Stochastic Models for Service Operations and Supply Chains

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL OF THE

UNIVERSITY OF MINNESOTA

BY

Yu (Rowan) Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Saif Benjaafar

JUNE, 2014

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor Professor Saif Benjaafar at the University of Minnesota, for his mentoring, guidance, support, and encouragement throughout my PhD studies. He has demonstrated to me the attitude and substance of a world-class scholar. His enthusiasm and patience have continually and persuasively conveyed to me the spirit of adventure in regard to scientific research. Without his excellent supervision, persistent help, stimulating suggestions, and tireless feedback, this dissertation would not have been possible. On a personal level, I enjoyed a lot the close friendship between us as well as our families. I would never forget the delicious backyard grill in Minnehaha and cruise dinner in Suzhou.

I am very grateful to Professor Shuzhong Zhang, for not only the research guidance but also the career advice. Being Chinese, he has showed me the rigorous scholarship and enormous charisma to be a master, which inspired me tremendously. I would also like to express my appreciation to the rest of my committee members, Professor William L. Cooper and Professor Karen Donohue, who have taken time to comment on this dissertation and given helpful suggestions over the years. I also want to thank them for letting my defense be an enjoyable

moment.

Sincere appreciation should go to Professor Oualid Jouini at Ecole Centrale Paris, France, for introducing me to the interesting ideas of finite arrival queueing systems and collaborating with me on my first ever paper. I have benefited a lot from his erudition and seriousness. I enjoyed the time we work together in the windowless basement lab at UMN, the non-air-conditioned smoking room at ECP, as well as through the blurred and intermittent phone calls. Special thanks must also go to my fellow student and coauthor David Chen, for every single discussion from "implicit function theorem" to "lost sales with lead time". I am used to sharing with him whenever I have new ideas. It is my best memory to be his matchmaker and the best man at his wedding.

I am grateful to Professor Shaun Kennedy, Dr. Morgan Hennessey, and Col. John Hoffman, for their meaningful suggestions and guidance during my time with NCFPD, and to Dr. Robert G. Haight at USDA Forest Service for extensive discussion and comments on my work.

I was very lucky to meet two senior students Dr. Yimin Yu and Dr. Le Li, when I came to Minnesota. I would not have learned dynamic programming without Laoyu, and this thesis would have been written in Word instead of Latex without Pangpang. They are the ones who helped me out during the hardest time in my life, and they have made our cold office a warm home. Many thanks must be given to my best friends and collaborators Dr. Yinghao Zhang and Yibin Chen, for our discussions on research topics and more importantly, the great times we had together. In addition, I appreciate the help from other ISYE, and in general UMN faculty, staff, as well as fellow students.

I would never get to my PhD studies without the grueling but fruitful three years of undergraduate study at the Hong Kong University of Science and Technology. I would like to express my sincere gratitude to Professor Min Yan, the first professor I met in college, who spent countless hours explaining to me mathematical analysis. He is the best professor I have ever met and my fine example of an educator. I would never have started my academic journey without his inspiration and encouragement. Special appreciation should also go to Professor Liming Liu, for introducing me to the glamorous area of operations research, and to Professor Rachel Q. Zhang for recommending me to my advisor Professor Saif Benjaafar.

Throughout my PhD studies, I have visited several universities and got to know many brilliant faculty members who have provided me generous suggestions on research, teaching, and career developments, in particular Professor Niyazi Taneri and Professor Lingjie Duan at the Singapore University of Technology and Design, and Professor Yanzhi Li at City University of Hong Kong. I indeed appreciate their help. I would also like to share this accomplishment with my colleagues in the academic world, in particular Dr. Qizheng Yin, Dr. Adel Omri, Dr. Chenren Xu, Dr. Li Chen, Lingtian Kong, Xiaobo Li, and Xiaowen Yu, for our hard work towards the faculty dream; and with my buddies in Minnesota, Singapore, Hong Kong, Shanghai, Hubei, and around the world, for our friendship which makes me never walk alone.

Last but not least, words cannot express my gratitude to my parents, Dongchuan Wang and Jing Xiao, for bringing me to the world in the first place, educating me well from childhood, and supporting me unregretfully throughout

my life. I am always proud to be their son. I am also glad and lucky to be in the warmest big family, having support and encouragement from grandparents and all the relatives whenever needed. At the end, I would like to extend my sincerest appreciation to my sweetheart Yichen Liu, for her love and company. I am willing to share with her my achievement, happiness, and the rest of my life.

Dedication

To my mom who stopped her PhD studies to take care of me.

To my dad who made the most important decision in his life to move to Shanghai.

To my grandparents and relatives who have loved, supported, and encouraged me.

Abstract

This thesis consists of three essays in service operations and supply chains. The first essay is on managing stochastic inventory systems with scarce resources. We study an inventory system where a firm is subject to an *allowance* (a limit) on either the amount of input it can use or the amount of output it can produce over a specified *compliance* period. With such an allowance constraint, the quantity produced in one period affects the quantity that can be produced in future periods. We formulate the problem as a stochastic dynamic program with a two-dimensional state space. Using a novel extended state-space analysis, we reduce the problem into one that is single-dimensional and easier to analyze. We show the optimal policy for this modified version and then use it to characterize the structure of the optimal policy for the original problem. We also consider an extended version of the problem where the firm decides the allowance amount at the beginning of the compliance period. Throughout, we draw several managerial insights.

The second essay is on service systems with finite and heterogeneous customer arrivals. We analyze a queueing system where a finite number of customer arrivals occur over a period of time. Customer inter-arrival times and service times are heterogeneous. Using an embedded Markov chain approach, we analytically characterize various performance measures of interest, including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all customers. Through numerical

experiments, we examine the effect of heterogeneity in inter-arrival and service times. We derive managerial insights and discuss implications for settings where inter-arrival and service time features can be induced. We also validate the numerical results using a fluid approximation that yields closed form expressions.

The third essay is on service systems with appointment-driven arrivals, non-punctual customers, and no-shows. We consider settings where a finite number of customers arrive to a service system based on appointments. However, customers are not necessarily punctual and may also not show up altogether. Customers' punctuality, show-up probabilities, and the time between previous and subsequent appointments are all heterogeneous. We develop an exact analytical approach to obtain various performance measures related to customer waiting time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis consists of five chapters. Chapter 2 describes a completed work on inventory systems with scarce resources. This chapter is based on the paper Benjaafar et al. (2014). Chapter 3 presents another completed work on queueing systems with finite arrivals. This chapter is based on the paper Wang et al. (2014). Chapter 4 contains preliminary results from an ongoing work on queueing systems with appointment-driven arrivals. Chapter 5 provides conclusions and discusses future extensions. The following paragraphs give a brief synopsis of each of the Chapters 2-4.

In Chapter 2, we consider the problem of managing production in a production-inventory system where a firm is subject to an *allowance* (a limit) on either the amount of input it can use or the amount of output it can produce over a specified *compliance* period. Examples of such settings are numerous and include those where limits are placed on the use of scarce natural resources as input or on the amount of waste or harmful pollution generated by production. In each production period, the firm must decide, in the face of stochastic demand,

how much to produce knowing that the quantity produced in one period affects the quantity that can be produced in future periods. We formulate the problem as a stochastic dynamic program with a two-dimensional state space: on-hand inventory level and remaining allowance. To simplify the analysis, we consider an *extended state-space* version of the problem and show that this modified version of the problem reduces to a one-dimensional problem. We then show that the structure of the optimal policy for this modified version is the same as the one for the original problem and has similar properties. In particular, we show that the optimal production policy is specified by dynamic thresholds that depend on both the on-hand inventory level and the remaining allowance but only via the sum of these two quantities. We also consider an extended version of the problem where the firm decides (purchases) the allowance amount at the beginning of the compliance period. Throughout, we draw several managerial insights. In particular, we show that it is possible to significantly reduce the allowance amount without significantly increasing cost.

In Chapter 3, we consider service systems with a finite number of customer arrivals, where customer inter-arrival times and service times are both stochastic and heterogeneous. Applications of such systems are numerous and include systems where arrivals are driven by events or service completions in serial processes, and systems where servers are subject to learning or fatigue. Using an embedded Markov chain approach, we characterize the waiting time distribution for each customer, from which we obtain various performance measures of interest, including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all

customers. We carry out extensive numerical experiments to examine the effect of heterogeneity in inter-arrival and service times. In particular, we examine cases where inter-arrival and service times increase with each subsequent arrival or service completion, decrease, increase and then decrease, or decrease and then increase. We derive several managerial insights and discuss implications for settings where such features can be induced. We validate the numerical results using a fluid approximation that yields closed form expressions.

In Chapter 4, we consider service systems where a finite number of customers arrive over time. The arrival of customers is driven by appointments, with a scheduled appointment time associated with each customer. However, customers are not necessarily punctual and may arrive either earlier or later than their scheduled appointment times. Customers may also not show up altogether. The arrival times of customers (relative to their scheduled appointments) are stochastic. Customers are not homogeneous in their punctuality, show-up probabilities, and time between previous and subsequent appointments. We develop an exact analytical approach to obtain various performance measures of interest.

Chapter 2

Managing Stochastic Inventory Systems with Scarce Resources

2.1 Introduction

In this work, we consider the problem of managing production in a production-inventory system where a firm is subject to an *allowance* (a limit) on either the amount of input it can use or the amount of output it can produce over a specified *compliance* period. The setting we consider is one where the compliance period is substantially longer than a production period, so that each compliance period may consist of multiple production periods. In each production period, the firm must decide, in the face of stochastic demand, how much to produce knowing that the quantity produced in one period affects the quantity that can be produced in future periods (producing in one period consumes some of the available allowance and affects the allowance available in future

periods). In doing so, the firm must balance inventory holding and shortage costs while taking into account the allowance constraint. In contrast to a traditional production capacity constraint, imposed independently on each production period, an allowance constraint over a compliance period introduces capacity dependencies across periods. Hence in deciding on production quantities, the firm is also deciding on how the allowance is allocated over time. We also consider settings where the amount of allowance is a decision variable, determined by the firm at the beginning of the compliance period. In that case, the firm must trade off the increased production flexibility that more allowance buys with the associated higher initial investment cost.

We are motivated, in part, by settings where firms face limits on access to key input materials, which in turn limit how much the firms can produce. For example, logging companies in protected forest areas are subject to annual allowances on how much wood can be harvested and processed (Beaudoin et al. 2007, Ouhimmou et al. 2009). Similarly, fish and seafood processing facilities are constrained by annual fishing allowances in countries where overfishing is a concern (Grimm et al. 2012). In regions where there is concern about water shortage (e.g., regions where water tables have dropped significantly), industrial and agricultural facilities are subject to allowances on water usage (Dudley and Musgrave 1988, Rogers et al. 2013). In several countries, access to rare minerals and metals by mining and processing operations is restricted and exports are subject to quotas.

We are also motivated by settings where firms may face, instead of limits on their consumption of input, direct limits on the production of their output. This is the case, for example, when such output is associated with the generation of

waste or harmful pollution (e.g., Chinese government sets direct limits on the annual production output of several polluting industries). This may also be the case when the product is associated with undesirable health effects (e.g., some countries place limits on the production of alcohol and tobacco).

In these examples, the affected firms are typically provided with the right to use input, or produce output, up to a maximum allowance amount over a specified compliance period (e.g., one year). In some cases, the amount of allowance is provided to the firms for free (e.g., fishing rights that are grandfathered). In others, it is a decision variable with the amount of allowance purchased by the firms at the beginning of the compliance period (e.g., logging rights). The constraint imposed by the allowance amount can in some instances be relaxed if the firm exerts effort to improve the efficiency with which it uses its input or to reduce its output of harmful byproducts. However, the firm in all cases would still be left with a constraint, albeit one that is less strict.

The presence of an allowance constraint over a compliance period (or, equivalently, a capacity constraint over a planning horizon) raises several important questions. How does such a constraint affect production decisions over the compliance period? How are these decisions different from those for a system without such a constraint or with a capacity constraint that applies to each production period? How are decisions affected in each production period by the remaining allowance and the time until the end of the compliance period? Should the optimal policy turn out to be complicated, and, could simpler heuristics be effective? How does the presence of such a constraint affect expected optimal cost and how is this affected by various problem parameters, including the inventory

holding and shortage costs? In settings where reducing the allowance amount is desirable (e.g., when the input material is scarce or when there are negative externalities associated with production), are there settings where it is possible to reduce the allowance amount without significantly increasing the cost?

In this work, we address these and other related questions. We formulate the problem as a stochastic dynamic program with a two-dimensional state space: on-hand inventory level and remaining allowance. The two-dimensionality of the problem makes it difficult to analyze and to describe the structure of the optimal policy. To simplify the analysis, we consider an *extended state-space* version of the problem and show that this modified version of the problem reduces to a one-dimensional problem. We describe various properties of the optimal policy for the modified version of the problem and then show that these properties also hold for the optimal policy for the original problem. We then use these properties to characterize the structure of the optimal policy for the original problem. In particular, we show that the optimal production policy is specified by dynamic thresholds that depend on both the on-hand inventory level and the remaining allowance but only via the sum of these two quantities.

In addition, we characterize the impact of the allowance constraint and provide numerical results that examine the tradeoff between the expected optimal cost and the expected cumulative amount produced. We evaluate the performance of three plausible heuristic policies that are simpler to compute and implement. We also consider an extended version of the problem where the firm decides (purchases) the allowance amount at the beginning of the compliance period. We examine how the optimal allowance amount and the *allowance usage* (the ratio of the expected

cumulative amount produced over the entire compliance period to the amount of allowance purchased) is affected by the price of the allowance.

Some of our key findings are highlighted below:

- The expected optimal cost is convex with respect to the remaining allowance, which implies that cost becomes increasingly insensitive to the allowance amount as the allowance amount increases. This result suggests that it might be possible to impose an allowance constraint (or to tighten an existing constraint) without significantly increasing cost.

- There is a range of values on the allowance amount, for which the percentage reduction in the cumulative amount produced is higher than the percentage increase in cost.

- While there are regions under which simple heuristics perform well, there is also a range of parameter values under which the performance of the optimal policy is significantly superior.

- A small initial increase in the price per unit of allowance can lead to a significant decrease in the amount of allowance purchased. That is, putting a modest price on the natural resource or a modest penalty on the harmful externality can significantly reduce the corresponding usage.

- If we charge a modest price for each unit of allowance, then the usage of the allowance increases significantly. This implies that scarce resources would be used much more efficiently if accessing these resources is not for free.

2.2 Related Literature

Although constraints on production due to limits on either input or output are common in practice, the literature on this topic is relatively limited. There is of course extensive literature on stochastic inventory systems where a constraint on capacity is applied independently to each production period (see, e.g., Federgruen and Zipkin 1986a,b, Kapuscinski and Tayur 1998). In that case, and for settings similar to ours, the problem is much simpler, has a single dimension, and admits a simple optimal policy specified by a modified base-stock policy (in each period, it is optimal to produce and bring inventory level as close as possible to a target threshold (the base-stock level) without exceeding the capacity constraint). Several variations on the problem have been studied including for systems with fixed costs (Deng and Yano 2006, Zhang et al. 2012), multiple demand classes (Zhou et al. 2011), and multiple echelons (Glasserman and Tayur 1994, Parker and Kapuscinski 2004).

There is also an extensive related literature on supply chain contracts with quantity commitments (see, e.g., Bassok and Anupindi 1997, Anupindi and Bassok 1998, Urban 2000, Bassok and Anupindi 2008). Such contacts specify a minimum or a maximum amount a buyer commits to purchasing from a supplier. In most cases, these commitments are specified for each production period. An exception is Bassok and Anupindi (1997) who study a problem with minimum order quantity commitment over the entire planning horizon. However, considering minimum order quantities leads to a very different problem from ours, which can be viewed as one involving maximum order quantity commitments. Our setting is mentioned in Bassok and Anupindi

(1997) as a future research direction. There are settings where constraints arise because of budget requirements. For example, Chao et al. (2008) consider a system with a cash flow constraint, where the total production cost in each period cannot exceed the budget constraint of that period. The revenue from sales in that period, together with the available capital and savings interest, determines the budget constraint in the next period. Such a setting is different from ours in how production decisions in one period affect production decisions in subsequent periods. In our case, the more is produced in one period, the lower is the available allowance in future periods. In the setting considered by Chao et al. (2008), producing more in one period could possibly generate more revenue and, therefore, allow for more production in future periods.

Another related stream of literature considers settings where inventory replenishment takes place only at fixed intervals, with each interval being a multiple of the periods in which demand takes places (see, e.g., Graves 1996, Chen and Samroengraja 2000, Chao et al. 2009). In other words, demand occurs in every period but ordering or production can take place only once every k periods, for some positive integer k (in a multi-echelon system, this may arise because deliveries from one echelon to the next occur only periodically). As a consequence, what is ordered or produced at the beginning of an interval becomes a constraint on how much demand can be fulfilled in the periods within that interval. This is different from our setting where, in addition to deciding on how much capacity to acquire at the beginning of an interval (the compliance period), we also decide on how much to produce in each production period. The fact that we also consider lost sales, in contrast to much of the existing literature which

treats backorders, makes our problem considerably more difficult to analyze.

Finally, there is related literature from economics that considers the impact of production input and output limits (see, e.g., Weitzman 1974, Baron and Myerson 1982, Cropper and Oates 1992). However, that literature relies on aggregate models of demand and supply and does not model operational decisions as we do in this work.

The rest of this chapter is organized as follows. In Section 2.3, we describe and formulate the problem. In Section 2.4, we analyze the modified version of the problem using our extended state-space approach. In Section 2.5, we describe the structure of the optimal policy for the original problem and provide some managerial insights. In Section 2.6, we discuss heuristics and compare their performances against the optimal policy. In Section 2.7, we consider the joint allowance optimization and inventory control problem.

2.3 Problem Description and Formulation

We consider a setting where production is managed over a finite planning horizon (the compliance period) consisting of T discrete time periods (production periods). Demand in each production period is a continuous and strictly positive random variable denoted by D. Demand in different periods are independently and identically distributed ($i.i.d.$) with cumulative distribution function (CDF) Φ and probability density function (pdf) ϕ. The decisions of whether or not, and how much to produce are made at the beginning of each period before the realization of demand, with a cost p incurred per unit produced. We assume that quantities produced in one period, if any, can be used to fulfill demand in that period (i.e.,

periods are sufficiently longer than unit production times). Each unit of positive inventory leftover at the end of a period incurs a holding cost h. Unfulfilled demand in any period is lost and a lost-sales penalty cost of l per unit lost is incurred. The one-period discount factor is denoted by $\alpha(\alpha \in (0, 1])$. All costs are assumed to be expressed at the beginning of each period. The salvage value per unit inventory at the end of the compliance period is assumed to be equal to the unit production cost, and unused allowance at the end of the compliance period is forfeited. We assume that $l \geq p$ (which also implies $l + h \geq \alpha p$). Periods are denoted by $t = 0, ..., T$, where T is the length of the planning horizon. We index periods in a backward fashion so that period t corresponds to the period that is t periods away form the end of the planning horizon. We denote the on-hand inventory level at the beginning of period t (prior to production) as x_t.

In each period a decision is made on the production quantity, or equivalently on the level to which the inventory level should be brought. We refer to this "produce-up-to" level as y_t. Clearly, $y_t \geq x_t$. In the absence of a constraint on the production allowance, y_t can be arbitrarily large. However, in our case, this produce-up-to level is constrained in period t by the remaining allowance c_t. In contrast to traditional notions of capacity, this allowance amount in each period varies and depends on the production decisions in previous periods. In particular, we assume that there is an allowance amount c_T available at the beginning of the planning horizon (this allowance corresponds to the maximum cumulative amount that could be produced over the entire planning horizon). We assume that c_T is exogenously specified. In Section 2.7, we consider the case where c_T is also a decision variable.

The above assumptions are consistent with the examples mentioned in the Section 2.1. For instance, in the case where logging is restricted, saw mills would typically have an assigned annual acreage which they can harvest. Because saw mills are located within a relatively short distance from the forested area, harvesting does not usually take place in a single shot, but is instead phased over the entire season which may consist of several months. This phasing out of the harvest allows saw mills to limit the storage costs and to prevent quality deterioration (Haight 2013). Such local saw mills would typically have limited access to supplies outside of their harvest area because transporting unprocessed logs long distances is cost-prohibitive. The saw mills respond to demand which can be stochastic from downstream buyers. In many cases, the assigned harvesting quotas are annual and cannot be banked (or borrowed against) over the years. Hence saw mills must manage harvesting and production with this quota in mind and knowing that any remaining balance carries no value; see Beaudoin et al. (2007) and Ouhimmou et al. (2009) for further discussion and details on forest operations.

Similar requirements arise in other applications, such as when water usage is restricted, water is drawn based on demand which can be stochastic. Supplementing the locally drawn water with water shipped from elsewhere is again cost-prohibitive. Quotas are allocated periodically and unused balances cannot typically be banked or borrowed against (see, e.g., Dudley and Musgrave 1988, Rogers et al. 2013).

Available allowance, along with on-hand inventory, is updated in each period

as follows:

$$\begin{aligned} x_{t-1} &= (y_t - d_t)^+, \\ c_{t-1} &= [c_t - (y_t - x_t)]^+, \end{aligned}$$

for $t = T, ..., 1$, where d_t is the realized demand in period t, and $x^+ = \max\{x, 0\}$. Note that to ensure feasibility, the produce-up-to level y_t must satisfy $x_t \leq y_t \leq x_t + c_t$. Together, x_t and c_t define the state of the system, and the knowledge of both is needed in making production decisions.

The objective is to determine in each period the optimal produce-up-to level such that the expected total discounted cost over the planning horizon is minimized. The problem can be formulated as a two-dimensional stochastic dynamic program, where the optimality equation for every period t, $t = T, ..., 1$, with state (x, c) is given by

$$\begin{aligned} f_t(x,c) = \min_{x \leq y \leq x+c} \{ & p(y-x) + L(y) + \alpha \int_0^y f_{t-1}(y-\xi, x+c-y)\phi(\xi)d\xi \\ & + \alpha \int_y^\infty f_{t-1}(0, x+c-y)\phi(\xi)d\xi \}, \end{aligned} \tag{2.1}$$

where

$$L(y) = \int_0^y h(y-\xi)\phi(\xi)d\xi + \int_y^\infty l(\xi - y)\phi(\xi)d\xi$$

corresponds to the expected one-period holding and shortage costs, and

$$f_0(x,c) = -px.$$

We can rewrite Equation (2.1) as follows:

$$f_t(x,c) = \min_{x \leq y \leq x+c} \{G_t(y, x+c)\} - px, \tag{2.2}$$

with

$$G_t(y, x+c) = py + L(y) + \alpha \int_0^y f_{t-1}(y-\xi, x+c-y)\phi(\xi)d\xi + \alpha \int_y^\infty f_{t-1}(0, x+c-y)\phi(\xi)d\xi. \quad (2.3)$$

Thus, starting with on-hand inventory level x and remaining allowance c, the optimal decision in period t is found by minimizing $G_t(y, x+c)$ over $\{y \mid x \leq y \leq x+c\}$.

In the absence of the allowance constraint, the problem is one-dimensional and admits a simple solution. In particular, the optimal production threshold in each period, which we denote by $\tilde{y}$, is fixed and given by $\tilde{y} = \Phi^{-1}(\frac{l-p}{h+l-\alpha p})$. The policy is said to be base-stock with base-stock level $\tilde{y}$ because, for each period t, if $x_t < \tilde{y}$, we produce up to $\tilde{y}$ ($y_t = \tilde{y}$), otherwise, we do not produce ($y_t = x_t$).

For our problem, we expect this simple policy not to hold since the feasible decision space can change from period to period, so that a fixed base-stock level may not always be attainable. More significantly, the fundamental cost tradeoffs in the two problems are different. In the unconstrained problem, the tradeoff is between inventory holding and shortage cost, and this tradeoff is the same in every period. In our constrained problem, there is now an additional concern. Whenever the amount of available allowance is tight, we must determine how best to allocate this allowance over the planning horizon to mitigate holding and shortage costs and also to ensure that available allowance does not unnecessarily go unused. This is, for example, the case when shortages would occur anyway over the planning horizon (i.e., cumulative demand over the periods is known with certainty to exceed the allowance). In that case, if too much allowance is used early on, we run the risk of incurring too much holding cost, and if we postpone consuming

allowance until later periods, we run the risk of incurring too much shortage cost. This allocation feature is absent from most traditional inventory problems and is what makes our problem different.

The fact that the problem is two-dimensional with a decision space that is state-dependent significantly complicates the analysis. However, examining Equation (2.3), we see that the sum $x + c$ is being treated as if it were a single variable. Note that $x + c$ specifies the maximum possible inventory level we can reach. We therefore refer to $x + c$ as the *effective capacity*. Unfortunately, we cannot entirely ignore the individual values of x and c since in Equation (2.2), the decision space is the interval $[x, x + c]$. In view of this difficulty, we introduce a new method, to which we refer as the extended state-space approach. We describe this approach in the next section.

2.4 Extended State-Space Analysis

In this section, we introduce an *extended* version of the original system, in which the action space is extended from $[x, x+c]$ to $[0, x+c]$. Doing so allows us to reduce the original two-dimensional problem into one with a single dimension. This one-dimensional problem is easier to analyze and enables us to identify properties that will serve as a basis for characterizing the structure of the optimal policy for the original problem.

In particular, define for $t = T, ..., 1$ (throughout this chapter, we use the overbar notation to denote functions or quantities associated with the extended

system),

$$\bar{f}_t(x,c) = \min_{0\le y\le x+c}\{\bar{G}_t(y,x+c)\} - px,$$

where

$$\bar{G}_t(y,x+c) = py + L(y) + \alpha\int_0^y \bar{f}_{t-1}(y-\xi, x+c-y)\phi(\xi)d\xi + \alpha\int_y^\infty \bar{f}_{t-1}(0,x+c-y)\phi(\xi)d\xi,$$

and

$$\bar{f}_0(x,c) = -px.$$

Notice that under the extended system, we can produce in such a way that the inventory level is set to any point $y \in [0, x+c]$. This means that we may either increase or decrease inventory at the beginning of each period. The cost of doing so is $p(y-x)$, which corresponds to either cost if $y-x>0$ or revenue if $y-x<0$.

Next, define

$$\bar{g}_t(z) = \min_{0\le y\le z}\{\bar{G}_t(y,z)\}. \tag{2.4}$$

Thus, we have $\bar{f}_t(x,c) = \bar{g}_t(x+c) - px$. We also get $\bar{G}_1(y,z) = py + L(y) - \alpha p\int_0^y(y-\xi)\phi(\xi)d\xi$, which is independent of z. Therefore, we can let

$$\bar{q}(y) = \bar{G}_1(y,\cdot) = py + L(y) - \alpha p\int_0^y (y-\xi)\phi(\xi)d\xi,$$

and then we have

$$\bar{G}_t(y,z) = \bar{q}(y) + \alpha[1-\Phi(y)]\bar{g}_{t-1}(z-y) + \alpha\int_0^y \bar{g}_{t-1}(z-\xi)\phi(\xi)d\xi. \tag{2.5}$$

From Equations (2.4) and (2.5), we can see that the extended system reduces to a one-dimensional system where the state is solely represented by z.

Next, we derive properties for the extended system.

Lemma 1. *For $1 \leq t \leq T$, the following holds*

(a) $\frac{\partial^2}{\partial y^2}\bar{G}_t(y,z) \geq 0 \ \forall y \ \forall z$,

(b) $p - l \leq \bar{g}_t'(z) \leq 0 \ \forall z$,

(c) $\bar{g}_t'(0) = p - l$, *and*

(d) $\bar{g}_t''(z) \geq 0 \ \forall z$.

Proof of Lemma 1: We prove Lemma 1 using induction. For $t = 1$, it is easy to show that $\frac{\partial^2}{\partial y^2}\bar{G}_1(y,z) \geq 0$. Let $\bar{y}_1(z)$ denote a minimizer of $\bar{q}(y)$ over $[0,z]$. From the convexity of $\bar{q}(y)$,

$$\bar{y}_1(z) = \begin{cases} \tilde{y} & \text{if } z \geq \tilde{y}, \\ z & \text{otherwise,} \end{cases}$$

where $\tilde{y} = \Phi^{-1}(\frac{l-p}{h+l-\alpha p})$, which is the base-stock level in the corresponding problem without the allowance constraint. Thus,

$$\bar{g}_1(z) = \begin{cases} \bar{q}(\tilde{y}) & \text{if } z \geq \tilde{y}, \\ \bar{q}(z) & \text{otherwise.} \end{cases}$$

It is then easy to verify that $\bar{g}_1'(0) = p - l$, $p - l \leq \bar{g}_1'(z) \leq 0$, and $\bar{g}_1''(z) \geq 0$.

