Three Contributions to the Theory and Practice of Operations Management

by

Anton Ovchinnikov

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Joseph L. Rotman School of Management
University of Toronto

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-27714-0
Our file Notre référence
ISBN: 978-0-494-27714-0

Three Contributions to the Theory and Practice of Operations Management

Anton Ovchinnikov
Doctor of Philosophy
Joseph L. Rotman School of Management
University of Toronto
2007

ABSTRACT

This dissertation is a collection of three essays in Operations Management and Management Science.

Chapter 1 considers a problem where consumers learn about the possibility of "last-minute" discounts (typical of travel industries) and strategically wait for them. We present a stylized model of aggregate consumer behavior where the firm puts a number of units on sale to maximize its current and future revenue, given that the fraction of customers waiting, and, hence, the revenue in the future, changes depending on the firm's decisions. We formulate the problem as a dynamic program and develop a novel solution approach. We consider several model variations and show that the firm's optimal policy depends on the learning behavior: it is either "passive", where the firm puts some units on sale and allows consumers to "self-regulate" future waiting, or it is an "active" "bang-bang" policy, where the firm intermixes the periods with many units on sale with those with none and thus manages consumer waiting. We discuss managerial insights and show that the firm can strategically allow overbooking to increase its revenue.

Chapter 2 studies the problem of constructing balanced work groups (i.e., containing approximately equal proportion of members with different gender, age, cultural backgrounds, and other relevant attributes), based on a practical problem of assigning MBA students to study groups. We view balancing requirements as constraints, develop efficient user-friendly software, discuss its implementation and report major improvements in all aspects of students' group work. We also discuss a problem of creating multiple lists of non-overlapping groups, which is unique to our work.

Chapter 3 explains the empirical phenomenon that the constraint programs resulting from the real-world problems in Chapter 2 always had solutions. From the worst-case perspective this need not be the case for in most cases there exist instances with few (e.g., three) attributes where balanced groups cannot be constructed. However, via a variety of techniques (dynamic programming combined with simulation, analytical upper bound and empirical lower bound) we find that the probability that a random instance similar to those observed in practice can be partitioned into balanced groups is effectively 100 percent.

Acknowledgements

I would like to express my gratitude to the members of my doctoral committee, Professors Dmitry Krass, Joseph Milner and Oded Berman at the Joseph L. Rotman School of Management, University of Toronto, and Professor Harry Groenevelt from the University of Rochester for their guidance and support during my research and preparation of this dissertation.

There are several other people that I would like to acknowledge. First are my parents, Tamara and Sergei Ovchinnikov, who with their personal examples of hard work and courage gave me strength to go to the Doctoral Program. This, however, would not be possible without Professors Hans Wissema and Rem Hlebopros. It was during my work with Prof. Wissema when I made a decision to shift from my then business job and to go to the academia; and it was Prof. Hlebopros who, during one of the worst times in my life, recommended my candidacy to Professor Krass, thus starting the process that lead to this work, as well as many other works that are not included in my dissertation. Finally, I would like to thank my wife, Mila, and my daughter, Katya, who through all these years have been a source of great inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Revenue Management through Last Minute Deals in the Presence of Customer Learning

1.1 Introduction

Over the last several years there has been rapid growth in online purchases of airline tickets and other travel-related products. Forrester Research estimates that "Web travelers now make up 79% of the U.S. travel population, and 55% of them buy leisure travel online" (Hartveldt et al. 2006). This growth has presented a number of opportunities and challenges for travel-related firms. These include the ability and need to rapidly change the prices and availability of inventory, to track and respond to competitor moves, and to address changes in consumer behavior, as the full array of options for consumers are more easily found through Web searches.

This increase in online business has also provided the capability to place inventory that is not selling at the expected rates (so-called "distressed" inventory) on sale in the days immediately prior to a departure of a flight or other product, such as a hotel room night, vacation package, or a weekend car rental. That is, the "last-minute" in the title may not literally be the last minute, but rather it is a moment of price reduction in the days leading to a departure. For consumers, such last-minute deals present an opportunity to purchase products at noticeably lower prices. For example, a quick search for a week-long all-inclusive vacation departing in an approaching weekend in the fall of 2006 produced choices for less than \$400; this compares with prices in excess of \$1,000 available months in advance.

Given that last-minute deals are so great, it was not long before "many savvy travelers ... noticed it" (Michael Sands, CMO of Orbitz, quoted in Stinger 2002). Increasingly, though, more than just savvy travelers are learning to expect last-minute deals and "prefer to book later in the hope of getting a good deal" (Fenton and Griffin 2004). According to research by American Express, "nearly half of all travelers say they intend to wait until the last minute to plan their vacations" (De Lisser 2002). Similarly, in private conversations, executives of a leading vacation tour operator noted that as a result of customer waiting for the deep discounts mentioned above, early bookings are "slow" and 27% of the bookings are made in the last 15 days. That is, because customers act strategically and increasingly come to expect last-minute sales, the discounts that were meant to help sell distressed inventory turned out to cause more units to be distressed. As a result firms sell more units at a discount and thus lose revenues, since in the absence of strategic waiting for discounts, some of these units could have been sold at higher prices. This suggests that firms should carefully consider strategic consumer response and incorporate it into their revenue management policies.

The goal of this paper is to develop a stylized model that incorporates strategic customer response to revenue management. The natural setting for such a model is a travel firm (a tour operator, car rental firm, hotel, airline, etc.) determining the number of units (seats, cars, rooms, etc.) to put on last-minute sale in the days prior to a departure. We consider cases with two and three customer classes purchasing inventory in a multiple-period (week, flight, etc.) setting. In each period, a fraction of the customers purchase at a regular, nondiscounted price and a fraction waits for a potential last-minute sale. This fraction changes as customers learn to expect such sales and adjust their behavior to take advantage. Customers who wait, but do not receive inventory at the discounted price, may be offered inventory for purchase at a higher price. The decision faced by the firm is to determine the number of units (if any) to put on last-minute sale in each period.

Our initial model with two customer classes (and two prices) reflects such industries as packaged vacations and performance events, where a common practice of prepublishing prices in catalogs effectively reduces the firms' ability to increase prices in the case of high demand. Our three classes (prices) model captures the examples of airlines, car rental firms, and hotels that can increase the price if there are customers willing to spend more for the product. We study the limited number of classes in order to derive optimal policies in this complicated multiple-period model. In practice, revenue management techniques with multiple fare classes would be undertaken prior to offering a last-minute discount.

In our general model, we allow both the total demand in a period and the number of customers waiting to be stochastic. We introduce two learning behaviors and, under some regularity conditions on customer demand and behavior, we show that the number of units to put on last-minute sale is uniquely determined. In particular, we show that for some demand levels and expectations of customer behavior, it is optimal to place no units on sale. We observe that the firm, in expectation, follows a pattern of offering last-minute deals to increase the number of customers waiting, and then periodically puts no units on sale, generating revenue from customers then forced to purchase at a higher price, and at the same time controlling future waiting.

Our work is different from previous research in a number of dimensions. First and foremost, we consider a series of a firm's revenue management decisions, which influence customers' behavior for the future periods. To the contrary, the vast majority of the research in revenue management considers a single selling period (flight, etc.), and effectively ignores the possible effects on future periods. Second, we incorporate the "double" uncertainty of both stochastic total demand and stochastic number of customers waiting. We show that the resulting dynamic programming model is not amenable to standard solution methodologies. We generalize our case to a subclass of dynamic programs and suggest a new solution approach. Third, we derive the optimal policy in the closed form for three simplified models. Finally, we discuss the effects of different patterns of consumer behavior with respect to the types and speed of learning, and with respect to overbooking.

The main contribution of this paper is the proposal and solution of a model of strategic response to revenue management that reflects both stochastic demand and stochastic customer behavior. Further, by restricting either of these stochastic elements to a deterministic form, we provide closed-form solutions for our problem while relaxing several of the previous assumptions. Through numerical studies we document the degree to which the optimal solution provides benefits over reasonable policies such as discounting excess inventory based on a coin flip or using naïve rule-based policies. In addition, we make a theoretical contribution by presenting a solution methodology to a subclass of dynamic programs in which the state of the system evolves nonmonotonically.

The remainder of the paper is as follows. In Section 1.2 we review the relevant literature. In Section 1.3 we introduce the model and describe the two types of learning behavior. In Section 1.4 we study the optimal policy for two customer classes under the assumption of "self-regulating" learning, and present the solution to the resulting dynamic program. The case of "smoothing" learning is discussed in Section 1.5, where we introduce two simplifications to our general model and present their optimal policies in the closed form. In Section 1.6 we extend our models to the case with three customer classes, and discuss the effects of overbooking on customer behavior and on the resulting optimal policy of the firm. Numerical results are presented in Section 1.7, followed by the conclusions and prospects for future research.

1.2 Literature Review

Revenue management has been an active area of research for some time. We review only the literature directly related to the current study. McGill and van Ryzin (1999), Bitran and Caldentey (2003) and the recent book, Talluri and van Ryzin (2004) provide comprehensive reviews of the broader literature. Most previous research focusses on pricing and inventory policy for either a single flight or product, or a network of flights or multiple products, where strategic response by customers to the determined policy is ignored. For example, the fundamental work of Belobaba (1989) assumes that demand depends only on the price for the current flight and not on any previous pricing policy. Similarly, Gallego and van Ryzin (1994) and Bitran and Mondschein (1997), who discuss determining the optimal pricing policy for demand over time, also do not consider the strategic response of customers to the stated pricing policy.

All of the literature mentioned above deal with a single selling season (flight, etc.). Relevant work that considers multiple selling seasons includes papers on intertemporal price discrimination and advance selling, such as Stokey (1979), Sobel (1984), and Conlisk et al. (1984). A summary of retail pricing can be found in Lazear (1986). Besanko and Winston (1990) and Gale and Holmes (1993) discuss the optimal price skimming by a monopolist. Dana (1999), Xie and Shugan (2001), and Tang et al. (2004) discuss advance selling. These works do not consider customer learning and in this regard are different from ours. Customer learning is often modeled through reference price effects; for example, consider Greenleaf (1995) and Popescu and Wu (2005). These models do not consider the internal dynamics of selling to several classes of customers within each selling season, which is a major feature of revenue management systems.

A newsvendor problem with two customer classes is considered in Sen and Zhang (1999), whose increasing prices model is similar to our model in the single period. We note that they do not study repetitive problems and customer learning.

Within the research on revenue management, only recent work directly relates to ours in considering of how customers strategically react to the pricing policy of the firm. Anderson and Wilson (2003), Aviv and Pazgal (2003), and Elmaghraby, Gulcu and Keskinocak (2004) consider a problem somewhat reverse to ours, where customers react strategically to a preset policy of the firm. Levin et al. (2006) examine a dynamic game between the firm and strategic consumers, and Su (2006) considers a case when a part of consumer population acts strategically. These papers consider a single selling season as well. Anderson and Wilson (2006), Liu and van Ryzin (2006) and Zhang and Cooper (2006) study the behaviors of the firm and strategic customers over two subperiods within a single selling season.

A multiple-period setting similar to ours is considered in Cooper, Homem-de-Melo and Kleywegt (2004), who model the "spiral-down" effect and demonstrate that in the multiple-period problem with customer learning the effect of otherwise optimal (single-period) revenue management policy could be significantly diluted. This suggests that the optimal multiple-period pricing policy is different and worth studying. The current paper develops such a policy.

1.3 Model

Next we present a stylized model followed by a discussion of our modeling approach and the assumptions we make. Figure 1.1 depicts the timeline of the model and main notation.

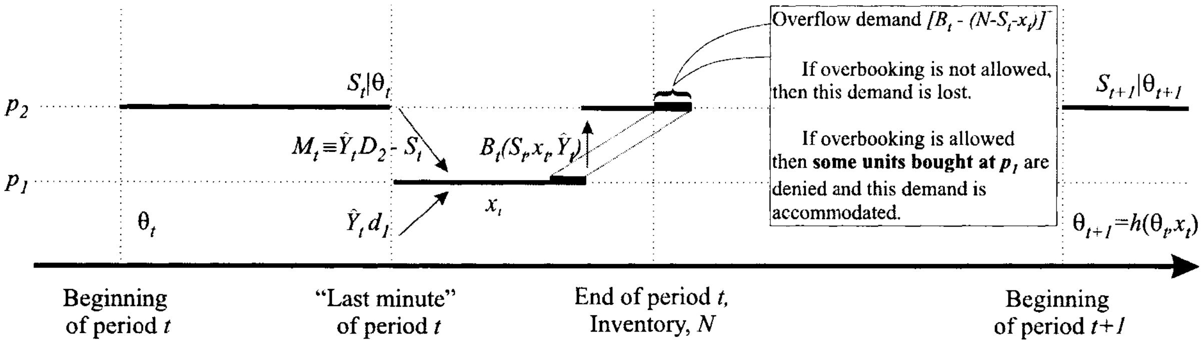

Overflow demand $[B_t - (N-S_t-x_t)]^+$
If overbooking is not allowed, then this demand is lost.
If overbooking is allowed then **some units bought at p_1** are denied and this demand is accommodated.
$S_t|\theta_t$
p_2
p_1
$M_t \equiv \hat{Y}_t D_2 - S_t$
$B_t(S_t, x_t, \hat{Y}_t)$
$\hat{Y}_t d_1$
x_t
θ_t
$S_{t+1}|\theta_{t+1}$
$\theta_{t+1}=h(\theta_t, x_t)$
Beginning of period t
"Last minute" of period t
End of period t, Inventory, N
Beginning of period $t+1$

Figure 1.1: Timeline of the model.

1.3.1 Formulation

Consider a sequence of identical offerings of a perishable product or service, for example, weekly all-inclusive vacations at the same resort, Wednesday morning flights from London to New York, or weekend car rentals. To distinguish between the copies of the product offerings, we assume that they are offered in different *periods*, and that there is one offering per period. In each period $t = 1, 2, ...T$, for finite T, there are N units of product available, and the firm decides whether it should offer a last-minute deal on a part of its inventory.

We initially assume that there are two customer classes: $i = 1, 2$. Let p_i be the highest price that class i is willing to pay for the unit of product, and, without loss of generality, we assume $p_2 \geq p_1$ (an extension to three prices/customer classes is presented in Section 1.6). Demand from class i in period t reflects a nominal demand, d_i, and an exogenous stochastic multiplier, Y_t, representing, for example, weather or exchange rate. We assume that the demand from class i in period t is $Y_t d_i$. Thus the total nominal demand at price i is $D_i = \sum_{j=i}^{2} d_j$, $i = 1, 2$, and the total demand at price i in period t therefore is $Y_t D_i$. We assume Y_t is a random variable with finite support on $[\underline{y}, \overline{y}]$, where $\overline{y} D_2 \leq N$. For simplicity we treat all demands and capacities as continuous variables, and for $\Delta > 0$, Δ units of inventory fill Δ units of demand.

At the start of period t the firm initially sets the price at p_2 and by some "last minute" observes the initial sales, $S_t \in [0, Y_t D_2]$, to its class-2 customers. The remaining $M_t \equiv Y_t D_2 - S_t$ class-2 customers wait for a discount; also waiting are all $Y_t d_1$ class-1 customers for a total of $Y_t D_1 - S_t$ customers waiting. We assume that S_t is a random variable whose cdf, $F_{S_t|\theta_t}(\cdot)$, is parameterized by a value, θ_t, that represents the propensity of customers to wait for a discount – the *waiting behavior.* For example, there could be a random fraction, α_t, of class-2 customers waiting (then $S_t = (1 - \alpha_t)Y_t D_2$) and θ_t could be the average fraction waiting. We refer to S_t as the *demand signal* and assume θ_t is known to the firm prior to observing S_t (e.g., through consumer research).

At the "last minute" of period t the firm determines x_t, the number of unsold units (possibly zero) to put on sale at price p_1. To do so, knowing θ_t and having observed S_t, the firm estimates the demand multiplier, Y_t, and, hence, the number of customers of each class waiting. Let $\hat{Y}_t \equiv Y_t|(\theta_t, S_t)$ be the random variable representing the demand multiplier Y_t conditional on (θ_t, S_t). If $x_t \geq \hat{Y}_t D_1 - S_t$ then all waiting demand is satisfied. Otherwise, some class-2 customers may still be unserved, and we assume that their number is given by the *allocation* function $B_t(S_t, x_t, \hat{Y}_t)$. To accommodate these $B_t(\cdot)$ class-2 customers the firm offers all the inventory remaining after the last-minute sale again at p_2 (in Section 1.6 we discuss an extension where the remaining units are offered at price $p_3 > p_2$). If the remaining class-2 demand exceeds the remaining capacity, the firm may overbook*. In this case the firm denies some of the units purchased at price p_1 and sells them at p_2 to the remaining class-2 customers (the overflow demand). For doing so the firm incurs a penalty, p_C, per unit. We assume that customers denied product because of overbooking cannot purchase a unit in the same period. We consider cases when the firm overbooks and when it does not, and comment on when the firm benefits from strategically allowing overbooking (Section 1.7.1).

*Here, the term "overbooking" refers to intentionally selling some units of capacity twice; it does not refer to the practice of selling more units in anticipation of a cancelation – for simplicity we do not consider cancelations.

The total revenue of the firm net of the overbooking cost for period t given $(S_t, x_t, \hat{Y}_t)$ is

$$\begin{aligned} g_t(S_t, x_t, \hat{Y}_t) &= p_2 S_t + p_1 \min[x_t, \hat{Y}_t D_1 - S_t] \\ &+ p_2 B_t(S_t, x_t, \hat{Y}_t) - p_C \left(B_t(S_t, x_t, \hat{Y}_t) - (N - S_t - x_t)\right)^+ \end{aligned} \quad (1.1)$$

noting that overbooking occurs only when $\hat{Y}_t D_1 - S_t \geq x_t$. The expected single-period revenue in period t given θ_t and observing S_t is

$$r_t(\theta_t, S_t, x_t) = \int_{\underline{y}}^{\overline{y}} g_t(S_t, x_t, y) \mathrm{d}F_{\hat{Y}_t}(y). \quad (1.2)$$

where $F_{\hat{Y}_t}(y) \equiv F_{Y_t|(\theta_t, S_t)}(y)$ is the cdf of $\hat{Y}_t$, demand multiplier Y_t conditional on (θ_t, S_t).

We refer to the 2-vector (θ_t, S_t) as to the *state of the system.* We assume that the system evolves based on the decision x_t according to a function $h(\theta_t, x_t)$, defining θ_{t+1}, and a random draw of S_{t+1} from the distribution of future sales, $F_{S_{t+1}|\theta_{t+1}}(\cdot)$. We refer to $h(\cdot)$ as the *learning function* because it reflects the changes in the waiting behavior of the customers as they learn about the policy of the firm. We define two types of the learning functions:

(i) *smoothing*, if $\frac{\partial h}{\partial x} \geq 0$ and $\frac{\partial h}{\partial \theta} \geq 0$, e.g., $h(\theta, x) = \lambda \frac{x}{N} + (1 - \lambda)\theta$ for $0 \leq \lambda \leq 1$;

(ii) *self-regulating*, if $\frac{\partial h}{\partial x} \geq 0$ and $\frac{\partial h}{\partial \theta} \leq 0$, e.g., $h(\theta, x) = \kappa + \lambda \frac{x}{N} - (1 - \lambda)\theta$ for $\{\kappa, \lambda > 0 | \kappa + \lambda \leq 1\}$.

The objective of the firm is to maximize the expected T-period revenue, discounted at a fixed rate $\delta \in (0, 1)$. Therefore,given the initial θ_1, the firm determines the number of units on sale, x_t, for each period $t = 1, 2, ...T$ by solving the following dynamic program

$$J_t(\theta_t, S_t, x_t) = r_t(\theta_t, S_t, x_t) + \delta E_{S_{t+1}|\theta_{t+1}} \left[J^*_{t+1}(\theta_{t+1}, S_{t+1})\right] \quad (1.3a)$$

where

$$\begin{aligned} J_t^*(\theta_t, S_t) \quad &= \quad \max\nolimits_{0 \le x_t \le N - S_t} J_t(\theta_t, S_t, x_t) \\ &\quad\ \text{subject to} \\ &\quad\ J_{T+1}^*(\theta_{T+1}, S_{T+1}) = 0 \qquad \text{for all } (\theta_{T+1}, S_{T+1}) \\ &\quad\ \theta_{t+1} = h_t(\theta_t, x_t) \end{aligned} \tag{1.3b}$$

We refer to (1.3.1) as to the *general* model, as it reflects the uncertainty in the overall demand as well as in the fraction of class-2 customers waiting in every period. In Section 1.5 we consider *simplified* models, where alternately one or the other of these uncertainties is removed.

Finally, we make several assumptions on the stochastic ordering of the random variables. A family of random variables X_θ with cdf $F_X(x;\theta)$ is stochastically increasing (concave, supermodular, etc.) in parameter θ iff $1 - F_X(x;\theta)$ is increasing (concave, supermodular, etc.) in θ for any x from the support of X. We assume that (i) S_t is stochastically decreasing in θ_t; (ii) $\hat{Y}_t$ is stochastically increasing in S_t; and (iii) $\hat{Y}_t$ is either stochastically increasing or decreasing in θ_t. These assumptions are consistent with the following intuitive observations. Since θ_t measures the propensity of customers to wait, the number of customers who purchase (i.e., do not wait), S_t, should decrease in θ_t. Similarly, more customers purchasing at the initial price implies higher overall demand, thus $\hat{Y}_t$ should increase in S_t. However, since the total demand from class-2 is $\hat{Y}_t D_2 = S_t + M_t$, the effects of increasing θ_t on the conditional demand multiplier, $\hat{Y}_t$ could be twofold. Specifically, if M_t increases faster than S_t decreases, then $\hat{Y}_t$ would increase in θ_t. Otherwise, $\hat{Y}_t$ could decrease in θ_t. In our analysis in Sections 1.4 - 1.6 we use these assumptions in order to establish monotonicity properties of revenue function. To do so we require the following fundamental theorem:

Theorem 1.1 (3.9.1 in Topkis 1998) *If T is a subset of R^m, $\{F_X(x;\theta) : \theta \in T\}$ is a collection of distribution functions, and $\mathcal{F}$ is a closed (in the topology of pointwise convergence), convex cone of real-valued functions on T, then for any increasing set S,*

$\int_S \mathrm{d}F_X(x;\theta)$ is in $\mathcal{F}$ iff $\int_S v(x)\mathrm{d}F_X(x;\theta)$ is in $\mathcal{F}$ for any increasing real-valued function $v(x)$.

1.3.2 Discussion

Aggregate Demand. We consider a model where rather than tracking the detailed arrival dynamics of individual customers, the firm focuses on the aggregate behavior of customer classes. In particular, we assume that the firm knows the value, θ_t, that parameterizes the fraction of customers in class-2 who wait for the last-minute deal. This value is then updated in each period based on the outcome of the previous periods.

Our modeling approach is motivated by many discussions with revenue management executives, who noted that in a multiple-period setting like ours, customers who purchase products in different periods are typically different individuals, and it is rather unclear how these individuals react to the pricing policy of the firm or even if they have accurate information about it at all. At the same time, the firm is predominantly concerned not with the behavior of each individual customer, but rather with an aggregate outcome of these individual behaviors – the aggregate demand. These executives further noted that there exists aggregate-level information, such as industry reports, news articles, or word-of-mouth, through which aggregate demand reacts to the pricing policy of the firm. Therefore in designing its pricing policy over multiple selling seasons, the firm could consider a model of aggregate demand that comes from classes of customers, where these aggregate classes and not the individual customers react to such information that in turn is updated to reflect pricing policies.

Waiting Parameter and Waiting Fraction. Our key differentiating assumption from previous work is that the fraction of class-2 customers who wait in period t is described by a parameter, θ_t, representing, for example, the average fraction of customers waiting. This waiting parameter is a proxy for the aggregate-level information about waiting. Customers learn based on the firm's decision, x_t, and θ_t changes over time, determining the fraction of customers who wait in the future.

In practice there are two factors that lead to the existence of such a waiting fraction, and, more generally, to the environment where some customers wait and some buy early: anxiety/risk and anticipation. Customers who wait for a last-minute deal may experience anxiety and risk because they are not guaranteed a product. If the firm does not overbook, class-2 customers who wait risk not obtaining a product if the total number of customers waiting exceeds the remaining inventory. If the firm does overbook, class-2 customers who wait and book at price p_1 are indistinguishable from the customers of class-1 and thus can be denied a product if the unit they were promised is resold at a higher price after the last-minute sale (i.e., overbooked). Therefore, for a class-2 customer who decides to wait or buy, the positive utility from potential savings of $p_2 - p_1 > 0$, is offset by a disutility from anxiety and risk caused by a possibility of not getting a product at all. Overbooking as we describe is common in the packaged vacation industry where overbooked customers have little recourse other than accepting the alternate arrangements proffered by the firm plus any accompanying compensation.

Also, researchers in marketing (e.g., Nowlis et. al. 2004) and economics (e.g., Loewenstein 1987) showed that for "pleasure" products, of which a vacation is a classical example, there exists a positive utility of anticipation. That is, provided that the price is unchanged, customers who purchase a product earlier obtain a higher utility from consuming it. We note that the utility of anticipation is widely recognized by executives in the vacation industry as one of the drivers of early purchases. In the context of an individual customer's wait-or-buy decision, utility of anticipation creates another tradeoff against waiting, in addition to that from anxiety and risk. Heterogeneity of the customer population with respect to valuing such time and risk tradeoffs (Chesson and Kip Viscussi 2000) naturally yields the aggregate outcome that some customers wait and some purchase early, which is observed in practice and is reflected in our model through waiting parameter θ_t.

Allocation of Discounted Units. In practice, discounted units are sold on the "first come, first served" basis and the class of the customer is not known. Thus $B_t(S_t, x_t, Y_t)$ is a result of a random draw in which, for example, all waiting customers could have equal probability of buying a discounted product; then $B_t(\cdot)$ would have a hypergeometric distribution. However, incorporating random allocation leads to an untractable model, in part, because it requires treating the demands and capacities as integers. Therefore, we consider deterministic allocation mechanisms. In Sections 1.4 and 1.5, respectively, we discuss two forms of proportional allocation depending on the nominal or realized demands (Talluri and van Ryzin 2004, pp. 330 call such mechanisms "proportional rationing"). Our proportional allocation based on realized demands simplifies $B_t(\cdot)$ to be equal to the expected outcome of the above-mentioned random allocation. In Section 1.7.3 through numerical simulations we document that the optimal policy obtained with such a simplification is very robust: that is, the revenue generated by such a policy is only marginally different from that resulting from the policy optimal under random allocation.

Fixed Discounted Price $\mathbf{p_1}$. In general, firms could determine both the quantity to discount and the sale price for each period. Analyzing both variables simultaneously, however, leads to a very complex multi-period model, since doing so would requires understanding the customer's perception of how quantity complements/substitutes for price. For example, we would need to make assumptions on the effect on future customer waiting for the case when the firm offered only few units on sale, but the price was very low, as opposed to the case when the firm put many units on sale, but the discount was small. Therefore, in attempt to create a stylized parsimonious model we assume that the p_i's are fixed for the entire T periods and concentrate on the quantity decision. Furthermore, research has shown that a heuristic that charges the properly chosen single price instead of a dynamic price often performs just marginally suboptimally (e.g., Gallego and van Ryzin 1994). In Section 1.7.2 we demonstrate how one might determine the optimal static discount price while dynamically optimizing the quantity to discount.

Learning Behaviors. Finally, we investigate both self-regulating and smoothing functions $h(\theta_t, x_t)$, in Sections 1.4 and 1.5, respectively. For the kind of "smoothing" functions that we presnet in the example, the next period's waiting parameter, θ_{t+1}, lies between θ_t and x_t, so that the decision x_t is "smoothed" into the previous belief, θ_t. Smoothing functions of such a form represent the standard moving average forecasting and are frequently used (e.g., Greenleaf 1995, Popescu and Wu 2005). Alternately, "self-regulating" functions reflect the following behavior: as the total number of waiting customers increases, the chances to obtain a product on sale decrease for an individual customer, which negatively affects the number of customers waiting. If few customers are waiting, then this individual's chances of obtaining a product at a discount increase, facilitating waiting; that is, customers "self-regulate".

In Section 1.4 we show that for the case with a self-regulating learning function, the revenue function is concave in the number of units on sale, x_t, for every t. To do so we develop the necessary methodology to prove concavity and show that in addition to being concave, the expected single-period revenue $r_t(\theta_t, S_t, x_t)$ is also required to be supermodular and increasing. We note that our approach to showing concavity differs from standard methodologies for showing optimality of monotone policies, such as Topkis (1998) Section 3.9.2, Putterman (1994) Section 4.7.3, or their recent extensions, e.g., Smith and McCardle (2002). In their models the state of the system evolves monotonically in the previous state and decision. That is, the components of the state vector either increase or decrease in the previous state and decision. To the contrary, in our general model (1.3.1), state transitions are not monotonic, since S_t is stochastically decreasing in θ_t, while θ_t is increasing in either

θ_{t-1} or x_{t-1}, or both.

In Section 1.5 we study the case with a smoothing learning function and show that in general concavity does not hold, unless the speed of customer learning is "slow." Then we consider two simplifications to the general model and derive their solutions in closed form. Section 1.6 presents a model with three customer classes, for which the type of learning behavior does not influence the determination of the optimal policy.

1.4 Optimal Policy for Self-Regulating Learning

In this section we derive the conditions under which the revenue-to-go function is concave when the learning function, $h(\cdot)$, is self-regulating. We organize this section as follows. First in Section 1.4.1 we assume that the single-period expected revenue function, $r(\theta, S, x)$, given in (1.2), is concave, supermodular and increasing, and we show that these properties hold for $J_t(\theta_t, S_t, x_t)$ for all periods. Then in Section 1.4.2 we discuss the properties of the revenue function $g(S, x, \hat{y})$, given in (1.1) and other parameters of the model, which ensure concavity, supermodularity and monotonicity of $r(\theta, S, x)$. We conclude by presenting an example.

1.4.1 Concavity in Dynamic Programs With Nonmonotonic State Transitions

Observe from (1.3a) that since $\theta_{t+1} = h_t(\theta_t, x_t)$ is independent of S_t, the expected future revenue, $E_{S_{t+1}|\theta_{t+1}}\left[J^*_{t+1}(\theta_t, S_t)\right]$, also does not depend on S_t and therefore, letting $\phi_{t+1}(h_t(\theta_t, x_t)) = E_{S_{t+1}|\theta_{t+1}}\left[J^*_{t+1}(\theta_t, S_t)\right]$ we can substitute

$$J_t(\theta_t, S_t, x_t) = r_t(\theta_t, S_t, x_t) + \delta\phi_{t+1}(h_t(\theta_t, x_t)) \qquad (1.4)$$

where the function ϕ_{t+1} can be interpreted as the expected future revenue.

We assume that $r_t(\theta_t, S_t, x_t)$ is **(A1)** jointly concave in (θ_t, x_t), **(A2)** supermodular in (θ_t, S_t, x_t), **(A3)** increasing in S_t, **(A4)** S_t is stochastically decreasing and concave in θ_t, and **(A5)** $h_t(\theta_t, x_t)$ is linear self-regulating, i.e. $\frac{\partial h}{\partial x}\frac{\partial h}{\partial \theta} \leq 0$, and $\frac{\partial^2 h}{\partial x^2} = \frac{\partial^2 h}{\partial \theta^2} = \frac{\partial^2 h}{\partial x \partial \theta} = 0$.

Concavity and supermodularity in (1.4) are related by the following lemma (all proofs are presented in the appendix):

Lemma 1.1 *If ϕ_{t+1} is concave in h_t, then J_t is concave in x_t and supermodular in (θ, S, x).*

Proof. Recall that by assumption (A5) h is linear.

(i) Concavity follows from

$$\frac{\partial^2}{\partial x^2} J(\theta, S, x) = \frac{\partial^2}{\partial x^2} r(\theta, S, x) + \delta \frac{\partial^2 \phi(h)}{\partial h^2} \left(\frac{\partial h}{\partial x} \right)^2 \leq 0$$

since r is concave in x by assumption (A1) and ϕ is concave in h by the condition of the lemma.

(ii) Supermodularity in (θ, x) follows from

$$\frac{\partial^2}{\partial x \partial \theta} J(\theta, S, x) = \frac{\partial^2}{\partial x \partial \theta} r(\theta, S, x) + \delta \frac{\partial^2 \phi}{\partial h^2} \frac{\partial h}{\partial x} \frac{\partial h}{\partial \theta} \geq 0$$

since r is supermodular in (θ, x) by assumption (A2), ϕ is concave in h by the condition of the lemma and h is self-regulating by assumption (A5).

(iii) Supermodularity in (θ, S) and (S, x) follows from (A2) because ϕ does not depend on S. Supermodularity in multiple dimensions is equivalent to supermodularity in each pair (Topkis (1998), Theorem 2.6.1). ■

Therefore in order to establish that J_t is concave in x_t it is sufficient to show that ϕ_{t+1} is concave in h_t, which in our original notation corresponds to $E_{S_{t+1}|\theta_{t+1}} \left[J^*_{t+1}(\theta_{t+1}, S_{t+1}) \right]$ being concave in θ_{t+1}. Let $f_{S_{t+1}}(y; \theta_{t+1})$ be the density of S_{t+1} given θ_{t+1}. Then $E_{S_{t+1}|\theta_{t+1}} \left[J^*_{t+1}(\theta_{t+1}, S_{t+1}) \right] = \int J^*_{t+1}(\theta_{t+1}, y) f_{S_{t+1}}(y; \theta_{t+1})\, \mathrm{d}y$. Concavity of this integral is established by the following lemma, which extends Theorem 1.1 to the case where the integrant depends on the parameter.

Lemma 1.2 *Let a family of univariate random variables X_θ with cdf $F_X(x;\theta)$ and density $f_X(x;\theta)$ be stochastically decreasing and concave in scalar parameter θ. Let $v(\theta,x)$ be supermodular in (θ,x), increasing in x and concave in θ. Then $\int v(\theta,x)\mathrm{d}F_X(x;\theta)$ is concave in θ.*

Proof. Let $\hat{\theta}$ be an arbitrary fixed value of θ.

Then

$$\begin{aligned}
&\frac{\partial^2}{\partial\theta^2}\left(\int v(\theta,x)\mathrm{d}F_X(x,\theta)\right)|_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta}\left(\int\left(\frac{\partial}{\partial\theta}\right)f_X(x;\theta)\mathrm{d}x + \int v(\theta,x)\frac{\partial}{\partial\theta}f_X(x;\theta)\mathrm{d}x\right)|_{\theta=\hat{\theta}}\\
= &\int\left(\frac{\partial^2}{\partial\theta^2}v(\theta,x)\right)|_{\theta=\hat{\theta}}f_X(x;\hat{\theta})\mathrm{d}x + 2\int\left(\frac{\partial}{\partial\theta}v(\theta,x)\right)\left(\frac{\partial}{\partial\theta}f_x(x;\theta)\right)|_{\theta=\hat{\theta}}\mathrm{d}x\\
&+\int v(\hat{\theta},x)\left(\frac{\partial^2}{\partial\theta^2}f_X(x;\theta)\right)|_{\theta=\hat{\theta}}\mathrm{d}x\\
\leq &\ 2\int\left(\frac{\partial}{\partial\theta}v(\theta,x)\right)\left(\frac{\partial}{\partial\theta}f_x(x;\theta)\right)|_{\theta=\hat{\theta}}\mathrm{d}x + \int v(\hat{\theta},x)\left(\frac{\partial^2}{\partial\theta^2}f_X(x;\theta)\right)|_{\theta=\hat{\theta}}\mathrm{d}x\\
\leq &\int v(\hat{\theta},x)\left(\frac{\partial^2}{\partial\theta^2}f_X(x;\theta)\right)|_{\theta=\hat{\theta}}\mathrm{d}x\\
= &\ \frac{\partial^2}{\partial\theta^2}\int v(\hat{\theta},x)\mathrm{d}F_X(x;\theta)\\
\leq &\ 0
\end{aligned}$$

The first inequality follows from the concavity of $v(\theta,x)$ in θ. The second inequality follows by Theorem 1.1 because $v(\theta,x)$ is supermodular (i.e. $\partial v/\partial\theta$ is increasing in x), while X_θ is stochastically decreasing. Finally, the third inequality results from Theorem 1.1 because $v(\theta,x)$ is increasing in x and X_θ is stochastically concave. ■

We also require the following two results:

Proposition 1.1 (follows from Theorem 5.3 in Rockafellar 1997) *If $f(x,\theta)$ is jointly concave in (x,θ), then $\sup_x f(x,\theta)$ is concave in θ.*

Proposition 1.2 (Theorem 2.7.6 in Topkis 1998) *If $f(x,\theta)$ is supermodular in (x,θ), then $\sup_x f(x,\theta)$ is supermodular in θ.*

Our main result in this Section is given by the following Theorem.

