy from each state

to a $(0, 1)$ Bernoulli distribution which can be estimated from the historical data. Out of each state-action pair (s, a), there is a one stage reward $r_{s,a}$. Given any policy π, the state evolution follows a well-defined Markov chain s_t^π with transition probabilites

$$P\left(s_{t+1}^\pi = j | s_t^\pi = i\right) = p_{ij}^\pi.$$

$r_{s,a}$ and p_{ij}^π are not provided a priori but need to be estimated from the data.

The "reward-to-go" of policy π starting from state s is defined as

$$J^\pi = E\left[\sum_{t=0}^{+\infty} \alpha r_{s_t,\pi(s_t)}\right], \tag{2.1}$$

where α is a discount factor satisfying $0 < \alpha < 1$. The optimal reward-to-go is defined by

$$J^*(s) = \max J^\pi(s).$$

It is well known from standard dynamic programming results [6] that there exists an optimal stationary policy π^* solves the above maximization problem for all state $s \in S$.

In the next few sections, we discuss the construction of the model in details. In section 2.4, we discuss the variables that we used to construct the state space **X**. After that in section 2.5, we present our argument of choosing a discrete state space and detailed discretization procedures. In Section 2.6 we will describe a dynamic programming algorithm capable of identifying the optimal policy and the aggregate discounted profits associated with this policy. As a benchmark the algorithm will also describe the aggregate discounted profits associated with the policy represented in the historical data (the company's current policy).

2.4 Summarizing Customers' History

We have claimed that the probability of a purchase is affected by the customers' stock of prior experiences. In this section we review the variables used to describe these expe-

riences. We begin by focusing on customers' purchase histories and then consider their purchase histories.

2.4.1 Purchase Histories

Customers' purchase histories are typically summarized using the RFM variables. However, there is an important limitation in the RFM measures. More recent information would seem to be more useful in prediction and decision-making, yet the frequency and monetary value measures do not discriminate according to the recency of the purchase (or mailing). For example, they cannot distinguish between two customers who both purchased twice, but one purchased in the last month and two years ago and the other customer purchased in each of the last two months. A solution is to use discounted aggregate stock measures: $p_t = \sum_{k \in K_t} \eta^{T_k} x_k$. Here K_t is the set of purchases by the customer prior to period t, $\eta \in [0, 1]$ is a decay rate per unit of time, T_k denotes the number of units of time between period t and the k^{th} purchase, and x_k describes the amount spent on the k^{th} purchase. An analogous stock measure can be constructed for frequency by omitting the x_k term. Both measures distinguish between the two customers described in the example above. By varying the choices of η and x_k, we can construct variables that summarize the purchase history of a customer into a more general range of variables than the RFM values.

2.4.2 Mailing Histories

Although maintaining a record of a customer's mailing history is no more difficult than maintaining a record of the customer's purchase history, few catalog retailers store the mailing history. We do not offer this as an explanation for why the mailing history is not used in the mailing policy. Indeed, the causation probably operates in the reverse; many firms do not store the mailing history because they do not use it. An alternative explanation is that the mailing history is highly correlated with the purchase history, so that the purchase history provides a sufficient statistic. However, in practice, variance in the mailing policy

ensures that the purchase history is not a sufficient statistic. This variance results from personnel changes, experimentation, seasonality and changes in the procedures that firms use to calculate the probability of a response. The variance is important; without it we could not estimate the effectiveness of alternative mailing strategies (see later discussion).

To describe each customer's mailing history we can use an analogous set of variables to those developed to describe the purchase history. In particular, the recency and frequency of past mailings are directly analogous to the recency and frequency variables of past purchases. A mailing frequency measure suffers from the same shortcoming as the purchase frequency measure in that it does not distinguish between more and less recent mailings. However, the proposed stock variables can also be used in the mailing context. In particular, define $m_t = \sum_{k \in K_t} \eta^{T_k}$ where K_t identifies the set of catalogs mailed to the customer prior to period t.

It is possible to design a wide range of variables to describe the complexity of customers' mailing and purchase histories. In practice, however, high dimensionality brings computational challenges. In the next section we propose a strategy for discretizing the state space that does not suffer from these problems.

2.5 Constructing the Discrete State Space

In this section we propose a method for discretizing the original vector space $\mathbf{X}$ to the discrete state space S, or more formally, a mapping $D : \mathbf{X} \rightarrow S$. Use of a discrete state space guarantee an exact and robust dynamic programming solution. However, it also results in loss of information.

The general approach to designing a discrete state space is to tile along the dimensions. This is what firms currently do when implementing the standard RFM model. They group customers into segments by simply discretizing the RFM measures using fixed demarcations of each variable. Each customer in each time period falls within a unique demarcation on each variable, and the interaction of these demarcations defines a set of discrete Markov

states. There are several difficulties with this approach. Notably, it can yield a large number of states, and observations are often unevenly distributed across these states (states are populated with few or no observations).

An alternative approach is to develop a predictive model of how likely customers are to respond to a catalog and to discretize the predictions from this model. The DMA reports that this approach, which will tend to yield fewer more evenly distributed segments, is used by approximately 28% of catalog firms (DMA 2001 [2]). However, while this approach is well-suited to a myopic mailing policy, it is not well suited to a dynamic policy. There is no guarantee that grouping customers according to the predicted response to the next catalog will allow the model sufficient discrimination in a dynamic context. In particular, a new customer with few prior purchases may have the same purchase probability as an established customer who has extensive experience with the catalog. Yet the long-term benefits of mailing the established customer may be different than the benefits of mailing the new customer.

Therefore we propose a new algorithm for constructing a finite Markov state space S from the original vector space $\mathbf{X}$. We adopt three objectives when designing the discrete states. First, the states should be "meaningful," so that each state $s \in S$ is visited with positive probability. Second, the states should be "representative," so that data points in the same state are geometrically close to each other (in $\mathbf{X}$ space). Finally, the states should be "consistent," so that observations within the state share a similar profit stream given an identical mailing policy. We can only apply the "meaningful" and "consistent" criteria to the policy represented in the historical data. However, we can validate these criteria *a posteriori* when an optimal policy is available.

We will begin by initially estimating a value function for each customer under the historical mailing policy. For a customer at point $x \in \mathbf{X}$, let function $\tilde{V}^{\pi_H}(x)$ estimate the present value of the discounted future profit stream, given the historical mailing policy embedded in the data. Here π_H indicates the historical mailing policy and the tilde denotes the initial estimation. If the period of time covered by the historical data is sufficiently long,

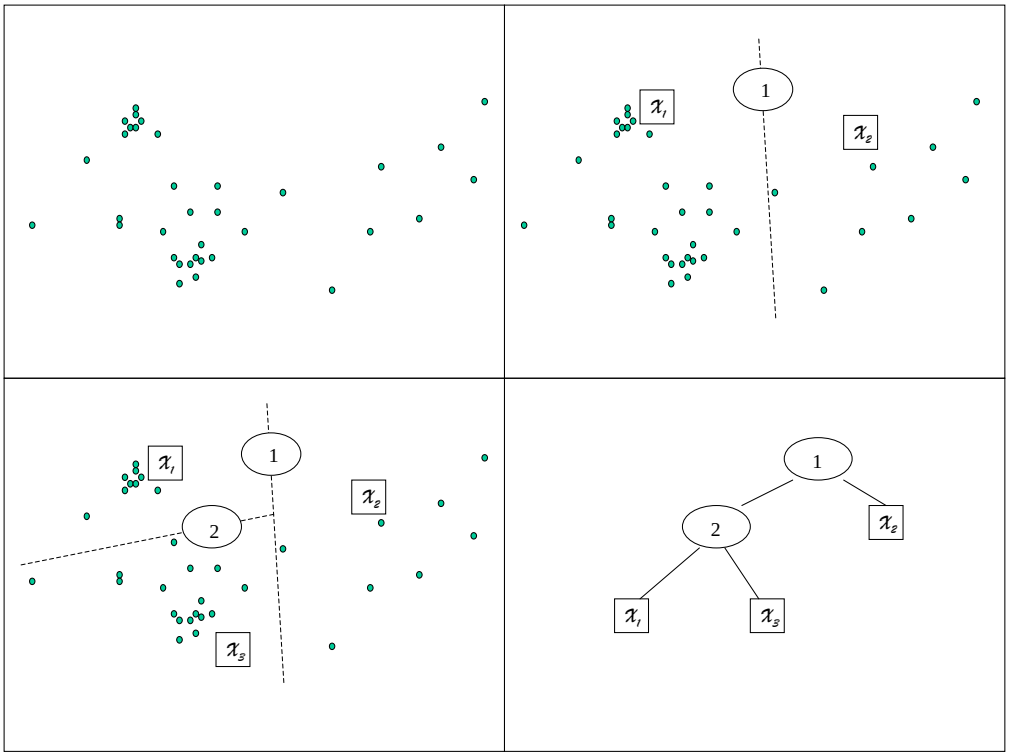

Figure 2-1: Discrete State Space Design

this estimate can be derived by fitting a function of the discounted aggregate profits earned for a representative sample of customers (see later discussion). Given the estimates of the value function for the historical policy we use a series of separating hyperplanes to divide the state space into pieces organized by a binary tree structure.

We illustrate this intuition in Figure 2-1 below. Assume that we describe customers' history using just two variables ($n = 2$). A sample of data represented in this two dimensional **X** space is portrayed in Figure 2-1(a). Line 1 represents a hyperplane in this **X** space that separates the sample in two sub-segments (Figure 2-1(b)). The next iteration begins by selecting the segment with the highest variance in $\tilde{V}^{\pi_H}(x)$ (not shown) and placing a second separating hyperplane (Line 2) through this segment. Following this second iteration there are a total of three segments (see Figure 2-1(c)). The process continues until a stopping rule is met, such as the desired number of segments.

The outcome is a tree-structure (Figure 2-1(d)), where the hyperplanes are branches on

the tree and the segments are the leaves. A state space with N segments requires a tree with $N-1$ hyperplanes.

Given the tree structure, the path from the root to each leave node defines a set of inequalities identifying each state. Aggregation of states is also easily accomplished by pruning a large tree structure to a smaller one. This use of a binary tree structure is similar in spirit to the decision tree methods for classification (Duda, Hart and Stork 2000 [11]) and the Chi-Squared Automatic Interaction Detection (CH-AID) methods in customer segmentation (see for example Bult and Wansbeek 1995 [9]). The primary difference between the methods is the design of the hyperplanes determining the branches.

The algorithm that we use for identifying the hyperplanes proceeds iteratively, where each iteration has two steps. First, we select the segment for which the variance in $\tilde{V}^{\pi_H}(x)$ is the largest. Formally, we select the segment $X_i \subset \mathbf{X}$ for which $\sum_{x \in X_i} \left(\tilde{V}^{\pi_H}(x) - \bar{V}_{X_i} \right)^2$ is largest. This criterion favors the selection of segments that are least consistent and/or have the most members. To prevent states with very few observations we only select from amongst segments with at least $1,000$ observations in them.

In the second step we divide this segment X_i into two segments X_i' and X_i''. To satisfy the consistent objective, we would like the observations within each sub-segment to have similar values on $\tilde{V}^{\pi_H}(x)$. To achieve this we might fit a step-size function to the $\tilde{V}^{\pi_H}(x)$ values in X_i. However computationally this is a difficult problem, and so we use a heuristic to approximate this step. The heuristic uses the following steps:

1 Use OLS to estimate $\hat{V}^{\pi_H} = \alpha + \beta^T x$ using all of the observations (x) in the selected segment X_i. That is, we find α and p that minimize

$$\sum_{x \in X_i} \left(\tilde{(V)}^{\pi_H}(x) - \alpha - \beta^T x \right)^2$$

2 Find the center of the observations in the segment $\bar{x} = \sum_{x \in X_i} x$ by calculating the average of the observations on each dimension of $\mathbf{X}$.

3 Compute α' such that $\alpha' + \beta^T \bar{x} = 0$ and divide segment X_i into two segments X_i'

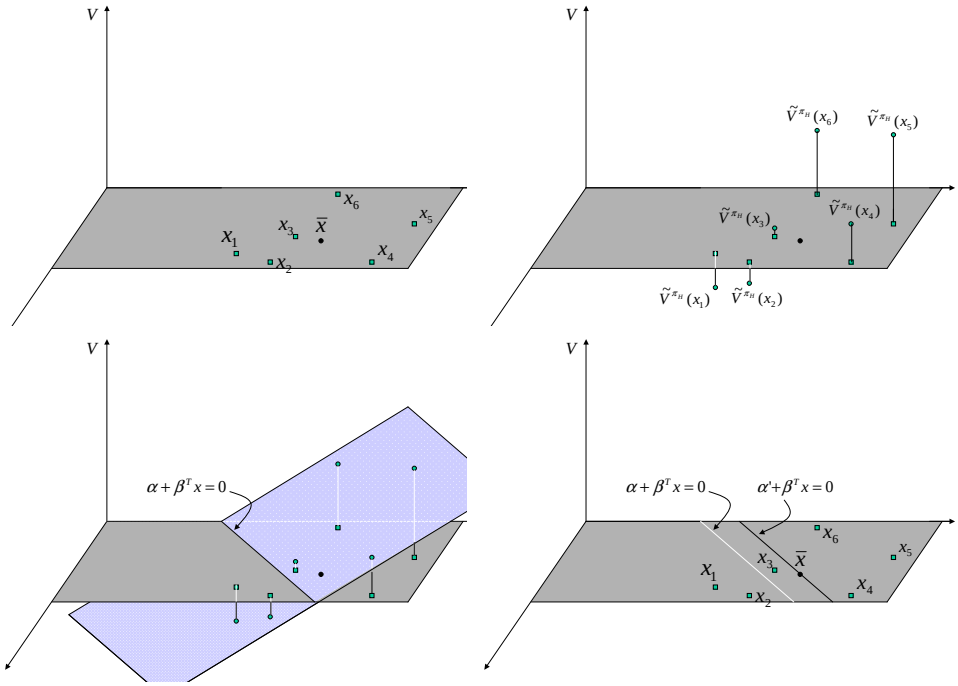

Figure 2-2: Dividing Segment

and X_i'' along the hyperplane defined by $\alpha' + \beta^T x = 0$.

We can again illustrate this process in Figure 2-2 using a $2-$dimensional **X** space.

In Figure 2-2(a) we depict the observations in a selected segment. The center of these observations is defined by $\bar{x}$ and each observation has an estimated $\tilde{V}^{\pi_H}$ (Figure 2-2(b)). We use OLS to regress $\tilde{V}^{\pi_H}$ on x, which we illustrate in Figure 2-2(c) as a plane intersecting with the **X** space. The intersection of the regression function and the **X** space defines a separating hyperplane ($\alpha + \beta^T x = 0$) that separate the chosen segment into two sub-segments. The slope of the hyperplane is given by β while its location is determined by α. To satisfy the "meaningful" objective, we locate the hyperplane so that it passes through the center of the observations in the segment (Figure 2-2(d)). We accomplish this by dividing along $\alpha' + \beta^T x = 0$.

The primary difference between this approach and other binary tree methods (such as CH-AID) is that the hyperplanes need not be perpendicular to the axes of the X-space. Instead, we allow the hyperplanes to be linear functions of the axes. The use of a re-

sponse measure and the continuous nature of this response variable also distinguishes this approach from both clustering and classification methods. Clustering methods generally do not include a response variable. They focus on the representative objective without regard to the consistent criterion. Classification methods do use a response measure, but require that the response measure is binary or discrete. The response measure in our approach is the continuous variable $\tilde{V}^{\pi_H}(x)$.

2.6 Dynamic Programming

In this section, we summarize the dynamic programming methodology that is used in the computation. Recall that the firm's objective is to maximize its discounted aggregate profits. The optimal profit-to-go function values satisfy the famous Bellman's Equation (Bellman 1957 [5]), the objective function for this problem can be formulated as:

$$V(s) = \max_{\pi} E_{r_{s,\pi(s)},T,s'} \left[r_{s,\pi(s)} + \delta^T V(s') | s, \pi(s) \right] \quad \forall s \in S \tag{2.2}$$

Here we use notations $r_{s,a}$ as the random variable representing the immediate profit from the Markov chain after visiting state s and taking mailing action a, δ as the discount factor per unit time and T as the length of the inter mailing time period after visiting state s. Notice here that since inter mailing time periods are not always the same, T is in fact a random variable, which is interdependent with s.

For a fixed policy π, the following equation provides the expected discounted aggregate profits (profit-to-go) when starting at state $s \in S$:

$$V^{\pi}(s) = E_{r_{s,\pi(s)},T,s'} \left[r_{s,\pi(s)} + \delta^T V^{\pi} | s, \pi(s) \right] \quad \forall s \in S$$

If we use term $\bar{r}_{s,a}$ to represent the expected rewards earned in a period from a customer in state s when the firm chooses mailing action a, the above equation system (in general) can be expressed as:

$$
\begin{aligned}
V^{\pi}(s) &= \bar{r}_{s,\pi(s)} + E_{T,s'}\left[\delta^{T} V^{\pi}(s') | s, \pi(s)\right] \\
&= \bar{r}_{s,\pi(s)} + \sum_{T}\sum_{s'} \delta^{T} p_{s,\pi(s)\to T,s'} V^{\pi} \\
&= \bar{r}_{s,\pi(s)} + \sum_{s'} V^{\pi}(s') \sum_{T} \delta^{T} p_{s,\pi(s)\to T,s'}
\end{aligned}
$$

Here term $p_{s,\pi(s)\to T,s'}$ represents the probability that a customer in state s after the mailing action a will transition to state s' after time period T. In the computation, we can directly estimate $p_{s,s',a} := \sum_{T} \delta^{T} p_{s,a\to T,s'}$ from the data, which takes care of both the transition probability and the discounting.