For $t \geq 2$, suppose that in period $t-1$, $\frac{\partial^2}{\partial y^2}\bar{G}_{t-1}(y,z) \geq 0$, $p - l \leq \bar{g}_{t-1}'(z) \leq 0$, $\bar{g}_{t-1}'(0) = p - l$ and $\bar{g}_{t-1}''(z) \geq 0$. Then in period t,

$$\frac{\partial}{\partial y}\bar{G}_t(y,z) = \bar{q}'(y) + \alpha[\Phi(y) - 1]\bar{g}_{t-1}'(z-y),$$

and

$$\frac{\partial^2}{\partial y^2}\bar{G}_t(y,z) = [h + l - \alpha p + \alpha\bar{g}_{t-1}'(z-y)]\phi(y) + \alpha[1 - \Phi(y)]\bar{g}_{t-1}''(z-y) \geq 0.$$

After some algebra, we can show that $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=0} \leq 0$, and

$$\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=z} = [h + (1-\alpha)l]\Phi(z) + (1-\alpha)(p-l).$$

Next, we again let $\bar{y}_t(z)$ denote a minimizer of $\bar{G}(y,z)$ over $[0,z]$. Since $\bar{G}_t(y,z)$ is convex in y, if $z \geq \check{y} = \Phi^{-1}(\frac{(1-\alpha)(l-p)}{h+(1-\alpha)l})$ which implies $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=z} \geq 0$, then there exists a minimizer $\bar{Y}_t(z) \in [0,z]$, such that $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} = 0$, and we have $\bar{y}_t(z) = \bar{Y}_t(z)$. Otherwise, $\bar{y}_t(z) = z$. That is

$$\bar{y}_t(z) = \begin{cases} \bar{Y}_t(z) & \text{if } z \geq \check{y}, \\ z & \text{otherwise.} \end{cases}$$

Thus,

$$\bar{g}_t(z) = \begin{cases} \bar{G}_t(\bar{Y}_t(z), z) & \text{if } z \geq \check{y}, \\ \bar{G}_t(z,z) & \text{otherwise,} \end{cases}$$

and

$$\bar{g}'_t(z) = \begin{cases} \frac{d}{dz}\bar{G}_t(\bar{Y}_t(z), z) & \text{if } z \geq \check{y}, \\ \frac{d}{dz}\bar{G}_t(z,z) & \text{otherwise.} \end{cases}$$

For $\frac{d}{dz}\bar{G}_t(\bar{Y}_t(z), z)$, since

$$\bar{G}_t(\bar{Y}_t(z), z) = \bar{q}(\bar{Y}_t(z)) + \alpha[1-\Phi(\bar{Y}_t(z))]\bar{g}_{t-1}(z-\bar{Y}_t(z)) + \alpha\int_0^{\bar{Y}_t(z)} \bar{g}_{t-1}(z-\xi)\phi(\xi)d\xi,$$

we have

$$\begin{aligned}
\frac{d}{dz}\bar{G}_t(\bar{Y}_t(z), z) &= \bar{q}'(\bar{Y}_t(z))\bar{Y}'_t(z) + \alpha[1-\Phi(\bar{Y}_t(z))](1-\bar{Y}'_t(z))\bar{g}'_{t-1}(z-\bar{Y}_t(z)) \\
&\quad + \alpha\int_0^{\bar{Y}_t(z)} \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\
&= \alpha[1-\Phi(\bar{Y}_t(z))]\bar{g}'_{t-1}(z-\bar{Y}_t(z)) + \alpha\int_0^{\bar{Y}_t(z)} \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\
&\geq \alpha[1-\Phi(\bar{Y}_t(z))](p-l) + \alpha\int_0^{\bar{Y}_t(z)} (p-l)\phi(\xi)d\xi \\
&= \alpha(p-l) \\
&\geq p-l,
\end{aligned}$$

where the second equality is due to the fact that $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} = 0$, which implies $\bar{q}'(\bar{Y}_t(z)) = \alpha[1-\Phi(\bar{Y}_t(z))]\bar{g}'_{t-1}(z-\bar{Y}_t(z))$. On the other hand, we have

$$\begin{aligned}\frac{d}{dz}\bar{G}_t(\bar{Y}_t(z),z) &= \alpha[1-\Phi(\bar{Y}_t(z))]\bar{g}'_{t-1}(z-\bar{Y}_t(z)) + \alpha\int_0^{\bar{Y}_t(z)}\bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\ &\leq 0.\end{aligned}$$

Now, for $\frac{d}{dz}\bar{G}_t(z,z)$, since

$$\bar{G}_t(z,z) = \bar{q}(z) + \alpha[1-\Phi(z)]\bar{g}_{t-1}(0) + \alpha\int_0^z \bar{g}_{t-1}(z-\xi)\phi(\xi)d\xi,$$

we have

$$\begin{aligned}\frac{d}{dz}\bar{G}_t(z,z) &= (h+l-\alpha p)\Phi(z) + (p-l) + \alpha\int_0^z \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\ &\geq [h+(1-\alpha)l]\Phi(z) + (p-l) \\ &\geq p-l.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\frac{d}{dz}\bar{G}_t(z,z) &= \bar{q}'(z) + \alpha\int_0^z \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\ &< \alpha[1-\Phi(z)](p-l) + \alpha\int_0^z \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi \\ &\leq 0,\end{aligned}$$

where the first inequality is due to the fact that $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=z} < 0$, which implies $\bar{q}'(z) < \alpha[1-\Phi(z)](p-l)$. Now we can conclude that $p-l \leq \bar{g}'_t(z) \leq 0$.

Since $0 < \check{y}$, we have $\bar{g}_t(0) = \frac{d}{dz}\bar{G}_t(z,z)|_{z=0} = p-l$.

Next, since $\bar{g}_t(z) = \min_{0\leq y\leq z}\{\bar{G}_t(y,z)\}$, $\bar{g}_t(z)$ is convex in z if $\bar{G}_t(y,z)$ is jointly convex in (y,z) (see proof on page 227 in Porteus 2002). Let $H(\bar{G}_t)$ refer to the

Hessian of $\bar{G}_t(y,z)$, then we have

$$\begin{aligned}
|H(\bar{G}_t)| &= \frac{\partial^2}{\partial y^2}\bar{G}_t\frac{\partial^2}{\partial z^2}\bar{G}_t - (\frac{\partial^2}{\partial z \partial y}\bar{G}_t)^2 \\
&= \{[h+l-\alpha p+\alpha \bar{g}'_{t-1}(z-y)]\phi(y)+\alpha[1-\Phi(y)]\bar{g}''_{t-1}(z-y)\} \\
&\{\alpha[1-\Phi(y)]\bar{g}''_{t-1}(z-y)+\alpha\int_0^y \bar{g}''_{t-1}(z-\xi)\phi(\xi)d\xi\} \\
&-\{\alpha[\Phi(y)-1]\bar{g}''_{t-1}(z-y)\}^2 \\
&\geq \{\alpha[1-\Phi(y)]\bar{g}''_{t-1}(z-y)\}^2-\{\alpha[\Phi(y)-1]\bar{g}''_{t-1}(z-y)\}^2 \\
&= 0,
\end{aligned}$$

which implies that $H(\bar{G}_t)$ is positive-semidefinite and therefore $\bar{g}_t(z)$ is convex in z. This completes the induction and the proof. □

Property (a) in Lemma 1 states that $\bar{G}_t(y,z)$ is convex in y. Properties (b)-(d) are important results for subsequent proofs. In particular, the second inequality of property (b) states that $\bar{g}_t(z)$ is nonincreasing in z. Since $\bar{g}_t(z)$ is the optimal value of the minimization problem (Equation (2.4)), the more effective capacity we have, the larger the feasible region becomes and the lower the optimal value would be. The first inequality of property (b) indicates that one unit increase of z can at most reduce $\bar{g}_t(z)$ by $l-p$. Note that l is the unit lost-sales penalty cost and p is the unit production cost. Having one more unit of effective capacity gives us the possibility to satisfy at most one more unit of demand and therefore reduces the cost by at most $l-p$. Property (c) is due to the fact that the demand is strictly positive. Property (d) is intuitively obvious since z generates the constraint and the constraint cannot be tight for large enough z.

The following lemma explores additional properties of the extended system.

Lemma 2. *For $1 \leq t \leq T$, the following holds*

(a) $\bar{y}_t(z) \leq \tilde{y}\ \forall z$,

(b) $\bar{y}_t(z) \geq \frac{z}{t}$ *for* $z < t\tilde{y}$,

(c) $0 \leq \bar{y}_t'(z) \leq 1\ \forall z$,

(d) $\bar{y}_{t+1}(z) \leq \bar{y}_t(z)\ \forall z$, *and*

(e) $\bar{g}_{t+1}'(z) \leq \bar{g}_t'(z)\ \forall z$.

Proof of Lemma 2: (a) For $t = 1$, this is obviously true. For $t \geq 2$, $\bar{y}_t(z) = \begin{cases} \bar{Y}_t(z) & \text{if } z \geq \check{y}, \\ z & \text{otherwise.} \end{cases}$ For $z \geq \check{y}$, we have $\bar{y}_t(z) = \bar{Y}_t(z)$. By definition, $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} = 0$, which implies

$$\begin{aligned} \Phi(\bar{Y}_t(z)) &= \frac{l - p + \alpha\bar{g}_{t-1}'(z - \bar{Y}_t(z))}{h + l - \alpha p + \alpha\bar{g}_{t-1}'(z - \bar{Y}_t(z))} \\ &\leq \frac{l-p}{h+l-\alpha p} \\ &= \Phi(\tilde{y}), \end{aligned}$$

where the inequality is due to the fact that $p - l \leq \bar{g}_{t-1}'(\cdot) \leq 0$ by Lemma 1. Thus, we have $\bar{y}_t(z) = \bar{Y}_t(z) \leq \tilde{y}$. On the other hand, for $z < \check{y}$, we have $\bar{y}_t(z) = z < \check{y}$. It is easy to check that $\check{y} \leq \tilde{y}$, which implies $\bar{y}_t(z) \leq \tilde{y}$.

(b) can be proved using induction. For $t = 1$, this is obviously true. For $t \geq 2$, $\bar{y}_t(z) = \begin{cases} \bar{Y}_t(z) & \text{if } z \geq \check{y}, \\ z & \text{otherwise.} \end{cases}$ Since $\bar{y}_t(z) = z \geq \frac{z}{t}$ when $z < \check{y}$, we only need to show $\bar{Y}_t(z) \geq \frac{z}{t}$ when $z < t\tilde{y}$. Suppose that in period $t-1$, $\bar{Y}_{t-1}(z) \geq \frac{z}{t-1}$ when $z < (t-1)\tilde{y}$. Then in period t,

$$\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} = \bar{q}'(\frac{z}{t}) + \alpha[\Phi(\frac{z}{t}) - 1]\bar{g}_{t-1}'(\frac{t-1}{t}z).$$

We consider two cases: $\frac{t-1}{t}z \le \check{y}$ and $\frac{t-1}{t}z > \check{y}$. If $\frac{t-1}{t}z \le \check{y}$, then for $t \ge 2$, we have $\frac{z}{t} \le \frac{t-1}{t}z \le \check{y}$ and $\Phi(\frac{z}{t}) \le \Phi(\check{y}) = \frac{(1-\alpha)(l-p)}{h+(1-\alpha)l}$. Therefore,

$$\begin{aligned}\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} &\le \bar{q}'(\frac{z}{t}) + \alpha[\Phi(\frac{z}{t}) - 1](p-l) \\ &= [h+(1-\alpha)l]\Phi(\frac{z}{t}) + (1-\alpha)(p-l) \\ &\le [h+(1-\alpha)l]\frac{(1-\alpha)(l-p)}{h+(1-\alpha)l} + (1-\alpha)(p-l) \\ &= 0.\end{aligned}$$

By definition $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} = 0$. So $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} \ge \frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}}$. From Lemma 1, $\frac{\partial}{\partial y}\bar{G}_t(y,z)$ is nondecreasing in y. Therefore we have $\bar{Y}_t(z) \ge \frac{z}{t}$.

If $\frac{t-1}{t}z > \check{y}$, then $\bar{y}_{t-1}(\frac{t-1}{t}z) = \bar{Y}_{t-1}(\frac{t-1}{t}z)$, and

$$\bar{g}'_{t-1}(\frac{t-1}{t}z) = \bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) + \alpha\int_0^{\bar{Y}_{t-1}(\frac{t-1}{t}z)} \bar{g}'_{t-2}(\frac{t-1}{t}z - \xi)\phi(\xi)d\xi.$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} = \bar{q}'(\frac{z}{t}) &+ \alpha[\Phi(\frac{z}{t}) - 1][\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \\ &+ \alpha\int_0^{\bar{Y}_{t-1}(\frac{t-1}{t}z)} \bar{g}'_{t-2}(\frac{t-1}{t}z - \xi)\phi(\xi)d\xi].\end{aligned}$$

Again, from Lemma 1, $\bar{g}'(\cdot)$ is nondecreasing, and therefore, for $\xi \in [0, \bar{Y}_{t-1}(\frac{t-1}{t}z)]$, $\bar{g}'_{t-2}(\frac{t-1}{t}z - \xi) \ge \bar{g}'_{t-2}(\frac{t-1}{t}z - \bar{Y}_{t-1}(\frac{t-1}{t}z))$. Consequently,

$$\begin{aligned}\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} &\le \bar{q}'(\frac{z}{t}) + \alpha[\Phi(\frac{z}{t}) - 1][\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \\ &\quad + \alpha\int_0^{\bar{Y}_{t-1}(\frac{t-1}{t}z)} \bar{g}'_{t-2}(\frac{t-1}{t}z - \bar{Y}_{t-1}(\frac{t-1}{t}z))\phi(\xi)d\xi] \\ &= \bar{q}'(\frac{z}{t}) + \alpha[\Phi(\frac{z}{t}) - 1][\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \\ &\quad + \alpha\bar{g}'_{t-2}(\frac{t-1}{t}z - \bar{Y}_{t-1}(\frac{t-1}{t}z))\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))].\end{aligned}$$

By definition, we have $\frac{\partial}{\partial y}\bar{G}_{t-1}(y, \frac{t-1}{t}z)|_{y=\bar{Y}_{t-1}(\frac{t-1}{t}z)} = 0$, or equivalently,

$$\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) + \alpha[\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z)) - 1]\bar{g}'_{t-2}(\frac{t-1}{t}z - \bar{Y}_{t-1}(\frac{t-1}{t}z)) = 0.$$

From (a), $\bar{Y}_{t-1}(\frac{t-1}{t}z) \leq \tilde{y}$, and therefore, $\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \leq \Phi(\tilde{y}) = \frac{l-p}{h+l-\alpha p} < 1$, which implies $1 - \Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z)) > 0$. Thus, we have $\alpha\bar{g}'_{t-2}(\frac{t-1}{t}z - \bar{Y}_{t-1}(\frac{t-1}{t}z)) = \frac{\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z))}{1-\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))}$. So,

$$\begin{aligned}
\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} \leq\ & \bar{q}'(\frac{z}{t}) + \alpha[\Phi(\frac{z}{t}) - 1] \\
& [\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) + \frac{\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z))}{1-\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))}\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))] \\
=\ & \frac{1}{1-\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))}\{\bar{q}'(\frac{z}{t})[1 - \Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))] \\
& + \alpha[\Phi(\frac{z}{t}) - 1]\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z))\}.
\end{aligned}$$

It is clear that $\bar{q}'(\cdot)$ is nondecreasing. Since $\bar{Y}_{t-1}(\frac{t-1}{t}z) \leq \tilde{y}$, we have $\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \leq \bar{q}'(\tilde{y}) = 0$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} \leq\ & \frac{1}{1-\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))}\{\bar{q}'(\frac{z}{t})[1 - \Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))] \\
& + [\Phi(\frac{z}{t}) - 1]\bar{q}'(\bar{Y}_{t-1}(\frac{t-1}{t}z))\} \\
=\ & \frac{[h + (1-\alpha)p][\Phi(\frac{z}{t}) - \Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))]}{1-\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z))}.
\end{aligned}$$

By the inductive assumption for period $t-1$, we have $\bar{Y}_{t-1}(\frac{t-1}{t}z) \geq \frac{1}{t-1}(\frac{t-1}{t}z) = \frac{z}{t}$, which means $\Phi(\bar{Y}_{t-1}(\frac{t-1}{t}z)) \geq \Phi(\frac{z}{t})$. This implies $\frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\frac{z}{t}} \leq 0$, and therefore $\bar{Y}_t(z) \geq \frac{z}{t}$. This completes the induction and the proof of (b).

(c) For $t = 1$, this is trivially true. For $t \geq 2$, applying the implicit function theorem, we have

$$\bar{Y}'_t(z) = \frac{\alpha[1 - \Phi(\bar{Y}_t(z))]\bar{g}''_{t-1}(z - \bar{Y}_t(z))}{[h + l - \alpha p + \alpha\bar{g}'_{t-1}(z - \bar{Y}_t(z))]\phi(\bar{Y}_t(z)) + \alpha[1 - \Phi(\bar{Y}_t(z))]\bar{g}''_{t-1}(z - \bar{Y}_t(z))}.$$

Thus, it is easy to see that $\bar{Y}_t'(z) \in [0,1]$, and therefore, $0 \leq \bar{y}_t'(z) \leq 1$.

(d) and (e) can be proved using induction. For $t = 1$, we have

$$\bar{y}_1(z) = \begin{cases} \tilde{y} & \text{if } z \geq \tilde{y}, \\ z & \text{otherwise,} \end{cases} \quad \text{and } \bar{y}_2(z) = \begin{cases} \bar{Y}_2(z) & \text{if } z \geq \check{y}, \\ z & \text{otherwise.} \end{cases} \quad \bar{y}_2(z) \leq z \text{ as}$$

$\bar{Y}_2(z) \in [0,z]$, and since $\bar{y}_2(z) \leq \tilde{y}$ by (a), we have $\bar{y}_2(z) \leq \bar{y}_1(z)$. Next, recall that

$$\bar{g}_1'(z) = \begin{cases} 0 & \text{if } z \geq \tilde{y}, \\ (h+l-\alpha p)\Phi(z) + (p-l) & \text{otherwise,} \end{cases}$$

and

$$\bar{g}_2'(z) = \begin{cases} \frac{d}{dz}\bar{G}_2(\bar{Y}_2(z), z) & \text{if } z \geq \check{y}, \\ \frac{d}{dz}\bar{G}_2(z,z) & \text{otherwise.} \end{cases}$$

Since $\frac{d}{dz}\bar{G}_2(\bar{Y}_2(z), z) \leq \bar{q}'(\bar{Y}_2(z)) = (h+l-\alpha p)\Phi(\bar{Y}_2(z)) + (p-l) \leq (h+l-\alpha p)\Phi(z) + (p-l)$, and $\frac{d}{dz}\bar{G}_2(z,z) \leq \bar{q}'(z) = (h+l-\alpha p)\Phi(z) + (p-l)$, we have $\bar{g}_2'(z) \leq (h+l-\alpha p)\Phi(z) + (p-l)$. By Lemma 1, $\bar{g}_2'(z) \leq 0$, and therefore $\bar{g}_2'(z) \leq \bar{g}_1'(z)$. This completes the proof for $t = 1$.

For $t \geq 2$, suppose that in period $t-1$, $\bar{y}_t(z) \leq \bar{y}_{t-1}(z)$ and $\bar{g}_t'(z) \leq \bar{g}_{t-1}'(z)$, then in period t, $\bar{y}_{t+1}(z) = z = \bar{y}_t(z)$ when $z < \check{y}$, and when $z \geq \check{y}$, we have $\bar{y}_{t+1}(z) = \bar{Y}_{t+1}(z)$ and $\bar{y}_t(z) = \bar{Y}_t(z)$. Since $\frac{\partial}{\partial y}\bar{G}_{t+1}(y,z)$ is nondecreasing in y and

$$\begin{aligned} \frac{\partial}{\partial y}\bar{G}_{t+1}(y,z)|_{y=\bar{Y}_t(z)} &= \bar{q}'(\bar{Y}_t(z)) + \alpha[\Phi(\bar{Y}_t(z)) - 1]\bar{g}_t'(z - \bar{Y}_t(z)) \\ &\geq \bar{q}'(\bar{Y}_t(z)) + \alpha[\Phi(\bar{Y}_t(z)) - 1]\bar{g}_{t-1}'(z - \bar{Y}_t(z)) \\ &= \frac{\partial}{\partial y}\bar{G}_t(y,z)|_{y=\bar{Y}_t(z)} \\ &= 0 \\ &= \frac{\partial}{\partial y}\bar{G}_{t+1}(y,z)|_{y=\bar{Y}_{t+1}(z)}, \end{aligned}$$

we have $\bar{Y}_{t+1}(z) \leq \bar{Y}_t(z)$. Thus, $\bar{y}_{t+1}(z) \leq \bar{y}_t(z)$. Next, since

$$\begin{aligned}
\bar{g}'_{t+1}(z) - \bar{g}'_t(z) &= (h+l-\alpha p)\Phi(\bar{y}_{t+1}(z)) + (p-l) + \alpha \int_0^{\bar{y}_{t+1}(z)} \bar{g}'_t(z-\xi)\phi(\xi)d\xi \\
&\quad - [(h+l-\alpha p)\Phi(\bar{y}_t(z)) + (p-l) + \alpha \int_0^{\bar{y}_t(z)} \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi] \\
&\leq (h+l-\alpha p)[\Phi(\bar{y}_{t+1}(z)) - \Phi(\bar{y}_t(z))] \\
&\quad - \alpha \int_{\bar{y}_{t+1}(z)}^{\bar{y}_t(z)} \bar{g}'_{t-1}(z-\xi)\phi(\xi)d\xi,
\end{aligned}$$

where the last inequality is due to the inductive assumption that $\bar{g}'_t(z) \leq \bar{g}'_{t-1}(z)$. Since $\bar{y}_{t+1}(z) \leq \bar{y}_t(z)$ and $p - l \leq \bar{g}'_{t-1}(z) \leq 0$ by Lemma 1, we have

$$\begin{aligned}
\bar{g}'_{t+1}(z) - \bar{g}'_t(z) &\leq (h+l-\alpha p)[\Phi(\bar{y}_{t+1}(z)) - \Phi(\bar{y}_t(z))] - \alpha \int_{\bar{y}_{t+1}(z)}^{\bar{y}_t(z)} (p-l)\phi(\xi)d\xi \\
&= [h + (1-\alpha)l][\Phi(\bar{y}_{t+1}(z)) - \Phi(\bar{y}_t(z))] \\
&\leq 0.
\end{aligned}$$

Thus, $\bar{g}'_{t+1}(z) \leq \bar{g}'_t(z)$. This completes the induction and the proof. □

Property (a) states that the optimal produce-up-to level in the extended system cannot exceed the base-stock level in the corresponding problem without the allowance constraint. Property (b) says that when the effective capacity ($z = x + c$) is not enough, the optimal produce-up-to level in the extended system would be greater than or equal to the value of the effective capacity divided by the number of remaining periods. Properties (c)-(e) show the monotonicity of the optimal produce-up-to level. Namely, property (c) states that the optimal produce-up-to level is nondecreasing in the effective capacity, and having one more unit of effective capacity, we will produce up to at most one more unit. Property (d) says that the closer we get to the end of the planning horizon, the

higher is the produce-up-to level. Property (e) says that the closer we get to the end of the planning horizon, the less is the value of having more capacity. This is because that the fewer periods we have, the less decision flexibility we have in optimizing the system.

2.5 Structure of Optimal Policy

In this section, we show how the properties we have shown for the extended system can be used to characterize the optimal policy for the original system.

First, since under the extended system, we have the option to reduce inventory to any target level, it should be true that the expected optimal cost in the extended system is a lower bound of the expected optimal cost of the original system. That is, for $1 \leq t \leq T$, $f_t(x,c) \geq \bar{f}_t(x,c)\ \forall x\ \forall c$. We omit the proof as it is straightforward. Next, by Lemma 2, for the extended system, the optimal produce-up-to level $\bar{y}_t(z)$ is nonincreasing in t (t is the number of remaining periods until the end of the planning horizon). Therefore it is not difficult to see that, if $x_\tau \leq \bar{y}_\tau(x_\tau + c_\tau)$ for a certain period τ, then $x_t \leq \bar{y}_t(x_t + c_t)$ for $t = \tau-1, \tau-2, ..., 1$. In other words, in the extended system, if the starting inventory level in any period is below the optimal produce-up-to level, then the starting inventory levels in all the remaining periods will not exceed the corresponding optimal produce-up-to levels, respectively. Thus, starting from that period, it is never optimal to reduce inventory (even if it is allowed), and all the optimal production quantities will be nonnegative. This implies that, the expected optimal cost in the extended system equals that in the original system when the starting inventory level is below the optimal produce-up-to level in the extended system. We rigorously state this in

the next lemma.

Lemma 3. *For* $1 \leq t \leq T$, $f_t(x,c) = \bar{f}_t(x,c)$ *when* $x \leq \bar{y}_t(x+c)$.

Proof of Lemma 3: We prove Lemma 3 using induction. For $t = 1$, we have

$$\begin{aligned} f_1(x,c) &= \min_{x \leq y \leq x+c} \{\bar{q}(y) - px\} \\ &= \bar{q}(\bar{y}_1(x+c)) - px \\ &= \min_{0 \leq y \leq x+c} \{\bar{q}(y)\} - px \\ &= \bar{f}_1(x,c). \end{aligned}$$

For $t \geq 2$, suppose that in period $t-1$, $f_{t-1}(x,c) = \bar{f}_{t-1}(x,c)$ when $x \leq \bar{y}_{t-1}(x+c)$. Then in period t, when $x \leq \bar{y}_t(x+c)$,

$$\begin{aligned} \bar{f}_t(x,c) &= \min_{0 \leq y \leq x+c} \{\bar{G}_t(y, x+c)\} - px \\ &= \bar{G}_t(\bar{y}_t(x+c), x+c) - px \\ &= p\bar{y}_t(x+c) + L(\bar{y}_t(x+c)) \\ &\quad + \alpha \int_0^{\bar{y}_t(x+c)} \bar{f}_{t-1}(\bar{y}_t(x+c) - \xi, x + c - \bar{y}_t(x+c))\phi(\xi)d\xi \\ &\quad + \alpha \int_{\bar{y}_t(x+c)}^{\infty} \bar{f}_{t-1}(0, x + c - \bar{y}_t(x+c))\phi(\xi)d\xi - px. \end{aligned}$$

Since $0 \leq \bar{y}_{t-1}(x + c - \bar{y}_t(x+c))$, we have $\bar{f}_{t-1}(0, x + c - \bar{y}_t(x+c)) = f_{t-1}(0, x + c - \bar{y}_t(x+c))$. Therefore,

$$\begin{aligned} \bar{f}_t(x,c) &= p\bar{y}_t(x+c) + L(\bar{y}_t(x+c)) \\ &\quad + \alpha \int_0^{\bar{y}_t(x+c)} \bar{f}_{t-1}(\bar{y}_t(x+c) - \xi, x + c - \bar{y}_t(x+c))\phi(\xi)d\xi \\ &\quad + \alpha \int_{\bar{y}_t(x+c)}^{\infty} f_{t-1}(0, x + c - \bar{y}_t(x,c))\phi(\xi)d\xi - px. \end{aligned}$$

Now, for the term $\int_0^{\bar{y}_t(x+c)} \bar{f}_{t-1}(\bar{y}_t(x+c)-\xi, x+c-\bar{y}_t(x+c))\phi(\xi)d\xi$, if we can show that $\bar{y}_t(x+c)-\xi \leq \bar{y}_{t-1}(x+c-\xi)$ for $\xi \in [0, \bar{y}_t(x+c)]$, we can then replace $\bar{f}_{t-1}$ by f_{t-1}. Define for $0 \leq \xi \leq \bar{y}_t(x+c)$,

$$h(\xi) = \bar{y}_{t-1}(x+c-\xi) - \bar{y}_t(x+c) + \xi.$$

Then,

$$h'(\xi) = -\bar{y}'_{t-1}(x+c-\xi) + 1.$$

By Lemma 2, we have $h'(\xi) \geq 0$, which implies $h(\xi) \geq h(0)$ for $\xi \in [0, \bar{y}_t(x+c)]$. As $h(0) = \bar{y}_{t-1}(x+c) - \bar{y}_t(x+c)$ and therefore is nonnegative by Lemma 2, we have $h(\xi) \geq 0$, and therefore $\bar{y}_t(x+c) - \xi \leq \bar{y}_{t-1}(x+c-\xi)$ for $\xi \in [0, \bar{y}_t(x+c)]$. This means that we can indeed replace $\bar{f}_{t-1}$ by f_{t-1}. As a consequence, we have $\bar{f}_t(x,c) = G_t(\bar{y}_t(x+c), x+c) - px$. Since $x \leq \bar{y}_t(x+c) \leq x+c$, we have $\bar{f}_t(x,c) \geq \min_{x \leq y \leq x+c} \{G_t(y, x+c) - px\} = f_t(x,c)$. Moreover, we have $f_t(x,c) \geq \bar{f}_t(x,c)$. Thus, we can conclude that $f_t(x,c) = \bar{f}_t(x,c)$ when $x \leq \bar{y}_t(x+c)$. This completes the induction and the proof. □

Next, we show that $G_t(y, x+c)$ in the original system is nondecreasing in y when $\bar{y}_t(x+c) < y \leq x+c$.