Theorem 1.2 *Under assumptions **A1** – **A5**, $J_t(\theta_t, S_t, x_t)$ is concave in x_t and supermodular in (θ_t, S_t, x_t) for all $t = 1, 2, ...T$.*

Proof. For $t = T$ the claim holds by assumptions (A1) and (A2) respectively. Let $1 \leq t \leq T$ and suppose that for every period $n \in [t+1, T]$: **(I1)** $J_n^*(\theta_n, S_n)$ is increasing in S_n; **(I2)** $J_n^*(\theta_n, S_n)$ is concave in θ_n and **(I3)** $J_n^*(\theta_n, S_n)$ is supermodular in (θ_n, S_n).

With assumptions (A4) and (I1)-(I3) following Lemma 1.2, ϕ_{t+1} is concave in h_t, and therefore with assumptions (A1), (A2) and (A5) by Lemma 1.1, $J_t(\theta_t, S_t, x_t)$ is concave in x_t and supermodular in (θ_t, S_t, x_t).

For period t: (I1) follows from (A3) since ϕ_{t+1} does not depend on S_t; (I2) follows by Proposition 1.1 from (A1), since (A5) implies that the Hessian of ϕ_{t+1} equals $\frac{\partial^2\phi}{\partial h^2}\left(\frac{\partial h}{\partial x}\right)^2 \frac{\partial^2\phi}{\partial h^2}\left(\frac{\partial h}{\partial \theta}\right)^2 - \left(\frac{\partial^2\phi}{\partial h^2}\left(\frac{\partial h}{\partial x}\frac{\partial h}{\partial \theta}\right)\right)^2 = 0$; (I3) follows by Proposition 1.2 since $J_t(\theta_t, S_t, x_t)$ is supermodular in (θ_t, S_t, x_t). ■

Problem (1.3.1) defines a subclass of dynamic programs for which a vector of the system state is not monotonic in the previous period's state and decision. In this section we have shown how to establish concavity for such dynamic programs. A similar logic with a different set of initial assumptions could be used to prove other properties, for example, convexity. We also note that to our knowledge there is no published research dealing with concavity (convexity) in the dynamic programs with nonmonotonic transitions. As such this is a technical contribution of our paper.

Next we discuss the underlying conditions on the customer behavior which ensure that concavity holds.

1.4.2 Allocation Functions

Recall that the allocation function, $B(S, x, \hat{Y})$, determines the number of class-2 customers that remain after the last-minute sale. In this section we study the properties of $B(S, x, \hat{Y})$ and the other parameters of the model which ensure that the expected single-period revenue function, $r(\theta, S, x)$, is concave, supermodular, and increasing. We show that if $B(S, x, \hat{Y})$ satisfies the assumptions (B1)-(B6) presented below, and the prices and demand multipliers satisfy some regularity conditions, then $r(\theta, S, x)$ is concave, supermodular and increasing, and therefore by Theorem 1.2 the revenue-to-go is concave for every period. Since the discussion relates to a single period, time indices are omitted.

We assume the following properties of the allocation function, $B(S, x, \hat{Y})$: **B1:** B is supermodular, increasing in $\hat{Y}$ and decreasing in S and x; **B2:** $\partial B/\partial x \geq -1$; **B3:** $\partial B/\partial S \geq -1$; **B4:** if $x \geq \hat{y}D_1 - S$ then $B(S, x, \hat{y}) = 0$; **B5:** if $S = 0$ and $x = 0$ then $B(0, 0, \hat{y}) = \hat{y}D_2$, and **B6:** $B(S, x, \hat{y})$ is piecewise concave in x on $[0, \hat{y}D_1 - S)$ and $[\hat{y}D_1 - S, N]$.

These are intuitive for an allocation function, since, respectively (B1): by the definition, B is the number of class-2 customers who were not allocated a discounted seat, which increases when demand multiplier increases, and decreases when more class-2 customers purchase the product either initially (S increases) or on sale (x increases); (B2): when additional Δ discounted units are allocated between class-1 and class-2 customers, the number of class-2 customers remaining can decrease by at most Δ (i.e., when all Δ units are allocated to a class-2 customers); (B3): increasing S by Δ decreases the number of waiting class-2 customers, but the number of class-1 customers waiting does not change, therefore fewer than Δ discounted units are allocated to class-2 customers, and so the number of customers remaining decreases by no more than Δ; (B4): if all waiting class-2 customers purchase discounted products, then no customers remain; and (B5): if no units are purchased at the regular price or at a discount, then all class-2 customers remain. Condition (B6) is technical.

We assume $\hat{Y}$ is stochastically increasing and concave in S and in θ and that $\hat{Y}$ is stochastically supermodular in (θ, S). Recall that these reflect the intuitive observations of consumer behavior as per the discussion in Section 1.3.

Recall that $r(\theta, S, x) = \int_{\underline{y}}^{\bar{y}} g(S, x, y)\mathrm{d}F_{\hat{Y}|(\theta,S)}(y)$ where $g(S, x, \hat{Y}) = p_2 s + p_1 \min[x, \hat{Y}D_1 - S] + p_2 B(S, x, \hat{Y}) - p_c\left(B(S, x, \hat{Y}) - (N - S - x)\right)^+$.

Let $y_L = (S + x)/D_1$, and let y_H be the largest solution to $B(S, x, y_H) = N - S - x$. If $\hat{y} < y_L$, then $x \geq \hat{y}D_1 - S$ and so by the assumption (B4) $B = 0$; i.e., overbooking cannot happen. If $\hat{y} > y_H$, then from assumption (B2), the demand from the waiting class-2 customers exceeds the inventory remaining after all x discounted units are sold, and overbooking occurs. Note $y_L \leq y_H$.

Lemma 1.3 *$r(\theta, S, x)$ is increasing in S and concave in θ.*

Proof. By the definition of y_L and y_H:

$$g(S, x, \hat{y}) = \begin{cases} p_2 S + p_1(\hat{y}D_1 - S), & \text{if } \hat{y} \leq y_L; \\ p_2 S + p_1 x + p_2 B(S, x, \hat{y}), & \text{if } y_L < \hat{y} \leq y_H; \\ p_2 S + p_1 x + (p_2 - p_c)B(S, x, \hat{y}) + (N - S - x)p_c, & \text{if } y_H < \hat{y}. \end{cases}$$

Therefore by assumptions (B1) and (B3), $g(S, x, \hat{y})$ is nondecreasing in $\hat{y}$ and S respectively.

Let $f_{\hat{Y}}(y; S)$ be the density of $\hat{Y}$ and let $\hat{S}$ be an arbitrary fixed value of S.

Then $r(\theta, S, x)$ is increasing in S because

$$\begin{aligned} \frac{\partial}{\partial S}\left(\int g(S, x, y)\mathrm{d}F_{\hat{Y}}(y; S)\right)|_{S=\hat{S}} &= \int \frac{\partial}{\partial S}(g(S, x, y)f_{\hat{Y}}(y; S))|_{S=\hat{S}}\mathrm{d}y \\ &= \int \left(\frac{\partial}{\partial S}g(S, x, y)\right)|_{S=\hat{S}}f_{\hat{Y}}(y, \hat{S})\mathrm{d}y + \int g(\hat{S}, x, y)\left(\frac{\partial}{\partial S}f_{\hat{Y}}(y; S)\right)|_{S=\hat{S}}\mathrm{d}y \\ &\geq \int g(\hat{S}, x, y)\left(\frac{\partial}{\partial S}f_{\hat{Y}}(y; S)\right)|_{S=\hat{S}}\mathrm{d}y \\ &= \frac{\partial}{\partial S}\int g(\hat{S}, x, y)\mathrm{d}F_{\hat{Y}}(y; S) \\ &\geq 0 \end{aligned}$$

The first inequality holds since $g(S, x, \hat{y})$ is nondecreasing in S and the second holds by Theorem 1.1 since $g(S, x, \hat{y})$ is increasing in $\hat{y}$ while $\hat{Y}$ is stochastically increasing in S.

Similarly, $r(\theta, S, x)$ is concave in θ by Theorem 1.1 because $g(S, x, \hat{y})$ is nondecreasing in $\hat{y}$ and $\hat{y}$ is stochastically concave in θ. ■

Lemma 1.4 *$r(\theta, S, x)$ is concave in x if $\frac{\partial B}{\partial x} \geq -p_1/p_2$.*

Proof. Let $\hat{x}$ be the solution to $B(S, \hat{x}, \hat{y}) = N - S - \hat{x}$. Since B is decreasing in x and $\partial B/\partial x \geq -1$ it follows that if $0 \leq x \leq \hat{x}$ then $B(S, x, \hat{y}) \leq N - S - x$, and conversely if $\hat{x} < x \leq N - S$ then $B(S, x, \hat{y}) > N - S - x$. Since $B(S, x, \hat{y})$ is nonnegative, $\hat{x} \leq N - S$.

Consider two cases:

Case 1: if $\hat{y}D_1 > N$ then $x \leq N - S \leq \hat{y}D_1 - S$. So from (1.1) we obtain

$$g(S, x, \hat{y}) = \begin{cases} p_2 S + p_1 x + p_2 B(S, x, \hat{y}), & \text{if } 0 \leq x \leq \hat{x}; \\ (p_2 - p_c)S + (p_1 - p_c)x + (p_2 - p_c)B(S, x, \hat{y}) + p_c N, & \text{if } \hat{x} < x \leq N - S. \end{cases}$$

which is concave because it consists of two concave segments (since B is concave in x on $[0, \hat{y}D_1 - S]$ by the assumption (B6) and $p_2 \geq p_c$), and $\frac{\partial g}{\partial x}|_{x \leq \hat{x}} = p_1 + p_2 \frac{\partial B}{\partial x} \geq (p_1 - p_C) + (p_2 - p_C)\frac{\partial B}{\partial x} = \frac{\partial g}{\partial x}|_{x > \hat{x}}$ (since $\partial B(S, x, \hat{y})/\partial x \geq -1$ by the assumption (B2) and $p_c \geq 0$).

Case 2: if $\hat{y}D_1 \leq N$ then no overbooking can occur (i.e. $B(S, x, \hat{y}) \leq N - S - x$), and so

$$g(S, x, \hat{y}) = \begin{cases} p_2 S + p_1 x + p_2 B(S, x, \hat{y}), & \text{if } 0 \leq x \leq \hat{y}D_1 - S; \\ p_2 S + p_1(\hat{y}D_1 - S), & \text{if } \hat{y}D_1 - S < x \leq N - S. \end{cases}$$

which consists of a concave segment (for $x \leq \hat{y}D_1 - S$) and a flat segment (for $\hat{y}D_1 - S < x \leq N - S$). If $\partial B/\partial x \geq -p_1/p_2$ then $g(S, x, \hat{y})$ is also increasing on $x \leq \hat{y}D_1 - S$, and so $g(S, x, \hat{y})$ is concave on $0 \leq x \leq N - S$.

Combining these two cases, $g(S,x,\hat{y})$ is concave in x, and since concavity is maintained under expectation over exogenous variable (recall distribution of $\hat{Y}$ is independent of x), $r(\theta,S,x)$ is also concave in x. ■

We note that concavity of $r(\theta,S,x)$ in x could hold under a milder condition than that of Lemma 1.4, which implies that $g(S,x,\hat{y})$ is concave in x. Specifically, g does not have to be everywhere concave. It can be shown that g is not concave for $\hat{y} \in \mathbf{Y}$, where $\mathbf{Y} = \left[\hat{y}|\hat{y}D_1 \leq N, \frac{\partial B}{\partial x} \leq -p_1/p_2\right]$. Thus, if the distribution of $\hat{Y}$ places small enough probability on such subset, $\mathbf{Y}$, then the expected revenue function, r, is still concave. We do not discuss this point any further.

Comparing $\frac{\partial B}{\partial x}$ to the ratio of prices is important since if $\frac{\partial B}{\partial x} \geq -p_1/p_2$ then (not taking overbooking into account) the firm gets higher revenue from putting more units on sale; however, this revenue is offset by paying overbooking penalties. As a result of this trade-off the optimal number of units on sale in the single period can be determined by solving the first-order condition $\partial r(\theta,S,x)/\partial x = 0$, and is given by the following corollary:

Corollary 1.1 *The single-period optimal number of units on last-minute sale, x^*, satisfies*

$$\begin{aligned} & p_1\left(1 - F_{\hat{Y}|(\theta,S)}(\frac{S+x^*}{D_1})\right) + p_2\int_{\frac{S+x^*}{D_1}}^{\overline{y}} \frac{\partial B(S,x,y)}{\partial x}|_{x=x^*}\mathrm{d}F_{\hat{Y}|(\theta,S)}(y) \\ = \; & p_C\left(1 - F_{\hat{Y}|(\theta,S)}(y_H(x^*)) + \int_{y_H(x^*)}^{\overline{y}} \frac{\partial B(S,x,y)}{\partial x}|_{x=x^*}\mathrm{d}F_{\hat{Y}|(\theta,S)}(y)\right). \end{aligned} \tag{1.5}$$

Proof. Substituting from (1.1) and (1.2) we obtain

$$\begin{aligned} r(\theta,S,x) \; = \; & \int_{\underline{y}}^{\frac{S+x}{D_1}} (p_2S + p_1(yD_1 - S))\,\mathrm{d}F_{\hat{Y}|(\theta,S)}(y) \\ + \; & \int_{\frac{S+x}{D_1}}^{y_H(x)} (p_2S + p_1x + p_2B(S,x,y))\,\mathrm{d}F_{\hat{Y}|(\theta,S)}(y) \\ + \; & \int_{y_H(x)}^{\overline{y}} ((p_2 - p_C)S + (p_1 - p_C)x + (p_2 - p_C)B(S,x,y) + p_CN)\,\mathrm{d}F_{\hat{Y}|(\theta,S)}(y) \end{aligned}$$

where the limits of the integration follows from the definitions of y_L and y_H.

Differentiating $r(\theta, S, x)$ in x we obtain

$$\begin{aligned}
\frac{\partial r(\theta,S,x)}{\partial x} &= \frac{1}{D_1}\left(p_2 S + p_1 x\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(\frac{S+x}{D_1}) && (1.6)\\
&+ \int_{\frac{S+x}{D_1}}^{y_H(x)} \left(p_1 + p_2 \frac{\partial B(S,x,y)}{\partial x}\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(y) && (1.7)\\
&+ \frac{\partial y_H(x)}{\partial x}\left(p_2 S + p_1 x + p_2 B(S,x,y_H(x))\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(y_H(x)) && (1.8)\\
&- \frac{1}{D_1}\left(p_2 S + p_1 x\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(\frac{S+x}{D_1}) && (1.9)\\
&+ \int_{y_H(x)}^{\overline{y}} \left((p_1 - p_C) + (p_2 - p_C)\frac{\partial B(S,x,y)}{\partial x}\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(y) && (1.10)\\
&- \frac{\partial y_H}{\partial x}\Bigg((p_2 S + p_1 x + p_2 B(S,x,y_H) && (1.11)\\
&\qquad + p_C(B(S,x,y_H) - (N-S-x))\Bigg)\mathrm{d}F_{\hat{Y}|(\theta,S)}(y_H)
\end{aligned}$$

Observe the terms in (1.6) and (1.9) are identical. The terms in (1.8) and (1.11) are also identical, since by the definition of y_H, $B(S,x,y_H(x)) = N - S - x$. Therefore the derivative simplifies to

$$\begin{aligned}
\frac{\partial r(\theta,S,x)}{\partial x} &= \int_{\frac{S+x}{D_1}}^{y_H(x)} \left(p_1 + p_2 \frac{\partial B(S,x,y)}{\partial x}\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(y) && (1.12)\\
&+ \int_{y_H(x)}^{\overline{y}} \left((p_1 - p_C) + (p_2 - p_C)\frac{\partial B(S,x,y)}{\partial x}\right) \mathrm{d}F_{\hat{Y}|(\theta,S)}(y) && (1.13)
\end{aligned}$$

By setting $\partial r / \partial x = 0$, and rearranging the terms we arrive to the equation in (1.5) ■

The provided first-order condition has an appealing intuitive interpretation in the light of the newsvendor model. The firm selects the number of units to place on last-minute sale such that to balance the revenues from selling an extra unit at price p_1 plus the revenues from selling at price p_2 to the remaining B class-2 customers, with the losses from paying the overbooking penalty to the overflow customers. Therefore x^* satisfies the following logic, which is reflected in (1.5):

"p_1Prob($S + x^* + 1^{st}$ unit is sold at p_1) + p_2Prob($B(S, x^* + 1, \hat{y})^{st}$ unit is sold at p_2)=

p_CProb($S + x^* + 1^{st}$ unit is overbooked)."

Next we discuss the supermodularity of $r(\theta, S, x)$. Supermodularity of the expectation of a function can be established by Theorem 1.1 given that the function itself is supermodular (see Section 3.10.1 in Topkis (1998)). Therefore since $\hat{Y}$ is stochastically supermodular in (θ, S), if $g(S, x, \hat{y})$ is supermodular in $(S, x, \hat{y})$ then $r(\theta, S, x)$ is also supermodular in (θ, S, x). Supermodularity of $g(S, x, \hat{y})$ is studied in the following lemma:

Lemma 1.5 *$g(S, x, \hat{y})$ is supermodular in $(S, x, \hat{y})$ if either (i) $p_C > 0$, $\underline{y}D_1 \geq N$, $\frac{\partial B}{\partial S} = -1$ and $\frac{\partial B}{\partial x} = -1$; (ii) $p_C > 0$, $\overline{y}D_1 \leq N$, $\frac{\partial B}{\partial S} = -p_1/p_2$ and $\frac{\partial B}{\partial x} \geq -p_1/p_2$; (iii) $p_C = 0$, $\underline{y}D_1 \geq N$ and $\frac{\partial B}{\partial S}$, $\frac{\partial B}{\partial x}$ are unrestricted; or (iv) $p_C = 0$, $\overline{y}D_1 \leq N$ and $\frac{\partial B}{\partial S} = -p_1/p_2$ and $\frac{\partial B}{\partial x} \geq -p_1/p_2$.*

Proof. Let $p_C > 0$ and consider two cases:

Case 1: if $\hat{y}D_1 > N$ then $x \leq N - S \leq \hat{y}D_1 - S$. From (1.1) we obtain

$$g(S, x, \hat{y}) = \begin{cases} p_2 S + p_1 x + p_2 B(S, x, \hat{y}), & \text{if } 0 \leq x \leq \hat{x}; \\ (p_2 - p_c)S + (p_1 - p_c)x + (p_2 - p_c)B + p_c N, & \text{if } \hat{x} < x \leq N - S. \end{cases}$$

and so

$$\frac{\partial g(S, x, \hat{y})}{\partial S} = \begin{cases} p_2 + p_2 \frac{\partial B}{\partial S}, & \text{if } 0 \leq x \leq \hat{x}; \\ (p_2 - p_c) + (p_2 - p_c)\frac{\partial B}{\partial S}, & \text{if } \hat{x} < x \leq N - S. \end{cases}$$

Supermodularity in (S, x) and (S, y) demands that the above derivative increases in x and y respectively. Therefore it implies $\partial B/\partial S \leq -1$. However, since $\partial B/\partial S \geq -1$ by the assumption (B3), it follows that necessarily $\partial B/\partial S = -1$. By the same logic, supermodularity in (x, y) requires $\partial g/\partial x$ to increase in y, leading to $\partial B/\partial x \leq -1$, which under the assumption (B2) results in $\partial B/\partial x = -1$.

Case 2: if $\hat{y}D_1 \leq N$ then

$$g(S,x,\hat{y}) = \begin{cases} p_2S + p_1x + p_2B(S,x,\hat{y}), & \text{if } 0 \leq x \leq \hat{y}D_1 - S; \\ p_2S + p_1(\hat{y}D_1 - S), & \text{if } \hat{y}D_1 - S < x \leq N - S. \end{cases}$$

and so

$$\frac{\partial g(S,x,\hat{y})}{\partial S} = \begin{cases} p_2 + p_2\frac{\partial B(S,x,\hat{y})}{\partial S}, & \text{if } 0 \leq x \leq \hat{y}D_1 - S; \\ p_2 - p_1, & \text{if } \hat{y}D_1 - S < x \leq N - S. \end{cases}$$

Observe that $\partial g/\partial S$ increases in $\hat{y}$ if $\partial B/\partial S \geq -p_1/p_2$, while $\partial g/\partial S$ increases in x if $\partial B/\partial S \leq -p_1/p_2$. Therefore in this case necessarily $\partial B/\partial S = -p_1/p_2$. Same as above, supermodularity in (x,y) requires $\partial g/\partial x$ to increase in y, leading to $\partial B/\partial x \geq -p_1/p_2$.

Finally, if $p_C = 0$ then in Case 1, $g(S,x,\hat{y}) = p_2S + p_1x + p_2B(S,x,\hat{y})$, which is supermodular following assumption (B1), and therefore the conditions on B resemble those of Case 2. ■

A direct corollary to the above Lemma is that to achieve supermodularity the model must be restricted to the cases of only "high" demand ($\underline{y}D_1 \geq N$) or only "low" demand ($\overline{y}D_1 < N$). If $p_C > 0$, and demand is high, then $B = f(\hat{y}) - S - x$ for some function $f(\cdot)$; i.e., the firm allocates the inventory on sale first to class-2 customers. If demand is low, then, for any p_C, the above conditions imply that the inventory is allocated to both classes, such that the proportion of class-2 customers is at least p_1/p_2. Finally, if $p_C = 0$ and demand is high, then there are no additional restrictions on allocation.

Joint concavity can be established by the following lemma:

Lemma 1.6 *If $g(S,x,\hat{y})$ is linear in $(x,\hat{y})$, then $r(\theta,S,x)$ is jointly concave in (θ,x).*

Proof. Since r is concave in θ and concave in x by Lemmas 1.3 and 1.4 respectively, joint concavity of $r(\theta,S,x)$ in (θ,x) requires the determinant of the Hessian (H) to be non-negative; where

$$H = \frac{\partial^2 r}{\partial \theta^2}\frac{\partial^2 r}{\partial x^2} - \left(\frac{\partial^2 r}{\partial \theta \partial x}\right)^2 \tag{1.14}$$

If $g(S, x, \hat{y})$ is linear in $(x, \hat{y})$, then $\frac{\partial g}{\partial x}$ is not a function of $\hat{y}$. So, from (1.2), $\frac{\partial^2 r}{\partial \theta \partial x} = \frac{\partial g}{\partial x} \int \mathrm{d}\partial F_{\hat{Y}|\theta}/\partial \theta = 0$ since $\int \mathrm{d}F_{\hat{Y}|\theta} = 1$ for all θ. With this from (1.14) $H = \frac{\partial^2 r}{\partial \theta^2}\frac{\partial^2 r}{\partial x^2} \geq 0$ ■

Note that since r is supermodular, x^* is nondecreasing in θ (see Section 2.8 in Topkis 1998). So r should be concave only when x and θ are changing in the same direction. Recognizing this, however, we were unable to derive easy interpretable conditions on the parameters of the model, which would ensure joint concavity in the above sense.

Summarizing, we have the following theorem:

Theorem 1.3 *If an allocation function $B(S, x, \hat{y})$, and the other parameters of the model satisfy conditions (B1) - (B6), and those of Lemmas 1.4 - 1.6, then the expected single-period revenue $r(\theta, S, x)$ satisfies assumptions (A1)-(A3) and Theorem 1.2 holds. That is the revenue-to-go function $J_t(\theta_t, S_t, x_t)$ is concave in x_t for all t.*

As an example of an allocation function that satisfies these assumptions, for the case with $p_C = 0$ and $\underline{y}D_1 \geq N$, consider $B(S, x, \hat{y}) \equiv \hat{y}D_2 - S - x\frac{D_2}{D_1}$. Here, the number of remaining class-2 customers reflects the total number of class-2 customers that wait, $\hat{y}D_2 - S$, net the number of class-2 customers that purchased product on last-minute sale. Further, the discounted units are allocated on proportion, which is constant and depends on the nominal demands, D_2/D_1. We note that letting the proportion depend on the realized demands, rather then nominal, would result in a more accurate representation of the actual allocation on proportion. However, it would complicate the model beyond tractability in the general setting. We use such realized proportions later in the simplified models of Section 1.5. With this, $g(S, x, \hat{y}) = \hat{y}p_2D_2 + x\left(p_1 - p_2\frac{D_2}{D_1}\right)$, which is supermodular in $(S, x, \hat{y})$, and linear in $(x, \hat{y})$. Therefore the conditions of Theorem 3 hold and the expected revenue-to-go function is concave for every period, and so the optimal number of units on sale is easy to find.

Summarizing our results for the case with a self-regulating learning function, we showed that the revenue function is concave and therefore the firm places some units on sale in every period. Furthermore, since the revenue function is supermodular, the number of units it places on sale increases in the waiting parameter. But, the self-regulating learning behavior controls the number of customers waiting in the subsequent period so that it does not continue to increase; that is, the firm takes a passive role, placing some units on sale, and relies on the consumer behavior to control future waiting. This is not be the case of smoothing learning function, where the firm must actively manage consumer waiting as we discuss below.

1.5 Optimal Policy for Smoothing Learning Function

In this section we assume that the learning function $h_t(\theta_t, x_t)$ is smoothing; that is, the next period's waiting parameter, θ_{t+1}, increases in both the current waiting parameter, θ_t, and the number of units on last-minute sale in period t, x_t. We show that in the general model the revenue-to-go function is not necessarily concave, unless the speed of consumer learning is "slow" as defined below. To address the problems with arbitrary speed of learning we present two simplified models and show that for either simplification, the optimal policy has a "bang-bang" structure where the firm alternately places a number, $\hat{x}_t$, or zero units on sale. We describe this optimal policy in the closed form.

In the general model recall that supermodularity of the revenue-to-go function is required by Theorem 1.2 to establish concavity. Since the learning function is linear, concavity and supermodularity of the revenue-to-go require, respectively,

$$\frac{\partial^2}{\partial x^2} J(\theta, S, x) = \frac{\partial^2}{\partial x^2} r(\theta, S, x) + \delta \frac{\partial^2 \phi}{\partial h^2} \left(\frac{\partial h}{\partial x} \right)^2 \leq 0 \text{ and} \tag{1.15}$$

$$\frac{\partial^2}{\partial x \partial \theta} J(\theta, S, x) = \frac{\partial^2}{\partial x \partial \theta} r(\theta, S, x) + \delta \frac{\partial^2 \phi}{\partial h^2} \frac{\partial h}{\partial x} \frac{\partial h}{\partial \theta} \geq 0. \tag{1.16}$$

In the case of a self-regulating learning function, both inequalities hold if $\frac{\partial^2 \phi}{\partial h^2} \leq 0$, since by definition a self-regulating learning function satisfies $\frac{\partial h}{\partial x} \frac{\partial h}{\partial \theta} \leq 0$. In the case of a smoothing learning function, however, $\frac{\partial h}{\partial x} \frac{\partial h}{\partial \theta} \geq 0$, and so concavity and supermodularity of the revenue function place contradictory requirements on $\frac{\partial^2 \phi}{\partial h^2}$. Therefore we conclude that the revenue function is not necessarily concave.

Observe that $\partial h / \partial x$ reflects the speed at which customers learn about the firm's decisions. Specifically, we say that customer learning is *slow* if $\partial h / \partial x$ is small; otherwise we say it is *fast*. From (1.15) if $\partial h / \partial x$ is small enough then J would be concave, regardless of ϕ, given that r is concave. Similarly, J would be supermodular, provided that r is supermodular; that is, if customer learning is slow enough, then concavity and supermodularity of the expected revenue function for every period are equivalent to the corresponding single-period property, and so can be established as per Section 1.4.2.

Slow learning has been documented in the works on reference price learning with respect to the sales promotions. Greenleaf (1995) and Hardie et al. (1993) studied point-of-sales data for such commodities as peanut butter and refrigerated orange juice, and reported an analog of our $\partial h / \partial x$ to be at 0.075 and 0.17 respectively. However, we know of no research regarding the speed of learning for the last-minute specials in services. This is of interest for future research.

1.5.1 Simplified Models

In this section we simplify the general model so that upon observing the signal, the firm can infer the *exact* number of customers waiting for period t. The future demand and purchasing behavior remain random. This simplification allows us to solve the problem in the closed form, at the same time utilizing a more realistic allocation function and relaxing the linearity assumption of the learning function.

Let $\alpha_t \in [\underline{\alpha}, \overline{\alpha}] \subseteq [0,1]$ be a fraction of the regular price demand that waits for the last-minute sale in period t. We refer to α_t as the *waiting fraction.* Then $M_t \equiv \alpha_t Y_t D_2$ and $S_t \equiv (1-\alpha_t)Y_t D_2$, where Y_t is the (unconditional) demand multiplier.

In this section we consider two simplifications:

(i) Deterministic waiting fraction model. In this model we assume that demand multiplier, Y_t, is stochastic, but its distribution does not depend on the waiting parameter; waiting fraction is deterministic with $\alpha_t \equiv \theta_t$, and it evolves according to a smoothing concave learning function $\alpha_{t+1} = h_t(\alpha_t, x_t)$;

(ii) Deterministic demand model. In this model we assume that $Y_t \equiv const$ for all $t = 1, 2, ...T$ and w.l.o.g. we set $Y_t \equiv 1$; the random waiting fraction, α_t, is stochastically increasing and concave in the waiting parameter θ_t, which evolves according to a smoothing concave learning function $\theta_{t+1} = h_t(\theta_t, x_t)$.

Let $A_t(\alpha_t, Y_t)$ be the actual demand for the discounted seats. There are M_t class-2 and all $Y_t d_1$ class-1 customers waiting (recall they wait since $p_2 > p_1$). Therefore $A_t(\alpha_t, Y_t) = Y_t(d_1 + \alpha_t D_2) = Y_t D_1 - S_t$. Since the firm puts x_t units on last-minute sale, $\min[x_t, A_t]$ units are sold at the discounted price p_1. Assuming that the discounted inventory is allocated proportionally between class-1 and class-2 customers based on their *realized* demands, the number of class-2 customers that purchase discounted packages is $\min[x_t, A_t]\frac{\alpha_t Y_t D_2}{A_t}$ and the corresponding allocation function is

$$B_t(\alpha_t, Y_t, x_t) = \alpha_t Y_t D_2 \left(1 - \frac{\min[x_t, A_t]}{A_t}\right). \tag{1.17}$$

The above allocation expresses proportional allocation based on realized demands.

With these from (1.1) the net single-period revenue is

$$\begin{aligned} g_t(\alpha_t, Y_t, x_t) &= p_2 S_t + p_1 \min[x_t, A_t] + p_2 \alpha_t Y_t D_2 \left(1 - \frac{\min[x_t, A_t]}{A_t}\right) \\ &- p_C \left(\alpha_t Y_t D_2 \left(1 - \frac{\min[x_t, A_t]}{A_t}\right) - (N - S_t - x_t)\right)^+ \end{aligned} \tag{1.18}$$

Observe that for either model by knowing θ_t and observing S_t the firm can determine the exact (realized) values for α_t and y_t, and therefore at the last minute there is no uncertainty for the current period. If $A_t + S_t = y_t D_1 < N$, then the firm has *excess* capacity and no overbooking can occur. Otherwise the capacity is *scarce*, and overbooking can occur if too many units are put on sale. Since the firm knows which case realizes with certainty, it forces an intuitive restriction $p_C \geq p_1$. Otherwise overbooking would imply *intentional* selling discounted products and later bumping class-1 customers (as overflow) for a premium of $p_1 - p_C > 0$.

1.5.2 Single-Period Solution for the Simplified Models

Recall that $A(\alpha, y) = y(d_1 + \alpha D_2)$ and let $\hat{x}(\alpha, y)$ be the maximum number of discounted units such that: (i) all class-2 customers are allocated a product without overbooking; and (ii) all $\hat{x}(\alpha, y)$ units are sold. In the case of excess capacity the firm cannot sell all available inventory, and so $\hat{x}(\alpha, y) = A(\alpha, y)$. Otherwise discounting too many units could result in overbooking. Therefore solving $\alpha y D_2 \left(1 - \frac{x}{A(\alpha,y)}\right) - (N - S - x) = 0$ leads $\hat{x} = A(\alpha, y) \frac{N - S - \alpha y D_2}{A(\alpha,y) - \alpha y D_2} < A(\alpha, y)$ and the solution is unique since the equation is linear in

x. Observe $\hat{x} \leq N - S$. In summary

$$\hat{x}(\alpha, y) = \begin{cases} A(\alpha, y), & \text{if } A \leq N - S \text{ (excess capacity case);} \\ A(\alpha, y)\frac{N-S-\alpha y D_2}{A(\alpha,y)-\alpha y D_2}, & \text{if } A > N - S \text{ (scarce capacity case).} \end{cases} \tag{1.19}$$

The single-period optimal policy for the simplified models is given by the following theorem:

Theorem 1.4 *In either simplified model there exists a threshold waiting fraction α^*, such that if $\alpha \geq \alpha^*$ then $x^* = 0$. Otherwise in the case of scarce capacity $x^* = \hat{x}(\alpha, y)$, and in the case of excess capacity any $x \in [\hat{x}(\alpha, y), N - S]$ is optimal.*

Proof. In the case of excess capacity if $x > \hat{x} = A$, then $g(\alpha, y, x) = p_2(1 - \alpha)yD_2 + p_1 y(d_1 + \alpha D_2)$ which is independent of x.

In the case of scarce capacity by definition $A > N - S \geq x$. If $x > \hat{x}$, then $g(\alpha, y, x) = p_2 S + p_1 x + (p_2 - p_C)\alpha y D_2 \left(1 - \frac{x}{A}\right) + p_C(N - S - x)$, and so $\frac{\partial g}{\partial x} = (p_1 - p_c) - (p_2 - p_C)\frac{\alpha y D_2}{A} \leq 0$ as $p_2 \geq p_C \geq p_1$. Therefore, $x^* \in [0, \hat{x}(\alpha, y)]$. On this interval $g(\alpha, y, x) = p_2 S + p_1 x + p_2 \alpha y D_2 \left(1 - \frac{x}{y(d_1 + \alpha D_2)}\right)$, which is linear in x, and so the optimal solution in on the boundary of the interval; that is $x^* \in \{0; \hat{x}(\alpha, y)\} = \Pi_t$.

We next show the existence and uniqueness of the threshold waiting fraction. Let $C(\alpha) = \frac{\partial g}{\partial x}$ and observe $C(\alpha) = p_1 - \frac{\alpha p_2 D_2}{d_1 + \alpha D2}$. Differentiating it we obtain $\frac{\partial C}{\partial \alpha} = -\frac{p_2 D_2 d_1}{(d_1 + \alpha D_2)^2} \leq 0$. Therefore α^* solves $C(\alpha^*) = 0$. Since $C(0) = p_1 > 0$ and C is monotonically decreasing, α^* is unique.

