Optimal control in a Netflix-like closed rental system

Achal Bassamboo* Northwestern University

Ramandeep Singh Randhawa† The University of Texas at Austin

Abstract

We analyze the problem of product allocation to customers following the introduction of a new product in a closed rental system, such as Netflix. We consider two types of customers who differ in their rental time distributions. All customers desire the newly introduced product, and if a customer's request for this product is denied, she receives a substitute product and requests for the new product upon return. We study the control problem of minimizing the mean delay encountered by customers before obtaining the new product in a large market setting. We show that asymptotically this problem is equivalent to solving a linear program that depends on the entire rental duration distribution as opposed to mean alone. The optimal policy turns out to be a mixed priority rule that prioritizes the slower customer class while maintaining a base allocation to the faster customer class to ensure quick return.

1 Introduction

Optimal control in queueing settings has a long history. A typical control problem in this setting involves multiple classes of customers differing on their behavior, arrival and service rates, and delay costs. The system manager is concerned with the allocation of scarce resources to these customers to achieve a socially acceptable outcome, usually the total cost of the customers is minimized. Under linear delay costs, the well known $c\mu$ rule is known to be optimal under fairly mild assumptions (see Cox & Smith (1961)). This rule gives priority to customers with small service requirements and high delay costs. In this paper, we study a similar control problem arising in the context of rental systems with a fixed base of customers or subscribers. Here, the optimal policy turns out to be a mixed priority rule.

The canonical example of rental firms we study is Netflix. Here, the firm limits the number of DVDs that a customer can have at any given time; for simplification, we restrict customers to renting *one* DVD at a time. Each customer provides a preference list, and upon the return of a DVD, the firm sends out a DVD that is highest on the customer's preference list among those available. It is reasonable to assume that the firm has sufficient variety and quantity of "classic" movies to ensure that there is always an available DVD that can be sent out when a customer returns a previous rental. It is also reasonable to assume that no customer wants to see the same movies more than once.

*Kellogg School of Management, e-mail: a-bassamboo@northwestern.edu

†McCombs School of Business, e-mail: rsr@mail.utexas.edu

Consider the scenario in which a new product is introduced that all customers wish to rent (upon completion of their current rental). One expects the demand for this product to typically increase with time, reach its peak and then die down (see, for example, Bassamboo et al. (2007)). If the firm stocks copies of this new product fewer in number than the customer population, a natural control problem of dynamic capacity allocation arises. The firm may choose to prioritize customers based on their "loyalty" or tendency to quit the system based on not receiving desired products. It had been speculated for a while that Netflix was prioritizing infrequent customers by means of "throttling" or "smoothing" the faster users. The following excerpt from the *terms of use* of Netflix acknowledges this fact:

> "In determining priority for shipping and inventory allocation, we may utilize many different factors, including without limitation, the number and type of DVDs you rent through our service, the subscription plan you select, as well as other uses of our service by you. For example, if all other factors are the same, we give priority to those members who receive the fewest DVDs through our service."[1]

At first glance, this policy seems to be a ploy of discouraging overuse of the unlimited renting ability and appealing to the loyal customer base. However, in this paper we show that in some cases such a rule can also optimize a social criterion; the objective we consider in this paper is that of minimizing the mean delay in obtaining the new product. We model this setting in line with the queueing literature by considering two customer classes which differ on their mean rental duration: fast customers and slow customers, where the fast customers return products faster than the slow customers. Noting that products can be allocated to customers only when they request for it, there is a natural trade-off between accepting requests of slow and fast customers. If a fast customer's request is accepted (as opposed to a slow customer), she does return the product faster, however, if her request is denied, her next request will also be sooner than that of a slow customer.

The absence of queueing and the transience in the system dynamics make this problem quite interesting and at the same time difficult to analyze. These complexities are manifested even under *deterministic* rental times. Leaving out variability from the analysis is often useful to get a better sense of the underlying structure of the problem. In our setting even this simplification is not straightforward. To see this, consider the setting where fast and slow customers have rental durations *exactly* m_f and m_s time units, with $m_f < m_s$. Suppose there is one unit of the new product stocked. In this case, it is not difficult to see that it is always optimal to give the new product to the fast customers before giving it to the slow customers, which is in some sense akin to the $c\mu$ rule. However, this initial impression quickly runs into trouble when multiple copies are stocked. Consider the setting of two copies of the new product stocked and three customers: two fast and one slow. Giving both copies to the fast customers incurs a mean delay of $\frac{m_s}{3}$ while giving one copy to the fast and one to slow incurs a lower mean delay of $\frac{m_f}{3}$, and hence prioritizing slow customers is optimal in this setting. The situation becomes further complicated when customers do not initially request for the new product at the same time, but are staggered. These examples illustrate that the optimal control policy cannot be a rule of the $c\mu$ form, i.e., that prioritizes one class over the other. The main aim of this paper is to provide an insight into the structure of the

[1] http://www.netflix.com/SettlementTermsOfUse

optimal policies in such settings.

We consider a rental firm with N_f fast customers with mean rental duration m_f and N_s slow customers with mean rental duration m_s. Within each type, customers are homogenous, i.e., have the same rental distribution. To facilitate our analysis we restrict attention to settings where rental durations are in discrete units, for example days. This amounts to assuming discrete, lattice rental distributions. We assume the system has achieved stationarity when the new product is introduced, and that the firm stocks C copies of the new product. We begin by formally modeling the control problem, and then noting its intractability move on to a large system asymptotic analysis where the number of customers and capacity grows without bound in a fixed proportion. The limiting "fluid" problem turns out to be a linear program that depends on the *entire rental distribution* of the customers. Further, an asymptotically optimal policy, with an error of order $o(n)$ where n denotes a measure of the system size, can easily be constructed from the solution to the limiting linear program. Thus, we focus for the most part on analyzing the optimal solution to the limiting problem.

We first study the limiting problem when customer rental distributions are deterministic in Section 4. In this case, the limiting problem is equivalent to an exact analysis of the system with finite size, albeit allowing for partial allocations. To focus on the interaction between the two classes rather than the effect of scarce capacity, we consider settings with capacity levels that ensure that if *only* one class of customers were part of the system, no customer request would be rejected, i.e., $C \geq \max(N_s, N_f)$. In this setting, we are able to explicitly compute the solution to the control problem. The optimal policy is one that maximizes the number of slow customer requests accepted until period m_f with the caveat that no customer request be rejected after period m_f (see Theorems 1-3). We show that this amounts to giving full priority to slow customers (slow customers' allocation equals that in absence of fast customers) if the slow customers have fairly long mean rental durations ($m_s \geq 3m_f$) or there are a relatively large proportion of slow customers in the system ($\frac{N_s}{m_s} \geq \frac{N_f}{m_f}$). Further, we show that in other settings the optimal policy rejects at the most only 25% of slow customer requests (see Proposition 8). In Section 4.3.1, we study the performance of different policies that prioritize fast customers, prioritize slow customers, allocate products in a FCFS (first come first serve) fashion. It turns out that the FCFS policy, which is implicitly a mixed priority rule, performs reasonably well.

In Section 5, we consider geometric rental distributions, the analog of the exponential distribution in our setting. In this case, though the limiting linear program can be easily solved numerically, we are unable to characterize an explicit solution. To get some insight in to the structure of the optimal policy, we compare different policies as in Section 4.3.1. It turns out that though no policy dominates the others, the FCFS policy again performs reasonably well.

The contributions of the paper can be summarized as follows:

- We model a dynamic control problem arising upon the introduction of a new product in a closed rental system. Here, denied customers do not wait to receive the product, and request for it again at a later time. The control problem is "transient" as the customers only rent the new product once, and thus demand for the new product dies down with time. To the best of our knowledge, there has not been any study of control problems in such settings.

- We propose a policy for this control problem that is near optimal in a large market setting. This policy is the solution to a linear program, that depends intricately on the entire rental distribution but can be easily solved. (See Section 3 for details.)

- For the case of deterministic rental distributions in a high quality-of-service setting, we explicitly characterize the solution to the linear program, and study its structure. The optimal solution is a mixed priority rule that prioritizes slow customers while allocating some base level capacity to the fast customers (see Section 4.2). In particular, the optimal solution maximizes the total capacity allocated to the slow customers, while allocating sufficient capacity to the fast customers so that *no* customer request is denied after period m_f (see Section 4.1 for details). In some settings, this amounts to giving full priority to the slow customers.

- For deterministic and geometric rental distributions, we observe that compared to fixed priority rules, FCFS performs reasonably well, thus suggesting that no-control could be a good option (see Section 4.3.1 and 5.1).

- The methodology used in this paper is based on a large market approximation akin to the many-server limiting regime introduced in Halfin & Whitt (1981), but with general distributions. The literature on optimal control in this regime is currently limited to exponential distributions, with the only exception we are aware of being Tezcan (2007), which considers phase-type distributions.

1.1 Literature survey

One of the most celebrated results in the control of queueing systems is the $c\mu$ rule which dates back to Smith (1956) and Cox & Smith (1961) and minimizes the average delay in a multi-class queueing system. Loosely speaking, this policy states that upon each service completion, the server next serves a customer from the class that has the highest per unit delay cost rate, $c_i\mu_i$, where c_i is the delay cost per unit time and $1/\mu_i$ is the mean service time. This result was extended in an asymptotic setting for non-linear convex costs by van Mieghem (1995) for a single server pool, and further by Mandelbaum & Stolyar (2004) for multi-server pools. Similar results have also been observed in the Halfin-Whitt asymptotic regime in the recent work of Armony (2005), Tezcan & Dai (2006) and Gurvich & Whitt (2007).

In addition to these papers, there is significant amount of literature that focuses on approximation based optimal control of open queuing systems (see for example, Ata & Kumar (2005), Bell & Williams (2001), Dai & Lin (2005), and references therein). The literature on optimal control in closed systems, which is the setting of the current paper, is relatively sparse with Kumar (2000) and Harrison & Wein (1990) as exceptions.

In this paper, we use a fluid-scale approximation to compute the optimal policy. Other papers that use a similar approach are Maglaras (2000), Savin et al. (2005), Bassamboo et al. (2006), and Nazarathy & Weiss (2007). The latter is perhaps the most similar to our paper. Here the authors study optimal policies in a transient (finite horizon) queueing setting. The authors show that the fluid solution solves a separated continuous linear program than can be solved by the techniques proposed in Weiss (2007).

The video rental industry in general has been the subject of a lot of interesting research. For instance,

Mortimer (2004) utilizes data collected at a large number of video rental stores to compare the stocking levels, rental prices, etc. Tang & Deo (2007) studies the competition between retailers on rental price and rental duration. In this paper, we use the rental system model introduced in Bassamboo et al. (2007), where the authors study the stocking of new products upon introduction in a Netflix-like setting with general rental distributions in large market settings. One of the insights of this paper is that highly variable rental duration distributions may allow the firm to achieve a high quality-of-service at fairly low capacity levels. We study the setting where the stock level is fixed and the manager wishes to allocate this scarce resource optimally among the customers so as to achieve a socially optimal outcome.

2 Model

Consider a system with two types of customers: type f denoting fast customers and type s denoting slow customers. There are N_f fast and N_s slow customers, each with a rental duration distribution F_f with mean m_f, and F_s with mean m_s, respectively. The fast customers have a smaller mean rental duration, i.e., $m_f < m_s$. We assume that F_f, F_s are discrete distributions with a lattice of one that do not charge the origin ($F_\alpha(1-) = 0$ for $\alpha = f, s$). In particular, the probability that the rental duration of a type α customer equals j is $p_{j,\alpha} = F_\alpha(j) - F_\alpha(j-1)$ for $\alpha = f, s$ for $j = 1, 2, \ldots$. As described in the introduction, customers always have one product with them (being rented) at any given time, and request for another product upon the return of the current. If they are not given their preferred product, they are given some other product to rent, this is a perfectly reasonable assumption in systems like Netflix. Customer i, $i = 1, \ldots, N_s + N_f$, holds onto her k-th rental for a random time v_{ik}, where $\{v_{ik}\}_{k=1}^{\infty}$ is a sequence of independent and identically distributed random variables, distributed according to the cumulative distribution function F_{α_i}, where $\alpha_i \in \{f, s\}$ denotes customer i's type. In particular, the rental durations do not depend on the product being rented. We assume that the initial rental duration v_{i0} is a random variable that is independent of v_{ik} for $k = 1, 2, \ldots$. We will discuss the distribution of v_{i0} shortly. The rental times are assumed to be independent across customers.