Lemma 4. *For* $1 \leq t \leq T$, $\frac{\partial}{\partial y} G_t(y, x+c) \geq 0$ *when* $\bar{y}_t(x+c) \leq y \leq x+c$.

Proof of Lemma 4: We prove Lemma 4 using induction. For $t = 1$, we have $G_1(y, x+c) = \bar{G}_1(y, x+c) = \bar{q}(y)$, and $\bar{y}_1(x+c) = \begin{cases} \tilde{y} & \text{if } x+c \geq \tilde{y}, \\ x+c & \text{otherwise.} \end{cases}$ Since $\bar{y}_1(x+c) < y \leq x+c$, it must be $\tilde{y} < y \leq x+c$. For $y \geq \tilde{y}$, we have $\bar{q}'(y) \geq 0$. Thus, $\frac{\partial}{\partial y} G_1(y, x+c) \geq 0$ when $\bar{y}_1(x+c) \leq y \leq x+c$.

For $t \geq 2$, $\bar{y}_t(x+c) = \begin{cases} \bar{Y}_t(x+c) & \text{if } x+c \geq \check{y}, \\ x+c & \text{otherwise.} \end{cases}$ We only need to prove $\frac{\partial}{\partial y}G_t(y, x+c) \geq 0$ when $\bar{Y}_t(x+c) \leq y \leq x+c$. Suppose that in period $t-1$, $\frac{\partial}{\partial y}G_{t-1}(y, x+c) \geq 0$ when $\bar{Y}_{t-1}(x+c) \leq y \leq x+c$. Then in period t,

$$G_t(y, x+c) = py + L(y) + \alpha \int_0^y f_{t-1}(y-\xi, x+c-y)\phi(\xi)d\xi + \alpha \int_y^\infty f_{t-1}(0, x+c-y)\phi(\xi)d\xi.$$

By Lemma 3, we have $f_{t-1}(0, x+c-y) = \bar{f}_{t-1}(0, x+c-y) = \bar{g}_{t-1}(x+c-y)$. If $y-\xi \leq \bar{Y}_{t-1}(x+c-\xi)$, which implies $y-\xi \leq \bar{y}_{t-1}(x+c-\xi)$, then $f_{t-1}(y-\xi, x+c-y) = \bar{f}_{t-1}(y-\xi, x+c-y) = \bar{g}_{t-1}(x+c-\xi) - p(y-\xi)$.

Define for $0 \leq \xi \leq y$,

$$k(\xi) = \bar{Y}_{t-1}(x+c-\xi) - y + \xi.$$

Then,

$$k'(\xi) = -\bar{Y}'_{t-1}(x+c-\xi) + 1.$$

By Lemma 2, $k'(\xi) \geq 0$, which means $k(\xi)$ is nondecreasing and therefore $k(\xi) \geq k(0)$ for $\xi \in [0, y]$, where $k(0) = \bar{Y}_{t-1}(x+c) - y$. We separate the condition $\bar{Y}_t(x+c) \leq y \leq x+c$ into two cases: $\bar{Y}_t(x+c) \leq y \leq \bar{Y}_{t-1}(x+c)$ and $\bar{Y}_{t-1}(x+c) < y \leq x+c$.

If $\bar{Y}_t(x+c) \leq y \leq \bar{Y}_{t-1}(x+c)$, then we have $k(0) \geq 0$ and $y-\xi \leq \bar{Y}_{t-1}(x+c-\xi)$ for $\xi \in [0, y]$. Therefore, $f_{t-1}(y-\xi, x+c-y) = \bar{g}_{t-1}(x+c-\xi) - p(y-\xi)$ for $\xi \in [0, y]$. Thus,

$$G_t(y, x+c) = py + L(y) + \alpha \int_0^y [\bar{g}_{t-1}(x+c-\xi) - p(y-\xi)]\phi(\xi)d\xi$$

$$+ \alpha \int_y^\infty \bar{g}_{t-1}(x+c-y)\phi(\xi)d\xi$$
$$= \bar{G}_t(y, x+c).$$

By Lemma 1, we have $\frac{\partial}{\partial y}G_t(y, x+c) = \frac{\partial}{\partial y}\bar{G}_t(y, x+c) \geq 0$.

If $\bar{Y}_{t-1}(x+c) < y \leq x+c$, we have $k(0) < 0$ and $k(y) = \bar{Y}_{t-1}(x+c-y) \geq 0$. Therefore, there exists a $\varepsilon(x+c, y) \in [0, y]$, such that $k(\xi) < 0$ for $\xi \in [0, \varepsilon(x+c, y))$ and $k(\xi) \geq 0$ for $\xi \in [\varepsilon(x+c, y), y]$. By the assumption for period $t-1$, if $\bar{Y}_{t-1}(x+c-\xi) \leq y-\xi \leq x+c-\xi$, then $\min_{y-\xi \leq \omega \leq x+c-\xi}\{G_{t-1}(\omega, x+c-\xi)\} = G_{t-1}(y-\xi, x+c-\xi)$, since $G_{t-1}(\omega, x+c-\xi)$ in nondecreasing in ω on that region. Thus,

$$f_{t-1}(y-\xi, x+c-y) = \min_{y-\xi \leq \omega \leq x+c-\xi}\{G_{t-1}(\omega, x+c-\xi) - p(y-\xi)\}$$
$$= G_{t-1}(y-\xi, x+c-\xi) - p(y-\xi).$$

Consequently, we have $f_{t-1}(y-\xi, x+c-y) = G_{t-1}(y-\xi, x+c-\xi) - p(y-\xi)$ for $\xi \in [0, \varepsilon(x+c, y))$ and $f_{t-1}(y-\xi, x+c-y) = \bar{g}_{t-1}(x+c-\xi) - p(y-\xi)$ for $\xi \in [\varepsilon(x+c, y), y]$. This leads to

$$G_t(y, x+c) = \bar{q}(y) + \alpha[1-\Phi(y)]\bar{g}_{t-1}(x+c-y)$$
$$+ \alpha \int_0^{\varepsilon(x+c,y)} G_{t-1}(y-\xi, x+c-\xi)\phi(\xi)d\xi$$
$$+ \alpha \int_{\varepsilon(x+c,y)}^{y} \bar{g}_{t-1}(x+c-\xi)\phi(\xi)d\xi,$$

and

$$\frac{\partial}{\partial y}G_t(y, x+c) = \bar{q}'(y) - \alpha\phi(y)\bar{g}_{t-1}(x+c-y) - \alpha[1-\Phi(y)]\bar{g}'_{t-1}(x+c-y)$$
$$+ \alpha\frac{\partial}{\partial y}\varepsilon(x+c, y)G_{t-1}(y-\varepsilon(x+c, y), x+c-\varepsilon(x+c, y))$$

$$\begin{aligned}
&\phi(\varepsilon(x+c,y)) \\
&+\alpha\int_0^{\varepsilon(x+c,y)}\frac{\partial}{\partial y}G_{t-1}(y-\xi,x+c-\xi)\phi(\xi)d\xi \\
&+\alpha\bar{g}_{t-1}(x+c-y)\phi(y) \\
&-\alpha\frac{\partial}{\partial y}\varepsilon(x+c,y)\bar{g}_{t-1}(x+c-\varepsilon(x+c,y))\phi(\varepsilon(x+c,y)).
\end{aligned}$$

Note that $G_{t-1}(y-\varepsilon(x+c,y),x+c-\varepsilon(x+c,y)) = f_{t-1}(y-\varepsilon(x+c,y),x+c-y)+p(y-\varepsilon(x+c,y)) = \bar{g}_{t-1}(x+c-\varepsilon(x+c,y))$. Therefore,

$$\begin{aligned}
\frac{\partial}{\partial y}G_t(y,x+c) &= \bar{q}'(y)+\alpha[\Phi(y)-1]\bar{g}'_{t-1}(x+c-y) \\
&\quad+\alpha\int_0^{\varepsilon(x+c,y)}\frac{\partial}{\partial y}G_{t-1}(y-\xi,x+c-\xi)\phi(\xi)d\xi \\
&= \frac{\partial}{\partial y}\bar{G}_t(y,x+c)+\alpha\int_0^{\varepsilon(x+c,y)}\frac{\partial}{\partial y}G_{t-1}(y-\xi,x+c-\xi)\phi(\xi)d\xi.
\end{aligned}$$

By Lemma 1, $\frac{\partial}{\partial y}\bar{G}_t(y,x+c) \geq 0$ for $y \geq \bar{Y}_t(x+c)$. For $\xi \in [0,\varepsilon(x+c,y))$, we have $\bar{Y}_{t-1}(x+c-\xi) < y-\xi$. By the assumption for period $t-1$, we have $\frac{\partial}{\partial y}G_{t-1}(y-\xi,x+c-\xi) \geq 0$ there. Thus, $\frac{\partial}{\partial y}G_t(y,x+c) \geq 0$. Therefore, we can conclude that $\frac{\partial}{\partial y}G_t(y,x+c) \geq 0$ when $\bar{Y}_t(x+c) \leq y \leq x+c$. This completes the induction and the proof. □

Corollary 5. *For* $1 \leq t \leq T$, $\underset{x\leq y\leq x+c}{\operatorname{argmin}}\{G_t(y,x+c)\} = \bar{y}_t(x+c)$ *when* $x \leq \bar{y}_t(x+c)$, *and* $\underset{x\leq y\leq x+c}{\operatorname{argmin}}\{G_t(y,x+c)\} = x$ *when* $x > \bar{y}_t(x+c)$.

Corollary 5 directly follows from Lemmas 1, 3, and 4. It shows that the optimal production thresholds in the original system are exactly the same as those in the extended system, and therefore, they share all the properties that we proved in the previous lemmas. We precisely describe the structure of the optimal policy of the original system in the following theorem.

Theorem 6. *The optimal policy is specified by thresholds $y_t^*(x_t, c_t)$ such that if $x_t < y_t^*(x_t, c_t)$, we produce up to $y_t^*(x_t, c_t)$, otherwise, we do not produce. Moreover, $y_t^*(x_t, c_t)$ depends only on the sum $x_t + c_t$; that is, $y_t^*(x_t, c_t) = y_t^*(x_t', c_t')$, for all (x_t, c_t) and (x_t', c_t') such that $x_t + c_t = x_t' + c_t'$. The optimal policy has the following additional properties*

(a) If $x_t + c_t \geq t\tilde{y}$, then $y_t^*(x_t, c_t) = \tilde{y}$,

(b) If $x_t + c_t < t\tilde{y}$, then $\frac{x_t + c_t}{t} \leq y_t^*(x_t, c_t) \leq \tilde{y}$,

(c) If $x_t + c_t < \check{y}$, then $y_t^*(x_t, c_t) = x_t + c_t$,

(d) $y_t^*(x_t, c_t)$ is nondecreasing in c_t,

(e) $\frac{\partial}{\partial c_t} y_t^*(x_t, c_t) \leq 1$, and

(f) $y_t^*(x_t, c_t)$ is nonincreasing in t.

($\tilde{y} = \Phi^{-1}(\frac{l-p}{h+l-\alpha p})$ is the base-stock level in the corresponding problem without the allowance constraint, and $\check{y} = \Phi^{-1}(\frac{(1-\alpha)(l-p)}{h+(1-\alpha)l})$.)

Theorem 6 is a consequence of the preceding statements. From Corollary 5, the optimal policy is specified by the dynamic threshold $y_t^*(x_t, c_t) = \bar{y}_t(x_t + c_t)$. In each period t, if the inventory level x_t is less than $y_t^*(x_t, c_t)$, we produce up to $y_t^*(x_t, c_t)$, otherwise, we do not produce. Property (a) is due to the fact that when $x_t + c_t \geq t\tilde{y}$, the allowance constraint is not active. Thus, the original optimal solution is feasible and therefore optimal. Properties (b), (d)-(f) are proved in Lemma 2. Property (c) is proved in Lemma 1.

Property (a) in Theorem 6 states that when the effective capacity $(x + c)$ is sufficiently high, we produce up to the base-stock level in the corresponding problem without the allowance constraint. Property (b) indicates that when the effective capacity is in the mid-region, we produce up to a threshold that

is upper-bounded by the base-stock level in the unconstrained problem, and lower-bounded by the value of the effective capacity divided by the number of remaining periods. Property (c) states that when the effective capacity is sufficiently low, we produce as much as possible. Property (d) shows that, in each period, the higher the remaining allowance amount is, the higher the production threshold is. Property (e) states that each unit increase in the remaining allowance will lead to at most one unit increase in the production threshold. Property (f) suggests that the closer we get to the end of the planning horizon, the higher the production threshold will be.

Property (a) confirms our intuition in Section 2.3 that when the allowance amount is high, we should not change the production decision from the one used in the unconstrained problem. However, when the allowance amount is in the mid-region, the decision becomes how best to allocate the allowance over the planning horizon, and the production thresholds are lower than those in the unconstrained problem. This is because we run the risk of incurring too much holding cost if too much allowance is used early on. On the other hand, the production thresholds are higher than the value of the effective capacity divided by the number of remaining periods. This is because we run the risk of incurring too much shortage cost (while having unused allowance) if we postpone using allowance until later periods. Property (f) confirms the intuition that, knowing shortages would occur anyway over the remaining planning horizon (i.e., cumulative demand over the remaining periods is known with certainty to exceed the remaining capacity), we should use the available allowance up as much as possible. Properties (d) and (e) are due to the fact that the allowance serves as

a constraint on our optimization problem. Property (c) intrinsically results from the discounted cost setting.

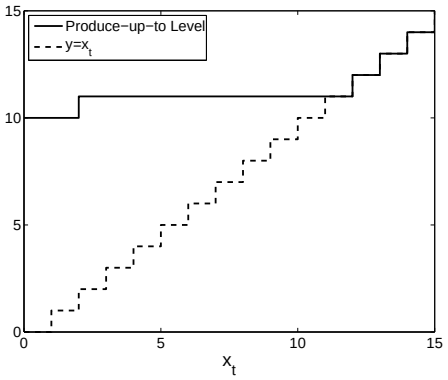
Next, we provide numerical results illustrating the structure of the optimal policy. Figure 2.1 illustrates the structure of the optimal policy and the properties stated in Theorem 6. (The figures are for an example where the demand is uniformly distributed on $[1, 15]$. We use this demand setting throughout all the numerical studies in this work, and the results are qualitatively the same for other common distributions we tested. $p = 3$, $h = 1$ and $l = 8$ in 2.1(a)-2.1(e); $\alpha = 0.8$ in 2.1(a)-2.1(d); $\alpha = 1$ in 2.1(e); $t = 5$ in 2.1(a)-2.1(d); $c_t = 30$ in 2.1(a); and $x_t = 0$ in 2.1(c)-2.1(e).)

Figures 2.1(a) and 2.1(b) illustrate the structure of the optimal policy. Figure 2.1(c) shows the production thresholds for different values of x_t+c_t, and properties (a)-(d) in Theorem 6 can be easily observed there. Namely, if $x_t + c_t \geq t\tilde{y}$, then $y_t^* = \tilde{y}$; if $x_t + c_t < t\tilde{y}$, then $\frac{x_t+c_t}{t} \leq y_t^* \leq \tilde{y}$; and if $x_t + c_t < \check{y}$, then $y_t^* = x_t + c_t$; besides, y_t^* is nondecreasing in c_t. Figure 2.1(d) illustrates property (e), namely, $\frac{\partial}{\partial c_t} y_t^*(x_t, c_t) \leq 1$. Figure 2.1(e) confirms property (f), namely, y_t^* is nonincreasing in t.

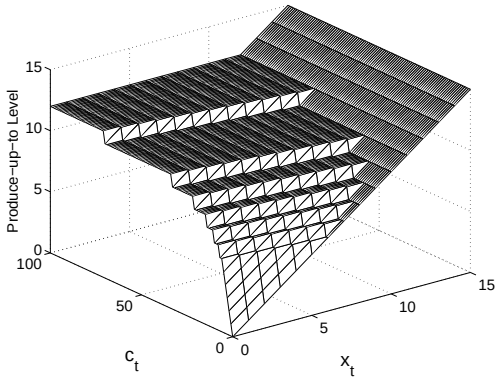
Next, we examine how the allowance constraint affects the expected optimal cost and the expected cumulative amount produced.

Corollary 7. *For $0 \leq t \leq T$, the following holds*
(a) $f_t(x, c)$ is convex in c $\forall x$, and
(b) $p - l \leq \frac{\partial}{\partial c} f_t(x, c) \leq 0$ $\forall x$ $\forall c$.

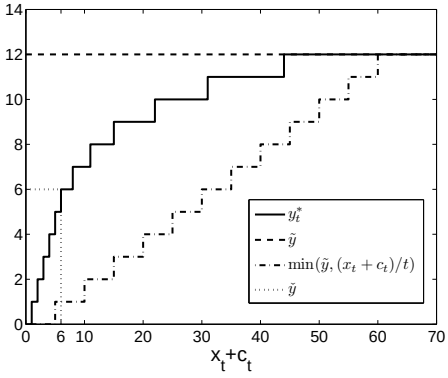
Proof of Corollary 7: We prove Corollary 7 using induction, and for part (a), we verify the convexity by showing that $\frac{\partial}{\partial c} f_{t-1}(x, c)$ is nondecreasing in c. For



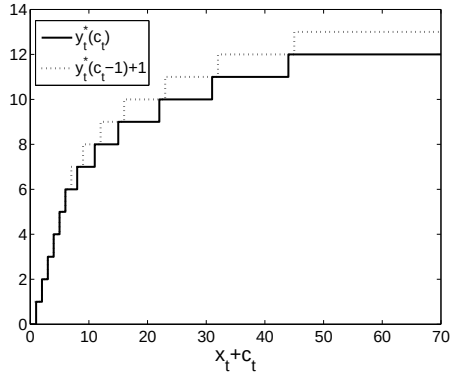
(a) Optimal Produce-Up-To Level as A Function of x_t for Fixed c_t



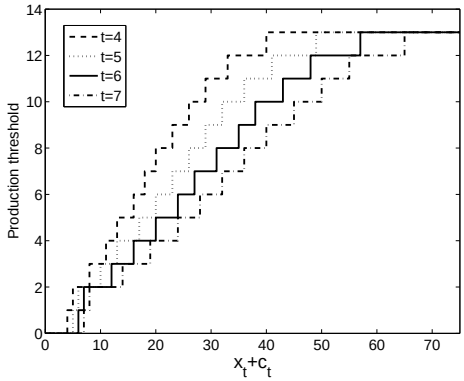
(b) Optimal Produce-Up-To Level as A Function of x_t and c_t



(c) Illustration of Properties a-d



(d) Illustration of Property e



(e) Illustration of Property f

Figure 2.1: An Illustration of the Structure of Optimal Policy

$t = 0$, this is obviously true.

For $t \geq 1$, suppose that in period $t-1$, $\frac{\partial}{\partial c} f_{t-1}(x,c)$ is nondecreasing in c, and $p - l \leq \frac{\partial}{\partial c} f_{t-1}(x,c) \leq 0$, then in period t, we consider three cases: $x \leq \bar{y}_t(x)$, $x \geq \tilde{y}$, and $\bar{y}_t(x) < x < \tilde{y}$.

If $x \leq \bar{y}_t(x)$, then by Lemma 2, $x \leq \bar{y}_t(x+c)$ $\forall c$. Thus, by Lemma 3, we have $f_t(x,c) = \bar{f}_t(x,c)$, and therefore $\frac{\partial^2}{\partial c^2} f_t(x,c) = \frac{\partial^2}{\partial c^2} \bar{f}_t(x,c) = \bar{g}_t''(x+c) \geq 0$, which implies that $\frac{\partial}{\partial c} f_t(x,c)$ is nondecreasing in c $\forall c$.

If $x \geq \tilde{y}$, then by Lemma 2, $x \geq \bar{y}_t(x+c)$ $\forall c$. Thus, by Lemma 4, we have

$$\begin{aligned} f_t(x,c) &= \min_{x \leq y \leq x+c} \{G_t(y, x+c)\} - px \\ &= G_t(x, x+c) - px \\ &= L(x) + \alpha \int_0^x f_{t-1}(x-\xi, c)\phi(\xi)d\xi + \alpha \int_x^\infty f_{t-1}(0,c)\phi(\xi)d\xi. \end{aligned}$$

Thus,

$$\frac{\partial}{\partial c} f_t(x,c) = \alpha \int_0^x \frac{\partial}{\partial c} f_{t-1}(x-\xi, c)\phi(\xi)d\xi + \alpha \int_x^\infty \frac{\partial}{\partial c} f_{t-1}(0,c)\phi(\xi)d\xi,$$

which is nondecreasing in c $\forall c$, by the inductive assumption for period $t-1$.

If $\bar{y}_t(x) < x < \tilde{y}$, since $\bar{y}_t(x+c)$ is nondecreasing in c for $x \in (\bar{y}_t(x), \tilde{y})$, $\exists$ a finite $\hat{c}(x) > 0$ such that $x > \bar{y}_t(x+c)$ for $c \in [0, \hat{c}(x))$, $x < \bar{y}_t(x+c)$ for $c \in (\hat{c}(x), \infty)$ and $x = \bar{y}_t(x + \hat{c}(x))$. Applying the above argument, $\frac{\partial}{\partial c} f_t(x,c)$ is nondecreasing in c for $c \in [0, \hat{c}(x)]$ and $c \in [\hat{c}(x), \infty)$. To show that $\frac{\partial}{\partial c} f_t(x,c)$ is nondecreasing in c on $[0, \infty)$, we still need to show that $\lim_{c \to \hat{c}(x)^-} \frac{\partial}{\partial c} f_t(x,c) = \lim_{c \to \hat{c}(x)^+} \frac{\partial}{\partial c} f_t(x,c)$, where

$$\begin{aligned} \lim_{c \to \hat{c}(x)^-} \frac{\partial}{\partial c} f_t(x,c) = \alpha &\int_0^x \frac{\partial}{\partial c} f_{t-1}(x-\xi, c)|_{c=\hat{c}(x)}\phi(\xi)d\xi \\ &+ \alpha \int_x^\infty \frac{\partial}{\partial c} f_{t-1}(0,c)|_{c=\hat{c}(x)}\phi(\xi)d\xi, \end{aligned}$$

and $\lim_{c\to\hat{c}(x)^+}\frac{\partial}{\partial c}f_t(x,c)=\lim_{c\to\hat{c}(x)}\bar{g}'_t(x+c)$. Note that since $\bar{y}_t(x+\hat{c}(x))=x<x+\hat{c}(x)$, $\exists\ \delta>0$ such that $\bar{y}_t(x+c)<x+c$ for $c\in[\hat{c}(x),\hat{c}(x)+\delta)$, and by the definition of $\bar{y}_t(x+c)$, we must have $x+c\geq\check{y}$, and $\bar{y}_t(x+c)=\bar{Y}_t(x+c)$, thus, from the proof of Lemma 1,

$$\begin{aligned}\lim_{c\to\hat{c}(x)^+}\frac{\partial}{\partial c}f_t(x,c)=&\lim_{c\to\hat{c}(x)}\{\alpha[1-\Phi(\bar{y}_t(x+c))]\bar{g}'_{t-1}(x+c-\bar{y}_t(x+c))\\&+\alpha\int_0^{\bar{y}_t(x+c)}\bar{g}'_{t-1}(x+c-\xi)\phi(\xi)d\xi\}\\=&\lim_{c\to\hat{c}(x)}[\alpha\int_0^{\bar{y}_t(x+c)}\bar{g}'_{t-1}(x+c-\xi)\phi(\xi)d\xi\\&+\alpha\int_{\bar{y}_t(x+c)}^{\infty}\bar{g}'_{t-1}(x+c-\bar{y}_t(x+c))\phi(\xi)d\xi].\end{aligned}$$

However, when $c\in[\hat{c}(x),\hat{c}(x)+\delta)$, we have $x\leq\bar{y}_t(x+c)$, and therefore $x\leq\bar{y}_{t-1}(x+c)$ by Lemma 2. It is easy to check (using the method in the proof of Lemma 3) that $x-\xi\leq\bar{y}_{t-1}(x+c-\xi)$ for all $\xi\in[0,\bar{y}_t(x+c)]$. Thus, $f_{t-1}(x-\xi,c)=\bar{f}_{t-1}(x-\xi,c)$, and therefore $\frac{\partial}{\partial c}f_{t-1}(x-\xi,c)=\frac{\partial}{\partial c}\bar{f}_{t-1}(x-\xi,c)=\bar{g}'_{t-1}(x+c-\xi)$, for all $\xi\in[0,\bar{y}_t(x+c)]$. Consequently,

$$\begin{aligned}\lim_{c\to\hat{c}(x)^+}\frac{\partial}{\partial c}f_t(x,c)=\lim_{c\to\hat{c}(x)}[&\alpha\int_0^{\bar{y}_t(x+c)}\frac{\partial}{\partial c}f_{t-1}(x-\xi,c)\phi(\xi)d\xi\\&+\alpha\int_{\bar{y}_t(x+c)}^{\infty}\frac{\partial}{\partial c}f_{t-1}(x-\bar{y}_t(x+c),c)\phi(\xi)d\xi].\end{aligned}$$

Now, notice that $x=\bar{y}_t(x+\hat{c}(x))$, thus,

$$\begin{aligned}\lim_{c\to\hat{c}(x)^+}\frac{\partial}{\partial c}f_t(x,c)=&\alpha\int_0^x\frac{\partial}{\partial c}f_{t-1}(x-\xi,c)|_{c=\hat{c}(x)}\phi(\xi)d\xi\\&+\alpha\int_x^{\infty}\frac{\partial}{\partial c}f_{t-1}(0,c)|_{c=\hat{c}(x)}\phi(\xi)d\xi\\=&\lim_{c\to\hat{c}(x)^-}\frac{\partial}{\partial c}f_t(x,c).\end{aligned}$$

This completes the induction for part (a).