Finally, for $\alpha \geq \alpha^*$, $C(\alpha) \leq 0$; that is $g(\alpha, x, y)$ is nonincreasing in x and so $x^* = 0$. Otherwise the maximal revenue is attained at $x^* = \hat{x}(\alpha, y)$. In the case of excess capacity, however, the revenue function is flat on $[\hat{x}(\alpha, y), N - S]$, and so in this case any $x \in [\hat{x}(\alpha, y), N - S]$ is optimal if $\alpha \leq \alpha^*$. ■

We note that $\alpha^* \leq 1$ if $D_1 p_1 \leq D_2 p_2$; that is, if the revenue from the regular-price segment is larger then from the low-price one, then the single-period optimal policy is "bang-bang": the optimal number of discounted units drops down to zero if too many customers wait (i.e., when $\alpha \geq \alpha^*$); it jumps up to $\hat{x}$ otherwise.

Next we prove that a similar "bang-bang" policy holds for every period.

1.5.3 Multiple-Period Solution for Simplified Models

Let $R_t(\theta_t, \alpha_t, y_t, x_t)$ be the expected revenue-to-go, given that for period t, the waiting parameter is θ_t, the realized waiting fraction is α_t, the observed demand multiplier is y_t and x_t units are put on last-minute sale. The optimal number of discounted units, x_t^*, can be found for each period, $t = 1, 2, ...T$, by solving the following dynamic program:

$$R_t(\theta_t, \alpha_t, y_t, x_t) = g_t(\alpha_t, y_t, x_t) + \delta E_{(\alpha_{t+1}, y_{t+1}) | \theta_{t+1} = h_t(\theta_t, x_t)} \left[R^*_{t+1}(\theta_{t+1}, \alpha_{t+1}, y_{t+1}) \right] \quad (1.20)$$

where the optimal revenue-to-go is given by

$$R^*_t(\theta_t, \alpha_t, y_t) = \max_{0 \leq x_t \leq N - (1-\alpha_t) y_t D_2} R_t(\theta_t, \alpha_t, y_t, x_t) \quad (1.21)$$

and

1. $g_t(\alpha_t, y_t, x_t)$ is given by (1.18)
2. $R^*_{T+1}(\theta_{T+1}, \alpha_{T+1}, y_{T+1}) = 0$ for all $(\theta_{T+1}, \alpha_{T+1}, y_{T+1})$
3. $\theta_{t+1} = h_t(\theta_t, x_t)$, is increasing and concave in either argument (smoothing)

As in Section 1.4.1, observe that $\theta_{t+1} = h_t(\theta_t, x_t)$ does not depend on the realized values of α_t and y_t. Therefore we can substitute $R_t(\theta_t, \alpha_t, y_t, x_t) = g_t(\alpha_t, y_t, x_t) + \delta \phi_{t+1}(h_t(\theta_t, x_t))$.

In our two simplified models ϕ_{t+1} takes the following specific forms:

- *In the deterministic waiting fraction (DW) model,* $\alpha_t \equiv \theta_t$ for all t by the assumption and the distribution of Y_t is independent of α_t. Therefore

$$\phi_{t+1}^{DW}(h_t(\alpha_t, x_t)) = E_{y_{t+1}}\left[R_{t+1}^*(h_t(\alpha_t, x_t), y_{t+1})\right] \tag{1.22}$$

- *In the deterministic demand (DD) model,* $Y_t \equiv 1$ for all t by the assumption, and so w.l.o.g. y can be dropped from the expectation of the future revenue, leading to

$$\phi_{t+1}^{DD}(h_t(\theta_t, x_t)) = E_{(\theta_{t+1}, \alpha_{t+1})|\theta_{t+1} = h_t(\theta_t, x_t)}\left[R_{t+1}^*(\theta_{t+1}, \alpha_{t+1})\right] \tag{1.23}$$

Let $\Pi_t = \{x_t : R_t(\theta_t, \alpha_t, y_t, x_t) = R_t^*(\theta_t, \alpha_t, y_t) \text{ and } 0 \leq x_t \leq N - (1-\alpha_t)y_y D_2\}$ be the set of "potentially optimal" solutions for period t. Our main result for the simplified models is that $\Pi_t = \{0; \hat{x}_t\}$ for all $t = 1, 2, ...T$. The concept of our proof is the following. Suppose that the expected future revenue, ϕ_{t+1}, is decreasing and convex in x_t and α_t. Since $g_t(\alpha_t, x_t, y_t)$ is piecewise linear in x, R_t consists of two adjacent and convex segments. Since g is also decreasing for $x_t > \hat{x}_t$, R_t is also decreasing if $x_t > \hat{x}_t$. Therefore $x_t^* \in \{0; \hat{x}_t\} \equiv \Pi_t$. We summarize this result in the theorem below.

Theorem 1.5 *In either simplified model,* $x_t^* \in \Pi_t = \{0; \hat{x}_t\}$ *for all periods* $t = 1, 2, ...T$.

Proof. Recall that we consider two simplified models: the one with a deterministic waiting fraction, given by (1.22), and the one with deterministic demand, given by (1.23). We first prove the theorem for the deterministic waiting case, and then extend it to deterministic demand case. We require the following three lemmas.

Lemma 1.7 $g(\alpha, y, x)$ *is decreasing convex in* α *for* $x \in \Pi_t$.

Proof. If $x = 0$ then $g(\alpha, y, 0) = p_2 y(1-\alpha)D_2 + p_2\alpha y D_2 = p_2 y D_2$, which is not a function of α, and therefore is decreasing convex in a weak sense.

If $x = \hat{x}$ then in the excess capacity case $g(\alpha, y, \hat{x}(\alpha, y)) = p_2(1-\alpha)yD_2 + p_1(d_1 + \alpha D_2)$ which is linear decreasing in α since $\frac{\partial g}{\partial \alpha} = -yD_2(p_2 - p_1) \leq 0$ because $p_2 \geq p_1$. In the scarce capacity case observe that $\frac{\hat{x}}{A} = \frac{N-S-\alpha y D_2}{yd_1 + \alpha y D_2 - \alpha y D_2} = \frac{N-S-\alpha y D_2}{yd_1}$, which is linear in α. Therefore $g(\alpha, y, \hat{x}(\alpha, y)) = p_2 y(1-\alpha)D_2 + p_1 x + (p_2 - p_C)\frac{N-S-\alpha y D_2}{yd_1} + p_C(N - S - x)$, which is linear decreasing in α since $\frac{\partial g}{\partial \alpha} = -\frac{D_2(p_2-p_1)(N-yD_2)}{d_1} \leq 0$ as $yD_2 \leq \bar{y}D_2 \leq N$ and $p_2 \geq p_1$ by the assumption, and $\frac{\partial^2 g}{\partial \alpha^2} = 0$. ■

Lemma 1.8 *If $f(x)$ is decreasing convex and $g(x)$ is increasing concave, then $f(g(x))$ is decreasing convex.*

Proof. follows by the chain rule.

Convex: $\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial g^2}\left(\frac{\partial g}{\partial x}\right)^2 + \frac{\partial f}{\partial g}\frac{\partial^2 g}{\partial x^2} \geq 0$, because f is decreasing convex and g is concave.

Decreasing: $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} \leq 0$, because f is decreasing, and g is increasing. ■

Lemma 1.9 *Let $f(x, y)$ be decreasing and(or) convex in x for all $y \in Y$, then*

(a) $\sum_{y \in Y} f(x, y)$ *is decreasing and(or) convex in x;*

(b) $\sup_{y \in Y} f(x, y)$ *is decreasing and(or) convex in x;*

Proof. follows from Theorems 5.2 and 5.5 in Rockafellar (1970). ■

(i) Deterministic waiting model. For period T the claim is implied by Theorem 1.4. For $x_T \in \Pi_T$, by Lemma 1.7, the single-period revenue, $g_T(\alpha_T, y_T, x_T)$, is decreasing and convex in α_T. Therefore since $R_T^*(\theta_T, \alpha_T, y_T) = \max_{x_T \in \Pi_T} g_T(\alpha_T, y_T, x_T)$, R_T^*, is also decreasing and convex in α_T by part (b) of Lemma 1.9.

In the deterministic waiting model $\alpha_T \equiv \theta_T = h_{T-1}(\theta_{T-1}, x_{T-1})$, and so by the above R_T^* is also decreasing and convex in $h_{T-1}(\cdot)$. Therefore its expectation over y_T, ϕ_T^{DW} as per (1.22), is decreasing and convex in $h_{T-1}(\cdot)$ by part (a) of Lemma 1.9.

Let $1 \leq t < T$ and suppose that for every $n \in [t+1, T]$, $\phi_n(h_{n-1}(\theta_{n-1}, x_{n-1}))$ is decreasing convex in h_{n-1}. Then by Lemma 1.8, in period $t+1$, ϕ_{t+1} is decreasing and convex in θ_t and x_t.

Recall that $g_t(\alpha_t, y_t, x_t)$ is piecewise linear in x_t on $[0, N - S_t]$ with the breakpoint at $x_t = \hat{x}_t$, and further recall that $g_t(\alpha_t, y_t, x_t)$ is decreasing in x_t on $x_t \in (\hat{x}_t, N - S_t]$. Since ϕ_{t+1} is decreasing and convex in x_t, $R_t(\theta_t, \alpha_t y_t, x_t)$ consists of two adjacent segments both convex in x_t, and $R_t(\theta_t, \alpha_t y_t, x_t)$ decreases in x_t on $x_t \in (\hat{x}_t, N - S_t]$. Thus $x_t^* \in [0, \hat{x}_t]$. Finally, since the revenue-to-go function is convex on this interval, $x^* \in \{0; \hat{x}_t\} \equiv \Pi_t$.

Therefore it is sufficient to prove that the induction assumption holds for period t; that is that ϕ_t is decreasing and convex in h_{t-1}.

For $x_t \in \Pi_t$, the single-period revenue, $g_t(\alpha_t, y_t, x_t)$, is decreasing and convex in α_t by Lemma 1.7. The future revenue, ϕ_{t+1}, is also decreasing and convex in α_t by the induction assumption, upon noting that in the deterministic waiting model $\alpha_t \equiv \theta_t$. Therefore $R_t(\theta_t, \alpha_t, y_t, x_t)$ is decreasing and convex in α_t. And so by part (b) of Lemma 1.9, R_t^* is also decreasing and convex in α_t. Finally, since in deterministic waiting model y_t is independent of α_t, by part (a) of Lemma 1.9 $\phi_t = E_{y_t}[R_t^*]$ is decreasing and convex in α_t, and therefore in h_{t-1} (since $\alpha_t \equiv \theta_t = h_{t-1}$).

(ii) Deterministic demand model. For period T the claim is implied by the single-period result, given in Theorem 1.4.

In the deterministic demand model, observe that $R_T^*(\theta_T, \alpha_T, 1) = \max_{x_T \in \Pi_T} g(\alpha_T, 1, x_T)$ is independent of θ_T. Thus, from (1.23), $\phi_T^{DD} = E_{\alpha_T | h_{T-1}}[R_T^*(\alpha_T)]$, which is decreasing and convex in h_{T-1} by Theorem 1.1, since by the above R_T^* is decreasing in α_T, and α_T is stochastically increasing and concave in θ_T by our assumption. Therefore by Lemma 1.8, ϕ_T is decreasing and convex in θ_{T-1} and x_{T-1}, since $h_{T-1}(\theta_{T-1}, x_{T-1})$ is increasing and concave.

Let $1 \leq t < T$ and suppose that for every $n \in [t+1, T]$, $\phi_n(h_{n-1}(\theta_{n-1}, x_{n-1}))$ is decreasing convex in h_{n-1}. Then by Lemma 1.8, in period $t+1$, ϕ_{t+1} is decreasing and convex in θ_t and x_t. With this $x_t^* \in \{0; \hat{x}_t\} \equiv \Pi_t$ by the same argument as in the proof of Theorem 1.5, and it remains to prove that ϕ_t is decreasing and convex in h_{t-1}.

Since α_t is stochastically increasing and concave in $\theta_t = h_{t-1}(\cdot)$, the vector (θ_t, α_t) is stochastically increasing and concave in h_{t-1}.

For $x_t \in \Pi_t$, single-period revenue $g_t(\alpha_t 1, x_t)$, is decreasing in α_t by Lemma 1.7 and independent of θ_t. Future revenue, ϕ_{t+1}, is independent of α_t, and is decreasing in h_t by the induction assumption, and therefore by Lemma 1.8 ϕ_{t+1} is also decreasing in θ_t. Thus $R_t(\theta_t, \alpha_t, y_t, x_t)$ is decreasing in (θ_t, α_t). And by part (b) of Lemma 1.9, R_t^* is also decreasing in (θ_t, α_t). From (1.23) $\phi_t = E_{(\theta_t, \alpha_t)}[R_t^*]$ is decreasing and convex in $h_{t-1}(\cdot)$ by Theorem 1.1 because (θ_t, α_t) is stochastically increasing and concave in h_{t-1}. ■

To summarize, for the case with a smoothing learning function for both simplified models, the optimal policy is "bang-bang"; it places either 0 or $\hat{x}_t$ units on last-minute sale depending on the realized waiting fraction, α_t. Furthermore, since the future revenue, ϕ_{t+1} is decreasing in x_t, and if $\alpha_t \geq \alpha^*$, then g_t is also decreasing in x_t, it follows that if $\alpha_t \geq \alpha^*$, then $x_t^* = 0$ for all t. That is the firm offers units on sale and increases the number of customers waiting, and then periodically holds no sale, withdrawing revenue from waiting class-2 customers and decreasing future waiting. By following such policy the firm simultaneously achieves high utilization of its capacity, and at the same time controls the number of customers waiting. This policy is quite different from that of the self-regulating case, because the firm actively manages the waiting, as opposed to relying on the consumers to control the waiting themselves.

Observe that since $\hat{x}_t \leq N - S_t$ overbooking is not optimal. This is because the marginal revenue p_2 per unit could as well be obtained from the initial sales, and since the future revenue is decreasing in x_t, the firm puts fewer units on sale and reduces the future waiting. This is not the case if the firm can obtain a marginal revenue in excess of p_2, as happens in the three price model that we study next.

1.6 Three Price Model with Deterministic Waiting

In this section we study the case where a firm may choose to offer some units for sale at p_1, while raising the price to a higher value for the remaining inventory. By doing so the firm can both capture the low-price demand, as well as the demand willing to pay extra for being accommodated at the last minute. The three-price model reflects frequently observed situations where the walk-up price is higher than the regular, while some units have been sold at a discount earlier (e.g., airlines, car rentals or hotels).

Let $p_3 \geq p_2$ be the "high" price, and let D_3 be the number of customers who are willing to pay p_3 (the nominal demand). Then the actual demand at price p_3 in period t is $Y_t D_3$. As before, at the last minute the firm decides x_t, the number of units to offer at the discounted price p_1. The remaining units are offered at price p_3.

To determine the revenue of the firm, recall that $S_t \equiv (1-\alpha_t)Y_t D_2$ units are sold at the initial price p_2, and $\min[x_t, A_t]$ units are sold at the discounted price p_1, where $A_t \equiv Y_t(d_1 + \alpha_t D_2)$.

Observe that only class-3 customers purchase at price p_3. Let $\psi_t(\alpha_t) \in [0,1]$ be the fraction of class-3 customers who wait for a deal, given that there is a fraction $\alpha_t \in [\underline{\alpha}; \overline{\alpha}]$ of class-2 and -3 customers waiting combined. That is the total number of class-3 customers waiting is $\psi_t(\alpha_t)Y_t D_3$, and the number of class-2 customers is $\alpha Y_t D_2 - \psi_t(\alpha_t)Y_t D_3$. Since the latter is non-negative, it is implied that $\psi_t(\alpha_t)Y_t D_3 \leq \alpha Y_t D_2$ for all $\alpha_t \in [\underline{\alpha}; \overline{\alpha}]$.

As before, we assume that the discounted units are allocated on proportion. That is, the number of discounted units that are sold to class-3 customers is $\min[x_t, A_t]\frac{\psi_t(\alpha_t)Y_tD_3}{A_t}$ and the net single-period revenue is $g_t(\alpha_t, Y_t, x_t) = p_2S_t + p_1\min[x_t, A_t] + p_3\psi_t(\alpha_t)Y_tD_3\left(1 - \frac{\min[x_t, A_t]}{A_t}\right) - p_C\left(\psi_t(\alpha_t)Y_tD_3\left(1 - \frac{\min[x_t, A_t]}{A_t}\right) - (N - S_t - x_t)\right)^+$.

In this section we assume that the demand multiplier, Y_t, is stochastic, and the waiting fraction, α_t, is deterministic[†]. We assume that the distribution of Y_t does not depend on α_t, and that in turn, α_t evolves according to a linear learning function $\alpha_{t+1} = h_t(\alpha_t, x_t)$. Note that we place no restriction whether $h(\cdot)$ is smoothing or self-regulating.

We consider two cases: one where overbooking is allowed and one where it is not. For each of these we define a representative waiting function, ψ.

If there is no overbooking, then since class-3 customers have a higher valuation for the product, they are more cautious to wait. Therefore we assume ψ is "small," compared with α, and class-3 customers do not wait unless many class-2 customers are already waiting. For example, if we assume that class-3 customers do not wait, unless *all* class-2 customers are already waiting, then $\psi(\alpha) = \max[0, 1 - \frac{D_2}{D_3} + \alpha\frac{D_2}{D_3}]$ for $\alpha \in [0, 1]$.

In the case with overbooking, class-3 customers do not have capacity concerns, as they are guaranteed a product at their reservation price, p_3 (since $y_tD_3 \leq \bar{y}D_3 \leq \bar{y}D_2 \leq N$). Therefore, they are more likely to wait, and we assume ψ is "large"; there may be a fraction of class-3 customers waiting even if no class-2 customers wait. For example $\psi(\alpha) = \frac{d_1}{D_1} + \alpha\frac{D_2}{D_1}$ for $\alpha \in [\frac{D_3(D_1-D_2)}{D_2(D_1-D_3)}, 1]$. In this case the fraction of class-3 customers that always wait is $\psi(\underline{\alpha}) = d_1/(d_1 + d_2)$.

Technically we assume that ψ is nondecreasing convex, and in the case with no overbooking $\psi = 0$ for $\alpha \leq \hat{\alpha}$ for some $\hat{\alpha} \in [\underline{\alpha}, d_2/D_2]$ and $\psi' D_3 = D_2$ otherwise, and in the case with overbooking $\psi' D_3 \leq D_2$ and $\psi' A(\alpha, y) \leq \psi y D_2$. Intuitively, the functions that satisfy

[†] Alternative models with stochastic demand and waiting, or deterministic Y and stochastic α, could in general be of interest as well. However, we found that in three price context they place restrictive and uninterpretable conditions on prices, demands and waiting. Therefore we do not present them.

these assumptions correspond to the "small" and "large" ψ as per the discussion above.

Then by redefining $\hat{x}$ to include the waiting of class-3 customers as

$$\hat{x}(\alpha, y) = \begin{cases} A(\alpha, y), & \text{if } A \leq N - S \text{ (excess capacity case);} \\ A(\alpha, y)\frac{N-S-\psi(\alpha)yD_3}{A(\alpha,y)-\psi(\alpha)yD_3}, & \text{if } A > N - S \text{ (scarce capacity case).} \end{cases} \tag{1.24}$$

by the same argument as in the proof of Theorem 1.4 we obtain that the single-period optimal policy resembles that of the two-price model.

Theorem 1.6 *In the three price model there exists a threshold waiting fraction α^*, such that if $\alpha \geq \alpha^*$ then $x^* = 0$. Otherwise in the case of scarce capacity $x^* = \hat{x}(\alpha, y)$, and in the case of excess capacity any $x \in [\hat{x}(\alpha, y), N - S]$ is optimal.*

Proof. Follows by the same argument as in the deterministic waiting case in the proof of Theorem 1.5. As in Lemma 1.7 we prove that $g_t(\alpha_t, y_t, x_t)$ is convex in α_t for $x_t \in \Pi_t$.

If $x = 0$ then $g(\alpha, y, 0) = p_2 S + p_3 y\psi(\alpha)D_3$, which is convex since ψ is convex.

If $x = \hat{x}$ then in the excess capacity case $g(\alpha, y, A(\alpha, y)) = p_2(1-\alpha)yD_2 + p_1(d_1 + \alpha D_2)$ which is linear in α. In the scarce capacity case, if overbooking is not allowed then g is piecewise linear convex. By our assumption, if $\alpha \leq \hat{\alpha}$ then $\psi(\alpha) = 0$ and so $\frac{\partial g}{\partial \alpha} = -yD_2(p_2 - p_1)$. If $\alpha > \hat{\alpha}$ then $\psi' = D_2/D_3$ and $\frac{\partial g}{\partial \alpha} = -yD_2(p_2 - p_1 E(\alpha, y)) + yD_2 p_3(1 - E(\alpha, y))$, where $E(\alpha, y) = \frac{N-S-\psi(\alpha)yD_3}{A(\alpha,y)-\psi(\alpha)yD_3} < 1$ and $\frac{\partial E}{\partial \alpha} = -\frac{(yD_1-N)(D_2-\psi'_\alpha D_3)}{y(A(\alpha,y)-\psi(\alpha)yD_3)^2} = 0$. The function is convex because $\frac{\partial g}{\partial \alpha}|_{\alpha \geq \hat{\alpha}} - \frac{\partial g}{\partial \alpha}|_{\alpha < \hat{\alpha}} = yD_2(p_3 - p_1)(1 - E(\alpha, y)) \geq 0$. If overbooking is allowed, then

$$\frac{\partial^2 g}{\partial \alpha^2} = \left(\frac{yD_3(yD_1 - N)(p_3 - p_1)}{(A(\alpha, y) - y\psi(\alpha)D_3)^3} \right) \tag{1.25}$$
$$(A(\alpha, y)\psi''_\alpha(A(\alpha, y) - y\psi(\alpha)D_3) + 2y(\psi'_\alpha D_3 - D_2)(\psi'_\alpha A(\alpha, y) - y\psi D_2)) \geq 0$$

because the numerator in the first term in (1.25) is positive since in the scarce capacity case $yD_1 \geq N$, and $p_3 > p_1$ by the definition. The denominator is positive since $A(\alpha, y) -$

$y\psi(\alpha)D_3 = yd_1 + y(\alpha D_2 - \psi(\alpha)D_3) \geq yd_1 \geq 0$ as $\alpha D_2 \geq \psi(\alpha)D_3$ (by the definition of ψ, since the number of waiting customers of class-2 is non-negative). In the second term, $A(\alpha, y)\psi''_\alpha(A(\alpha, y) - y\psi(\alpha)D_3) \geq 0$ since ψ is convex and by the above $A(\alpha, y) - y\psi(\alpha)D_3 > 0$. And $2y(\psi'_\alpha D_3 - D_2)(\psi'_\alpha A(\alpha, y) - y\psi D_2) \geq 0$ because by the assumptions on ψ both elements of the product are negative.

If $x = N - S$ then in the case of the excess capacity $N - S > A$ and therefore $g(\alpha, y, N - S) = p_2(1-\alpha)yD_2 + p_1 y(d_1 + \alpha D_2)$, which is linear in α. In the case of scarce capacity $N - S > \hat{x}$ and therefore $g(\alpha, y, N - S) = p_2(1-\alpha)yD_2 + p_1(N - (1-\alpha)yD_2) + (p_3 - p_C)\psi(\alpha)yD_3\,(1 - (N - (1-\alpha)yD_2)/A(\alpha, y))$. If overbooking is not allowed then $p_C = p_3$ and g is linear. Otherwise

$$\frac{\partial^2 g}{\partial \alpha^2} = \left(\frac{yD_3(yD_1 - N)(p_3 - p_C)}{A(\alpha, y)^3}\right) \qquad (1.26)$$
$$\left(A(\alpha, y)^2\psi''_\alpha + 2yD_2(y\psi D_2 - \psi'_\alpha A(\alpha, y))\right) \geq 0$$

because the first term in (1.26) is positive since in the scarce capacity case $yD_1 \geq N$, and $p_3 > p_C$ by the definition. And in the second term, ψ is convex, and by the assumption $y\psi D_2 \geq \psi'_\alpha A(\alpha, y)$. Therefore $g(\alpha, y, x)$ is convex in α for $x \in \Pi_t$, and the result follows.

■

In this case, $\alpha^* \leq 1$ if $D_1 p_1 \leq D_3 p_3$. That is, if the revenue from the high-price segment is higher than from the low-price one, then the single-period optimal policy is "bang-bang."

By redefining $\Pi_t = \{0; \hat{x}_t; N - S_t\}$ the multiple-period optimal solution is given below.

Theorem 1.7 *In the three-price model (either with or without overbooking), $x_t^* \in \Pi_t = \{0; \hat{x}_t; N - S_t\}$ for all periods $t = 1, 2, ...T$.*

In summary, in the model with three prices (customer classes), the optimal solution is the same for both the model with and without overbooking, provided that we account for the changes in the behavior of class-3 customers caused by allowing (or not allowing) overbooking and the corresponding availability concerns. The solution does not have the exact same "bang-bang" structure as in the two-price model, since the future revenue is not everywhere decreasing. Numerically, however, we observed that there still exists α^* such that if $\alpha_t > \alpha^*$, then $x_t^* = 0$. That is the optimal policy follows a pattern of increasing the fraction of customers waiting, and then periodically offering no sale, withdrawing revenue from waiting class-3 customers and decreasing future waiting.

The difference between the two- and the three-price models is that in the latter it could be optimal to overbook. This is because the firm has a possibility to obtain a high revenue, p_3, per unit, which cannot be obtained if the customers do not wait. Therefore the firm wants a number of customers to wait, and so puts more units on sale, even though it may result in paying overbooking penalties.

1.7 Numerical Studies

In this section we provide several examples to illustrate the value of making decisions optimally as compared with several heuristics managers use in real-life situations, and examine how this value and the optimal policy itself change in different situations. We also analyze numerically two extensions to our model: selecting the optimal discount price p_1^*, and allowing random allocation of discounted units rather than proportional. We use the three price model since it allows for all types of consumer behaviors that we study in the current paper.

We set $N = 100$, $\delta = 0.95$, $D_2 = 50$, $p_1 = 100$, $p_2 = 300$ and $p_3 = 500$, and consider four families of instances: smoothing overbooking (MB), smoothing nonoverbooking (MN), self-regulating overbooking (RB) and self-regulating nonoverbooking (RN). For each family of instances we study four demand curves, with $D_1 = 150$ or $D_1 = 100$ of class-1 customers, and $D_3 = 30$ or $D_3 = 10$ of class-3 customers. We denote these demand curves as "150-50-

30," "150-50-10," "100-50-30," and "100-50-10," respectively. For cases with overbooking we study two penalties: $p_C = 150$ and $p_C = 450$.

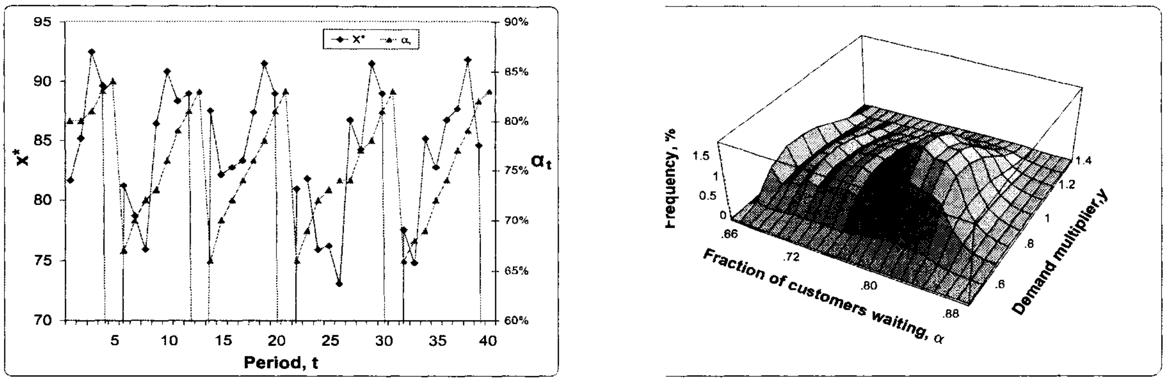
We use functions $h(\alpha, x) = \lambda \frac{x}{N} + (1-\lambda)\alpha$ and $h(\alpha, x) = (1-\lambda) + \lambda \frac{x}{N} - (1-\lambda)\alpha$, for $0 < \lambda < 1$ for the smoothing and self-regulating learning, respectively. We use $\psi(\alpha) = \frac{d_1}{D_1} + \alpha \frac{D_2}{D_1}$ and $\psi(\alpha) = \max[0; 1 - \frac{D_2}{D_3} + \alpha \frac{D_2}{D_3}]$ for waiting functions with and without overbooking, respectively.

We set $\underline{y} = 0.6$ and $\overline{y} = 1.4$, so that the demand multiplier $Y_t \in [0.6, 1.4]$, and examine three distributions: truncated $Normal[1, 0.2$[‡]$]$, $Beta[1.75, 3]$ and $Beta[0.8, 2]$, with the expected values and CVs, respectively, (1, 0.913, 0.841) and (0.181, 0.191, 0.237).

For each instance we compute the expected infinite horizon discounted revenue (the *revenue*), assuming that the system starts from steady state – intuitively, this is the revenue that the firm will generate starting at an arbitrary time in the future. To compute the revenue we discretize α_t as $\{0; 0.01; 0.02; \ldots 1\}$ and discretize y_t as $\{0.6; 0.7; \ldots 1.4\}$ for a total of 909 states. We use successive approximations with error bounds to compute the infinite horizon expected revenue function value (Bertsekas (1987), pp. 188-193). Then we determine the subset of recurrent states and the steady state probabilities, and obtain the expected revenue as the weighted sum (Puterman 1994, pp.589-594).

In Figure 1.2 (a) we present a sample path for the optimal decision, x^*, and the fraction of customers waiting, α_t, and in (b) we present the recurrent states and the frequencies with which they are visited in the steady state. We observe that the optimal decision and the fraction of customers waiting follow the cycles of variable length, see (a). This expresses the "bang-bang" structure of the optimal policy: if in period t the fraction of waiting customers, α_t, gets large enough, then the firm puts $x_t^* = 0$ units on sale and so α_{t+1} drops. Hence in period $t+1$ the firm puts $\hat{x}(\alpha_{t+1}, y_{t+1}) > 0$ units on sale and α_{t+2} gets up again; note 1-period time lag between x and α in (a). Following such cycles, the optimal policy with

[‡] We use 2σ limits so that the endpoints have the nonnegligible probabilities.



(a)

(b)

Figure 1.2: In (a): sample path of the optimal decision, x_t^*, and the fraction of customers waiting, α_t. In (b): frequencies of visiting different states by the optimal policy. Both examples are for the MN instance with 150-50-30 demand curve and Beta(0.8,2) demand multipliers; $\lambda = 0.2$.

nonzero probability visits a variety of states; see (b). Because of the uncertainty in demand the length of cycles is random; thus the customers cannot anticipate the transitions and hence the decision of the firm.

1.7.1 Performance of the Optimal Policy and Managerial Insights

To better understand the performance of the optimal policy we compare it with four heuristics that appeal to managers. In each heuristic we determine the number of units to put on sale through different methods. We consider the following:

Do-nothing: puts $x_t = N - S_t$ or $x_t = 0$ units on sale for all t, whichever is better;

BestP: puts $x_t = N - S_t$ units on sale with probability P^* and $x_t = 0$ with probability $1 - P^*$, where the value of P^* is the one that results in the highest revenue;

S*: selects $x_t = 0$ if $S_t > S^*$ and $x_t = N - S_t$ otherwise, where the value of S^* is the one that results in the highest revenue;

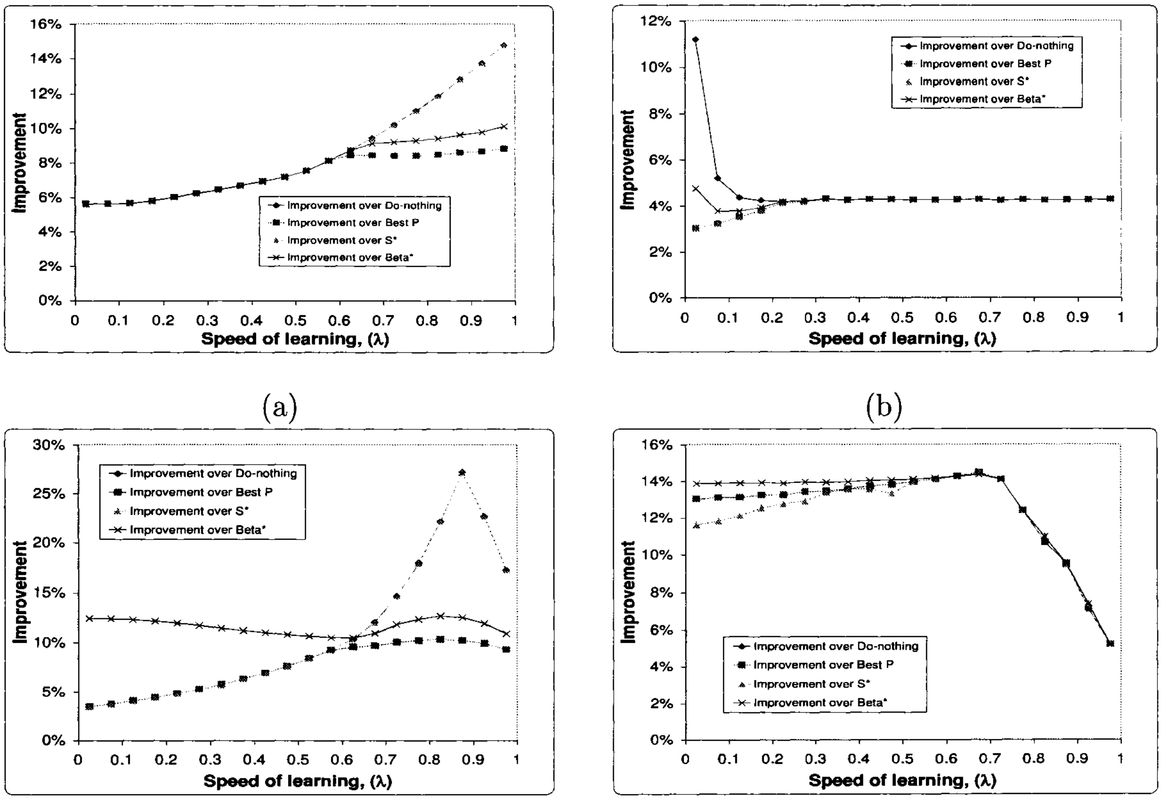
Beta*: puts $x_t = N - S_t$ with probability $\beta^* \frac{N-S_t}{N}$, and $x_t = 0$ with probability $1 - \beta^* \frac{N-S_t}{N}$, where the value of β^* is the one that results in the highest revenue.

The rationale behind a do-nothing heuristic is straightforward. We observed that there is no consistent "do-nothing" activity across the different instances, and therefore our do-nothing is the best of "all" or "nothing." The BestP heuristic attempts to prevent consumers from guessing if a sale will occur in a given period. A variation of the BestP heuristic is used by a car rental company with whom we discussed our work: they have a deal almost every week, but to access it, customers need a promotion code. These codes are e-mailed to a subset of their registered webmail customers, where a customer is included on the mailing list for a given week with some probability. Heuristics S* and Beta* represent a naïve managerial approach that holds that discounts should be offered in periods with low regular price sales (i.e., when S_t is small). The former heuristic does so when a threshold is crossed, whereas the latter places units on sale based on a linear probabilistic rule. We find P^*, S^* and β^* through numerical search. Given the revenues of the optimal and heuristic policies, we compute the *relative improvement* of the optimal policy over a particular heuristic as (optimal revenue − heuristic revenue)÷(heuristic revenue).

Figure 1.3 presents the relative improvements over the heuristics for the four families of instances with 150-50-30 demand curve and Beta(0.8;2) demand multipliers.

Our main observation is that in all cases the optimal policy generates five to fifteen percent additional revenue over the best heuristic. This value changes depending on the speed of learning and the type of consumer behavior. It also depends on which heuristic is the best.