At any given time period j, the residual time $R_i(j)$ of customer i represents the time remaining on her current rental (before return). Since she obtains the next product only on the return of her current rental, her request for the next product occurs at time $j + R_i(j)$. Let $A_i(j)$ denote the counting process that counts the number of products rented by customer i by time j. That is, $A_i(j) = \sup\{\ell : \sum_{k=0}^{\ell} v_{ik} \leq j\}$. Let $T_i(\ell)$ denote the time instant at which the ℓ^{th} rental of customer i began, i.e., $T_i(\ell) = \sum_{k=0}^{\ell-1} v_{ik}$. Thus, we can write $R_i(j)$ as follows:

$$R_i(j) = v_{iA_i(j)} + T_i(A_i(j)) - j.$$

Now, suppose at some arbitrary time t the new product is introduced. We assume that each customer desires the new product upon returning the product she is currently renting. Each customer rents the new product for a duration identical in distribution to that of the old product, which is defined to be any product other than the new one. We assume that the rental duration of the product the customer is renting when the new product is introduced is not affected by the introduction of the new product. In this setting, under constrained capacity (copies of the new product), a natural control problem of

allocating the new product arises. A fair objective criterion is to minimize the total delay encountered by the customers in obtaining the new product. This criterion is akin to minimizing the mean delay in queueing settings, however a major difference is that customers in our setting do not wait. If the new product is not allocated to a customer upon request, the customer rents an old product and requests the new product again upon returning the old product. The optimal allocation is with respect to a social objective, in particular, the average delay encountered by customers between their first request for the new product and the time of obtaining it. As we assume that customers do not rent the new product more than once, the control problem is in fact transient.

The following result, which follows from Chapter 2-16 in Wolff (1989), allows us to make a convenient assumption that frees our analysis from dependence on the introduction time T.

Proposition 1. *For each customer i, $i = 1, \ldots, N_s + N_f$, $x > 0$ and $\alpha = f, s$, we have:*

1. *(Stationarity) If $\mathbb{P}(v_{i0} > x | \alpha_i = \alpha) = \frac{1}{m_\alpha} \sum_{s=x}^{\infty} [1 - F_\alpha(s)]$, then $\mathbb{P}(R_i(j) > x | \alpha_i = \alpha) = \frac{1}{m_\alpha} \sum_{s=x}^{\infty} [1 - F_\alpha(s)]$ for all $j > 0$.*
2. *(Steady-state) If v_{i0} has a finite mean, we obtain*

$$\lim_{j \to \infty} \frac{\sum_{s=0}^{j} \mathbb{P}(R_i(s) > x | \alpha_i = \alpha)}{j} = \frac{\sum_{s=x}^{\infty} [1 - F_\alpha(s)]}{m_\alpha}.$$

In contexts such as Netflix, it is reasonable to assume that the system has been operating for a long time when the new product is introduced, and hence achieved stationarity. The proposition above allows us to make this mathematically precise, with the following assumption: $v_{i,0}$ is distributed according to $F_{\alpha_i,e} \equiv \frac{\sum_{s=0}^{x}[1-F_{\alpha_i}(s)]}{m_{\alpha_i}}$. Then using Proposition 1, we conclude that the residual rental duration for customer i at the time of the new product introduction has a distribution that is independent of the time of introduction t. That is, $R_i(t) \stackrel{d}{=} R_i$, where we abuse the notation R_i to denote the random variable with distribution $F_{\alpha_i,e}$. We shall henceforth use V_i to denote customer i's rental duration of the new product, note that $V_i \stackrel{d}{=} v_{i1}$. Thus, the introduction time t becomes redundant for our analysis, and we simply drop it from our notation.

We now formally describe the control problem. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space on which all random elements are defined. An allocation policy π can be represented by a stochastic process taking values in $\mathbb{Z}_+^2$. For convenience, we suppress the dependence on sample path in the notation, and write $\pi = \{(X_\pi, Y_\pi) \in \mathbb{Z}_+^\infty \times \mathbb{Z}_+^\infty\}$ where $X_\pi(j)$ and $Y_\pi(j)$ denote the (random) number of requests accepted of fast and slow customers in period j, respectively.

For a fixed policy π, let $\mathcal{R}_\pi(j)$ be the set of customers who request the new product at time j and $\mathcal{A}_\pi(j) \subseteq \mathcal{R}_\pi(j)$ be the set of customers whose request is accepted at time j. A customer i belongs to $\mathcal{R}_\pi(j)$ if she has not been allocated the new product prior to period j and her residual time in period $j-1$ equals 1, i.e.,

$$i \in \mathcal{R}_\pi(j) \text{ iff (if and only if) } i \notin \cup_{k=1}^{j-1} \mathcal{A}_\pi(k) \text{ and } R_i(j-1) = 1. \tag{1}$$

In each period customer requests are accepted in the order of increasing index numbers. Thus, for customer i of type $\alpha_i = f$, we have

$$i \in \mathcal{A}_\pi(j) \text{ iff } i \in \mathcal{R}_\pi(j) \text{ and } \#\{k : k \in \mathcal{R}_\pi(j),\ k \leq i,\ \alpha_k = f\} \leq X_\pi(j). \tag{2}$$

Similarly, for customer i of type $\alpha_i = s$, we have

$$i \in \mathcal{A}_\pi(j) \text{ iff } i \in \mathcal{R}_\pi(j) \text{ and } \#\{k : k \in \mathcal{R}_\pi(j),\ k \leq i,\ \alpha_k = s\} \leq Y_\pi(j). \tag{3}$$

Let $D_{\pi,f}(j)$ and $D_{\pi,s}(j)$ represent the number of requests for the new product in period j from fast and slow customers, respectively. For $\alpha = f, s$, we can write

$$D_{\pi,\alpha}(j) = \#\{k : k \in \mathcal{R}_\pi(j), \alpha_k = \alpha\}, \tag{4}$$

and we obtain the following natural constraint on the allocations implied by the demand.

$$\begin{aligned} X_\pi(j) &\leq D_{\pi,f}(j) \\ Y_\pi(j) &\leq D_{\pi,s}(j). \end{aligned} \tag{5}$$

Let $\mathcal{B}_\pi(j)$ be the set of customers who return the new product in period j. We have that

$$i \in \mathcal{B}_\pi(j) \text{ iff there exists } k < j \text{ such that } i \in \mathcal{A}_\pi(k) \text{ and } v_{i,A_i(k)+1} = j - k. \tag{6}$$

Thus, noting that the sum of allocations until period j must be less than the capacity C and the returns until period j, we can write the capacity constraint as:

$$\sum_{k=1}^{j}(X_\pi(k) + Y_\pi(k)) \leq C + \sum_{k=1}^{j} \#\mathcal{B}_\pi(k) \tag{7}$$

for all $j = 1, 2, \ldots$.

We say that an allocation policy π is admissible if there exist $\mathcal{R}_\pi$, $\mathcal{A}_\pi$, $\mathcal{B}_\pi$, $D_{\pi,f}$ and $D_{\pi,s}$ that satisfy the conditions (1-7). Denote the set of admissible policies by Π. The expected cost associated with a policy $\pi \in \Pi$ can be computed as

$$V(\pi) = \mathbb{E}\left[\sum_{j=1}^{\infty}(N_f - \sum_{k=1}^{j} X_\pi(k)) + \sum_{j=1}^{\infty}(N_s - \sum_{k=1}^{j} Y_\pi(k))\right] - N_f m_f - N_s m_s. \tag{8}$$

The first term on the right hand side represents the expected total time elapsed before obtaining the new product summed for all customers. This includes the time until the first request for the new product which equals m_α in expectation for each customer, and thus needs to be subtracted from the first term to obtain the cost. Thus, the optimization problem can be written as

$$\min_{\pi \in \Pi} V(\pi). \tag{9}$$

The reader may reasonably object that our problem formulation does not rule out clairvoyance on the part of the system manager. A realistic formulation would require the allocation policy to be non-anticipating in an appropriate sense. The focus of this paper is on the exact analysis of deterministic rental durations, and on the 'fluid' limits derived from strong laws for non-deterministic rental durations. In both these cases, dropping the non-anticipating assumption has no impact on our analysis. In fact, we shall show that the proposed policy will be optimal even among clairvoyant policies.

3 Asymptotic Analysis

In this section, we formally develop the large system asymptotics. We consider a sequence of systems indexed by $n = 1, 2, \ldots$. The n^{th} system is characterized by: nN_f fast and nN_s slow customers, and nC new products. The rental distribution of the fast and slow customers does not change with n.

For the purpose of minimizing the expected cost, it suffices to consider policies that satisfy the following conditions: (a) if at some time τ the capacity available on-hand exceeds the number of customers who have not yet rented the product, then no requests are denied beyond τ, and (b) eventually allocate the new product to all customers. That is, we consider policies $\pi \in \hat{\Pi}$ which are defined as $\pi \in \Pi$ that satisfy $\tau_\pi := \inf\{\ell : C + \sum_{k=1}^{\ell} \#\mathcal{B}_\pi(k) - \sum_{k=1}^{\ell}(X_\pi(k) + Y_\pi(k)) \geq (N_f - \sum_{k=1}^{\ell} X_\pi(k)) + (N_s - \sum_{k=1}^{\ell} Y_\pi(k))\}$, and for $j > \tau_\pi$, $X_\pi(j) = D_{\pi,s}(j)$ and $Y_\pi(j) = D_{\pi,f}(j)$, and $\sum_{j=1}^{\infty} X_\pi(j) = N_f$ and $\sum_{j=1}^{\infty} Y_\pi(j) = N_s$. This restriction is without loss of optimality as for any $\pi \in \Pi$ which does not lie in $\hat{\Pi}$, one can easily construct an admissible policy in $\hat{\Pi}$ that has a lower expected cost.

Consider a sequence of policies $\{\pi^n : n = 1, 2, \ldots\}$, where π^n is an admissible policy for the n^{th} system, i.e., $\pi^n \in \hat{\Pi}^n$, the set of admissible policies for the n^{th} system as defined above. For ease of notation, henceforth, we will use the subscript π^n to denote quantities pertaining to the n^{th} system obtained while implementing policy π^n. For example, (X_{π^n}, Y_{π^n}) denotes the allocation vector corresponding to the policy π^n. We assume that the allocation vectors when scaled converge in the following sense:

$$\lim_{n\to\infty} \frac{X_{\pi^n}(j)}{n} = \bar{x}_\pi(j), \text{a.s., and } \lim_{n\to\infty} \frac{Y_{\pi^n}(j)}{n} = \bar{y}_\pi(j), \text{ a.s., for each } j \geq 1, \tag{10}$$

where $(\bar{x}_\pi, \bar{y}_\pi) \in \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty$ and we use the subscript π to denote the corresponding policy. (Note that even if this convergence does not hold for the actual sequence, it will hold along some subsequence.) Then, as $n \to \infty$, we obtain the following almost sure convergence for each $j \geq 1$ (this is proved in Proposition 2):

$$\frac{D_{\pi^n,f}(j)}{n} \to \bar{d}_{\pi,f}(j), \tag{11}$$

$$\frac{D_{\pi^n,s}(j)}{n} \to \bar{d}_{\pi,s}(j), \tag{12}$$

$$\frac{\#\mathcal{A}_{\pi^n}(j)}{n} \to \bar{a}_\pi(j), \tag{13}$$

$$\frac{\#\mathcal{B}_{\pi^n}(j)}{n} \to \bar{b}_\pi(j), \tag{14}$$

$$\frac{V(\pi^n)}{n} \to \bar{V}(\pi), \tag{15}$$

where

$$\bar{d}_{\pi,f}(j) = N_f p^e_{j,f} + \sum_{k=1}^{j-1} p_{j-k,f}(\bar{d}_{\pi,f}(k) - \bar{x}_\pi(k)) \tag{16}$$

$$\bar{d}_{\pi,s}(j) = N_s p^e_{j,s} + \sum_{k=1}^{j-1} p_{j-k,s}(\bar{d}_{\pi,s}(k) - \bar{y}_\pi(k)) \tag{17}$$

$$\bar{a}_\pi(j) = \bar{x}_\pi(j) + \bar{y}_\pi(j) \tag{18}$$

$$\bar{b}_\pi(j) = \sum_{k=1}^{j-1} \bar{x}_\pi(k) p_{j-k,f} + \sum_{k=1}^{j-1} \bar{y}_\pi(k) p_{j-k,s}, \tag{19}$$

$$\bar{V}(\pi) = \sum_{j=1}^{\infty}(N_f - \sum_{k=1}^{j} \bar{x}_\pi(k)) + \sum_{j=1}^{\infty}(N_s - \sum_{k=1}^{j} \bar{y}_\pi(k)) - N_f m_f - N_s m_s, \tag{20}$$

and for $\alpha = f, s$, $p^e_{j,\alpha} = \frac{1}{m_\alpha}\sum_{k=j+1}^{\infty} p_{k,\alpha}$ denotes the 'excess' distribution. (Note that we use the convention that $\sum_{j=1}^{0} \gamma(j) = 0$ for any $\gamma(\cdot) \in \mathbb{R}^\infty$.) The expressions in (16) and (17) can be interpreted by noting that the number of customers who request for the new product for the first time is distributed according to the excess distribution (see Proposition 1). Thus, applying the law of large numbers, in the limit, there will be exactly $N_f p^e_{j,f}$ ($N_s p^e_{j,s}$) such customers in period j. Using the same logic, the second term represents the number of customers whose requests have been rejected in the past and request again for the new product in period j. Analogously, we obtain (18-20). Thus, we obtain the following result.