With a slight modification of notation we can express the above equation in vector form. Let $\mathbf{P}^{\pi}$ denote a transition probability matrix for a given policy such that $\mathbf{P}^{\pi}_{i,j} = E\left[\delta^{T} p_{i,j,\pi(i)}\right]$, let $\bar{\mathbf{r}}^{\pi}$ denote the vector of expected rewards (with each element $\bar{r}_{i,\pi(i)}$), and let $\mathbf{v}^{\pi}$ denote the vector with elements $V^{\pi}(i)$. Given these notations we have: $\mathbf{v}^{\pi} = \bar{\mathbf{r}}^{\pi} + \mathbf{P}^{\pi}\mathbf{v}^{\pi}$, which yields $\mathbf{v}^{\pi} = \left(\mathbf{I} - \mathbf{P}^{\pi}\right)^{-1} \bar{\mathbf{r}}^{\pi}$ as the value function under policy π.

The stochastic historical policy π_H can be observed directly from the data and so we can directly compute the value function under this historical policy: $\mathbf{v}^{\pi_H} = \left(\mathbf{I} - \mathbf{P}^{\pi_H}\right)^{-1} \bar{\mathbf{r}}^{\pi_H}$.

Having v^{π_H} in hand, we use the classical policy iteration algorithm to compute the optimal mailing policy. The algorithm iterates between policy evaluation and policy improvement. In particular, the algorithm begins with a policy for which we calculate the profit-to-go function. We then use this profit-to-go function to improve the policy, which yields a new policy with which to begin the next iteration. The sequence of policies improves strictly monotonically until the current policy is optimal. It is well known that the policy iteration algorithm converge to a stationary policy that is optimal for the finite state infinite time horizon Markov Decision Process (Bertsekas 1995 [6]). In practice, the speed of convergence is surprisingly fast (Puteman 1994 [20]).

2.7 Implementation and Computational Results

2.7.1 Variables

We implemented the model on a dataset provided by a nationally distributed mail-order catalog company. The company is a medium sized firm that sells a range of durable products through separate divisions. In this study we focus on the women's apparel division. Apparel is one of the largest product categories sold through direct mail catalogs, representing between 40% and 50% of total household dollars spent on purchases from catalogs (DMA 2001 [2]). The women's apparel sold by this firm is in the moderate to high price range and almost all carry the company's own brand name. They are distributed through the company's own catalogs, and sold through both the company's own retail stores and some independent retailers.

We received data describing the purchasing and mailing history for approximately 1.8 million customers who had purchased at least one item of women's apparel. The purchase history data included each customer's entire purchase history. The mailing history data was complete for the six-year period from 1996 through 2002 (the company did not maintain a record of the mailing history prior to 1996). In this six-year period catalogs containing women's clothing were mailed on approximately 120 occasions, so that on average a mailing decision in this category occurred every 2-3 weeks. The company also mails catalogs for other product categories and the historical data received from the company contained a complete record of mailing and purchasing records for these other product categories.

The firm has historically used each customer's purchase history to make its mailing decisions but has not used the mailing history. It has also occasionally used two other data sources, although we will delay discussion of these data sources until a discussion of potential biases. The firm relies on a customized statistical model to predict the likelihood that a customer will respond to a catalog and, if they respond, the amount that they will purchase. It uses this model to make mailing decisions that maximize the expected response to a specific catalog (less mailing costs).

With the assistance of the catalog firm we identified a range of explanatory variables to describe each customer's mailing and purchase histories. Preliminary analysis of the data led to the inclusion of the following variables in the final model:

Women's Clothing Purchase History

Purchase Recency$_{it}$	Number of days since customer i's most recent purchase prior to period t.
Purchase Frequency$_{it}$	Number of orders placed by customer i prior to period t.
Monetary Value$_{it}$	Average size in dollars of orders placed by customer i prior to period t.
Monetary Value Stock$_{it}$	$p_{it} = \sum_{k \in K_{it}} \eta^{T_k} x_k$.
Customer Age$_{it}$	The number of days between period t and the customer i's first purchase.

Purchase History For Other Categories

NW Purchase Frequency$_{it}$	Number of orders placed by customer i prior to period t for items outside the women's clothing category.

Women's Clothing Mailing History

Mailing Frequency Stock$_{it}$	$m_{it} = \sum_{k \in K_{it}} \eta^{T_k}$.

We considered a variety of other variables describing customers' mailing and purchase histories from other product categories, but these variables had little effect on estimates of V or the optimal mailing policies. The *Monetary Value Stock*$_{it}$ and *Mailing Frequency Stock*$_{it}$ variables require that we specify values for the decay variables. In preliminary analysis we considered different values for these decay variables. This led to inclusion of two *Monetary Value Stock*$_{it}$ and *Mailing Frequency Stock*$_{it}$ variables with different decay rates. The decay rates for the *Monetary Value Stock*$_{it}$ variables were set at 0.9 and 0.8 per month, while for the *Mailing Frequency Stock*$_{it}$ variables the values were set at 0.9 and 0.8 per week. These values were chosen because they yielded greater variance in the optimal mailing policies (across different values of the stock variables). The final estimates of V were

relatively stable to different values for these decay rates.

Analysis of the raw data confirmed the presence of seasonality in both the purchasing and mailing histories. To capture seasonality in the purchase history we calculated the average number of orders received in each week of a calendar year (calculated across all of the years in the historical data). Because orders are received from a catalog for up to four months after the catalog is mailed, we calculated a weighted average of the number of orders received across subsequent weeks. In particular we weighted the subsequent weeks using data reported by the Direct Marketing Association (DMA 2001 [2]) describing the proportion of total orders that are received in each of the weeks after a catalog is mailed. The number of catalogs mailed affects the amount of revenue received and so we constructed a second seasonality measure to describe the variance in the historical mailing policy throughout a calendar year. It is calculated as a centered five week moving average of the number of catalogs mailed in corresponding weeks in the historical data.

Finally, we also included a third seasonality variable to capture the tendency amongst some customers to purchase at specific times of the year. In particular, we calculated the number of purchases made by a specific customer in the same quarter in previous years. We gave greater weight to more recent purchases by decaying prior purchases using an exponential weighting function (using a decay rate of 0.9). These three seasonality variables can be summarized as follows:

Seasonality	
Purchase Seasonality$_t$	The average number of orders received in the corresponding week across all years in the dataset.
Mailing Seasonality$_t$	The average number of catalogs mailed in the corresponding week across all years in the dataset.
Individual Seaonality$_{it}$	The discounted sum of the number of purchases by customer i in the same quarter in prior years.

We added one additional variable to control for the variation in the length of each mailing period. This variable was labeled Period Length$_t$ and was defined as the number of

weeks in mailing period t.

2.7.2 State Space

Having defined the vector space **X**, we discretized it using the approach described in Section 2.5. To simplify computation we focused on the transaction and purchase histories for a random sample of $100,000$ of the 1.8 million customers. The first year for which we had complete mailing and purchase history was 1994 and so we used data for this year (and prior years) to initialize the mailing and purchase stock measures. This estimation period comprised a total of 107 mailing periods, yielding approximately 9.5 million observations. An observation is defined as a specific mailing period for a specific customer (the missing observations result from customers whose first purchase occurred after 1994).

To obtain initial estimates of the value function for the historical policy ($\tilde{V}^{\pi_H}(x)$) we randomly selected a mailing period in 1995 for each of the $100,000$ customers and calculated the discounted profits earned from this customer in the subsequent four years. We focused on sales of the women's clothing division and so only considered mailing decisions and purchases from this category. The randomization process ensured that we obtained estimates of $\tilde{V}^{\pi_H}(x)$ for all values of the seasonality variables. Stochasticity in customers' purchase rates resulted in considerable variance in the resulting estimates and so we smoothed the estimates by regressing $\tilde{V}^{\pi_H}$ on a quadratic function of the explanatory variables. To ensure that the estimates were robust to the randomization process we repeated this process one hundred times and averaged the resulting parameter estimates to derive final estimates for $\tilde{V}^{\pi_H}$.

Having discretized the state space we used the dynamic programming methodology discussed in section 2.6 to estimate the value function for both the historical and optimal policies. The transition probabilities and expected rewards were calculated directly from the mailing and purchase histories using the same sample of $100,000$ customers. We again used 1994 to initialize the variables and so only considered mailing periods from 1995 on. The company supplements its purchase history data with additional information from other

Table 2.1: Average profit-to-go and mailing rates with discount rate

Monthly	Average Profit-to-Go($)		Mailing Rate	
Discount Rate	Historical	Optimal	Historical	Optimal
15%	11.64	13.52	58%	31%
10%	28.45	21.71	58%	43%
5%	37.39	48.23	58%	62%
3%	59.75	86.69	58%	71%
0.87%	159.17	343.22	58%	78%

sources to make mailing decisions for inactive customers (defined as customers who have not purchased within the last three years). Because we do not have access to this additional data, this introduces the potential for bias in the dynamic programming estimates (see later discussion). For this reason we focus on estimating the optimal policy for active customers by reverting to the actual mailing decision (reflecting the historical policy) for observations involving inactive customers. This restriction involved approximately 4.5 million of the 9.5 million observations in the training sample.

2.7.3 Results

For ease of exposition we will refer to the improved policy as the "optimal" policy. However, we caution that the optimality of the policy is conditional on the definition of the Markov decision problem, including the design of the discrete state space.

The optimal policy varies depending upon the rate at which future earnings are discounted. In Table 2.1, we report estimates of the historical and optimal policy profit-to-go for different discount rates in a state space with 500 states. The discount rates are monthly discount rates, with a rate of 0.87% corresponding to an annual discount rate of 10%. We restrict attention to active customers and weight the estimates for each state by the number of visits to each state in the training sample. The table also contains information about the mailing policy; we report the average percentage of (active) customers mailed a catalog in each mailing period.

There are several findings of interest. First, the average profit-to-go for the historical policy varies across discount rates. Although the policy does not vary, the rate at which future transactions are discounted affects the profit-to-go. Second, the average profit-to-go estimates for the optimal policy also increase with the discount rate. However, this variance reflects both the change in the rate at which future transactions are discounted and differences in the optimal policy. At higher discount rates it is optimal to mail a high proportion of customers because the model gives more weight to the favorable impact that mailing has on future purchasing.

At monthly discount rates higher than 10% the profit-to-go function for the optimal policy is similar to that of the current policy. At these high discount rates the objective function is relatively myopic, giving little weight to transactions that occur in later periods. The findings indicate that the improvement on the current policy is relatively small in these conditions. This is perhaps unsurprising given the myopic focus of the current policy and the extensive feedback that the firm receives about the immediate response to its mailing policies. However, as the discount rate decreases, so that more value is attributed to future earnings, the difference in the estimated profit-to-go functions increases.

In Figure 2-3 we graphically summarize the path of the profit flows under the different policies. In the figure we track the undiscounted profit earned from a sample of $100,000$ customers over 200 periods (almost eight years). The figure was constructed by drawing a random sample of $100,000$ customers from the 1.8 million customers in the database and starting the customers in the state they were in on January 1, 2001. We simulated the dynamic path of each customer using the transition probabilities and expected rewards.

For all policies the profits eventually decrease as a growing proportion of the $100,000$ customers become inactive. The rate at which this occurs varies across policies. In the optimal policies with lower discount rates the rate is slowed by more aggressive mailing policies in earlier periods. However, at the very start of the path, these more aggressive mailing policies yield lower profits than the historical policy. In these initial periods the firm is yet to realize the full benefits of its investments in additional mailings. This illustrates the

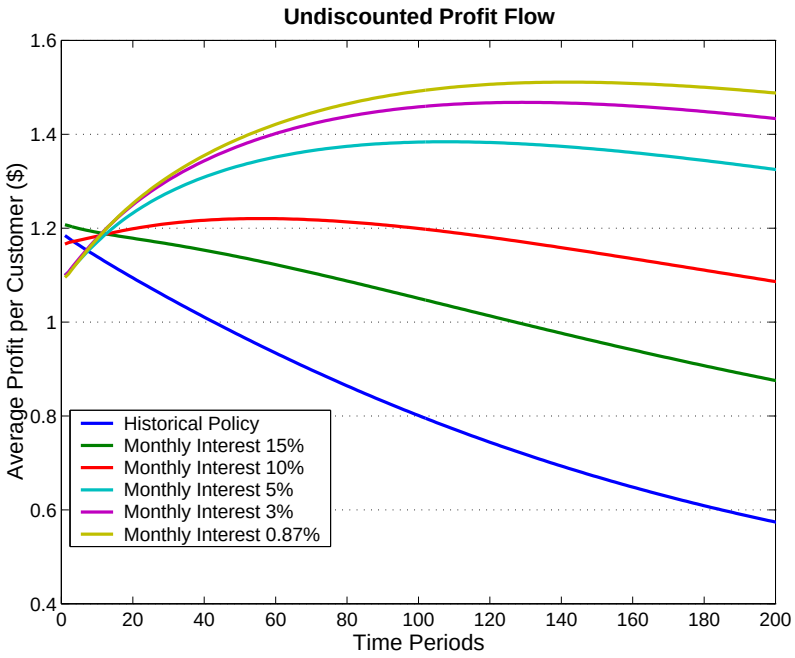

Figure 2-3: Average Profits (Undiscounted) Per Period

trade-off facing the firm. Investments in mailing catalogs cost the firm immediate profits but yield long-term payoffs. Varying the discount rates varies how the model resolves this tradeoff.

2.7.4 Comparison on Mailing Policies

We can further illustrate the difference between the current and optimal policies by comparing how the mailing rates vary as a function of the explanatory variables. In Figure 2-4 we report the proportion of times a catalog was mailed in the historical data (the current policy), together with the proportion mailed under the optimal policy, for different values of the *Purchase Recency* and *Mailing Stock* variables. The optimal policy in the figures uses a 3% per month discount rate and we only consider active customers. When customers become inactive (the recency measure exceeds 36 months) the optimal policy reverts to the current policy, which mails to approximately 14% of inactive customers.

Recall that *Purchase Recency* measures the number of months since the customers' last

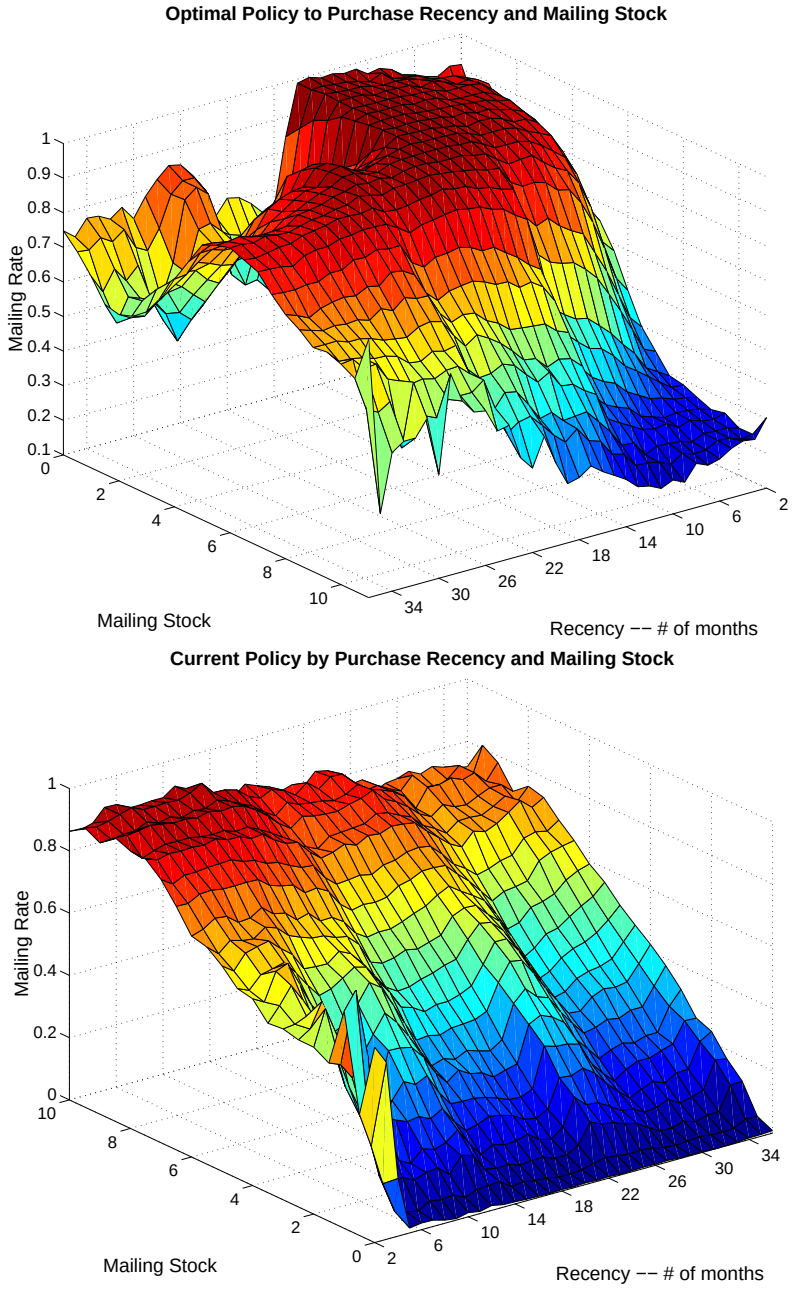

Figure 2-4: Optimal and Current Mailing Policies by Months Since Last Purchase

purchase, while the *Mailing Stock* measure is a discounted sum of the number of catalogs mailed to customers in previous periods (using a decay rate of 0.9 per week). A mailing stock measure of 10 indicates a customer who has recently been inundated with catalogs, while a mailing stock of 1 indicates a customer who has received very few recent catalogs. The figures reveal several important differences between the optimal and current policies:

1. The current policy figure confirms that the firm does not consider a customers' mailing history in its current policy. It keeps mailing to some customers even though they have received a lot of recent catalogs (high mailing stock), and does not mail to other customers even if they have not received any catalogs for a long time (low mailing stock).