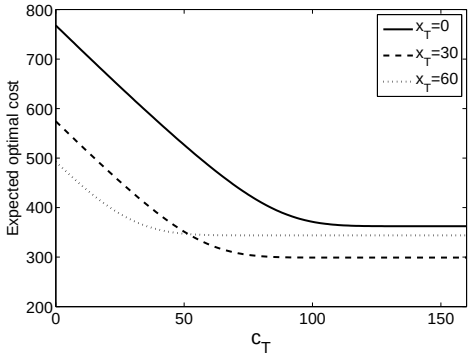
For part (b), we have shown that in period t, either $\frac{\partial}{\partial c} f_t(x,c) = \bar{g}_t'(x+c)$, or

$$\frac{\partial}{\partial c} f_t(x,c) = \alpha \int_0^x \frac{\partial}{\partial c} f_{t-1}(x-\xi,c)\phi(\xi)d\xi + \alpha \int_x^\infty \frac{\partial}{\partial c} f_{t-1}(0,c)\phi(\xi)d\xi.$$

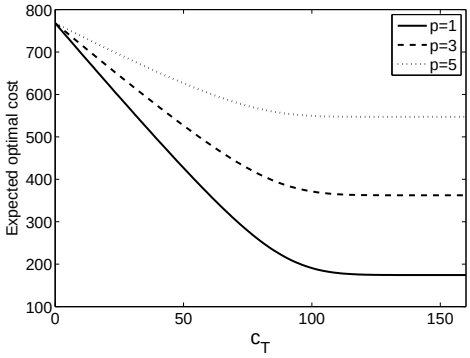
In the first case, we have $p - l \leq \frac{\partial}{\partial c} f_t(x,c) \leq 0$ by Lemma 1. In the second case, the inequalities directly follow the inductive assumption in period $t-1$. This completes the whole proof. □

Property (a) states that the expected optimal cost is convex with respect to the remaining allowance. This implies that there is diminishing value to increasing the allowance amount, or equivalently, that cost becomes increasingly insensitive to the allowance amount as the allowance amount increases. Property (b) states that a unit increase in the allowance amount leads to at most $p - l$ unit decrease in cost, or equivalently, a unit decrease in the allowance amount leads to at most $l - p$ unit increase in the expected optimal cost, further confirming the relative insensitivity of cost to the allowance amount. These results are illustrated in Figure 2.2 for an example system with 12 periods ($T = 12$). For all the cases shown, the allowance amount can be decreased by up to 50% from the maximum value of 160 with cost increasing by at most 15%. Note that an upper bound on the maximum amount of allowance that would ever be consumed under an optimal policy is $T\tilde{y}$, where $\tilde{y}$ is the base-stock level in the corresponding problem without the allowance constraint.

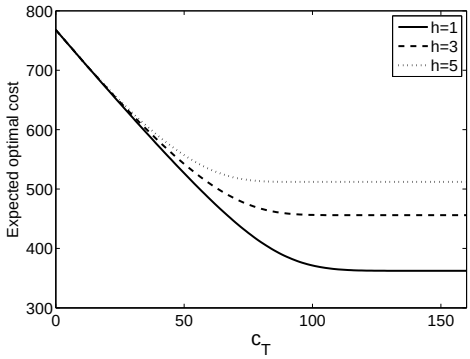
The above results suggest that in some cases it might be possible to impose an allowance constraint (or to tighten an existing constraint) without significantly increasing cost. This is of particular relevance to settings where the constraint is due to a scarce natural resource used as input in production or due to a limit on waste or harmful externalities associated with production.



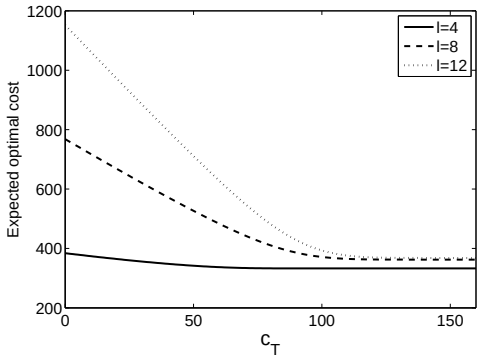
(a) $p = 3$, $h = 1$, $l = 8$, $\alpha = 1$



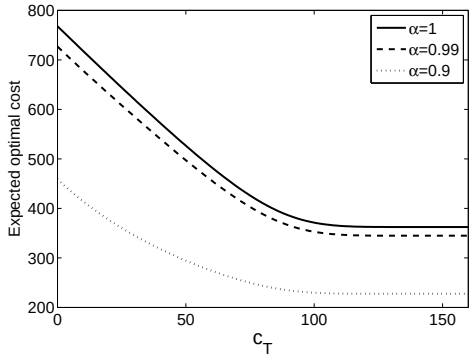
(b) $x_T = 0$, $h = 1$, $l = 8$, $\alpha = 1$



(c) $x_T = 0$, $p = 3$, $l = 8$, $\alpha = 1$



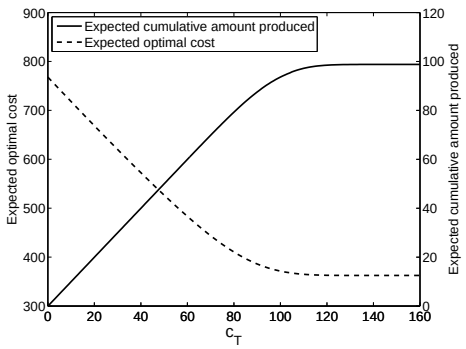
(d) $x_T = 0$, $p = 3$, $h = 1$, $\alpha = 1$



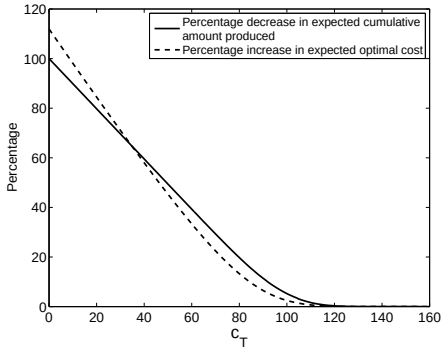
(e) $x_T = 0$, $p = 3$, $h = 1$, $l = 8$

Figure 2.2: Impact of Allowance on Optimal Cost

Figure 2.3(a) illustrates, for an example system, how the expected cumulative amount produced over the entire planning horizon under the optimal policy and the expected optimal cost are affected by the allowance constraint. Figure 2.3(b) shows the percentage differences in cumulative amount produced and cost between a constrained and an unconstrained system. The results confirm that, there is indeed an opportunity to significantly reduce the cumulative amount produced (in some applications this corresponds to the scare natural resources used or harmful externalities generated), without significantly increasing cost. Figure 2.3(b) also suggests that there is a range of values for c_T, for which the percentage reduction in cumulative amount produced is higher than the percentage increase in cost.



(a) Expected Cumulative Amount Produced and Expected Optimal Cost



(b) Constrained and Unconstrained Systems

Figure 2.3: Impact of Allowance on Cumulative Amount Produced and Expected Optimal Cost ($T = 12$, $x_T = 0$, $p = 3$, $h = 1$, $l = 8$, $\alpha = 1$)

2.6 Heuristics

In this section, we examine the benefit of using the optimal policy relative to using simpler heuristics. In particular, we consider three plausible heuristics

that account in different ways for the production allowance constraint. Below, we describe each heuristic and then provide numerical results that compare its performance relative to that of the optimal policy.

Heuristic U: This heuristic is motivated by Properties (a) and (b) in Theorem 6, which indicate that $y_t^*(x_t, c_t)$ is upper-bounded by $\tilde{y}$ (the base-stock level in the corresponding problem without the allowance constraint). Under this heuristic, if the on-hand inventory level at the beginning of a period is below $\tilde{y}$, we produce to bring the inventory level as close as possible to $\tilde{y}$; otherwise, we do not produce. In other words, the production thresholds are specified as follows:

$$\bar{y}_t^U(x_t, c_t) = \begin{cases} \tilde{y} & \text{if } x_t + c_t \geq \tilde{y}, \\ x_t + c_t & \text{otherwise.} \end{cases}$$

Heuristic L: This heuristic is motivated by the first inequality of Property (b) in Theorem 6, which indicates that $y_t^*(x_t, c_t)$ is lower-bounded by $\frac{x_t+c_t}{t}$ (the effective capacity divided by the number of remaining periods) when $x_t + c_t < t\tilde{y}$. Under this heuristic, if effective capacity in a period is sufficiently high, we produce up to $\tilde{y}$; otherwise, we produce up to a level which equals the effective capacity divided by the number of remaining periods. In particular, the production thresholds are given by:

$$\bar{y}_t^L(x_t, c_t) = \begin{cases} \tilde{y} & \text{if } x_t + c_t \geq t\tilde{y}, \\ \frac{x_t+c_t}{t} & \text{otherwise.} \end{cases}$$

Heuristic P: This heuristic is mimicked by inventory systems where there is a production capacity constraint applied independently to each period. Under this heuristic, allowance c_T is divided equally among the different periods, so that in each period, the production threshold is given by $\min\{\frac{c_T}{T}, \tilde{y}\}$.

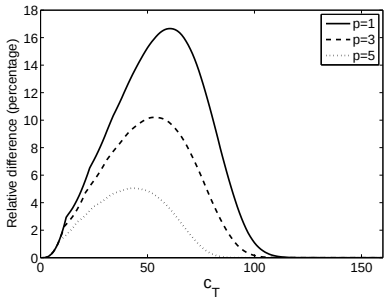
We provide numerical results that illustrate the relative difference between the performance of heuristics U and L and the performance of the optimal policy (we comment on the performance of heuristic P at the end of the section). The relative difference, which we denote by δ, is defined as follows:

$$\delta = \frac{\text{expected cost under the heuristic } - \text{ expected cost under the optimal policy}}{\text{expected cost under the optimal policy}}.$$

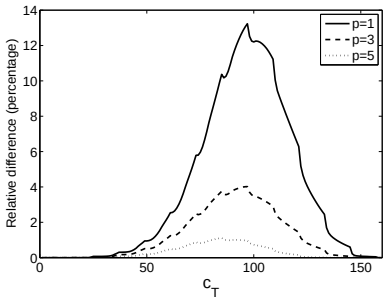
Figure 2.4 illustrates the effect of various problem parameters, namely c_T, p, α, and σ, where $\sigma = \frac{l-p}{h}$. (The effect of h and l can be evaluated by evaluating the effect of $\sigma = \frac{l-p}{h}$ since $\tilde{y} = \Phi^{-1}(\frac{l-p}{h+l-p})$ when $\alpha = 1$.) The following observations can be made.

Observation 1: *For both heuristics, the relative difference is (generally) first increasing and then decreasing in c_T.*

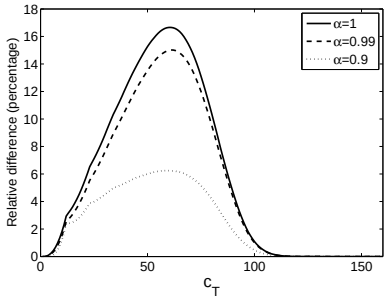
For heuristic U, the difference in cost relative to the optimal policy is mainly a consequence of the higher holding cost incurred under the heuristic. When c_T is small, lost-sales cost is large and dominates the holding cost under both the heuristic and the optimal policy. Thus, the relative difference is small. As c_t initially increases, the difference in holding costs between the heuristic and the optimal policy increases. However, with further increases, this difference diminishes as both the heuristic and the optimal policy are able to produce up to the same base-stock level over an increasing number of periods. For the extreme case where $c_T \geq T\tilde{y}$, heuristic U is optimal and the difference between the heuristic and the optimal policy vanishes. Similar explanations apply to heuristic L, except that for heuristic L, the difference in cost relative to the optimal policy is mainly a consequence of the higher lost-sales cost. (Notice that, the non-monotonicity in the increase and decrease observed in the figures for heuristic L is due to the



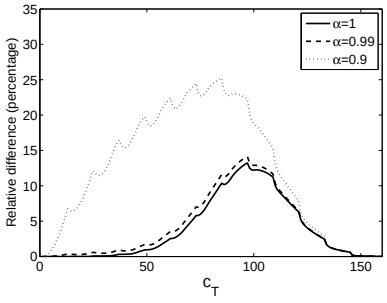
(a) Heuristic U ($h = 3$, $l = 8$, $\alpha = 1$)



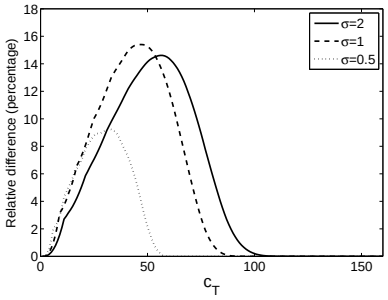
(d) Heuristic L ($h = 1$, $l = 8$, $\alpha = 1$)



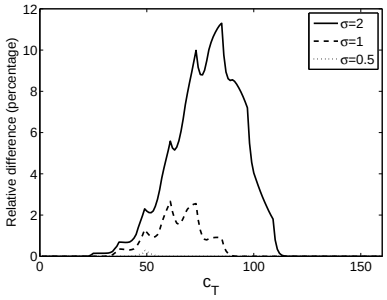
(b) Heuristic U ($p = 1$, $h = 3$, $l = 8$)



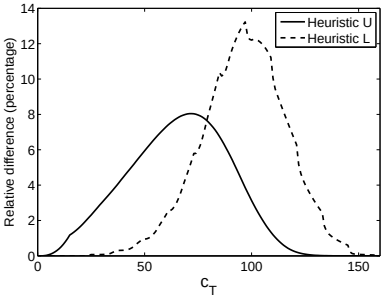
(e) Heuristic L ($p = 1$, $h = 1$, $l = 8$)



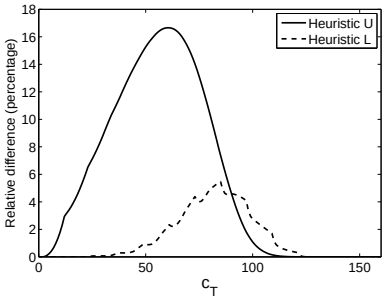
(c) Heuristic U ($p = 1$, $l = 5$, $\alpha = 1$)



(f) Heuristic L ($p = 1$, $l = 5$, $\alpha = 1$)



(g) Heuristics U and L ($p = 1$, $h = 1$, $l = 8$, $\alpha = 1$)



(h) Heuristics U and L ($p = 1$, $h = 3$, $l = 8$, $\alpha = 1$)

Figure 2.4: An Illustration of the Performance of Heuristics U and L ($T = 12$, $x_T = 0$)

discreteness of demand in the numerical experiments and the resulting rounding off of $\frac{x_t+c_t}{t}$.)

Observation 2: *For both heuristics, the relative difference is decreasing in p.*

As p increases, the fraction of expected cost due to production cost increases and gradually dominates the fractions due to inventory holding and lost-sales cost. This is true for both heuristics, explaining the decrease in the relative difference as p increases.

Observation 3: *The relative difference is increasing in α for heuristic U, but decreasing for heuristic L.*

As α gets smaller, future cost becomes less important. Thus, heuristic U which is myopic performs better. Heuristic L performs worse since under this heuristic, the effective capacity is divided equally among all the periods, and therefore all periods are treated equally.

Observation 4: *The relative difference is not monotone in the ratio σ for heuristic U, but is increasing for heuristic L.*

For heuristic U, when σ is large, the holding cost is dominated by the lost-sales cost. Thus, for the reasons discussed earlier, the relative difference is small. When σ is small, $\tilde{y}$ is small, and from Theorem 6, either heuristic U is optimal (if $c_T \geq T\tilde{y}$), or the difference between $\tilde{y}$ and y_t^* is small (the production quantity under heuristic U would be very close to the production quantity under the optimal policy), resulting in a small relative difference. As we can see from the figure, the largest relative difference appears when $\sigma = 1$. For heuristic L, as σ increases, the contribution of the lost-sales cost to total cost increases and, as a consequence, the relative difference increases.

Observation 5: *When the holding cost is low or when the available allowance is high, heuristic U performs better than heuristic L, and the reverse is true otherwise.*

As explained above, heuristic U performs better when the holding cost is low relative to lost-sales cost, and the opposite is true for heuristic L. The effect of the allowance can be explained as follows. When the allowance amount is low, lost sales are inevitable, but there is an opportunity to reduce holding cost (thus the superior performance of heuristic L). When the allowance amount is high, there is an opportunity to reduce lost sales (thus the superior performance of heuristic U).

Note that there are situations for which the heuristics perform much worse than what is shown in the figures. Consider for example a two-period problem where demand in each period takes value of 5 or 10 with equal probability. Suppose $c_2 = 10$, $x_2 = 0$, $p = 0.01$, $h = 10$, $l = 10.02$, and $\alpha = 1$. It is easy to check that it is optimal to produce 5 in each period, which results in an expected cost of 50.2. In contrast, under heuristic U, we produce 10 in period 1 and 0 in period 2, resulting in an expected cost of 75.2 and a corresponding relative difference of nearly 50%.

Finally, we note that the performance of heuristic P is similar to the performance of heuristic L. For example, the relative difference between heuristic P and the optimal policy is decreasing in p, decreasing in α, and increasing in the ratio σ. Under most of the settings considered, this period-based constraint policy performs worse than heuristic L.

In summary, while there are regions under which these simple heuristics

perform well, there is also a range of parameter values under which the performance of the optimal policy is significantly superior.

2.7 Joint Allowance Optimization and Inventory Control

We have so far assumed that the allowance amount c_T is exogenously set. In some applications, this may be a decision that would have to be made at the beginning of the compliance period. For example, guaranteeing access to a raw material or a critical natural resource that is in short supply may require paying a fee that is increasing in the amount to be secured. Similarly, securing the ability to order from a supplier up to a certain quantity may require paying a reservation fee in advance. In settings, where production is associated with negative environmental externalities, a regulating agency may also require the production firms to purchase pollution permits before production takes place.

In each of these examples, the firm must decide on how much allowance to purchase, knowing that not all of the allowance may eventually be used. The problem can thus be viewed as consisting of two stages. In the first stage (the investment stage), c_T is determined while, in the second stage (the operating stage), production decisions are made over the compliance period subject to the allowance constraint. Let w denote the price per unit of allowance. Then, the joint allowance optimization and inventory control problem can be formulated as

$$\min_{c \geq 0} F_T(w, x, c),$$

where

$$F_T(w, x, c) = wc + f_T(x, c),$$

and $f_T(x, c)$ is defined as in Section 2.3. $F_T(w, x, c)$ represents the expected total cost (the sum of investment cost and operating cost) in a T-period problem with allowance price w, allowance decision c, and starting inventory level x.

Proposition 8. *The expected total cost $F_T(w, x, c)$ is convex in c.*

The above result follows directly from the fact that $f_T(x, c)$ is convex in c per Corollary 7. This result is important because it implies that the optimal allowance amount can be computed efficiently using standard convex optimization methods.

Now, let $c_T^*(w, x)$ denote the minimum value of $c \geq 0$ that minimizes $F_T(w, x, c)$, which represents the optimal allowance amount to purchase. Then, let $F_T^*(w, x) = F_T(w, x, c_T^*(w, x))$ denote the corresponding optimal expected total cost.

Proposition 9. *The following holds*

(a) $c_T^*(w, x)$ is nonincreasing in w,

(b) $F_T^*(w, x)$ is nondecreasing and concave in w,

(c) $c_T^*(w, x) = 0$ when $w \geq l - p$, and

(d) $\frac{\partial c_T^*}{\partial w}(w, x)|_{w=0} = -\infty$.

Proof of Proposition 9: $\frac{\partial F_T}{\partial c}(w, x, c) = w + \frac{\partial f_T}{\partial c}(x, c)$. If $\frac{\partial f_T}{\partial c}(x, c)|_{c=0} \geq -w$, then by the convexity of $f_T(x, c)$ in c per Corollary 7, we have $c_T^*(w, x) = 0$ and $F_T^*(w, x) = f_T(x, 0)$. Otherwise, we have $w + \frac{\partial f_T}{\partial c}(x, c)|_{c=c_T^*(w,x)} = 0$. Thus, $1 + \frac{\partial^2 f_T}{\partial c^2}(x, c)|_{c=c_T^*(w,x)} \frac{\partial c_T^*}{\partial w}(w, x) = 0$, and $\frac{\partial c_T^*}{\partial w}(w, x) = -1/\frac{\partial^2 f_T}{\partial c^2}(x, c)|_{c=c_T^*(w,x)} \leq 0$.

Also, $\frac{\partial F_T^*}{\partial w}(x,w) = c_T^*(w,x) + w\frac{\partial c_T^*}{\partial w}(w,x) + \frac{\partial f_T}{\partial c}(x,c)|_{c=c_T^*(w,x)}\frac{\partial c_T^*}{\partial w}(w,x) = c_T^*(w,x) \geq 0$, and $\frac{\partial^2 F_T^*}{\partial w^2}(x,w) = \frac{\partial c_T^*}{\partial w}(w,x) \leq 0$. This proves (a) and (b).

(c) directly follows Corollary 7. To prove (d), note that $w + \frac{\partial f_T}{\partial c}(x,c)|_{c=c_T^*(w,x)} = 0$, which implies $\frac{\partial f_T}{\partial c}(x,c)|_{c=c_T^*(0,x)} = 0$. Then, applying Corollary 7, leads to $\frac{\partial^2 f_T}{\partial c^2}(x,c)|_{c=c_T^*(0,x)} = 0$. Noting that $\frac{\partial c_T^*}{\partial w}(w,x) = -1/\frac{\partial^2 f_T}{\partial c^2}(x,c)|_{c=c_T^*(w,x)}$, we conclude $\frac{\partial c_T^*}{\partial w}(w,x)|_{w=0} = -\infty$. This completes the proof. □

Two results from the above proposition are worth highlighting. First, the concavity of the optimal total cost in w stated in Property (b) implies that there is a diminishing effect to higher allowance prices, with the firm increasingly choosing lower allowance at the expenses of fulfilling demand; in the limit, the firm chooses not to purchase any allowance. Second, Property (d) indicates that a small initial increase in the unit price of allowance can lead to a significant decrease in the amount of allowance purchased. This result is of particular relevance to applications where the availability of the allowance is constrained because of limits on natural resources used in production or of negative environmental externalities associated with production. In such cases, putting a modest price on the natural resource or a modest penalty on the environmental externality can significantly reduce the corresponding usage. Figure 2.5 illustrates these effects for an example system. In this example, an increase in price from 0 to 0.1 leads to a 25% drop in the allowance amount, but remarkably only a 0.1% increase in the total cost.

Next, we examine how expected *allowance usage* is affected by the price of the allowance, where expected allowance usage refers to the ratio of the expected cumulative amount produced over the entire compliance period to the amount of the allowance purchased. This notion is useful because it indicates the degree

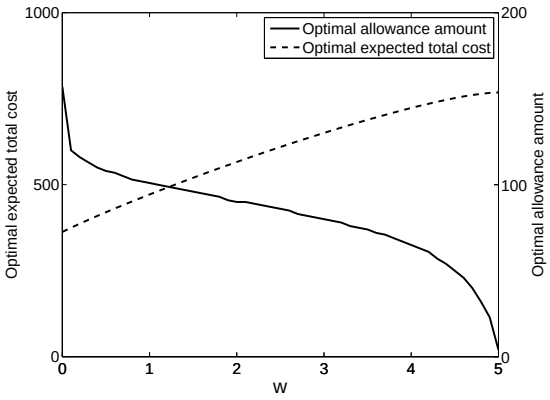


Figure 2.5: Impact of Allowance Price on Optimal Allowance Amount and Total Cost ($T = 12$, $x_T = 0$, $p = 3$, $h = 1$, $l = 8$, $\alpha = 1$)

to which some allowance is wasted or goes unused. For example, in settings where the allowance amount purchased corresponds to the amount of natural resources purchased, a high expected allowance usage indicates that most of the resources purchased would on average be used. This is desirable if the resource is in short supply and allocating allowances for one firm might be at the detriment of allocating the allowances to other firms.

Let $Q_t(x, c)$ denote the expected cumulative amount produced, under the optimal policy, from period t to the end of the planning horizon with starting inventory level x and remaining allowance c. Then, $Q_t(x, c)$, for $t = T, ..., 1$, can be computed recursively as

$$Q_t(x, c) = \int_0^\infty Q_{t-1}((x - \xi)^+, c)\phi(\xi)d\xi,$$

if $x \geq y_t^*(x, c)$, and

$$\bar{y}_t(x + c) - x + \int_0^\infty Q_{t-1}((y_t^*(x, c) - x - \xi)^+, x + c - y_t^*(x, c))\phi(\xi)d\xi,$$

otherwise; with $Q_0(x, c) = 0$. Let $u_T(w, x)$ denote the expected allowance usage.

Then,

$$u_T(w,x) = \frac{Q_T(x, c_T^*(w,x))}{c_T^*(w,x)}.$$

Proposition 10. $\frac{\partial u_T}{\partial w}(w,x)|_{w=0} = \infty$.

Proof of Proposition 10: We can easily prove that, for $1 \le t \le T$, $\bar{g}_t'(z) < 0$ when $z < t\tilde{y}$, and $\bar{g}_t'(z) = 0$ when $z \ge t\tilde{y}$. We can then show that $c_T^*(0,x) = \begin{cases} T\tilde{y} - x & \text{if } x < \tilde{y}, \\ (T-1)\tilde{y} & \text{otherwise,} \end{cases}$ and $\frac{\partial Q_T}{\partial c}(x,c)|_{c=c_T^*(0,x)} = 0$. Therefore, using the results from Proposition 9, we have

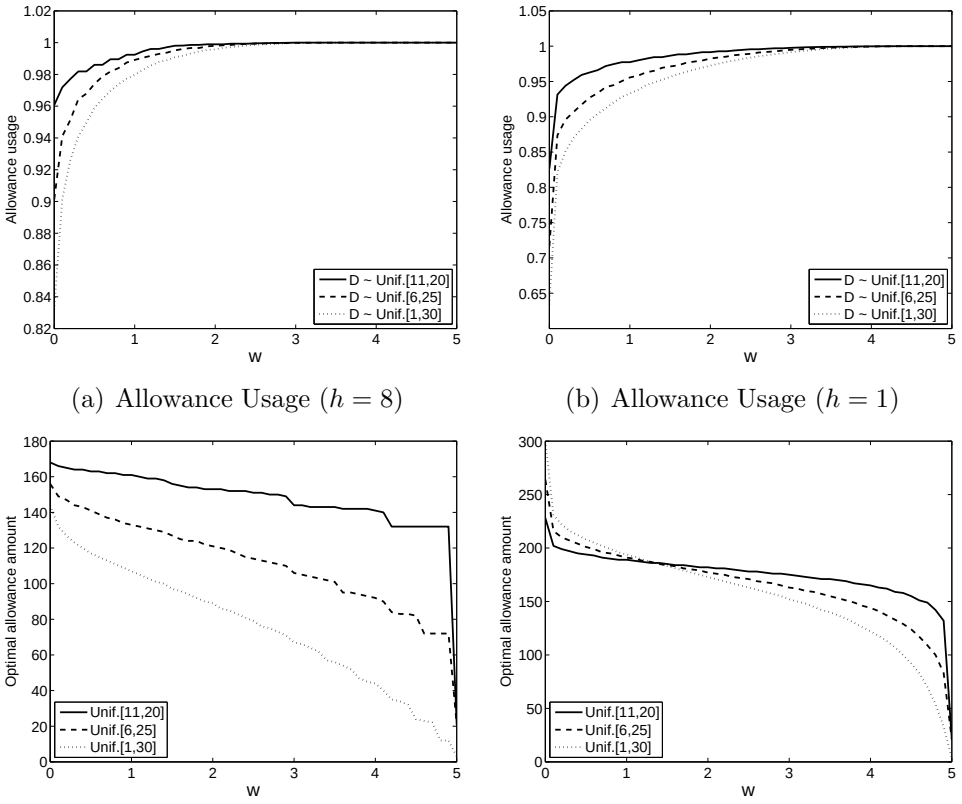
$$\begin{aligned} &\frac{\partial u_T}{\partial w}(w,x)|_{w=0} \\ &= \frac{\frac{\partial Q_T}{\partial c}(x,c)|_{c=c_T^*(w,x)} \frac{\partial c_T^*}{\partial w}(w,x) c_T^*(w,x) - \frac{\partial c_T^*}{\partial w}(w,x) Q_T(x, c_T^*(w,x))}{(c_T^*(w,x))^2}|_{w=0} \\ &= \infty. \end{aligned}$$

□

Proposition 10 implies that if the resource supplier charges a modest price per unit of the resource, then the resource increases significantly. In other words, resources would be used much more efficiently if they are not available for free, even if the associated price is very low. Figures 2.6(a) and 2.6(b) illustrate this effect.

The results in Figures 2.6 are shown for demand distributions with the same mean but different variances, and with the same lost-sales cost but different values of holding cost. We do so to illustrate the joint effect of demand variability and the ratio of holding cost to lost-sales costs.