Proposition 2. *For a sequence $\{\pi^n : n = 1, 2, \dots\}$ such that $\pi^n \in \hat{\Pi}^n$ and (10) holds, as $n \to \infty$*

1. *The convergences in (11-14) hold.*
2. *If the rental distributions satisfy $\mathbb{E}[v_{i1} - j | v_{i1} > j, \alpha_i = \alpha] \leq M$ for a constant $M < \infty$, $\alpha = f, s$, and all $j \geq 1$, the convergence in (15) holds.*

It is worth noting that the convergence of the cost function in (15) does not immediately follow from the convergences in (11-14). We need the continuity of the cost function V with respect to the metric induced by point-wise convergence on $\mathbb{R}^\infty$ for this convergence to hold. We prove this continuity under a mild condition on the rental distribution in part 2 of the result. Note that this condition holds for many distributions, in particular the deterministic and geometric distributions which will be discussed in detail in the paper.

Thus, we can formulate the asymptotic cost minimization problem as follows:

$$\begin{aligned}
&\min_{\pi \in (\mathbb{R}^\infty_+ \times \mathbb{R}^\infty_+)} \bar{V}(\pi) \\
&\text{s.t.} \\
&\text{(Demand Constraints)} \ \bar{x}_\pi(j) \leq \bar{d}_{\pi,f}(j), \ \bar{y}_\pi(j) \leq \bar{d}_{\pi,s}(j), \\
&\text{(Capacity Constraint)} \ \sum_{k=1}^{j}(\bar{x}_\pi(k) + \bar{y}_\pi(k)) \leq C + \sum_{k=1}^{j} \bar{b}_\pi(k), \\
&\sum_{j=1}^{\infty} \bar{x}_\pi(j) = N_f, \sum_{j=1}^{\infty} \bar{y}_\pi(j) = N_s, \\
&\bar{x}_\pi(j), \bar{y}_\pi(j) \geq 0.
\end{aligned} \tag{21}$$

In the above linear program, the demand and capacity constraints follow from the definition of admissibility and Proposition 2. Our restriction to policies that allocate the product to all customers, and the non-negativity constraint on allocation lead to the other constraints. These constraints characterize an admissible policy in this asymptotic regime, and we define $\bar{\Pi}$ to be the set of policies that satisfy all the constraints of this linear program.

Suppose π^* denotes a solution to (21). We use this policy as a means of computing an asymptotically optimal sequence of policies for the pre-limit $\{\pi^{*n} : n = 1, 2, \dots\}$. For any n, define the policy π^{*n} such that

$$\begin{aligned} X_{\pi^{*n}}(j) &= \min\left(n\bar{x}_{\pi^*}(j), D_{\pi^{*n},f}(j), C + \sum_{k=1}^{j} \#\mathcal{B}_{\pi^{*n}}(k) - \sum_{k=1}^{j-1}(X_{\pi^{*n}}(k) + Y_{\pi^{*n}}(k)) \right), \\ Y_{\pi^{*n}}(j) &= \min\left(n\bar{y}_{\pi^*}(j), D_{\pi^{*n},s}(j), C + \sum_{k=1}^{j} \#\mathcal{B}_{\pi^{*n}}(k) - \sum_{k=1}^{j-1}(X_{\pi^{*n}}(k) + Y_{\pi^{*n}}(k)) - X_{\pi^{*n}}(j) \right). \end{aligned} \tag{22}$$

That is, the allocation in period j is the minimum of the scaled version of the allocation in policy π^* with the arriving demand and available capacity. Note that we arbitrarily allocate the remaining capacity first to fast, and then to slow customers. Any other form of allocation will serve just as well. It is easy to see that $\pi^{*n} \in \hat{\Pi}^n$ for each n. The following result proves the asymptotic optimality of this policy sequence.

Proposition 3. *The sequence of policies* $\{\pi^{*n} : n = 1, 2, \dots\}$ *defined by* (22) *is asymptotically optimal, i.e., we have*

$$\liminf_{n\to\infty} V(\pi^n) \geq \limsup_{n\to\infty} V(\pi^{*n}) = \bar{V}(\pi^*).$$

for any sequence of policies $\{\pi^n : n = 1, 2, \dots\}$ *with* $\pi^n \in \hat{\Pi}^n$.

This result implies that once we solve the limiting optimization problem (21), we can construct good approximations to the optimal solution for large systems. Thus, henceforth, we focus on the limiting problem (21) alone. Note that this control problem depends intricately on the entire rental duration distribution, and not just the first few moments. We begin by solving this problem exactly for the case of a deterministic distribution in the following section. In Section 5, we study this problem for geometric distributions. Unfortunately, we are unable to completely characterize the asymptotically optimal policy in this setting. However, we study the structure of this policy via a numerical study.

4 Deterministic Rental Distributions

In this section, we focus on the case of deterministic distributions, i.e., for $\alpha = f, s$, $p_{j,\alpha} = 1$ if $j = m_\alpha$ and $p_{j,\alpha} = 0$ otherwise. Here, the equations (16-20) describe the system under a policy π: in an asymptotic sense, where the asymptotics are due to the approximation of integers by real numbers, or in an exact sense if customers are 'atomistic' or infinitely divisible.

Let $\pi = \{(\bar{x}_\pi, \bar{y}_\pi) \in \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty\}$ denote the manager's allocation policy, where $\bar{x}_\pi(j)$ and $\bar{y}_\pi(j)$ represents the number of new products allocated to customers of type f and s in period j. Noting that the excess distribution is given by $p^e_{j,\alpha} = \frac{1}{m_\alpha}$ for $j = 1, \dots, m_\alpha$ for $\alpha = f, s$, once the control is fixed, we

can compute the number of customer of each type requesting the new product in time period j, $\bar{d}_{\pi,f}(j)$ and $\bar{d}_{\pi,s}(j)$, as follows:

$$\bar{d}_{\pi,f}(j) = \begin{cases} \frac{N_f}{m_f} & \text{if } j \le m_f \\ \bar{d}_{\pi,f}(j-m_f) - \bar{x}_\pi(j-m_f) & \text{otherwise,} \end{cases} \tag{23}$$

$$\bar{d}_{\pi,s}(j) = \begin{cases} \frac{N_s}{m_s} & \text{if } j \le m_s \\ \bar{d}_{\pi,s}(j-m_s) - \bar{y}_\pi(j-m_s) & \text{otherwise.} \end{cases} \tag{24}$$

That is, for $\alpha = f, s$, requests of type α customers in any period j beyond m_α consist of customers denied the product in period $j - m_\alpha$.

Noting that the capacity currently in use is that allocated in the previous m_f and m_s periods to the fast and slow customers, respectively, we obtain the following capacity constraint analogous to that in (21) in this setting.

$$\sum_{\ell=1\vee(j-m_f+1)}^{j} \bar{x}_\pi(\ell) + \sum_{\ell=1\vee(j-m_s+1)}^{j} \bar{y}_\pi(\ell) \le C. \tag{25}$$

As each denied customer request incurs a delay of m_f or m_s depending on the customer type, the total delay encountered can be rewritten as $\bar{V}(\pi) \equiv \sum_{j=1}^{\infty} m_f \left(\bar{d}_{\pi,f}(j) - \bar{x}_\pi(j)\right) + \sum_{j=1}^{\infty} m_s \left(\bar{d}_{\pi,s}(j) - \bar{y}_\pi(j)\right)$. Thus, we obtain the following version of (21) for this setting:

$$\begin{aligned} \min_\pi \bar{V}(\pi) &= \sum_{j=1}^{\infty} m_f \left(\bar{d}_{\pi,f}(j) - \bar{x}_\pi(j)\right) + \sum_{j=1}^{\infty} m_s \left(\bar{d}_{\pi,s}(j) - \bar{y}_\pi(j)\right) \\ \text{s.t.,} & \sum_{\ell=1\vee(j-m_f+1)}^{j} \bar{x}_\pi(\ell) + \sum_{\ell=1\vee(j-m_s+1)}^{j} \bar{y}_\pi(\ell) \le C, \\ & \sum_{j=1}^{\infty} \bar{x}_\pi(j) = N_f, \sum_{j=1}^{\infty} \bar{y}_\pi(j) = N_s, \\ & 0 \le \bar{x}_\pi(j) \le \bar{d}_{\pi,f}(j),\ 0 \le \bar{y}_\pi(j) \le \bar{d}_{\pi,s}(j),\ \text{for } j \ge 1. \end{aligned} \tag{26}$$

By appropriately truncating the above linear program, one can envisage numerically solving this problem for different problem parameters. However, our goal in this paper is to develop insights into the structure of the solution, which given the form of the problem is not straightforward. As one can expect, if the capacity in the system is quite low, the linear program will extend over a large number of periods and the interplay between recurring demand (from denied requests) and product returns becomes extremely complicated. To alleviate this issue slightly, we assume the system has sufficient capacity to ensure immediate acceptance of requests of either class when arriving in isolation, i.e., $C \ge \max(N_s, N_f)$. This can be thought of as a *high quality* regime when the product is offered to one class alone. There are two benefits of this assumption: first, as mentioned earlier it allows us to focus on fewer periods, and second, as the capacity level is large enough to cater to requests of one class, it captures the features of an optimal policy with respect to the relative requests accepted between the two classes, and thus allowing us to weigh the operational benefits, if any, of prioritizing one class over the other.

4.1 An equivalent problem

For any policy π, let $k_\pi(j) = \sum_{\ell=1\vee(j-m_f+1)}^{j} \bar{x}_\pi(\ell) + \sum_{\ell=1\vee(j-m_s+1)}^{j} \bar{y}_\pi(\ell)$ denote the current usage of capacity, i.e., the number of new products currently rented out by customers. We shall show that all optimal policies maintain the capacity usage at m_f, $k_\pi(m_f)$, at a constant level, and further ensure that no customer request is denied beyond the period m_f. Thus, the cost of the optimal policy is determined by the weighted sum of number of rejections of fast and slow customers until period m_f, where the weights are m_f and m_s respectively. Using the fact that $k_\pi(m_f) = \gamma$, a constant that depends on the problem parameters, the rejections of fast and slow customers satisfy the following linear relation

$$\sum_{\ell=1}^{m_f}(\bar{d}_{\pi,f}(\ell) - \bar{x}_\pi(\ell)) + \sum_{\ell=1}^{m_f}(\bar{d}_{\pi,s}(\ell) - \bar{y}_\pi(\ell)) = N_f + \frac{N_s}{m_s}m_f - \gamma.$$

The corresponding cost function can then be written as $V(\pi) = (m_s - m_f)\sum_{j=1}^{m_f}\left(\bar{d}_{\pi,s}(j) - \bar{y}_\pi(j)\right) + m_f(N_f + \frac{N_s}{m_s}m_f - \gamma)$. Thus, an optimal policy will maximize the allocation to the slower customers until period m_f with the constraint that no customer is rejected after period m_f.

In order to write out this equivalent formulation precisely, we introduce some notation. Let $\Theta(\pi) \equiv \sum_{j=1}^{m_f} \bar{y}_\pi(j)$ denote the number of requests of slow customers accepted until period m_f. Let $\bar{\Pi}_\theta$ be the set of admissible policies $\pi \in \bar{\Pi}$ such that $\Theta(\pi) = \theta$, i.e., $\bar{\Pi}_\theta$ is the set of all policies that accept the same number of slow customer requests, θ, until period m_f. Define $\Delta(\pi)$ as the indicator that the policy π does not reject any request after period m_f. Noting that all fast (or slow) customers with requests denied before period m_f will request again before period $2m_f$ (or $m_f + m_s$), we obtain $\Delta(\pi) = 1$ iff $\bar{x}_\pi(j) = 0$ for all $j > 2m_f$ and $\bar{y}_j(\pi) = 0$ for all $j > m_f + m_s$ with $\sum_{j=1}^{2m_f} \bar{x}_\pi(j) = 1$ and $\sum_{j=1}^{m_f+m_s} \bar{y}_\pi(j) = 1$, otherwise $\Delta(\pi) = 0$.

We now formally write our equivalent control problem as

$$\max_{\{\pi\in\bar{\Pi}:\Delta(\pi)=1, k_\pi(m_f)=\gamma\}} \Theta(\pi). \tag{27}$$

We explicitly prove this equivalence via a construction of the optimal policy in Theorems 1-3.

4.2 Optimal policy: construction

We begin by first dividing the parameter space into three different regions: (a) $\frac{N_s}{m_s} < \frac{N_f}{m_f}$ with $C \geq N_f + \frac{N_s}{m_s}m_f$, (b) $\frac{N_s}{m_s} < \frac{N_f}{m_f}$ with $C < N_f + \frac{N_s}{m_s}m_f$, and (c) $\frac{N_s}{m_s} \geq \frac{N_f}{m_f}$. Case (a) corresponds to relatively few slow customers and there being sufficient capacity so that the policy which does not reject any request is admissible, and thus trivially optimal. Case (b) corresponds to relatively few slow customers and scarce capacity. The analysis of this case is involved and the construction of the optimal policy depends on the ratio of the mean rental durations of the customers $\frac{m_f}{m_s}$. In particular, if $\frac{m_f}{m_s} \leq \frac{1}{3}$, the policy that accepts all slow customer requests and allocates the remaining capacity to fast customers is optimal. For the cases $\frac{1}{3} < \frac{m_f}{m_s} \leq \frac{1}{2}$ and $\frac{1}{2} < \frac{m_f}{m_s}$, we explicitly construct two sets of optimal policies, one for each case; here, the optimal policy is a mixed priority rule. Finally, case (c) corresponds to the setting with relatively high number of slow customers. In this case, the optimal policy does not reject any slow customer request.