2. The optimal policy mails more catalogs, particularly to customers who have not received many recent catalogs (low mailing stock). Although the company judges that there is a low probability that the customers will respond immediately (and hence does not mail to them), the optimal policy judges that mailing to these customers increases the probability of future orders.

3. The optimal policy does not mail to customers who have received a lot of recent catalogs (high mailing stock values). Note that after not mailing to these customers in one mailing period the mailing stock will drop (by approximately 2), so that many of them will receive a catalog in the next mailing period.

4. Mailing rates in the optimal policy are particularly low for customers who have recently purchased and who have received a lot of recent catalogs (low recency, high mailing stock). Discussions with the company revealed that the current policy of mailing to customers who have recently purchased is somewhat arbitrary, and is not always supported by their statistical response model. It appears that this arbitrary policy is not always optimal.

5. The small peak on the left hand side of the optimal policy figure indicates that this

policy mails to many customers who have not purchased for two years, particularly if these customers have not received many catalogs recently. This is consistent with customers purchasing in annual cycles. Interestingly, in the current policy figures, there is also a ridge indicating higher mailing rates if customers have not purchased for two years.

We caution that the values of the *Mailing Stock* variable reflect the mailing history under the current policy. As such, the figures answer the following question: how would the first mailing decision differ under the two policies? After implementing the optimal policy for several periods the shape of the figures would change because customers will have different values for the *Mailing Stock* variable. We also caution readers not to conclude that the optimal policy is stochastic. Observations that have the same value on one of the RFM variables generally have different values on other variables, so that the observations are distributed across multiple states. The policy within a state is deterministic, but when aggregating across states it appears stochastic.

Comparison of mailing policies for customers who have placed a larger number of prior orders (not shown) is also interesting. The firm's current policy is to mail often to these customers, while the optimal policy is to mail less often, particularly if the customer has received a lot of recent catalogs. One interpretation is that these customers are likely to respond even if they do not receive the current catalog. This highlights another important difference in the design of the two policies. Like other firms in the industry, this company designs its mailing policy to maximize the response to a specific catalog. They evaluate how many customers (in a given segment) ordered from catalogs mailed to them in the past, and then mail to all customers for whom the expected response justifies the incremental mailing cost. This treats the probability that a customer will respond if they are not mailed a catalog as zero. Although it may appear that receiving the current catalog is a necessary condition for making a purchase, this is not true in practice. A customer who is ready to purchase will often purchase from a previous catalog if they do not receive the most recent catalog. Measuring the response to a specific catalog ignores potential cannibalization from prior

catalogs. This problem does not arise in the proposed model, where we measure profits earned over time, irrespective of which catalog the response came from.

2.8 Conclusions

We have presented a model that seeks to improve catalog-mailing decisions by explicitly considering the dynamic implications of those decisions. The findings have important implications for the catalog industry. The industry's current focus on maximizing the immediate profits earned from the current catalog results in sub-optimal policies, in which firms systematically mail too few catalogs. Broadening the firm's objectives to also consider the long-term benefits of mailing decisions has the potential to greatly increase their long-run discounted profits. The findings also have important implications for the Operations Research literature. They demonstrate the feasibility of using state-of-the-art optimization techniques developed primarily for physical science applications to address important social science problems.

The application reveals two important sources of bias that have received little attention elsewhere in the literature. The first source of bias results from imprecision in the estimates of the outcomes from each action. The optimization algorithm exploits this imprecision by tending to choose actions for which the imprecision leads to an overly favorable estimate of the outcome. The result is inflated profit-to-go function estimates and potentially sub-optimal policies. We propose and implement a method for detecting this bias. The second source of bias results from the endogeneity of the historical policy in the training data. If some of the information used to design this policy is unavailable then the model may misattribute differences in outcomes to actions rather than to the unobserved information. Further discussions on these problems are in Chapter 3 and 4 of the thesis.

Both sources of bias result from breaches of assumptions that are implicit in the techniques. These assumptions have not previously received attention in the literature because they are rarely breached in the applications for which the techniques have previously been

used. However, in a social science setting, the assumptions both become relevant. We demonstrate that the issues are important, even where the sample of training data is very large.

Although the findings are promising, further research is required to determine whether the findings survive a test in the field. A large-scale field test of the proposed model is currently underway with a catalog retailer. The field test involves a total of $60,000$ customers, randomly assigned to Treatment and Control groups. Mailing decisions for customers in the Control group use the firm's current mailing policy, while mailing decisions for customers in the Treatment group use the proposed model. Current plans are for the test to continue for at least six months, at the end of which we will compare the trajectory of the profits earned in the two conditions, together with the final distribution of customers across states. The details of the field test are presented in Chapter 5 of the thesis.

Chapter 3

Hidden Information, the Endogeneity Problem and Batch Online Learning

3.1 Introduction

From Chapter 2, it is clear that each state $\tilde{s}$ in the discrete state space $\tilde{S}$ groups customers with similar historical information. However, potentially each state $\tilde{s}$ could be further divided more finely into many smaller states. In applications where we have to construct discrete state space through approximation, potentially the constructed state space is an aggregation of a "true" state space S consisting of many smaller states. In general we are not able to identify the true state space either because of its huge size or lack of necessary information.

The fact that each state $\tilde{s}$ is an aggregation of many true states in the state space S potentially results in an "endogeneity problem" — the computed optimal policy depends on the current policy from which data are collected. More precisely, the current policy determines how data is distributed in the true state space; the distribution of data in the true states determines the estimation of model parameters in the aggregated state space; and the optimal policy on the aggregated state space is computed according to the estimated model parameters.

One special case of this endogeneity problem is what we label in Section 3.2, "attribution error". When attribution error arises, we tend to reach a suboptimal policy but the profit-to-go estimation is biased upwards. In Section 3.2, we justify this claim in theory and illustrate the evidence in the catalog mailing problem.

Suppose the discrete state space construction is exogenous. We want to find a policy based on the aggregated state space whose profit-to-go performance is close to optimal. An intuitive approach is to iteratively improve the policy according to the current model parameter estimation and collect more data for improving estimates of the model parameters. We call such an iterative procedure a "batch online learning" procedure. Will a carefully designed "batch online learning" algorithm converge to something of interest? Before answering this question, we discuss a related question in this chapter — is there a self-enforcing policy such that data collected according to this policy provide no motivation for policy improvement (or stopping the batch online learning procedure)? In Section 3.3, we show the existence of self-enforcing policies. However, in general such policies are not unique. As shown in an example towards the end of the Chapter, multiple self-enforcing polices may lead to very different profit-to-go estimation. This might undermine the potential of the batch online learning idea.

In this chapter, we first present the attribution error problem through a numerical example and discuss the theoretical justification. Then we show the existence of self-enforcing policies motivated by the possibility that batch online learning procedures may help to resolve the endogeneity problem.

3.2 The Attribution Error

Preliminary analysis from the catalog mailing application identified a potential source of bias in the dynamic programming estimates. Intuitively the model estimates the profit-to-go function associated with mailing and not mailing by dividing the observations in each state into two samples based on the mailing decision for each observation. It then compares

the average outcome for the mailed sample with the not mailed sample. In this manner, the model learns from natural experiments in the data through variance in the historical mailing policy. An implicit assumption is that the customers in the mailing versus no mailing samples within each state are statistically identical. If this is not the case then there is the potential for error.

The risk of this occurring is high when the firm uses data to determine its mailing policy that is hidden from the dynamic optimization model. For example, the firm may use data to identify which customers are most likely to respond, and only mail to those customers. If this data is hidden from the model, then the model may incorrectly infer that mailing to the customers who were not mailed would have led to the same outcome as that observed for the customers who were mailed.

It is helpful to illustrate this problem using an example. Assume for the moment that there are two groups of customers. Mailing to a customer in the first group yields a profit of $\$100$, while not mailing yields a profit of $\$0$. For the second group, mailing yields a profit of -$\$10$ and not mailing yields a profit of $0. If the firm can observe which group a customer is in it will mail to customers in the first group, but not mail to customers in the second group. Now assume that the model does not have access to the same information and so treats all customers as members of the same group. The model observes that the firm earned $\$100$ when mailing to customers from this group and $\$0$ when not mailing and so recommends mailing to everyone in the group. In doing so, the model incorrectly attributes the effect of the hidden information to the mailing action.

The outcome is an upward distortion in the profit-to-go estimates ($\mathbf{v}$) and sub-optimal policies. Moreover, this outcome is not limited to the states affected by the hidden information. The dynamic programming algorithm propagates the upward bias to other states that transition to the problematic states, and so the entire state space is potentially affected.

This example is motivated by actual findings observed in preliminary analysis of the data. The optimal mailing policy initially recommended mailing to almost all inactive customers. Discussions with the firm revealed that it uses two additional sources of in-

formation to decide which inactive customers to mail to. These two data sources include purchases from competing catalogs and the appearance of these customers on mailing lists rented to identify prospective customers. This preliminary analysis suggests that these two data sources are effective at discriminating between the firm's inactive customers. However, neither information sources was available when building this model. It is for this reason that we focus in this analysis on active customers. Observations for inactive customers remain in the model as the outcome for these customers affects the value function for active customers (who eventually become inactive). However, when a customer becomes inactive, the model continues to implement the historical mailing policy and does not attempt to improve upon this policy.

In what follows, we first present a toy example to show how serious the attribution error can potentially be. Then we provide theoretical justification for the upward bias that results.

3.2.1 A Toy Example

Assume the state space of a Markov process is $S = \{1, 2, 3, 4\}$ and at each state there are two possible actions $U = \{0, 1\}$. The transition probability matrix from each state-action pair to another state is

$$P = \begin{bmatrix} .25 & .25 & .25 & .25 \\ .01 & .49 & .01 & .49 \\ .05 & .45 & .05 & .45 \\ .01 & .49 & .01 & .49 \\ .25 & .25 & .25 & .25 \\ .01 & .49 & .01 & .49 \\ .05 & .45 & .05 & .45 \\ .01 & .49 & .01 & .49 \end{bmatrix}.$$

$P(j|i, a)$, the transition probability from state i taking action a to state j, is represented as component $P_{i*2+a-1,j}$ of the above matrix P. The expected immediate reward out of

state-action pair (i, a) to state j is represented as component $R_{i*2+a-1,j}$ of the following matrix

$$R = \begin{bmatrix} 3 & -1 & 3 & -1 \\ 4 & 0 & 4 & 0 \\ 3 & -1 & 3 & -1 \\ 4 & 0 & 4 & 0 \\ 3 & -1 & 3 & -1 \\ 4 & 0 & 4 & 0 \\ 3 & -1 & 3 & -1 \\ 4 & 0 & 4 & 0 \end{bmatrix}.$$

After running an infinite horizon DP with decay rate $\alpha = 0.9$, we get the optimal policy $\pi^*(1) = \pi^*(3) = 1$ and $\pi^*(2) = \pi^*(4) = 0$ with the corresponding optimal profit-to-go to be

$$J^* = \begin{bmatrix} 2.7113 \\ 1.0915 \\ 2.7113 \\ 1.0915 \end{bmatrix}.$$

Notice the above model can be treated as a simplified "catalog mailing problem" – for example if we mail to a customer at state 1, with probability $.25$, the customer will purchase, go to state 3 and bring $\$3$ profit, etc.

Now suppose historically the catalog mailer was able to observe all 4 states and did follow the optimal mailing action. In the case that later we are not able to differentiate between states 1 and 2, nor can we differentiate between state 3 and 4, we are only able to build a DP model with two states $S' = \{1', 2'\}$ and use the transition probability model and reward given as $P' = \begin{bmatrix} .5 & .5 \\ .02 & .98 \\ .5 & .5 \\ .02 & .98 \end{bmatrix}$ and $r' = \begin{bmatrix} 3 & -1 \\ 4 & 0 \\ 3 & -1 \\ 4 & 0 \end{bmatrix}$. In this case the resulting "optimal" policy would be $\pi'^*(1) = \pi'^*(2) = 1$ and the "optimal" profit-to-go is $J'^* = 10$

for both states.

The toy example presented here is not an individual case. In general, attribution errors tend to cause the upward biased profit-to-go estimation. Next we provide some theoretical justifications for it.

3.2.2 Theoretical justifications

First, we introduce some notations for this section and the rest of the chapter.

Notations

Assume nature knows the true state space $S : |S| = m$. Out of each state $s \in S$, we can take an action a out of an action set $U : |U| = n$. Also suppose that we can only observe data from an aggregated state space $\tilde{S} : |\tilde{S}| = \tilde{m} < m$ such that state $\tilde{s} \in \tilde{S}$ is the aggregation of a subset $S_{\tilde{S}} \subset S$ of the states in the original state space S. In what follows, the tilde sign is used to represent parameters in the aggregated state space $\tilde{S}$.

Assume historical data were generated according to some policy π. Under this policy, the distribution of data on the state space S follows a probability distribution $p : p \in [0,1]^m, p^T e = 1$. Denote the $m \times m$ transition probability matrix to be P and immediate reward out of each state to be an m dimensional vector g. In the aggregated state space $\tilde{S}$, we will evaluate the $\tilde{m} \times \tilde{m}$ transition probability matrix $\tilde{P}$ and the reward vector $\tilde{g}$.

We use notations J to express the profit-to-go vector in the original state space such that

$$J = (I_m - \alpha P)^{-1} g.$$

In order to express parameters in the aggregated state space, i.e., $\tilde{P}$ and $\tilde{g}$, we need to define two linear transformation matrices – $m \times \tilde{m}$ matrix A and $\tilde{m} \times m$ matrix $B(p)$. Having these two matrices, we can link the model parameters as well as the profit-to-go estimation in the aggregated state space with the parameters in the original state space.

Matrix A is a $0-1$ matrix. All the elements $A_{s,\tilde{s}}$'s are 1 if $s \in S_{\tilde{s}}$, i.e., state s is one

of the sub-states to be aggregated in to $\tilde{s}$. Other elements in matrix A are all 0's. In other words, each row of matrix A has one element 1 and all other elements 0 as in the following:

$$A = \begin{bmatrix} 1 & & & & & \\ 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & \vdots & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & 1 & \end{bmatrix}$$

The structure of matrix $B(p)$ is similar to that of matrix A in the sense that each column has at most one positive element ($B_{\tilde{s},s}$ for $s \in S_{\tilde{s}}$) and all other elements are 0's. The values of the positive elements in $B(p)$ depend on a probability distribution p in the true state space S following which data are distributed. In other words, element $B(p)_{s,\tilde{s}} = \frac{p_s}{\sum_{s' \in S_{\tilde{s}}} p_{s'}}$ if $s \in S_{\tilde{s}}$. The following matrix is an example of $B(p)$:

$$B(p) = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{p_i}{\sum_{j=i}^{i'} p_j} & \cdots & \frac{p_{i'}}{\sum_{j=i}^{i'} p_j} & \\ & & & & & & \cdots \\ & & & & & \frac{p_{n-1}}{\sum_{j=i''}^{n} p_j} & \frac{p_n}{\sum_{j=i''}^{n} p_j} \end{bmatrix}$$

There are some helpful properties of matrix A and $B(p)$ that we can exploit. For example, their product $B(p)A$ equals the $\tilde{m}$ dimensional identity matrix $I_{\tilde{m}}$, and p is an eigenvector of matrix $AB(p)$, i.e., $p^T AB(p) = p^T, \ \forall p$.

If sample trajectories are collected from the original state space according to a transition

probability matrix P and data distributes in the state space S according to a probability distribution p, the corresponding observed aggregated transition probability matrix $\tilde{P}$ is $\tilde{P} = B(p)PA$ and the aggregated cost vector $\tilde{g} = B(p)g$.