First, regarding the optimal allowance amount, from Figures 2.6(c) and 2.6(d), we see that when h is large (or l is small), higher demand variability always leads



(a) Allowance Usage ($h = 8$)

(b) Allowance Usage ($h = 1$)

(c) Optimal Allowance Amount ($h = 8$)

(d) Optimal Allowance Amount ($h = 1$)

Figure 2.6: Impact of Allowance Price and Demand Variability ($T = 12$, $x_T = 0$, $p = 3$, $l = 8$, $\alpha = 1$)

to lower optimal allowance amount. However, when h is small, higher demand variability leads to higher optimal allowance amount for small value of w, but lower optimal allowance amount for large value of w. Notice that for the unconstrained version of the problem, which corresponds to $w = 0$, higher demand variability always leads to lower (higher) production quantity for large (small) values of h. However, this is not the case in Figure 2.6(d), and the reason is that, larger w leads to lower unit profit and therefore relatively lower penalties for shortage.

Next, from Figures 2.6(a) and 2.6(b), higher variability always leads to lower expected allowance usage regardless of the price of allowance. This is perhaps surprising given the above observation that higher variability does not always lead to lower investment in allowance. This result is due to the fact that higher variability affects in the same way both the optimal allowance amount and the expected cumulative amount produced. Moreover, note that the percentage increase in allowance usage due to an initial increase in the allowance price is always higher with higher demand variability. The result suggests that initial increases in allowance price can be particularly desirable when demand variability is high.

Chapter 3

Service Systems with Finite and Heterogeneous Customer Arrivals

3.1 Introduction

This work is motivated by systems where a finite number of customer arrivals occur over a period of time followed by few or no arrivals for an extended period thereafter. During the period over which arrivals take place, inter-arrival times between consecutive customers can be different and so can be their service times. Examples of such systems are numerous.

Consider, for example, settings where arrivals are triggered by the start of an event or a service (e.g., the arrival of passengers to check-in for or to board a flight), the total number of arrivals is finite (and determined by the number of tickets sold). Passengers may belong to different classes (e.g., early, on-time, and late) or are assigned to different groups (e.g., priority boarding zones), so

that arrivals occur in waves with each wave drawing from the population of the corresponding class or group.

Another example is one where a finite number of jobs go through a sequence of production stages. The arrival process to each stage (other than the first one) corresponds to the departure process from the preceding one. Because production times at a particular stage are stochastic and can vary in distribution from job to job, the inter-arrival times to the subsequent stage are also stochastic and vary from job to job.

A third example is one where arrivals are driven by appointments (e.g., patient appointments at a health clinic). Assuming customers are punctual (or nearly punctual), inter-arrival times coincide with time between appointments. Depending on how appointments are scheduled, the inter-arrival times between customers can vary. For example, spacing appointments equally leads to uniform inter-arrival times, while other rules, such as those that schedule more appointments at the beginning and at the end, and fewer in between, lead to increasing and then decreasing inter-arrival times.

All the above examples share four common characteristics: (1) a finite number of customers; (2) heterogeneous (and possibly stochastic) inter-arrival times; (3) heterogeneous (and possibly stochastic) service times; and (4) inter-arrival and service times that depend on the position of the customers in the arrival process.

Accounting for heterogeneity in arrival and service times is important in settings where inter-arrival and service times exhibit distinctive features that make it difficult to justify the common assumption of identically distributed inter-arrival and service times. Such features include (1) arrivals that decrease in intensity with

each subsequent arrival; (2) arrivals that increase in intensity with each subsequent arrival; and (3) arrivals that exhibit the combinations of both the increasing and decreasing features. They also include (1) service times that increase with each subsequent service completion, typical of settings where servers are subject to fatigue; (2) service times that decrease with each subsequent service completion, typical of systems where learning takes place; and (3) service times that exhibit the combinations of both the increasing and decreasing features (e.g., initial learning by the servers that is followed by eventual fatigue).

The modeling and analysis of systems with finite arrivals and varying inter-arrival and service times raise several important questions: (1) What is the impact of different inter-arrival and service time features on system performance (for example, does system performance deteriorate with increased heterogeneity in inter-arrival or service times)? (2) For a fixed number of arrivals, are there features which lead to better performance than others (for example, given a target time window for arrivals, is it best to have more arrivals early on, in the middle, or at the end of the arrival time window)? (3) How are the answers to the above questions affected by other problem parameters such as the overall arrival intensity and the total number of arrivals (for example, do higher levels of the parameters favor certain arrival features over others)? (4) Does the heterogeneity in service times affect performance the same way that the heterogeneity in inter-arrival times does, or are there fundamental differences between these two?

In this work, we address these and other related questions. In particular, we consider a system with a finite number of arrivals, where the inter-arrival time between the m^{th} and $(m+1)^{th}$ customer is described by a random variable that has

a general distribution which can be different from the distributions that describe the inter-arrival times between other consecutive customers. Customer service times are described by exponential distributions; however, the mean service times (or service rates) of different customers can be different. We consider systems with both single and multiple servers. Using an embedded Markov chain approach, in each case, we are able to characterize analytically the probability distribution of the number of customers seen by each arrival. This allows us to characterize the waiting time distribution for each customer, from which we obtain various performance measures of interest, including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all customers (makespan). These characterizations further simplify for several special cases of interest, including systems with exponential and deterministic inter-arrival times.

We carry out extensive numerical experiments to examine the effects of heterogeneity in inter-arrival and service times. In particular, we examine cases where, with each subsequent arrival or service completion, inter-arrival and service times (1) increase, (2) decrease, (3) increase and then decrease, or (4) decrease and then increase. We derive several managerial insights and discuss implications for settings where such features can be induced. We validate the numerical results using a fluid approximation that yields closed form expressions. Some of our key findings are highlighted below:

- Arrival processes with different features can lead to significantly different expected waiting times. There is a considerable difference in performance between systems with homogeneous inter-arrival times and those with

heterogeneous inter-arrival times. Therefore, ignoring the heterogeneity in arrival process can lead to significant errors in performance evaluation.

- Arrival processes with homogeneous inter-arrival times may not lead to the lowest waiting time. In fact, for a wide range of parameter values, systems with homogenous inter-arrival times perform poorly.

- Although there is no strict ordering in terms of performance among the arrival processes considered, for systems with homogeneous service times, arrival processes where inter-arrival times decrease, or increase and then decrease, lead to lower waiting time than those where inter-arrival times increase, or decrease and then increase, suggesting that it is generally better to postpone the busy (or peak) period.

- When inter-arrival times are homogeneous, systems in which customers with short service times arrive early (at the beginning of the arrival period) have lower waiting time than those in which such customers arrive later. This is perhaps consistent with results about the optimality of processing customers with shorter processing times first. However, this is not true when inter-arrival times are heterogeneous.

- Inter-arrival and service time features that lead to lower waiting time may not lead to lower makespan.

These insights show that there might be opportunities for system managers to improve system performance by inducing certain arrival features and by differentiating between customers or jobs with different service requirements. We illustrate how arrivals could be affected using two examples. The first one involves

the sequencing of a finite number of jobs through two production stages in series. The second one involves the grouping of passengers into multiple boarding zones. For systems where arrivals cannot be controlled, we examine how arrival processes with different features affect the capacity needed to guarantee a specified level of performance (e.g., a maximum expected waiting time or makespan).

3.2 Related Literature

Although systems with a finite number of arrivals and distinct features in inter-arrival or service times are prevalent and perhaps even pervasive in practice, they have received relatively little attention in the service operations management literature (and more generally in the broader queueing literature). This appears to be, in part, due to the difficulty of analyzing these systems using standard queueing methodology which relies on steady state analysis (and therefore assumes an infinite number of arrivals) or requires homogenous inter-arrival and service times (see, e.g., Kleinrock 1975, Hall 1991).

There is an extensive literature that deals with finite population systems (see, e.g., Takagi 1993, Haque and Armstrong 2007). However, in that case, the finite population of customers cycles indefinitely through two phases of not needing service and needing service (e.g., machines that require repairs). The analysis typically assumes homogeneity in both arrival and service processes. Hence, this literature does not capture the essential features of the problem we consider here.

There is also an extensive literature on systems with time-dependent/state-dependent arrival or service processes (see, e.g., Courtois and Georges 1971, Ross 1978, Green et al. 1991) where the arrival

or service rates may depend on either time, the number of customers in system, or the evolution of certain exogenous stochastic processes. This literature does not capture the settings we describe here where inter-arrival and service times depend on the order in which a particular customer arrives to the system and where the number of customers is finite.

The literature which is most related to ours is the one on transient analysis of queueing systems (see, e.g., Kelton and Law 1985, Parthasarathy and Moosa 1989, Griffiths et al. 2006). However, this literature typically assumes homogenous inter-arrival and service time distributions and the existing results are for systems with Markovian arrivals. Other related papers include Hu and Benjaafar (2009), which treats a special case of our problem where all customers arrive at once (they refer to this as the rush hour regime). Parlar and Moosa (2008) also consider a special case of our problem where the arrivals are Markovian and determined by a pure death process so that the arrival rates are linearly decreasing. In our case, we allow for non-Markovian arrivals and arbitrary arrival rates. Hassin and Mendel (2008) consider a system with a single server and finite arrivals, but customer arrivals are determined by appointment times. Customers are assumed to be punctual and therefore there is no uncertainty regarding arrival times. The service times are exponentially and identically distributed.

There is an extensive body of literature in the area of scheduling which shares features of the problem we consider in this work; namely a finite number of customers (or jobs) that are processed through one or more machines. The jobs are available for processing at specified release times. Jobs may vary in their processing times, delay costs, and due dates. In some cases the release

and service times are stochastic. The focus of much of this literature is, on developing efficient algorithms for generating optimal job sequences, or on identifying structural properties of optimal sequences; see Pinedo (2012) and Emmons and Vairaktarakis (2013) for a discussion of important results and a review of relevant literature. Some of the literature treats the *online* version of the problem where jobs arrive over time and a decision on which job to process next is made with each job arrival and job completion (in the case where preemption is allowed); see, for example, Chou et al. (2006), Chen and Shen (2007), and Ouelhadj and Petrovic (2009). This literature is generally not concerned with developing performance evaluation models as we are in this work.

Finally, there is a growing body of literature which deals with the scheduling of appointments, particularly in healthcare settings. A review of this literature can be found in Preater (2001) and Cayirli and Veral (2003). We also refer the reader to Mondschein and Weintraub (2003), Gupta and Denton (2008), and Jouini et al. (2014). Most of this literature assumes that customers are punctual and the objective is to identify the optimal spacing between appointments where the optimality is determined by a weighted measure of patient's delay, physician's idleness, and tardiness. Note that when customers are punctual and service times are exponential, the performance of a specified schedule can be evaluated using the approach described in this work.

Some of the literature considers no-shows which introduces a particular form of stochasticity in patients' inter-arrival times. For example, Kaandorp and Koole (2007) develop a local search algorithm to identify optimal schedules in the presence of no shows and show that a so-called *dome-shaped* form where more

appointments are scheduled at the beginning and at the end of the schedule, is particularly effective (see related discussion in Section 3.7). Zeng et al. (2010) extend Kaandorp and Koole (2007) to include heterogeneous no-show rates. Koeleman and Koole (2012) also generalize the model by considering both scheduled and emergency arrivals. Some recent papers consider patient scheduling based on an open access model with same day appointments; see Robinson and Chen (2010) and the references therein.

The rest of the chapter is organized as follows. In Section 3.3, we describe the model and provide analysis for the single server system. In Section 3.4, we extend the analysis to the multi-server case. In Section 3.5, we present numerical results and discuss insights. In Section 3.6, we describe the fluid approximation. In Section 3.7, we discuss example applications.

3.3 Problem Description and Analysis

We consider a queueing system with a single server and a finite number of customers arriving randomly over time. The total number of customers is M. We index customers by the order of their arrivals, so that customer m for $m = 1, ..., M$, is the m^{th} customer to arrive. The inter-arrival time between customer $m-1$ and customer m has a general distribution with a finite mean $\frac{1}{\lambda_m}$ for $m = 2, ..., M$. No other specific assumptions are made concerning inter-arrival times except that they are independent. Customer service times are independent and exponentially distributed with a strictly positive and finite mean $\frac{1}{\mu_m}$ for customer m. We make the exponential assumption regarding the distribution of service times for mathematical tractability, as it allows us to formulate the problem as an embedded

Markov chain. This assumption is also useful in approximating the behavior of systems where service time variability is high. Doing away with this assumption without losing tractability is difficult, given the generality of the model otherwise (i.e., the heterogeneity in inter-arrival and service times). Upon arrival, a customer goes immediately into service if the server is available. If not, the customer joins the queue and waits. Customers waiting in queue are served on a first-come, first-served (FCFS) basis.

Note that the inter-arrival and service times are indexed by the position of the customer in the arrival sequence ($m = 1, ..., M$) and not by time, as in a time-dependent process. This is because we are interested in settings, such as the ones we describe in Section 3.1, where the characteristics of the arrival and service processes are affected by the number of customers that have already arrived and not by the amount of time that has already elapsed. This is apparent for example when customers, who are drawn from a finite population, arrive independently from each other, when arrivals correspond to service completions from a preceding process, or when service times are affected by the number of customers previously processed, as in situations in which learning and fatigue can take place.

We are interested in characterizing customer waiting time. Our approach consists of first computing the probabilities of the system states seen by a new arrival. We then compute the conditional waiting time, given the system state. Finally, we characterize the unconditional waiting time by averaging over all possibilities. We denote A_m as the random variable that describes the arrival time of customer m, and R_m as the random variable that describes the number of customers found in system by customer m, upon her arrival at A_m. This means

that the total number of customers in system immediately after A_m is R_m+1. We let $p_{m,i} = \Pr\{R_m = i\}$ refer to the probability that the m^{th} customer finds, upon arrival, i customers already in system (in queue or in service) for $i = 0, ..., m-1$ and $m = 1, ..., M$.

In what follows, we first characterize the probabilities $p_{m,i}$. Let T_m be the random variable describing the inter-arrival time between customers $m-1$ and m, and let $f_m(.)$ be its probability density function. We have $T_m = A_m - A_{m-1}$ for $m = 2, ..., M$. Without loss of generality, we assume the first customer arrives at time 0 ($T_1 = 0$). For $m = 1$, we have $p_{1,0} = 1$ and $p_{1,i} = 0$ for $i \neq 0$, because the first customer always finds the system empty. For $2 \leq m \leq M$, we separate the two cases, $1 \leq i \leq m-1$ and $i = 0$. Let us first consider the case $1 \leq i \leq m-1$. Conditioning on the number of customers found, upon arrival, by customer $m-1$, we obtain

$$p_{m,i} = \sum_{j=i-1}^{m-2} p_{m-1,j} \Pr\{R_m = i \mid R_{m-1} = j\} \tag{3.1}$$

for $2 \leq m \leq M$. Note that we must have $i-1 \leq j \leq m-2$. Let us now characterize the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $1 \leq i \leq m-1$ and $i-1 \leq j \leq m-2$. We again separate the analysis into two cases, $i \leq j \leq m-2$ and $j = i-1$. Firstly, when $i \leq j \leq m-2$, for customer m to find i customers given that customer $m-1$ finds j, there must be exactly $j-i+1$ service completions during the time period $(A_{m-1}, A_m]$. It is easy to see that the $j-i+1$ customers who have finished their service are customers $m-j-1, m-j, ..., m-i-1$, and the one under service at time A_m is customer $m-i$. Let us define $B_{m,i,j}$ as the random variable describing the total duration of those $j-i+1$ service completions, and let $f_{B_{m,i,j}}(.)$ and $F_{B_{m,i,j}}(.)$ be its probability density function and cumulative distribution function,

respectively. Noting that the underlying process is a pure death process, we can see that $B_{m,i,j}$ is equal to the summation of exponential random variables, and thus, it is hypoexponentially distributed with parameters $\mu_{m-j-1}, \mu_{m-j}, ..., \mu_{m-i-1}$. From Ross (2009), we have (in the case where all the rates are distinct) $f_{B_{m,i,j}}(t) = \sum_{l=m-j-1}^{m-i-1} \mu_l o_{m,i,j,l} e^{-\mu_l t}$ and $F_{B_{m,i,j}}(t) = 1 - \sum_{l=m-j-1}^{m-i-1} o_{m,i,j,l} e^{-\mu_l t}$ for $t \geq 0$, where $o_{m,i,j,l} = \prod_{n=m-j-1,\, n\neq l}^{m-i-1} \frac{\mu_n}{\mu_n - \mu_l}$. (By convention, an empty product equals 1.) We denote by ε_{m-i} the exponential random variable that describes the service time of the $(m-i)^{th}$ (yet to complete service) customer, and let $f_{\varepsilon_{m-i}}(.)$ be its probability density function, then we have $f_{\varepsilon_{m-i}}(t) = \mu_{m-i} e^{-\mu_{m-i} t}$ for $t \geq 0$. Let us now define the random variable $C_{m,i,j}$ by $C_{m,i,j} = B_{m,i,j} + \varepsilon_{m-i}$. One may easily see that $\Pr\{R_m = i \mid R_{m-1} = j\} = \Pr\{B_{m,i,j} < T_m < C_{m,i,j}\}$. Due to the independence between T_m, $B_{m,i,j}$ and ε_{m-i}, we have

$$\begin{aligned}
&\Pr\{R_m = i \mid R_{m-1} = j\} \\
&= \mu_{m-i} \sum_{l=m-j-1}^{m-i-1} \mu_l o_{m,i,j,l} \int_0^\infty \int_0^\infty \int_y^{y+z} f_m(x) e^{-\mu_l y - \mu_{m-i} z} dx dy dz
\end{aligned}$$

for $i \leq j \leq m-2$. Similarly, for $j = i-1$, we have

$$\Pr\{R_m = i \mid R_{m-1} = i-1\} = \mu_{m-i} \int_0^\infty \int_0^z f_m(x) e^{-\mu_{m-i} z} dx dz,$$

which leads to

$$\begin{aligned}
p_{m,i} = \mu_{m-i} &\sum_{j=i}^{m-2} \sum_{l=m-j-1}^{m-i-1} \mu_l p_{m-1,j} o_{m,i,j,l} \int_0^\infty \int_0^\infty \int_y^{y+z} f_m(x) e^{-\mu_l y - \mu_{m-i} z} dx dy dz \\
&+ p_{m-1,i-1} \mu_{m-i} \int_0^\infty \int_0^z f_m(x) e^{-\mu_{m-i} z} dx dz \qquad (3.2)
\end{aligned}$$

for $1 \leq i \leq m-1$. As for the quantity $p_{m,0}$, it is simply given by

$$p_{m,0} = 1 - \sum_{i=1}^{m-1} p_{m,i} \qquad (3.3)$$

for $2 \leq m \leq M$. Using Equations (3.2) and (3.3), the probabilities $p_{m,i}$ for $1 \leq m \leq M$ and $0 \leq i \leq m-1$ can be recursively computed starting with $m = 1$.

Next we show how the above probabilities can be used to characterize various performance measures. Let X_m, a random variable, denote the waiting time in queue of customer m, and let $E(X_m^k)$ be the corresponding k^{th} moment for $k \geq 1$. (For the rest of the chapter, we use $E(Z^k)$ to denote the k^{th} moment of a random variable Z for $k \geq 1$.) Note that $X_1 = 0$ with probability 1, since it corresponds to the waiting time of the first customer. For $2 \leq m \leq M$, we have

$$E(X_m^k) = \sum_{i=1}^{m-1} p_{m,i} E(X_{m,i}^k),$$

where $X_{m,i}$ is the conditional random variable denoting the waiting time in queue for customer m, given that customer m finds, upon arrival, i customers in system. Obviously, $X_{m,0} = 0$ with probability 1. For $1 \leq i \leq m-1$, the i customers seen by the m^{th} arrival are customers $m-1, m-2, ..., m-i$. For the $(m-i)^{th}$ customer who is currently in service, the remaining service time is still exponentially distributed with rate μ_{m-i}. Since their service times are independent and exponentially distributed, $X_{m,i}$ has a hypoexponential distribution with parameters $\mu_{m-1}, \mu_{m-2}, ..., \mu_{m-i}$. Hence, the quantities $E(X_{m,i}^k)$ for $k \geq 1$ can be easily computed. For example, we have $E(X_{m,i}) = \sum_{l=m-i}^{m-1} \frac{1}{\mu_l}$ and $E(X_{m,i}^2) = \sum_{l=m-i}^{m-1} \frac{1}{\mu_l^2} + (\sum_{l=m-i}^{m-1} \frac{1}{\mu_l})^2$.

Let the random variable X denote the waiting time in queue of an arbitrary customer among the M ones. Then, we obtain $E(X^k) = \frac{1}{M} \sum_{m=2}^{M} E(X_m^k) = \frac{1}{M} \sum_{m=2}^{M} \sum_{i=1}^{m-1} p_{m,i} E(X_{m,i}^k)$ for $k \geq 1$. In particular, we have

$$E(X) = \frac{1}{M} \sum_{m=2}^{M} \sum_{i=1}^{m-1} \sum_{l=m-i}^{m-1} \frac{p_{m,i}}{\mu_l}$$

and

$$Var(X) = \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1} p_{m,i}\left[\sum_{l=m-i}^{m-1}\frac{1}{\mu_l^2} + \left(\sum_{l=m-i}^{m-1}\frac{1}{\mu_l}\right)^2\right] - \frac{1}{M^2}\left(\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_l}\right)^2.$$

From the probabilities $p_{m,i}$, we can also characterize the distribution of X. Specifically, $\Pr\{X \le t\} = \frac{1}{M}(1 + \sum_{m=2}^{M}\Pr\{X_m \le t\}) = \frac{1}{M} + \frac{1}{M}\sum_{m=2}^{M}(p_{m,0} + \sum_{i=1}^{m-1} p_{m,i}\Pr\{X_{m,i} \le t\})$ for $t \ge 0$. In case all the rates are distinct, we have

$$\Pr\{X \le t\} = 1 - \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1} p_{m,i}o_{m,0,i-1,l}e^{-\mu_l t}.$$

In addition to waiting time, an important performance measure for systems with finite arrivals is *makespan*, namely, the time it takes the system to complete serving all customers. Since the server starts working at time zero, makespan can be computed as the departure time of the last customer (customer M). We define D_m as the random variable describing the departure time of customer m. Then $D_M = A_M + X_M + \varepsilon_M$, which leads to

$$E(D_M) = E(A_M) + E(X_M) + E(\varepsilon_M) = \sum_{m=2}^{M}\frac{1}{\lambda_m} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_l} + \frac{1}{\mu_M}.$$

Other measures of interest, such as those discussed in Cayirli and Veral (2003), can also be easily obtained. For example, the expected total time in system (waiting time + service time) for an arbitrary customer is given by $\frac{1}{M}(\sum_{m=1}^{M} E(X_m) + \frac{1}{\mu_m})$, or equivalently $\frac{1}{M}\sum_{m=1}^{M}\frac{1}{\mu_m} + \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_l}$; the expected server idle time is given by $E(D_M) - \sum_{m=1}^{M}\frac{1}{\mu_m}$, or equivalently $\sum_{m=2}^{M}\frac{1}{\lambda_m} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_l} - \sum_{m=1}^{M-1}\frac{1}{\mu_m}$; and the expected server

utilization is given by $(\sum_{m=1}^{M} \frac{1}{\mu_m})/E(D_M)$, which can also be rewritten as $(\sum_{m=1}^{M} \frac{1}{\mu_m})(\sum_{m=2}^{M} \frac{1}{\lambda_m} + \sum_{i=1}^{M-1} \sum_{l=M-i}^{M-1} \frac{p_{M,i}}{\mu_l} + \frac{1}{\mu_M})^{-1}$. Various service level measures can also be obtained, including the probability that a customer waits more than a specified threshold or that makespan exceeds a certain threshold.

In some applications where the arrival process can be controlled, another useful performance measure is the amount of time, starting from time zero, until a customer arrives. This can be viewed as the indirect or offline waiting time. The expected arrival time of an arbitrary customer is given by $\frac{\sum_{m=2}^{M} \sum_{i=2}^{m} E(T_i)}{M}$.

Next, we consider three special cases for which the analysis simplifies further. **The Case of Exponential Inter-arrival Times:** In this case, computing the probability $p_{m,i}$ simplifies by noting that, the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $1 \leq i \leq m-1$ and $i-1 \leq j \leq m-2$, can now be expressed as

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1} \frac{\mu_{m-l}}{\mu_{m-l} + \lambda_m} \right) \frac{\lambda_m}{\mu_{m-i} + \lambda_m}. \tag{3.4}$$

The Case of Deterministic Inter-arrival Times: In this case, T_m is constant and equal to $\frac{1}{\lambda_m}$ for $2 \leq m \leq M$. The probability density function $f_m(t)$ is now a Dirac delta function at $\frac{1}{\lambda_m}$, which leads to

$$\Pr\{R_m = i \mid R_{m-1} = j\} = e^{-\frac{\mu_{m-i}}{\lambda_m}} \sum_{l=m-j-1}^{m-i-1} o_{m,i,j,l} \frac{\mu_l}{\mu_{m-i} - \mu_l} (e^{\frac{\mu_{m-i} - \mu_l}{\lambda_m}} - 1)$$

for $i \leq j \leq m-2$, and $\Pr\{R_m = i \mid R_{m-1} = i-1\} = e^{-\frac{\mu_{m-i}}{\lambda_m}}$.

The case of deterministic inter-arrival times is of interest in applications where arrivals are determined by appointments and customers are punctual. In this case, arrival times correspond to appointment times. Note that the above allows for heterogeneous service time distributions and generalizes earlier treatments

that consider service times with homogenous rates (see, e.g., Kaandorp and Koole 2007, Hassin and Mendel 2008).

The Case of Instantaneous Arrivals: An extreme case of the arrival process is one where customers arrive all at once. In this case, the expected waiting time of the m^{th} customer corresponds to the sum of the expected service times of customers $1, 2, ..., m-1$, i.e. $E(X_m) = \sum_{i=1}^{m-1} \frac{1}{\mu_i}$. This leads to $E(X) = \frac{1}{M} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \frac{1}{\mu_l}$ and $E(D_M) = \sum_{m=1}^{M} \frac{1}{\mu_m}$.

3.4 Multi-Server Case

In this section, we consider the case of a queueing system with multiple servers. We assume that there are s parallel and identical servers. For tractability, we focus on the case where service times are independent and exponentially distributed with rate μ. An arriving customer immediately begins service if there is an available server. Otherwise, she waits in queue and will be served by the first available server. All other assumptions are the same as those for the single server case in Section 3.3, and we continue to use similar notations.

As in the single server case, let us first characterize the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $2 \leq m \leq M$, $1 \leq i \leq m-1$, and $i \leq j \leq m-2$. For customer m to find i customers given that customer $m-1$ finds j customers, there must exactly be $j - i + 1$ service completions during the time period $(A_{m-1}, A_m]$. We distinguish the following three cases.

Case 1, $s \leq i \leq j+1$**:** Once customer $m-1$ arrives, she joins the queue (if $j+1 > s$) or occupies the last available server (if $j+1 = s$). In both cases, customer m joins the queue once she arrives, and all the servers are busy during

the time period $(A_{m-1}, A_m]$. When all servers are busy, the departure process is Poisson with rate $s\mu$. The probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that $j - i + 1$ customers finish their service during T_m. So we may write

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \int_0^\infty \frac{(s\mu x)^{j-i+1}}{(j-i+1)!} e^{-s\mu x} f_m(x) dx.$$

Case 2, $1 \le i \le j + 1 < s$**:** In this case, there is no queue. Both customer $m-1$ and m immediately enter service once they arrive, and $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that exactly $j-i+1$ among $j+1$ customers finish their service during T_m. Noticing that $\binom{j+1}{j-i+1} = \binom{j+1}{i}$, this leads to

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \int_0^\infty \binom{j+1}{i} (1 - e^{-\mu x})^{j-i+1} e^{-\mu x i} f_m(x) dx.$$

Case 3, $1 \le i < s \le j+1$**:** In this case, the system starts busy with $j-s+1$ queued customers immediately after A_{m-1}. The probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that, within T_m, the first $j-s+1$ queued customers leave the queue and enter service (which implies that $j-s+1$ customers finish their service) and then $s-i$ customers finish their service afterwards, i.e., $j-i+1$ service completions in total. We denote by I the random variable that describes the time needed to complete those $j - s + 1$ services, then I has an Erlang distribution with $j - s + 1$ stages and parameter $s\mu$. Thus, the probability density function of I, say $f_I(t)$, is given by $f_I(t) = \frac{(s\mu)^{j-s+1} t^{j-s} e^{-s\mu t}}{(j-s)!}$ for $t \ge 0$. This leads to

$$\begin{aligned}
&\Pr\{R_m = i \mid R_{m-1} = j\} \\
&= \int_0^\infty \int_0^x \binom{s}{i} (1 - e^{-\mu(x-t)})^{s-i} e^{-\mu(x-t)i} \frac{(s\mu)^{j-s+1} t^{j-s} e^{-s\mu t}}{(j-s)!} f_m(x) dt dx.
\end{aligned}$$

As for the single server case, using Equations (3.1) and (3.3), we can obtain $p_{m,i}$ for $2 \le m \le M$ and $1 \le i \le m-1$ recursively.