4.2.1 Case $\frac{N_s}{m_s} < \frac{N_f}{m_f}$ and $C \geq N_f + \frac{N_s}{m_s} m_f$

In this case there is sufficient capacity to meet all demand. The optimal policy π^* can be defined as follows:

1. $\bar{x}_{\pi^*}(j) := \frac{N_f}{m_f}$ for $j = 1, \ldots, m_f$, and $\bar{x}_{\pi^*}(j) := 0$ for $j > m_f$.
2. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$, and $\bar{y}_{\pi^*}(j) := 0$ for $j > m_s$.

It can be easily verified that π^* is admissible, and hence trivially optimal. Thus, we immediately obtain the following equivalence.

Theorem 1. *The two control problems* (26) *and* (27) *with* $\gamma = N_f + \frac{N_s}{m_s} m_f$ *are equivalent, i.e., a policy* π *solves* (26) *if and only if it solves* (27).

4.2.2 Case $\frac{N_s}{m_s} < \frac{N_f}{m_f}$ and $C < N_f + \frac{N_s}{m_s} m_f$

We begin by noting that all optimal policies fully allocate the capacity by period m_f.

Lemma 1. *For any policy* π *that solves* (26)*, i.e., is optimal, we have* $k_\pi(m_f) = C$.

Thus, the equivalence result can formally be stated in this setting using the constant $\gamma = C$ as follows:

Theorem 2. *The two control problems* (26) *and* (27) *with* $\gamma = C$ *are equivalent, i.e., a policy* π *solves* (26) *if and only if it solves* (27).

We prove this result via an explicit construction of the optimal solution to (27) with $\gamma = C$. We need to consider three separate cases based on the ratio of the mean rental durations of the two classes: I. $\frac{1}{2} < m_f/m_s$; II. $\frac{1}{3} < m_f/m_s \leq \frac{1}{2}$; and III. $\frac{1}{3} \geq m_f/m_s$.

The following lower bound on the cost function $\bar{V}$ will serve useful in proving this result.

Lemma 2. *1. For a policy* $\pi \in \bar{\Pi}$*, the function* $\bar{V}$ *satisfies the following inequality:*

$$\bar{V}(\pi) \geq m_f(N_f - C) + N_s m_f - \Theta(\pi)(m_s - m_f).$$

Further, if $\Delta(\pi) = 1$ *and* $k_\pi(m_f) = C$, $\bar{V}(\pi) = m_f(N_f - C) + N_s m_f - \Theta(\pi)(m_s - m_f)$.

2. If there exists an admissible policy π *such that* $\Theta(\pi) = \frac{N_s}{m_s} m_f$, $\Delta(\pi) = 1$, *and* $k_\pi(m_f) = C$, *then* π *is a solution to* (26)*, and a policy* $\hat{\pi}$ *solves* (26) *if and only if* $\Theta(\hat{\pi}) = \frac{N_s}{m_s} m_f$, $\Delta(\hat{\pi}) = 1$, *and* $k_\pi(m_f) = C$.

I. Case $\frac{1}{2} < m_f/m_s$: We construct a policy that maximizes the allocation to the slow customers until period m_f while ensuring that no requests are rejected after period m_f. Formally, we propose the following policy π^* as a candidate optimal solution:

1. $\bar{x}_{\pi^*}(j) := \frac{1}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right)$ for $1 \leq j \leq m_s - m_f$.
2. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ and $\bar{x}_{\pi^*}(j + m_s - m_f) := \frac{1}{2}\left(\frac{N_f}{m_f} - \frac{N_s}{m_s}\right)$ for $1 \leq j \leq 2m_f - m_s$.

3. Let $\kappa := \sum_{\ell=1}^{m_f} \bar{x}_{\pi^*}(\ell) + \sum_{\ell=1}^{2m_f - m_s} \bar{y}_{\pi^*}(\ell)$ denote the amount of capacity currently allocated between periods 1 and m_f.

   (a) Set $j := 2m_f - m_s + 1$.

   (b) Set $\bar{y}_{\pi^*}(j) := \min(C - \kappa, N_s/m_s)$ (the allocation is minimum of available capacity and the maximum allocation possible).

   (c) $\kappa := \kappa + \bar{y}_{\pi^*}(j)$.

   (d) If $j < m_f$, set $j := j + 1$ and go to Step (b).

4. Let $\kappa := \sum_{\ell=1}^{m_f} \bar{x}_{\pi^*}(\ell) + \sum_{\ell=1}^{m_f} \bar{y}_{\pi^*}(\ell)$ denote the amount of capacity currently allocated between periods 1 and m_f. If $\kappa < C$

   (a) Set $j := 1$.

   (b) Set $\zeta := \min(C - \kappa, N_f/m_f - \bar{x}_{\pi^*}(j))$ and $\bar{x}_{\pi^*}(j) := \bar{x}_{\pi^*}(j) + \zeta$ (the allocation is minimum of available capacity and the maximum allocation possible).

   (c) $\kappa := \kappa + \zeta$.

   (d) If $j < m_f$, set $j := j + 1$ and go to Step (b).

5. $\bar{x}_{\pi^*}(j) := \frac{N_f}{m_f} - \bar{x}_{\pi^*}(j - m_f)$ for $m_f < j \leq 2m_f$ and $\bar{x}_{\pi^*}(j) := 0$ for $j > 2m_f$, i.e., no fast customer requests are denied beyond period m_f.

6. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $m_f < j \leq m_s$, $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s} - \bar{y}_{\pi^*}(j - m_s)$ for $m_s < j \leq m_f + m_s$ and $\bar{y}_{\pi^*}(j) := 0$ for $j > m_f + m_s$, i.e., no slow customer requests are denied beyond period m_f.

In Step 1, we allocate the minimal capacity to fast customers in periods $1, \ldots, m_s - m_f$ to ensure there are no rejections of either customer class requests between periods $m_f + 1, \ldots, m_s$. The amount of capacity allocated follows naturally from the fact that the requests arriving in periods $m_f + 1, \ldots, m_s$ consists of new (first-time) requests of slow customers and repeat requests of fast customers who have been denied in periods $1, \ldots, m_s - m_f$. Further, only the products that have been allocated to fast customers in periods $1, \ldots, m_s - m_f$ will be returned, and thus available for allocation. E.g., in period $m_f + 1$, we have $\frac{N_f}{m_f} - \bar{x}_{\pi^*}(1)$ requests of fast customers and $\frac{N_s}{m_s}$ requests of slow customers arriving, and we have $\bar{x}_{\pi^*}(1)$ number of products being returned. The policy π^* matches these terms so that $\bar{x}_{\pi^*}(1) = (\frac{N_f}{m_f} - \bar{x}_{\pi^*}(1)) + \frac{N_s}{m_s}$.

In Step 2, we allocate between the fast and slow customers to ensure that no request is denied between period $m_s + 1$ and $2m_f$. Note that there will be repeat requests from fast customers who were denied in periods $m_s - m_f + 1, \ldots, m_f$ and slow customers denied in periods $1, \ldots, 2m_f - m_s$. As our aim is to maximize the allocation to the slow customers we do not reject any of their requests. At the same time we allocate sufficient capacity to the fast customers to match the returns from both types of customers in periods $m_s + 1, \ldots, 2m_f$ with the arriving requests from the fast customers whose requests were denied in periods $m_s - m_f + 1, \ldots, m_f$. E.g., in period $m_s + 1$, the number of repeat requests from fast customer equals $\frac{N_f}{m_f} - \bar{x}_{\pi^*}(m_s - m_f + 1)$, while that from the slow customers equals $\frac{N_s}{m_s} - \bar{y}_{\pi^*}(1) = 0$. The number

of products being returned is $\bar{x}_{\pi^*}(m_s - m_f + 1)$ and $\bar{y}_{\pi^*}(1)$ from the fast and slow customers respectively. The policy π^* matches these terms so that $\bar{x}_{\pi^*}(m_s - m_f + 1) + \bar{y}_{\pi^*}(1) = \frac{N_f}{m_f} - \bar{x}_{\pi^*}(m_s - m_f + 1)$.

Lemma 1 implies that the entire capacity must be allocated by period m_f, and thus in Steps 3 and 4 we allocate the remaining capacity, i.e., C minus the amount currently allocated in Steps 1 and 2, to the slow customers who request in periods $2m_f - m_s + 1, \ldots, m_f$ and fast customers who requests in periods $1, \ldots, m_f$. We begin by allocating as much of the available capacity to slow customers requesting in period $2m_f - m_s + 1$, and then continue allocating capacity in this fashion sequentially in periods $2m_f - m_s + 2, \ldots, m_f$. Following this, we allocate the left over capacity to the fast customers starting from period 1 and then continue allocating in periods $2, \ldots, m_f$ eventually allocating all the capacity.

Steps 5 and 6 ensure that all customer requests arriving after period m_f are immediately accepted.

Note that π^* is an admissible policy. Further, $\pi^* \in \bar{\Pi}_{\theta^*}$, where

$$\theta^* := \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j) = \min\left[C - \sum_{j=1}^{m_f} \bar{x}_{\pi^*}(j), \frac{N_s}{m_s} m_f\right] = \min\left[C - \left(\frac{N_f}{2} + \frac{N_s}{m_s}\frac{(2m_s - 3m_f)}{2}\right), \frac{N_s}{m_s} m_f\right]. \quad (28)$$

The following result proves the optimality of π^* and Theorem 2 for this setting.

Proposition 4. *1. A policy $\hat{\pi}$ is a solution to (26) if and only if $\Theta(\hat{\pi}) = \theta^*$ and $\Delta(\hat{\pi}) = 1$. Thus, the policy π^* is a solution to (26).*

2. π^* is a solution to (27) with $\gamma = C$.

II. Case $\frac{1}{3} < m_f/m_s \leq \frac{1}{2}$: Consider the following policy π^*:

1. $\bar{x}_{\pi^*}(j) := \frac{1}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right)$ for $1 \leq j \leq m_s - 2m_f$.
2. Let $S := \max\left(0, \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4 m_f m_s}\right)$ and $j := m_s - 2m_f + 1$.
   (a) Set $\bar{x}_{\pi^*}(j) := \min\left(\frac{1}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right) + S, \frac{N_f}{m_f}\right)$
   (b) Set $S := \max\left(0, S - \left[\frac{N_f}{m_f} - \frac{1}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right)\right]\right)$.
   (c) If $j < m_f$, set $j := j + 1$ and go to Step (a).
3. Set $\kappa := \sum_{\ell=1}^{m_f} \bar{x}_{\pi^*}(\ell)$ denote the amount of capacity currently allocated between periods 1 and m_f.
   (a) Set $j := 1$.
   (b) Set $\bar{y}_{\pi^*}(j) := \min(C - \kappa, N_s/m_s)$ (the allocation is minimum of available capacity and the maximum allocation possible).
   (c) $\kappa := \kappa + \bar{y}_{\pi^*}(j)$.
   (d) If $j < m_f$, set $j := j + 1$ and go to Step (b).
4. Let $\kappa := \sum_{\ell=1}^{m_f} \bar{x}_{\pi^*}(\ell) + \sum_{\ell=1}^{m_f} \bar{y}_{\pi^*}(\ell)$ denote the amount of capacity currently allocated between periods 1 and m_f. If $\kappa < C$
   (a) Set $j := 1$.

(b) Set $\zeta := \min(C - \kappa, N_f/m_f - \bar{x}_{\pi^*}(j))$ and $\bar{x}_{\pi^*}(j) := \bar{x}_{\pi^*}(j) + \zeta$ (the allocation is minimum of available capacity and the maximum allocation possible).

(c) $\kappa := \kappa + \zeta$.

(d) If $j < m_f$, set $j := j + 1$ and go to Step (b).

5. $\bar{x}_{\pi^*}(j) := \frac{N_f}{m_f} - \bar{x}_{\pi^*}(j - m_f)$ for $m_f < j \leq 2m_f$ and $\bar{x}_{\pi^*}(j) := 0$ for $j > 2m_f$.

6. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $m_f < j \leq m_s$, $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s} - \bar{y}_{\pi^*}(j - m_s)$ for $m_s < j \leq m_f + m_s$ and $\bar{y}_{\pi^*}(j) := 0$ for $j > m_f + m_s$.

Steps 5 and 6 as before ensure that no requests are rejected after period m_f.