We can also express the profit-to-go estimation in the aggregated state space, vector $\tilde{J}$ to be

$$\begin{aligned} \tilde{J} &= \left(I_{\tilde{m}} - \alpha\tilde{P}\right)^{-1}\tilde{g} \\ &= \left(I_{\tilde{m}} - \alpha B(p)PA\right)^{-1}B(p)g \end{aligned}$$

Since the matrices A and $B(p)$ only depend on the structure of the state space aggregation and the distribution of data on the original state space, here we express all the model parameters and profit-to-go estimation in the aggregated state space using parameters in the original state space.

We use e_k to express k-dimensional vector of 1's. When the context is clear, we sometimes omit the subscription k.

Results

Using the notations from above, it is very easy to establish the following lemma:

Lemma 1 *$\tilde{p} := A^T p$ is the steady state probability of a Markov process with transition probability matrix $\tilde{P} := B(p)PA$.*

Proof. *Obviously $\tilde{p} = A^T p \geq 0$ and $\tilde{p}^T e_{n-1} = p^T A e_{n-1} = p^T e_n = 1$. We also have*

$$\tilde{p}^T\tilde{P} = p^T AB(p)PA = p^T PA = p^T A = \tilde{p}.$$

■

Lemma 1 shows that if the historical data distributed on the original state space S according to the steady state probability p, the observed distribution of data $A^T p$ on the aggregated state space matches the steady state probability according to the observed aggregated

transition probability matrix.

The following proposition shows that the average profit-to-go weighted by the steady state probabilities, in both the original and the aggregated state spaces, are the same.

Proposition 2 $p^T J = \tilde{p}^T \tilde{J}$

Proof. *Since* $p^T J = p^T g + \alpha p^T P g + \alpha^2 p^T P^2 g + ... = p^T g + \alpha p^T g + \alpha^2 p^T g + ... = \frac{1}{1-\alpha} p^T g$ *and* $\tilde{p}^T \tilde{J} := \tilde{p}^T \tilde{g} + \alpha \tilde{p}^T \tilde{P} \tilde{g} + \alpha^2 \tilde{p}^T \tilde{P}^2 \tilde{g} + ... = \tilde{p}^T \tilde{g} + \alpha \tilde{p}^T \tilde{g} + \alpha^2 \tilde{p}^T \tilde{g} + ... = \frac{1}{1-\alpha} \tilde{p}^T \tilde{g}$ *we just need to show* $p^T g = \tilde{p}^T \tilde{g}$*, which naturally follows from the fact that* $\tilde{p}^T \tilde{g} = p^T AB(p) g = p^T g$. ■

Now suppose that in the state space aggregation, two particular states $i, j \in S$ are aggregated into state $\tilde{i} \in \tilde{S}$, i.e., $S_{\tilde{i}} = \{i, j\}$. Also assume that according to the historical policy π that was used to generate the historical data, actions taken on states i and j are different, i.e., $\pi(i) \neq \pi j$. Now consider two other policies on the original state space S, $\pi_{(i)}$ and $\pi_{(j)}$ such that

$$\pi_{(i)}(k) = \pi(k) \; \forall k \neq j \text{ and } \pi_{(i)}(j) = \pi(i)$$

while

$$\pi_{(j)}(k) = \pi(k) \; \forall k \neq i \text{ and } \pi_{(j)}(i) = j$$

In other words, we consider policies in the aggregated state space $\tilde{S}$ such that a unique action is taken on state $\tilde{i}$.

Let $\tilde{P}_{(i)}$ and $\tilde{P}_{(j)}$ be the "evaluated" transition probability matrices in the aggregated state space following policies $\pi_{(i)}$ and $\pi_{(j)}$. Notice the "evaluation" is conducted according to historical data that were collected according to policy π. In this sense, matrices $\tilde{P}_{(i)}$, $\tilde{P}_{(j)}$ and $\tilde{P}$ only differ in the row representing the transition probability out of the aggregated state $\tilde{i}$. And matrix $\tilde{P}$ is a convex combination of matrices $\tilde{P}_{(i)}$ and $\tilde{P}_{(j)}$, i.e., $\tilde{P} = \frac{p_i}{p_i + p_j} \tilde{P}_{(i)} + \frac{p_j}{p_i + p_j} \tilde{P}_{(j)}$.

According to the evaluated model parameters $\tilde{P}_{(i)}$ and $\tilde{P}_{(j)}$ for the two policies $\pi_{(i)}$ and $\pi_{(j)}$, we define $\tilde{J}_{(i)}$ and $\tilde{J}_{(j)}$ to be the respective computed profit-to-go. The following

proposition shows that dynamic programming algorithm would always improve the profit-to-go by converting the historical policy on the aggregated state $\tilde{i}$ to be either one of the historical actions taken on sub-components i or j. However, this improvement cannot be justified via implementing the corresponding policy — it only appears in "computation" according to the evaluated model parameters.

Proposition 3 *Either $\tilde{J}_{(i)} \geq \tilde{J}$ or $\tilde{J}_{(j)} \geq \tilde{J}$. If $\tilde{P}_{(i)}\tilde{J} \neq \tilde{P}_{(j)}\tilde{J}$ then the above inequalities hold strictly for some components.*

Proof. *Obviously we have either*

$$g + \alpha\tilde{P}_{(i)}\tilde{J} \geq g + \alpha\tilde{P}\tilde{J}$$

or

$$g + \alpha\tilde{P}_{(j)}\tilde{J} \geq g + \alpha\tilde{P}\tilde{J}$$

Without loss of generality, we assume that

$$g + \alpha\tilde{P}_{(i)}\tilde{J} \geq g + \alpha\tilde{P}\tilde{J}$$

and define $J^1 := g + \alpha\tilde{P}_{(i)}\tilde{J}$ thus $J^1 \geq \tilde{J}$. If we define $J^{t+1} := g + \alpha\tilde{P}_{(i)}J^{(t)}$ for $t = 1, 2, ...$, we get an increasing and converging sequence $\{J^t\}$ and $\tilde{J}_{(i)} = J^\infty \geq \tilde{J}$

If $\tilde{P}_{(i)}\tilde{J} \neq \tilde{P}_{(j)}\tilde{J}$, it follows naturally that the strict inequality hold for some components. ■

We do not have access to the true model parameters for action $\pi(j)$ out of state i as well as action $\pi(i)$ out of state j. Thus we are not able to calculate the "true" profit-to-go following policies $\pi_{(i)}$ or $\pi_{(j)}$, either in the original state space S or in $\tilde{S}$. Without prior knowledge, the policy improvement step of the dynamic programming algorithm would lead towards an upward biased profit-to-go estimation.

3.2.3 Mitigating the attribution error in solving the dynamic catalog mailing problem

In this subsection, we describe the approach that we took to mitigate the attribution errors in the catalog mailing problem.

As mentioned in Chapter 2, the catalog mailing company used additional hidden information sources to help differentiate inactive customers (customers who had not purchased for more than 3 years). This information identifies inactive customers that have a higher potential to purchase. This means that in the historical data, inactive customers who received a catalog are in fact different from those who did not, even though they were in the same state (with similar historical information). According to the discussion above, we would tend to get an upward biased average profit-to-go estimation. Indeed, the reason that we were able to identify this problem was because originally we obtained unrealistically promising profit-to-go estimation and a mailing policy that we should mail to almost every inactive customer.

Because we do not have all of the information required to evaluate actions for inactive customers, we are not able to make mailing decisions for them. However we cannot eliminate inactive customers from the data. We thus tried to minimize the attribution error by only making decisions for the active customers and allowing decisions for inactive customers to follow the company's current mailing policy. In other words, when we set up the Markov decision process, the actions out of states corresponding to inactive customers are fixed. The dynamic programming algorithm only optimizes actions taken out of active customer states. This way, the final optimal policy from the DP algorithm only works on the active customer states and the profit-to-go estimation for all states are correct.

3.3 Fixed Points to Batch Online Learning Procedures

In the catalog mailing application, the state space aggregation is not limited to inactive customers. Because the state space is set up as a summary of customers' historical infor-

mation, each discrete state consists of many small "true" states. Even if historical data for active customers contains information on all actions out of each hidden state, there are still potential problems in the profit-to-go estimation and the optimal policy. This is because the model parameter estimation on the aggregated state space depends on how data are distributed on the hidden states. We describe this as an "endogeneity problem" because the historical policy affects the optimal policy by affecting the distribution of data on the hidden states and the resulting model parameter estimates.

As indicated earlier in the thesis, one way to address the problem is through "batch online learning". The procedure iterates the data collection, the model parameter estimation and the policy update steps. We call it a "batch" procedure because in each iteration, we obtain the optimal policy according to the most recent model parameter estimates.

If we conduct a batch online learning procedure, we can potentially fall into a loop — if we collect data according to some policy π_1, the estimated model parameters might indicate that another policy π_2 is optimal; however if we implement policy π_2 and collect data, the model parameters might indicate that π_1 is optimal. Naturally we are interested in knowing if there exist some "fixed point policies" such that data collected following the policy indicate that the policy is optimal, thus providing no motivation for continuing the batch online learning procedure. Rather than furthering the investigation on specific algorithms to update the model parameter estimation and policy, in this section, we focus on potential "final products" of the batch online learning procedure.

Some previous work has focused on the existence (or non-existence) of fixed points for different approximate dynamic programming algorithms. For example, de Farias and Van Roy 2000 [10] discussed the non-existence for a version of the "approximate value iteration algorithm" and showed that fixed points exist for an exploration version of the approximate value iteration algorithm that uses a δ-greedy policy. Gordon 2000 [15] shows that two popular reinforcement learning algorithms with function approximation, SARSA(0) and V(0), oscillate within a bounded region if not converging to a point. Unlike previous work on the convergence issues on specific algorithms, in this chapter, we discuss the existence

of a fixed point policy that satisfies some "optimality" conditions so that it provides no motivation for furthering the policy improvement. The idea of the self-enforcing policies is similar to network equilibrium, see, for example Florian and Hearn 1995 [13]. If we consider randomized stationary policies, we show that self enforcing fixed point policies do exist. However, they are not unique and multiple fixed points may perform quite differently, as illustrated by an example.

3.3.1 One Aggregated State

To make the analysis easier to explain, in this subsection we assume that $\tilde{m} = 1$. In other words, the m states in S are aggregated into one single state. In this setting, a policy in the aggregated state $\tilde{S}$ takes the same action (or combination of actions) on all the m states in the original state space S. In the following we show that there exists a randomized stationary policy such that data generated under the steady state probability of this policy provides no motivation for further policy improvement.

For a Markov decision process with m states and n actions, again we define probability distribution $p \in [0,1]^m, p^T e = 1$ to be the probabilities among the states that data were collected. Define a (randomized) policy $\lambda \in [0,1]^n, \lambda^T e = 1$ to be a probability distribution, with λ_a representing the probability that action a is taken.

Assume nature knows all the information about transition probabilities and immediate rewards in the original state space S. For each action a, denote $m \times m$ matrices P_a to be the transition probability matrix such that entry $P_a(s, s')$ is the probability of the system transiting to state s' out of state-action pair (s, a). Similarly denote m dimensional vector g_a such that entry $g_a(s)$ represents the immediate reward out of state-action pair (s, a). Obviously, under policy λ, the transition probability matrix and immediate reward for the

system are

$$\begin{aligned} P_\lambda &:= \sum_a \lambda_a P_a \\ g_\lambda &:= \sum_a \lambda_a g_a \end{aligned}$$

For transition probability matrix P_λ, we use notation p_λ to express its steady state probability distribution, i.e.,

$$p_\lambda^T = p_\lambda^T P_\lambda.$$

Notice since $\tilde{S}$ has only one state, the profit-to-go estimation is just $\frac{g}{1-\alpha}$. In the rest of this subsection, when we need to compare the profit-to-go estimation, we only need to consider the immediate reward g.

Equilibrium-Optimality Condition

We want to find a policy λ^* such that data generated according to its steady state probability p_{λ^*} self-enforce the optimality of policy λ^*. More formally, we define index sets $I_\lambda := \{a|\lambda_a > 0, a \in U\}$ to represent the set of actions that are taken according to policy λ^* (active actions) and $\bar{I}_\lambda := \{a|\lambda_a = 0, a \in U\}$ to represent the set of inactive actions. A self-enforcing policy has to satisfy the following *"equilibrium-optimality" condition*:

$$\begin{aligned} & p_{\lambda^*} g_a = p_{\lambda^*} g_{a'}, \ \ \forall a, a' \in I_{\lambda^*}, \\ & p_{\lambda^*} g_a \geq p_{\lambda^*} g_{a'}, \ \ \forall a \in I_{\lambda^*}, a' \in \bar{I}_{\lambda^*} \text{ and} \\ & \lambda^{*T} e = 1 \end{aligned}$$

The above condition requires that under the policy λ^*, if we collect data according to the current steady state probability distribution, the active actions have the same immediate reward estimation (so that there is no motivation for changing the probability each active action is taken) and this immediate reward estimation is higher than that of the inactive actions (so that there is no motivation to bring inactive actions to be active).

We can express the above *"equilibrium-optimality" condition* equivalently as the following:

E-O Condition

$$(p_{\lambda^*} g_a - g^*)\, \lambda_a^* = 0, \ \forall a = 1, \ldots, n \tag{3.1}$$

$$p_{\lambda^*} g_a - g^* \leq 0, \ \forall a = 1, \ldots, n \tag{3.2}$$

$$\lambda^{*T} e = 1 \tag{3.3}$$

Apparently, possible scalar value g^* belongs to a closed compact set $[\underline{g}, \bar{g}]$ that depends on the values in vectors g_a. Without loss of generality, we assume $0 \notin [\underline{g}, \bar{g}]$. This condition can be satisfied, e.g., by increasing each element in g_a by the same amount.

Variational Inequality formulation

In order to show there exists a λ^* that satisfies the above E-O Condition, we present the following Variational Inequality (VI) formulation and show that it is equivalent to the E-O Condition. The existence of a solution to the VI is easy to establish, which implies that self-enforcing policies exist.

First we want to assume that each P_a is an irreducible matrix. This means each action correspond to a recurrent Markov chain. Obviously P_λ is also irreducible for any policy λ.

According to the famous Perron-Frobenius Theorem, P_λ has a unique eigenvector p_λ according to its largest eigenvalue 1. This implies that p_λ is a continuous function of λ. The technical details for the following remark are shown in the Appendix A.

Remark 4 *Under the assumption that P_a is irreducible for all $a = 1, \ldots, n$, the eigenvector p_λ, determined by $p_\lambda^T = p_\lambda^T P_\lambda$, is a unique continuous function of λ.*

In order to formulate the variational inequality, we define a mapping $F : [0,1]^m \times$

$[\underline{g}, \bar{g}] \to R^{m+1}$ to be such that

$$F(\lambda, g) := \begin{pmatrix} | \\ p_\lambda^T g_a - g \\ | \\ \lambda^T e - 1 \end{pmatrix} \tag{3.4}$$

Now we present the following *Variational Inequality Formulation*:

VI Formulation: $F(\lambda^*, g^*) \cdot ((\lambda^*, g^*) - (\lambda, g)) \geq 0, \ \ \forall (\lambda, g) \in [0,1]^m \times [\underline{g}, \bar{g}]$

In the following we prove the VI formulation is equivalent to the E-O Condition.

Proposition 5 *Any λ^*, g^* satisfying the E-O Condition also satisfies the VI Formulation; on the other hand, if there exists λ^*, g^* satisfying the VI Formulation, the E-O Condition also has a solution.*

Proof. *First, we prove that if λ^*, g^* satisfies the E-O Condition it also satisfies the VI Formulation.*

According to (3.1), $\left(p_{\lambda^*}^T g_a - g^*\right) \lambda_a^* = 0$. For any $\lambda \in [0,1]^m$ and $g \in [\underline{g}, \bar{g}]$, according to (3.2), $\left(p_{\lambda^*}^T g_a - g^*\right) \lambda_a \leq 0$ and $\left(e^T \lambda^* - 1\right) g = 0$. Thus VI formulation holds.

Next, we assume that λ^*, g^* satisfies the VI formulation.

If we take $\lambda = \lambda^*$ and $g = g^* \pm \varepsilon$ in which $\varepsilon > 0$ is small enough such that $g \in [\underline{g}, \bar{g}]$, obviously (3.3) holds. To prove (3.1) and (3.2) hold, in the following we set $g = g^*$.

For $a \in \bar{I}$, i.e., $\lambda_a^* = 0$, (3.1) holds obviously. Take $\lambda = \lambda^* + e_a$ in which e_a is a n dimensional vector with all components 0 except the a^{th} component to be 1. VI implies that $-\left(p_{\lambda^*}^T g_a - g^*\right) \geq 0$ and thus (3.2) holds.

For $a \in I$ and $0 < \lambda_a^* < 1$, take $\lambda^1 = \lambda^* + \varepsilon e_a$ and $\lambda^2 = \lambda^* - \varepsilon e_a$. Here scalar $\varepsilon > 0$ is small enough such that both λ^1 and λ^2 belongs to $[0,1]$. VI implies that $\pm\left(p_{\lambda^*}^T g_a - g^*\right) \varepsilon \geq 0$ and thus $p_{\lambda^*}^T g_a - g^* = 0$. Thus both (3.1) and (3.2) hold. Thus in the case that $\nexists a$ such that $\lambda^* = 1$, (λ^*, g^*) satisfies the E-O Condition.