In the instances without overbooking (Figure 1.3 (a) and (c)), the best heuristic is BestP, as it outperforms heuristics S* and Beta*. At a first glance this might seem slightly counterintuitive, since the latter are based on the intuitive managerial approach to put more units on sale when S_t is small. However, recall that the optimal policy suggest exactly the



(a) (b)

(c) (d)

Figure 1.3: The relative improvements for the (a) MN , (b) MB, (c) RN and (d) RB instances with 150-50-30 demand curve and Beta(0.8,2) demand multipliers. In (b) $p_C = 150$ and in (d) $p_C = 450$.

opposite to this naïve approach. Specifically, $x_t^* = 0$ if $\alpha_t > \alpha^*$, and since $S_t = (1-\alpha_t)y_tD_2$, it follows that (in expectation) it is "optimal" *not* to put units on sale in the periods with small S_t.

The improvement over the BestP heuristic depends on the speed of learning. This is because the optimal policy determines when to offer a discount (the timing), and if one is offered, then how many units to discount (the number). By choosing the best probability, the BestP heuristic "optimizes" the long-run average number of units on sale, but cannot achieve the right timing of sales. In the cases of slow learning in order to change waiting behavior, the firm must have consistent series of periods with and without discounts. The BestP heuristic cannot ensure such consistency, and therefore chooses to do nothing (indeed

Instances where it is optimal to offer last-minute discounts	Instances where it is optimal to do nothing
MN, RN 100-50-30	MB, RB 100-50-30
MB, RB 150-50-10	MN, RN 150-50-10
MB, RB 100-50-10	MN, RN 100-50-10
MN, MB, RN, RB 150-50-30	

Table 1.1: Strategic use of overbooking in order to increase value from offering last-minute discounts.

$P^* = 0$ for $\lambda < 0.625$ on Figure 1.3 (a) and for $\lambda < 0.575$ on Figure 1.3 (c)). For faster speeds of learning, consistency is not required as customers readily change their waiting behavior, and therefore the timing of sales is less important than the average number of units on sale.

In the instances with overbooking (Figure 1.3 (b) and (d)), the best heuristic is S*. This is because the S* heuristic puts units on sale only when it is appropriate, and so has a direct control over the timing, as opposed to the probabilistic heuristics which do not. Better timing implies lower overbooking penalty, so the S* heuristic returns higher revenue. As the total penalty is proportional to p_C, the improvement is also larger in the cases with large p_C (compare (b) and (d) in Figure 1.3).

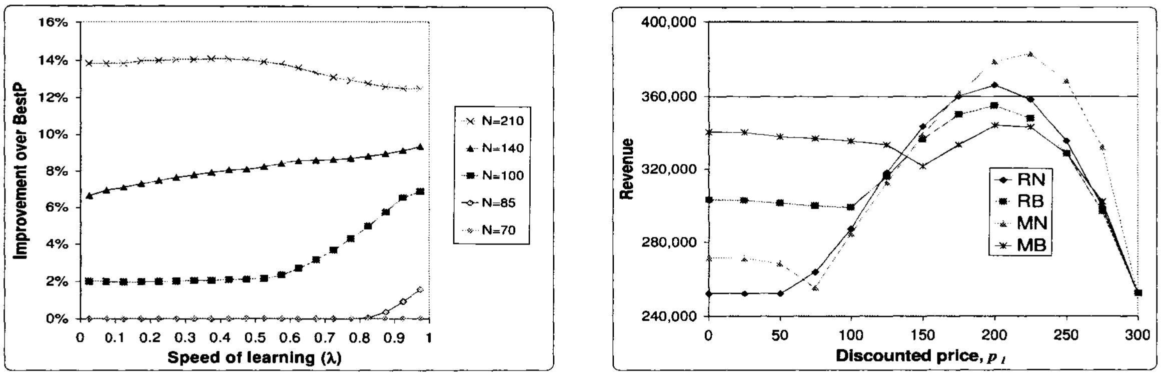
Next we study the factors that influence whether the strategic revenue management as we discuss in this paper will be effective. Table 1.1 classifies different instances into those where the firm benefits from offering last-minute discounts and those where it is optimal to do nothing. Observe that the firm can increase the value of its revenue management policy by strategically allowing or not to overbook. In the cases with few class-1, but many class-3 customers (row 1 in Table 1.1), it is advantageous for the firm not to allow overbooking, while in the cases with few class-3 customers (rows 2 and 3 in Table 1.1), overbooking is preferable. Such distinction can be made, because in order to generate value from last-minute sale the firm must ensure that the number of discounted units sold to high-value class-3 customers is not disproportionately large. This can happen when there are many class-3, but few class-1 customers. Thus in such a case the firm must force more of class-3 customers to buy early, which it does by not allowing overbooking. This allows the firm to simultaneously offer discounts to achieve high utilization and lose little revenue since few discounted units are sold to high-value customers, while occasionally sell inventory at p_3 to waiting class-3 customers. In the reverse case with few high-value customers, because of proportional allocation, even if all class-3 customers wait, they are displaced by customers of class-1, and hence only few end up buying at a discount, while most buy at p_3. Thus the firm provides an incentive for more class-3 customers to wait by allowing overbooking. In the remaining cases the sizes of classes are balanced and, regardless of overbooking, the firm can take advantage of periodic discounts to generate extra revenue.

We considered alternative distributions of demand multipliers: truncated Normal (1;0.2), Beta(1.75;3) and Beta(0.8;2). In our experiments, the instances with Beta(0.8;2) resulted in the largest improvements because this distribution has the lowest expected demand and the highest CV.

Figure 1.4(a) presents the relative improvements under different capacity scenarios. Observe that the value of offering discounts optimally increases in capacity. For $N = 70$ it is optimal not to put any units on sale, and therefore the improvement is zero. It is intuitive, since no discounts should be offered if the (expected) utilization of capacity is high enough. As capacity increases (i.e., the expected utilization decreases) the firm offers discounts, customers wait, and it becomes more valuable to manage this waiting optimally.

1.7.2 Selecting the Optimal Discount Price, p_1^*

Our model assumes that the discount price, p_1, is fixed for the entire horizon of T periods, and as we argue in the introduction, building a model where customers react to both



(a) (b)

Figure 1.4: The effects of capacity, N, and discounted price, p_1, on the optimal revenue. In (a): relative improvement over the BestP heuristic at different capacity levels for MN instance with 150-50-30 demand curve and truncated Normal(1,0.2) demand multipliers. In (b): Optimal revenue as a function of discounted price, p_1, in the instances with Beta(0.8,2) demand multipliers.

price and availability of discounted units is nontrivial. As research in dynamic pricing shows, however, a heuristic that charges an optimally selected single price (as opposed to optimizing it dynamically) often performs only marginally suboptimally. Therefore, as a heuristic policy, the firm could search for the optimal "static" discounted price, p_1^*, charge it in every period, and then determine the number of discounted units to put on sale following our optimal policy.

We search for such optimal static price, p_1^*, numerically over its domain, $[0, p_2]$; see Figure 1.4 (b). In this example, we impose a demand curve $D(p) = 200 - 0.5p$ (so that $D(300) = 50$ and $D(100) = 150$ as in previous examples). We also assume that the value of a discount influences the rate at which customers are willing to change their behavior, i.e., the speed of learning. In particular we assume $\lambda(p) = 0.3 - 0.001p$. We experimented with other functions for demand and speed of learning, but observed no qualitative differences from the case presented.

Two observations are evident from Figure 1.4 (b). First, the optimal revenue is not concave and often not quasiconcave, therefore it may not be possible to advance in determining the optimal discount analytically. There is an obvious optimal point, however. Such a point exists because the firm uses class-1 demand to achieve two goals. On one hand it wants p_1 to be high, since then it obtains larger revenue from the sale of each discounted unit. On the other hand, it wants $D(p_1)$ to be high, because under proportional allocation class-1 customers displace some waiting class-3 customers, such that they purchase at even larger price, p_3. Since price negatively affects demand, these goals conflict and thus at some price level, further increase in price becomes disadvantageous.

Figure 1.4 (b) also provides a neat illustration to the earlier point that firms can strategically use overbooking to increase revenue. Observe that when the discounted price is small, hence, class-1 demand is high, more revenue is obtained when the firm does overbook. Conversely, when the price is large, hence, class-1 demand is small, the firm benefits from not overbooking. This observation further supports our findings presented in Table 1.1 and the discussion after it.

1.7.3 Random Allocation of Discounted Units

Our second extension replaces allocation on proportion with a more realistic random allocation that reflects the "first come, first served" practice of selling discounted products. In particular, we assume that all waiting customers have equal probability of arrival. By construction, there are $\psi(\alpha)yD_3$ customers of class-3 and $A - \psi(\alpha)yD_3$ customers of classes 1 and 2 waiting for the x units of discounted product. Therefore, $B = \psi(\alpha)yD_3 - \mathit{Hypergeometric}(x, \psi(\alpha)yD_3, A)$; here we assume that all quantities are integers. Further, since $E[B] = \psi(\alpha)yD_3(1 - \frac{\min[x,A]}{A})$, our proportional allocation based on realized demands substitutes B with its expected value (disregarding rounding).

The results of such random allocation are not known to the firm. Therefore in order to determine the optimal number of units on sale, the firm must evaluate the expected revenue over such random allocations both in single period and in multiple periods (note that single-period revenue is not random under proportional allocation). Hence, for each state we simulate 1,000[§] random allocation trials and select the decision that yields the highest average revenue.

We compare such "optimal" decisions for random allocation derived using simulation with our optimal policy (derived analytically assuming proportional allocation). We observed average differences between the number of units put on sale by the analytical and simulated policies for $p_C = 200, 300, 400$ to be 2.2, 3.6 and 4.2 percent (1.6, 2.6 and 3.1 units), respectively. That is, the simulated decision is more conservative. This is intuitive, since, if, due to randomness, the number of class-3 customers that remain waiting exceeds the available capacity then the firm must pay overbooking penalties or lose potential revenue from sale at p_3. Our simulations confirmed this intuition: in general, random allocation causes fewer units to be put on sale[¶]. At the same time, the difference is minimal, often within one unit, which could be attributed to rounding. That is, substituting the random variable with its expectation in our simulations does not lead to a significant error. This is because in the analytical optimal policy when $x^* > 0$ then $x^* = \hat{x}$ and is relatively large (e.g., in Figure 1.2 (b) typically $\hat{x} = 75...90$ units). Since B, is a sum over x^* random trials determining whether a unit on sale is allocated to a customer of class-3 or not, by the central limit theorem the actual realizations of B are scattered closely around its average value, assumed by the (deterministic) proportional allocation. Further, since $\hat{x} < N - S$, decreasing the number of units on sale by typically just one or two the firm ensures that large overbooking penalties occur rarely. Since the overbooking cost increases

[§] We selected 1,000 trials, because this number provides a good balance between computational efficiency and convergence.

[¶] There are several exceptions, where the firm puts more units on sale under random allocation than under proportional, however, we believe they are caused by the random scatter around a positive mean difference.

in p_C, the difference between the simulated and analytically optimal numbers of units on sale also increases with overbooking penalty.

Because the simulated "optimal" policy is so close to the analytical, the revenue generated by following the analytical optimal policy when the discounted units are actually allocated randomly is also close to that when they are allocated on proportion. In our simulations the difference between these revenues never exceeded three percent. As such, for the examples we studied, simplifying the actual random allocation with proportional allocation results in a robust policy.

1.8 Conclusions

Our work is motivated by the concern that given the increased ability to search for better prices for travel related products (flights, vacation packages, etc.), consumers will learn to expect last-minute deals and will strategically wait for them. In this paper we consider how a firm should offer last-minute discounts over a series of selling periods, taking into account that future customer behavior is influenced by the firm's decisions. We present a model that incorporates both stochastic demand and stochastic customer waiting behavior. We consider two alternate waiting behaviors, one in which customers interpolate between their previous waiting likelihood and their observation of the firm's policy (the smoothing case), and a second in which they anticipate other customers' behavior and the likelihood that they will receive a unit on sale (the self-regulating case). We study the problem for cases of two and three customer classes. The two-class problem represents the case where a list price is given (as in the cruise or vacation packages industries); the three-class problem reflects typical airline pricing where prices may decrease or increase in the days prior to departure. We formulate the problem as a dynamic program and develop a solution approach amenable to the novel structural properties we find in the problem.

For the case of two customer classes, we show that under self-regulating customer behavior, the revenue function is concave in the number of units placed on sale at the last minute when the discounted units are allocated in a reasonable manner. That is, the firm in general will set some units on sale in each period and allow the customer behavior to limit the number receiving the benefit of the reduced price inventory. In contrast, in the case of smoothing customer behavior, we show the firm should follow a "bang-bang" sale policy, either placing most of the remaining units on sale or none. Thus, the firm takes a more active role, adjusting the customers' expectations by alternately increasing the number of customers waiting until a threshold is crossed, upon which the firm places no units on sale. By doing so, the firm is able to regulate the number of customers waiting and to increase its revenue by increasing utilization, allowing some units that would otherwise not be sold to be purchased by the lower-value customers.

In the model with three customer classes, we consider the effects of overbooking. If overbooking is not allowed (e.g. in guaranteed sales such as cruises), customers must balance potential price discounts with potential stockouts of inventory. If overbooking is allowed (e.g., for airlines where customers denied boarding are compensated), customers balance waiting for reduced prices with the risk of potentially higher prices. We show that a similar policy is optimal, provided that one appropriately accounts for the allowance or not of overbooking. Through numerical simulations we find that following the optimal policy the firm can obtain a benefit of five to fifteen percent more revenue over several reasonable heuristics that firms might follow. We show that allowing overbooking increases this benefit when there are few high-value customers and similarly show that disallowing overbooking increases the benefit when there are many high value customers. We also test the effects of relaxing modeling assumptions about fixed prices and deterministic proportional allocation and show that, with respect to these, the derived optimal policy is very robust.

We acknowledge that the model formulated here does not account for all factors that may influence strategic customer behavior in revenue management. Future studies should consider more explicit formulations of customer utility, competition between firms and gaming by both firms and customers. In addition, empirical work is needed to better understand consumer learning behavior with regards to travel-related discounts.

Bibliography

[1] Anderson, C. K., J. G. Wilson. 2003. Wait or Buy? The Strategic Consumer: Pricing and Profit Implications. J. Oper. Res. Soc. 54 (3) 299-306.

[2] Anderson, C. K., J. G. Wilson. 2006. Optimal Booking Limits in the Presence of Strategic Consumer Behavior. Working paper, University of Western Ontario.

[3] Aviv, Y., A. Pazgal. 2003. Optimal Pricing of Seasonal Products in the Presence of Forward-Looking Consumers. Working paper, Washington University, St. Louis.

[4] Belobaba, P. P. 1989. Application of Probabilistic Decision Model to Airline Seat Inventory Control. Operations Research. 37(2) 183-197.

[5] Bertsekas, D. P. 1987. Dynamic Programming. Prentice-Hall, New Jersey.

[6] Besanko, D., W. L. Winston. 1990. Optimal Price Skimming By a Monopolist Facing Rational Consumers. Management Science 36(5) 555-567.

[7] Bitran, G. R., R. Caldentey. 2003. An Overwiev of Pricing Models for Revenue Management. M&SOM. 5(3) 203-229.

[8] Bitran, G. R., S. V. Mondschein. 1997. Periodic Pricing of Seasonal Products in Retailing. Management Science. 43(1) 64-79.

[9] Chesson, H., W. Kip Viscussi. 2000. The Heterogenetiy of Time-risk Tradeoffs. J. Beh. Dec. Making. 13 251-258.

[10] Conlisk, J., E. Gerstner, J. Sobel. 1984. Cyclic Pricing by a Durable Goods Monopolist. The Quarterly Journal Of Economics. 99(2) 489-505.

[11] Cooper, W. L., T. Homem-de-Mello, A. J. Kleywegt. 2004. Models of the Spiral-Down Effect in Revenue Management. Forthcoming in Operations Research.

[12] Dana, J. D. Jr. 1999. Using Yield Management to Shift Demand When the Peak Time is Unknown. The RAND J. of Eco. 30(3) 456-474.

[13] De Lisser, E. 2002. Cranky Consumer: Booking a Last-Minute Ticket. Wall Street Journal. July 2, 2002. D2.

[14] Elmaghraby, W., A. Gülcü, P. Keskinocak. 2003. Optimal Markdown Mechanisms in the Presence of Rational Consumers with Multi-unit Demands. Working paper, Georgia Institute for Technology.

[15] Fenton, B. and Z. Griffin. 2004. 4pm Peak Holiday Deals Go Begging. The Daily Telegraph. July 14 2004. 01.

[16] Gale, I. L., T. J. Holmes. 1993. Advance Purchase Discounts and Monopoly Allocation of Capacity. American Eco. Rev. 83 (1) 135-146.

[17] Gallego, G., G. van Ryzin. 1994. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons, Mgt. Sci. 40(8) 999-1020.

[18] Greenleaf, E. A. 1995. he Impact of Reference Price Effects On The Profitability of Price Promotion. Marketing Science. 14(1) 82-104.

[19] Hardie, B. G. S., E. J. Johnson, P.S. Fader. 1993. Modelling Loss Aversion and Reference Dependence Effects on Brand Choice. Mkt. Sci. 12 (4) 378-394.

[20] Hartveldt, H. H., S. R. Epps, B. Tesch, T. McHarg 2006. The Mainstreaming Of The Web Traveler. Forrester Research Report

[21] Lazear, E. P. 1986. Retail Pricing and Clearance Sales. The American Economic Review. 76(1) 14-32.

[22] Levin, Y., J. McGill, M. Nediak. 2006. Optimal Dynamic Pricing of Perishable Items by a Monopolist Facing Strategic Consumers. Working paper, Queens University.

[23] Liu Q., G. van Ryzin. 2005. Strategic Capacity Rationing to Induce Early Purchases. Working paper. Columbia University.

[24] Loewenstein, G. 1997. Anticipation and the Valuation of Delayed Consumption. Eco. J. 97 666-684

[25] McGill, J. I., G. J. Van Ryzin. 1999. Revenue Management: Research Overview and Prospects. Transportations Science. 33(2) 233-256.

[26] Nowlis, S. M., N. Mandel, D. B. McCabe. 2004. The Effect of a Delay between Choice and Consumption on Consumption Enjoyment. J. Cons. Res. 31 502-510

[27] Popescu, I., Y. Wu. Dynamic Pricing Strategies under Repeated Interactions. Working paper, INSEAD.

[28] Puterman M. L. 1994. Markov Decision Processes. John Wiley & Sons, New York.

[29] Rockafellar, T. R. 1997. Convex Analysis. Princeton University Press, New Jersey.

[30] Sen A., A. X. Zhang. 1999. The Newsboy Problem With Multiple Demand Classes. IIE Transactions. 31 (5) 431-444.

[31] Smith, J., K. F. McCardle. 2002. Structural Properties of Stochastic Dynamic Programs. Operations Research. 50 (5) 769-809

[32] Sobel, J. 1984. The Timing of Sales. Rev. of Eco. Studies. 51(3) 353-368.

[33] Stokey, N. L. 1979. Intertemporal Price Discrimination. The Quarterly Journal of Economics. 93(3) 355-371.

[34] Stringer, K. 2002. Airlines Now Offer 'Last Minute' Fare Bargains Weeks Before Flights. Wall Street Journal, March 15, 2002. B1.

[35] Su, X. 2006. Inter-temporal Pricing with Strategic Customer Behavior. Working paper. University of California, Berkeley.

[36] Talluri, K. T., G. J. van Ryzin. 2004. The Theory and Practice of Revenue Management. Springer, New York.

[37] Tang, C. S., K. Rajaram, A. Alptekinoglu, J. Ou. The Benefits of Advance Booking Discount Programs: Models and Analysis. Mgt Sci 50(4) 465-478.

[38] Topkis, D. M. 1998. Supermodularity and Complimentarity. Princeton U. Press. New Jersey.

[39] Xie, J., S. M. Shugan. 2001. Electronic Tickets, Smart Cards and Online Prepayments: When and How to Advance Sell. Mkt. Sci. 20(3) 219-243.

Chapter 2

Constructing Balanced Work Groups

2.1 Introduction

On many occasions organizations face series of diversified tasks, and the exact details of the future tasks are frequently unknown. In order to better address the solvability of these tasks, organizations often create stable heterogenous work groups, which are known to perform better then homogenous groups on complex projects and problem solving tasks (Hackman 1990, Kirchmeyer and McLellan 1990, McShane 1992).

In this two-chapter essay we study management science techniques that can be used to construct such groups. Initially this work was motivated by an applied project of constructing effective MBA study groups. Therefore the essay parents both the theoretical and applied results.

The first part of this essay, Chapter 2, describes the applied group construction project. In contrast to a more common approach of trying to minimize deviations from perfectly designed ('balanced') groups, we, from the start, enforce 'perfect balance' with constraints. The idea was that these constraints will be relaxed when a perfectly balanced group design cannot be found. To our surprise, in the practical applications, these constraints did not have to be relaxed in four years of actual use of our group creation software - perfectly balanced groups were always found.

The second part of this essay, Chapter 3, is an attempt to provide an explanation to this counter intuitive phenomenon. After all, examples where the perfect balance cannot be achieved are relatively easy to construct.

2.2 Motivation: Creating MBA Study Groups at Rotman School

Ranked among the best business schools in the world, the University of Toronto's Joseph L. Rotman School of Management offers a number of research and degree programs, including the full-time and part-time MBA, a one-year Executive MBA, and several other graduate and undergraduate programs. In many of these programs, students are assigned to groups to work on group assignments and projects. Rotman faculty and staff view these groups as important learning tools preparing students for future teamwork environments.

The Rotman School is hardly unique in this regard. Indeed, the goal of business education is to produce professionals who are capable of driving future economic growth; to do so, they must collaborate consistently and productively with other people and organizations. Thus, developing the effective teamwork skills is important, comparable, in some cases, to developing the classical skills taught at the business school, such as proficiency in marketing or accounting. In Rotman MBA programs, nearly 40 percent of the course work is group based, on average. Therefore the school must ensure that the composition of the student groups leads to effective group-based learning.

2.3 Group Work and the Creation of Balanced Groups - the Background

Researches in organizational behavior (OB) have found that heterogeneous well-balanced groups, whose members possess diverse personal characteristics and backgrounds, are more effective than homogenous groups on complex projects and problem solving tasks (Hackman 1990, Kirchmeyer and McLellan 1990, McShane 1992).

The management science and decision analysis literature includes a large number of studies of the problem of creating well-balanced student groups. Baker and Powell (2002) provide an excellent overview and cite studies that provide a theoretical analysis of the problem and describe the methods that could be used to construct student groups. However, it is often not clear how these methods could be used to create practical decision support systems. For example, Desrosiers, Mladenovic and Villeneuve (2005) describe a sophisticated solution method with hardware requirements that would be infeasible in a typical MBA office. A notable exception is Weitz and Jelassi (1992), who developed and implemented a group assignment system at INSEAD that was later implemented at NYU in a modified version. Their system is a variation of the people sequencing heuristic of Beheshtian-Arkedani and Mahmood (1986).

2.4 Rotman Study Groups

The Rotman School has used study groups for years. Initially, the faculty members allowed students to choose their own partners for the study groups. Over the years, in an effort to make groups more effective, the school centralized the process, with the MBA office personnel assigning students to study groups upon their enrollment in the program. They expect students to stay in these groups for their first (compulsory) year of the MBA studies.

Over time a number of concerns have arisen, which can be loosely classified into two issues: group composition and work splitting.

The first issue arises from the potential lack of balance among groups. It is generally desirable to create heterogeneous groups that include students of different genders, different cultural and different academic backgrounds. The experience at Rotman seems to conform (at least anecdotally) to the OB theory that the more diverse the groups are, the deeper perspective each student can gain from his or her peers. Students whose first language is not English are a special concern, because they may find courses with heavy writing requirements difficult; they should be spread evenly among the groups. Students with degrees in technical disciplines (engineering, mathematics, computer science) may have an advantage in quantitative courses, such as statistics, while students with degrees in humanities and social sciences may have an advantage in other courses, such as organizational behavior. Ideally the groups should be balanced with respect to all such factors. In other words, we would like to see homogeneity (balance) between groups and heterogeneity (diversity) within groups.

The self-selection approach of letting the students pick their own group partners tends to lead to the opposite results. Students tend to form groups with people they knew as undergraduates, people of the same cultural background, or people similar to them in other ways, leading to within-group homogeneity (it was not uncommon to see an all-Chinese group or an all-engineers group when students were allowed to form their own groups) and between-group heterogeneity.

Work splitting was another major concern, gaining its importance as the weight of group-based assignments increased in the Rotman curriculum. Students tend to split group projects within the group, assigning work from different classes to the group members with perceived strength in that area. For example, a group faced with projects in statistics and organizational behavior would often split up the work, assigning the statistics project to the one or two students with strong statistics backgrounds, and the organizational behavior project to the students who were psychology majors. They would thus diminish the supposed benefits of group work (since individuals or very small homogeneous subgroups would do the projects), and risk subverting the learning process: in many cases, the students who do the assignments have the strongest prior background in that particular area and thus gain the least benefit by doing the additional work. This is a particular concern in an MBA program, where the diversity of student backgrounds can be staggering, students' educational backgrounds in a particular subject area often range from an introductory undergraduate course to a PhD.

These issues have been further exacerbated by the rapid growth in the Rotman MBA program over the last few years: 275 students entered the full-time MBA program in 2003 (the class of 2005), an increase of 21 percent over the class of 2004 and 66 percent over the class of 2003. Apart from being large, the recent incoming classes are also very diverse: the class of 2005 includes students from 25 different countries, who vary in academic and industrial backgrounds, languages, religious and cultural norms.

2.5 Creating Multiple Balanced Groups

During the summer of 2002, the administration of the Rotman MBA program examined the way student groups were functioning and the group-formation process. In spite of the problems, they decided that the emphasis on group work should be retained because of its strong benefits; however the administration had to find creative ways to deal with the problems related to group composition and work splitting. In particular, they needed to come up with a strategy to ensure fairness, to prevent certain groups from having a priori advantages (either real or perceived) with respect to workload in particular classes. They decided that the school needed to revamp the group-creation process to ensure that all students are assigned to multiple, well-balanced groups.

The concept of balancing is based on the premise that all groups should contain roughly the same proportion of students meeting particular criteria; these criteria may include gender, English proficiency and educational background. Once the school selects the criteria the goal is to create groups that are as similar or well-balanced as possible with respect to all of the criteria.

What does a well-balanced group look like? Let us illustrate with two criteria: group size and gender. The class of 2005 consists of 275 students, and the administrators decided that 46 groups should be created. Hence, a perfectly balanced group with respect to group size should consist of 275/46=5.978 students. Because the number of students in a group cannot be fractional, we say that the school achieves perfect balancing with respect to group size if all groups have either five or six students. Similarly, the class consists of 84 females and 191 males. Thus, perfect balancing by gender would imply that each group has $84/46 = 1.826$ females, which we would translate into a balancing requirement that each group contains either one or two females. By applying similar logic to all the other balancing criteria, we obtain upper and lower bounds (balancing constraints) on the number of students meeting certain characteristics to include in each group. If it is impossible to meet all of the balancing constraints simultaneously (if the number of criteria gets too large, for example), we can widen the upper or lower bound constraints beyond the ideal levels.

To follow this procedure, we must first determine whether individual students meet particular criteria. For some criteria, such as gender, this determination is trivial. For other criteria, for example, citizenship, we must first translate the criteria into binary properties (Canadian or non-Canadian, for example) before we develop balancing constraints, and as a result, one criterion may lead to several properties and hence to several constraints (Table 2.1).

To address the issue of splitting work within groups, the MBA administrators decided to implement a multiple groups policy. Under this policy, they assigned each student to several study groups, with different groups employed by different courses. The intention was to discourage from trading assignments, with some members of the group doing an assignment for one class in return for other doing an assignment for another class. Another advantage of this approach is that we can adopt the set of balancing criteria to a particular class; for example, a student's quantitative skills may be important balancing criterion for a quantitative methods course but not for an organizational behavior course. For this strategy to be effective, we must ensure that the groups are nonoverlapping, that is, ideally, each student should have a completely different set of partners in each of his or her study groups. In practice, the perfect nonoverlapping assignment is very difficult and often impossible to achieve. Nevertheless, we seek to minimize the overlapping.

To summarize, our goal is to create multiple sets of groups so that groups within each set are balanced and overlapping between sets is minimal.

2.6 Rotman's Previous Group Creation Process

Until about 1990, Rotman had no centralized system for creating groups; it permitted students to pick their own partners, with no coordination of the resulting groups among courses. A group of close friends could form a study group and use it for all courses in the first year of the MBA studies. As group work became formally enshrined in the MBA curriculum in 1990s, the MBA office administrators took over the process of creating groups.

The process was manual, with office employees using their own judgment in assigning students to groups. Over time, as the balanced, multiple, nonoverlapping group policy took shape, and as MBA enrollments increased, this process became increasingly onerous to execute. By 2002, two employees were spending about a week creating the study groups. Even so, they had difficulty ensuring that the groups were well balanced with respect to several criteria; they could realistically consider only a few criteria. Creating balanced and non-overlapping groups was nearly impossible: typically, the employees would create first set of groups with some balancing in place and then they would create the other sets (groups to be used in other courses) to minimize the degree of overlapping with the first set but with no attention to the balancing criteria.

By 2003, when Rotman admitted its largest and most international first-year class ever, there were many complaints from both students and instructors about group composition and internal dynamics within groups. Students tended to blame the MBA office employees for putting them in inappropriate study groups, and dealing with these complaints took another toll on the limited resources of the MBA office.

2.7 The Management Science Approach

Underlying our approach is the recognition of the connection between the group creation problem and the classical management science assignment problem with side constraints.

Indeed, to create a single set of study groups, we need to assign students to groups, with each student assigned to only one group, and with each group satisfying upper and lower bounds on the group size - the standard assignment problem framework. In addition, each group must satisfy group-balancing constraints. Formally we define them as follows. The decision maker specifies balancing criteria $j = 1, 2, ...C$, and the property matrix with elements a_{ij} such that $a_{ij} = 1$ if student i possesses property j, and $a_{ij} = 0$ otherwise. In addition, the decision maker sets the minimal and maximal group composition values or bounds (min_j, max_j) so that a group is considered balanced with respect to criterion j if the number of students within the group that possess the corresponding property falls within the range specified by these bounds. One would normally start with the values of (min_j, max_j) corresponding to perfect balancing with respect to all properties $j = 1, 2, ...C$, and then relax some of the bounds if no feasible perfectly balanced assignment can be found.

Formally we must find a feasible solution to the constrained integer problem below. Let $y_{ig} = 1$ if student i is assigned to group g for $i = 1, 2, ...N$ and $g = 1, 2, ...G$.

$$\begin{aligned} min_j &\leq \sum_{i=1}^{N} y_{ig} a_{ij} && (2.1) \\ max_j &\geq \sum_{i=1}^{N} y_{ig} a_{ij} && (2.2) \\ \sum_{g=1}^{G} y_{ig} &= 1 && (2.3) \\ y_{ig} &\geq 0 \text{ and integer} && (2.4) \end{aligned}$$

for $i = 1, 2, ...N$, $j = 0, 1, 2, ...C$, $g = 1, 2, ...G$.

In this problem for notational convenience we say that attribute 0 denotes "membership in the group", and $a_{i0} = 1$ for all objects $i = 1, 2, ...N$. Then each perfectly balanced group must consist of either $min_0 = \lfloor \frac{N}{G} \rfloor$ or $max_0 = \lceil \frac{N}{G} \rceil$ objects.

The C balancing constraints are not of the conservation of flow type. Thus, the resulting set of linear constraints is no longer of pure assignment problem form. In particular, when we add balancing constraints, the total unimodularity property that ensures all-integer solutions in a standard assignment problem is no longer satisfied, and we must solve the problem as an integer program.

In principle, the problem of finding a single set of balanced groups is a feasibility problem, rather than an optimization problem, because any feasible solution to (2.1) - (2.4) defines a set of well-balanced groups. It is theoretically possible that no feasible solution satisfying the entire set of group-balancing constraints exists. We can take two approaches to overcome the possible lack of feasibility. First, we can treat the balancing constraints as soft constraints and use the goal programming framework with the deviation variables representing the degree of violation of each balancing criterion, and an objective function that minimizes the sum of the violations or some similar objective; for example we can give higher weights to violations of criteria judged to be more important. Second, we can simply alert the decision maker to the problem infeasibility and suggest relaxing some of the group-composition bounds. In the Study Group Adviser software we implemented at Rotman we used the second approach.

Our computational tests of the comprehensive list of properties (Table 2.1) applied to the data for the classes of 2003 and 2004 showed that no infeasibilities arise. We could always create perfectly balanced groups with respect to all criteria of interest to the MBA administrators; in fact many group assignments satisfied all of the balancing constraints. The absence of any infeasible solutions continued after the administration started using the system to create many sets of groups for MBA students in the classes of 2005 and 2006 (even though they expanded the set of properties). This could be partially due to the great care the Rotman admissions committee took to ensure sufficient diversity among the incoming students. In addition, in some cases the existence of perfectly balanced groups can be guaranteed. We study such cases in Section 3.1. Bhadurya et al. (2000) obtained similar results for a related problem.

For this data we could find a feasible solution to our formulation above without much difficulty using standard integer programming solvers on a reasonably powerful PC (we used AMPL CPLEX on a Dell PC with 512 MB of RAM). The solver found a solution within a few minutes for instances with 160 to 280 students and 30 to 50 groups.

Developing multiple sets of nonoverlapping balanced groups is much more difficult. In effect, we must simultaneously create Q sets of groups, with the groups in each set satisfying all of our constraints, while minimizing the degree of overlap between groups in different sets. An overlap occurs when we assign two students to the same group in different sets. For example, suppose that we must assign four students, numbered $1, 2, 3, 4$, to two groups, and we must create two sets of nonoverlapping groups. Then the assignment $[(1,2),(3,4)],[(1,3),(2,4)]$ represents nonoverlapping groups, while the assignment $[(1,2),(3,4)],[(1,2),(3,4)]$ contains four overlaps. In general, it may be impossible to avoid overlaps entirely; rather the goal is to minimize the incidence of overlaps. The direct approach is to formulate a model that assigns each student to multiple groups (one in each set), while maintaining the balancing constraints for each group in each set, with the objective function measuring the degree of overlapping between different sets of groups. We present such model below.

Let $y_{igs} = 1$ if student $i = 1,2,...N$ is assigned to group $g = 1,2,...G$ in set $s = 1,2,...Q$, and $y_{igs} = 0$ otherwise. Let $O_{ijps} = 1$ if students i and j overlap (i.e. are assigned to the same group) in sets p and s. The model follows:

$$Q - sets \qquad \min \sum_{p=1}^{Q} \sum_{s=p+1}^{Q} \sum_{i=1}^{N} \sum_{j=i+1}^{N} O_{ijps}$$

$$s.t. \qquad O_{ijps} \geq y_{igs} + y_{jgs} + y_{ikp} + y_{jkp} - 3 \tag{2.5}$$

$$min_j \leq \sum_{i=1}^{N} y_{igs} a_{ij} \tag{2.6}$$

$$max_j \geq \sum_{i=1}^{N} y_{igs} a_{ij} \tag{2.7}$$

$$\sum_{g=1}^{G} y_{igs} = 1 \tag{2.8}$$

$$y_{igs} \text{ binary, } O_{ijps} \geq 0$$

for all $i, j = 1,2,...N$, $g, k = 1,2,...G$ and $p, s = 1,2,...Q$.

Unfortunately the dimensionality of the resulting integer program is excessive. For example, if there are 250 students and two sets of 50 groups, the resulting model has 275,000 integer variables and over 600 million constraints. Problems of this size cannot be handled with the off-the-shelf software and the type of computing equipment available in the Rotman MBA office.