In Step 1, we use the same construction as in the first step of the case $m_f/m_s > \frac{1}{2}$ for the allocation of the fast customers for periods $1, \ldots, m_s - 2m_f$. Note that using the same logic, the allocation to fast customers in periods $m_s - 2m_f + 1, \ldots, m_f$ must be at least $\frac{1}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right)$. As slow customers return products for the first time only in period $m_s + 1 > 2m_f$, to ensure no rejections after period $2m_f$, we need to increase the allocation to the fast customers in periods $m_s - 2m_f + 1, \ldots, m_f$ to obtain sufficient returns to meet the demand (of slow customers) in periods $2m_f + 1, \ldots, m_s$. We do this in Step 2. The total capacity that will be returned in periods $2m_f + 1, \ldots, m_s$ consists of the allocation to fast customers in periods $m_f + 1, \ldots, m_s - m_f$ given by $\sum_{j=m_f+1}^{m_s - m_f} \bar{x}_{\pi^*}(j)$, and the capacity remaining in periods $m_s - m_f + 1, \ldots, m_s$ after allocating to the fast and slow customers given by $\sum_{j=m_s-m_f+1}^{m_s} (\bar{x}_{\pi^*}(j - m_f) - \bar{x}_{\pi^*}(j) - \bar{y}_{\pi^*}(j))$. Denoting S as the total excess capacity allocated to fast customers (beyond that allocated in Step 1) in periods $m_s - 2m_f + 1, \ldots, m_f$, we will have $2S = \sum_{j=m_s-m_f+1}^{m_s} (\bar{x}_{\pi^*}(j - m_f) - \bar{x}_{\pi^*}(j) - \bar{y}_{\pi^*}(j))$, and thus we can write.

$$\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_s-m_f+1}^{m_s} (\bar{x}_{\pi^*}(j - m_f) - \bar{x}_{\pi^*}(j) - \bar{y}_{\pi^*}(j)) = \sum_{j=1}^{m_s-2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j)\right) + 2S$$
$$= \frac{(m_s - 2m_f)(N_f m_s - N_s m_f)}{2m_f m_s} + 2S, \tag{29}$$

where we use $\bar{x}_{\pi^*}(j) = \frac{N_f}{m_f} - \bar{x}_{\pi^*}(j - m_f)$ for $m_f + 1 \leq j \leq m_s - m_f$.

Demand arriving between periods $2m_f + 1, \ldots, m_s$ is from the slow customers alone and equals $(m_s - 2m_f)\frac{N_s}{m_s}$. Equating this demand with the available capacity computed in (29), we obtain the amount of the surplus required, $S = \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4m_f m_s}$. Note that this calculation only gives us a match between the total demand and available capacity in periods $2m_f + 1, \ldots, m_s$. To ensure a period by period match, we begin by allocating as much of the surplus S possible to the fast customers arriving in period $m_f - 2m_s + 1$, and then continue allocating the remaining surplus to fast customers arriving in periods $m_f - 2m_s + 1, \ldots, m_f$ sequentially in this manner.

Lemma 1 implies that the entire capacity must be allocated by period m_f, and thus in Steps 3 and 4 we allocate the remaining capacity, i.e., C minus the amount currently allocated in Steps 1 and 2, to the slow and fast customers of periods $1, \ldots, m_f$.

We note that $\pi^* \in \bar{\Pi}_{\theta^*}$, where

$$\begin{aligned}
\theta^* &:= \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j) \\
&= \min\left[C - \sum_{j=1}^{m_f} \bar{x}_{\pi^*}(j), \frac{N_s}{m_s} m_f\right] \\
&= \min\left[C - \left[\frac{m_s - 2m_f}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right) + \min\left(\frac{3m_f - m_s}{2}\left(\frac{N_f}{m_f} + \frac{N_s}{m_s}\right) + S, \frac{N_f}{m_f}(3m_f - m_s)\right)\right], \frac{N_s}{m_s} m_f\right] \\
&= \min\left[C - \left[\min\left(\frac{N_f}{2} + \frac{N_s}{m_s}\frac{m_f}{2} + S, \frac{4m_f - m_s}{2}\frac{N_f}{m_f} + \frac{m_s - 2m_f}{2}\frac{N_s}{m_s}\right)\right], \frac{N_s}{m_s} m_f\right].
\end{aligned} \tag{30}$$

The following result proves the optimality of π^* and Theorem 2 for this setting.

Proposition 5. *1. A policy $\hat{\pi}$ is a solution to* (26) *if and only if $\Theta(\hat{\pi}) = \theta^*$ and $\Delta(\hat{\pi}) = 1$. Thus, the policy π^* is a solution to* (26).

2. π^* is a solution to (27) *with $\gamma = C$.*

III. Case $m_f/m_s \leq \frac{1}{3}$: full priority to slow customers. In this case, we consider a policy π^* that does not reject any request by a slow customer. This leaves us with a capacity of $C - \frac{N_f}{m_s} m_f$ to allocate to the fast customers in periods $1, \ldots, m_f$. Beginning with fast customers' requests in the first period, we allocate as much capacity as we can to these customers subject to available capacity, and then continue allocating capacity in this fashion sequentially in periods $2, \ldots, m_f$ eventually allocating all the capacity. Formally, the policy π^* is given as follows:

1. Set $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$.
2. Set $\kappa := \sum_{\ell=1}^{m_f} \bar{y}_{\pi^*}(\ell) = \frac{N_s}{m_s} m_f$ and $j := 1$.
   (a) Set $\bar{x}_{\pi^*}(j) := \min\left(C - \kappa, \frac{N_f}{m_f}\right)$ and $\kappa := \kappa + \bar{x}_{\pi^*}(j)$.
   (b) If $j < m_f$, set $j := j + 1$ and go to Step (a).
3. $\bar{x}_{\pi^*}(j) := \frac{N_f}{m_f} - \bar{x}_{\pi^*}(j - m_f)$ for $m_f < j \leq 2m_f$ and $\bar{x}_{\pi^*}(j) := 0$ for $j > 2m_f$.

Defining

$$\theta^* := \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j) = \frac{N_s}{m_s} m_f, \tag{31}$$

we have $\pi^* \in \bar{\Pi}_{\theta^*}$. Noting that for any policy π, we must have $\Theta(\pi) \leq \theta^*$, applying Lemma 2.1 and arguing as in the proof of Propositions 4 and 5, we obtain the following optimality result.

Proposition 6. *1. A policy $\hat{\pi}$ is a solution to* (26) *if and only if $\Theta(\hat{\pi}) = \theta^*$ and $\Delta(\hat{\pi}) = 1$. Thus, the policy π^* is a solution to* (26).

2. π^* is a solution to (27) *with $\gamma = C$.*

This completes the proof of Theorem 2 for this setting.

4.2.3 Case $\frac{N_s}{m_s} \geq \frac{N_f}{m_f}$: full priority to slow customers

Consider the following policy π^* as a candidate for the optimal solution.

1. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$.
2. Let $\kappa := \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j)$ denote the amount of capacity currently allocated between periods 1 and m_f.
   (a) Set $j := 1$.
   (b) Set $\bar{x}_{\pi^*}(j) := \min(C - \kappa, N_f/m_f)$ (the allocation is minimum of available capacity and the maximum allocation possible).
   (c) If $j \leq m_s - m_f$ then $\kappa := \kappa + \frac{N_s}{m_s}$ else $\kappa := \kappa + \bar{x}_{\pi^*}(j)$.
   (d) If $j < m_f$, set $j := j + 1$ and go to Step (b).
3. $\bar{x}_{\pi^*}(j) := \frac{N_f}{m_f} - \bar{x}_{\pi^*}(j - m_f)$ for $m_f < j \leq 2m_f$ and $\bar{x}_{\pi^*}(j) := 0$ for $j > 2m_f$.
4. $\bar{y}_{\pi^*}(j) := \frac{N_s}{m_s}$ for $m_f < j \leq m_s$ and $\bar{y}_{\pi^*}(j) := 0$ for $j > m_s$.

Note that in this case, unlike the previous setting, we have the per period requests of slow customers, $\frac{N_s}{m_s} \geq \frac{N_f}{m_f}$, the per period demand of fast customers. Thus, in Step 2, as we allocate capacity to the fast customers 'maximally', we need to 'withhold' some capacity to ensure sufficient capacity for allocation to slow customer requests who request beyond period m_f. This withholding takes place in part (c) of Step 2, where we increment the counter κ which tracks the allocated capacity by $\frac{N_s}{m_s}$ for each of the first $m_s - m_f$ periods.

We again define

$$\theta^* := \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j) = \frac{N_s}{m_s} m_f. \tag{32}$$

Further, defining $\beta(\pi) = C - k_\pi(m_f)$ as the amount of capacity withheld by period m_f for any policy π, we obtain the following lower bound on the cost associated with π.

Lemma 3. *For any policy $\pi \in \bar{\Pi}$, we have $\bar{V}(\pi) \geq m_f(N_f + \beta(\pi) - C + \Theta(\pi)) + m_s(\frac{N_s}{m_s} m_f - \Theta(\pi)) + m_s(\beta(\pi^*) - \beta(\pi))^+$.*

This result immediately gives us the following optimality of π^*.

Proposition 7. *1. A policy $\hat{\pi}$ is a solution to (26) if and only if $\Theta(\hat{\pi}) = \frac{N_s}{m_s} m_f$, $\Delta(\hat{\pi}) = 1$ and $\beta(\hat{\pi}) = \beta(\pi^*)$. Thus, the policy π^* is a solution to (26).*

2. π^* is a solution to (27).

Applying Proposition 7, we obtain the following result analogous to Theorem 2.

Theorem 3. *The two control problems (26) and (27) (with $\gamma = C - \beta(\pi^*)$) are equivalent, i.e., a policy π solves (26) if and only if it solves (27).*

4.3 Discussion

We study the implications of the optimal control policy constructed in the previous section. We begin by studying the percentage of requests rejected of the slow and fast customers as a function of the problem parameters. We fix the capacity level at the critical amount $\max(N_s, N_f)$ and vary the rental durations and the number of customers in each class. Using the value of θ^* computed for the optimal policy in each case, we can calculate the percentage of slow customer requests rejected as $\rho_s = \frac{\frac{N_s}{m_s} m_f - \theta^*}{N_s}$, and those of the fast customers as $\rho_f = \frac{N_f - \min(C-\theta^*, N_f)}{N_f}$. Noting that these parameters only depend on the ratios N_s/N_f and m_s/m_f, we normalize $N_s = 1$ and vary $N_f \in \mathbb{R}_+$ and $0 < m_f < m_s$. We present plots of these quantities in Figures 1-5 for values of $N_f = 0.5, 0.65, 1, 1.35, 2$. In each plot, we vary m_f from zero to m_s. Figure 3 represents the critical case with $N_f = 1$. A commonality in all these plots is that as m_f increases, the rejections of the slower customers at first increases with a peak at $m_f/m_s = 0.5$, and then decreases. Note that as m_f approaches m_s, almost all requests of the fast customers are rejected. Note that as N_f moves away from one (in either direction), the rejections of the slow customers decrease (see Figures 2 and 4) and eventually no slow customer request is denied (see Figures 1 and 5). Further, the maximum fraction of rejections for the slow customers is at a low 25% for the case $N_f = N_s = C$ and $m_s = 2m_f$. This is formalized in the following result.

Proposition 8. *If the capacity $C \geq \max(N_s, N_f)$ then the fraction of slow customer requests which are rejected is bounded above by $\frac{1}{4}$.*

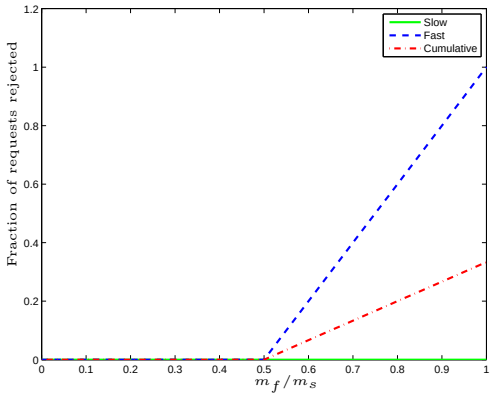

Figure 1: Rejection fractions, $N_f = 0.5$

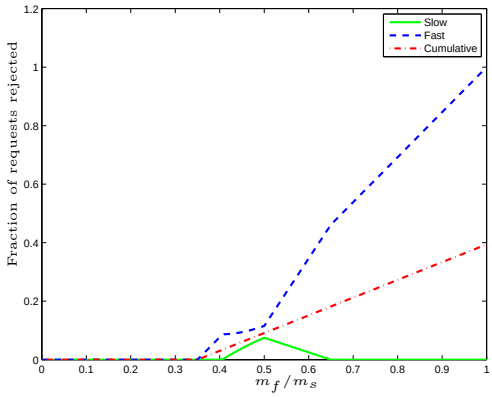

Figure 2: Rejection fractions, $N_f = 0.65$

4.3.1 Comparison with other policies

In this section, we study the performance of static priority rules (of prioritizing fast or slow customers) relative to the optimal policy. Specifically, we are interested in understanding the settings in which these policies yield near optimal performance. We begin by describing the implementation of the priority rules.