In the case that there exists a such that $\lambda_a^* = 1$, since we just proved that (3.3) holds, all other $a' \neq a$ belongs to $\bar{I}$, i.e., $\lambda_{a'}^* = 0$. If we take $\lambda = 0$, the VI Formulation implies that $p_{\lambda^*} g_a - g^* \geq 0$ while $p_{\lambda^*} g_{a'} - g^* \leq 0, \ \forall a' \neq a$. Thus $p_{\lambda^*} g_{a'} \leq g^* \leq p_{\lambda^*}, \ \forall a' \neq a$. If we define $\hat{g}^* = p_{\lambda^*} g_a$ and without loss of generality assume that $\hat{g}^* \in [\underline{g}, \bar{g}]$, it is obvious that $(\lambda^*, \hat{g}^*)$ satisfy (3.1) and (3.2). and thus is a solution to the E-O Condition. ■

Since the mapping F is continuous and is defined on a compact set, we know that the VI Formulation has a solution.

Theorem 6 *There exists a policy λ^* such that the E-O Condition (3.1 - 3.2) holds.*

Now we have shown that there exists a self-enforcing policy λ^* in the case that all m states are aggregated into a single state. Next, we show that the result holds more generally.

3.3.2 Multiple States

Now assume that nature knows the transition probabilities in S. Because of the state space aggregation, what matters to us are those transition probability matrices for each state $\tilde{s}$ and action a. Similar to matrices P_a in the last subsection, we use $m \times m$ matrices $P_{\tilde{s},a}$'s to represent the transition probabilities out of state-action pair $s \in S_{\tilde{s}}$ and a. Thus $P_{\tilde{s},a}$ has non-negative entries only in rows corresponding to $s \in S_{\tilde{s}}$.

We also assume that nature knows the immediate rewards in the original state space S. We use m dimensional vector $g_{\tilde{s},a}$ to represent the immediate reward out of each state $s \in S_{\tilde{s}}$ when an action a is taken. Notice here only entries corresponding to $s \in S_{\tilde{s}}$ in $g_{\tilde{s},a}$ are potentially non-zeros.

Now we define a randomized-policy on the $\tilde{S}$ state space. We express such a policy using an $\tilde{m} \times n$ matrix λ, with each entry $\lambda_{\tilde{s},a}$ representing the probability that action a is taken out of state $\tilde{s}$. Given policy λ, the transition probability matrix in the original state space S is

$$P_\lambda := \sum_a P_{\tilde{s},a} \lambda_{\tilde{s},a} \tag{3.5}$$

and the immediate reward vector is

$$g_\lambda := \sum_a g_{\tilde{s},a}\lambda_{\tilde{s},a} \tag{3.6}$$

which are both linear functions of λ.

Under a policy λ, if data are collected according to the steady state probability $p(\lambda)$: $p(\lambda)^T P_\lambda = p(\lambda)^T$, we define the aggregation matrix $B_\lambda := B\left(p(\lambda)\right)$. The observed transition probability matrix in state space $\tilde{S}$ is thus $\tilde{P} = B_\lambda P_\lambda A$ and immediate reward is $\tilde{g} = B_\lambda g_\lambda$. The corresponding profit-to-go vector in the aggregated state space $\tilde{S}$ is thus

$$\tilde{J}_\lambda := \left(I - \alpha\tilde{P}_\lambda\right)^{-1}\tilde{g} = \left(I - \alpha B_\lambda P_\lambda A\right)^{-1} B_\lambda g_\lambda = B_\lambda g_\lambda + \alpha B_\lambda P_\lambda A\tilde{J}_\lambda. \tag{3.7}$$

After collecting data under the steady state probability distribution of policy λ for a while, the policy evaluation part of the dynamic programming algorithm would evaluate profit-to-go out of each state-action pair and improve the policy based upon the following profit-to-go function:

$$\bar{J}_\lambda(\tilde{s},a) := B_\lambda^{(\tilde{s})} g_{\tilde{s},a} + \alpha B_\lambda^{(\tilde{s})} P_{\tilde{s},a} A\tilde{J}_\lambda \tag{3.8}$$

in which $B_\lambda^{(\tilde{s})}$ is the $\tilde{s}^{\text{th}}$ row of matrix B_λ.

We want to discuss the possibility of the existence of such a policy λ^* that there is no motivation for a policy improvement step to change it according to $\bar{J}$. In other words, if for each state $\tilde{s}$ we define index sets $I_\lambda^{\tilde{s}} \subset U$ to express the set of active actions according to a policy λ and $\bar{I}_\lambda^{\tilde{s}} \subset U$ the set of inactive actions, the following condition is necessary for the existence of the policy λ^*

$$\begin{aligned}
&\bar{J}_{\lambda^*}(\tilde{s},a) = \bar{J}_{\lambda^*}(\tilde{s},a'), \quad \forall a,a' \in I_{\lambda^*}^{\tilde{s}}, \forall \tilde{s} \in \tilde{S} \\
&\bar{J}_{\lambda^*}(\tilde{s},a) \geq \bar{J}_{\lambda^*}(\tilde{s},a'), \quad \forall a \in I_{\lambda^*}^{\tilde{s}}, a' \in \bar{I}_{\lambda^*}^{\tilde{s}}, \forall \tilde{s} \in \tilde{S} \text{ and} \\
&\lambda^{*T} e_n = e_{\tilde{m}}
\end{aligned}$$

Similar to the self-enforcing policy discussed in the one aggregated state case, a policy λ^* satisfying the above condition means

1. For each state, the actions that are taken in λ^* have the same estimated profit-to-go — there is no motivation to change the probability that each action is taken;

2. For each state, actions that are taken have higher profit-to-go estimation then actions that are not taken — there is no motivation to bring in other actions.

If we define the value of the profit-to-go for the active actions out of each state $\tilde{s}$ to be $\hat{J}_{\tilde{s}}$, the above *"equilibrium-optimality" condition* can be express equivalently as the following:

E-O Condition

$$\left(\bar{J}_{\lambda^*}(\tilde{s},a) - \hat{J}^*_{\tilde{s}}\right)\lambda^*_{\tilde{s},a} = 0 \ \ \forall \tilde{s} \in \tilde{S}, a \in U \tag{3.9}$$

$$\bar{J}_{\lambda^*}(\tilde{s},a) - \hat{J}^*_{\tilde{s}} \leq 0, \ \ \forall \tilde{s} \in \tilde{S}, a \in U \tag{3.10}$$

$$\lambda^{*T} e_n = e_{\tilde{m}} \tag{3.11}$$

Apparently possible $\hat{J}^{(\tilde{s})}$ values are bounded with upper and lower bound L and U determined by the values in $P_{\tilde{s},a}$ and $g_{\tilde{s},a}$. Without loss of generality, we can also assume that $0 \notin [L,U]$.

We define mapping $F : [0,1]^{\tilde{m}\times n} \times [L,U]^{\tilde{m}} \to R^{\tilde{m}^2 n}$ defined as the following.

The first $\tilde{m} \times n$ components of F are as the following:

$$F_{\tilde{s},a}(\lambda, \hat{J}) := \bar{J}_{\lambda}(\tilde{s},a) - \hat{J}^{(\tilde{s})} \ \ \forall \tilde{s} \in \tilde{S}, a \in U \tag{3.12}$$

The last $\tilde{m}$ components of F are as the following

$$F_{\tilde{s}}(\lambda, \hat{J}) := \sum_a \lambda_{\tilde{s},a} - 1 \ \ \forall \tilde{s} \in \tilde{S} \tag{3.13}$$

Having the definition of the mapping F and use notation $x^* := (\lambda^*, \hat{J}^*)$, we can formulate the following *Variational Inequality* formulation.

VI Formulation: $F(x^*) \cdot (x^* - x) \geq 0$ for all $x \in [0,1]^{\tilde{m} \times n} \times [L,U]^{\tilde{m}}$.

Similar to the last subsection, we have

Proposition 7 *If $(\lambda^*, \hat{J}^*)$ satisfies the E-O Condition, it also satisfies the VI Formulation and $\hat{J}^* = \tilde{J}_{\lambda^*}$. On the other hand, if $(\lambda^*, \hat{J}^*)$ satisfies the VI Formulation, $(\lambda^*, \tilde{J}_{\lambda^*})$ satisfies the E-O Condition.*

Proof. *The idea of the proof is very similar to the proof of Proposition 5. Detailed proof is in the Appendix 3.4 in the end of the Chapter.* ■

Under the assumption that for any policy λ, transition probability matrix $P_\lambda = \sum_{\tilde{s},a} P_{\tilde{s},a} \lambda_{\tilde{s},a}$ is irreducible, we have the following lemma

Lemma 8 *$F(\lambda, \hat{J})$ is continuous on λ and $\hat{J}$.*

Proof. *This follows naturally from the Perron-Frobenius Theorem with same idea as in the proof of Remark 4.* ■

Thus we have the existence of solution $(\lambda^*, \hat{J}^*)$ that satisfies the E-O Condition.

Theorem 9 *There exists a policy λ^* and its corresponding profit-to-go $\tilde{J}_{\lambda^*}$ that satisfies the E-O Condition and the VI Formulation.*

So far we have shown that self-enforcing policies do exist if we are only able to observe and evaluate an aggregated state space. Notice we have not presented any algorithms leading towards such self-enforcing policies. In the next subsection, we present an example showing that the self-enforcing policy is not unique.

3.3.3 Multiple Fixed Points

In this section, we present a case where there exists multiple self-enforcing policies and these policies lead to very different profit-to-go estimations.

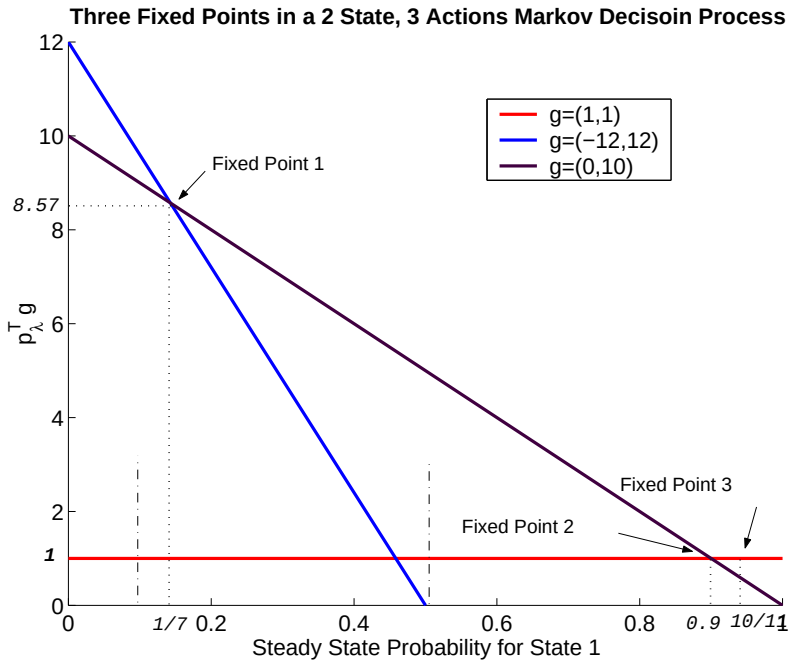

Figure 3-1: Three Fixed Points in the Aggregation of a 2 State, 3 Actions Markov Decision Process: The x-axis shows the probability p on state 1. Each line in the figure represents the aggregated immediate reward $p_\lambda g$ for one action.

Now assume we have Markov decision process with a two state $S = \{1, 2\}$ and three actions $U = \{1, 2, 3\}$. The state space is aggregated into one state. Following the notations in Section 3.3.1,

$$g_1 = (1, 1); g_2 = (-12, 12); g_3 = (0, 10)$$

$$P_1 = \begin{bmatrix} \frac{19}{20} & \frac{1}{20} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}; P_3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{14} & \frac{13}{14} \end{bmatrix}.$$

There are three self-enforcing policies. Table 3.1 summarizes these policies.

Figure 3-1 shows the three self-enforcing policies' steady state probabilities and immediate rewards. The $x-$axis is the steady state probability of state 1. Each one of the three fixed points in the Figure indicates the steady state probabilities as well as the immediate reward estimation under each self-enforcing policy. Details of each self-enforcing policy are shown in Table 3.1.

Table 3.1: Three self-enforcing policies

	Policy_1	Policy_2	Policy_3
λ	$\left(0, \frac{1}{36}, \frac{35}{36}\right)$	$\left(\frac{620}{627}, 0, \frac{7}{627}\right)$	$(1, 0, 0)$
p_λ	$\left(\frac{1}{7}, \frac{6}{7}\right)$	$\left(\frac{9}{10}, \frac{1}{10}\right)$	$\left(\frac{10}{11}, \frac{1}{11}\right)$
g_λ	$\left(1, \frac{60}{7}, \frac{60}{7}\right)$	$\left(1, -\frac{48}{5}, 1\right)$	$\left(1, -\frac{108}{11}, \frac{1}{11}\right)$

From Table 3.1, we see that according to Policy_1, actions 2 and 3 are taken and the aggregated immediate reward is $\frac{60}{7}$. According to Policy_2 and Policy_3, in which actions $1, 3$ and 1 are taken respectively, the aggregated immediate reward estimation are both 1. The immediate reward estimation following Policy_1 is almost eight times more than that of Policy_2 and Policy_3. This example shows that although self-enforcing policies do exist, they should not be the ultimate goal of an algorithm as their profit-to-go performances may be very different. Rather, we should try to design batch online learning procedures that approach "the best" self-enforcing policy if possible, or do not pursue a self-enforcing policy as long as.

3.3.4 Further Discussions

The current problem setting is that we only have access to an aggregated state space. Ideally, we want to find a policy on the aggregated state space $\tilde{S}$ such that the profit-to-go $\tilde{J}$ is maximized. In other words, ideally we want to solve the following problem

$$\lambda^* := \arg\max_{\lambda} \left(I - \alpha B(p_\lambda) P_\lambda A\right)^{-1} B(p_\lambda) g_\lambda \tag{3.14}$$

However, in the current problem setting we are not able to solve the above optimization problem because it requires knowledge of the true transition probability matrix P_a in the original state space. Even if we assume that P_a is available, the above optimization problem is still hard to solve because p_λ, the steady state probability of P_λ, is a nonlinear function of λ, which makes the objective function highly nonlinear (neither concave nor convex).

It is a conjecture that following results in [18], we may show that finding the above λ^* is NP-complete.

This is why we turn to analyze self-enforcing policies, whose optimality can be justified by the data collected following the policy. So far in this section, we have discussed the existence of stationary randomized self-enforcing policies. However, the self-enforcing policies discussed above also have other issues that need further investigation.

The first issue is a stability problem. According to the definition of self-enforcing policies, if the system is set up according to the steady state probability of some self-enforcing policy λ_s, and we collect some data according to another policy λ', the model parameter estimation would indicate that the policy λ' does not lead to a better profit-to-go estimation. However, if the system is set up according to the steady state probability of policy λ' and we collect some data according to this policy, it is not clear that the model parameter estimation will indicate that moving towards a self-enforcing policy is the direction to improve the policy. In order for the batch online learning algorithm to lead towards a self-enforcing policy λ_s, we need the policy to be at least locally "stable", i.e., if we implement a perturbed policy λ around λ_s, we need the data to indicate that the direction $(\lambda_s - \lambda)$ is a policy improvement direction. This problem is not addressed in this thesis.

The second problem is about multiple self-enforcing policies and their relationship. What determines the performances of different self-enforcing policies remains a problem.

The third issue is about the "best possible" policies. Although we claim that we are not able to obtain the best possible policy λ^* as defined in (3.14), it is interesting to consider whether we can establish some relationship between the profit-to-go under a self-enforcing policy and a best possible policy.

The fourth question is the most fundamental one. We have not proposed an algorithm for finding self-enforcing policies. The research in this section is to further the understanding of batch online learning algorithms. The ultimate goal is to design and understand learning algorithms from continuously cumulated data.

3.4 Appendix: Proof of Proposition 7

Lemma 10 *If $0 \leq \lambda^* \leq 1$ (component wise) and $\hat{J}^* \in [L, U]^{\tilde{m}}$ satisfy the E-O Condition, it also satisfies the VI Formulation*

Proof. *Denote $x^* := (\lambda^*, \hat{J}^*)$. First, we show that $F(x^*) \cdot x^* = 0$. Notice*

$$F(x^*) \cdot x^* = \sum_{\tilde{s},a} F_{\tilde{s},a}(\lambda^*, \hat{J}^*)\lambda^*_{\tilde{s},a} + \sum_{\tilde{s}} F_{\tilde{s}}(\lambda^*, \hat{J}^*)\hat{J}^*_{\tilde{s}}$$

in which $F_{\tilde{s},a}$ and $F_{\tilde{s}}$ are defined as (3.12) and (3.13).