We developed a simple heuristic-based approach instead, based on the classical column-generation approach (Bazaara et al. 1977), recognizing the fact that the problem of creating several sets of balanced nonoverlapping groups has an easy sub-problem, the creation of a single set of balanced groups. We can treat each set of groups as one column, with the new columns added to the solution if they improve the objective function. However, instead of implementing the formal column-generation scheme (where we would select the new column to insert based on the computation of reduced costs or a similar rule) we initially generate a reasonably large number of columns and then obtain the best subset of columns by explicitly enumerating all possible subsets. We base this approach on the fact that the number (Q) of sets of groups required is typically small (usually $Q = 2$ or 3), and thus as long as the initial population of candidate columns is not too large (we found 20 to 30 to be adequate), we can evaluate all possible subsets fairly quickly. However, we must ensure that the initial population of columns is likely to contain some good candidates (with a low degree of overlap). To this end, we employ an objective function when generating feasible solutions to the single set model. Each objective function coefficient determines the suitability of assigning student i to group g. While, in principle, we could use an arbitrary objective function (since the single-set problem requires only the solution be feasible), we found that running the single-set model with two different objectives where the vectors of coefficients are uncorrelated tends to produce sets of groups with little overlap (uncorrelated coefficients tend to ensure that students end up in different groups in the two sets). Thus, to generate the initial population of columns for our heuristic, we run the single-set assignment model repeatedly with the objective function coefficients generated randomly for each run. This approach proved to be very effective, leading to low levels of overlapping, while maintaining the desired (usually perfect) group balancing. We used this heuristic as the analytical engine for our Study Group Adviser software package.

Our approach differs from the previous approaches to group balancing. A common methodology is to enforce balance by using weights, where each weight represents the relative importance of the group being balanced with respect to the corresponding property. In an approach based on mathematical programming (for example, Desrosiers et al. 2005), one defines a balancing goal for each group with respect to each property (for example, an ideal group should contain 1.85 females) and then forms the objective function consisting of the weighted sum of deviations from the corresponding goals. In a heuristic-based approach, we may use weights to assign a single score to each student (Beheshtian-Ardekani and Mahmood, (1986), where the score is the weighted sum of students' criteria values), or to pairs of students (Weitz and Jelassi (1992), where the score represents the weighted difference of criteria values). Baker and Benn (2001), Weitz and Lakshminarayanan (1998) and Wright (2005) use similar weight-based approaches; Baker and Powell (2002) use weighted objective in seven out of the nine formulations they examine.

In our view, this approach suffers from several shortcomings. First, deviation-based objective functions typically result in very large, often nonlinear, models, that can be only solved with specialized algorithms or heuristics (Desrosiers et al. 2005). In contrast, our feasibility model is smaller and is quite easy to solve using standard off-the-shelf software. Second, using weights forces the decision maker (1) to set the values of the weights and (2) to understand the effects of these values on the resulting group assignment. The core of these problems is that the values of the weights have no physical meaning to the decision maker. As a result, he or she may have difficulty establishing the right values for the weights and understanding how they influence the quality of a particular group assignment (for example, if the assignment causes an unacceptable deviation with respect to a particular criterion, would doubling the corresponding weight solve the problem?) (Weitz and Jelassi 1992). A related problem is interpreting the value of the objective function in the approaches based on mathematical programming. For example, Desrosiers et al. (2005) describe an instance in which the objective function value is 255. Should the decision-maker interpret this to mean that the corresponding group assignment is sufficiently well balanced?

In contrast, our feasibility-based approach is very intuitive, giving the decision maker direct control of the solution quality (through balancing constraints). If the upper and lower bounds for the balancing constraints reflect acceptable levels of deviation for the criteria, then any feasible solution is acceptably balanced. In particular, when a feasible solution is found for the problem with the ideal bounds, which happened in all cases for our application, we know that the current assignment is, in fact, perfectly balanced and cannot be further improved. To the contrary, the weights-based model may fail to find a perfectly balanced group assignment even when one exists, because of the nonintegrality of the goals. Indeed, if the goal is to have 1.85 females per group, the deviation-based approach may bypass a perfectly balanced solution (with one or two females in each group) and keep searching for a better solution. This behavior could explain why using our approach it takes minutes to create the groups for a large class of 275 students on a standard PC, while the studies listed above either rely (predominantly) on heuristics or on using powerful hardware for smaller problems.

The feasibility-based approach seems best suited for situations in which one can represent all balancing criteria through binary properties (which a student either has or does not have). In some situations this is not the case. Cutshall et al. (2005) seek to balance groups with respect to the average historical grade per group, which is clearly not a 0/1 measure. In such a case, the weight-based approach seems suitable (Cutshall et al. (2005) proposed a combination of the weighted objective and a feasibility-based model similar to ours).

The multiple group aspect of our model is novel. Desrosiers et al. (2005) seem to be the only other attempt to create multiple sets of groups. They use a variable neighborhood search heuristic.

2.8 The Study Group Adviser Software

Our ultimate goal was to implement our ideas and algorithms in a software tool for creating groups that would be versatile and simple to use. Versatility is necessary because many of the parameters of the group-creation process fluctuate constantly (including the number of students, groups, and lists of groups, and the criteria to use for balancing). Thus, a software package that could create groups based only on a hard-coded list of criteria would be of little use. The software had to be simple to use, because the MBA office employees who create groups are generally not familiar with management science concepts, and may not be sophisticated computer users.

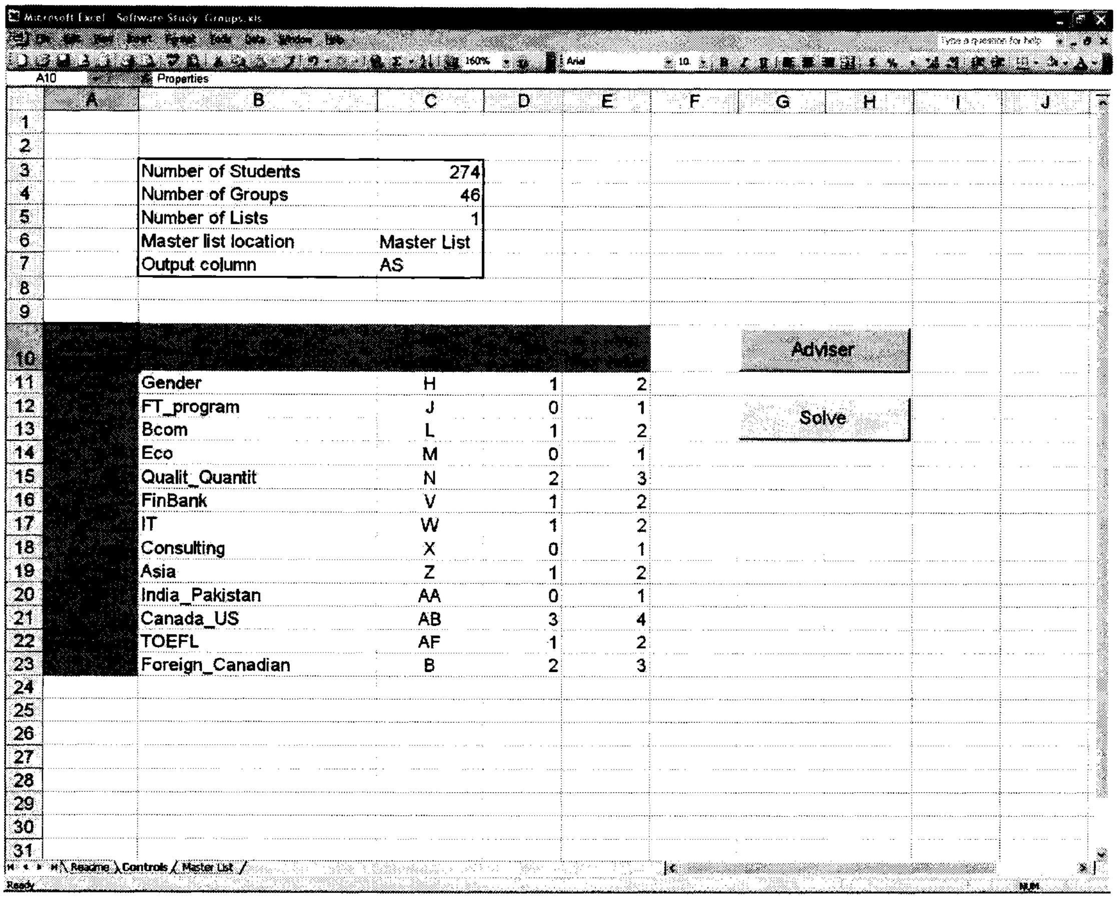
We implemented the Study Group Adviser program as a Microsoft Excel macro (written in the Visual Basic for Applications language) since the potential users are familiar with the Microsoft Excel environment and use it for a variety of MBA program scheduling and management tasks. Users supply two types of inputs (Figure 2.1):

1. Basic parameters, such as number of students, number of groups, multiple lists, and data locations and output range;
2. Names, column identifiers, and upper and lower bounds for the balancing properties (ideal values for the bounds are computed automatically).

In addition, the program requires an input data table consisting of an Excel work sheet with a row for each student and a column for each balancing property; the 0/1 entries in each cell indicate whether a student satisfies the corresponding criterion.

Once the user supplies all the parameter values, he or she presses the Solve button, which invokes the macro that implements our heuristic (we use AMPL with the CPLEX solver to solve the integer programs).

We designed Study Group Adviser to detect various errors and suggest possible solutions



Number of Students	274
Number of Groups	46
Number of Lists	1
Master list location	Master List
Output column	AS

Gender	H	1	2
FT_program	J	0	1
Bcom	L	1	2
Eco	M	0	1
Qualit_Quantit	N	2	3
FinBank	V	1	2
IT	W	1	2
Consulting	X	0	1
Asia	Z	1	2
India_Pakistan	AA	0	1
Canada_US	AB	3	4
TOEFL	AF	1	2
Foreign_Canadian	B	2	3

Figure 2.1: Rotman Study Group Adviser software interface. The upper block contains basic parameters, and the lower block lists names, locations, and bounds for the balancing constraints. Pressing Adviser button calculates the bounds corresponding to the perfect balance. Pressing Solve button generates the groups.

to the user. For example, if the assignment problem is infeasible, it advises the user to relax some of the bounds for the balancing constraints. The current version of the software does not specify which constraints to relax (in principle, one could iteratively relax constraints until a feasible solution is found). The MBA administrators frequently know what kinds of groups they want, which usually means achieving perfect balance with respect to only a few criteria and allowing some flexibility with respect to the rest. Thus, we left it to the users to determine which constraint(s) to relax in the event of infeasibility. So far feasible solutions have been found for all instances solved to date without relaxing balancing constraints.

The software was well received by the MBA office personnel. Initially, the office managers worried that the software might be hard to use and designated one employee to be trained as a specialist. However, after a demonstration and training session, nearly all of the office employees creating student groups were quite comfortable using the package. They first used it to create student groups for the 2005 class in August 2004, creating two sets of groups. They subsequently used it to reassign students to the two new sets of groups for the second semester of the 2004-2005 academic year.

The software proved to be sufficiently adaptable to deal with some group emergencies. In one case, the members of one group proved to be incompatible, and the administrators decided to split up this group. By creating a new property called Incompatible, which was satisfied only by the members of the original group (that is, they received a 1 in the corresponding data column) and including this property as one of the balancing criteria, the employees created a new set of groups in which no members of the problem group overlapped.

2.9 Comparing Manual and Computer Generated Groups for the Class of 2004

To assess software performance, we applied the Study Group Adviser package to the data for the class of 2004 and compared the quality of the resulting assignments with the original manually created groups. For this class, the MBA office staff originally created two sets of groups: the first set served as the main study groups and the office staff made an attempt to satisfy the balancing criteria. Then they generated the second set (called the alternate groups) to be nonoverlapping with the first set but without taking the balancing criteria into account.

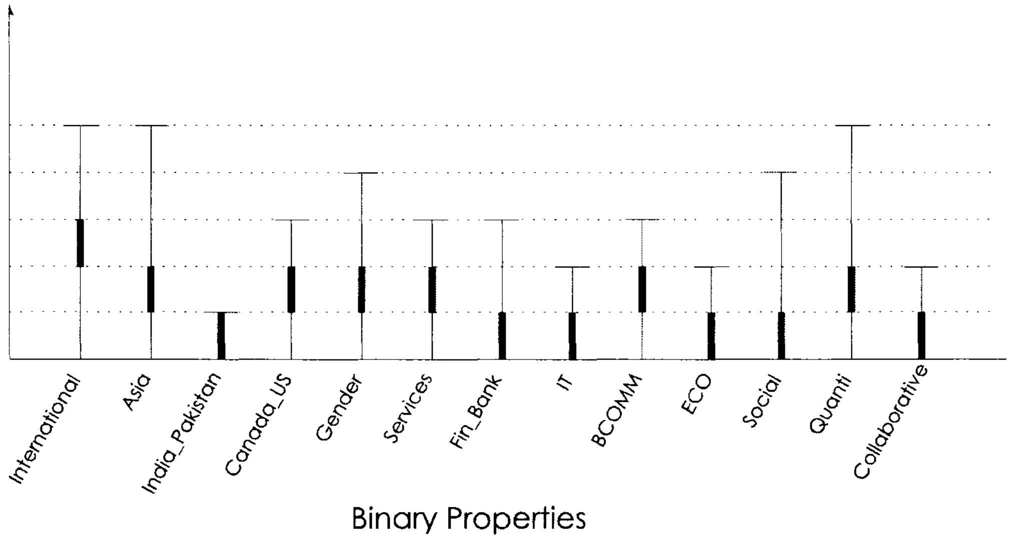


Figure 2.2: Quality of manually (hairline bars) and automatically created groups (bold bars) for the class of 2004. Descriptions of binary properties and corresponding criteria are given in Table 2.1.

We used our software to generate two sets of groups, both required to satisfy all of the balancing criteria, while we minimized the degree of overlap.

We first compared our groups to the main study groups with respect to the balancing criteria (Figure 2.2). We used 13 balancing criteria, and the software-generated groups outperformed the manually created ones with respect to all but one of the criteria (for which the performance was equal). The results were even stronger when compared against the alternate groups, which were not subject to balancing.

With respect to the nonoverlapping objective, the software-generated groups scored slightly worse than those produced manually; they had five overlaps, while the manual groups had none.

The MBA office managers much preferred the software-generated groups to the manually created ones, because they view good group balance as much more important than the presence of a few overlaps. As one of the managers put it, "Poor balancing affects a lot of students, while overlaps affect only a few."

The time required to create the groups was another important consideration. It took the MBA office employees eight person/days to create the group assignments for the class of 2004; the Study Group Adviser software handled this task in just 20 minutes.

The strong performance of our software on the class of 2004 test was an important factor in the decision to move forward with its implementation.

2.10 Well-Balanced Groups in Action: The Class of 2005

The switch to software-based groups was quite smooth, and the MBA office personnel accepted the software quickly. The most serious issue that arose during the implementation concerned defining the balancing criteria. For a characteristic to be a good balancing criterion it must be common in the student population, otherwise many groups will contain no students with this characteristic. To create a common characteristic, we had to aggregate several basic characteristics into a new aggregate property. For example, one of the balancing criteria of interest was membership in a collaborative program; we wanted to balance the student groups with respect to the number of students from such programs. The Rotman MBA program has ongoing collaborative programs with six other departments in the university, but the number of students in each of these programs is quite small. Thus, if we set up separate balancing criteria for the six collaborative programs, most of the groups would have no representative from any one program. On the other hand, one group could end up with representatives from several different collaborative programs since the balancing requirement would apply to each criterion separately - an undesirable outcome. To resolve this issue, one should define a new aggregate property, membership in any of the collaborative programs, and balance with respect to this property instead. Similarly if we define citizenship criteria for individual foreign countries, the resulting number of students from most countries would be too small to achieve true balancing, while could create a group with a large number of foreign students, all from different countries. Once we aggregate the citizenship property sufficiently (Table 2.1), this problem disappears. It did not take long for the MBA office employees to learn to resolve these issues on their own.

Preparing the input data table has not been a problem because the property membership data are recorded during the admission process and are available electronically. User can either import this data into the Excel work sheet directly or code it as columns of 0's and 1's using the standard Excel tools and functions. We have been pleasantly surprised by how painless the software adoption process has been; over a year we received no service calls. The MBA office personnel can quickly create groups and instructors now routinely ask them to set up special groups for particular projects, and they create new group assignments within a few hours.

A somewhat unexpected side benefit of the software was the reaction of the MBA students to the new system: a drastic drop in the number of complaints about being placed in the wrong group. Apparently, the fact that a sophisticated computer tool created the assignments to study groups made them more acceptable to the students. This effect was important to the MBA office and to the Rotman School in general, since student complaints about group composition had been a sore point in the past.

More important issue is the overall impact of multiple well-balanced groups on the educational process at Rotman. Certainly, students see group-based work now as more fair (since no group is composed solely of engineers or foreign-language speakers). Intragroup dynamics seem more sensitive since the school implemented group balancing. Students are forced to collaborate with partners they did not choose. When students were allowed to form their own groups, the groups were more homogeneous, and their problems had to do with intergroup differences. The current approach has eliminated most of the intergroup

Table 2.1: Balancing Criteria, Corresponding Binary Properties and Group Composition Values for the Class of 2005.

Balancing Criteria	**Property name and description**	**Min and Max composition values**
Size of the group	Equals 1 for every student	5 − 6
Gender	Equals 1 for females	1 − 2
Collaborative program	Equals 1 if a student is from a collaborative program	0 − 1
Academic background	BCOMM: equals 1 if a student has a degree in business/commerce	1 − 2
	ECO: equals 1 if a student has a degree in economics	0 − 1
	QUANTI: equals 1 if a student has a degree in a quantitative discipline (e.g. engineering, mathematics)	2 − 3
	SOCIAL: equals 1 if a student has a degree in a social discipline (e.g. psychology, law)	2 − 3 0 − 1
Industrial background	FINBANK: equals 1 if a student has work experience in finance or banking	1 − 2
	IT: equals 1 if a student has work experience in information technology or telecommunications	1 − 2
	CONSULTING: equals 1 if a student has work experience in consulting	0 − 1
	SERVICES: equals 1 if a student has work experience in services	1 − 2
Citizenship	Canada/US	3 − 4
	India/Pakistan	0 − 1
	Asia	1 − 2
International	Equals 1 if a student too TOEFL test (test of Enlish as a foreign language	1 − 2
Cultural	Equals 1 if a student has Canadian or US citizenship and is not a new immigrant.	2 − 3

problems, but the enforced within-group heterogeneity brings about its own problems.

Assigning students to multiple nonoverlapping groups seems to have reduced the amount of work splitting among groups members, at least with respect to group-based assignments from different classes. Group members splitting work on major assignments for a particular class is still quite common; group creation strategy cannot accomplish much in this regard. In addition, the school established only two sets of groups for 2004-2005, and students typically take four to six classes each semester, which means that the same groups were typically used in several classes each term, increasing students' opportunities for work splitting. Perhaps the school should consider increasing the number of sets of groups to match the number of classes with major group assignments.

Do the heterogeneous groups produce better quality work than the homogenous groups of the past, as the organizational behavior theory would suggest? This question is hard to answer and would require a separate study. In particular, Rotman has tightened its admission standards progressively over the last several years, improving the quality of the incoming students. Thus, while most instructors would agree that the average quality of student work (both individual and group based) has been improving, which, if any, part of this improvement can be traced to better group composition is not clear. We note that researchers seeking to evaluate the impact of group balancing on learning outcomes in business schools found it has little or no effect (Donohue and Fox 1993, Muller 1989).

Overall, however, the students, faculty, and the MBA office staff seem to be happy with the new system for creating groups. The school has no plans to revert to self-selected or to manually created groups.

2.11 Conclusions

In this chapter we studied how organizations can create efficient work groups or teams that would be maximally successful in performing diverse and complex tasks. Coming up with the groupings is usually hard, since one must simultaneously consider a variety of different factors, and as the number of groups and factors increase, manual creation of balanced groups becomes inferior. Non-surprisingly this problem attracted considerable attention in the management science and decision analysis literature. By far the most frequent approach is to set up balancing goals and search for the partitions that minimize some measure of deviations from these goals, such as sum of squared deviations. The major problem with implementing such approach in practice is that such objective has no physical meaning to decision maker and causes a lot of confusion; this problem has been reported by a number of researchers. The approach we take is different. In effect it brings the thinking of the decision maker into the model.

Our approach is based on constraining balanced groups with the minimal and maximal amount of each attribute per group - precisely what decision maker is trying to achieve in her mind. We formulate group balancing problem as a constraint integer program and search for the feasible solution with arbitrary objective. This approach appears to be very successful. It has been implemented and is used at Rotman School of Management for several years. We report major improvements in various issues involving group work. In addition we report a heuristic that creates multiple lists of minimally overlapping groups. Such feature is often desirable if individual group members must work in several groups.

Even tough the constrained approach seems to work well, theoretically it is quite surprising: one would typically expect that no feasible solution would exist or it would be hard to find, especially when the program becomes very constrained. To better understand the limitations of our approach as well as find the reason for why it works well in practice in the next Chapter we study the conditions that guarantee the existence of balanced groups, and quantify the probability to find a partition.

Bibliography

[1] Baker M.B., C. Benn 2001. Assigning pupils to tutor groups in a comprehensive school. Journal of the Operational Research Society 52(6) 623-629.

[2] Baker, K.R., S.G. Powell 2002. Methods for assigning students to groups: a study of alternative objective functions. Journal of the Operational Research Society 53(4) 397-404.

[3] Bazaraa, M.S., J.J. Jarvis, J.J., H.D. Sherali 1977. Linear Programming and Network Flows. 2nd ed. John Wiley and Sons, New York.

[4] Beheshtian-Ardekani, M., M., A. Mahmood 1986. Development and validation of a tool for assigning students to groups for class projects. Decision Sciences 17(1) 92-113.

[5] Bhadurya, J., E.J. Mighty, H. Damar 2000. Maximizing workforce diversity in project teams: a network flow approach. Omega 28(2) 143-153.

[6] Cutshall, R., S. Gavirneni, K. Shultz, 2005. Kelley School of Business uses integer programming to form equitable, cohesive case teams in integrated core (I-Core). Working paper. Kelly School of Business, Indiana University.

[7] Desrosiers, J., N. Mladenovic, D. Villeneuve 2005. Design of balanced MBA student teams. Journal of the Operational Research Society 56(1) 60-66.

[8] Donohue, J., J. Fox 1993. An investigation into the people-sequencing heuristic method of group formation. Decision Sciences 24(2) 493-508.

[9] Hackman, J.R., ed. 1990. Groups That Work (and Those That Don't). Jossey-Bass Publishers, San Francisco.

[10] Kirchmeyer C., J. McLellan 1990. Managing ethnic diversity: utilizing the creative potential of a diverse workforce to meet challenges of the future. Proceedings of the Annual ASAC Conference, Organizational Behaviour Division. 11 pt. 5 120-129.

[11] McShane, S.L. 1992. Canadian Organizational Behavior. Richard D. Irwin, Homewood.

[12] Muller, T. E. 1989. Assigning students to groups for class projects: An explanatory test of two methods. Decision Sciences 20(3) 623-634.

[13] Weitz, R. R., M.T. Jelassi 1992. Assigning students to groups: A multi-criteria decision making support system approach. Decision Sciences 23(3) 746-757.

[14] Weitz, R.R., S. Lakshminarayanan 1998. An empirical comparison of heuristic methods for creating maximally diverse groups. Journal of the Operational Research Society 49(6) 635-646.

[15] Wright, M. 2005. Experiments with plateau-rich solution space. Working paper. Lancaster University Management School.

Chapter 3

Balanced Multiple-Attribute Set Partitioning Problem

3.1 Introduction

The problem of creating MBA study groups described above is an application of the following more general problem. Consider a set $\mathbf{S}$ of N objects $\{s_1, s_2, ...s_N\}$, where each object is characterized by a vector of C attributes $\{1, 2, ...C\}$ through a matrix $A : \{a_{ij}, i = 1, 2, ...N, j = 1, 2...C\}$ such that $a_{ij} = 1$ implies that object i possesses attribute j, and $a_{ij} = 0$ implies the reverse. The problem is to partition these objects into G perfectly balanced groups as described below.

Let $c_j = \sum_{i=1}^{N} a_{ij}$, $0 \leq c_j \leq N$, be the column sum for the j^{th} column of A. Observe that c_j represents the total amount of attribute j contained in set $\mathbf{S}$. Let $min_j = \lfloor \frac{c_j}{G} \rfloor$ and $max_j = \lceil \frac{c_j}{G} \rceil$ denote the minimal and maximal amount of attribute j that each group is allowed to possess to be balanced. To define the number of objects per group (size of the group), augment A by a column of ones, i.e. let $a_{i0} = 1$ for all $i = 1, 2, ...N$. Then let

$min_0 = \lfloor\frac{N}{G}\rfloor$ and $max_0 = \lceil\frac{N}{G}\rceil$ denote the number of objects that each group is allowed to possess.

Recall $y_{ig} = 1$ if object i is assigned to group g for $i = 1, 2, ...N$ and $g = 1, 2, ...G$. Perfectly balanced multiple attribute set partitioning problem (BMASP) is to partition set **S** into G groups such that

$(BMASP)$

$$\begin{aligned} min_j &\leq \sum_{i=1}^{N} y_{ig}a_{ij} \text{ for all } j = 0, 1, 2, ...C, g = 1, 2, ...G && (3.1)\\ max_j &\geq \sum_{i=1}^{N} y_{ig}a_{ij} \text{ for all } j = 0, 1, 2, ...C, g = 1, 2, ...G && (3.2)\\ \sum_{g=1}^{G} y_{ig} &= 1 \text{ for all } i = 1, 2, ...N && (3.3)\\ y_{ig} &\in \{0; 1\} \text{ for all } i = 1, 2, ...N, g = 1, 2, ...G. \end{aligned}$$

We use the term "perfectly" balanced to stress out that the values of min_j and max_j defined above are as tight as possible (recall the discussion of 1.82 females per group). In general one could seek a partition in which the group composition values min_j and max_j could not be this tight; for example, each group could contain 1 to 3 females. Since the existence of a perfectly balanced partition obviously implies the existence of the other types of balanced partitions, we only consider the former, and hereinafter refer to them as balanced (that is we omit the adjective "perfectly"), leading to the BMASP problem.

There are several approaches that could be used to search for balanced partitions. In the Rotman project we took a feasibility approach: we set up an arbitrary objective and searched for a feasible solution to the BMASP formulation. As discussed above, such approach is more appealing to the decision makers and we see this as the primary reason for its successful implementation. However, feasibility approach would be of little use if the resulting integer program (3.1)-(3.3) would often be infeasible (i.e. balanced partitions

a_{ij}	Gender	Siblings	Vegas
Ross	1	1	1
Chandler	1	0	0
Monica	0	1	0
Rachel	0	0	1

Table 3.1: An example where balanced groups do not exist

would fail to exist in many instances) or if feasible solutions would be hard to find. As noted, this was not the case in our implementation: many feasible solutions existed and were easy to find for all instances solved to date. That is our feasibility based approach appears to be quite practical.

Theoretically, however, the existence of such partitions is quite surprising. Intuitively, as we add more attributes the integer problem becomes increasingly more constrained, increasing the chances that no feasible solution exists. So why is that feasible solutions appear to be so frequent in practice? In the remainder of the chapter we present some results explaining why and when feasible groups exist.

To further motivate our work we present a simple example where balanced groups do not exist. Consider four friends, whom we want to split into two groups: Ross, Rachel, Monica and Chandler; and consider three attributes: gender, siblings, and "married in Las Vegas and then divorced". The attribute matrix $\{a_{ij}\}$ is given in Table 3.1.

In this example $N = 2$, $G = 2$, $C = 3$, $c_j = 2$ and $min_j = max_j = 1$ for $j = 1, 2, 3$. Verbally, each balanced group should consist of two friends, one of which should possess one unit of each attribute.

It is easy to see that balanced groups do not exist in this example. Indeed, the first attribute implies that Ross must be in a group with either Monica or Rachel. The second attribute implies that Ross cannot be in a group with Monica because they are brother and sister, while the third attribute implies that neither he can be in a group with Rachel, since they married in Vegas and then got divorced. Therefore not two, but even one group cannot be created. In fact, we shall see that the subproblem of creating one group plays an important role in our analysis.

To characterize the conditions when balanced partitions exist we develop the following equivalence relations.

3.2 Equivalence Relations between BMASP Instances

We present two kinds of equivalence relations. The first one allows us to assume without loss of generality that the number of objects per group, N/G, is an integer. The second allows to draw equivalences among the problems with different number of objects and groups, but with the attributes of the same types (the concept of type is described below).

3.2.1 Number of Objects Per Group

Consider an arbitrary BMASP instance and the corresponding matrix A such that N/G is not integral. In such instance each balanced group should contain either $min_0 = \lfloor \frac{N}{G} \rfloor$ or $max_0 = \lceil \frac{N}{G} \rceil$ objects. We refer to this instance as the *original* instance.

For such original instance with A and G construct an *induced* instance $BMASP'$ with $A' = N' \times C'$ and the same G groups in the following way. Let $N' = \lceil \frac{N}{G} \rceil G$ and let $C' = C + 1$. Create one dummy attribute C' and $N' - N$ dummy objects such that real (not dummy) objects possess the same real attributes in both instances, dummy objects possess one unit of dummy attribute and no real attributes, while real objects do not possess dummy attribute. That is for $i = N + 1, N + 2, ...N'$, $a'_{i,C'} = 1$ and $a'_{ij} = 0$, and for $i = 1, 2, ...N$, $a'_{i,C'} = 0$ and $a'_{ij} = a_{ij}$ for $j = 1, 2, ...C$. Observe $c'_j = c_j$ for $j = 1, 2, ...C$, N'/G is an integer, $c_{C'} = N' - N < G$ and hence $min'_{C'} = 0$ and $max_{C'} = 1$. Finally observe that $min_0^{'} = max_0^{'} = N^{'}/G = max_0$.

Theorem 3.1 *A balanced partition in the original instance of $BMASP$ exists if and only if one exists in the induced instance $BMASP^{'}$.*

Proof. Let p^{min} and p^{max} be the solutions to the diophantine equation $p^{min}min_0 + p^{max}max_0 = N$. Observe that balancing of the original instance with respect to size implies that there are p^{min} groups of size min_0 and p^{max} groups of size max_0. By the construction of $BMASP^{'}$, $p^{min} = N^{'} - N = c_{C^{'}}$, i.e. there are p^{min} dummy objects.

In addition by the construction of $BMASP^{'}$, $c_j^{'} = c_j$ for $j = 1, 2, ...C$, and hence $min_j^{'} = min_j$ and $max_j^{'} = max_j$ for all real attributes. Therefore, since dummy objects do not possess real attributes, adding (removing) dummy objects to (from) a balanced group does not affect its balance with respect to real attributes.

Therefore, if there exists a feasible partition in $BMASP$, then adding one of p^{min} dummy object to each of the p^{min} groups with min_0 objects results in a balanced partition in $BMASP^{'}$ since each group contains $N^{'}/G$ objects, min_j or max_j of which possess real attributes $j = 1, 2, ...C$ and each group possess $min_{C^{'}}^{'} = 0$ or $max_{C^{'}}^{'} = 1$ dummy objects, which, by definition, possess one unit of dummy attribute each.

Likewise, if there exists a feasible partition in $BMASP^{'}$, then since $max_{C^{'}}^{'} = 1$, each group of $N^{'}/G = max_0$ objects contains at most one dummy object. Since there are p^{min} dummy objects and for all real attributes $min_j^{'} = min_j$ and $max_j^{'} = max_j$, removing the dummy objects from the corresponding groups results in p^{min} groups of size $max_0 - 1 = min_0$ and p^{max} groups of size max_0 which are all balanced with respect to real attributes. By definition such partition is a balanced partition in $BMASP$. ■

Following Theorem 3.1 in the remainder of the chapter we assume N/G is an integer and for notational convenience we let $k = N/G$. That is without loss of generality we consider only the instances where each balanced group should consist of exactly k objects. In the BMASP formulation, we therefore can rewrite constraints (3.1) and (3.2) for $j = 0$ as $\sum_{i=1}^{N} y_{ig} = k$, $g = 1, 2, ...G$.

We further note that balancing with respect to an attribute is equivalent to balancing with respect to its negation. For example, in Table 3.1 for the first attribute ("Gender") we could as well have used 1 to denote females; likewise we could have used 1 to denote "not siblings" in the second attribute, etc. Therefore without loss of generality we make the following assumption:

Assumption 1 $kG \geq 2\max_{j=1,...C} c_j$. *Equivalently,* $k \geq 2max_j$ *for all* $j = 1, ...C$.

3.2.2 Equivalence Classes

Consider two instances $BMASP_1$ and $BMASP_2$ and the corresponding matrices $A_1 = (N_1 \times C_1)$ and $A_2 = (N_2 \times C_2)$ that have to be partitioned into G_1 and G_2 groups of size k_1 and k_2 respectively. We use subscripts 1 and 2 to distinguish between various parameters of these instances.

Let the pair (min_j, max_j) denote the type of the attribute. We say that attributes j and j' are of the same type if $min_j = min_{j'}$ and $max_j = max_{j'}$. Let $\mathbf{T}_i$ be the set of all types of attributes that instance i has and let $t_{(h,\overline{h}),i}$ be the number of attributes of type $(h, \overline{h})$ that instance i has, $(h, \overline{h}) \in \mathbf{T}_i$, $i = 1, 2$. Observe $\sum_{(h,\overline{h}) \in \mathbf{T_i}} t_{(h,\overline{h}),i} = C_i$.

We say that instances $BMASP_1$ and $BMASP_2$ are the same equivalence class if:

1. $k_1 = k_2 = k$ for some k ;
2. $G_1 = G_2 = G$ for some G;
3. $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{T}$ for some set $\mathbf{T}$;

4. $t_{(h,\bar{h}),1} = t_{(h,\bar{h}),2} = t_{(h,\bar{h})}$ for some $t_{(h,\bar{h})}$ for all $(h, \bar{h}) \in \mathbf{T}$

We denote such equivalence class by $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$, where $\{t_{(h,\bar{h})}\}$ is a vector of $t_{(h,\bar{h})}$ values.

In words, two instances of BMASP are equivalent if they have the same number of objects, the same number of attributes of the same types and are required to be partitioned into the same number of groups of the same size.

The following observation is intuitive:

Observation 1 *Consider two instances with attribute matrices A and A' in equivalence classes $\boldsymbol{E}(G, k, \boldsymbol{T}, \{t_{(h,\bar{h})}\})$ and $\boldsymbol{E}$'$(G, k, \boldsymbol{T}$', $\{t'_{(h,\bar{h})}\})$ respectively. If $A' \subseteq A$ and instance A' cannot be partitioned into balanced groups, then a balanced partition also cannot exist for A.*

We refer to the attributes of types (i, i) for some i as of the fixed type (i.e. if attribute j is of the fixed type then $min_j = max_j$). Otherwise, the type is variable.

Observe that if an equivalence class contains only fixed attributes, then with respect to the attribute of a given type, all instances in such a class have the same c_j values, because for fixed types $c_j = Gmin_j = Gmax_j$. To the contrary, if an equivalence class contains variable attributes, then even with respect to the attributes of the same type, c_j values can be different, because for variable types $c_j \in \{Gmin_j + 1, ..., Gmax_j - 1\}$. Therefore to distinguish between the instances in the same equivalence class, $\mathbf{E}$, but with different c_j values, we say that an instance is in the sub-class $\bar{c}_{\mathbf{E}}$, where $\bar{c}_{\mathbf{E}} = \{c_1, c_2, ...c_C\}$ is a vector of some given c_j values of instances in $\mathbf{E}$.

The remainder of the chapter is organized as follows. In Section 3.3 we present the worst-case analysis and characterize the the properties of equivalence classes for which a balanced partition may not exist. Then in Section 3.4 we present probabilistic analysis of the likelihood that a random instance in a given equivalence class can be partitioned into balanced groups.

3.3 Worst-case Analysis

In this Section we study the properties of equivalence classes which lead to the existence of the instances for which balanced partitions do not exist. While we are generally interested in the existence of partitions, a partition obviously cannot exist in the cases when even a single group does not exist. Therefore we study both a single group sub-problem and a problem of creating a partition. We suggest two complementary approaches, one based on exploring block structures in the attribute matrices, and another based on constructing and solving a cover problem.