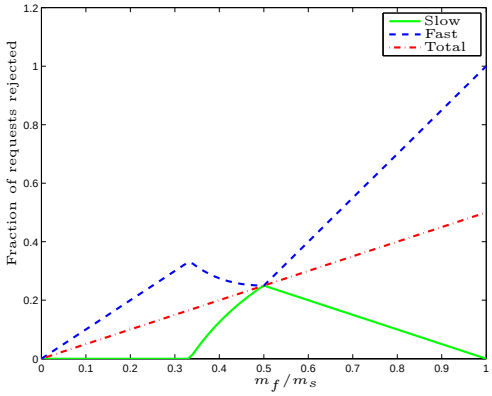

Figure 3: Rejection fractions, $N_f = 1$

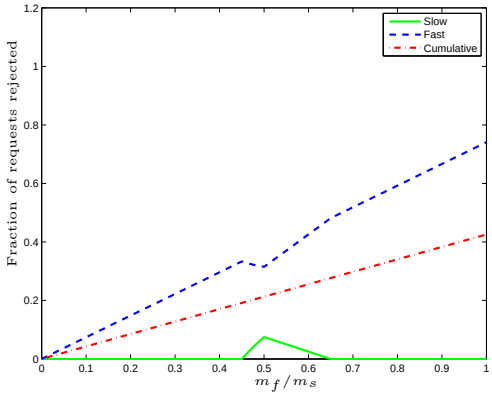

Figure 4: Rejection fractions, $N_f = 1.35$

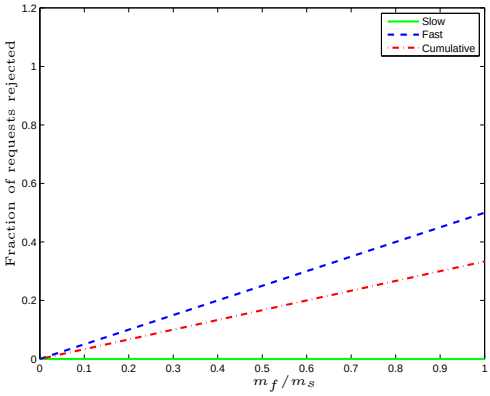

Figure 5: Rejection fractions, $N_f = 2$

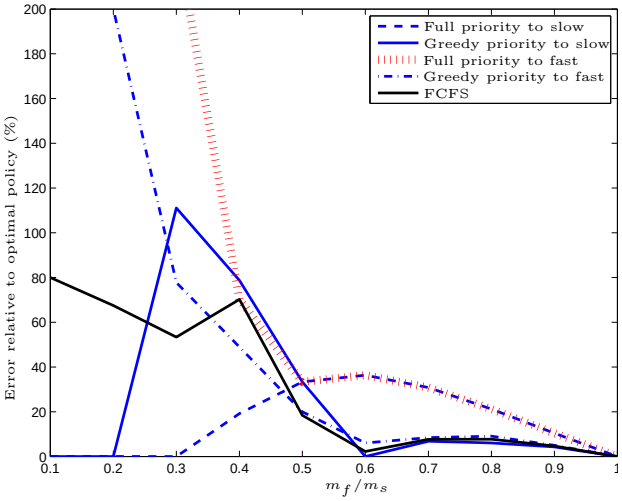

Figure 6: Percentage errors for different policies

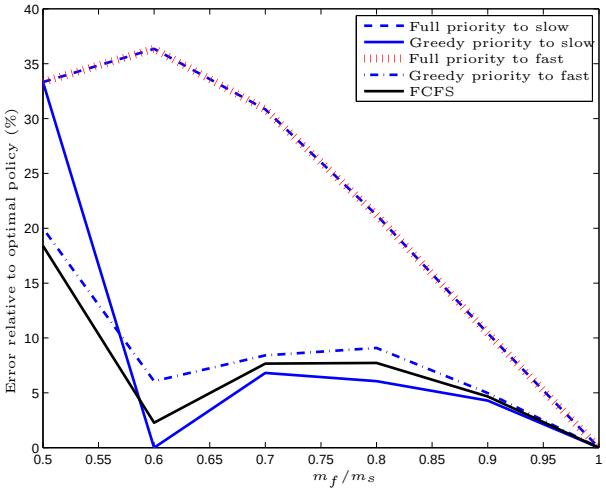

Figure 7: Enlarged version for $0.5 \leq \beta \leq 1$

Priority to fast customers It is a well known result in queueing systems that when encountered with multiple customer classes and a single type of server, the optimal scheduling rule that minimizes the average customer delay is the rule that prioritizes customer classes in decreasing order of μ_i, where μ_i is their service rate. One can envisage two extensions of this rule to the setting in this paper:

- **Full priority to fast customers**: This rule amounts to giving 'full' priority to the *fast* customers in the sense that these customers will have all their requests accepted. From the fast customer perspective this rule is equivalent to the setting in which there are no slow customers. For the case $N_s = N_f = C$, this rule simply allocates the new product to the fast customers until period m_f, and then to the slow customers after period m_f.
- **Greedy priority to fast customers**: This rule prioritizes fast customers, but in a period by period greedy fashion. In particular, in each period, this rule first accepts as many fast customers requests as possible, after which the remaining capacity is allocated to the slow customers.

Analogously, we obtain the following rules that prioritize slow customers.

Priority to slow customers

- **Full priority to slow customers**: This rule amounts to giving *full* priority to the *slow* customers. Note that giving full priority to the slow capacity still leaves us a capacity equal to $C - \frac{N_s}{m_s} m_f$ available to allocate to the fast customers in periods $1, \ldots, m_f$. This allocation can be computed by solving an optimization problem that is similar to the one in (26) with the allocations to the slow customers fixed at $\frac{N_s}{m_s}$ for periods $1, \ldots, m_s$.
- **Greedy priority to slow customers**: This rule is the analog of the greedy priority to fast customers and prioritizes slow customers in a period by period greedy fashion.

In addition to these policies, we also study the FCFS policy that allocates capacity on a first come first serve basis. We fix $m_s = 10$, $N_s = N_f = C = 1$, and vary m_f from 1 to 10. For each set of parameters, we compute the cost for each policy and compute the percentage error in the policy relative to the optimal value. Figure 6 displays the errors in each policy, while Figure 7 presents an enlarged view of the region $5 \leq m_f \leq 10$. Based on these results, we note:

1. Prioritizing fast customers can lead to arbitrarily large costs (compared to the optimal), especially when the ratio $\frac{m_f}{m_s}$ is small.
2. The policies that prioritize slow customers seem to perform fairly well. In fact, the full priority to slow policy always does better than the full priority to fast policy and has a maximum error in value function of 40%.
3. The FCFS policy performs reasonably well.
4. As expected, when m_f is close to m_s, all policies perform well.

It is worth noting that in most cases the greedy policies perform better than their non-greedy counterparts. The greedy policies can thought as analogs of *non-preemptive*, *non-idling* policies in queueing settings in the sense that available capacity is always allocated to customers and never withheld (for future use), and allocated capacity cannot be forced away from customers, i.e., once allocated, the capacity will be available for reuse only upon return by the customer. The non-greedy policies are the analogs of *preemptive* policies. Typically, preemptive policies are expected to perform better than their non-preemptive analogs. However, we observe the opposite in this setting.

Having investigated the deterministic setting, we now turn our attention to the case of geometric rental distributions, which are the analog of the exponential distributions in our setting.

5 Geometric Rental Distributions

In this section, we focus on the case of geometric distributions, i.e., for $\alpha = f, s$, $p_{j,\alpha} = (1-p_\alpha)^{j-1}p_\alpha$ for $j \geq 1$. We focus on the limiting control problem (21). Let $\pi = \{(\bar{x}_\pi, \bar{y}_\pi) \in \mathbb{R}_+^\infty \times \mathbb{R}_+^\infty\}$ denote the manager's allocation policy. Noting that the excess distribution is identical to the rental distribution, i.e., $p^e_{j,\alpha} = p_{j,\alpha}$, once the control is fixed, we can compute the number (fraction) of customers of each type $\alpha = f, s$ requesting the new product in time period j as:

$$\bar{d}_{\pi,\alpha}(j) = N_\alpha(1-p_\alpha)^{j-1}p_\alpha + \sum_{k=1}^{j-1}(1-p_\alpha)^{j-k-1}p_\alpha(\bar{d}_{\pi,\alpha}(k) - \bar{x}_\pi(k)). \tag{33}$$

Using the form of the asymptotic delay cost as given in (20), the limiting control problem is the following version of (21):

$$\begin{aligned}
&\min_\pi \bar{V}(\pi) = \sum_{j=1}^{\infty}(N_f - \sum_{k=1}^{j}\bar{x}_\pi(k)) + \sum_{j=1}^{\infty}(N_s - \sum_{k=1}^{j}\bar{y}_\pi(k)) - N_f m_f - N_s m_s,\\
&\text{s.t. } \sum_{k=1}^{j}(\bar{x}_\pi(k) + \bar{y}_\pi(k)) \leq C + \sum_{k=1}^{j}\bar{b}_\pi(k),\\
&\quad \sum_{j=1}^{\infty}\bar{x}_\pi(j) = N_f, \sum_{j=1}^{\infty}\bar{y}_\pi(j) = N_s,\\
&\quad 0 \leq \bar{x}_\pi(j) \leq \bar{d}_{\pi,f}(j),\ 0 \leq \bar{y}_\pi(j) \leq \bar{d}_{\pi,s}(j),\ \text{for } j \geq 1.
\end{aligned} \tag{34}$$

where $\bar{b}_\pi(k) = \sum_{\ell=1}^{k-1}\bar{x}_\pi(\ell)(1-p_f)^{k-\ell-1}p_f + \sum_{\ell=1}^{k-1}\bar{y}_\pi(\ell)(1-p_s)^{k-\ell-1}p_s$ denotes the number of products returned in period k.

This optimization problem despite being in a simplified form (as compared to the pre-limit problem) is intractable. However, noting that solving this problem amounts to solving a linear programming problem, we numerically study properties of the optimal solution.

5.1 Numerical results

We shall solve (34) numerically, and compare its solution with the four fixed priority rules discussed in Section 4.3.1 and the FCFS policy. We choose four values for $p_f = 0.25, 0.5, 0.75, 1$, and for each value of p_f, we set $p_s = \gamma p_f$ for $\gamma = 0.25, 0.5, 0.75, 0.99$. We fix $N_f = N_s = 1$, and in each case set the capacity to

					Percentage Error				
					Priority to Fast		Priority to Slow		
p_f	p_s	$\frac{m_f}{m_s}$	C	Optimal Cost	Full	Greedy	Full	Greedy	FCFS
0.25	0.0625	0.25	0.42	5.00	2.33	3.01	105.09	105.09	28.09
	0.125	0.5		7.09	4.30	1.15	69.84	43.71	15.34
	0.1875	0.75		7.30	9.88	0.54	39.64	14.47	6.57
	0.2475	0.99		7.23	14.04	0.02	15.22	0.54	0.27
0.5	0.125	0.25	0.50	1.70	17.44	17.44	40.49	40.49	23.66
	0.25	0.5		2.80	1.68	1.68	46.05	46.05	21.07
	0.375	0.75		3.29	0.20	0.20	27.48	18.27	9.63
	0.495	0.99		3.49	0.00	0.00	1.31	0.75	0.38
0.75	0.1875	0.25	0.75	0.28	255.54	255.54	0.00	0.00	42.22
	0.375	0.5		0.92	8.33	8.33	7.48	7.48	10.91
	0.5625	0.75		1.25	6.67	6.67	11.95	11.95	6.59
	0.7425	0.99		1.49	0.22	0.22	0.22	0.22	0.22
1	0.25	0.25	1.00	0.25	300.00	300.00	0.00	0.00	60.00
	0.5	0.5		0.50	100.00	100.00	0.00	0.00	33.33
	0.75	0.75		0.83	20.00	20.00	5.00	5.00	2.86
	0.99	0.99		0.99	0.51	0.51	0.49	0.49	0.01

Table 1: Comparison of different priority rules.

the lowest level such that if customers of only one type joined the system, no requests would be rejected. Table 1 displays the percentage error of each fixed priority rule with respect to the optimal cost.

Based on these results, we note:

1. No rule uniformly dominates the others.
2. The policies that prioritize fast customers perform well for high m_f (low p_f) values, but have very high error levels when m_f and the ratio $\frac{m_f}{m_s}$ are both low.
3. The policies that prioritize slow customers perform well for low m_f (high p_f) values, but have high error levels when m_f is high and the ratio $\frac{m_f}{m_s}$ is low.
4. Greedy policies perform better than those that give full priority as in the case of deterministic distributions.
5. The FCFS rule performs reasonably well. Though it dominates all other policies only in a few settings, the errors are of a moderate level in most settings.

6 Conclusion

This paper studies optimal dynamic capacity allocation of a newly introduced product in a closed Netflix-like rental system. Netflix tends to prioritize infrequent customers, who can be construed to be their loyal customers. This priority rule is quite the opposite of the conventional wisdom in queueing literature that prioritizes fast customers. In this paper, we demonstrate that prioritizing slow customers can be socially optimal in the sense of minimizing the mean delay encountered in obtaining the new release.