Having the probabilities $p_{m,i}$ on-hand, we can now compute various performance measures. In particular, we have

$$E(X_m^k) = \sum_{i=s}^{m-1} p_{m,i} E(X_{m,i}^k)$$

for $1 \leq m \leq M$. Obviously $X_{m,i} = 0$ with probability 1 for $i \leq s-1$. For $i \geq s$, $X_{m,i}$ is Erlang distributed with $i-s+1$ stages and parameter $s\mu$. Consequently, we have

$$E(X_m) = \sum_{i=s}^{m-1} p_{m,i} \frac{i-s+1}{s\mu} \tag{3.5}$$

and $E(X_m^2) = \sum_{i=s}^{m-1} p_{m,i} \frac{(i-s+1)(i-s+2)}{s^2\mu^2}$. Higher moments can be similarly computed. Since $E(X_m) = 0$ for $m \leq s$, we have

$$E(X^k) = \frac{1}{M} \sum_{m=s+1}^{M} E(X_m^k).$$

From the cumulative distribution function of Erlang distribution, we obtain $\Pr\{X_{m,i} \leq t\} = 1 - \sum_{l=0}^{i-s} \frac{(s\mu t)^l}{l!} e^{-s\mu t}$ and then $\Pr\{X_m \leq t\} = 1 - \sum_{i=s}^{m-1} \sum_{l=0}^{i-s} p_{m,i} \frac{(s\mu t)^l}{l!} e^{-s\mu t}$. This leads to

$$\Pr\{X \leq t\} = 1 - \frac{1}{M} \sum_{m=s+1}^{M} \sum_{i=s}^{m-1} \sum_{l=0}^{i-s} p_{m,i} \frac{(s\mu t)^l}{l!} e^{-s\mu t}.$$

As in Section 3.3, we can also characterize the makespan. However, in contrast to the single server case, makespan in the multi-server system no longer necessarily coincides with the departure time of customer M. The reason is that, if there are other customers under service at the time when customer M enters service, since service times are random, customer M may finish service and leave the system earlier than someone else. But, note that, although customer M may not be the last one to leave the system, she is still the last one to enter service by assumption

(FCFS). Therefore, makespan equals the sum of the time it takes customer M to enter service, and the time it takes to empty the system after she enters service. When customer M arrives, seeing i customers in system, there are two possibilities. The first possibility is $i \leq s-1$, which implies that there is at least one idle server, and customer M immediately enters service without waiting. In this case, the time to empty the system corresponds to the longest completion time among the $i+1$ services. This time has the hypoexponential distribution with parameters $(i+1)\mu, i\mu, ..., \mu$. Thus, if customer M finds i customers in system upon her arrival and $i \leq s-1$, then the expected makespan is given by $\sum_{m=2}^{M} \frac{1}{\lambda_m} + \sum_{l=1}^{i+1} \frac{1}{l\mu}$.

The second possibility is $i \geq s$, which implies that customer M has to wait in queue before being served. In this case, the waiting time of customer M is Erlang distributed with $i-s+1$ stages and parameter $s\mu$, and the time to empty the system has the hypoexponential distribution with rates $s\mu, (s-1)\mu, ..., \mu$. Thus, if customer M finds i customers in system upon her arrival and $i \geq s$, then the expected makespan is given by $\sum_{m=2}^{M} \frac{1}{\lambda_m} + \frac{i-s+1}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu}$.

Putting it all together, the unconditional expected makespan can be obtained as

$$E[Makespan] = \sum_{m=2}^{M} \frac{1}{\lambda_m} + \sum_{i=0}^{s-1} \left(p_{M,i} \sum_{l=1}^{i+1} \frac{1}{l\mu} \right) + \sum_{i=s}^{M-1} p_{M,i} \left(\frac{i-s+1}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu} \right).$$

Other performance measures can be similarly obtained, and we omit the details for the sake of brevity.

The Case of Exponential Inter-Arrival Times: Using similar arguments as in the single server case and noting that, when there are l customers in system,

the service rate is $\mu\ min(l, s)$, we obtain

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1} \frac{\mu \min(l, s)}{\mu \min(l, s) + \lambda_m} \right) \frac{\lambda_m}{\mu \min(i, s) + \lambda_m}.$$

The Case of Deterministic Inter-Arrival Times: This also follows the approach used for the single server case by setting $f_m(t)$ as a Dirac delta function at $\frac{1}{\lambda_m}$ and then computing $\Pr\{R_m = i \mid R_{m-1} = j\}$ using the corresponding equations.

The Case of Instantaneous Arrivals: In this case, the first s customers have zero waiting time, and customer $s + i$ $(1 \leq i \leq M - s)$ waits for i service completions to start service. This leads to $E(X) = \frac{(M-s)^2+(M-s)}{2s\mu M}$ and $E(D_M) = \frac{M-s}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu}$.

3.5 Numerical Experiments

In this section, we describe results from the numerical experiments we carried out to examine the impact of features that are unique to the systems we consider, namely, the finite number of arrivals, the heterogeneity in inter-arrival times, and the heterogeneity in service times. Our objective is three-fold: (1) to draw insights into how these specific features affect system performance, (2) to show that models which do not explicitly account for these features can lead to significant errors in performance evaluation, and (3) to illustrate how the models we present in this chapter can be used to support operational decision making, particularly as it pertains to capacity planning (see Section 3.7 for discussions on additional applications). In Sections 3.5.1 and 3.5.2, we consider respectively the impact of heterogeneity in inter-arrival times and service times,

on various performance measures. In Section 3.5.3, we consider the joint impact of heterogeneity in inter-arrival and service times. In Section 3.5.4, we discuss the impact of heterogeneity on capacity levels. Throughout this section, we focus on the single server setting. We also studied the multi-server setting and obtained similar results; we omit the details for the sake of brevity.

3.5.1 Impact of Heterogeneity in Inter-Arrival Times

To examine the impact of heterogeneity in inter-arrival times, we investigate five arrival processes with different inter-arrival time features that may arise naturally in practice (see our earlier discussion in Section 3.1). These five processes are described in Table 3.1. To allow for a fair comparison between different processes, we maintain the same number of customers and the same average expected inter-arrival time (equals $\frac{1}{\lambda}$) across processes. The first process corresponds to a setting where the expected inter-arrival times decrease with each subsequent arrival. Specifically, we let $E(T_m) = \frac{M-m+1}{M}\frac{2}{\lambda}$ for $m = 2, ..., M$. The other processes correspond similarly to settings where expected inter-arrival times (1) increase with each subsequent arrival, (2) decrease and then increase, (3) increase and then decrease, and (4) are constant. Note that $\{E(T_m)|m = 2, ..., M\}$ in the four heterogeneous processes are indeed four specific permutations of the sequence $\{\frac{1}{M}\frac{2}{\lambda}, ..., \frac{M-1}{M}\frac{2}{\lambda}\}$.

A representative sample from an extensive set of numerical results on expected waiting time is shown in Figure 3.1. (Additional results are available upon request.) The results are shown for systems where inter-arrival times are exponentially distributed and service times are independently, identically, and

Inter-arrival Time Features	Expected Inter-arrival Times	
Decreasing	$E(T_m) = \frac{M-m+1}{M}\frac{2}{\lambda}$	for $m = 2, ..., M$
Increasing	$E(T_m) = \frac{m-1}{M}\frac{2}{\lambda}$	for $m = 2, ..., M$
Decreasing/Increasing	$E(T_m) = \frac{M-2m+3}{M}\frac{2}{\lambda}$	for $m = 2, ..., \frac{M+2}{2}$
	$E(T_m) = \frac{2m-M-2}{M}\frac{2}{\lambda}$	for $m = \frac{M+4}{2}, ..., M$
Increasing/Decreasing	$E(T_m) = \frac{2m-2}{M}\frac{2}{\lambda}$	for $m = 2, ..., \frac{M}{2}$
	$E(T_m) = \frac{2M-2m+1}{M}\frac{2}{\lambda}$	for $m = \frac{M+2}{2}, ..., M$
Constant	$E(T_m) = \frac{1}{\lambda}$	for $m = 2, ..., M$

Table 3.1: Inter-Arrival Time Features

exponentially distributed. (The results are qualitatively the same for other common inter-arrival time distributions we tested.) Note that by varying λ for fixed M and μ, the workload in system (i.e. the traffic intensity or the utilization of server) over the arrival period, as measured by $\rho = \frac{\lambda}{\mu}$, is varied. On the other hand, by varying M for fixed λ and μ, the workload remains constant, but the period of arrivals, as measured by the expected time until the last customer arrives, is varied.

The following observations can be made regarding system performance in terms of the expected waiting time of an arbitrary customer.

- Arrival processes with different features can lead to significantly different expected waiting times. Moreover, there is a considerable difference between the performance of systems with constant expected inter-arrival times and those with heterogeneous expected inter-arrival times. Clearly, ignoring the heterogeneity in the arrival process can lead to significant errors in performance evaluation.

- Arrival processes with constant expected inter-arrival times does not guarantee better performance. In other words, arrivals with a fixed *intensity*

may not necessarily be preferable to arrivals with variable intensity.

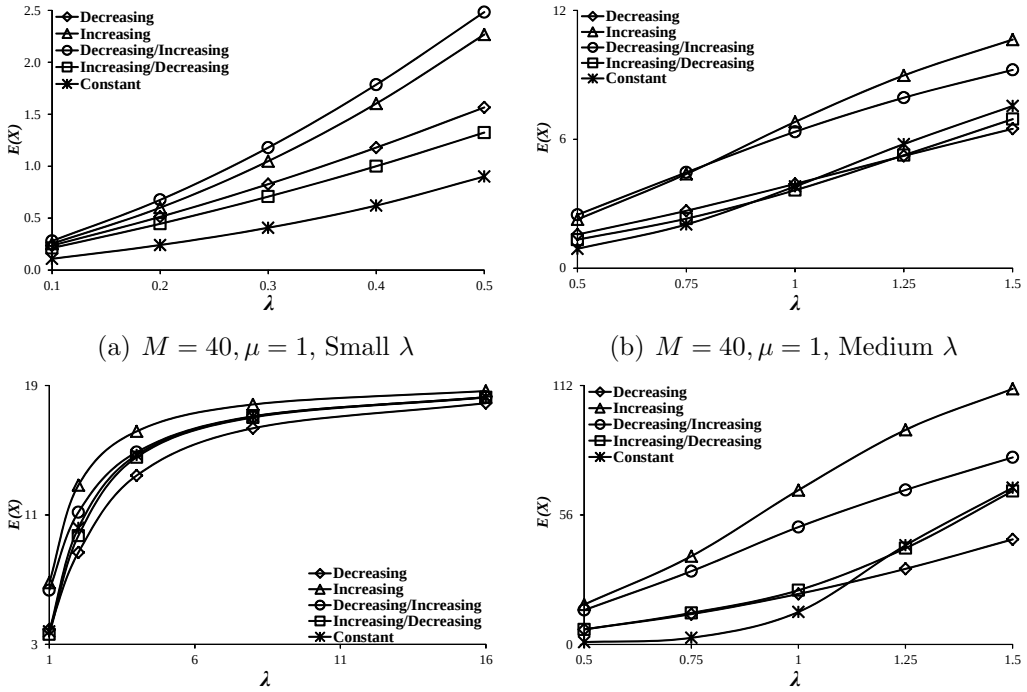
- Arrival processes with "Decreasing" inter-arrival times always perform better than processes with "Increasing" and "Decreasing/Increasing" inter-arrival times. In other words, processes where arrivals peak later leads to better performance than those where arrivals peak earlier. This is due to the fact that a peak in arrivals that occurs early in the process can delay all customers that arrive subsequently.

- The relative performance of different arrival processes depends on problem parameter values. For example, when ρ is small ($\rho << 1$), "Constant" is the best as it spreads out arrivals, reducing the possibility of congestion. On the other hand, when ρ is large ($\rho >> 1$), congestion is inevitable. In that case, arrival processes, with features that can limit the number of customers affected by congestion, become more preferable, explaining, for example why "Decreasing" is the best.

- The difference in performance between different arrival processes decreases as λ increases. The performances become indistinguishable as λ gets very large, in which case, all customers arrive nearly instantaneously.

- The threshold on ρ that determines the relative performance of different arrival processes is affected by M. For example, the larger M is, the larger is the value of ρ under which "Constant" performs the best. In Section 3.6, we provide an approximation that allows us to specify these thresholds in closed form.

In addition to the expected waiting time, we also obtained results for the



(a) $M = 40, \mu = 1$, Small λ

(b) $M = 40, \mu = 1$, Medium λ

(c) $M = 40, \mu = 1$, Large λ

(d) $M = 400, \mu = 1$, Medium λ

Figure 3.1: Impact of Inter-Arrival Time Features on Expected Waiting Time

impact of different arrival processes on the variance of waiting time. For brevity, we omit these results (available upon request) and note the following.

- Most of the observations on expected waiting time continue to hold. For example, arrival processes with different features lead to significantly different variances, with the "Constant" inter-arrival time feature not always leading to the lowest variance. The difference in variances induced by different arrival processes decreases as λ increases, with the threshold on ρ that determines the relative performance of different processes affected by M.

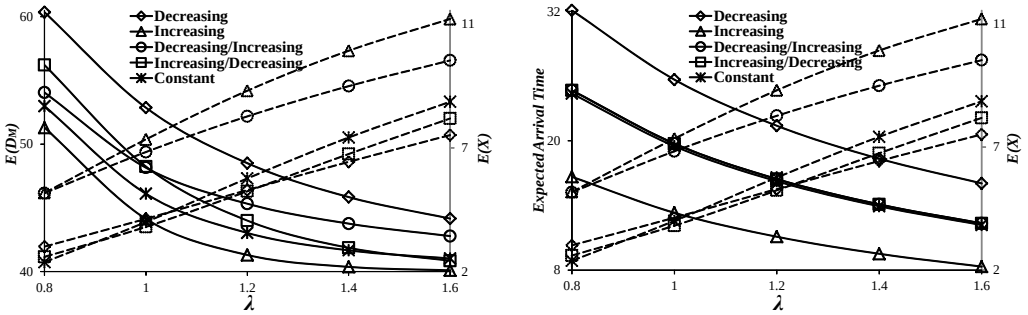
- Systems with "Constant" and "Increasing/Decreasing" inter-arrival times always perform better than the others. In particular, for small ρ, "Constant" performs the best as it smoothes the arrival process and reduces the possibility of congestion. However, for large ρ, congestion is inevitable, and "Increasing/Decreasing" performs the best since it separates the arrival process into two sub-processes with each one having a lower peak value of congestion.

In Figures 3.2(a) and 3.2(b), we present results that illustrate the impact of different arrival processes on the expected makespan and the expected arrival time, with solid lines representing expected makespan and expected arrival time, respectively, and dashed lines representing expected waiting time. Here too, arrival processes with different inter-arrival time features can lead to significantly different expected makespans, with "Constant" not necessarily being the best. While the average expected inter-arrival time stays the same for all processes, makespan is minimized by minimizing the expected waiting time in queue of

the last customer (or equivalently minimizing idleness of the server). This is achieved by maximizing the number of customers that arrive early, explaining why "Increasing" performs the best and "Decreasing" performs the worst. The relative performance of other processes depends on system utilization. For example, when utilization is low, "Decreasing/Increasing" performs better than "Increasing/Decreasing". Although the peak of arrivals occurs later under "Decreasing/Increasing", there is enough capacity in the system to ensure that most customers would clear before the last customer arrives. This is not the case when utilization is high. There, it is preferable to have the peak of arrivals occur as early as possible to minimize the idleness of the server, explaining why "Increasing/Decreasing" is more preferable. Same as for the expected waiting time, the difference in the expected makespan induced by different arrival processes decreases as λ increases. This difference approaches zero as λ becomes large. Similar to the expected makespan, the expected arrival time is lower when more customers arrive earlier. Therefore, the relative performance of different arrival processes on the expected arrival time coincides with the one observed for the expected makespan.

3.5.2 Impact of Heterogeneity in Service Times

In results (the details of which are not shown here for the sake of brevity), we examine the impact of heterogeneity in service times. Here again, we investigate five service processes with different service time features, as shown in Table 3.2. These include settings where expected service times (1) decrease with each subsequent service completion, (2) increase, (3) decrease and then increase, (4)



(a) Makespan and Waiting Time (b) Arrival Time and Waiting Time

Figure 3.2: Impact of Inter-Arrival Time Features on Makespan and Arrival Time ($M = 40$, $\mu = 1$)

increase and then decrease, and (5) are constant. To allow for a fair comparison between different processes, we maintain the same number of customers and the same average expected service time (equals $\frac{1}{\mu}$) across processes.

Service Time Features	Expected Service Times	
Decreasing	$E(\varepsilon_m) = \frac{M-m+1}{M+1}\frac{2}{\mu}$	for $m = 1, ..., M$
Increasing	$E(\varepsilon_m) = \frac{m}{M+1}\frac{2}{\mu}$	for $m = 1, ..., M$
Decreasing/Increasing	$E(\varepsilon_m) = \frac{M-2m+1}{M+1}\frac{2}{\mu}$	for $m = 1, ..., \frac{M}{2}$
	$E(\varepsilon_m) = \frac{2m-M}{M+1}\frac{2}{\mu}$	for $m = \frac{M+2}{2}, ..., M$
Increasing/Decreasing	$E(\varepsilon_m) = \frac{2m}{M+1}\frac{2}{\mu}$	for $m = 1, ..., \frac{M}{2}$
	$E(\varepsilon_m) = \frac{2M-2m+1}{M+1}\frac{2}{\mu}$	for $m = \frac{M+2}{2}, ..., M$
Constant	$E(\varepsilon_m) = \frac{1}{\mu}$	for $m = 1, ..., M$

Table 3.2: Service Time Features

Similar to what we have observed for the arrival process, service processes with different service time features can lead to significantly different expected waiting times, with the "Constant" service time feature again not necessarily being the best. Service processes with features that postpone congestion are preferable when utilization is high ($\rho >> 1$) (e.g., "Increasing" tends to perform the best). This is perhaps also consistent with known results from the scheduling literature

regarding the optimality of the "shortest processing time first" scheduling rule. However, when utilization is low ($\rho << 1$), this is not the case, and "Constant" performs the best for reasons similar to those explained for the arrival process.

With regard to the variance of waiting time, again for the same reasons as explained in the previous section, when utilization is high, "Decreasing/Increasing" performs the best, and when utilization is low, "Constant" performs the best. For expected makespan, the order of preference tends to be reversed, with features that reduce congestion later in the arrival process being preferable (in other words, for the expected makespan, it is preferable that arrivals with shorter service times occur later in the arrival process).

3.5.3 Joint Impact of Heterogeneity in Inter-Arrival and Service Times

In this section, we consider settings where both the inter-arrival and service times are heterogeneous. In particular, we consider the four scenarios shown in Table 3.3. Numerical results for the expected waiting time are shown in Figure 3.3 (results for the expected makespan are omitted for the sake of brevity). We find that combinations of different inter-arrival and service time features lead to significantly different waiting times and makespans. Thus, it is important to explicitly account for both. Again, we find that there are two distinct regimes of operation. When utilization is high ($\rho >> 1$), and a peak in congestion is unavoidable, a combination of inter-arrival and service time features that delays the peak until later in the arrival process reduces expected waiting time the

most, which explains why the combination of "Decreasing" inter-arrival times and "Increasing" service times is most preferable, and the combination of "Increasing" inter-arrival times and "Decreasing" service times is least preferable. This ordering is reversed for makespan. On the other hand, when utilization is low ($\rho << 1$), those combinations such as "Decreasing" inter-arrival times and "Decreasing" service times, and "Increasing" inter-arrival times and "Increasing" service times, which avoid peak congestion, tend to reduce expected waiting time the most.

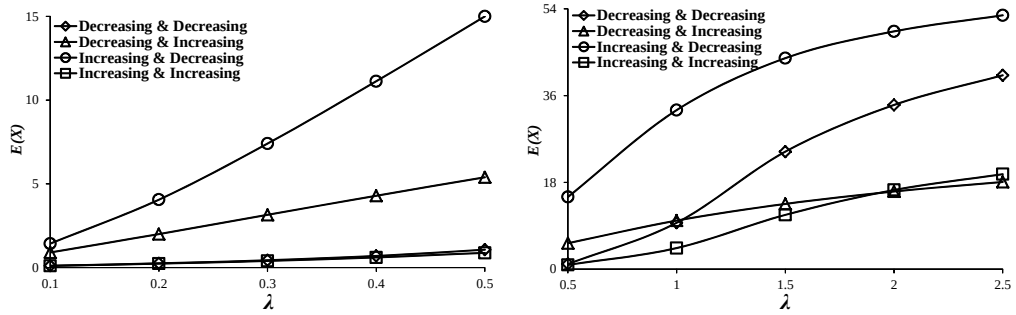
Inter-arrival & Service Time Features	Expected Inter-arrival & Service Times
Inter-arrival Time: Decreasing	$E(T_m) = 2\frac{M-m+1}{M}\frac{1}{\lambda}$ for $m = 2, ..., M$
Service Time: Decreasing	$E(\varepsilon_m) = 2\frac{M-m+1}{M+1}\frac{1}{\mu}$ for $m = 1, ..., M$
Inter-arrival Time: Decreasing	$E(T_m) = 2\frac{M-m+1}{M}\frac{1}{\lambda}$ for $m = 2, ..., M$
Service Time: Increasing	$E(\varepsilon_m) = 2\frac{m}{M+1}\frac{1}{\mu}$ for $m = 1, ..., M$
Inter-arrival Time: Increasing	$E(T_m) = 2\frac{m-1}{M}\frac{1}{\lambda}$ for $m = 2, ..., M$
Service Time: Decreasing	$E(\varepsilon_m) = 2\frac{M-m+1}{M+1}\frac{1}{\mu}$ for $m = 1, ..., M$
Inter-arrival Time: Increasing	$E(T_m) = 2\frac{m-1}{M}\frac{1}{\lambda}$ for $m = 2, ..., M$
Service Time: Increasing	$E(\varepsilon_m) = 2\frac{m}{M+1}\frac{1}{\mu}$ for $m = 1, ..., M$

Table 3.3: Inter-Arrival and Service Time Features

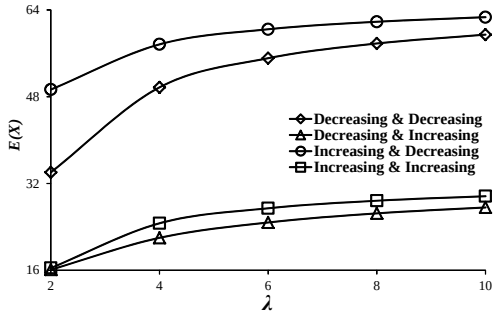
3.5.4 On the Impact on Capacity Levels

In this section, we examine how arrival processes with different features affect the capacity needed to guarantee a specified level of performance (e.g., a maximum expected waiting time or makespan). For single server systems, determining this capacity requires determining the minimum processing rate. For systems with multiple servers, this requires determining the minimum number of servers.

In Figure 3.4, we show the minimum service rate μ needed under each of the four heterogeneous arrival processes described in Table 3.1 to meet a specified minimum expected waiting time target. In this case, the specified target is



(a) $M = 100, \mu = 1$, Small λ (b) $M = 100, \mu = 1$, Medium λ



(c) $M = 100, \mu = 1$, Large λ

Figure 3.3: Impact of Inter-Arrival and Service Time Features on Expected Waiting Time

the expected waiting time obtained under the arrival process with "Constant" inter-arrival times at $\mu = 1$. As we can see, the difference in the capacity levels needed under different arrival processes can be dramatically different. Ignoring the heterogeneity in inter-arrival times (and similarly in service requirements) can therefore lead to significant under or over investments in capacity, resulting in either poor service quality or unjustified additional capacity cost.

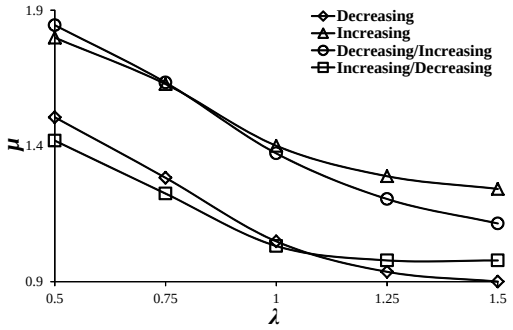


Figure 3.4: Impact of Inter-Arrival Time Features on Capacity Level ($M = 100$)

3.6 Fluid Approximation

Although the performance analysis given in Sections 3.3 and 3.4 is exact, we resorted to numerical analysis in order to draw the conclusions in Section 3.5. This is because the exact results are not in closed form and therefore difficult to use to characterize structural results. To provide further support for the numerical results, we discuss in this section a deterministic *fluid* approximation that does yield closed form expressions and allows us to capture key features of our setting. The objective from this approximation is of course not to substitute for the exact analysis which is easy to implement, but to analytically confirm the

numerical findings of Section 3.5 and provide evidence of their robustness. The approximation may also be useful in investigating additional structural results and as a first step in examining first order effects. The approximation does not require the assumption of exponential service times and, therefore, is useful for the study of more general systems. For the sake of brevity, we describe the approximation in the context of the single server model. However, extending the treatment to the multi-server case is relatively straightforward.

We treat all customer inter-arrival and service times as being deterministic and replace all corresponding random variables by their expected values. (For every quantity Z defined in Section 3.3 for the original model, we define a corresponding quantity Z^F for the fluid approximation). We treat the arrival of customers as fluid, one unit per customer, that is "pumped-in" to the system at a constant rate λ_m over the time period $(A^F_{m-1}, A^F_m]$ for $m = 2, ..., M$. Since $T_1 = 0$ in the original model, we assume all the fluid associated with the first customer is present in the system at time 0. Similarly, we treat the service process as fluid, also one unit per customer, that is "pumped-out" at a constant rate μ_m over the time period $(D^F_{m-1}, D^F_m]$ for $m = 2, ..., M$, and at the rate μ_1 over the time period $(0, D^F_1]$, where $D^F_m = \max(D^F_{m-1}, A^F_m) + \frac{1}{\mu_m}$ with $D^F_1 = \frac{1}{\mu_1}$. By induction, it is straightforward to show that $D^F_m = \max_{1\le i\le m}\{\sum_{j=2}^{i}\frac{1}{\lambda_j} + \sum_{j=i}^{m}\frac{1}{\mu_j}\}$ for $m = 1, ..., M$ (by convention, an empty sum equals 0).