We start with a 'positive' result.

Theorem 3.2 *Every BMASP instance with 2 attributes can be partitioned into balanced groups.*

Proof. Suppose arbitrarily that $c_1 \leq c_2$. Observe it implies $max_1 \leq max_2 \leq min_2 + 1$. Let q_j for $j = 1, 2$ be the solution to diophantine equation $q_j min_j + (G - q_j) max_j = c_j$, i.e. a balanced partition with respect to attribute j consists of q_j groups with min_j of this attribute and $G - q_j$ groups with max_j.

Observe that in an attribute matrix with $C = 2$ columns there are four unique rows: $\{1, 1\}$, $\{1, 0\}$, $\{0, 1\}$ and $\{0, 0\}$.

Take the first min_1 rows (objects) with '1' in the first column and assign them to group 1, repeat q_1 times to create q_1 groups with min_1 of attribute 1. Then assign the remaining $(G - q_1)max_1$ rows (objects) with '1' in column 1 to the remaining $(G - q_1)$ groups with max_1 of attribute 1 each.

By construction, each group contains min_1 or max_1 of attribute 1, and at most $max_1 \leq max_2$ of attribute 2. Furthermore, every row $\{1, 1\}$ is already assigned to some group, and so every unassigned row that has '0' in column 1. Therefore, assigning rows $\{0, 1\}$ in a greedy manner to each group such that it has min_2 or max_2 of attribute 2 results in groups that are balanced with respect to both attributes, and the only unassigned rows are $\{0, 0\}$.

Since each group contains at most $2max_2$ objects, and by Assumption 1 $2max_2 \leq k$, assigning rows $\{0, 0\}$ in a greedy manner to each group results in a balanced partition. ■

Interestingly, for all $C \geq 3$ there exist instances in which balanced partitions do not exist. We study such instances next.

3.3.1 Block-matrix approach

To build the complexity gradually we study three versions of BMASP problems: two restricted problems, one with identical fixed attributes and the second with non-identical fixed attributes, and a general BMASP as per Section 3.1.

The Case with Identical Fixed Attributes

In this subsection we consider equivalence classes where all attributes are of type (m, m), i.e. with $min_j = max_j = m$ for all $j = 1, ..., C$ for some $m, C \geq 1$. For a given number of groups, G, and a number of objects per group, k, this equivalence class is $\mathbf{E}(G, k, \{m\}, C)$. In light of Assumption 1 throughout this Section we assume $k \geq 2m$.

We require the following definition:

Definition 1 *A* <u>*block*</u> (C, p, q)*, where* $C, p, q \geq 1$*, is a* $(p \times C)$ *matrix* $D, d_{ij} = \{0; 1\}$ *such that* $\sum_{i=1,...p} d_{ij} = q$ *for all* $j = 1, ...C$ *and* $\sum_{j=1,...C} d_{ij} \geq 1$ *for all* $i = 1, ...p$*.*

In words, a block is a $0-1$ matrix with positive row sums and equal column sums.

If in a block (C, p, q) there exists a subset of rows that form a (smaller) block (C, p', q'), where $p' \leq p-1$ then we say that (C, p', q') is a sub-block (note that since column-sums in a block are positive we must have $q' \leq q-1$). Observe that by definition of a block, the rows of (C, p, q) that are not in (C, p', q') also form a block (sub-block). A block is indivisible if it contains no sub-blocks.

Consider an arbitrary instance in $\mathbf{E}(G, k, \{m\}, C)$. This instance is described by the matrix $A = (kG \times C)$ with the column sums $c_j = mG$ for all columns (attributes) $j = 1, 2, \ldots C$. Therefore for some $P \leq kG$, A consists of a (C, P, mG) block, augmented by $kG - P$ all-zero rows. Likewise, in a balanced partition each group of k objects contains m objects that possess attribute j and $k-m$ objects that do not possess attribute j, for all $j = 1, 2, \ldots C$. Thus for some $p \leq k$ a balanced group is a (C, p, m) block, augmented with $k-p$ all-zero rows.

Let A' be the matrix obtained from A by deleting all-zero rows. We call A' the attribute matrix block. The following observation follow immediately from the discussion above.

Observation 2 *Suppose A consists of a block A' and p all-zero rows. Then*

- *if A' can be partitioned into G balanced group sub-blocks, (C, p_g, m), $p_g \leq k, g = 1, \ldots G$, then G groups exist;*
- *if A' contains one balanced group sub-block, (C, p, m), $p \leq k$ and $k-p$ or more all-zero rows, then at least one balanced group exists;*
- *if A' does not contain a group sub-block or there are less than $k-p$ all-zero rows, then no balanced groups exist,*

For $r \geq 2$ consider a square $(r \times r)$ matrix with 0s on the main diagonal and 1s in all off-diagonal entries. Note that it forms a block $(r, r, r-1)$; we refer to such a block as $D(r, r, r-1)$. Blocks $D(r, r, r-1)$ are indivisible for any r; see Proposition 3.1 in the Appendix.

Let j_m be the largest integer such that m is divisible by all integers $i \in (1, ..., j_m)$. Next we present our main result for the case with identical fixed attributes.

Theorem 3.3 *For any $G \geq j_m{+}1$ and $C \geq j_m{+}2$ there exists an instance in $\mathbf{E}(G, k, \{m\}, C)$ for which a partition into G balanced groups does not exist.*

Proof. By Observation 1 it is sufficient to consider only the equivalence classes with $C = j_m + 2$. Consider two cases:

Case 1: G is divisible by $j_m + 1$. Let $i = \frac{G}{j_m+1}$ and $q = mi$. Note that $(j_m + 1)q = mG$ and $kG - q(j_m + 2) \geq kG - 2q(j_m + 1) \geq (k - 2m)G \geq 0$ because $k \geq 2m$ by Assumption 1. Create $(q(j_m+2) \times (j_m+2))$ matrix A' by adding (one below another) q blocks $D(j_m + 2, j_m + 2, j_m + 1)$. Augment A' with $kG - q(j_m + 2) \geq 0$ zero rows; call the resulting matrix A. By construction, A is a $(kG \times (j_m + 2))$ matrix with column sums $(j_m + 1)q = mG$. That is A is an attribute matrix of an instance in $\mathbf{E}(G, k, \{m\}, j_m + 2)$.

Next we show that for every subset of rows in A with equal column sums (a sub-block or a sub-block augmented by all-zero rows), the column sum is divisible by $j_m + 1$.

Since every non-zero row in A comes from $D(j_m + 2, j_m + 2, j_m + 1)$, let row q be of type i if $d_{qi} = 0$, $q, i = 1, ...j_m + 2$. Observe that a collection of rows, one of each type, is a block $D(j_m + 2, j_m + 2, j_m + 1)$.

From Observation 2 non-zero rows of a balanced group in $\mathbf{E}(G, k, \{m\}, j_m + 2)$ must form a block $(j_m + 2, p, m)$, $p \leq k$. Suppose such a block exists and call it D'. Let k_i be the number of rows of type i in D'; note $\sum_{i=1}^{j_m+2} k_i = p$. By construction, column i of D' has column sum $c_i = p - k_i$. Therefore, since D' is a $(j_m + 2, p, m)$ block all column sums equal to m and so $k_i = k_{i'}$ for all $i, i' = 1, \ldots j_m + 2$. Hence D' contains equal number of rows of each type, i.e. D' is an integral number of copies of $D(j_m + 2, j_m + 2, j_m + 1)$ added one below another. Thus m divides by $j_m + 1$, which contradicts the definition of j_m.

Case 2: G **is not divisible by** $j_m + 1$**.** Let $q = \lfloor \frac{mG}{j_m+1} \rfloor$; note $mG = q(j_m + 1) + g$, where $g \in \{1, \ldots j_m\}$. Create a $(q(j_m + 2) \times C)$ matrix A' by adding (one below another) q indivisible blocks $D(j_m + 2, j_m + 2, j_m + 1)$; note $C = j_m + 2$. Augment A' with g all-ones rows and $kG - q(j_m+2) - g$ all-zero rows; call the resulting matrix A. Observe $kG - q(j_m + 2) - g \geq kG - 2q(j_m + 1) - 2g = kG - 2(q(j_m + 1) + g) = G(k - 2m) \geq 0$. By construction, A is a $(kG \times C)$ matrix with column sums mG. That is A is an attribute matrix of an instance in $\mathbf{E}(G, k, \{m\}, C)$.

Recall that vector $r = \{r(1), \ldots r(j_m + 2)\}$ (a row in A) is of type i if $r(i) = 0$, $i = 1, \ldots j_m + 2$ and $r(i') = 1$ otherwise. Observe that A contains q rows of each of the types $1, \ldots j_m + 2$ and g all-ones rows.

By the same argument in the Case 1, every block that contains only rows of types $1, \ldots j_m + 2$ has column sums divisible by $j_m + 1$. Therefore, each balanced group must contain at least one all-ones row. Since by construction there are $g \leq j_m < j_m + 1 \leq G$ such rows in A, G balanced groups do not exist. ■

We note that if G is divisible by $j_m + 1$ then not only the balanced partition, but even a single balanced group cannot be constructed (for example, for odd m and even G).

To visualize the 'worst-cases' with identical fixed attributes, in Table 3.2 we present the the number of attributes, C, for different $G, m, k = 2m$ for which balanced partition may not exist. It is easy to observe that in the most cases* relatively few attributes are required.

*Some of the entries that correspond to the "small" G cannot be established using our analytical results presented in the current Section. We suggest an approach to finding such values in Section 3.3.2.

	m											
	1	2	3	4	5	6	7	8	9	10	11	12
$G=2$	3		3		3		3		3		3	
$G=3$	3	4	3	4	3		3	4	3	4	3	
$G=4$	3	4	3	4	3	5	3	4	3	4	3	
$G=5$	3	4	3	4	3	5	3	4	3	4	3	6
$G=6$	3	4	3	4	3	5	3	4	3	4	3	6

Table 3.2: Number of attributes, C, for which balanced partition does not exist, for different $G, m, k = 2m$.

Next we show that this observation and the the block-matrix approach in general can be extended to the equivalence classes with non-identical fixed and variable attributes.

The Case with Non-identical Fixed Attributes

In this section we consider BMASP problems in which all attributes are of the fixed, but not necessarily of identical types (recall that the type of an attribute is a pair (min_j, max_j)). For notational convenience we refer attribute of type (h, h) as type h. Note that in the previous subsection, all attributes were assumed to be of the same type, m.

Using the notation from Section 3.2, $T \subset \{1, 2, ...H\}$ denotes the set of attribute types, and $t_h \in [0, C]$ denotes the number of attributes of type h, $h \in T$. For notational convenience we assume that $T = \{1, 2, ...H\}$, and if an instance does not have any attributes of a certain type, e.g. h, then $t_h = 0$. This way H is the largest number of objects per group that possess the same attribute. Consistent with this, Assumption 1 implies $k \geq 2H$. Thus the corresponding equivalence class is $\mathbf{E}(G, k, \{1...H\}, \{t_1...t_H\})$.

It follows from Theorem 3.3 and Observation 1 that if there exists an attribute type $h \in T$ such that $t_h \geq j_h + 2$ and $G \geq j_h + 1$, then we can construct an instance which cannot be partitioned into G balanced groups. Below we provide a stronger result by showing that even when $t_h \leq j_h + 1$ holds for all h, balanced partitions may not exist when the number of attributes of some type exceeds the j_h-value of another, "less divisible" type (i.e., the j_h value of which is smaller).

Theorem 3.4 *Suppose in* $\boldsymbol{E}(G, k, \{1...H\}, \{t_1...t_H\})$ *there exist two attribute types, i and i', $t_i, t_{i'} \geq 1$, such that $i \leq i'$, $j_i \leq j_{i'}$, $t_{i'} \geq j_i + 1$. Then for $G \geq j_i + 1$ there exists an instance in* $\boldsymbol{E}(G, k, \{1...H\}, \{t_1...t_H\})$ *for which a balanced partition into G groups does not exist.*

Proof. By Observation 1 it is sufficient to consider only the equivalence classes with $t_i = 1$, $t_{i'} = j_i + 1$. Suppose $G = j_i + 1$.

Create $((i(j_i + 2) + (i' - i)G) \times (j_i + 2))$ matrix A by adding (one below another) i blocks $D(j_i + 2, j_i + 2, j_i + 1)$ and add (from below) $(i' - i)G$ rows of type 1 (recall that we say that a row is of type i if its i^{th} coordinate is zero and all other coordinates are ones). By construction A has $(j_i + 2)$ columns and column sum $i(j_i + 1) = iG$ in the first column and $i(j_i + 1) + (i' - i)G = i'G$ in other $j_i + 1 = t_{i'}$ columns. Note that $kG - i'G - i \geq (k - H)G - H \geq H(G - 1) \geq 0$ because by Assumption 1 $k \geq 2H$ and by construction $i, i' \leq H$. Therefore augmenting A from below with $kG - i'G - i$ all-zeros rows results in matrix A of an instance in $\mathbf{E}(G, k, \{i, i'\}, \{1, j_i + 1\})$.

With respect to an attribute of type i, each balanced group must contain i rows of types $2, ...j_i + 2$ (because rows of type 1 do not have 1 in the first coordinate). Let k_p be the number of rows of type $p = 2, ...j_i + 2$ that such a group contains. Then in columns $p = 2, ...j_i + 2$, this group has column sums $i - k_p$. Thus in order to have equal column sums with respect to attributes of type i' the group should contain equal number of rows of each of the types $p = 2, ...j_i + 2$ ($j_i + 1$ types in total). Hence, i should be divisible by $j_i + 1$, which contradicts the definition of j_i.

Finally, if $G > j_i + 1$ then, modify A' as in the Case 2 of the proof of Theorem 3.3; the claim holds by the same argument. ■

Type 1	Type 2	Type 2
0	1	1
1	0	1
1	1	0
0	1	1
0	1	1

Table 3.3: An instance in $\mathbf{E}(2, k, \{1, 2\}, \{1, 2\})$ for which partition does not exist (all-zero rows are omitted)

To visualize the difference between Theorems 3.3 and 3.4 consider the equivalence class with one attribute of type 1 and two attributes of type 2. None of these attributes alone could prevent the existence of balanced partition, since, by Theorem 3.3 there has to be $j_1 + 2 = 3$ or more attributes of type 1 or $j_2 + 2 = 4$ or more attributes of type 2. At the same time, an equivalence class containing both these types must contain an instance where balanced partition does not exist, since by Theorem 3.4 with $i = 1, i' = 2$ we have $t_2 = 2 \geq 2 = j_1 + 1$. An example of such instance is shown on Table 3.3 for the case with $G = 2$ (all-zeros rows are omitted).

Note that Theorem 3.4 assumes that a less divisible attribute must be of a "lower" type (condition $i \leq i'$ in the statement of the Theorem). Next we discuss a more general case with two attribute types: t_i of type i and $t_{i'}$ of type i', and no assumption about the ordering. For the ease of exposition assume that $j_i \leq j_{i'}$ and $i \leq i'$. We also assume that $t_i \leq j_i + 1$, since otherwise by Theorem 3.3 a partition may not exist; and that $t_i \geq 2$, since if $t_i = 1$ then $t_{i'} = j_{i'} + 2 - t_i = j_{i'} + 1 \geq j_i + 1$ and a thus a partition may not exist by Theorem 3.4.

Theorem 3.5 *Suppose in $\mathbf{E}(G, k, \{1...H\}, \{t_1...t_H\})$ there exist two attribute types, i and i', $t_i, t_{i'} \geq 1$, such that $t_i + t_{i'} \geq j_{i'} + 2$, $i \neq t_{i'}$ and $i' < i\frac{t_i}{t_i - 1}$. Then for G divisible by $j_i + 1$ there exists an instance in $\mathbf{E}(G, k, \{1...H\}, \{t_1...t_H\})$ for which a balanced partition does not exist.*

Proof. In light of Observation 1 suppose $t_i + t_{i'} = j_{i'} + 2$ and w.l.o.g. assume $G = j_{i'} + 1$. Create matrix A' by adding one below another i blocks $D(j_{i'}+2, j_{i'}+2, j_{i'}+1)$ and augment it with $G(i' - i)$ row vectors of length $t_i + t_{i'}$ with zeros in the first t_i components and ones otherwise. For notational convenience we denote such row as 0...01...1. By construction A' has column sums Gi in the first t_i columns and sums Gi' in columns $t_i + 1, ...t_i + t_{i'}$. Thus for $k = 2i'$, augment A' with $kG - i'G - i \geq 0$ all zero rows; the resulting matrix (refer to it as A) is an attribute matrix for an instance in $\mathbf{E}(G, k, \{i, i'\}, \{t_i, t_{i'}\})$. For notational convenience we refer columns $1, ...t_i$ as type i and columns $t_i + 1, ...t_i + t_{i'}$ as type i'.

Observe that a group that is balanced with respect to attributes of type i must contain either of the following rows (we refer to such groups as types 1,2,3 respectively):

1. i rows of types $t_i + 1, ...t_i + t_{i'}$ and no rows of types $1, ...t_i$;

2. $\frac{i}{t_i - 1}$ rows of each of the types $1, ...t_i$ and no rows of types $t_i + 1, ...t_i + t_{i'}$. Note $\frac{i}{t_i-1}$ is an integer, because $t_i \geq 2, t_i \leq j_i + 1$.

3. p copies of rows of each of the types $1, ...t_i$, $p \in [1, ...\frac{i}{t_i-1} - 1]$, plus $i - p(t_i - 1)$ rows of types $t_i + 1, ...t_i + t_{i'}$. Note $i - p(t_i - 1) \geq 1$.

For a group of type 1, observe that unless $i = t_{i'}$, by the same argument as in the proofs of Theorems 3.3 and 3.4 there exist two columns of type i' with unequal column sums. Thus if $i \neq t_{i'}$, since all remaining rows have ones in the columns of type i', such group cannot be balanced with respect to attributes of type i'.

A group of type 2 has column sums $i\frac{t_i}{t_i-1}$ in every column of type i'. Thus if $i' < i\frac{t_i}{t_i-1}$ then such group cannot be balanced; otherwise we must add $i' - i\frac{t_i}{t_i-1}$ rows 0...01...1.

A group of type 3, in order to have equal column sums, must contain equal number of zeros in each column of type i'. Since there are in total $i - p(t_i - 1)$ zeros in columns of type i' and there are $t_{i'}$ such columns it follows that $p = \frac{i - lt_{i'}}{t_i - 1}$ for $l = 1, ...\lfloor\frac{i}{t_{i'}}\rfloor$. Since

Type 2	Type 2	Type 3	Type 3
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0
0	0	1	1
0	0	1	1
0	0	1	1

Table 3.4: An instance in $\mathbf{E}(3, k, \{2, 3\}, \{2, 2\})$ for which balanced partition does not exist (all-zero rows are omitted)

a balanced group must have columns sums i' in every column of type i', we must add $i' - (i - p(t_i - 1) - 1) > i' - i$ rows 0...01...1.

Finally observe that under the conditions of the Theorem, the only balanced groups are that of type 3 and thus G balanced groups of type 3 require $G(i' - (i - p(t_i - 1) - 1)) > G(i' - i)$ rows 0...01...1. ■

To visualize the difference between Theorems 3.4 and 3.5 consider equivalence class with 2 attributes of type 2 and 2 attributes of type 3; attribute matrix of an instance with $G = 3$ for which balanced partition does not exist is presented in Table 3.4. Theorem 3.4 cannot be applied to this example, since a less divisible attribute is of the higher type. At the same time, $i = 3, i' = 2$ and $3 \neq t_2 = 2$, $2 < 3 * 2$, thus a partition does not exist by Theorem 3.5. Note, that a single balanced group can be constructed in this case, but a partition into three groups cannot.

General Case with Fixed and Variable Attributes

In subsection 3.3.1 and Theorem 3.3 we considered the case when all attributes are of the same fixed type (i, i) for some i and showed that there exists a threshold $j_i + 2$ such that if the number of attributes of type (i, i) exceeds it, then balanced partitions may not exist. Then in Section 3.3.1 and Theorem 3.4 we extended this result by showing balanced partition may not exist when this threshold is exceeded by a combination of two fixed attributes. Next we show that this result extends to a combination of fixed and variable attributes.

Suppose that an instance has attributes of two types, $(i, i + 1)$ and (i, i). That is for attribute of the latter type $min_j = max_j = i$, while for the former $min_j = i$ and $max_j = i + 1$. Denote such equivalence class for some k, G as $\mathbf{E}$.

Theorem 3.6 *If $t_{(i,i)} \geq j_i + 1$, $G > i$ and is divisible by $j_i + 1$ and $j_{i+1} \leq j_i$ then there exists an instance in $\boldsymbol{E}$ that cannot be partitioned into G balanced groups.*

Proof. If the claim is true for $t_{(i,i)} = j_i + 1$ and $t_{(i,i+1)} = 1$ then it is also true for $t_{(i,i)} \geq j_i + 1$ and $t_{(i,i+1)} \geq 1$ by Observation 1. Thus assume $t_{(i,i)} = j_i + 1$, $t_{(i,i+1)} = 1$ and $G = j_i + 1$; an extension to G divisible by $j_i + 1$ is straightforward. Let $c_{(i,i+1)}$ denote the column sum for attribute of type $(i, i + 1)$.

Create attribute matrix A by adding i blocks $D(j_i + 2, j_i + 2, j_i + 1)$ and $c_{(i,i+1)} - Gi$ rows with one in the first coordinate and zeros otherwise (we refer to such rows as $(1, 0, ...0)$. Augment A by $kG - c_{(i,i+1)} \geq 0$ all-zero rows. By construction A is an attribute matrix of an instance in $\mathbf{E}$.

Suppose a balanced partition of $\mathbf{E}$ exists. Then since by construction there are $i < G$ rows of type 1, there must exist a group that does not contain a row of type 1. By the same argument as in the proof of Theorem 3.3, in order to be balanced with respect to the attributes of type (i, i) such a group must contain equal number of copies of rows of each of the types $2, ..., j_i + 2$ ($j_i + 1$ types in total). Thus with respect to the attribute of type $(i, i + 1)$ it has column sum divisible by $j_i + 1$. By conditions of the Theorem $j_{i+1} \leq j_i < j_i + 1$ and therefore such a group is not balanced. ■

In words, a BMASP problem that contains a mixture of attributes of fixed and variable types could be as hard as its fixed 'subproblem'. Indeed, by comparing Theorems 3.4 and 3.6 substituting some fixed attributes with variable, but keeping 'enough' fixed attributes in place still allows us to characterize the cases when balanced partitions do not exist.

Next we present an alternative approach to searching for BMASP instances in which balanced partitions do not exist.

3.3.2 Cover Problem Approach

Consider sub-class $\bar{c}_{\mathbf{E}}$ in equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$. Suppose that balance partitions exist in all instances in this sub-class. How many attributes of a certain type can be added before there would exist an instance for which a balanced partition does not exist? Likewise, in the opposite case when a sub-class contains an instance for which balanced partition does not exist, by how much do we have to decrease the number of attributes of a given type in order to guarantee that all instances in the resulting subclass can be partitioned into balanced groups?

In this subsection we suggest an approach to answer these questions numerically. It further allows us to obtain a theoretical result that compares the existence of groups in the classes with fixed and variable attributes.

Observe that an attribute (column in A) with column sum c_j is a binary column-vector of length kG with c_j ones and $kG - c_j$ zeros. For instance, attribute "Gender" in the example from Table 3.1 is a binary vector $(1, 1, 0, 0)$. In a similar way, a group is a binary column-vector of length kG with k ones and $kG - k$ zeros. For example, group "Monica and Rachel" is a binary vector $(0,0,1,1)$. Since all these vectors are of length kG we use "$x-$vector" to denote a binary vector of length kG with x ones and $kG - x$ zeros. That is an attribute is an c_j-vector and a group is a $k-$vector.

For given c_j let $\mathbf{U}_{c_j}$ be the set of all non-identical attribute c_j-vectors; clearly, there are $|\mathbf{U}| = \binom{kG}{c_j}$ ("kG choose c_j") elements in $\mathbf{U}_{c_j}$. Note that if an instance in $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\overline{h})}\})$ has an attribute with column sum c_j then the corresponding c_j-vector is a member of $\mathbf{U}_{c_j}$. Similarly, let $\mathbf{V}$ be the set of all possible groups of k objects from the available kG objects; there are $|\mathbf{V}| = \binom{kG}{k}$ non-identical group $k-$vectors.

We say that an attribute vector $\mathbf{u} \in \mathbf{U}_{c_j}$ of an attribute j of type $(h, \overline{h})$ covers group vector $\mathbf{v}$ if

$$\begin{array}{rcl} \mathbf{u}^T\mathbf{v} & \leq & h - 1 \\ & or & \\ \mathbf{u}^T\mathbf{v} & \geq & \overline{h} + 1 \end{array}$$

that is too many or too few objects from group $\mathbf{v}$ contain attribute j and thus group $\mathbf{v}$ is not balanced with respect to this attribute.

Definition 2 *Let* $\{b_{u(c_j)v}\}$ *be a* cover matrix, *with elements* $b_{u(c_j)v} = 1$ *if attribute* $\boldsymbol{u} \in \boldsymbol{U}_{c_j}$ *with column sum* c_j *covers group* $\boldsymbol{v} \in \boldsymbol{V}$, *and* $b_{u(c_j)v} = 0$ *otherwise.*

To visualize the concept of cover matrix consider the example from Table 3.1. In this case $N = 4, G = 2, k = 2$ and $c_j = 2$ for $j = 1, 2, 3$. There are six group vectors: $(1,1,0,0)$, $(1,0,1,0)$, $(1,0,0,1)$, $(0,1,0,1)$ and $(0,0,1,1)$; likewise there are the same six attribute vectors. Table 3.5 (a) presents all possible attribute vectors (there are more attributes than that presented in the actual example in Table 3.1). Consider group vector #1, $(1,1,0,0)$, and attribute vectors # 1 and 2, $(1,1,0,0)$ and $(1,0,1,0)$ respectively. In the first case

a_{ij}	Attribute vectors					
Objects	(1,1,0,0)	(1,0,1,0)	(1,0,0,1)	(0,1,1,0)	(0,1,0,1)	(0,0,1,1)
1	1	1	1	0	0	0
2	1	0	0	1	1	0
3	0	1	0	1	0	1
4	0	0	1	0	1	1

(a)

b_{uc_jv}	Attribute vectors					
Group vectors	(1,1,0,0)	(1,0,1,0)	(1,0,0,1)	(0,1,1,0)	(0,1,0,1)	(0,0,1,1)
(1,1,0,0)	1	0	0	0	0	1
(1,0,1,0)	0	1	0	0	1	0
(1,0,0,1)	0	0	1	1	0	0
(0,1,1,0)	0	0	1	1	0	0
(0,1,0,1)	0	1	0	0	1	0
(0,0,1,1)	1	0	0	0	0	1

(b)

Table 3.5: Attribute vectors (a) and cover matrix (b) for 'Friends' example from Table 3.1.

$b_{1,(2),1} = 1$ because $(1,1,0,0)^T \times (1,1,0,0) = 2 > 1$ (in words, group "Ross and Chandler" is not balanced with respect to attribute "Gender" because they are both males). In the second case $b_{1,(2),2} = 0$ because $(1,1,0,0)^T \times (1,0,1,0) = 1$ (in words, group "Ross and Chandler" is balanced because with respect to attribute "Siblings"). Refer to Table 3.5 (b) for the cover matrix $\{b_{u(c_j)v}\}$; note that attributes in this example have the same c_j value and thus index (2) is omitted.

Let $x_{uc_j} = 1$ if attribute $\mathbf{u} \in \mathbf{U}_{c_j}$ with column sum c_j is selected from $\mathbf{U}_{c_j}$ and $x_{uc_j} = 0$ otherwise.

For a given sub-class $\bar{c}_{\mathbf{E}}$ in equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$ let $\tilde{c} \in \bar{c}_{\mathbf{E}}$ be a column sum of some column. Note that there may be several columns with column sum $\tilde{c}$. Let $Z_1^{\tilde{c}}(\mathbf{E}, \bar{c}_{\mathbf{E}})$ be the smallest number of attributes with column sum $\tilde{c}$ adding which and keeping the number and column sums of the other attributes (which sum is not $\tilde{c}$) unchanged, results in an instance for which a single balanced group cannot be constructed. Note that even though we are interested in the existence of balanced partitions into G groups, such a partition obviously cannot exist in the cases when even a single groups does not exist. Therefore such a single-group subproblem is of interest as well.

$Z_1^{\bar{c}}(\mathbf{E}, \bar{c}_{\mathbf{E}})$ can be found by solving the following set cover problem.

$(Z_1^{\bar{c}}(\mathbf{E}, \bar{c}_{\mathbf{E}}))$

$$\min \sum_{c_j \in \bar{c}} \sum_{u \in U_{c_j}} x_{uc_j} \tag{3.4}$$

subject to

$$\sum_{c_j \in \bar{c}} \sum_{u \in U_{c_j}} b_{u(c_j)v} x_{uc_j} \geq 1 \text{ for all } v \in \mathbf{V} \tag{3.5}$$

$$\sum_{u \in U_{c_j}} x_{uc_j} = t_{c_j} \text{ for all } c_j \in \bar{c}_{\mathbf{E}}, c_j \neq \tilde{c} \tag{3.6}$$

$$x_{uc_j} \in \{0; 1\}$$

where t_x is the number of attributes with $c_j = x$.

If $Z_1^{\bar{c}}(\mathbf{E}, \bar{c}_{\mathbf{E}}) > t_{\tilde{c}}$ then a single balanced group can be constructed in all instances in sub-class $\bar{c}_{\mathbf{E}}$ in equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$. Otherwise, let $X^* = \bigcup_{c_j \in \bar{c}} \{u | x^*_{uc_j} = 1, u \in \mathbf{U}_{c_j}\}$ be a set of binary column-vectors that are selected in the optimal solution. Construct matrix A by including the corresponding X^* selected attribute column-vectors. By definition A is the attribute matrix of an instance in sub-class $\bar{c}_{\mathbf{E}}$ in $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$ for which a balanced group cannot be created.

Table 3.6 presents Z_1 values obtained by solving the cover problem $Z_1^{\bar{c}}(\mathbf{E}, \bar{c}_{\mathbf{E}})$ for both the cases with fixed and variable attribute types for $G = 2$ and various k, c_j combinations (all attributes have equal c_j values). It is easy to see that, as suggested by Theorem 3.6, attributes of variable type indeed require substantially more columns in order to ensure that groups do not exist, e.g. for $k = 6$, as little as 3 attributes of type $(1, 1)$ can prohibit the existence of balanced groups[†], while we need at least 7 attributes of type $(1, 2)$.

[†] In fact, we established this result analytically in Theorem 3.3

	c_j					
	2	3	4	5	6	7
Type	(1,1)	(1,2)	(2,2)	(2,3)	(3,3)	(3,4)
k=3	3	10				
k=4	3	7	4			
k=5	3	7	3	7		
k=6	3	7	4	7	3	
k=7	3	7	4	10	3	11

Table 3.6: $Z_1^{c_j}(\mathbf{E}, \bar{c}_{\mathbf{E}})$ values for $G = 2$, $k = 3, ...7$ and $c_j = 2, ...7$ (all attributes have the same c_j value).

Does this observation imply that a BMASP problem with variable attributes is necessarily easier than that with fixed? Next we show that this in indeed the case for some sub-classes.

Theorem 3.7 *Consider sub-class $\bar{c}$ in the equivalence class $\boldsymbol{E}(G, k, \boldsymbol{T}, \{t_{(h,\bar{h})}\})$, in which there exists an attribute (column) j, of variable type $(i, i+1) \in \boldsymbol{T}$ such that $c_j = iG + 1$. Let $\boldsymbol{E}^1$ and $\bar{c}^1$ be the equivalence class and sub-class obtained from $\boldsymbol{E}$ by substituting j with an attribute (column) of type (i, i).*

Then $Z_1^{\bar{c}}(\boldsymbol{E}, \bar{c}) \geq Z_1^{\bar{c}}(\boldsymbol{E}^1, \bar{c}^1)$.

Proof. Let A be the attribute matrix of an instance in sub-class $\bar{c}$ in the equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$. For some row r such that $a_{rj} = 1$, construct matrix A^1 such that $a^1_{rj} = 0$ and $a^1_{pq} = a_{pq}$ otherwise. That is, A and A^1 are identical, except for element (r, j). By construction, A^1 is an attribute matrix of an instance in sub-class $\bar{c}^1$ in $\mathbf{E}^1$.

Consider two problems of covering group vectors in $\mathbf{E}$ and $\mathbf{E}^1$ by only the attributes (columns) from A and A^1 respectively. By construction group vectors in both classes are identical and so are column vectors except for column j. Therefore, $b_{u(\cdot)v} = b^1_{w(\cdot)v}$ for all column vectors $u = w$, $u, w \neq j$ and group vectors v, where $\{b_{u(\cdot)v}\}$ and $\{b^1_{w(\cdot)v}\}$ be the corresponding cover matrices.

In column j observe that for an instance in $\mathbf{E}$ (with attribute matrix A), group vectors are covered by a $iG+1-$vector (we refer to it as u), while for an instance in $\mathbf{E}^1$ with attribute matrix A^1, by a $iG-$vector (we refer to it as w) with the same components equal 1, except for component r.

Next we show that $b_{j(iG+1)v} \leq b^1_{j(iG)v}$, that is, if a group vector is covered in an instance with attribute matrix A then it is also covered in an instance with matrix A^1. Intuitively this happens because in the former case a balanced group could contain either min_j or $min_j + 1 = max_j$ of attribute j, but in the latter $min_j = max_j$.

Formally, let $v(r)$ denote the r^{th} component of group vector v (recall r by definition is the index of the row in which in column j we substituted 1 by 0). Each group vector v can be of either of the two types:

v(r)=0 Then by construction $u^T v = w^T v$. Hence:

- if $u^T v \leq i-1$ then $b_{j(iG+1)v} = b^1_{j(iG)v} = 1$;
- if $u^T v = i$ then $b_{j(iG+1)v} = b^1_{j(iG)v} = 0$;
- if $u^T v = i+1$ then $b_{j(iG+1)v} = 0 \leq b^1_{j(iG)v} = 1$;
- if $u^T v \geq i+2$ then $b_{j(iG+1)v} = b^1_{j(iG)v} = 1$.

v(r)=1 Then by construction $u^T v = w^T v + 1$. Hence:

- if $u^T v \leq i-1$ then $w^T v \leq i-2$ and so $b_{j(iG+1)v} = b^1_{j(iG)v} = 1$;
- if $u^T v = i$ then $w^T v = i-1$ and so $b_{j(iG+1)v} = 0 \leq b^1_{j(iG)v} = 0$;
- if $u^T v = i+1$ then $w^T v = i$ and so $b_{j(iG+1)v} = b^1_{j(iG)v} = 0$;
- if $u^T v \geq i+2$ then $w^T v \geq i+1$ and so $b_{j(iG+1)v} = b^1_{j(iG)v} = 1$.

Therefore from (3.5) and (3.6) the feasible region for the cover problem representation of the instance in $\mathbf{E}^1$ (with attribute matrix A^1) is contained in the feasible region for $\mathbf{E}$

(with attribute matrix A), and hence by the principle of optimality $Z_1^{\bar{c}}(\mathbf{E},\bar{c}) \geq Z_1^{\bar{c}}(\mathbf{E}^1,\bar{c}^1)$.

■

By considering a similar argument, Theorem 3.7 can be extended to equivalence classes with variable attributes with $c_j = i'G - 1$. We summarize it in the following corollary:

Corollary 3.1 *Consider sub-class $\bar{c}$ in the equivalence class $\boldsymbol{E}(G, k, \boldsymbol{T}, \{t_{(h,\bar{h})}\})$, in which there exists an attribute (column) j, of variable type $(i, i+1) \in \boldsymbol{T}$ such that $c_j = (i+1)G-1$. Let $\boldsymbol{E}^2$ and $\bar{c}^2$ be the equivalence class and sub-class obtained from $\boldsymbol{E}$ by substituting j with an attribute (column) of type $(i+1, i+1)$.*

Then $Z_1^{\bar{c}}(\boldsymbol{E},\bar{c}) \geq Z_1^{\bar{c}}(\boldsymbol{E}^2,\bar{c}^2)$.