We study this control problem in a large system asymptotic regime. The limiting 'fluid' control problem is a linear program that depends on the entire customer rental distribution. We are able to write out the optimal solution explicitly for deterministic distributions; unlike most deterministic problems, the solution is non-trivial. The optimal solution maximizes the number of slow customer requests accepted until period m_f with the constraint that no request (of slow or fast customers) is rejected after period m_f. This solution gives full priority to slow customers (no request rejected) for various model parameters, for example, $m_s \geq 3m_f$ and $\frac{N_s}{m_s} \geq \frac{N_f}{m_f}$. In all settings, slow customers have at the most 25% of their requests rejected. We also study the case of geometric distributions. In this case, the optimal solution is again a mixed priority rule. We solve the problem numerically and contrast it with various other policies. We note there are settings where prioritizing fast customers performs very well, and settings where doing the opposite does quite well. However, we note that the FCFS rule performs reasonably well in all settings.

A Proof of Lemmas

Proof of Lemma 1. Suppose $k_\pi(m_f) < C$. Define $\tau = \arg\min\{\ell : \bar{x}_\pi(\ell) < N_f/m_f\}$. We consider two cases:

Case I. $\tau \leq m_f$**:** Consider the policy π' such that $\bar{y}_{\pi'}(j) = \bar{y}_\pi(j)$ and $\bar{x}_{\pi'}(j) = \bar{x}_\pi(j)$ for $j \geq 1$, $j \neq \tau, m_f + \tau$. Let $\bar{x}_{\pi'}(\tau) = \bar{x}_\pi(\tau) + \epsilon$ and $\bar{x}_{\pi'}(m_f + \tau) = \max(0, \bar{x}_\pi(m_f + \tau) - \epsilon)$, where $\epsilon = \min(C - k_\pi(m_f), N_f/m_f - \bar{x}_\pi(\tau))$. Clearly, $\bar{V}(\pi') \leq \bar{V}(\pi) + \epsilon m_f$.

Case II. $\tau = m_f + 1$**:** Define $\tau' = \min\{\ell : \bar{y}_\pi(\ell) < N_s/m_s\}$. Note that as $C < N_f + \frac{N_s}{m_s} m_f$, we have $\tau' \leq m_f$. Consider the policy π' defined as follows: $\bar{y}_{\pi'}(j) = \bar{y}_\pi(j)$ for $j \geq 1$, $j \neq \tau', m_s + \tau'$, and $\bar{x}_{\pi'}(j) = \bar{x}_\pi(j)$ for $j > 1$, $j \neq m_f + 1$, $x_{\pi'}(1) = \frac{N_f}{m_f} - \epsilon$ where $0 < \epsilon < \min\left(C - k_\pi(m_f), \frac{1}{2}\left(\frac{N_f}{m_f} - \frac{N_s}{m_s}\right), \frac{N_s}{m_s} - \bar{y}_\pi(\tau')\right)$, $y_{\pi'}(\tau') = y_\pi(\tau') + \epsilon$, $x_{\pi'}(m_f + 1) = \epsilon$, and $y_{\pi'}(m_s + \tau') = y_\pi(m_s + \tau') - \epsilon$. Under the policy π', in the first period, we reduce the allocation to the fast customers by ϵ and increase the allocation of the slow customers in the period τ' by ϵ. The fact that $\epsilon < \frac{1}{2}\left(\frac{N_f}{m_f} - \frac{N_s}{m_s}\right)$ and the fact that $\bar{y}_{\pi'}(m_f + 1) \leq \frac{N_s}{m_s}$ ensures that the allocation $\bar{y}_{\pi'}(m_f + 1)$ and $\bar{x}_{\pi'}(m_f + 1)$ is still feasible in the modified policy. Thus, under this modified policy the number of fast customer requests denied increases by ϵ, whereas the number of slow customer requests denied decreases by ϵ, and we obtain $\bar{V}(\pi') = \bar{V}(\pi) - (m_s - m_f)\epsilon < \bar{V}(\pi)$. Thus, the policy π cannot be optimal. This completes the proof. □

Proof of Lemma 2. We begin by proving part 1. Fix any policy $\pi \in \bar{\Pi}_\theta$. The definition of θ implies that

$\sum_{j=1}^{m_f} \bar{y}_\pi(j) = \theta$. Further, using the capacity constraint at period m_f, we also have

$$\sum_{j=1}^{m_f} \bar{x}_\pi(j) \leq C - \theta. \tag{35}$$

Then, using the definition of the objective function (26), we have

$$\begin{aligned} \bar{V}(\pi) &\overset{(a)}{\geq} \sum_{j=1}^{m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_\pi(j) \right) + \sum_{j=1}^{m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_\pi(j) \right) \\ &\overset{(b)}{\geq} m_f(N_f - C + \theta) + m_s \left(\frac{N_s}{m_s} m_f - \theta \right) \\ &= m_f(N_f - C) + N_s m_f - \theta(m_s - m_f), \end{aligned} \tag{36}$$

where (a) follows by truncation and (b) by using (35). Thus, the first part of the result follows.

Consider any policy $\pi \in \Pi_\theta$ such that $\Delta(\pi) = 1$ and $k_\pi(m_f) = C$. Note that $\Delta(\pi) = 1$ ensures that the first inequality (a) in (36) actually holds as equality and $k_\pi(m_f) = C$ ensures that (35) holds with equality, and thus the inequality (b) in (36) holds as equality as well. This implies that $\bar{V}(\pi) = m_f(N_f - C) + N_s m_f - \theta(m_s - m_f)$ and completes the proof of part 1. Part 2 follows immediately from part 1. □

Proof of Lemma 3. Pick any admissible policy π. Then, we have the following bound on the cost function

$$\begin{aligned} \bar{V}(\pi) &\overset{(a)}{\geq} \sum_{j=1}^{m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_\pi(j) \right) + \sum_{j=1}^{m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_\pi(j) \right) + \sum_{j=m_f+1}^{m_s} m_s \left(\frac{N_s}{m_s} - \bar{y}_\pi(j) \right) \\ &\overset{(b)}{\geq} m_f(N_f - (k_\pi(m_f) - \Theta(\pi))) + m_s(\frac{N_s}{m_s} m_f - \Theta(\pi)) + m_s(\beta(\pi^*) - \beta(\pi))^+ \\ &= m_f(N_f + \beta(\pi) - C + \Theta(\pi))) + m_s(\frac{N_s}{m_s} m_f - \Theta(\pi)) + m_s(\beta(\pi^*) - \beta(\pi))^+, \end{aligned} \tag{37}$$

where (a) and the relation for the first two terms in (b) follow as in Lemma 2.1. The bound on the third term in (b) follows with the observation that the returns during $m_f + 1$ and m_s from the fast customers is bounded above by $\frac{N_f}{m_f}(m_s - m_f)$, the number of requests from slow customers between $m_f + 1$ and m_s equals $\frac{N_s}{m_s}(m_s - m_f)$, which by the construction of π^* equals $\beta(\pi^*) + \frac{N_f}{m_f}(m_s - m_f)$. As the capacity on hand after m_f periods under policy π is $\beta(\pi)$, we obtain the bound $\sum_{j=m_f+1}^{m_s} m_s \left(\frac{N_s}{m_s} - \bar{y}_\pi(j) \right) \geq m_s(\beta(\pi^*) - \beta(\pi))^+$. □

B Proof of Propositions

Proof of Proposition 2. **1.** These convergences follow by the use of the strong law of large numbers. Consider the convergence in (11). For convenience, we relabel the random variables corresponding to the customers' rental durations $\{v_{ik}\}_{k=1}^{\infty}$ for $i = 1, \ldots, nN_f + nN_s$ as $\{U_\ell^a(j), U_\ell^r(j), V_\ell^a(j), V_\ell^r(j)\}_{j=1}^{\infty}$, where in each period j, $U_\ell^a(j)$ denotes the rental duration of the ℓ^{th} fast customer to be allocated the new product in this period, while $U_\ell^r(j)$ denotes the rental duration of the ℓ^{th} fast customer whose request is rejected (denied) in this period. Thus, $U_\ell^a(j)$ and $U_\ell^a(j)$ are distributed according to the distribution F_f for all $j, \ell \geq 1$. $V_\ell^{\cdot}(j)$ are defined analogously for the slow customers.

We denote $\{\tilde{R}_\ell\}_{\ell=1}^{nN_f}$ as the initial residual rental durations of the fast customers. Further, denote $\mathcal{N}_{\pi^n,f}(j)$ as the set of fast customers whose request was rejected in period j. Using this notation for any period j, we can write

$$\frac{D_{\pi^n,f}(j)}{n} = \sum_{\ell=1}^{nN_f} \frac{\mathbb{I}\{\tilde{R}_\ell = j\}}{n} + \sum_{k=1}^{j-1} \sum_{\ell=1}^{\#\mathcal{N}_{\pi^n,f}(k)} \mathbb{I}\{U_\ell^r(k) = j-k\}. \tag{38}$$

We shall prove the result by induction on j. Note that for $j=1$, the convergence (11) follows immediately by the strong law. Suppose (11) holds for all $j \le m$ for some $m \ge 1$. We now show that (11) holds for $j = m+1$. Note that for each $k \le m$, we can write $\#\mathcal{N}_{\pi^n,f}(k) = D_{\pi^n,f}(k) - X_{\pi^n}(k)$. Thus, by the inductive hypothesis and (10), we obtain $\frac{\mathcal{N}_{\pi^n,f}(k)}{n} \to \bar{d}_{\pi,f}(k) - \bar{x}_\pi(k)$. Using this convergence in (38), the convergence in (11) immediately follows. Arguing similarly, we obtain the convergences in (12) and (14). Note that (13) follows immediately from (10).

2. We first consider the case $\sum_{j=1}^\infty \bar{x}_\pi(j) + \sum_{j=1}^\infty \bar{y}_\pi(j) < N_f + N_s$, i.e., 'asymptotically' all customers are not allocated the new product. Then, we have that $\bar{V}(\pi) = \infty$ and the existence of an $\epsilon > 0$ such that

$$\bar{\tau} := \inf\left\{\ell : N_f - \sum_{k=1}^{\ell} \bar{x}_\pi(k) \le \epsilon,\ N_s - \sum_{k=1}^{\ell} \bar{y}_\pi(k) \le \epsilon\right\} = \infty. \tag{39}$$

Defining

$$\tau^n = \inf\left\{\ell : nN_f - \sum_{k=1}^{\ell} X_{\pi^n}(k) \le n\epsilon,\ nN_s - \sum_{k=1}^{\ell} Y_{\pi^n}(k) \le n\epsilon\right\},$$

the following result holds.

Lemma 4. *We have* $\liminf_{n\to\infty} \tau^n = \infty$ *a.s.*

We also have

$$V(\pi^n) + nN_s m_s + nN_f m_f \ge (\epsilon n - 1)\mathbb{E}\tau^n,$$

where the left hand side is the expected mean time (instead of delay) of obtain the new product, and the inequality follows as in period τ^n there are at least $\lfloor \epsilon n \rfloor$ remaining customers, each of whom have not received the new product by period τ^n.

Thus, appealing to Lemma 4 and Fatou's lemma we have that

$$\liminf_{n\to\infty} \frac{V(\pi^n)}{n} = \infty.$$

As we have $\bar{V}(\pi) = \infty$, this completes the proof.

Next we consider the case $\sum_{j=1}^\infty \bar{x}_\pi(j) = N_f$ and $\sum_{j=1}^\infty \bar{y}_\pi(j) = N_s$. Fix an $\omega \in \Omega$ and any $\epsilon > 0$ such that (10) holds. We modify the definition of $\bar{\tau}$ as follows

$$\bar{\tau} = \inf\left\{\ell : N_f - \sum_{k=1}^{\ell} \bar{x}_\pi(k) < \epsilon,\ N_s - \sum_{k=1}^{\ell} \bar{y}_\pi(k) < \epsilon \text{ and } C + \sum_{k=1}^{\ell} \bar{b}_{\pi^n}(k) - \sum_{k=1}^{\ell} (\bar{x}_\pi(k) + \bar{y}_\pi(k)) > 2\epsilon\right\}.$$

Let $\mathcal{Q}_{\pi^n}(j) = \left(\cup_{\ell=1}^j \mathcal{A}_{\pi^n}(\ell)\right)^c$ denote the set of customers who have not been allocated the new product by time j. Noting the convergence in (10) and (14), for large n we have $nN_f - \sum_{k=1}^{\bar{\tau}} X_{\pi^n}(k) < \epsilon n$, $nN_s - \sum_{k=1}^{\bar{\tau}} Y_{\pi^n}(k) < \epsilon n$ and $nC + \sum_{k=1}^{\bar{\tau}} \#\mathcal{B}_{\pi^n}(k) - \sum_{k=1}^{\bar{\tau}-1}(X_{\pi^n}(k) + Y_{\pi^n}(k)) > 2\epsilon n$. Then, the delay

incurred beyond period $\bar{\tau}$ in the n^{th} system consists of the residual rental durations of the customers who have not rented the new product by period $\bar{\tau}$, and can be written as follows:

$$\sum_{j=\bar{\tau}+1}^{\infty}(nN_f - \sum_{k=1}^{j} X_{\pi^n}(k)) + \sum_{j=\bar{\tau}+1}^{\infty}(nN_s - \sum_{k=1}^{j} Y_{\pi^n}(k)) = \sum_{i\in\mathcal{Q}_{\pi^n}(\bar{\tau})} R_i(\bar{\tau}). \tag{40}$$

Dividing both sides by n and taking limits as $n \to \infty$, we obtain

$$\limsup_{n\to\infty}\left[\sum_{j=\bar{\tau}+1}^{\infty}(N_f - \sum_{k=1}^{j}\frac{X_{\pi^n}(k)}{n}) + \sum_{j=\bar{\tau}+1}^{\infty}(N_s - \sum_{k=1}^{j}\frac{Y_{\pi^n}(k)}{n})\right] = \limsup_{n\to\infty}\sum_{i\in\mathcal{Q}_{\pi^n}(\bar{\tau})}\frac{R_i(\bar{\tau})}{n}. \tag{41}$$

We have the following bound on the term on the right hand side of this relation.