From (3.9), it is obvious that $\sum_{\tilde{s},a} F_{\tilde{s},a}(\lambda^*, \hat{J}^*)\lambda^*_{\tilde{s},a} = 0$. According to (3.11), obviously $\sum_{\tilde{s}} F_{\tilde{s}}(\lambda^*, \hat{J}^*)\hat{J}^*_{\tilde{s}} = 0$.

For any $\lambda \geq 0$ and $\hat{J} \in [L, U]^{\tilde{m}}$, according to (3.10), $F_{\tilde{s},a}(\lambda^*, \hat{J}^*) \leq 0$ thus $F_{\tilde{s},a}(\lambda^*, \hat{J}^*)\lambda_{\tilde{s},a} \leq 0$. And we have $F_{\tilde{s}}(\lambda^*, \hat{J}^*) = 0$ thus $F_{\tilde{s}}(\lambda^*, \hat{J}^*)\hat{J}_{\tilde{s}}$. It is then obvious that

$$\begin{aligned} & F(x^*) \cdot (x^* - x) \\ = \; & F(x^*) \cdot x^* - F(x^*) \cdot x \\ = \; & 0 - \sum_{\tilde{s},a} F_{\tilde{s},a}(\lambda^*, \hat{J}^*)\lambda_{\tilde{s},a} - \sum_{\tilde{s}} F_{\tilde{s}}(\lambda^*, \hat{J}^*)\hat{J}_{\tilde{s}} \\ \geq \; & 0 \end{aligned}$$

■

Lemma 11 *If $0 \leq \lambda^* \leq 1$ (component wise) and $\hat{J}^* \in [L, U]^{\tilde{m}}$ satisfy the VI Formulation, the E-O Condition holds.*

Proof. *First, we show that (3.11) holds with λ^*. This is because if we take $\lambda = \lambda*$, $\hat{J}_{\tilde{s}} = \hat{J}^*_{\tilde{s}} \pm \varepsilon$ and $\hat{J}_{\tilde{s}'} = \hat{J}^*_{\tilde{s}'}\ \forall \tilde{s}' \neq \tilde{s}$ in which $\varepsilon > 0$ is small enough such that $\hat{J}_{\tilde{s}} \in [L, U]$. According to the VI Formulation, we have that $\sum_a \lambda^*_{\tilde{s},a} - 1 = 0$ and this is true for every $\tilde{s} \in \tilde{S}$. Thus (3.11) holds.*

From now on we set $\hat{J} = \hat{J}^*$ and establish that (3.9) and (3.10) hold for a fixed $\tilde{s}$.

For $a \in \bar{I}^{\tilde{s}}_{\lambda^*}$, i.e., $\lambda^*_{\tilde{s},a} = 0$, (3.9) holds. Set λ to be the same as λ^* except one component

$\lambda_{\tilde{s},a} = \varepsilon$ *in which value* $\varepsilon \in (0,1)$. *According to the VI Formulation,* $-\varepsilon F_{\tilde{s},a}(\lambda, \hat{J}) \geq 0$ *thus (3.10) holds.*

For $a \in I^{\tilde{s}}_{\lambda^*}$ *and* $0 < \lambda^*_{\tilde{s},a} < 1$, *take* λ^1 *and* λ^2 *to be the same as* λ^* *except the* $\tilde{s}, a$ *component. Set* $\lambda^1_{\tilde{s},a} = \lambda^*_{\tilde{s},a} + \varepsilon$ *and* $\lambda^2_{\tilde{s},a} = \lambda^*_{\tilde{s},a} - \varepsilon$. *Here scalar* $\varepsilon > 0$ *is small enough such that both* $\lambda^1_{\tilde{s},a}$ *and* $\lambda^2_{\tilde{s},a}$ *are between* 0 *and* 1. *The VI Formulation implies that* $\pm\varepsilon F_{\tilde{s},a}(\lambda, \hat{J}) \geq 0$ *thus* $F_{\tilde{s},a}(\lambda, \hat{J}) = 0$. *So (3.9) and (3.10) both hold.*

In the case that there exist a $\lambda^*_{\tilde{s},a} = 1$, *since we just showed that (3.11) holds for* $\tilde{s}$, $\lambda^*_{\tilde{s},a'} = 0,\ \forall a' \neq a$. *If we take* λ *such that it is the same as* λ^* *except* $\lambda_{\tilde{s},a} = 0$, *the VI Formulation implies that* $\bar{J}_\lambda(\tilde{s}, a) - \hat{J}^{(\tilde{s})} \geq 0$ *while* $\bar{J}_\lambda(\tilde{s}, a') - \hat{J}^{(\tilde{s})} \leq 0\ \forall a' \neq a$. *Thus* $\bar{J}_\lambda(\tilde{s}, a') \leq \hat{J}^{(\tilde{s})} \leq \bar{J}_\lambda(\tilde{s}, a)\ \forall a' \neq a$. *If we take* $\hat{J}'^{(\tilde{s})} = \bar{J}_\lambda(\tilde{s}, a)$ *and without loss of generality assume that* $L \leq \hat{J}'^{(\tilde{s})} \leq U$, *we know that* $(\lambda^*, \hat{J}'^{(\tilde{s})})$ *satisfies (3.9) and (3.10)*

Since the above proof holds for any $\tilde{s}$, *we have shown that if VI has a solution* $(\lambda^*, \hat{J}^*)$, *an solution can be constructed for the E-O Condition as well.* ■

Lemma 12 *If* $(\lambda^*, \hat{J}^*)$ *satisfies the E-O Condition,* $\hat{J}^* = \tilde{J}_{\lambda^*}$.

Proof. *(3.9) implies that for a fixed* $\tilde{s}$,

$$\hat{J}^*_{\tilde{s}} = \bar{J}_{\lambda^*}(\tilde{s}, a)\ \ \forall a \in I^{\tilde{s}}_{\lambda^*}$$

Use notation $\tilde{J}^{(\tilde{s})}_{\lambda^*}$ *to express the* $\tilde{s}^{th}$ *component of* $\tilde{J}_{\lambda^*}$. *According to (3.7) and (3.8), the above equation implies that*

$$\begin{aligned}
\tilde{J}^{(\tilde{s})}_{\lambda^*} &= B^{(\tilde{s})}_{\lambda^*} g_{\lambda^*} + \alpha B^{(\tilde{s})}_{\lambda^*} P_{\lambda^*} A \tilde{J}_{\lambda^*} \\
&= \sum_a \bar{J}_{\lambda^*}(\tilde{s}, a) \lambda^*_{\tilde{s},a} \\
&= \sum_{a \in I^{\tilde{s}}_{\lambda^*}} \bar{J}_{\lambda^*}(\tilde{s}, a) \lambda^*_{\tilde{s},a} \\
&= \bar{J}_{\lambda^*}(\tilde{s}, a) \\
&= \hat{J}^*_{\tilde{s}}
\end{aligned}$$

■

According to the above lemmas, Proposition 7 naturally follows.

Chapter 4

Effects of Random Noise in Model Parameter Estimation

4.1 Introduction

In this chapter, we present another source of bias when constructing a dynamic programming model using off policy sample trajectories. First, we present empirical evidence from the catalog mailing problem. Then we provide a theoretical investigation of the sources of these biases.

4.2 Empirical Evidence

4.2.1 Varying the Number of States

In Table 4.1 we report the profit-to-go estimates from the model described in Chapter 2 when varying the number of states. The profit-to-go estimates for the historical policy are relatively invariant to the number of discrete states used. However, the estimates for the optimal policy increase monotonically with the number of states. There are two possible explanations for this phenomenon: one favorable and the other unfavorable. The favorable

Table 4.1: Average profit-to-go with discount rates and the cardinalities of state spaces

Monthly Discount Rate	Optimal Profit-to-Go ($) Number of States		
	500	1000	2000
15%	13.52	14.02	14.52
10%	21.71	22.47	23.28
5%	48.23	49.88	51.42
3%	86.69	90.00	92.86
0.87%	343.22	362.22	377.568

explanation is that classifying the observations more finely by using a larger number of states offers additional degrees of freedom with which to optimize. The models with fewer states are nested versions of the larger models, and therefore represent more restricted optimizations.

The unfavorable interpretation is that the optimization step in the dynamic programming algorithm exploits stochasticity in the training data. Recall that the transition probabilities and expected rewards are calculated directly from the data. The optimization algorithm chooses actions to maximize future discounted returns. This favors actions for which the errors in the expected rewards are positive and errors in the transition probabilities lead towards more valuable states. The outcome is an upwards bias in the profit-to-go estimates. The potential for bias is stronger when estimates of the transition probabilities and expected returns are less precise. This will tend to occur when the fixed sample of training observations are distributed across a larger number of states.

We can test for this bias by re-estimating the profit-to-go function for the optimal policy derived from one data set on a separate sample of data. The stochastic errors should vary across datasets, so that evaluating a policy designed using one dataset on a second data set should offer an unbiased estimate of the profit-to-go function for that policy. Following notations in Chapter 2, we took the optimal policy from the calibration data set and calculated the transition probabilities ($\mathbf{P}$) and expected returns ($\bar{\mathbf{r}}$) under this policy for a validation dataset. The profit-to-go for the validation data set is then given by: $\mathbf{v} = (\mathbf{I} - \mathbf{P})^{-1} \bar{\mathbf{r}}$.

Table 4.2: Average Profit-to-Go Estimates from a Separate Validation Sample by Discount Rate and Number of States

Monthly Discount Rate	Average Profit-to-Go ($) Optimality 500	Optimality 1000	Optimality 2000	Validation 500	Validation 1000	Validation 2000	Historical
15%	13.52	14.02	14.52	12.11	12.08	11.92	11.18
10%	21.71	22.47	23.28	19.29	19.26	19.07	17.75
5%	48.23	49.88	51.42	42.17	42.11	41.68	36.07
3%	86.69	90.00	92.86	75.18	75.38	75.15	57.73
0.87%	343.22	362.22	377.57	271.83	272.96	277.27	153.77

In Table 4.2 we report the validated profit-to-go function evaluated on a random sample of $100,000$ customers. This second validation sample was drawn randomly from the remaining 1.7 million customers in the original sample after removing the $100,000$ customers used to design the policies. As a benchmark we also report the profit-to-go estimates under the current policy for the validation data set (these estimates were almost invariant to the number of states and so we report the estimates for 500 states).

In Figure 4-1 we show the profit-to-go estimates for the optimal policy on both datasets when varying the number of states. When we increase the number of states, the profit-to-go estimates from the calibration dataset monotonically increase while the estimates from the validation dataset hold steadily and even slightly decreases.

The findings reveal little evidence that using $2,000$ states rather than 500 states yields a more valuable policy. We conclude that the increase in the profit-to-go function estimates when there are more states (Table 4.1) appears to reflect bias due to the DP algorithm taking advantage of stochastic errors in the data. If the increases were due to additional degrees of freedom we would expect them to survive when re-evaluating the policy on a different sample of data.

The observation that stochasticity can bias the profit-to-go function has received little attention in the literature. This illustration confirms that the issue is not just of theoretical interest, but may also have practical importance. We were able to detect the phenomenon despite the very large sample used in this application. Of the 9.5 million observations in

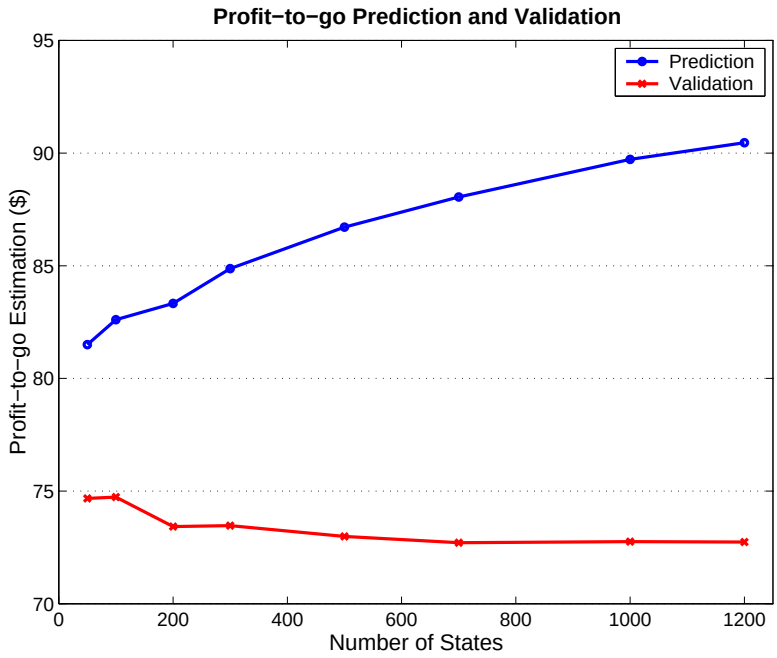

Figure 4-1: Predicted versus Validated Profit-to-go Estimation

the training sample, just under 5 million of them involved active customers. Distributing these observations across $2,000$ states yields an average of $2,500$ observations per state, or $1,250$ observations for either "mail" or "not mail" actions in each state.

In the next section, we further investigate the effects of inaccuracy in model parameter estimation.

4.3 Inaccuracy in Parameter Estimation

Because the parameters in the DP models are estimated from data, random noise in the model parameter estimates is inevitable. If we assume that the "true" immediate reward is a vector g and the "true transition probability matrix is P, the observed reward vector and transition probability matrix are $g + \Delta g$ and $P + \Delta P$, in which Δg and ΔP are zero mean random vector and matrix representing the estimation error. If we consider the number of states in the finite Markov chain to be n, the number of parameters to be estimated is in the

order of n^2.

4.3.1 Δg

In this subsection, we discuss the effect of estimation error in g, assuming we have a perfect estimate of the transition probability model P. We show that although the estimation error has zero mean Δg, the effect in the computed DP profit-to-go is that the computed DP profit-to-go is upward biased.

We define $m := \sum_{k \in S} |A_k|$ to be the total number of state-action pairs and $n := |S|$ the size of the state space and $m \times n$ matrix

$$E := \begin{bmatrix} 1 & & & & \\ 1 & & & & \\ & 1 & & & \\ & 1 & & & \\ & 1 & & & \\ & & \cdots & & \\ & & & & 1 \end{bmatrix}.$$

It is well known (see, e.g., Bertsekas 1995 [6]) that the linear optimization formulation for the discounted infinite time horizon DP is

$$\begin{aligned} (P)\ Z = \max_{J} \quad & c^T J \\ \text{s.t.} \quad & EJ \quad \leq g + \alpha P J \end{aligned}$$

in which the objective function value is a convex combination of J values according to the probability distribution c.

Assume Δg is component wise zero mean. From linear programming sensitivity anal-

ysis, $Z(\Delta g)$ defined as

$$(P_{\Delta g})\ Z(\Delta g) := \max_{J} \quad c^T J$$
$$\text{s.t.} \quad (E - \alpha P)J \ \leq g$$

is a convex function of Δg. From Jensen's inequality, we know that

Proposition 13 $E_{\Delta g}\left[Z(\Delta g)\right] \geq Z$

If we use J^* to express the optimal solution to (P) and $J^*(\Delta g)$ to $(P(\Delta g))$, we can show that the above effect holds component-wise such that

Proposition 14 $E_{\Delta g}\left[J^*(\Delta g)\right] \geq J^*$

Proof. *Define $\tilde{g} := g + \Delta g$ and series*

$$\begin{aligned} J_s^0 &:= \max_a\{g_{s,a}\} \\ J_s^{k+1} &:= \max_a\{g_{s,a} + \alpha P_{s,a} J^k\} \ \ \forall k = 0, 1, ... \\ \tilde{J}_s^0 &:= \max_a\{\tilde{g}_{s,a}\} \\ \tilde{J}_s^{k+1} &:= \max_a\{\tilde{g}_{s,a} + \alpha P_{s,a} \tilde{J}^k\} \ \ \forall k = 0, 1, ... \end{aligned}$$

Obviously, the value iteration algorithm generates either $\{J^k\}$ or $\{\tilde{J}^k\}$, depending on the model. It is also well known that J^k (and $\tilde{J}^k$) converges to the optimal profit-to-go as $k \to \infty$.

Since $E\left[\tilde{g}\right] = g$, it is obvious that $E\left[\tilde{J}^0\right] \geq J^0$. If $E\left[\tilde{J}^k\right] \geq J^k$ and a^* is the optimal action in $\tilde{J}_s^{k+1}$, we have

$$E\left[\tilde{J}_s^{k+1}\right] - J_s^{k+1} \geq \alpha P_{s,a^*}\left(E\left[\tilde{J}^k\right] - J^k\right) \geq 0$$

■

The above propositions show that when estimation error exists for the immediate reward g, the effect on the profit-to-go estimation for each state is upward biased.

4.3.2 ΔP

In this subsection we assume the estimation of g is accurate and consider the effect of ΔP.