For $G = 2, 3$ for all variable attributes $c_j = min_j G + 1$ or $c_j = max_j G - 1$. Therefore BMASP problem with variable attributes in these cases is always 'easier'. However, for $G \geq 4$ there exist variable attributes with c_j−values that do not satisfy conditions of Theorem 3.7 and Corollary 3.1, e.g. when $c_j = iG + 2$, and therefore we could not prove that BMASP problems with variable attributes are necessarily easier for such G. This is because for small G, (i.e. under the conditions of Theorem 3.7 and Corollary 3.1) all groups in the instance with a fixed attribute can be covered by modifying the attribute matrix of an instance with variable attribute (by changing some 1 to 0). For large G such modified attribute vectors may not cover all groups; however, it could very well be that some other $C' < Z_1^{\bar{c}}(\mathbf{E},\bar{c})$ attributes would cover all groups. For example, for $k = 3$ and $G = 4, 5$ and equivalence classes where all attributes have equal c_j−values we established numerically that the number of attributes of a variable type required to cover all groups is always larger than the number of the corresponding fixed attributes; see Table 3.7. This was the case in all our numerical experiments and therefore we conjecture that such a property holds for all c_j values.

To conclude the worst-case analysis we apply our results to an example with $k = 6$,

$Z_{\bar{1}}^{\tilde{c}}(\mathbf{E},\bar{c})$	$\tilde{c}$					
	2	3	4	5	6	7
G=4	30	11	3	8	9	
G=5	49	18	12	4	7	7

Table 3.7: $Z_{\bar{1}}^{\tilde{c}}(\mathbf{E},\bar{c})$ values for $G = 4, 5$, $k = 3$ and $\tilde{c} = 2, ...7$ (all attributes have the same c_j values).

which is most representative of the MBA study groups problem that motivated our work.

3.3.3 Example: equivalence classes with $k = 6$

In this Section we illustrate our findings by considering several equivalence classes with $k = 6$. Note, that in light of Assumption 1 it is sufficient to consider only the classes with attributes of types $(h, \bar{h})$ with $\bar{h} \leq 3$. We further restrict the example to the case with $G = 2$ whenever the results are obtained numerically. Therefore we do not consider attributes of type $(0, 1)$; for $G = 2$ such an attribute has $c_j = 1$ and hence it does not influence the existence of balanced groups. Ten illustrative equivalence classes are presented in Table 3.8.

Classes 1-5 contain only fixed attributes. Class 1 contains too many attributes of a single type, and thus groups do not exist with respect to this attribute alone by Theorem 3.3. Classes 2 and 3 contain less than the critical number of attributes of each type, but their combination is such that groups do not exist by Theorems 3.4 and 3.5 respectively. Classes 4 and 5 illustrate the importance of the combination of different attributes; indeed, the total number of attributes is the same in the classes 1,2,4 and 5, yet in some classes balanced partition is guaranteed to exist while in others it is not.

Classes 6-10 contain both fixed and variable attributes. Class 6 illustrates the extension of the block-building concept of Theorems 3.3 - 3.5 to the classes with variable attributes. Class 7 illustrates the covering result of Theorem 3.7. Finally, classes 8-10 illustrate how

Eq. class	$t_{(1,1)}$	$t_{(2,2)}$	$t_{(3,3)}$	$t_{(1,2)}$	$t_{(2,3)}$	Must partition exist?	Reason
$\mathbf{E}_1$	3	0	0	0	0	No	Theorem 3.3
$\mathbf{E}_2$	1	2	0	0	0	No	Theorem 3.4
$\mathbf{E}_3$	0	2	2	0	0	No	Theorem 3.5
$\mathbf{E}_4$	1	1	1	0	0	Yes	$Z_1^{(3,3)}(\mathbf{E}_4, \bar{c}) = 2$
$\mathbf{E}_5$	2	1	0	0	0	Yes	$Z_1^{(2,2)}(\mathbf{E}_5, \bar{c}) = 2$
$\mathbf{E}_6$	0	3	0	0	1	No	Theorem 3.6
$\mathbf{E}_7$	2	0	0	1	0	No	Theorem 3.7
$\mathbf{E}_8$	1	0	0	5	0	Yes	$Z_1^{(1,2)}(\mathbf{E}_8, \bar{c}) = 6$
$\mathbf{E}_9$	1	1	0	3	0	No	$Z_1^{(1,2)}(\mathbf{E}_9, \bar{c}) = 3$
$\mathbf{E}_{10}$	1	0	1	3	0	Yes	$Z_1^{(1,2)}(\mathbf{E}_{10}, \bar{c}) = 4$

Table 3.8: The existence of balanced partitions for different equivalence classes with $k = 6$. In classes 4,5,8,9 and 10 $G = 2$, in classes 1 and 7 $G \geq 3$, otherwise $G \geq 4$.

different combinations of fixed and variable attributes influence the existence of balanced partitions when the total number of attribute is the same (similar to Classes 4 and 5 for only fixed attributes).

We note that due to the large dimensionality of the covering problems, we were unable to solve similar examples for larger G. However, as the discussion of the probabilistic analysis (below) shows, such probabilities very quickly stabilize for $G \leq 5$, thus, we believe that the qualitative conclusions of this example hold for all G.

3.4 Probabilistic Analysis

From the preceding discussion we know that there exist equivalence classes that contain instances which cannot be partitioned into balanced groups. However, such instances may be relatively rare within the corresponding equivalence class. This leads to the following questions: for a given equivalence class for which it cannot be guaranteed that a balanced partition exists in all its instances, how practical is the BMASP approach to group balancing? The answer depends on the probability that a partition exists in a randomly selected instance. In this Section we estimate a probability that an arbitrary instance in a given class can be partitioned into balanced groups.

We describe three complementary approaches. First, we present a general network-based representation of group creation problem, and use it to estimate the probability that a balanced partition exists through numerical simulations. These simulations are based on the covering representation of the problem of creating one group described in Section 3.3.2. Our second approach is a distribution-free upper bound, where to obtain a rigorous bound we note that the distribution of the number of groups that exist is difficult to find and so we build our analysis not relying on knowing such a distribution. Finally our third approach presents a lower bound, based on empirically estimating a fraction of instances for which a balanced partition has been found by solving BMAPS integer programs numerically.

3.4.1 Network Representation

Consider a given subclass $\bar{c}_{\mathbf{E}}$ in equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$. We view the problem of creating balanced groups as a sequential process. We create one balanced group and delete the objects in this group. Then we create the second group from the remaining objects, delete its objects, and continue doing so until, hopefully, all objects are assigned to balanced groups, i.e., a balanced partition is constructed. As we discuss next, there are many different kinds of groups and they may not always exist. Therefore in constructing groups sequentially we must account for both these issues. It is convenient to do so using the following random network; an example if provided on Figure 3.1.

The nodes of the network correspond to the different subproblems that occur in constructing groups sequentially. The nodes are grouped into *stages*, where stage $g = G, G - 1, ..., 2, 1$, signifies the number of groups yet to be created. Each node at stage g represents a subclass that contains all instances that can be obtained if the objects that form the first

$G-g$ balanced groups of certain types (see below) created at stages $G, G-1, ...g+1$ are deleted from the sets of objects of all instances in $\bar{c}_{\mathbf{E}}$. At stage G there is only a single node representing all instances in $\bar{c}_{\mathbf{E}}$. At other stages there may be multiple nodes. Let $\overline{c(g)} = \{c_1(g), c_2(g), ..., c_C(g)\}$ denote a node at stage g. Thus, the node is the vector of column sums at stage g (the amounts of each attribute that must be used in creating the last g groups), i.e., a sub-class. We set $\overline{c(G)} = \bar{c}_{\mathbf{E}}$. To construct all nodes that may arise at different stages for a given $\bar{c}_{\mathbf{E}}$, let $l_j(g)$ and $u_j(g)$, $j = 1, ..., C$ denote the lower and upper bounds such that given $c_j(G), min_j, max_j$ at stage g, $c_j(g) \in [l_j(g), u_j(g)]$. Let q_j is the solution to diophantine equation $q_j min_j + (G-q_j)max_j = c_j(G)$, i.e. a balanced partition with respect to attribute j consists of q_j groups with min_j of this attribute and $G-q_j$ groups with max_j. Then $l_j(g) = c_j - \min[G-g, G-q_j]max_j - \max[q_j - g, 0]min_j$ and $u_j(g) = c_j - \min[G-g, q_j]min_j - \max[G-g-q_j, 0]max_j$. Therefore, the nodes at stage g can be obtained by enumerating all permutations of $c_j(g) \in [l_j(g), u_j(g)]$ values for all $j = 1, ..., C$.

The arcs of the network correspond to group creation and represent a *transition* from an instance at stage g to an instance at stage $g-1$. To describe these transitions we say that a group I is of *type* f if a binary representation of integer f, a binary vector of length C, $Bin_f = \{Bin_f(1), ..., Bin_f(C)\}$, is such that $Bin_f(j) = 0$ if group I has min_j objects that possess attribute j and $Bin_f(j) = 1$ if I possesses max_j such objects. For example, a group that contains min_j objects with all attributes $j = 1, 2, ..., C$ has $Bin_f = \{0, 0, ..., 0\}$ and thus is of type $f = 0$. There are $O(2^C)$ group types and there is one arc exiting every node per each type that is balanced at this node. Observe that not all group types that are balanced at node $\bar{c}_{\mathbf{E}}$ are balanced at all other nodes; for example, if for some attribute j and stage g such that $q_j \leq G-g$, all groups that were constructed at stages $G, G-1, ...G-g$ had min_j of objects with attribute j, then the remaining g groups must all have max_j objects with attribute j, i.e., a group with min_j such objects is not balanced at stages $g, g-1, ...1$, see node $(3, 6)$ on Figure 3.1. We say that an arc is of type f if it corresponds to a group of this type.

Let $\overline{c^f(g)}$ be the end node at stage $g-1$ of the arc of type f that exists node $\overline{c(g)}$ at stage g. In words, because nodes represent subclasses, $\overline{c^f(g)}$ is a subclass that contains all instances that remain after the objects comprising a group of type f are removed from any instance in $\overline{c(g)}$. Note, that any group of type f exiting a given $\overline{c(g)}$ points to the same end node $\overline{c^f(g)}$; in realistic subclasses there are typically thousands of different combinations of objects that lead to a group of a given type. Binary representation for a group of type f implies:

$$c_j^f(g) = c_j(g) - (Bin_f(j)max_j + (1 - Bin_f(j))min_j) \tag{3.7}$$

for all $j = 1, ..., C, g = G, ..., 2$.

Note, that for a given instance in $\overline{c(g)}$ it may not be possible to create a balanced group that would result in the remaining problem belonging to $\overline{c^f(g)}$ – such a transition may be possible from only some of the instances in $\overline{c(g)}$. Therefore, for the arc of type f, $\left(\overline{c(g)}, \overline{c^f(g)}\right)$, we define a quantity, $P(g, f, \overline{c(g)})$, measuring the fraction of instances in $\overline{c(g)}$ in which a balanced group r esulting in such a transition exists – the probability that arc (balanced group) of type f, exists at node (subclass) $\overline{c(g)}$ at stage g.

Figure 3.1 depicts an example of such a *transition network.* There are $G = 4$ groups and $C = 2$ attributes with $\overline{c(4)} = \{5, 9\}$, i.e., $min_1 = 1, max_1 = 2$ and $min_2 = 2, max_2 = 3$. The first group (at stage 4), could be either of four types, 0,1,2 or 3. In our notation, for example, for a group of type 0, $Bin_0 = (0, 0)$, meaning that it contains $min_1 = 1$ object with attribute 1 and $min_2 = 2$ objects with attribute 2. Thus to associate a physical meaning with this type on the figure we mark it as $(1, 2)$ – see the upper leftmost arc on the Figure. Suppose a group of type 0 is created at node $(5, 9)$; then from (3.7), $c_1^0(4) = 5 - (0 * 2 + (1 - 0) * 1) = 4$ and $c_2^0(4) = 9 - (0 * 3 + (1 - 0) * 2) = 7$. Therefore by creating a group of type 0 at node $(5, 9)$ at stage 4 we transition to the node $(4, 7)$ at stage 3, meaning that 4 objects with attribute 1 and 7 with attribute 2 should be used in

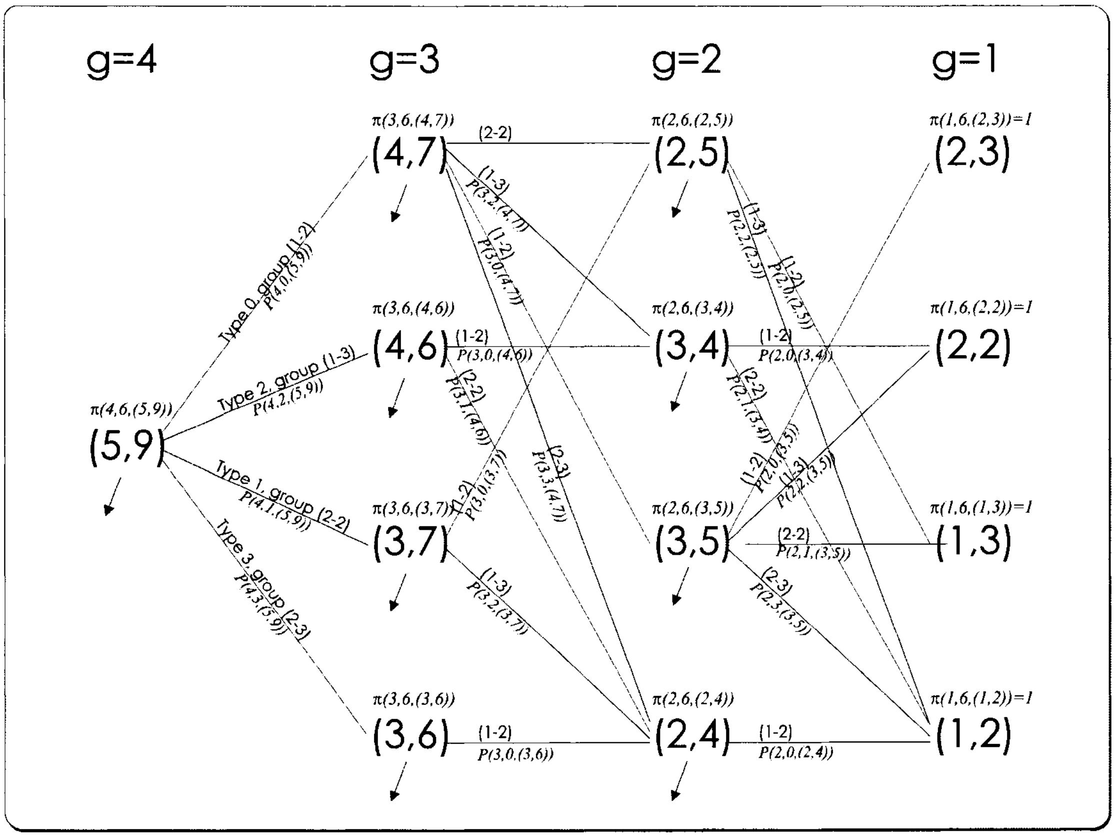


Figure 3.1: Transition network for $C = 2, G = 4$, $c_E = (5, 9)$. The nodes are vectors $\overline{c(g)}$, the marks on the arcs denote the group types and probabilities that such type exist. Probabilities on nodes are the probabilities of reaching any node at stage 1. Exit arcs at nodes represent the cases when no groups can be created.

the remaining 3 groups to be constructed at stages 3,2 and 1.

Let $\pi(g, k, \overline{c(g)})$ be the probability that an arbitrary instance in subclass $\overline{c(g)}$ can be partitioned into g groups of size k. For notational convenience we denote the probabilities on arcs and nodes as the $P-$ and $\pi-$ probabilities respectively.

With this, being at node $\overline{c(g)}$ with arcs of types $f = 0, 1, ...2^C - 1$ with probabilities $P(g, f, \overline{c(g)})$ on these arcs, leading to nodes $\overline{c^f(g)}$ at stage $g - 1$ with probabilities $\pi(g - 1, k, \overline{c^f(g)})$ of building the remaining $g - 1$ groups from these nodes, $\pi(g, k, \overline{c(g)})$ is determined by the following recursion:

$$\begin{aligned}\pi(g,k,\overline{c(g)}) &= \sum_{f=0}^{2^C-1} P(g,f,\overline{c(g)})\pi(g-1,k,\overline{c^f(g)}) \\ &\quad \times \prod_{i=0}^{f-1}\{1-P(g,i,\overline{c(g)})\pi(g-1,k,\overline{c^i(g)})\} \qquad (3.8)\end{aligned}$$

To illustrate recursion (3.8) suppose that in the example on Figure 3.1, $P(4,0,(5,9)) = 0.8$, $P(4,2,(5,9)) = 0.3$, $P(4,1,(5,9)) = P(4,3,(5,9)) = 0$ and $\pi(3,6,(4,7)) = 0.6$, $\pi(3,6,(4,6)) = 0.1$, $\pi(3,6,(3,7)) = 0.5$, $\pi(3,6,(3,6)) = 0.5$. In words, two upper arcs leaving node $(5,9)$ have $P-$probabilities 0.8 and 0.3 and the corresponding nodes have $\pi-$probabilities 0.6 and 0.1; other arcs have zero $P-$probabilities, other $\pi-$probabilities are 0.5. Then $\pi(4,6,(5,9)) = 0.8*0.6+0*0.5*(1-0.8*0.6)+0.3*0.1*(1-0*0.5)*(1-0.8*0.6)+0*0.5*(1-0.3*0.1)*(1-0*0.5)*(1-0.8*0.6) = 0.4956$.

We refer to the $\pi(G,k,\overline{c(G)})$ obtained by recursion (3.8) as the *N-estimate* (network estimate). Note that while in (3.8) we evaluate the nodes at stage $g-1$ in the order of our binary representation for f, (3.7), any other numbering of nodes would lead to the same result. This follows from observing that the probability of forming g groups at some node $\overline{c(g)}$ is 1-probability that groups do not exist.

In order to follow recursion (3.8) we require the knowledge of $P(g,f,\overline{c(g)})$ values, which are the probabilities that a group of type f can be created in an arbitrary instance in subclass $\overline{c(g)}$ with some given number of objects per group, k. Next we discuss now to find such $P-$probabilities. We first clarify what we mean by an 'arbitrary instance'.

Assumption 2 *An arbitrary BMASP instance in the sub-class $\bar{c}_{\boldsymbol{E}}$ in equivalence class $\boldsymbol{E}(G,k,\boldsymbol{T},\{t_{(h,\bar{h})}\})$ is obtained by sampling the corresponding number of attribute column-vectors with column sums c_j from the sets $\boldsymbol{U}_{c_j}$ uniformly at random (i.e., each column-vector in a given set has equal probability to appear) with replacement for each $c_j \in \bar{c}_{\boldsymbol{E}}$.*

That is, the arbitrary instance is not the one which has the entries of its attribute matrix are generated randomly; rather the columns of the attribute matrix are sampled at random with replacement from the corresponding sets of all such columns. In practical applications replacement may apply since correlated (identical, or negatively correlated, e.g., see Table 3.5 (b)) column-vectors may arise.

With this assumption we can evaluate probability measure $P(g, f, \overline{c(g)})$. We were not successful in deriving it analytically; its distribution seem to depend on complicated combinatorial relationships between the column-vectors, and as a result, is does not reflect any standard parametric family (we discuss this issue in detail in subsection 3.4.3). In principle, one could use complete enumeration: create all instances in a given sub-class $\overline{c(g)}$ and then check all groups in the group-vector set $\mathbf{V}$ of its equivalence class (recall that as defined in Section 3.3.2 $\mathbf{V}$ is a set of all group-vectors in a given equivalence class). Then for each group type f, $P(g, f, \overline{c(g)})$ would be given by the ratio of the number of instances in which there exists an uncovered group of type f to the total number of instances in $\overline{c(g)}$. Unfortunately, such an approach is computationally prohibitive in effectively all cases of interest. For example, for $k = 6, g = 2, c(g) = \{6, 6, 6\}$, there are $\binom{12}{6} = 924$ group vectors and the same number attribute vectors. Thus there are $\binom{924}{3} = 131054924$ different instances that must be checked.

Alternative approach is to evaluate $P(g, f, \overline{c(g)})$ values using simulation. To do so in each trial of the simulation we create a random BMASP instance as per assumption 2. Then we test all group-vectors in $\mathbf{V}$ and determine whether in this instance there exists at least one group vector of type f that is balanced (in light of the cover problem analogy from subsection 3.3.2, we could also say 'not covered' to mean being balanced in a given instance). If y denotes the number of such instances (where at least one uncovered group-vector exists) and z is the total number of simulation trials, then $P(g, f, \overline{c(g)}) \approx \frac{y}{z}$.

Since computational complexity of such a simulation depends largely on the size of set $\mathbf{V}$, which is $\binom{kg}{k}$, it can be used very effectively for small k, g. In particular, evaluating

g	C																					
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	1	.98	.92	.85	.71	.58	.41	.31	.2	.12	.09	.03	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	.97	.94	.88	.78	.67	.44	.27	.11	.06	.04	.01	0	0	0	0	0	0
4	1	1	1	1	1	1	.99	.98	.97	.86	.66	.48	.28	.19	.12	.07	.05	.03	0	0	0	0
5	1	1	1	1	1	1	1	1	1	.99	.93	.78	.57	.29	.19	.07	.04	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	1	1	.96	.86	.65	.39	.24	.05	.02	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	.98	.77	.62	.4	.2	.1	.04	.03	.01	0

$(m = 1)$

g	C																					
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	1	1	.96	.8	.58	.3	.14	.09	.03	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	.99	.79	.39	.15	.03	.01	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	.92	.62	.34	.22	.06	.01	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	.95	.7	.42	.17	.1	.06	.02	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	.94	.66	.24	.12	.05	.02	.02	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	.99	.93	.52	.22	.09	.05	.03	0	0	0	0	0	0

$(m = 2)$

g	C																					
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	1	.97	.89	.72	.48	.25	.1	.06	.01	.01	.01	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	.95	.69	.29	.11	.03	.01	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	.99	.88	.47	.21	.08	.01	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	.82	.51	.21	.1	.07	.01	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	.99	.81	.42	.17	.06	.02	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	.97	.7	.34	.15	.06	.01	0	0	0	0	0	0	0

$(m = 3)$

Table 3.9: Estimates for $P(g, \cdot, \overline{c(g)})$ obtained by simulation for the equivalence classes where all attributes are of the same fixed type (i.e. $\overline{c(g)} = \{mg, mg, ..., mg\}$ of length C) for $m = 1, 2, 3$, $g = 2, ..., 7$, $C = 4, ..., 25$ and $k = 6$.

$P(g, f, \overline{c(g)})$ for $g = 2, 3, 4$ is computationally inexpensive and quick for effectively any k of interest; thus we can run thousands of trials and obtain very accurate approximations of the underlying probability measure. For large k, g simulation cannot be used, since the size of the set of group-vectors, $\binom{kg}{k}$, grows exponentially in k, g. In our experiments it took approximately 24 hours to run 256 trials for the cases with $k = 4, g = 10$ and $k = 6, g = 7$, which were the 'largest' cases we considered in simulation.

Running such simulations we noticed that P-probabilities (for a given number of attributes of given types) are the largest when g is large. Table 3.9 presents P-values for the equivalence classes with identical fixed attributes (described in detail in the next subsection), i.e. the type of the group is determined by the number of objects per group, m. For example, consider the case with $m = 1$ and suppose we are interested in the probability that a partition exists in the equivalence class with $g = 7$ and $C = 15$ attributes. Then, at first stage for $g = 7$ the probability that one group exists is 100%, hence we transition to the stage with $g = 6$. Here the probability that one group exists is 96% and so we reach the stage with $g = 5$ with probability 96%. However, even if the process reaches stage with $g = 2$, this happens with probability $\approx$ 10%, the last group almost never exists (3% probability).

Therefore, the probability that a partition exists is effectively determined by the P-values of the last stages (with small g). Since we can accurately approximate them through simulation, the possible inaccuracy of the estimates for large g has little effect on $\pi(G, k, \overline{c(G)})$.

Recognizing this we can choose some $\hat{g}$ such that all $P(g, f, \overline{c(g)})$'s for $g \leq \hat{g}$ can be evaluated using simulation, and substitute $P(g, f, \overline{c(g)}) = 1$ for all $g > \hat{g}$. Then from (**??**) we obtain an estimate of the upper bound on $\pi(G, k, \overline{c(G)})$. We note that in our numerical experiments such assumption made virtually no difference for large g; in effect, the probability of creating a partition is determined by the P-values for small g. We illustrate this issue on examples in the following subsections, where we discuss some interesting observations based on solving the recursion with simulated $P-$probabilities for different subclasses in certain equivalence classes. In particular, we discuss classes with only fixed attributes (both identical and not), as well as the classes with only variable attributes.

3.4.2 Probabilities for the equivalence classes with fixed attributes

Equivalence classes with only fixed attributes are fundamentally simpler to analyze, because at each state there could exist a balanced group of only one type with $min_j = max_j$ for all $j = 1, 2, ..., C$. Correspondingly, the problem of ordering groups is non-existent and transition network is a path. Therefore, for notational convenience we omit group type index, f, from $P(g, f, \overline{c(g)})$. By substituting it into (??) and rearranging we obtain:

$$\begin{aligned}\pi(G, k, \overline{c(G)}) &= P(G, \overline{c(G)}) - \\ &\sum_{i=1}^{G-2} \left\{ (1 - P(G-i, \overline{c(G-i)}) \times \prod_{l=0}^{i-1} \left(P(G-l, \overline{c(G-l)}) \right) \right\}, \quad (3.9)\end{aligned}$$

where $c_j(g) = c_j(G) - (G-g)min_j = c_j(G) - (G-g)max_j$, $j = 1, ..., C$.

Next we numerically estimate the probability that balanced partition exists for various equivalence classes. All computations are done in Mathematica on a 3Hz desktop PC. We first consider the case when all attributes are of the same type.

Probabilities for the classes with identical type fixed attributes

In this subsection we consider equivalence classes where in addition to being of the fixed types, all attributes are further assumed to be of the same type, m. That is every group should contain m objects that possess each of the C attributes. Our goal is to examine how $\pi(G, k, \overline{c(G)})$ is affected by the number of groups, G, the number of objects per group, k, density of attributes, m, and their number, C.

We first demonstrate that π's very quickly stabilize in the number of groups, since in effect P probabilities for large g do not influence the resulting π probability. This should not be surprising since as we argued, whenever for small g P-probability is close to 0, π is largely determined by it, and thus it does not matter what happens when g is large; conversely, when for small g P is not close to 0, then for large g it is close to 1.

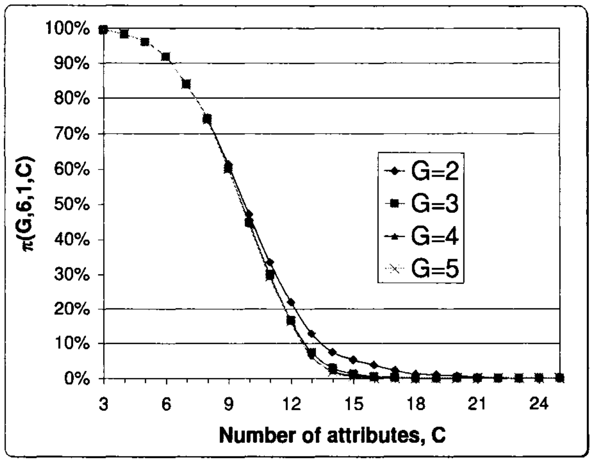


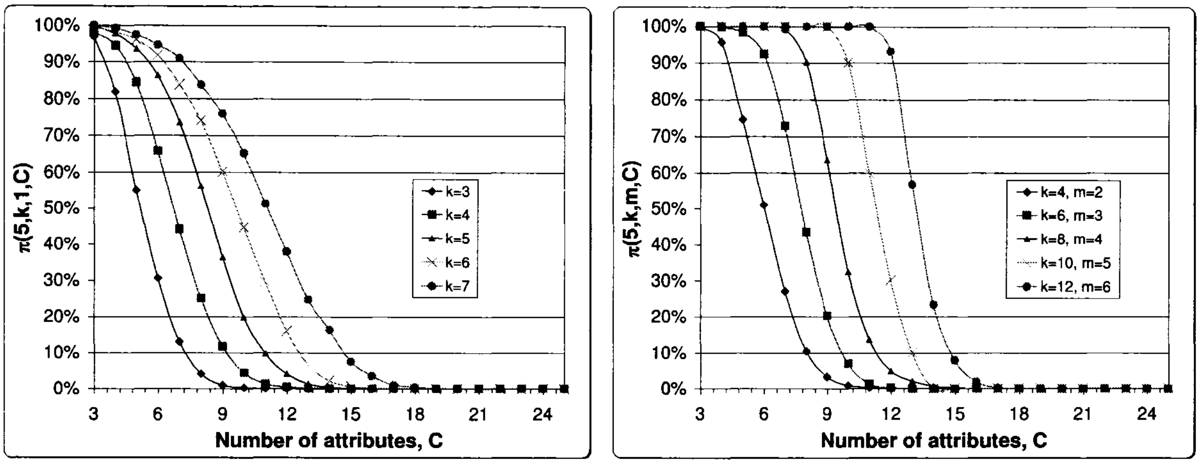
Figure 3.2: Demonstration that N-estimates for the probability that a balanced partition exists, $\pi(G, k, m, C)$, converge in G for all C. In (a) $m = 1$ we use simulation based P; $k = 6$ in both cases.

In full agrement with this intuition, Figure 3.2 suggest that for the groups of size 6, the corresponding π values are virtually identical for $G \geq 3$ for any number of attributes. We note that these observations hold for different k, m.

We next study how π depends on the size of the group, k, and relative attribute density, m/k. Figure 3.3 presents the corresponding probabilities for different k and fixed $m = 1$, figure (a), and for different m such that $m/k = const$, figure (b).

Figure 3.3 (a) shows that the probability that groups exist increases with group size (for fixed $m = 1$). We note this is rather different from the worst-case analysis, where the size of the group did not matter. At the same time, such result is supported by our other simulations, distribution-free and empirical bounds discussed below. There we see that probability decreases with an increases in m. Since an increase in the size of the group, k, is analogous to a decrease in the relative density k/m, we believe that these behaviors should be similar.

Three more observations are evident from Figure 3.3. First is that the increase in probability for different k could be substantial. For example, for $C = 11$ $\pi(k = 4) \approx 0$



(a) (b)

Figure 3.3: N-estimates for the probabilities that a balanced partition exists, $\pi(5, k, m, C)$, for different k and m. In (a) $m = 1$ and in (b) $m/k = 0.5$ - the largest possible m for a given k.

while $\pi(k = 7) \approx 0.5$. Second, the number of attributes that results in a certain probability that groups exist is approximately linear in group size when the density decreases, see (a). Further, it is also approximately linear when density is fixed, see (b). Such behavior suggests that what matters more to the existence of balanced partitions is not the number of objects per group that must possess a certain attribute, m, but rather the flexibility to include other objects in the group that do not possess this attribute. Such flexibility could be informally expressed by $k - m$, which increases in k in both cases (a) and (b) and thus the probability increases. Further, since the increase in $k - m$ is linear in k, the approximately linear relationship in one case, (a), leads to a similar relationship in the other case, (b). We conjecture that these relationships can be established analytically, but have not been able to do so.

Probabilities for the classes with non-identical type fixed attributes

Next we compute the probabilities to construct balanced groups for the classes with non-identical type fixed attributes and discuss the differences and similarities to the case with

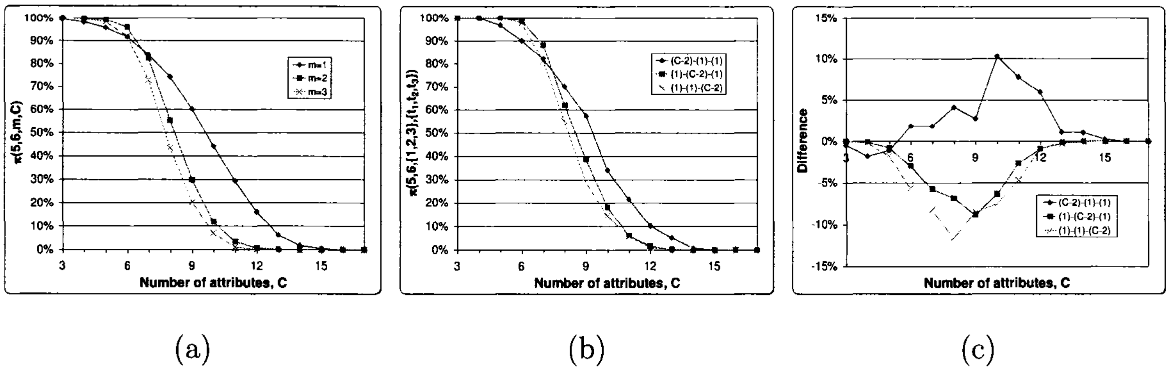
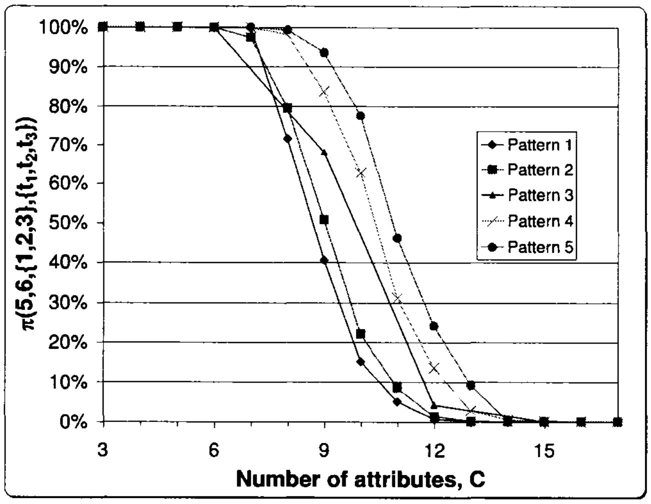


Figure 3.4: Comparing the N-estimates for the probabilities that balanced partitions exist for $k = 6, G = 5$ for the case with C identical attributes (a), and the case with different attributes, where $C - 2$ attributes are of the same type, plus one attribute of the remaining two types (b), as well as the difference in the corresponding probabilities (identical minus different), (c).

identical attributes. We set $k = 6$ and $G = 5$ in our experiments because the case with $k = 6$ is most relevant for our motivating example of creating MBA study groups, and same as in the case with identical attributes we observed that the probabilities quickly converge in G.

First consider Figure 3.4. Figure (a) presents the cases where all C attributes are of type $m = 1, 2, 3$ respectively (from Section 3.4.2). Observe that for the same C we observe that the probability that groups exist is lower for larger densities, m. Indeed, for the problem with 8 attributes, ($C = 8$) we have $\pi = 0.741$, $\pi = 0.551$ and $\pi = 0.434$ for $m = 1, 2, 3$ respectively. Similar probabilities for $C = 12$ are 0.161, 0.006 and 0.0009. This observation is quite intuitive: as more objects per group must possess certain properties, one must pick them with greater care since the same object must possess certain multiple attributes. This result contrasts with the worst-case bounds from Section 3.3.1, where only the divisibility of m matters, not its value.