We define $A^F(t)$ and $D^F(t)$ as the cumulative arrivals to the system and the cumulative departures from the system by time t, respectively (with $A^F(0) = 1$). It is not difficult to see that, $A^F(t)$ and $D^F(t)$ are piecewise linear functions (see Figure 3.5 for an illustration). The area between $A^F(t)$ and $D^F(t)$ over the interval

$[0, D_M^F]$ corresponds to the total time in system for all customers, which, when divided by the total number of customers, yields the expected time in system of an arbitrary customer. Let us denote the expected time in system of an arbitrary customer by $E^F(Y)$. Then, we have

$$E^F(Y) = \frac{\int_0^{D_M^F} [A^F(t) - D^F(t)]dt}{M}.$$

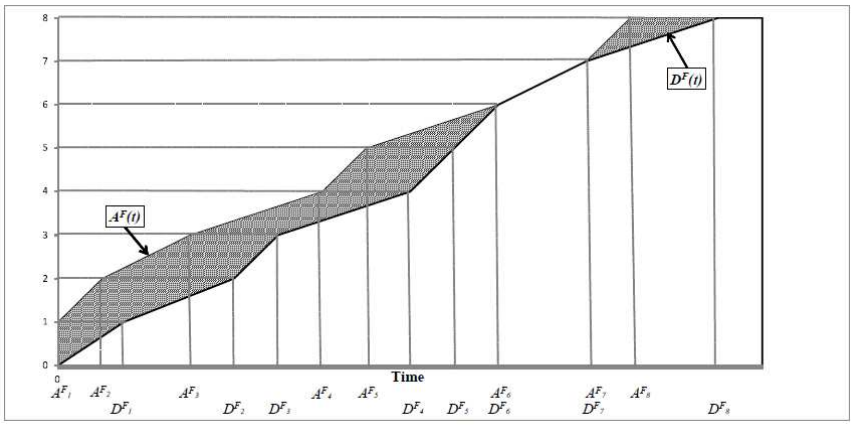


Figure 3.5: An Illustration of the Fluid Approximation

The area under $A^F(t)$ over the interval $[0, D_M^F]$ is the sum of the areas of $M-1$ trapezoids and one rectangle. If we define $S_m^F(A)$ as the area of the m^{th} trapezoid from left, then $S_m^F(A) = (m + \frac{1}{2})\frac{1}{\lambda_{m+1}}$ for $m = 1, ...M-1$, and the area of the rectangle, which we denote by $S_r^F(A)$, equals $M(D_M^F - A_M^F)$.

We now let $S^F(A)$ denote the total area under $A^F(t)$ for $t \in [0, D_M^F]$. Then, we can show that $S^F(A) = \sum_{m=1}^{M-1} S_m^F(A) + S_r^F(A) = \sum_{m=2}^{M} m\frac{1}{\lambda_m} - (M+\frac{1}{2})A_M^F + MD_M^F$.

Similarly, we denote $S^F(D)$ as the area under $D^F(t)$ over the interval $[0, D_M^F]$. This is the sum of the areas of one triangle and $M-1$ trapezoids. The area of the triangle, which we denote by $S_t^F(D)$, equals $\frac{1}{2}D_1^F$. The area of the m^{th} trapezoid

from left, which we denoted by $S_m^F(D)$, is given by $S_m^F(D) = (m+\frac{1}{2})(D_{m+1}^F - D_m^F)$ for $m = 1, ..., M-1$. This implies that $S^F(D) = \sum_{m=1}^{M-1} S_m^F(D) + S_t^F(D) = (M-\frac{1}{2})D_M^F - \sum_{m=1}^{M-1} D_m^F$.

Putting it together, the expected time in system can be written as

$$E^F(Y) = \frac{S^F(A) - S^F(D)}{M} = \frac{\sum_{m=2}^{M} m\frac{1}{\lambda_m} - (M+\frac{1}{2})A_M^F + \sum_{m=1}^{M} D_m^F - \frac{1}{2}D_M^F}{M}.$$

Using the above explicit expressions, we can evaluate each of the arrival and service time processes considered in the numerical study of the previous sections. For the sake of brevity, we focus on the relative performance of different arrival processes. Without loss of generality, we scale time such that $\mu_m = 1$ for $m = 1, ..., M$, and the sequences $\{\frac{1}{\lambda_m} | m = 2, ..., M\}$ are as those sequences in Table 3.1. For the four arrival processes with heterogeneous inter-arrival times, $\frac{1}{\lambda_m} \in \{\frac{1}{M}\frac{2}{\lambda}, ..., \frac{M-1}{M}\frac{2}{\lambda}\}$, and for the process with constant inter-arrival times, we have $\frac{1}{M}\frac{2}{\lambda} \leq \frac{1}{\lambda} \leq \frac{M-1}{M}\frac{2}{\lambda}$. In what follows, we consider the average time in system instead of the average waiting time in queue. Since the total service times of all customers are the same among all the arrival processes, the ordering of processes will not be affected by using time in system instead of waiting time in queue. Let $E^F(Y)_{(C)}$, $E^F(Y)_{(D)}$, $E^F(Y)_{(I)}$, $E^F(Y)_{(DI)}$, and $E^F(Y)_{(ID)}$ refer respectively to the expected time in system for the arrival processes with "Constant", "Decreasing", "Increasing", "Decreasing/Increasing", and "Increasing/Decreasing" inter-arrival times.

We distinguish three different cases: Case 1 ($\frac{1}{M}\frac{2}{\lambda} \geq 1$); Case 2 ($\frac{M-1}{M}\frac{2}{\lambda} \leq 1$); and Case 3 ($\frac{1}{M}\frac{2}{\lambda} < 1 < \frac{M-1}{M}\frac{2}{\lambda}$).

Case 1: This is an obvious case. We have $D_M^F = \frac{M+(\lambda-1)}{\lambda}$ for all the processes. Therefore, it is easy to show that $E^F(Y)$ is the same for all the processes.

Case 2: In this case, $D_M^F = M$ for all the processes. After some algebra, we obtain $E^F(Y)_{(C)} = \frac{(\lambda-1)M^2+2M-1}{2\lambda M}$, $E^F(Y)_{(D)} = \frac{(3\lambda-4)M^2+9M-5}{6\lambda M}$, $E^F(Y)_{(I)} = \frac{(3\lambda-2)M^2+3M-1}{6\lambda M}$, $E^F(Y)_{(DI)} = \frac{(2\lambda-2)M^2+3M}{4\lambda M}$, and $E^F(Y)_{(ID)} = \frac{(2\lambda-2)M^2+3M}{4\lambda M}$. Then, we can easily show that

$$E^F(Y)_{(D)} < E^F(Y)_{(ID)} = E^F(Y)_{(DI)} < E^F(Y)_{(C)} < E^F(Y)_{(I)},$$

which is consistent with the results in Section 3.5.1.

Case 3: Denote $D^F_{M(C)}$, $D^F_{M(D)}$, and $D^F_{M(I)}$ as the makespan for the arrival processes with "Constant", "Decreasing", and "Increasing" inter-arrival times, respectively.

Constant: For this process, we distinguish two cases, $\lambda \geq 1$ and $\lambda < 1$. In the first case, we have $\frac{1}{\lambda_m} \leq 1$ for $m = 2, ..., M$, and therefore $D_m^F = m$. In the second case, we have $\frac{1}{\lambda_m} > 1$ for $m = 2, ..., M$, and therefore $D_m^F = \frac{m+(\lambda-1)}{\lambda}$. Thus,

$$D^F_{M(C)} = \begin{cases} \frac{M+(\lambda-1)}{\lambda} & \text{for } \lambda \in (2\frac{1}{M}, 1), \\ M & \text{for } \lambda \in [1, 2\frac{M-1}{M}), \end{cases} \quad \text{and}$$

$$E^F(Y)_{(C)} = \begin{cases} \frac{2\lambda M-\lambda}{2\lambda M} & \text{for } \lambda \in (2\frac{1}{M}, 1), \\ \frac{(\lambda-1)M^2+2M-1}{2\lambda M} & \text{for } \lambda \in [1, 2\frac{M-1}{M}). \end{cases}$$

Decreasing: For this process, the inter-arrival times are such that $\frac{1}{\lambda_m} \geq 1$ for $m \in [2, \frac{2-\lambda}{2}M + 1]$ (assuming $\frac{\lambda M}{2}$ takes integer values), and $\frac{1}{\lambda_m} < 1$ for $m \in [\frac{2-\lambda}{2}M + 2, M]$. Therefore, $D_m^F = \sum_{j=2}^{m} \frac{1}{\lambda_j} + 1$ for $m \in [2, \frac{2-\lambda}{2}M + 1]$, and $D_m^F = \sum_{j=2}^{\frac{2-\lambda}{2}M+1} \frac{1}{\lambda_j} + (m - \frac{2-\lambda}{2}M)$ for $m \in [\frac{2-\lambda}{2}M + 2, M]$. That is,

$$D_m^F = \begin{cases} \frac{(2m+\lambda-2)M-m(m-1)}{\lambda M} & \text{for } m \in [2, \frac{2-\lambda}{2}M + 1], \\ \frac{(\lambda-2)^2M+2(\lambda-2)}{4\lambda} + m & \text{for } m \in [\frac{2-\lambda}{2}M + 2, M]. \end{cases} \quad \text{Thus,}$$

$D^F_{M(D)} = \frac{(\lambda^2+4)M+(2\lambda-4)}{4\lambda}$, which finally gives $E^F(Y)_{(D)} = \frac{\lambda^3M^2-(3\lambda^2-24\lambda)M-10\lambda}{24\lambda M}$.

Increasing: For this process, the inter-arrival times are such that $\frac{1}{\lambda_m} < 1$ for

$m \in [2, \frac{\lambda}{2}M]$, and $\frac{1}{\lambda_m} \geq 1$ for $m \in [\frac{\lambda}{2}M + 1, M]$. We can then see that $\max_{1\leq i\leq m}\{\sum_{j=2}^{i}\frac{1}{\lambda_j} + \sum_{j=i}^{m}\frac{1}{\mu_j}\}$ is equal to either $\sum_{j=2}^{m}\frac{1}{\lambda_j} + \frac{1}{\mu_m}$ or $\sum_{j=1}^{m}\frac{1}{\mu_j}$. Therefore, $D_m^F = \max_{1\leq i\leq m}\{\sum_{j=2}^{m}\frac{1}{\lambda_j}, \sum_{j=1}^{m}\frac{1}{\mu_j}\} = (m-1)\max\{\frac{m}{\lambda M}, 1\} + 1$. Let us now distinguish two cases, $\lambda \geq 1$ and $\lambda < 1$. In the first case, $\frac{m}{\lambda M} \leq 1$ for $m = 2, ..., M$, and therefore $D_m^F = m$. In the second case, $D_m^F = \begin{cases} m & \text{for } m \in [2, \lambda M], \\ \frac{\lambda M + m(m-1)}{\lambda M} & \text{for } m \in [\lambda M + 1, M]. \end{cases}$ Therefore, we can obtain

$$D_{M(I)}^F = \begin{cases} \frac{M+\lambda-1}{\lambda} & \text{for } \lambda \in (2\frac{1}{M}, 1), \\ M & \text{for } \lambda \in [1, 2\frac{M-1}{M}), \end{cases} \text{ and}$$

$$E^F(Y)_{(I)} = \begin{cases} \frac{\lambda^3M^2-(3\lambda^2-6\lambda)M-\lambda}{6\lambda M} & \text{for } \lambda \in (2\frac{1}{M}, 1), \\ \frac{(3\lambda-2)M^2+3M-1}{6\lambda M} & \text{for } \lambda \in [1, 2\frac{M-1}{M}). \end{cases}$$

Applying the implicit function theorem, it is easy to show that there exits an $\alpha^F(M) \in (1, 2\frac{M-1}{M})$ increasing in M such that

$$E^F(Y)_{(C)} < E^F(Y)_{(D)} < E^F(Y)_{(I)} \text{ for } \lambda \in \left(2\frac{1}{M}, \alpha^F(M)\right), \text{ and}$$
$$E^F(Y)_{(D)} < E^F(Y)_{(C)} < E^F(Y)_{(I)} \text{ for } \lambda \in \left(\alpha^F(M), 2\frac{M-1}{M}\right),$$

which is again consistent with the results in Section 3.5.1. (We can obtain similar expressions for the expected time in system for the arrival processes with "Decreasing/Increasing" and "Increasing/Decreasing" inter-arrival times. For the sake of brevity, we omit the details. The relative ordering also coincides with the one observed in the previous section.)

Other results from Section 3.5.1 can also be confirmed using the fluid approximation. For example, the difference in performance between different arrival processes decreases as λ increases and approaches 0 as $\lambda \to \infty$. The limit case of $\lambda \to \infty$ corresponds to the case of instantaneous arrivals. In

that case, the expression for the expected time in system reduces to $E^F(Y) = \frac{1}{M}\sum_{m=2}^{M}\sum_{j=1}^{m-1}\frac{1}{\mu_j}+\frac{1}{2M}\sum_{m=1}^{M}\frac{1}{\mu_j}$. It is straightforward to show that this expression converges asymptotically to the expression from the exact analysis in Section 3.3 as $M \to \infty$, with $\lim_{M\to\infty}\frac{E^F(Y)}{E(Y)} = 1$.

3.7 Example Applications

In this section, we describe example applications where the results from our analysis can be used to support operational decision making.

3.7.1 A Job Sequencing Problem

Consider the job sequencing problem described in Section 3.1. In particular, consider a system with M jobs to be sequenced on two production stages (e.g., a manufacturing stage and an inspection stage) in series, with a single server at each stage (the extension to multiple servers is straightforward). All M jobs are available at time 0. The processing time of job h for $h = 1, ..., M$, at stage r for $r = 1, 2$, is exponentially distributed with rate $\mu_{(h),r}$. Once a sequence is selected, the jobs are processed in that sequence on both stages without idling (i.e., a server never idles if there is a job available to be processed). For a given sequence, the expected waiting time of an arbitrary job at the first stage equals $\frac{1}{M}\sum_{m=2}^{M}\sum_{l=1}^{m-1}\frac{1}{\mu_{l,1}}$, where $\mu_{l,1}$ is the processing rate of the job assigned to position l (the l^{th} to process), and the corresponding total time in that stage equals $\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{m}\frac{1}{\mu_{l,1}}$. To characterize the performance at the second stage, we must first characterize the inter-arrival time distributions to

that stage. This can be done by recognizing that, given a job sequence, the distributions of inter-arrival times to the second stage are simply the distributions of processing times at the first stage. In particular, if job h is assigned position m ($m \geq 2$) in the sequence, then the time between the $(m-1)^{th}$ and m^{th} arrivals to the second stage is exponentially distributed with rate $\mu_{(h),1}$. Consequently, the expected waiting time for an arbitrary job at the second stage is given by $\frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_{l,2}}$, where $p_{m,i}$ can be computed via the analysis we developed in Section 3.3, with λ_l and μ_l in Equation (3.4) replaced by $\mu_{l,1}$ and $\mu_{l,2}$ for all l, respectively. This leads to the expected total waiting time in system of an arbitrary job as $\frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}(\frac{1}{\mu_{i,1}} + \sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_{l,2}})$. Other performance measures can be similarly obtained. In particular, the expected makespan is given by $\sum_{m=1}^{M}\frac{1}{\mu_{m,1}} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_{l,2}} + \frac{1}{\mu_{M,2}}$.

From the above analysis, we can see that by controlling the job sequence, the system manager can control the distributions of inter-arrival times at the second stage, and therefore the corresponding system performance. Next, we present numerical results for an example system where $\mu_{(h),1} = \frac{M+1}{h}\frac{\varepsilon}{2}$ and $\mu_{(h),2} = \mu$, for $h = 1, ..., M$ and constants ε and μ. We evaluate four different sequences (four permutations of the sequence $\{\frac{M+1}{1}\frac{\varepsilon}{2}, ..., \frac{M+1}{M}\frac{\varepsilon}{2}\}$) as described in Table 3.4 (to be consistent with the other sections, we name the sequences according to the expected service times instead of the service rates). The first sequence corresponds to an ordering of the jobs in decreasing expected service times at stage 1, which implies an ordering of the jobs in decreasing expected inter-arrival times at stage 2. The second sequence corresponds to an ordering in increasing expected inter-arrival times at stage 2, while the third and fourth correspond respectively

to, decreasing and then increasing, and, increasing and then decreasing, orderings of the expected inter-arrival times at stage 2.

Job Sequences	Expected Service Times at Stage 1
Decreasing	$E(\varepsilon_m) = \frac{M-m+1}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., M$
Increasing	$E(\varepsilon_m) = \frac{m}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., M$
Decreasing/Increasing	$E(\varepsilon_m) = \frac{M-2m+1}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., \frac{M}{2}$ $E(\varepsilon_m) = \frac{2m-M}{M+1}\frac{2}{\varepsilon}$ for $m = \frac{M+2}{2}, ..., M$
Increasing/Decreasing	$E(\varepsilon_m) = \frac{2m}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., \frac{M}{2}$ $E(\varepsilon_m) = \frac{2M-2m+1}{M+1}\frac{2}{\varepsilon}$ for $m = \frac{M+2}{2}, ..., M$

Table 3.4: Job Sequences

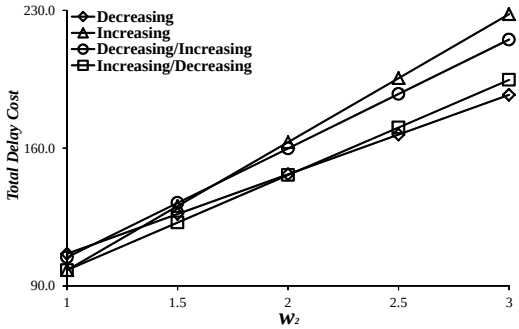


Figure 3.6: Impact of Job Sequence on Delay Cost ($M = 100$, $\varepsilon = 1$, $\mu = 0.5$)

Figure 3.6 provides comparisons of the four job sequences under different values of delay costs (consistent with the job scheduling literature, we assign a delay cost, w_r per job per unit time at stage r for $r = 1, 2$; without loss of generality, we let $w_1 = 1$ and vary w_2; the case of $w_1 = w_2 = 1$ allows us to compare the expected total delay in system for the four different job sequences). As we can see, the four job sequences lead to significantly different total delay costs. Perhaps surprisingly, the "Increasing" sequence which minimizes the delay cost at stage 1 does not necessarily minimize the expected total delay cost. In fact, for sufficiently large w_2, such a sequence performs the worst. This can be explained as follows.

The "Increasing" sequence generates the "Increasing" inter-arrival times at stage 2, which, as discussed in Section 3.5.1, results in long waiting times. On the other hand, the "Decreasing" sequence, although leading to long waiting times at stage 1, generates the "Decreasing" inter-arrival times at stage 2 and therefore results in short waiting times at that stage. The net effect, when w_2 is large, is lower total delay cost.

Additional results (the details of which are not shown here for the sake of brevity) indicate that the four job sequences also lead to significant differences in makespan, with the "Increasing" sequence always performing the best. Note that characterizing the optimal sequence is difficult in general (even for the deterministic setting, the problem is strongly NP-hard; see discussions from Pinedo 2012) and is outside the scope of this work.

3.7.2 A Flight Boarding Problem

Consider the flight boarding problem described in Section 3.1. There are M passengers waiting to board a flight, and they are grouped into K equal size zones, each consisting of $\frac{M}{K}$ passengers (assuming M is divisible by K). Passengers from a zone are called to embark only after all the passengers from a higher ranked zone have finished embarking. The announcement of each zone results in arrivals to the gate drawn from a population of $\frac{M}{K}$ passengers. Assuming each passenger takes an exponentially distributed amount of time to arrive, independent of other customers, then the arrival process for each zone corresponds to a pure death process, with the inter-arrival time between customer $m-1$ and customer m being exponentially distributed with rate $(\frac{M}{K}+1-m)\lambda$ for $m = 2, ..., \frac{M}{K}$ (the arrival

time of the first customer is exponentially distributed with rate $\frac{M}{K}\lambda$). This also implies that the expected inter-arrival times within a zone is strictly increasing. Assuming that service times are exponentially distributed with rate μ, the results of Section 3.3 can be readily applied to obtain various measures of performance. In particular, the expected waiting time of an arbitrary passenger can be obtained by setting $\lambda_m = (\frac{M}{K} + 1 - m)\lambda$ for $m = 2, ..., \frac{M}{K}$ and $\mu_m = \mu$ for $m = 1, ..., \frac{M}{K}$ in Equation (3.4), and the expected makespan (the expected boarding completion time of all zones) is given by $K[\frac{1}{\lambda}\sum_{m=1}^{\frac{M}{K}}\frac{1}{m} + \frac{1}{\mu}(\sum_{i=1}^{\frac{M}{K}-1} i\ p_{\frac{M}{K},i} + 1)]$.

As we can see, by controlling the number of zones, the system manager can control the distributions of inter-arrival times and therefore the corresponding system performance. Two extreme cases are worth highlighting. The first is when $K = M$; in this case, the expected inter-arrival times are constant. The second is when $K = 1$; in that case, the expected inter-arrival times are strictly increasing. In between, the expected inter-arrival times exhibit a cyclical pattern of being strictly increasing within a cycle (a zone) and having a step decrease between cycles (the start of boarding of each zone). Fewer zones reduce makespan while more zones reduce waiting time. The system manager would typically want to balance the costs associated with these two measures; customers prefer to wait less while boarding (and there is an implied delay cost) while the airline would like to reduce the total boarding time (and there is an implied resource usage cost). There is of course indirect waiting time related to customers waiting for their zones to be called, but the cost of that waiting is lower since customers are less inconvenienced in that case than when they are waiting to board.

In Figure 3.7, we present numerical results for an example system with 120

passengers. The solid line represents the expected waiting times of an arbitrary customer, and the dashed line represents the expected makespan of the boarding process. It is interesting to note the diminishing value of having more zones. An initial increase in the number of zones significantly reduces expected waiting time while further increases lead to only marginal further reduction. Given that the increase in makespan due to more zones does not exhibit a similar diminishing effect, the optimal number of zones would generally be relatively small.

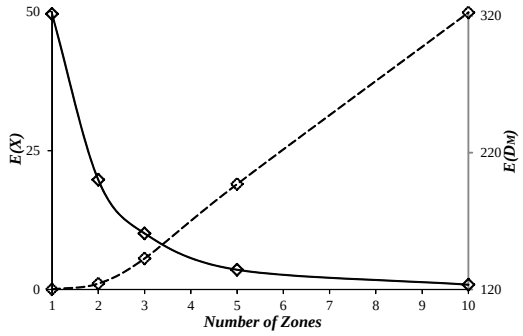


Figure 3.7: Impact of the Number of Zones on Expected Waiting Time and Makespan ($M = 120$, $\lambda = 0.1$, $\mu = 1$)

It is worth to note that results from the above examples, as well as those from the previous sections, show that in general, inter-arrival or service time features that reduce waiting time do not reduce makespan (in fact, the reverse is typically true). Thus, there is a need to trade off the benefit of lower waiting time against shorter makespan, in making decisions about which features to induce.

There are other related settings where arrivals exhibit features that are similar to the ones observed in the flight boarding problem. As mentioned in the introduction, this can be the case when the arrival of customers is triggered by the start of an event (e.g., the arrival of passengers to check-in for a flight or the

arrivals of fans to a concert), and customers may belong to different classes that are differentiated by their risk attitudes toward being late for the event (with some classes preferring to arrive earlier than others). The arrival of customers within the same class can be modeled as a pure death process, which again leads to increasing mean inter-arrival times. Although controlling the number of customers within each class is more difficult in this case than in the flight boarding case, it may be possible, with sufficient incentives, to induce customers to arrive earlier or later. More importantly, recognizing the heterogeneity in inter-arrival times allows the system manager to plan for the necessary capacity (e.g., to meet target service levels as discussed in Section 3.5.4).

We conclude this section by noting that the insights provided so far also apply to settings where arrivals can be controlled in a more direct way, such as when arrivals to a particular process can be specified. This is the case, as we mentioned in the introduction, when arrivals are determined by appointment times. Assuming customers are punctual, inter-arrival times would be deterministic and would correspond to the time between appointment times. Depending on how appointments are scheduled, inter-arrival times may exhibit different features. For example, scheduling more (fewer) appointments early on and then progressively fewer (more) leads to increasing (decreasing) inter-arrival times. Scheduling appointments differently could lead to inter-arrival times that exhibit combinations of both the increasing and decreasing features.

To evaluate the impact of different arrival and service time features, we carried out extensive experiments similar to those in Section 3.5 (for the sake of brevity, we omit the details). The results obtained are qualitatively

consistent with those described there. Hence, our observations also provide insights into desirable features of appointment schedules for such settings. We note that some of these are consistent with the results from the appointment scheduling literature. For example, we observe that arrival processes with the "Increasing/Decreasing" inter-arrival time feature, although not always performing the best, do perform relatively well for all the performance measures considered. This "Increasing/Decreasing" feature is consistent with the "dome-shaped" appointment schedule shown in Kaandorp and Koole (2007), performing well when the performance measure is a weighted cost of waiting time, idle time, and tardiness.

Chapter 4

Service Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows

4.1 Introduction

There are numerous service systems where the arrivals of customers are driven by scheduled appointments. Examples include arrivals to healthcare facilities, government agencies (e.g., immigration, social services, and internal revenue), and academic advising offices at universities, just to name a few. Despite this prevalence, analytical tools for the performance evaluation of these systems are relatively limited. Existing approaches from queueing theory cannot be

readily applied because of several important differences between system with appointment-driven arrivals and standard queueing systems. Systems with appointment-driven arrivals are characterized by (1) a finite number of customers (e.g., the set of patients that have been scheduled at a clinic in a given day), so that steady state analysis cannot be applied, (2) arrivals that are in part determined by known scheduled appointment times, (3) appointment times that may not be equally spaced, and (4) the possibility of customer non-punctuality and no-shows. The difficulty of the analysis can be further compounded in settings where customers are heterogeneous in their punctuality and no-show probabilities.

In this work, we consider a system with a finite number of customers, where each customer has a scheduled appointment. However, customers are not necessarily punctual and may arrive earlier or later than their appointment times. Customers may also not show up. We allow for the arrival time distributions (relative to the appointment time) and the probability of show-up to be customer-specific. We also allow for appointments to be arbitrarily spaced. Under a relatively mild condition on customer arrivals, namely that customers arrive in the order of their appointment times, we develop an exact analytical approach and obtain various performance measures related to customer waiting time. To our knowledge there are no papers that consider simultaneously appointment driven arrivals, non-punctuality, and no-shows, and do so for a setting as general as ours.

There is of course an extensive literature on systems where arrivals are determined by appointment times with typical application in healthcare (see, e.g. the reviews in Preater 2001, Cayirli and Veral 2003, Mondschein and Weintraub

2003, Gupta and Denton 2008). However, in nearly all of that literature, customers are assumed to arrive on time. In most of these papers, performance evaluation is carried out using simulation or traditional queueing analysis where steady state behavior, with an infinite number of arrivals and independent and identically distributed inter-arrival times, is assumed. There are few papers that consider no-shows. Examples include Kaandorp and Koole (2007) and Hassin and Mendel (2008) and the references therein. However, in all of this literature, customers that do show up are assumed to arrive on time. The treatment in this literature is also limited to single server systems. Mercer (1973) considers a system with appointment-driven arrivals but in that case, the number of arrivals is infinite, the appointment times are equally spaced and customers have identical show up probabilities and lateness distributions. Green and Savin (2008) consider a single server model with finite buffer, Poisson arrivals, and deterministic processing times. In their setting, a customer could cancel the appointment creating a no-show but rejoin the queue with a certain probability. Their model does not capture punctuality (arrivals are not driven by specified appointments) and assumes an infinite number of arrivals. Parlar and Moosa (2008) consider a model with a finite number of arrivals motivated by the arrival process of customers to check-in counters for a particular flight. However, in their case, customers arrive independently of each other, with arrivals modeled as a "death process" from the population of a finite number of travelers booked on the flight. Our analysis approach in this work is related to the approach used in the transient analysis of queueing systems (see, e.g., Kelton and Law 1985, Parthasarathy and Moosa 1989, Griffiths et al. 2006). However, transient

analysis of queueing systems typically assumes homogeneous inter-arrival time distributions and the results that exist are mostly for systems with Markovian arrivals.