The analysis on ΔP is more complicated. We are not able to draw an "upward bias" conclusion as in the Δg case. Here is an example. Assume there are two policies available. One policy has transition probability matrix $\begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}$ and with probability 0.5 perturb to $\begin{bmatrix} 0.5 & 0.5 \\ 0.6 & 0.4 \end{bmatrix}$ or $\begin{bmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{bmatrix}$. The discount rate $\alpha = 0.9$ and the immediate reward out of the two states are $(3, 5)$. The nominal profit-to-go is $(39.15, 40.85)$ and the mean profit-to-go with respect to the perturbation on the transition probability is $(39.09, 40.79)$, which is lower.

Suppose there is no perturbation in the parameters for the other policy and the profit-go-go is far away enough from the first policy so that there is no cross-over effect. Then the effect of the parameter perturbation is downward bias.

In short, the "upward bias" (or Jensen's Inequality) effect happens when random noise affects the policy improvement step. Imprecision in the transition probabilities may have complicated effects on the profit-to-go for a fixed policy, because the function

$$J(\Delta P) := \left[I - \alpha \left(P + \Delta P\right)^{-1}\right] g \tag{4.1}$$

is neither a convex, nor a concave function of ΔP. Notice that deflating the profit-to-go estimates using a validation sample, as in Table 4.1, does not resolve errors resulting from the non-linearity of this function.

In the case that ΔP is small enough so that $o(\Delta P)$ is negligible, (4.1) can be linearized to

$$J(\Delta P) = \left(I - \alpha P\right)^{-1} g + \alpha \left(I - \alpha P\right)^{-2} \Delta P g \; . \tag{4.2}$$

In this case

$$E_{\Delta P}\left[J(\Delta P)\right] = J(0) \; .$$

Thus the effect of ΔP is reduced to the Δg case, which is upward biased.

4.4 Conclusion Remarks

Initially it seems surprising that the model parameter estimation error poses a problem given the volume of data that we used. The analysis in this chapter indicates that the number of model parameters to be estimated is not negligible and the effect tends to be upward biased. Further analysis needs to be done on bounding the profit-to-go estimation error as a function of the model parameter estimation error. The ultimate goal is to design robust dynamic programming procedures that overcome the estimation error problem.

Chapter 5

Field Test

5.1 Introduction

The goal of this research is to develop a model for catalog mailing companies to make the right dynamic mailing decisions using their business data. We need to test the practicability of the proposed approach in a field test. As discussed previously in the thesis, the model parameters are estimated from historical data. Errors and limitations in the historical data may distort the model and the mailing policy. For example, the following issues cannot be fully resolved using historical data.

1. *The endogeneity problem caused by hidden information.* Because we do not have access to the hidden state information, historical data are not enough to evaluate the distribution of data within a state and the change of model parameters under a different policy. In a field test, because we collect data under the new policy, we are able to directly observe the changes and their effects.

2. *Model parameter estimation error.* As discussed in Chapter 4, when we use a hold out sample from the historical data to evaluate the new policy, we are able to overcome the upward bias problem caused by Jensen's Inequality. However, we cannot overcome the errors that result from non-linearities in (4.1).

3. *Non-Stationarity.* We estimated the model parameters using a group of customers' data in the past 6 years. Using data from a long historical period introduces a potential problem. The overall economic environment could have changed so dramatically that the model parameter estimation does not accurately reflect the current (and future) reality any more. An alternative way of sampling historical data is to sample a larger number of customers over a shorter period of time. We did not follow this alternative approach because of potential over fitting problems. However, a field test provides an effective way of testing whether the model works in a changing environment.

In short, through a field test, we want to check whether the proposed approach is robust.

In this chapter, we describe the procedure used to design the field test and present some preliminary empirical results and analysis. Because the data is not yet all available, the empirical results and analysis focus on checking the accuracy of the model parameters, profit-to-go estimates and the distribution of the data.

5.2 Experiment Design

The field test extends over a 6-month period, during which our model makes mailing decisions for a group of the mailing company's customers. In this section, we describe the customer samples and the implementation process.

5.2.1 Before the Field Test Starts

Customer sample

We began by randomly selecting $60,000$ customers whose *purchase recency* was less than 2 and half years. This guarantees that these customers would be active during the entire field test time period.

The sample of $60,000$ customers was randomly assigned to 2 groups of equal size. Our

model makes mailing decisions for one group of customers (*the treatment group*) while the company continues to make mailing decisions for the other group of customers (*the control group*).

Before the field test started, we obtained the whole transaction history and mailing history for the treatment group.

Our Model

In order to obtain the optimal mailing policy, we constructed the model following the procedures described in Chapter 2, Section 2.7, using $100,000$ customers. There are two issues worth mentioning at this point. The first issue is "lead time". The second is the initial model parameters.

Mailing decisions are made about 8 weeks before the scheduled "in house dates" (the date on which customers receive the catalogs) to provide time for an outsourced publisher to print and distribute the catalog books. Thus, the transaction history for the mailing decision is only available 56 days prior to the in house date. In other words, in the model described in Chapter 2, we use transaction history information 56 days before the mailing time period to generate the n variables formulating the $\mathbf{X}$ space. The lead time issue is not a problem for the mailing history and seasonality variables. This is because we know all of the previous mailing decisions.

In the computational results shown in Chapter 2 and Chapter 4, we used different discount rates δ (as in equation 2.2). After discussing this issue with the catalog mailing company, we realized that the catalog managers' view of long term risk is different than what is suggested by the capital markets. Indeed, their primary concern is annual profit. Trading off the above concern and the long term profit maximization goal, we agreed to use a 3% monthly interest rate, which is approximately a 30% annual interest rate. As for the number of states in the state space, we need to trade off degrees of freedom versus parameter estimation accuracy. After discussions with the company we used 500 states for active customers and 500 states for inactive customers.

Table 5.1: Mailing Dates

Decision Date	Mailing Date
15-Nov-2002	10-Jan-2003
29-Nov-2002	24-Jan-2003
13-Dec-2002	7-Feb-2003
27-Dec-2002	21-Feb-2003
17-Jan-2003	7-Mar-2003
31-Jan-2003	21-Mar-2003
14-Feb-2003	4-Apr-2003
28-Feb-2003	18-Apr-2003
14-Mar-2003	2-May-2003
28-Mar-2003	16-May-2003
18-Apr-2003	6-Jun-2003
9-May-2003	27-Jun-2003

After constructing the state space and running the dynamic programming algorithm, we obtain a lookup table, which tells us which action to take at each state. In the field test, at each mailing period we map each customer to a specific state according to his/her historical information. The lookup table then reveals the mailing action for each customer.

5.2.2 Decision Process

Over the 6-month period, there are 12 female clothing catalog mailings. The inter-mailing period varies from 2 weeks to 4 weeks. Table 5.1 lists the mailing dates (in-house dates) as well as the decision dates. In general, each decision date is about 8 weeks before the mailing date.

At each mailing period, we cumulate the transaction and mailing history information for the customers in the treatment group in order to map them into states.

Transaction history information

Every Monday we receive from the company the complete transaction history for the $30,000$ customers in the treatment group. The reason we require the complete transac-

tion history is because a “de-duplication” procedure is conducted frequently to identify duplicate accounts for individual customers. Customer “de-duplication” is not conducted on the mailing history files, as mailing information is not used in the company’s decision making.

Mailing history information

As indicated earlier, before the field test starts, we have access to the past 6 years of mailing history for each customer. After the field test starts, we update the mailing history for the treatment group by recording our own mailing decisions.

5.2.3 Ideal Tests

Below, we list the comparisons that could be conducted for the purpose of checking the robustness of the model.

1. Profit

   The predicted un-discounted profit flow is predicted in Figure 2-3. It is important to see if the profit flow initially goes up over time in the treatment group.

2. Policy

   We do not expect the aggregated mailing rate on the treatment group to follow the exact predictions in Chapter 2. This is due to variations in customer distributions over the period of the test. We will analyze the changes in the policy over time and compare the effects of different mailing costs in the treatment and control groups.

3. Distribution of customers

   Because of the problems mentioned in Section 5.1, the distributions of customers following the two mailing policies will not exactly follow what is predicted in the model. It is important to monitor the changing distributions of customers in the treatment and control groups to gain insights into the effects of different policies. At

the end of the field test, we are interested in observing which policy positions more customers in "favorable" states, measured by profit-to-go estimates.

4. Profit-to-go estimation

   It is important to know how accurate the profit-to-go estimates are. Profit-to-go is measured over an infinite time horizon. However, we can observe how well the Bellman Equation is fitted as a proxy for testing the profit-to-go estimates.

5. Model parameter estimation

   The field test provides a good opportunity to understand the performance of the model in an "on-policy" scheme. Because of the endogeneity problem and the non-stationary problem, it is understandable that the observed parameters P and g will vary from the estimates in the model. In the field test we want to compare the model parameters over time between the control and treatment groups. This comparison will provide insights into the effects of different policies on the model parameters.

5.3 Preliminary Empirical Results

In this section, we present some preliminary empirical results from the field test. We do not currently have any information about the control group. Instead, we compare results from the treatment group with what happened last year to a group of customers with similar characteristics. More specifically, we draw a sample of customers from the customer pool whose *purchase recency* measurement was also no more than two and half years in the beginning of 2002 (the same criterion used to draw customers for the field test). We then observe what happened to this group of customers in the last year as a comparison benchmark with what happened to the control group. Notice this approach has limitations. The most notable limitation is that the differences we observe embed differences in the macro economic situations between this year and the last.

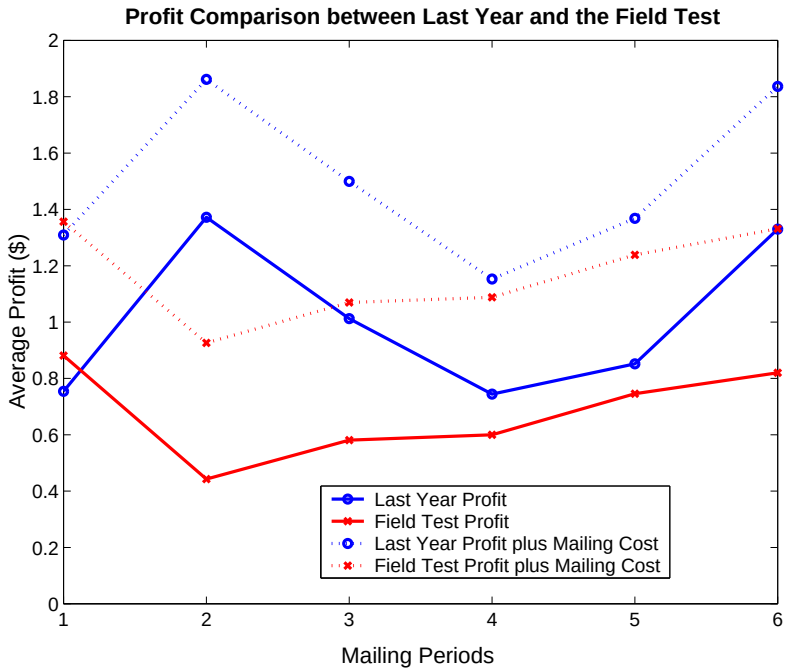

Figure 5-1: Comparing the profit obtained in each mailing period early last year versus in the field test.

5.3.1 Profit

Figure 5-1 compares the profit obtained per customer in each mailing period in the last year versus the result for the treatment group. Solid lines represent the average discounted profits per customer (immediate reward) in each mailing period. We also present the dotted lines for this profit plus mailing cost, reflecting the actual sales.

From the figure, we see that purchases during the same time period last year was stronger than this year in the field test, especially during the second mailing period. This difference could be a result of different mailing policies or different macroeconomic situations, such as stronger overall retail spending last year and the pending war this year. The control group will provide an explicit control for these differences.

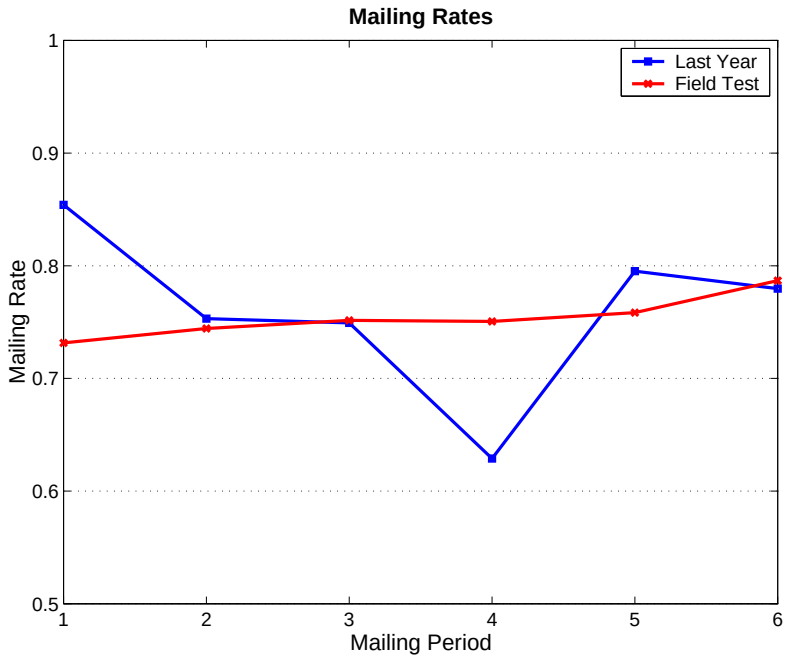

Figure 5-2: Comparing the mailing rates early last year versus in the field test.

5.3.2 Policy

Figure 5-2 compares the aggregated mailing rates. From the figure, we see that the overall mailing rates are similar. Mailing rates in the historical policy varied more than in our policy, indicating that the company may take into consideration catalog heterogeneity, while we treat all mailings as the same. From the figure we can conclude that at least in the first few mailings, mailing cost does not contribute much to the difference in profits.

In order to show the changes in policies and distribution of customers over time, we introduce Figure 5-3. Red represents high values and blue low values. We aggregate the treatment group customers' information according to the *Purchase Recency* variable and a *Mailing Stock* variable. Sub-Figures 5-3(c) and 5-3(d) depict the mailing rates in the first versus the sixth mailing period. It turns out that 5-3(c) (the first mailing) looks similar to the mailing policy figure in Chapter 2 (upper Sub-Figure 2-4), indicating that the distribution of customers in the beginning of the field test was still similar to the distribution of historical data in the state space. However, by the 6^{th} mailing, the mailing rate figure 5-3(d) is

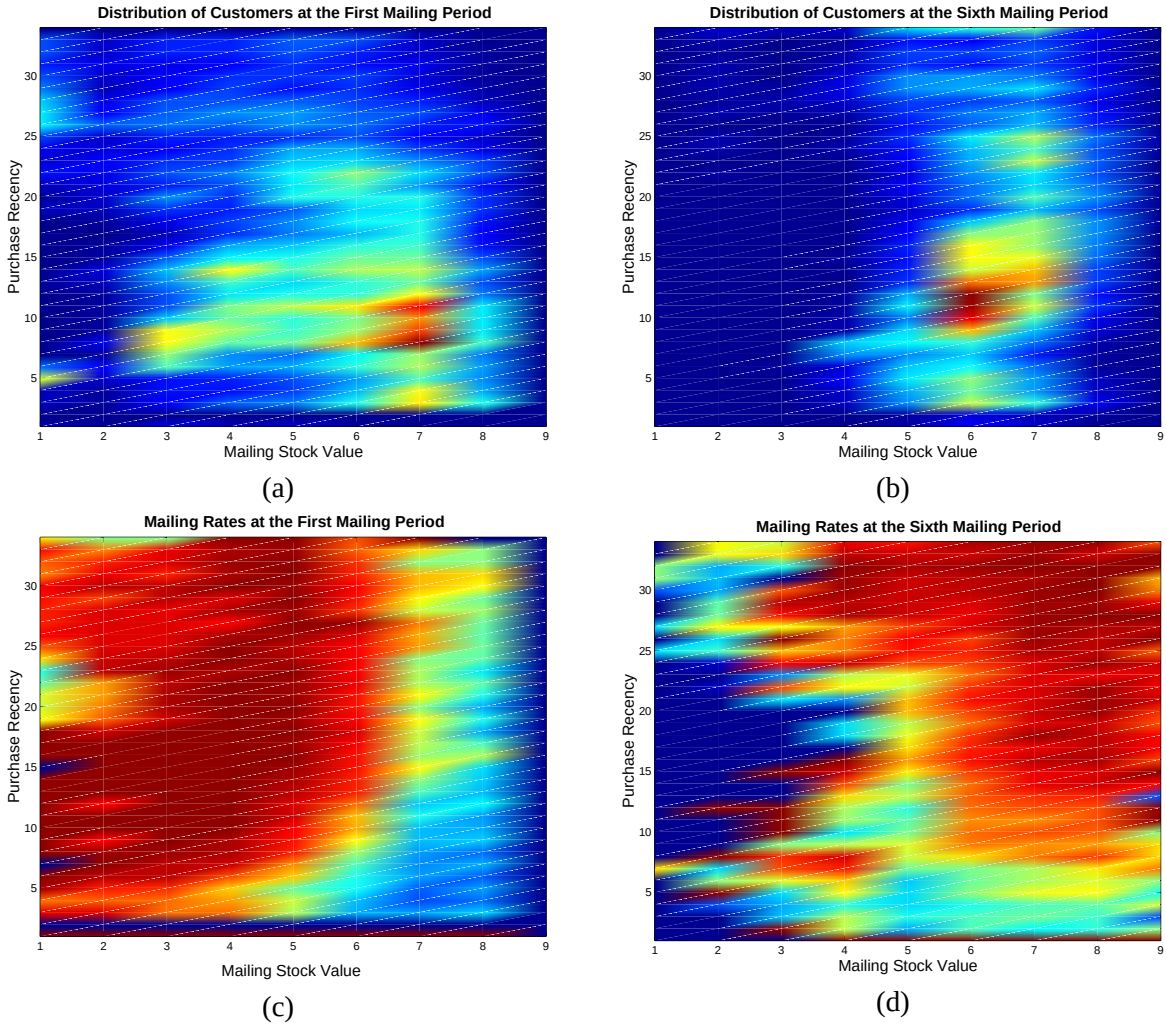

Figure 5-3: Distribution of customers and mailing rate.