In Figure 3.4 (b) we plotted a nearly identical case, where $C - 2$ attributes are of the same type as in (a), plus there is one attribute of each of the remaining two types. While the figures look seemingly alike, the difference between these probabilities (identical minus



	α_1	α_2	α_3
Pattern 1	1/31	4/31	16/31
Pattern 2	1/7	2/7	4/7
Pattern 3	1/3	1/3	1/3
Pattern 4	4/7	2/7	1/7
Pattern 5	16/31	4/31	1/31

Figure 3.5: N-estimates for the probabilities that balanced partitions exist, $k = 6, G = 5$ for $t_i = \alpha_i C$, where α_i values, $i = 1, 2, 3$ for each pattern (1 through 5) are given in the adjacent table.

non-identical) is significant; Figure 3.4 (c). For example, in $(C - 2) - (1) - (1)$ case, substituting two attributes with $m = 1$ with one attribute with $m = 2$ and one with $m = 3$ could decrease the probability of constructing balanced groups by approximately 10 percent Similarly, substituting 2 attributes with $m = 3$ by one with $m = 1$ and one with $m = 2$ could increase the probability by 10 percent. Note also, that keeping the same attribute density "on average", i.e. substituting two attributes with $m = 2$ with one with $m = 1$ and one with $m = 3$, (case $(1) - (C - 2) - (1)$), for some C could lead to an increase (about 10 percent) in the probability to construct balanced groups.

In our second experiment we test different compositions of the attributes, in particular those with many high density attributes and those with few. To do so, for a given total number of attributes, C, we design attribute patterns such that $t_i = \mathit{Round}(\alpha_i C)$, $\sum \alpha_i = 1, i = 1, 2, 3$, where function $\mathit{Round}(\cdot)$ rounds a number to the nearest integer (recall that t_i is the number of attributes of fixed type (i, i)). We test five patterns; see Figure 3.5. In pattern 1 (pattern 5) the number of attributes with $m = 3$ is four times larger (smaller) than that with $m = 2$, which in turn is four times larger (smaller) than that with $m = 1$. In pattern 2 (pattern 4) the same logic propagates but with twofold differences. Pattern 3 represents the case with equal number of attributes of each type. Observe that the density of attributes decreases in pattern number, most attributes in pattern 1 are with $m = 3$, while in pattern 5, with $m = 1$.

For these patterns Figure 3.5 supports the initial observation that having attributes with small densities is less restrictive than having those with large densities. Furthermore, the difference could be very substantial: for example for $C = 9$, instance with pattern 1 has a 40 percent probability of the existence of a balance partition, while the instance with pattern 5 has 95 percent probability. This is because in the instance with pattern 1 most attributes have high density ($m = 3$), while in pattern 5 most attributes have low density ($m = 1$).

Probabilities for the classes with variable attribute types

In this subsection we use transition networks to compute probability estimates for equivalence classes with variable attribute types, i.e. the classes when each group could contain min_j or $max_j == min_j + 1$ objects with each attribute $j = 1, 2, ..., C$. We use our random network approach and estimate P-probabilities for $g \leq 4$ using simulations and otherwise we assume they equal one. We present $\pi-$probabilities for the case with $G = 5$, $k = 6$ in Figure 3.6. In this simulation we generate 100 subclasses containing C attributes of the same class $(i, i+1)$, $i = 0, 1, 2$, by sampling column sums, c_j, uniformly at random from $[iG + 1, (i+1)G - 1]$.

Our main observation is that a balanced partition in an instance with variable attribute types are much more likely to exist than in an instance with fixed. For example, comparing Figures 3.6 and 3.4 (a) it is easy to see that for the same k, G in the equivalence classes with fixed types balanced partition is unlikely to exist for $C >\approx 15$, while in comparable instances in the classes with variable types groups nearly always exist.

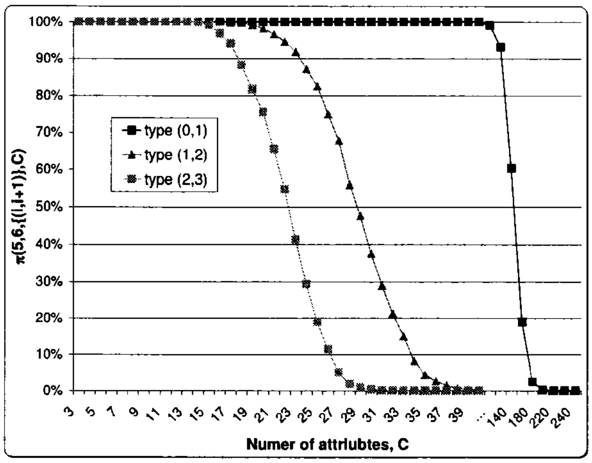


Figure 3.6: N-estimate for the probability that a balanced partition exists in the equivalence classes with attributes of variable types ($k = 6, G = 5$ all C attributes are of the same type with c_j sampled uniformly at random from the corresponding range).

Another observation is that the differences in probabilities for different types of attributes are magnified to a greater extent. Indeed, while from 3.4 (a) instances in the class with attributes of type $(1, 1)$ are only somewhat more likely to be partitioned than instances with types $(2, 2)$ or $(3, 3)$, while from 3.6 the corresponding differences are very large. In particular, the class should contain in excess of a hundred of attributes of type (0,1) for the $\pi-$probability to start to decrease; this decrease is also very slow (note the scale for large C).

In summary, we suggested a network-based recursive approach to estimate the probability that a balanced partition can be constructed in an arbitrary instance in a given equivalence class and its subclass. We further demonstrated how this approach can be used in conjunction with simulation, and that a hybrid approach, where the transition probabilities for final stages are simulated, while the probabilities for earlier stages are assumed to be equal one, in practice leads to a rather tight (upper) bound. We note, however, that this estimate cannot be rigorously considered as an upper bound because it is based on estimating $P(g, f, \overline{c(g)})$'s using simulation. Therefore next we present an approach that even though results in a less tight estimate, is a rigorous upper bound.

3.4.3 Distribution-free bound

Our key observation in this subsection is that a partition obviously cannot exist in an instance with fewer than G uncovered group-vectors. Therefore, the probability that there are fewer than G uncovered group-vectors is a lower bound on the probability that a partition does not exist, and hence its complement is an upper bound for $\pi(G, k, \bar{c}_{\mathbf{E}})$. Below we discuss how to estimate such bound analytically.

Observe that Assumption 2 (sampling with replacement), implies that in an arbitrary instance, the probability that a column-vector covers an arbitrary group-vector is independent of whether this group-vector has already been covered by other attribute vectors. As a result, we can easily determine the probability that an arbitrary group-vector is covered as well as the expected number of covered group-vectors. Note, however, that the probability that an arbitrary group is covered is not independent of whether other group vectors have already been covered by a given attribute vector. Therefore, the distribution of the number of covered group vectors is not trivial to determine.

Let $Y_{\bar{c}_{\mathbf{E}}}$ be the random variable representing the number of group-vectors covered by the attribute vectors in an arbitrary instance in sub-class $\bar{c}_{\mathbf{E}}$ in equivalence class $\mathbf{E}(G, k, \mathbf{T}, \{t_{(h,\bar{h})}\})$.

We first evaluate the distribution of $Y_{\bar{c}_{\mathbf{E}}}$ through simulation. We note that this distribution is 'very uneven', i.e. only relatively few values occur with non-zero frequencies, moreover these values are interspersed with intervals of zero mass. For example, in the equivalence class with $k = 4, G = 2, C = 2, m = 1$ (both attributes have $c_j = 2$) the total number of group vectors is 70, 30 of which are covered by any single attribute column vector alone (we discuss this in detail next). Thus values from 30 to 70 could have non-zero mass, i.e. $Y \in [30, 70]$. However, through simulation we observed that $Y \in \{30, 46, 50\}$ and all

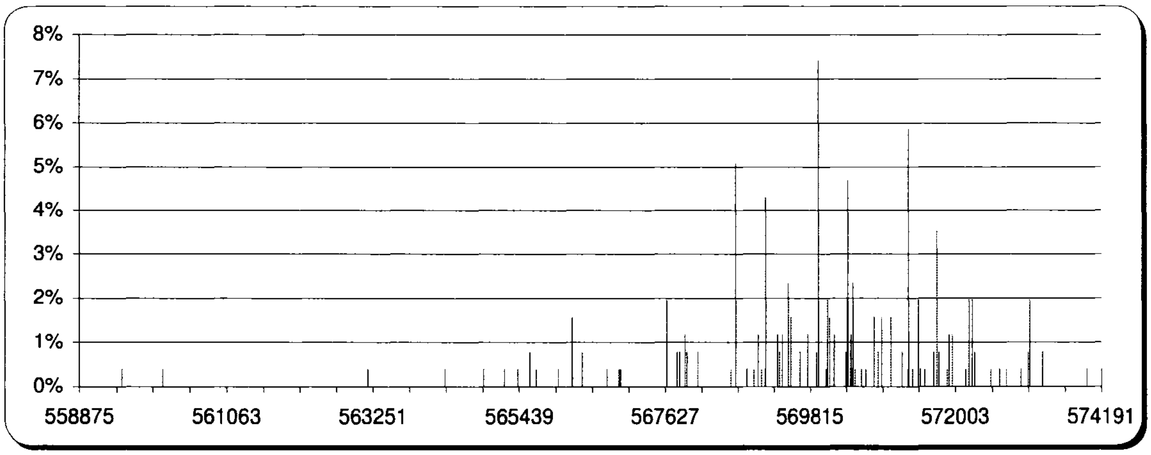


Figure 3.7: Distribution of the number of covered group vectors in the equivalence class with $k = 6, G = 5, m = 1, C = 4$ (all attributes are of the same fixed type). Total number of group vectors is 574191.

other values have zero frequencies. To confirm this observation we ran 10000 simulations for this equivalence class, but it did not change the result. Another example is presented on Figure 3.7 for the equivalence class with $k = 6, G = 5$ and $\overline{c_E} = \{5, 5, 5, 5\}$. By the same logic as we obtained 30 and 70 above it could be verified that in this equivalence class $Y \in [328125, 593775]$. However, again, there are only 88 values (out of 265650) that have non-zero frequencies, some of which are very likely (7.8% mass falls on 565600 group-vectors to be covered).

We believe that such unevenness is caused by the complicated combinatorial relationships between the column-vectors, which we have not been able to characterize in a general case. As a result the distribution of $Y_{\bar{c}_E}$ cannot be estimated by a standard parametric distribution family, because any such family would assign a positive probability on many $Y-$values for which the actual probability is zero due to the unevenness. Therefore we develop a distribution-free bound.

Observe that we can always renumber objects such that any given attribute c_j-vector becomes a vector with first c_j components equal one and the remaining components equal zero. Therefore, for a given c_j, each attribute vector covers the same number of group vectors.

Let $H(G, k, c_j)$ be the number of group $k-$vectors (from set $\mathbf{V}$) covered by the attribute c_j-vector in an instance with G groups. To compute H, consider an attribute vector $(1, 1, ...1, 0, 0, ...0)$ with ones in the first c_j components and zeros otherwise. By counting the number of group-vectors that contain 0, 1, 2, ... $min_j - 1$, $max_j + 1$, ...k elements from $1, 2, ...c_j$ we obtain:

$$H(G, k, c_j) = \sum_{\substack{i=0 \\ i \neq min_j \\ i \neq max_j}}^{\min[k, c_j]} \binom{c_j}{i} \binom{kG - c_j}{k - i}, \tag{3.10}$$

and therefore the probability that an arbitrary c_j-vector covers an arbitrary group vector equals:

$$P_1(G, k, c_j) = \frac{H(G, k, c_j)}{K}. \tag{3.11}$$

where for notational convenience we let $K = \binom{kG}{k}$.

Thus for $C = 1$, i.e. when $\bar{c}_{\mathbf{E}} = \{\tilde{c}\}$ for some $\tilde{c}$, $Y_{\bar{c}_{\mathbf{E}}} = H(G, k, \tilde{c})$ and is not random. However, for $C \geq 2$, $Y_{\bar{c}_{\mathbf{E}}}$ could attain different values, and as we argued its distribution is hard to find analytically. At the same time, $E[Y_{\bar{c}_{\mathbf{E}}}]$ is easy to find as we show next.

Let $P_{\bar{c}_{\mathbf{E}}}$ denote the the probability that at least one attribute vector in an arbitrary instance in subclass $\bar{c}_{\mathbf{E}} = \{c(1), ...c(C)\}$ covers an arbitrary group vector in equivalence class $\mathbf{E}$. Since Assumption 2 implies that attribute vectors cover group vectors independently[‡]

$$P_{\bar{c}_{\mathbf{E}}} = \sum_{i=1}^{C} P_1(G, k, c(i)) \prod_{j=1}^{i-1} \left[1 - P_1(G, k, c(j))\right], \tag{3.12}$$

Let $y^i_{\bar{c}_{\mathbf{E}}}$ be the random variable representing whether group vector $i = 1, ..., K$ is covered

[‡] For notational convenience we assume $\prod_{j=1}^{0} f(j) = 1$.

in an arbitrary instance in $\bar{c}_{\mathbf{E}}$. Then its distribution is given by:

$$y^i_{\bar{c}_{\mathbf{E}}} = \begin{cases} 1, & \text{with probability } P_{\bar{c}_{\mathbf{E}}}; \\ 0, & \text{with probability } 1 - P_{\bar{c}_{\mathbf{E}}}. \end{cases} \tag{3.13}$$

With these, since the expectation of a sum of random variables (not necessarily independent) equals to the sum of expectations (e.g. see Proposition 2.4 (a) in Knight 2000), upon noting that from (3.11) $H = P_1 K$ we obtain

$$\begin{aligned} E[Y_{\bar{c}_{\mathbf{E}}}] &= E[\sum_{i=1}^{K} y^i_{\bar{c}_{\mathbf{E}}}] = \sum_{i=1}^{K} E[y^i_{\bar{c}_{\mathbf{E}}}] = P_{\bar{c}_{\mathbf{E}}} K \\ &= \sum_{i=1}^{C} H(G,k,c(i)) \prod_{j=1}^{i-1} [1 - P_1(G,k,c(j))] \end{aligned} \tag{3.14}$$

Observe that $K - Y_{\bar{c}_{\mathbf{E}}}$ represents the number of uncovered group vectors. If $K - Y_{\bar{c}_{\mathbf{E}}} \leq G - 1$ then a balanced partition clearly cannot exist. Therefore:

$$\pi(G,k,\bar{c}_{\mathbf{E}}) \leq 1 - Prob(K - Y_{\bar{c}_{\mathbf{E}}} \leq G - 1) \tag{3.15}$$

Note that the reverse is not necessarily true, since not every set of G group vectors forms a valid partition.

Next we provide a distribution-free bound for the probability that $K - Y_{\bar{c}_{\mathbf{E}}} \leq G - 1$.

Theorem 3.8 *In an arbitrary instance in sub-class $\bar{c}_{\mathbf{E}}$ in equivalence class $\boldsymbol{E}(G, k, \boldsymbol{T}, \{t_{(h,\bar{h})}\})$*

$$Prob(K - Y_{\bar{c}_{\mathbf{E}}} \leq G - 1) \geq \frac{E[Y_{\bar{c}_{\mathbf{E}}}] + G - K}{G}. \tag{3.16}$$

Proof. By the definition of the expected value:

$$\begin{aligned} E[Y_{\bar{c}_{\mathbf{E}}}] &= \sum_{y=0}^{K-G} Prob(Y_{\bar{c}_{\mathbf{E}}} = y)y + \sum_{y=K-G+1}^{K} Prob(Y_{\bar{c}_{\mathbf{E}}} = y)y \\ &\leq (K-G) \sum_{y=0}^{K-G} Prob(Y_{\bar{c}_{\mathbf{E}}} = y) + K \sum_{y=K-G+1}^{K} Prob(Y_{\bar{c}_{\mathbf{E}}} = y) \\ &= K - G \sum_{y=0}^{K-G} Prob(Y_{\bar{c}_{\mathbf{E}}} = y) = K - G(1 - Prob(Y_{\bar{c}_{\mathbf{E}}} \geq K - G + 1)). \end{aligned}$$

Equivalence class	95%	50 %	5 %
Identical fixed with $m = 1$, type (1,1)	14	16	19
Identical fixed with $m = 2$, type (2,2)	11	13	17
Identical fixed with $m = 3$, type (3,3)	11	12	14
Equal number of fixed with $m = 1, 2$, types (1,1) and (2,2)	12	14	16
Equal number of fixed with $m = 1, 3$, types (1,1) and (3,3)	12	14	16
Equal number of fixed with $m = 2, 3$, types (2,2) and (3,3)	12	14	14*
Equal number of fixed with $m = 1, 2, 3$, types (1,1), (2,2) and (3,3)	15	15*	18
Same number of variable with $c_j = 3$, type (0,1)	120	126	150
Same number of variable with $c_j = G + 3$, type (1,2)	33	35	42
Same number of variable with $c_j = 2G + 3$, type (2,3)	27	28	33

Table 3.10: Critical numbers, $\overline{c_j}$, for probability thresholds of 95, 50 and 5 percent for different equivalence classes with $k = 6, G = 5$. Values with asterisk represent the cases where the probability instantaneously drops below the next threshold thus not allowing us to differentiate between the two adjacent thresholds.

The claim follows by rearranging the terms in the inequality above. ■

Thus, when $K - E[Y_{\bar{c}_E}] \leq G - 1$, we can establish an upper bound on the probability that a balanced partition exists. For example:

Corollary 3.2 *If $K - E[Y_{\bar{c}_E}] \leq G/2$ then $\pi(G, k, \bar{c}_E) \leq 50\%$.*

In particular, since from (3.14) the expected number of covered group vectors increases in the total number of attributes, C, (hence the number of uncovered group vectors decreases), for attributes with a given c_j and for any probability threshold (e.g. 50%), there exists a 'critical number' of such attributes (in addition to the other attributes contained in $\bar{c}_E$), $\overline{c_j}$, beyond which the probability that a partition exists is less than this threshold.

Table 3.10 presents such critical numbers for probability thresholds of 95, 50 and 5 percent for different equivalence classes. Several observations can be made. First, for equivalence classes with fixed attribute types, substituting some attributes with the same number of attributes with larger m decreases the probability of the existence of partition (compare the cases with $m = 1$ with the class where half have $m = 1$ and half have $m = 2$). Second, the number of attributes of variable types needed to reduce the probability is much larger than that of fixed types. Such behavior fully agrees with simulation-based approach.

We note that the distribution-free bound is less tight than the simulation-based estimate. At the same time, it provides a guaranteed upper bound. Next we discuss a lower bound.

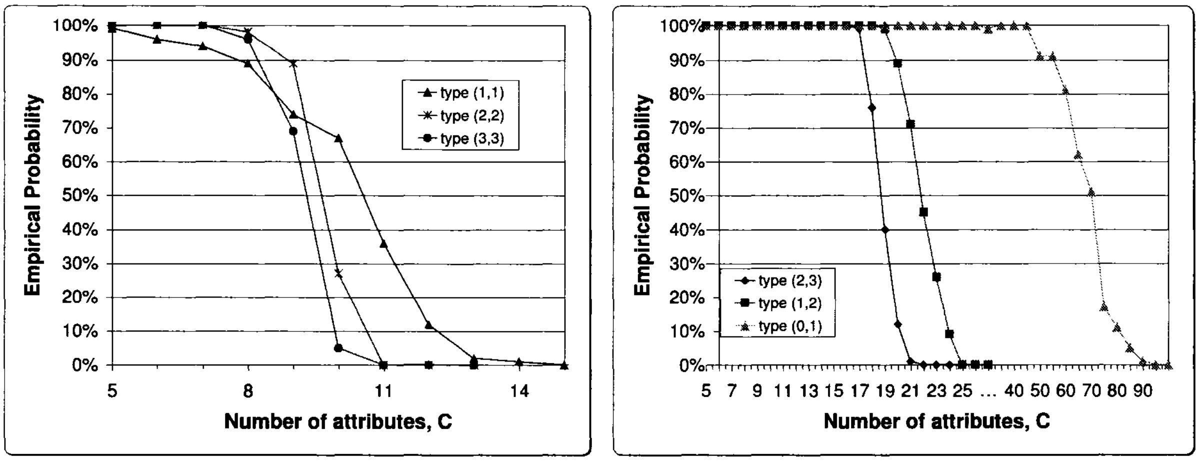
3.4.4 Empirical Lower Bound

Observe that a probability that balanced partition exists in a given equivalence class can be estimated as a fraction of randomly generated instances for which feasible solutions were found for the BMASP integer programs. Such a fraction obviously is an estimate for the lower bound on the corresponding $\pi-$probability (this is a lower bound because the IP solver may fail to find a balanced partition even when one exists). Therefore the lower boundary of the confidence interval for this fraction is a probabilistic lower bound.

Let n be the number of instances generated and let $r \in [0, 1]$ be the fraction of these n instances in which a balanced partition has been found. Let z be the required $1 - \alpha/2$ percentile point of the standard Normal distribution (i.e., for the 99% confidence interval for the lower boundary, $\alpha = 0.02$ and so $z = 2.3263$). We use Wilson's "score" method using asymptotic variance with no continuity correction. This method has been reported to perform the best on the instances where little is known about the random process causing r (Newcombe 1998), as is in our case. The lower and upper limits of the confidence interval are given by:

$$\frac{2nr + z^2 \pm \sqrt{z^2 + 4nr(1-r)}}{2(n + z^2)}.$$

In order for such an empirical approach to be effective, 'sufficiently many' instances have to be generated for every equivalence class of interest. Also, because integer programs are NP-hard, in practice it is only possible to find out whether a feasible solution has been



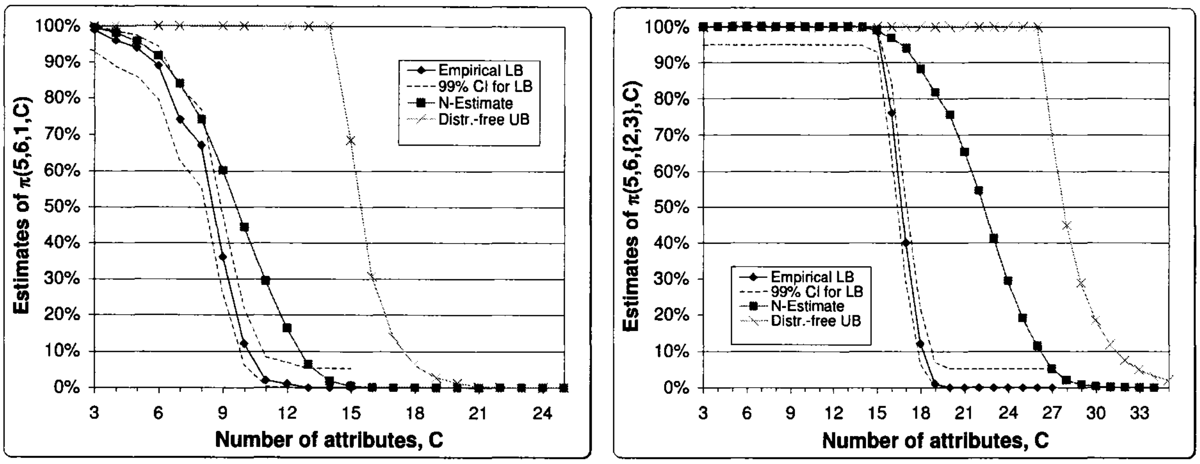
(a) (b)

Figure 3.8: 99% lower limit of the confidence interval for the empirical probability: the fraction of instances for which a balanced partition was found by solving BMASP IP (a) for the equivalence classes with fixed attributes, (b) for the equivalence classes with variables attributes.

found within a pre-specified time limit. Thus the time limit has to be selected such that it is rarely reached in the cases when other estimates suggested that groups should exist. In our experiments for every equivalence class that we studied we generated $n = 100$ random BMASP instances , and set time limit to 900 seconds (in our experiments such time limit satisfies the abovementioned property). We use Mathematica to generate random attribute matrices with given column sums, and we use AMPL CPLEX to solve integer programs; all computations are done on a 3GHz desktop PC.

We present lower boundaries for the 99% confidence intervals for such empirical fractions in Figures 3.8 (a) and (b) for the cases with fixed and variable attribute types respectively. The general pattern is the same as in the DP estimates: for small C there is a 'plateau' for which probability is effectively 100%, then probability sharply drops and from some C onwards it is basically zero.

We finally compare all three approaches we discuss in this section: DP simulation-based probability estimates, distribution-free upper bound and the empirical lower bound;



(a) (b)

Figure 3.9: Comparison between the empirical lower bound and its 98% double-sided confidence intervals, N-estimate and distribution-free upper bound for the equivalence classes with $G = 5, k = 6$ and C attributes. In (a), for the fixed type $(1,1)$, in (b), for the variable type $(2,3)$.

see Figure 3.9.

For fixed type, (a), three observations are evident. First, the gap between lower and upper bounds is large, and since DP estimate is closer to the lower bound, this is likely because the upper bound is not very tight. This is expected, since in the situation where the exact distribution of number of covered groups cannot be found, in attempt to to construct a rigorous upper bound we used the worst-case distribution, which may be quite different from the actual.

Second, for small and medium C, the DP estimate is a rather tight upper bound for the empirical fraction - it is within the condifence intervals for the empirical bound. This suggests that our hybrid simulation-based approach indeed produces a good estimate. Thus, in practice, it can be used instead of a rigorous, yet very conservative distribution-free upper bound.

Finally, for larger C that still result in non-zero probabilities, DP estimate is not close to the empirical fraction and outside of the confidence interval. We believe that these differences could be attributed to two factors. The DP estimate suggests that as C increases the probability decreases, and thus in a given instance it becomes harder to find a feasible partition. At the same time, when C grows, the BMASP integer program becomes larger and harder to solve (within a fixed time-limit). Thus, feasible partitions are less frequently found, while they could in fact exist.

This logic is even more evident in Figure 3.9 (b), where we compare these three estimates for classes with variable types. In particular, since for variable types the equivalence class must include more attributes in order for the probability that a partition exist to start to decrease, integer program becomes even larger and hence even harder, and so the difference between DP bound and empirical bound is larger than in the case with fixed types.

In other words, these differences highlight the limitations of the empirical approach to estimating probabilities that partitions exist and proves the that our analytical approaches are worth taking. In particular, from the standpoint of solving group balancing problems in practice using BMASP formulation, if DP estimate suggests high probability of the existence of partition, but it cannot be found, then perhaps one should try increasing time limit or using specialized constraint programming software or algorithms.

3.4.5 Analysis of Rotman Classes of 2004 and 2005

This work was motivated by a successful implementation of a BMASP-based software package, Advisor, to the problem of creating MBA study groups at the Rotman School of Management, University of Toronto. In particular, Advisor was able to find balanced groups in all instances solved to date. Hence, our question was, to what extent can we expect that balanced groups can be found in the instances similar to those at Rotman. Now, when we developed our methodologies we can finally answer it.

In the Rotman problem, classes of 260-280 students must be partitioned into 40-50

	$t_{(0,1)}$	$t_{(1,2)}$	$t_{(2,3)}$	C	G	k
Class of 2004	6+1	6	1	13+1	46	6
Class of 2004	5+1	7	4	16+1	49	6

Table 3.11: Equivalence classes for the Rotman examples of classes of 2004 and 2005.

groups. In addition, Rotman school typically first splits the entire class into 4 sections of 65-70 students and then partitions each section into 10-13 groups. The equivalence classes corresponding to the Rotman classes of 2004 and 2005 are described in Table 3.11. We note that in the actual application $k \in \{5, 6\}$, and therefore from Theorem 3.1 we add one attribute of type $(0, 1)$ to convert these to the classes with $k = 6$.

In the class of 2004, from the worst-case perspective, the amount of each individual attribute is too small to prohibit the existence of balanced groups. From Table 3.8 (at least for $G = 2$) there have to be at least 7 or more attributes of types $(1, 2)$ or $(2, 3)$. At the same time, the combination of different attributes can result in an instance where a balanced partition does not exist. However, the empirical probability for $G = 5, ...40$ is 100%, see Table 3.12; DP probability estimate and distribution free UB are also 100% for all G (recall that to compute DP estimate we set transition probabilities to 1 from $G \geq 5$, thus increasing the number of groups beyond 5 does not affect this estimate). This is not surprising, from Figure 3.6 an instance must contain two/three times more attributes in order to have a probability of the existence of a partition different from 100%.

In the class of 2005, there are 7 attributes of type $(1, 2)$, which could result in an instance in which a partition does not exist (Table 3.8). As a result, empirical probabilities are lower; see Table 3.12. For $G = 5, 10$ empirical probabilities show that balanced partitions are always found within the 15 minutes time limit, however, for $G = 20, 40$ the time limit was frequently reached, therefore we believe that the actual probability is larger. DP estimate and distribution-free UB are also 100% for all G for the Rotman class of 2005, since, still the number of attributes is about twice as small as needed to see the probabilities decrease.

	Class of 2004	Class of 2005
$G = 5$	100 %	100 %
$G = 10$	100 %	100 %
$G = 20$	100 %	72 %
$G = 40$	100 %	20 %

Table 3.12: Empirical probabilities for the Rotman classes of 2004 and 2005.

3.5 Conclusions

This theoretical work complements our practical work in constructing MBA study groups and creating Advisor group balancing software at the Rotman School of Management. In particular, we seek to what extent the success of the constraint-based Advisor in creating perfectly balanced groups can be attributed to a pure luck, and to what to the properties and relationships internal to a general constraint multiple attribute program.

We generalize the model underlying the Advisor to a general constraint problem, BMASP, such that any feasible solution to it describes the set of balanced groups, and study the worst case and probabilistic aspects of its performance.

We find that the worst-case could be indeed quite bad. As little as three attributes could lead to an instance in which a balanced partition does not exist; even further, even one balanced group may not exist in many cases, see Theorem 3.3. Such cases arise in the instances with the attributes of fixed type (i.e. with respect to which each group should contain the same number of objects with such attributes). This very restrictive conditions propagates to the the instances with the attributes of both fixed and variable types (where the number of objects with a given attribute per group may vary ±1) as well, see Theorem 3.6. However, in such cases the combination of attributes becomes important, see Theorems 3.4 and 3.5. All these results are obtained using the block-matrix approach, where we suggest a framework for constructing the instances that cannot be partitioned into balanced groups.

We also suggest an alternative approach to worst-case analysis, based on cover problem representation. We use it to demonstrate that, as intuition suggests, BMASP problems with variable attributes are less constrained than those with fixed; that is on many occasions when in the fixed case the partition may not exist, it always exists in the variable case. We prove this claim analytically for specific attribute types, see Theorem 3.7 and otherwise demonstrate it numerically.

The cover problem representation also plays a central role in our probabilistic analysis. We suggest three approaches: (i) a recursion that can be used to estimate the probability that a balanced partition exists; (ii) an analytical distribution-free upper bound, and (iii) an empirical lower bound based on solving BMAPS integer programs.

We study these approaches on different variations of BMASP problem with fixed and variable types, and make several major observations. First, and quite intuitively, as the number of attributes grows, the probability declines. Further, the problem becomes more constrained as the relative density of attributes increases (that is, more objects per group must possess given attribute) and hence the probability that groups exists declines. Finally, the cases with variable attributes are indeed much less constrained than those with fixed, typically if for some number of attributes in the case with fixed types, the probability is zero, then for the same number with variable types it is still one.

Lastly, we study the Rotman problem instances of the classes of 2004 and 2005. Most importantly, our results suggest that the probability that balanced groups exist in the instances similar to those observed at Rotman is effectively equal one. Further, substantially more attributes (of variable types) could be added without decreasing this probability. However, in terms of finding such partitions in practice, Rotman instances are quite close to the 'limit' for which the groups can be found by a 'brute-force' computation used in the Advisor, which generates and solves BMASP integer program. This suggests that there could exist more elaborate specialized algorithms that could search for the balanced multiple-attribute partitions. Creating such algorithms is of interest for future research.

Bibliography

[1] Bertsekas, Dimitri. 1987. Dynamic Programming. Prentice-Hall, Inc. New Jersey, USA.

[2] Knight, Keith. 2000. Mathematical Statistics. Chapman & Hall/CRC Boca Raton, USA.

[3] Newcombe, Robert G. 1998. Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. Statitics in Medicine Vol. 17() pp. 857-872.

(a)

1	1	1	1	1
1	1	1	1	1

(b)

0	1	1	1	1
1	**0**	1	1	1
1	**1**	0	0	0

(c)

0	1	1	1	1
1	**0**	**0**	1	1
1	**1**	0	0	0
0	0	**1**	0	0

(d)

0	1	1	1	1
1	**0**	**0**	**0**	1
1	**1**	0	0	0
0	0	**1**	0	0
0	0	0	**1**	0

Figure 3.10: Constructing indivisible block matrices D for $C = 5$ and $j = 2$. (a) - initialization step for $q = 2$; (b) - block $(5, 3, 2)$; (c) - block $(5, 4, 2)$ and (d) - block $(5, 5, 2)$.

Appendix

Building indivisible blocks

Next we study indivisible blocks. Note that Definition 1 implies $q \leq p$. Further, if $p = q$ then the block matrix has all entries equal 1. Thus such a block is divisible, since it consists of p blocks $(C, 1, 1)$. For $q \leq p - 1$ we obtain the following result.

Proposition 3.1 *For every $C, p \geq 1$, $p \leq C$ there exist indivisible blocks (C, p, j) for all $j = 1, 2, ...p - 1$.*

Proof. Consider arbitrary C and $j \leq p - 1 \leq C - 1$. If $j = 1$ then the proof is trivial: consider $(p \times C)$ matrix D where the only non-zero entries are $d_{i,i} = 1$ for $i = 1, 2, ...p$ and $d_{pj} = 1$ for $j = p + 1, ...C$. By definition it is a block, which is indivisible since any subset of rows of cardinality less than p contains an all-zero column.

For $j \geq 2$ we provide an algorithm for constructing indivisible blocks (C, p, j) for $p = j + 1, ...C$. The algorithm is illustrated in Figure 3.10.

To initialize the algorithm, create j row vectors $(1 \times C)$ with norm C; i.e. let $d_{st} = 1$ for $t = 1, 2, ...C$, $s = 1, 2, ...j$, (see (a) in Figure 3.10).

To create block (C, p, j) for $p = j + 1$ let $d_{j+1,i} = 1$, $d_{j+1,x} = 0$ otherwise, and let $d_{ii} = 0$ for $i = 1, 2, ...j$. Intuitively this procedure "drags and drops" i^{th} "1" from row i to the i^{th}

component of $j+1^{st}$ row (see (b) in Figure 3.10; corresponding entries are highlighted in bold). The resulting $((j+1) \times C)$ matrix D is by definition a (C, p, j) block for $p = j+1$.

To create block (C, p, j) for $p = j+2, ...C$ take block $(C, j+1, j)$ (created above) and let $d_{i,i-1} = 1$, $d_{ix} = 0$ otherwise and let $d_{j,i-1} = 0$ for $i = j+2, ...p$. Intuitively this procedure "drags and drops" i^{th} "1" form row j into a new unit norm row with "1" in i^{th} component (see (c) and (d) in Figure 3.10 for $p = 4, 5$ respectively). The resulting matrix by construction is a (C, p, j) block.

To observe that blocks created this way are indivisible, first consider the case with $p = j+1$. Let r_i, $i = 1, 2, ...j+1$ be the rows in D. Let D' be the subset of rows that form a sub-block. By definition there exists another sub-block, $D'' \neq \emptyset$, such that $D' \cap D'' = \emptyset$ and $D' \cup D'' = D = \{1, 2, ...j+1\}$. Suppose $r_{j+1} \in D'$; note this can be done without loss of generality. Then by construction for all row indices, i, such that $r_i \in D''$, $r_i(i) = 0$, while $r_i(i') = 1$ for all $i' \geq j+1$. Thus columns $j+1, ...C$ of D'' have norms $|D''|$, while columns $1, ...j$ have norms $\leq |D''| - 1$. Hence D'' is not a block, which is a contradiction.

Now consider the case with $p = j+2, ...C$. Same as above consider sub-blocks D' and D'', where $r_{j+1} \in D'$. Since by construction $d_{iq} = 0$ for $q = 1, ...j$, but $d_{i,i-1} = 1$ for $i = j+2, ...p$, any subset of rows $j+2, ...p$ cannot form a sub-block. Therefore D'' contains $1 \leq Q \leq j$ rows from $1, ...j$. Then columns $1, ...j$ of D'' have norms $\leq Q-1$, but column C has norm Q, which is a contradiction. Therefore (C, p, j) is indivisible. ■