4.2 Problem Description and Preliminary Results

We consider a queueing system with a single server and M customers arriving over time. Each customer is assigned a time to arrive (the appointment time), and we index the customers by their appointment times. We denote by a_m, for $m = 1, ..., M$, the appointment time of the m^{th} customer, and we have $a_m \leq a_{m+1}$. Customer m has a probability α_m of showing up, independent of all other events. If a customer shows up, she may do so earlier or later than her appointment time. More precisely, we assume that each customer has a finite number of possible arrival times. We denote by K_m the total number of possible arrival times for customer m, and we denote by $t_{m,k}$ the k^{th} possible arrival time of customer m for $k = 1, ..., K_m$ and $m = 1, ..., M$. We have $t_{m,k} < t_{m,k+1}$. For customer m, we also assume her appointment time to be the K_m^*-th possible arrival time (i.e., $t_{m,K_m^*} = a_m$) for an integer $K_m^* \in [1, K_m]$. Customer m may show up at $t_{m,k}$ with probability $q_{m,k}$ ($\sum_{k=1}^{K_m} q_{m,k} = 1$). We denote by A_m the random variable that describes the arrival time of customer m if she shows up. Then, A_m has a finite support on $[t_{m,1}, t_{m,K_m}]$, and its probability mass function (pmf) is specified by

$$f_{A_m}(t) = \begin{cases} q_{m,k} & \text{if } t = t_{m,k} \text{ for } k = 1, ..., K_m, \\ 0 & \text{otherwise.} \end{cases}$$

We model customer arrival times using discrete random variables. However, our analysis and results can also be applied to settings with continuous arrival time distributions noticing that continuous time arrivals can always be approximated by discrete time arrivals with sufficiently small time intervals. For mathematical tractability, we assume that customers arrive in the same order as their appointment times, so that $A_m \leq A_{m+1}$ almost surely (*a.s.*), or equivalently $t_{m,K_m} \leq t_{m+1,1}$. In other words, customer arrival times are non-overlapping. This assumption is reasonable for settings where the length of time between successive appointments is long relative to the total length of the time between possible arrival times of each customer (for example, two successive appointment times are 30 minutes apart but customers are at most 15 minutes early or late).

Customer service times are independent, identically and exponentially distributed with a strictly positive and finite mean $\frac{1}{\mu}$. We make the exponential assumption regarding the distribution of service times for mathematical tractability, as it allows us to formulate the problem as an embedded Markov chain. This assumption is reasonable for systems with high service time variability where service times are typically small but there are occasionally long service times. We assume that the server (e.g., the physician in a healthcare clinic) is available to work starting exactly at a_1 (the appointment time of the first customer).

Upon arrival, a customer goes immediately into service if the server is available. If not, the customer joins the queue and waits. We assume that customers are processed in the order of their appointment times. We also assume that the system is work-conserving with the server never idling when there are customers in queue.

We are interested in characterizing customer waiting time. Our approach consists of first recursively computing the probabilities of the system state seen by a new arrival at a possible arrival time. We then compute the conditional waiting time, given the system state. Finally, we characterize the unconditional waiting time by averaging over all possibilities.

Without loss of generality, we assume the earliest possible arrival time of the first customer is time 0 ($t_{m,1} = 0$). We denote by $R_{m,k}$ the random variable that describes the number of customers found (would have been found) in system by customer m, if she shows up at her k^{th} possible arrival time. We let $p^i_{m,k} = \Pr\{R_{m,k} = i\}$ refer to the probability that the m^{th} customer finds (would have found), upon her arrival at k^{th} possible arrival time, i customers already in system (in queue or in service) for $i = 0, ..., m-1$, $k = 1, ..., K_m$, and $m = 1, ..., M$.

In what follows, we first characterize the probabilities $p^i_{m,k}$. For $m = 1$, we have $p^0_{m,k} = 1$ and $p^i_{m,k} = 0$ for $i \neq 0$, for all k, since the first customer always finds the system empty if she shows up. For $m = 2$, let us first compute $p^1_{2,k}$. For customer 2 to find one customer in system upon arrival, she has to arrive before customer 1 completes service. We distinguish the following three cases; Case 1: customer 1 arrives earlier than her appointment time; Case 2: customer 1 arrives at or later than her appointment time; and Case 3: customer 1 does not show up.
Case 1: If customer 1 arrives at $t_{1,l}$ for $l = 1, ..., K^*_1 - 1$, which is earlier than her appointment time, then she has to wait until a_1 (the starting time of the server) to start service. We denote by ε the random variable that describes the service time of a customer. It is exponentially distributed with mean $\frac{1}{\mu}$. Then,

$p^1_{2,k}$ corresponds to the probability that $\varepsilon > t_{2,k} - a_1$, and thus

$$p^1_{2,k} = e^{-\mu(t_{2,k}-a_1)}.$$

Case 2: If customer 1 arrives at $t_{1,l}$ for $l = K_1^*, ..., K_1$, which is equal to or later than her appointment time, then she starts service immediately after she arrives. Then, $p^1_{2,k}$ corresponds to the probability that $\varepsilon > t_{2,k} - t_{2,l}$, and thus

$$p^1_{2,k} = e^{-\mu(t_{2,k}-t_{2,l})}.$$

Case 3: This is a trivial case. If customer 1 dose not show up, customer 2 will find the system empty when she arrives, and $p^1_{2,k} = 0$.

Putting it all together, we get

$$p^1_{2,k} = \alpha_1 \left(\sum_{l=1}^{K_1^*-1} q_{1,l} e^{-\mu(t_{2,k}-a_1)} + \sum_{l=K_1^*}^{K_1} q_{1,l} e^{-\mu(t_{2,k}-t_{2,l})} \right) \tag{4.1}$$

for $k = 1, ..., K_2$. And $p^0_{2,k} = 1 - p^1_{2,k}$.

For $3 \leq m \leq M$, we first compute $p^i_{m,k}$ for $1 \leq i \leq m-1$, conditioning on the number of customers found (would have been found) in system by customer $m-1$, upon her arrival (if she showed up). It is easy to verify that, in order for customer m to find i $(1 \leq i \leq m-1)$ customers upon arrival, the number of customers found (would have been found) by customer $m-1$ must be at least $i-1$ (i). Thus,

$$\begin{aligned} p^i_{m,k} = \alpha_{m-1} \sum_{l=1}^{K_{m-1}} q_{m-1,l} \sum_{j=i-1}^{m-2} p^j_{m-1,l} \Pr\{R_{m,k} = i \mid R_{m-1,l} = j\} \\ + (1-\alpha_{m-1}) \sum_{l=1}^{K_{m-1}} q_{m-1,l} \sum_{j=i}^{m-2} p^j_{m-1,l} \Pr\{R_{m,k} = i \mid R_{m-1,l} = j\}. \end{aligned}$$

The key quantity to compute then is $\Pr\{R_{m,k} = i \mid R_{m-1,l} = j\}$. To do so, we distinguish the following two cases; Case 1: customer $m-1$ shows up; and Case 2: customer $m-1$ does not show up.

Case 1: Suppose customer $m-1$ arrives at $t_{m-1,l}$ for $l = 1, ..., K_{m-1}$, and finds j $(i-1 \leq j \leq m-2)$ customers in system, then the total number of customers in system immediately after $t_{m-1,l}$ is $j+1$. In order for customer m to find i customers upon arrival at $t_{m,k}$, there must be exactly $j-i+1$ service completions during the time period $(t_{m-1,l}, t_{m,k}]$. Since the server is always busy during this period and service times are exponential distributed, the number of customers served during $(t_{m-1,l}, t_{m,k}]$ follows a Poisson distribution with parameter $\mu(t_{m,k} - t_{m-1,l})$, and therefore we have

$$\Pr\{R_{m,k} = i \mid R_{m-1,l} = j\} = \frac{e^{-\mu(t_{m,k}-t_{m-1,l})}[\mu(t_{m,k} - t_{m-1,l})]^{j-i+1}}{(j-i+1)!}$$

for $i-1 \leq j \leq m-2$.

Case 2: Suppose customer $m-1$ had arrived at $t_{m-1,l}$, she would have found j $(i \leq j \leq m-2)$ customers in system. Then, for customer m to find i customers upon arrival at $t_{m,k}$, there must be exactly $j-i$ service completions during the time period $(t_{m-1,l}, t_{m,k}]$. Following the same analysis as in Case 1, we obtain

$$\Pr\{R_{m,k} = i \mid R_{m-1,l} = j\} = \frac{e^{-\mu(t_{m,k}-t_{m-1,l})}[\mu(t_{m,k} - t_{m-1,l})]^{j-i}}{(j-i)!}$$

for $i \leq j \leq m-2$.

Putting the two cases together leads to

$$p^i_{m,k} = \alpha_{m-1} \sum_{l=1}^{K_{m-1}} q_{m-1,l} \sum_{j=i-1}^{m-2} p^j_{m-1,l} \frac{e^{-\mu(t_{m,k}-t_{m-1,l})}[\mu(t_{m,k} - t_{m-1,l})]^{j-i+1}}{(j-i+1)!}$$
$$+ (1-\alpha_{m-1}) \sum_{l=1}^{K_{m-1}} q_{m-1,l} \sum_{j=i}^{m-2} p^j_{m-1,l} \frac{e^{-\mu(t_{m,k}-t_{m-1,l})}[\mu(t_{m,k} - t_{m-1,l})]^{j-i}}{(j-i)!}$$

$$= \sum_{l=1}^{K_{m-1}} q_{m-1,l} \sum_{j=i}^{m-2} p_{m-1,l}^{j} \frac{e^{-\mu(t_{m,k}-t_{m-1,l})}[\mu(t_{m,k}-t_{m-1,l})]^{j-i}}{(j-i)!}$$
$$\left(1-\alpha_{m-1}+\frac{\alpha_{m-1}\mu(t_{m,k}-t_{m-1,l})}{j-i+1}\right)$$
$$+\alpha_{m-1}\sum_{l=1}^{K_{m-1}} q_{m-1,l}\; p_{m-1,l}^{i-1}\; e^{-\mu(t_{m,k}-t_{m-1,l})} \tag{4.2}$$

for $i = 1, ..., m-1$, $k = 1, ..., K_m$, and $m = 3, ..., M$. (By convention, an empty sum equals 0.)

We can now compute $p_{m,k}^0$ as

$$p_{m,k}^0 = 1 - \sum_{i=1}^{m-1} p_{m,k}^i \tag{4.3}$$

for $k = 1, ..., K_m$, and $m = 2, ..., M$.

Using Equations (4.1)-(4.3), the probabilities $p_{m,k}^i$ for $i = 1, ..., m-1$, $k = 1, ..., K_m$, and $m = 1, ..., M$ can be computed recursively starting from $m = 1$.

Next we show how the above probabilities can be used to characterize various performance measures. Let us first compute the average waiting time in queue. Clearly, as service times are exponentially distributed, the waiting time of a customer only depends on the number of customers she finds upon arrival, but not her exact arrival time (due to the memoryless property). We denote by R_m the expected number of customers found (would have been found) in system by customer m, upon arrival. We also let $p_m^i = \Pr\{R_m = i\}$ refer to the probability that the m^{th} customer finds (would have found), upon arrival, i customers already in system for $i = 0, ..., m-1$ and $m = 1, ..., M$. Then, we have

$$p_m^i = \sum_{k=1}^{K_m} q_{m,k}\; p_{m,k}^i$$

for $i = 0, ..., m-1$ and $m = 2, ..., M$, and $p_1^0 = 1$.

Now, let W_m, a random variable, denote the waiting time in queue of the m^{th} customer, if she shows up, and let $E[(W_m)^n]$ be the corresponding n^{th} moment for $n \geq 1$. (For the rest of the chapter, we use $E[(X)^n]$ to denote the n^{th} moment of a random variable X for $n \geq 1$.) Then, we have

$$E[(W_m)^n] = \sum_{i=1}^{m-1} p_m^i E[(W_m^i)^n]$$

for $2 \leq m \leq M$, where W_m^i is the random variable denoting the waiting time in queue for the m^{th} customer, given that she shows up and finds i customers in system upon arrival. Since service times are independent, identically and exponentially distributed with parameter μ, W_m^i has an Erlang distribution with shape i and rate μ. Hence, the quantities $E[(W_m^i)^n]$ for $n \geq 1$ can be easily computed. For example, we have $E[(W_m^i)] = \frac{i}{\mu}$ and $E[(W_m^i)^2] = \frac{i^2+i}{\mu^2}$.

For the first customer, if she arrives earlier than her appointment time, then she has to wait for the server to start service. We treat this as the waiting time in queue for customer 1, then we have $E[(W_1)^n] = \sum_{k=1}^{K_1^*-1} q_{1,k}(a_1 - t_{1,k})^n$.

Now, let the random variable W denote the waiting time in queue of an arbitrary customer among the M customers. Then, we have

$$\begin{aligned} E[(W)^n] &= \frac{1}{M} \sum_{m=1}^{M} \alpha_m E[(W_m)^n] \\ &= \frac{1}{M} \left[\sum_{m=2}^{M} \alpha_m \sum_{i=1}^{m-1} p_m^i E[(W_m^i)^n] + \alpha_1 \sum_{k=1}^{K_1^*-1} q_{k,l}(a_1 - t_{1,k})^n \right]. \end{aligned}$$

In particular,

$$E[W] = \frac{1}{M\mu} \sum_{m=2}^{M} \alpha_m \sum_{i=1}^{m-1} i \; p_m^i + \frac{\alpha_1}{M} \sum_{k=1}^{K_1^*-1} q_{1,k}(a_1 - t_{1,k})$$

and

$$Var[W] = \frac{1}{M\mu^2}\sum_{m=2}^{M}\alpha_m\sum_{i=1}^{m-1}(i^2+i)p_m^i + \frac{\alpha_1}{M}\sum_{k=1}^{K_1^*-1}q_{1,k}(a_1-t_{1,k})^2 - \left[\frac{1}{M\mu}\sum_{m=2}^{M}\alpha_m\sum_{i=1}^{m-1}i\ p_m^i + \frac{\alpha_1}{M}\sum_{k=1}^{K_1^*-1}q_{1,k}(a_1-t_{1,k})\right]^2.$$

From the probabilities p_m^i, we can also characterize the distribution of W. First, notice that, for the first customer,

$$\Pr\{W_1 \le t\} = \sum_{k=K_1^t}^{K_1} q_{1,k}$$

for $t \ge 0$, where K_1^t is defined as the minimum value of k such that $a_1 - K_1^t \le t$.

For $2 \le m \le M$, since W_m^i is Erlang distributed with shape i and rate μ, we have

$$\Pr\{W_m^i \le t\} = 1 - e^{-\mu t}\sum_{n=0}^{i-1}\frac{(\mu t)^n}{n!},$$

and therefore

$$\Pr\{W_m \le t\} = p_m^0 + \sum_{i=1}^{m-1}p_m^i\Pr\{W_m^i \le t\} = 1 - e^{-\mu t}\sum_{i=1}^{m-1}p_m^i\sum_{n=0}^{i-1}\frac{(\mu t)^n}{n!}.$$

Together, this leads to

$$\begin{aligned}\Pr\{W \le t\} &= \frac{1}{M}\sum_{m=1}^{M}[(1-\alpha_m)1 + \alpha_m\Pr\{W_m \le t\}] \\ &= 1 - \frac{1}{M}\left[\sum_{m=2}^{M}\alpha_m e^{-\mu t}\sum_{i=1}^{m-1}p_m^i\sum_{n=0}^{i-1}\frac{(\mu t)^n}{n!} - \alpha_1\sum_{k=K_1^t}^{K_1}q_{1,k} + \alpha_1\right].\end{aligned}$$

Chapter 5

Conclusions and Future Research Directions

In this chapter, we provide conclusions and future research directions on the work presented in Chapters 2 and 3. We also discuss extensions and future plans for the work presented in Chapter 4.

5.1 On Managing Stochastic Inventory Systems with Scarce Resources

In Chapter 2, we studied the problem of managing production in a production-inventory system where a firm is subject to an allowance constraint on either the amount of input it can use or the amount of output it can produce over a specified compliance period. We considered an extended state-space version of the problem and showed that this modified version of the problem reduces to a

one-dimensional problem. We described various properties of the optimal policy for the modified version of the problem and then showed that these properties also hold for the optimal policy for the original problem. We then used these properties to characterize the structure of the optimal policy for the original problem. In particular, we showed that the optimal production policy is specified by dynamic thresholds that depend on both the on-hand inventory level and the remaining allowance but only via the sum of these two quantities.

We examined how the optimal allowance amount and the allowance usage are affected by the price of the allowance. In particular, we showed that the expected cost is convex in the allowance amount, implying that cost becomes increasingly insensitive to the allowance amount as the allowance amount increases. From numerical experiments, we observed that in some settings the amount of allowance can be tightened significantly without significantly increasing cost. We also observed that putting even a small price on the allowance can lead to a significant decrease in the amount purchased and an increase in its effective usage, implying that pricing scarce resources can lead to a more efficient usage of these resources but without significantly affecting cost.

There are several possible avenues for future research. It would be useful to generalize the results to a broader class of systems, including supply chains with multiple firms where each firm may be subject to its own allowance constraint (e.g., a constraint on water usage or pollution). In such a setting it would be useful to examine how the constraints imposed on one firm (or alternatively the conservation efforts of one firm) affect the production decisions and cost of other firms. It would also be interesting to consider settings where a firm can buy and

sell allowance in each period based on the realization of the price, which may be stochastic and determined by the dynamics of a trading market for allowances (as in a cap-and-trade system). In that case, a firm would make both production and allowance trading decisions in each period, taking into account the randomness of both demand and allowance prices. Finally, it would be useful to consider situations where multiple firms compete at the beginning of the compliance period for a share of a finite amount of total available allowance and a central planner (e.g., an environmental regulator) decides on a rule for allocating this total allowance.

5.2 On Service Systems with Finite and Heterogeneous Customer Arrivals

In Chapter 3, we studied service systems with a finite number of customer arrivals, where customer inter-arrival times and service times are both stochastic and heterogeneous. We characterized the waiting time distribution for each customer, from which we obtained various performance measures of interest including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all customers. We carried out extensive numerical experiments to examine the effect of heterogeneity in inter-arrival and service times. We validated the numerical results using a fluid approximation that yields closed form expressions.

The results of this work highlight the importance of accounting for the heterogeneity in customer inter-arrival and service times, when the number of

customers is finite and customer inter-arrival or service times depend on their positions in the arrival sequence. This heterogeneity arises naturally in many service systems, but could also be engineered into how these systems are designed and managed. Accounting for this heterogeneity is important because different inter-arrival and service time features, even if resulting in the same total workload for the system, can lead to different levels of performance.

There are several possible avenues for future research. It would be useful to generalize our results to a broader class of systems (including queueing networks, systems with general service time distributions, and systems with customer priorities) and to investigate additional applications where systems with the type of features we studied arise naturally. It would also be interesting to study systems with other types of arrival processes such as those with time-dependent arrival rates. Moreover, it would be useful to explore other types of approximations (e.g., diffusion approximations). Finally, it would be meaningful to revisit principles that have been shown to be effective in the design and operation of service systems under steady state assumptions, and to determine whether or not they continue to be effective in systems with finite arrivals and heterogeneous inter-arrival and service times. One such principle is the benefit of pooling of servers and queues in systems with multiple servers.

5.3 On Service Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows

We are currently extending the analysis to more general settings, including systems with multiple servers, Erlang service times, and heterogeneous service time distributions. We are also developing a fluid approximation that yields closed form expressions for the performance measures of interest. We plan to examine the impact of not accounting for non-punctuality and no-shows. We also intend to study how our approach can be used to support individualized appointment scheduling (scheduling that takes into account the punctuality and no-show behavior of each individual customer), and to investigate the extent to which such a scheduling improves performance relative to the more standard schemes where all appointment times are equally spaced.

References

Anupindi, R., Bassok, Y. (1998). Supply Contracts with Quantity Commitments and Stochastic Demand. Tayur, S., Magazine, M., Ganeshan, R. (Editors) *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers.

Baron, D.P., Myerson, R.B. (1982). Regulating a Monopolist with Unknown Costs. *Econometrica*. 50(4):911-930.

Bassok, Y., Anupindi, R. (1997). Analysis of Supply Contracts with Total Minimum Commitment. *IIE Transactions*. 29(5):373-381.

Bassok, Y., Anupindi, R. (2008). Analysis of Supply Contracts with Commitments and Flexibility. *Naval Research Logistics*. 55(5):459-477.

Beaudoin, D., LeBel, L., Frayret, J-M. (2007). Tactical Supply Chain Planning in the Forest Products Industry through Optimization and Scenario-Based Analysis. *Canadian Journal of Forest Research*. 37(1):128-140.

Benjaafar, S., Chen, D., Wang, R. (2014). Managing Stochastic Inventory Systems with Scarce Resources. *Working Paper, University of Minnesota*.

Cayirli, T., Veral, E. (2003). Outpatient Scheduling in Health Care: A Review of Literature. *Production and Operations Management*. 12(4):519-549.

Chao, X., Chen, J., Wang, S. (2008). Dynamic Inventory Management with Cash Flow Constraints. *Naval Research Logistics*. 55(8):758-768.

Chao, X., Zhou, S.X. (2009). Optimal Policy for a Multiechelon Inventory System with Batch Ordering and Fixed Replenishment Intervals. *Operations Research*. 57(2):377-390.

Chen, F., Samroengraja, R. (2000). A Staggered Ordering Policy for One-Warehouse, Multiretailer Systems. *Operations Research*. 48(2):281-293.

Chen, G., Shen, Z.M. (2007). Probabilistic Asymptotic Analysis of Stochastic Online Scheduling Problems. *IIE Transactions*. 39(5):525-538.

Chou, M.C., Liu, H., Queyranne, M., Simchi-Levi, D. (2006). On the Asymptotic Optimality of a Simple On-Line Algorithm for the Stochastic Single-Machine Weighted Completion Time Problem and Its Extensions. *Operations Research*. 54(3):464-474.

Courtois, P.J., Georges, J. (1971). On a Single-Server Finite Queuing Model with State-Dependent Arrival and Service Processes. *Operations Research*. 19(2):424-435.

Cropper, M.L., Oates, W.E. (1992). Environmental Economics: A Survey. *Journal of Economic Literature*. 30(2):675-740.

Deng, S., Yano, C.A. (2006). Joint Production and Pricing Decisions with Setup Costs and Capacity Constraints. *Management Science*. 52(5):741-756.

Dudley, N.J., Musgrave, W.F. (1988). Capacity Sharing of Water Reservoirs. *Water Resources Research*. 24(5):649-658.

Emmons, H., Vairaktarakis. G. (2013). *Flow Shop Scheduling: Theoretical Results, Algorithms, and Applications*. Springer.

Federgruen, A., Zipkin, P. (1986a). An Inventory Model with Limited Production Capacity and Uncertain Demands I. The Average-Cost Criterion. *Mathematics of Operations Research*. 11(2):193-207.

Federgruen, A., Zipkin, P. (1986b). An Inventory Model with Limited Production Capacity and Uncertain Demands II. The Discounted-Cost Criterion. *Mathematics of Operations Research*. 11(2):208-215.

Glasserman, P., Tayur, S. (1994). The Stability of a Capacitated, Multi-Echelon Production-Inventory System under a Base-Stock Policy. *Operations Research*. 42(5):913-925.

Graves, S.C. (1996). A Multiechelon Inventory Model with Fixed Replenishment Intervals. *Management Science*. 42(1):1-18.

Green, L., Kolesar P., Svoronos, A. (1991) Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research*. 39(3):502-511.

Green, V., Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*. 56(6):1526-1538.

Griffiths, J.D., Leonenko, G.M., Williams, J.E. (2006) The Transient Solution to $M/E_k/1$ Queue. *Operations Research Letters*. 34(3):349-354.

Grimm, D., Barkhorn, I., Festa, D., Bonzon, K., Boomhower, J., Hovland, V., Blau, J. (2012). Assessing Catch Shares' Effects Evidence from Federal United States and Associated British Columbian Fisheries. *Marine Policy*. 36(3):644-657.

Gupta, D., Denton, B. (2008). Appointment Scheduling in Health Care: Challenges and Opportunities. *IIE Transactions*. 40(9):800-819.

Haight, R.G. (2013). Personal Communication with Robert G. Haight. *USDA Forest Service, St. Paul, MN*. Oct 2013.

Hall, R.W. (1991). *Queueing Methods: For Services and Manufacturing*. Prentice Hall.

Haque, L., Armstrong, M.J. (2007). A Survey of the Machine Interference Problem. *European Journal of Operational Research*. 179(2):469-482.

Hassin, R., Mendel, S. (2008). Scheduling Arrivals to Queues: A Single-Server Model with No-Shows. *Management Science*. 54(3):565-572.

Hu, B., Benjaafar, S. (2009). Partitioning of Servers in Queueing Systems During Rush Hour. *Manufacturing & Service Operations Management*. 11(3):416-428.

Jouini, O., Wang, R., Benjaafar, S. (2014). Service Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows. *Working Paper, University of Minnesota*.

Kaandorp, G.C., Koole, G. (2007). Optimal Outpatient Appointment Scheduling. *Health Care Management Science*. 10(3):217-229.

Kapuscinski, R., Tayur, S. (1998). A Capacitated Production-Inventory Model with Periodic Demand. *Operations Research*. 46(6):899-911.

Kelton, W.D., Law, A.M. (1985). The Transient Behavior of the $M/M/s$ Queue, with Implications for the Steady-State Simulation. *Operations Research*. 33(2):378-396.

Kleinrock, L. (1975). *Queueing Systems, Volume 1: Theory*. Wiley-Interscience.

Koeleman, P.M., Koole, G.M. (2012). Optimal Outpatient Appointment Scheduling with Emergency Arrivals and General Service Times. *IIE Transactions on Healthcare Systems Engineering*. 2(1):14-30.

Mercer, A. (1973). Queues with Scheduled Arrivals: A Correction, Simplification and Extension. *Journal of the Royal Statistical Society: Series B*. 35(1):104-116.

Mondschein, S.V., Weintraub, G.Y. (2003). Appointment Policies in Service Operations: A Critical Analysis of the Economic Framework. *Production and Operations Management*. 12(2):266-286.

Ouelhadj, D., Petrovic, S. (2009). A Survey of Dynamic Scheduling in Manufacturing Systems. *Journal of Scheduling*. 12(4):417-431.

Ouhimmou, M., D'Amours, S., Beauregard, R., Ait-Kadi, D., Chauhan, S.S. (2009). Optimization Helps Shermag Gain Competitive Edge. *Interfaces*. 39(4):329-345.

Parker, R.P., Kapuscinski, R. (2004). Optimal Policies for a Capacitated Two-Echelon Inventory System. *Operations Research*. 52(5):739-755.

Parlar, M., Moosa, S. (2008). Dynamic Allocation of Airline Check-In Counters: A Queueing Optimization Approach. *Management Science*. 54(8):1410-1424.

Parthasarathy, P.R., Moosa, S. (1989). Transient Solution to the Many-Server Poisson Queue: A Simple Approach. *Journal of Applied Probability*. 26(3):584-594.

Pinedo, M.L. (2012). *Scheduling: Theory, Algorithms, and Systems*. Springer.

Porteus, E. (2002). *Foundations of Stochastic Inventory Theory*. Stanford University Press.

Preater, J. (2001). A Bibliography of Queues in Health and Medicine. *Keele Mathematics Research Report, Keele University*.

Robinson, L.W., Chen, R.R. (2010). A Comparison of Traditional and Open-Access Policies for Appointment Scheduling. *Manufacturing & Service Operations Management*. 12(2):330-346.

Rogers, J., Averyt, K., Clemmer, S., Davis, M., Flores-Lopez, F., Frumhoff, P., Kenney, D., Macknick, J., Madden, N., Meldrum, J., Overpeck, J., Sattler, S., Spanger-Siegfried, E., Yates, D. (2013). *Water-Smart Power: Strengthening the U.S. Electricity System in A Warming World*. Cambridge, Union of Concerned Scientists.

Ross, S.M. (1978). Average Delay in Queues with Non-Stationary Poisson Arrivals. *Journal of Applied Probability*. 15(3):602-609.

Ross, S.M. (2009). *Introduction to Probability Models*. Academic Press.

Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evaluation, Volume 2: Finite Systems*. North-Holland.

Urban, T.L. (2000). Supply Contracts with Periodic, Stationary Commitment. *Production and Operations Management*. 9(4):400-413.

Wang, R., Jouini, O., Benjaafar, S. (2014). Service Systems with Finite and Heterogeneous Customer Arrivals. *Manufacturing & Service Operations Management*. Forthcoming.

Weitzman, M.L. (1974). Prices vs. Quantities. *The Review of Economic Studies*. 41(4):477-491.

Zeng, B., Turkcan, A., Lin, J., Lawley, M. (2010). Clinic Scheduling Models with Overbooking for Patients with Heterogeneous No-Show Probabilities. *Annals of Operations Research*. 178(1):121-144.

Zhang, J., Chen, J., Lee, C. (2012). Coordinated Pricing and Inventory Control Problems with Capacity Constraints and Fixed Ordering Cost. *Naval Research Logistics*. 59(5):376-383.

Zhou, W., Lee, C., Wu, D. (2011). Optimal Control of a Capacitated Inventory System with Multiple Demand Classes. *Naval Research Logistics*. 58(1):43-58.