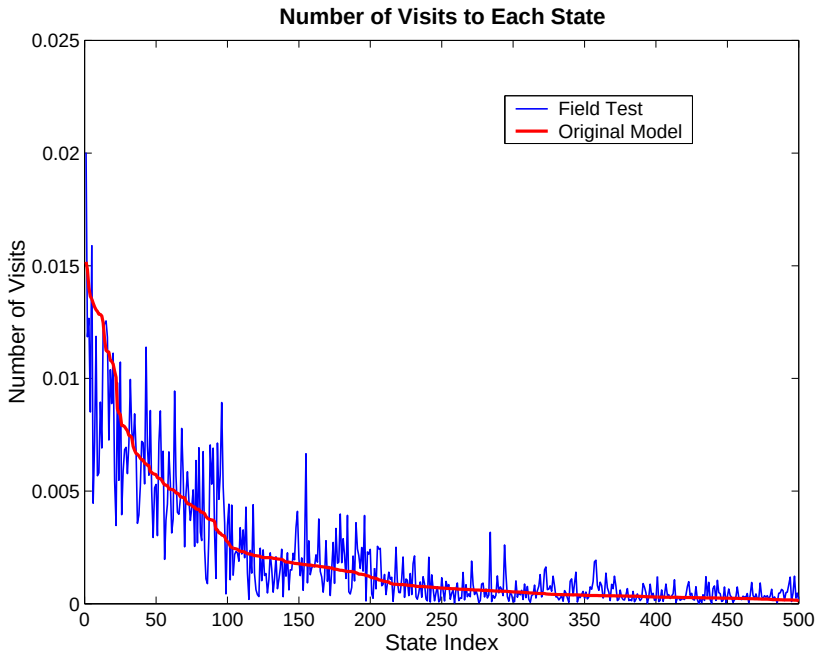

Figure 5-4: Comparing the number of visits to each state in the original model versus in the field test.

already very different. The mailing rate is much more highly correlated with the *Mailing Stock* variable value. This mailing policy is still very different from the historical mailing policy. The most recent purchasers (with low *Purchase Recency* values) get fewer mailings than those how have not purchased for a long time.

5.3.3 Distribution of customers

When we compare how the distribution of customers evolves in the field test, we compare the percentage of visits to each active state during the first five mailing periods in the field test versus the distribution of historical data in the state space. Figure 5-4 indicates this comparison. There is a clear positive correlation (0.8777) between the two curves in the Figure. This indicates that the distribution of customers has not diverged from the historical distribution of customers in the state space very much. However, as we can imagine, over time, this correlation keeps decreasing, as shown in Table 5.2. There are two reasons for

Table 5.2: Correlation between the percentage of visit to each state in the Treatment group and the prediction in the original model.

Mailing Period	1	2	3	4	5
Correlation	0.8968	0.8935	0.8826	0.8611	0.8518

this pattern. One is because the system follows a different policy in the field test. Because the state space records information on the mailing actions, merely changing the policy could potentially change the distribution of data in the state space. A second reason is that the treatment group is a fixed group of customers. Overtime customers become "older" compared to the distribution of customers in the historical data. The control group will provide a means of controlling for this second effect. We will be particularly interested in observing which policy pushes customers towards higher profit-to-go states.

In Figure 5-3(a)(b), we see that the distribution of data moves towards slightly higher *Purchase Recency* values, reflecting the fact that many customers do not purchase in any given time period (overall purchase recency values go up). Another observation is that data tend are initially spread more widely along the *Mailing Stock* variable value. A potential explanation is that our mailing policy (partly reflected in the *Mailing Stock* variable) depends less on the *Purchase Recency* measurement compared to the historical policy.

5.3.4 Fitting the Bellman Equation

The profit-to-go estimation is defined as the discounted total reward as shown in equation 2.1. Since we only have a finite number of time periods to observe the immediate reward, we are not able to test the accuracy of the profit-to-go estimation by observing infinite trajectories. One way to conduct the test is to check whether the following Bellman Equation –

$$V^{\pi}(s) = \mathrm{E}\left[r_{s,\pi(s)} + \delta^{T} V^{\pi}(s')|s, \pi(s)\right] \quad \forall s \in S \tag{5.1}$$

is satisfied according to the observed distribution of customers and the immediate rewards.

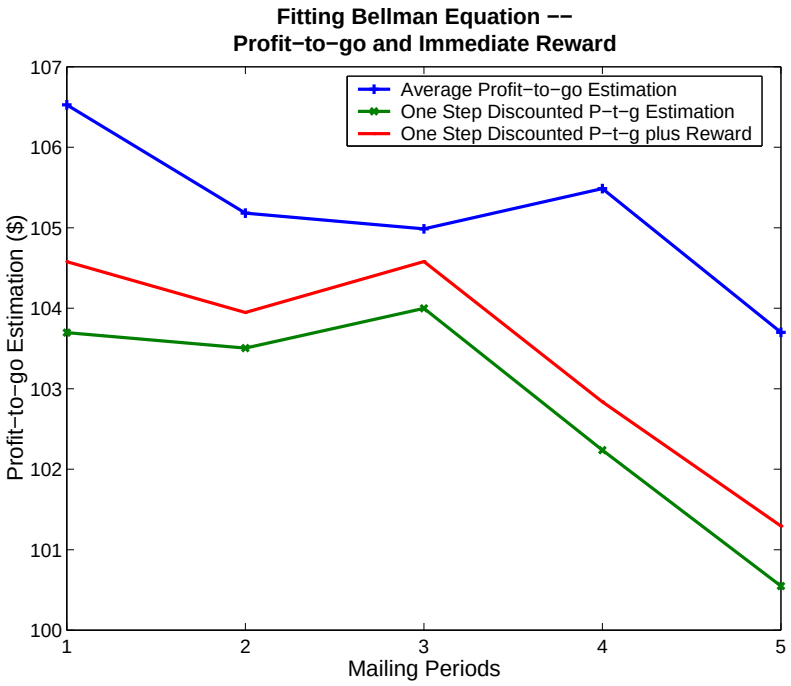

Figure 5-5: How does the actual profit-to-go and immediate reward fit the Bellman Equation?

Focusing on a mailing period in the field test, we can estimate the "average life time value" (V_0) of all customers in the beginning of the time period by averaging the validated profit-to-go estimations of these customers. Similarly, average life time value at the end of the mailing period V_1 can be estimated as well. If the Bellman Equation holds, the difference between V_0 and $\delta^T V_1$ should be the average immediate reward. The realized immediate reward can be directly estimated from the field test data.

Figure 5-5 illustrates the above analysis. We show the V_0 and $\delta^T V_1$ values from five mailing periods. The gap between the middle curve and the bottom curve shows the realized immediate reward. From the figure, it is clear that at least in the first five mailing periods, the observed immediate reward cannot make up for the drop in the profit-to-go estimates. In other words, the validated profit-to-go estimation from our model is biased upwards in the first five mailing periods. Part of the bias comes from the model inaccuracies discussed earlier. Observing the results for the control group will enable us to further investigate this

issue.

5.4 Conclusion

This chapter presents the field test procedures, the ideal test design and some preliminary empirical results based on the currently available data.

In the beginning of the Chapter, we listed potential problems in merely using off policy sample trajectories to build and test the learning model. Those problems are not only important motivations for the field test, but also provide directions for experiment design and explanations for potential gaps between the model predictions and observations from the field test.

Because we do not yet have updated information for the control group, we can only conduct a limited range of tests. From the preliminary empirical results, we see that the model parameter estimates do not perfectly match the field test in the first few mailing periods and the profit-to-go estimation from the model is over optimistic. The difference comes from two potential sources. One is systematic error in the model as discussed in previous chapters and indicated in the beginning of this Chapter. The other source is intervening variance specific to this time period. Data from the control group will provide a benchmark for separating the two reasons. We expect to have further results soon.

Chapter 6

Conclusions

In this thesis, we discussed an important application and some theoretical investigations around the theme of constructing dynamic decision making models from data.

In the first part (Chapter 2) of the thesis, we present an important application, the catalog mailing problem. We build an infinite horizon Markov decision process, use historical business data to estimate the model parameters and implement dynamic programming algorithms to solve for near optimal mailing policies. The main challenge is that we need to summarize relevant information into the state space and there is no natural way of constructing the state space. We present a new way of using data to construct dynamic programming state spaces. There is an important trade off between having a more refined state space and a computationally tractable model.

In Chapter 3, we discuss an important problem, the endogeneity problem caused by hidden information. When we do not have access to hidden state information, the model parameter estimates and the "optimal" policy depend on the historical data. We show that this may result in "attribution errors" if there exist hidden information that determined the historical policy and not all actions are evaluated in the historical data. When "attribution error" occurs, the dynamic programming algorithms tend to attribute the effect of the hidden information to different actions that were taken and propagate the effect to the whole state space. This leads to upward biased profit-to-go estimation and a suboptimal policy.

More generally, the endogeneity problem results from state space aggregation. When we are only able to observe the aggregated state space, model parameter estimation depends on the distribution of data on the true state space. In the second part of Chapter 3, we investigate the use of "batch online learning" procedures to overcome this problem. We focus this discussion around self-enforcing policies. We show that self-enforcing policies exist but in general are not unique.

Chapter 4 of the thesis discusses another important problem related to constructing learning model from data – the estimation error problem. We show in this chapter that estimation error in the model parameter estimates tents to upward bias profit-to-go estimates. Empirically, we show that in the catalog mailing application this problem is aggravated when the state space is very refined. The bias from model inaccuracy overcomes the advantages from more degrees of freedom. We also conduct a theoretical investigation of the upward bias problem.

An important component of this research is to build a decision making model to facilitate catalog companies' daily business decisions. In order to investigate the effectiveness of the proposed approach, we present preliminary results from a large scale field test. In the field test we build a model using a large dataset provided by a catalog mailing company. This model makes dynamic mailing decisions for a group of customers over a 6-month period. In Chapter 5, we present initial findings from this field test.

Appendix A

Perron-Frobenius Theory and the Continuity of $F(\lambda, g)$

In this Appendix, we present the well known Perron-Frobenious Theorem for irreducible matrices and use that to show that the F functional that we presented in Chapter 3 of the thesis is indeed continuous assuming that the Markov chain for each action is recurrent. In the rest of this Appendix, $\geq, \leq, >, <$ are component-wise when used between vectors or matrices.

Definition 15 (Primitive and Irreducible Matrices) *Let T be an $n \times n$ nonnegative matrix. T is called primitive if for some k, $T^k > 0$; T is called irreducible if for all i, j, there is a k such that $\left(T^k\right)_{ij} > 0$.*

Apparently, the transition probability matrix for a recurrent Markov chain is a irreducible matrix.

Theorem 16 (Perron-Fronbenious Theorem) *Let $T \geq 0$ be irreducible. Then there is a unique real number θ_0 with the following properties:*

1. *There is a real vector $x_0 > 0$ with $Tx_0 = \theta_0 x_0$.*
2. *θ_0 has geometric and algebraic multiplicity one.*

3. *For each eigenvalue θ of T we have $|\theta| \leq \theta_0$. If T is primitive, then $|\theta| = \theta_0$ implies $\theta = \theta_0$. In general, if T has period d, then T has precisely d eigenvalues θ with $|\theta| = \theta_0$, namely $\theta = \theta_0 e^{2\pi ij/d}$ for $j = 0, 1, ..., d-1$. In fact the entire spectrum of T is invariant under rotation of the complex plane over an angle $2\pi/d$ about the origin.*

4. *Any (nonzero) nonnegative left or right eigenvector of T has eigenvalue θ_0. More generally, if $x \geq 0,\ x \neq 0$ and $Tx \leq \theta x$, then $x > 0$ and $\theta \geq \theta_0$; moreover, $\theta = \theta_0$ if and only if $Tx = \theta x$.*

5. *If $0 \leq S \leq T$ or if S is a principal minor of T, and S has eigenvalue σ, then $|\sigma| \leq \theta_0$; if $|\sigma| = \theta_0$, then $S = T$.*

∎

If we define y to be the left eigenvector of a irreducible matrix T corresponding to the eigenvalue θ_0, such that $y^T T = \theta_0 y^T$, next we show that y is continuous at T.

Corollary 17 *Each component of vector y defined as $y^T T = \theta_0 y^T$ and $y^T e = 1$, with θ_0 defined as in the Perron-Fronbenious (P-F) Theorem, is a continuous function at any irreducible matrix T.*

Proof. *According to part 2 of the P-F Theorem, the solution to $y^T T = \theta_0 y^T$ and $y^T e = 1$ is unique. Notice in order to satisfy the above condition as a unique solution, y should be the last row of matrix $\begin{pmatrix} (T-\theta_0 I)' & e \end{pmatrix}^{-1}$, in which $(T-\theta_0 I)'$ is defined as an $n \times (n-1)$ matrix which is $T - \theta_0 I$ taken out the last column.*

Each component of y is $\frac{(-1)^t}{\det\left[(T-\theta_0 I)'\ e\right]}$ for some value of t. Since matrix $\left[(T-\theta_0 I)'\ e\right]$ is invertible, $\det\left[(T-\theta_0 I)'\ e\right] \neq 0$ and thus y is continuous at T. ∎

Since for a transition probability matrix $P \geq 0$ we have $Pe = e$, from part 1 of the P-F Theorem, $\theta_0 = 1$. Following Corollary 17, it is obvious that the steady state probability p defined as $p^T P = p$ is continuous under perturbation for any irreducible matrix P, which corresponds to a recurrent markov chain. Since obviously P_λ defined in (3.1) is irreducible given any P_a is irreducible, we have the following conclusion for the functional F defined in (3.4).

Proposition 18 *F is continuous.*

∎

Bibliography

[1] E. Anderson and D. Semester. Does promotion depth affect long-run demand. Working paper, MIT, 2002.

[2] Direct Marketing Association. *Statistical Fact Book, 23^{rd} Edition*. DMA, New York, NY, 2001.

[3] L. C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37, 1995.

[4] J. Banslaben. *The Direct Marketing Handbook, E. L. Nash (Eds.)*, chapter Predictive Modeling. McGraw-Hill, New York, NY, 1992.

[5] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[6] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Massachusetts, 1995.

[7] D.P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Optimization and Computation Series. Athena Scientific, Belmont, Massachusetts, 1996.

[8] G.R. Bitran and S. V. Mondschein. Mailing decisions in the catalog sales industry. *Management Science*, 42(9), Sep. 1996.

[9] J.R. Bult and T. Wansbeek. Optimal selection for direct mail. *Marketing Science*, 14(4), 1995.

[10] D. P. de Farias and B. Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 105(3), 2000.

[11] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, New York, NY, 2000.

[12] A-M. Fiore and H. Yu. Effects of imagery copy and product samples on responses toward the product. *Journal of Interactive Marketing*, 15(2):36–46, 2001.

[13] M. Florian and D. Hearn. *Handbooks in OR & MS, M.O. Ball et al. (Eds.)*, volume 8, chapter Network Equilibrium Models and Algorithms. Elsevier Science B.V., 1995.

[14] F. Gönül and M. Z. Shi. Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Science*, 44(9), Sep. 1998.

[15] G. J. Gordon. Reinforcement learning with function approximation converges to a region. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 1040–1046. MIT Press, 2001.

[16] L. Hayes. Catalog age special report; the 6th annual analysis of trends and practices in catalog business. *Catalog Age*, 9(12), 1992.

[17] J. D. Hess and G. E. Mayhew. Modeling merchandise returns in direct marketing. *Journal of Direct Marketing*, 11(2):20–35, 1997.

[18] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3), 1987.

[19] E. Pednault, N. Abe, and B. Zadrozny. Sequential cost-sensitive decision making with reinforcement learning. In *Proceddings of the Eighth ACM SIGDKK International Conference on Knowledge Discovery and Data Mining*, pages 259–268, August 2002.

[20] M. L. Puterman. *Markov Decision Problems*. Wiley, New York, NY, 1994.

[21] D. D. Schönbachler and G. L. Gordon. Trust and customer willingness to provide information in database-driven relationship marketing. *Journal of Interactive Marketing*, 16(3):2–16, 2002.

[22] R. Sutton and A. Barto. *Reinforcement Learing, An Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, 2000.

[23] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690, 1997.