Lemma 5. *We have* $\limsup_{n\to\infty}\sum_{i\in\mathcal{Q}_{\pi^n}(\bar{\tau})}\frac{R_i(\bar{\tau})}{n} \le 2M\epsilon$.

Thus, applying Fatou's Lemma, we obtain,

$$\limsup_{n\to\infty}\mathbb{E}\left[\sum_{j=\bar{\tau}+1}^{\infty}(N_f - \sum_{k=1}^{j}\frac{X_{\pi^n}(k)}{n}) + \sum_{j=\bar{\tau}+1}^{\infty}(N_s - \sum_{k=1}^{j}\frac{Y_{\pi^n}(k)}{n})\right] \le 2M\epsilon. \tag{42}$$

Using this result, we obtain

$$\begin{aligned}
&\limsup_{n\to\infty}\left[\frac{V(\pi^n)}{n} - \left(\mathbb{E}\left[\sum_{j=1}^{\bar{\tau}}(N_f - \sum_{k=1}^{j}\frac{X_{\pi^n}(k)}{n}) + \sum_{j=1}^{\bar{\tau}}(N_s - \sum_{k=1}^{j}\frac{Y_{\pi^n}(k)}{n})\right] - N_f m_f - N_s m_s\right)\right] \\
&= \limsup_{n\to\infty}\mathbb{E}\left[\sum_{j=\bar{\tau}+1}^{\infty}(N_f - \sum_{k=1}^{j}\frac{X_{\pi^n}(k)}{n}) + \sum_{j=\bar{\tau}+1}^{\infty}(N_s - \sum_{k=1}^{j}\frac{Y_{\pi^n}(k)}{n})\right] \\
&\le 2M\epsilon.
\end{aligned} \tag{43}$$

In addition, using (10) and dominated convergence we have

$$\lim_{n\to\infty}\mathbb{E}\left[\sum_{j=1}^{\bar{\tau}}(N_f - \sum_{k=1}^{j}\frac{X_{\pi^n}(k)}{n}) + \sum_{j=1}^{\bar{\tau}}(N_s - \sum_{k=1}^{j}\frac{Y_{\pi^n}(k)}{n})\right] = \left[\sum_{j=1}^{\bar{\tau}}(N_f - \sum_{k=1}^{j}\bar{x}_\pi(k)) + \sum_{j=1}^{\bar{\tau}}(N_s - \sum_{k=1}^{j}\bar{y}_\pi(k))\right]. \tag{44}$$

Lemma 6. *We have*

$$\left[\sum_{j=\bar{\tau}+1}^{\infty}(N_f - \sum_{k=1}^{j}\bar{x}_\pi(k)) + \sum_{j=\bar{\tau}+1}^{\infty}(N_s - \sum_{k=1}^{j}\bar{y}_\pi(k))\right] \le 2M\epsilon. \tag{45}$$

Combining (43-45) and definition of $\bar{V}(\pi)$ in (20) we have

$$\limsup_{n\to\infty}\left|\bar{V}(\pi) - \frac{V(\pi^n)}{n}\right| \le 4M\epsilon,$$

Noting the choice of ϵ was arbitrary, the result follows.

□

Proof of Proposition 3. We first note that we must have $\limsup_{n\to\infty} V(\pi^n) \ge \bar{V}(\pi^*)$. To see why this relation holds, assume the contrary. Then, we obtain a subsequence $\{\pi^{n_m}\}_{m=1}^{\infty}$ such that $\lim_{m\to\infty} V(\pi^{n_m}) < \bar{V}(\pi^*)$. Noting that $\frac{X_{\pi^{n_m}}(j)}{n} \le N_f$ and $\frac{Y_{\pi^{n_m}}(j)}{n} \le N_s$ for $j = 1, 2, \ldots$, we can apply a diagnolization argument to obtain a convergent subsubsequence $\{\pi^{n_{m_\ell}}\}_{\ell=1}^{\infty}$. Let $\bar{\pi}$ denote the limiting policy. Then, an

application of Proposition 2.2 gives us $\lim_{\ell\to\infty} V(\pi^{n_{m_\ell}}) = \bar{V}(\bar{\pi})$. It can be verified that $\bar{\pi}$ satisfies the constraints in (21). Thus, we obtain a contradiction to the optimality of π^*.

To complete the proof we need to prove that $\limsup_{n\to\infty} V(\pi^{*n}) = \bar{V}(\pi^*)$. An iterated application of the strong law, as in the proof of Proposition 2.1, gives us $\lim_{n\to\infty} \frac{X_{\pi^{*n}(j)}}{n} = \bar{x}_{\pi^*}(j)$ and $\lim_{n\to\infty} \frac{Y_{\pi^{*n}(j)}}{n} = \bar{y}_{\pi^*}(j)$, and thus applying Proposition 2.2, we obtain $\lim_{n\to\infty} \frac{V(\pi^{*n})}{n} = \bar{V}(\pi^*)$.

□

Proof of Proposition 4. We begin by noting that the following observations follow immediately by the construction of π^*.

Lemma 7. *The allocation rule π^* has the following properties:*

1. *π^* is an admissible policy, i.e., $\pi^* \in \bar{\Pi}_{\theta^*}$.*
2. *$\Delta(\pi^*) = 1$.*
3. *$k_{\pi^*}(m_f) = C$.*
4. *Either $k_{\pi^*}(j) = C$ for $m_f \le j \le 2m_f$ or $\bar{y}_{\pi^*}(j) = \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$.*

Note that part 4 of this result states that if any slow customer request is rejected, it must be the case that all capacity is in use between periods m_f and $2m_f$. We will break the proof into two parts based on the condition that holds in Lemma 7.4. For the case $\bar{y}_{\pi^*}(j) = \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$, the result follows immediately from Lemma 2.2.

We now turn to the case $k_{\pi^*}(j) = C$ for $m_f \le j \le 2m_f$. We first prove part 1 of the result. We begin by proving that $\pi' \in \bar{\Pi}_\theta$ where $\theta \neq \theta^*$ cannot be optimal.

Pick any $\pi' \in \bar{\Pi}_\theta$ for $\theta < \theta^*$. Applying Lemmas 2.1 and 7, we have $\bar{V}(\pi') \ge m_f(N_f - C) + N_s m_f - \theta(m_s - m_f) > m_f(N_f - C) + N_s m_f - \theta^*(m_s - m_f) = \bar{V}(\pi^*)$.

Now, pick any $\pi' \in \bar{\Pi}_\theta$ for $\theta > \theta^*$. Define $\epsilon \equiv \theta - \theta^*$. Using Lemma 1, we can restrict our attention to the case $k_{\pi'}(m_f) = C$ without loss of generality. As $\pi' \in \Pi_{\theta^*+\epsilon}$ and $\pi^* \in \bar{\Pi}_{\theta^*}$ with $k_{\pi^*}(m_f) = C$ by construction, we must have

$$\sum_{j=1}^{m_f} \bar{x}_{\pi'}(j) = \sum_{j=1}^{m_f} \bar{x}_{\pi^*}(j) - \epsilon \tag{46}$$

$$\sum_{j=1}^{2m_f - m_s} \bar{y}_{\pi'}(j) = \sum_{j=1}^{2m_f - m_s} \bar{y}_{\pi^*}(j) - \epsilon_1 \tag{47}$$

$$\sum_{j=2m_f - m_s + 1}^{m_f} \bar{y}_{\pi'}(j) = \sum_{j=2m_f - m_s + 1}^{m_f} \bar{y}_{\pi^*}(j) + \epsilon_2, \tag{48}$$

where $\epsilon_2 - \epsilon_1 = \epsilon$, and we must have $\epsilon_1 \ge 0$ as $\bar{y}_{\pi^*}(j) = N_s/m_s$ for $j \le 2m_f - m_s$. Using (47) and the

fact that no requests are denied after period m_f, we also have

$$\begin{aligned}\sum_{j=m_f+1}^{2m_f} \bar{y}_{\pi'}(j) &\leq \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi'}(j) + \sum_{j=1}^{2m_f-m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j)\right) \leq \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi^*}(j) + \sum_{j=1}^{2m_f-m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi^*}(j)\right) + \epsilon_1 \\ &= \sum_{j=m_f+1}^{2m_f} \bar{y}_{\pi^*}(j) + \epsilon_1,\end{aligned} \tag{49}$$

where we use $\bar{y}_{\pi'}(j) \leq \frac{N_s}{m_s} = \bar{y}_{\pi^*}(j)$, for $j = m_f + 1, \ldots, m_s$.

Using Lemma 7.4, we have $k_{\pi^*}(2m_f) = C$, and thus noting $k_{\pi'}(2m_f) \leq C = k_{\pi^*}(2m_f)$, we obtain

$$\sum_{j=m_f+1}^{2m_f} \bar{x}_{\pi'}(j) + \sum_{j=2m_f-m_s+1}^{2m_f} \bar{y}_{\pi'}(j) \leq \sum_{j=m_f+1}^{2m_f} \bar{x}_{\pi^*}(j) + \sum_{j=2m_f-m_s+1}^{2m_f} \bar{y}_{\pi^*}(j). \tag{50}$$

Adding (46) and subtracting (48) from this relation, we obtain

$$\sum_{j=1}^{2m_f} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{2m_f} \bar{y}_{\pi'}(j) \leq \sum_{j=1}^{2m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{2m_f} \bar{y}_{\pi^*}(j) - \epsilon_2 - \epsilon. \tag{51}$$

Using the convention $\bar{x}.(0) = \bar{y}.(0) = 0$, we can write

$$\begin{aligned}\bar{V}(\pi') &= \sum_{j=1}^{\infty} m_f \left(\frac{N_f}{m_f} - \sum_{\ell=0}^{\lfloor j/m_f \rfloor} \bar{x}_{\pi'}(j - \ell m_f)\right) + \sum_{j=1}^{\infty} m_s \left(\frac{N_s}{m_s} - \sum_{\ell=0}^{\lfloor j/m_s \rfloor} \bar{y}_{\pi'}(j - \ell m_s)\right) \\ &\geq \sum_{j=1}^{2m_f} m_f \left(\frac{N_f}{m_f} - \sum_{\ell=0}^{\lfloor j/m_f \rfloor} \bar{x}_{\pi'}(j - \ell m_f)\right) + \sum_{j=1}^{2m_f} m_s \left(\frac{N_s}{m_s} - \sum_{\ell=0}^{\lfloor j/m_s \rfloor} \bar{y}_{\pi'}(j - \ell m_s)\right) \\ &= \sum_{j=1}^{m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j)\right) + \sum_{j=m_f+1}^{2m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j - m_f) - \bar{x}_{\pi'}(j)\right) \\ &+ \sum_{j=1}^{m_s} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j)\right) + \sum_{j=m_s+1}^{2m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j - m_s) - \bar{y}_{\pi'}(j)\right) \\ &\overset{(a)}{=} -m_f \sum_{j=1}^{m_f} \bar{x}_{\pi'}(j) + \sum_{j=1}^{2m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j)\right) + \sum_{j=m_f+1}^{2m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j)\right) \\ &+ \sum_{j=1}^{m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j)\right) - m_s \sum_{j=1}^{2m_f-m_s} \bar{y}_{\pi'}(j) \\ &\overset{(b)}{\geq} \epsilon m_f - m_f \sum_{j=1}^{m_f} \bar{x}_{\pi^*}(j) + \sum_{j=1}^{2m_f} m_f \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j)\right) + \sum_{j=m_f+1}^{2m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi^*}(j)\right) + m_f(\epsilon + \epsilon_2) - \epsilon_1(m_s - m_f) \\ &+ \sum_{j=1}^{m_f} m_s \left(\frac{N_s}{m_s} - \bar{y}_{\pi^*}(j)\right) - \epsilon m_s - m_s \sum_{j=1}^{2m_f-m_s} \bar{y}_{\pi^*}(j) + \epsilon_1 m_s \\ &\overset{(c)}{=} \bar{V}(\pi^*) + (2\epsilon + \epsilon_2 + \epsilon_1)m_f - \epsilon m_s \\ &\overset{(d)}{>} \bar{V}(\pi^*),\end{aligned} \tag{52}$$

where (a) follows by rearranging terms, (b) follows by bounding each of the terms in the relation (a). We use (46) for the first term, the relation (51) and (49) along with the fact that $m_s \geq m_f$ for the sum

of the second and third terms, the sum of (47) and (48) for the fourth term, and finally (47) for the last term. The relation (c) is a consequence of the fact that

$$\bar{V}(\pi^*) = \sum_{j=1}^{2m_f} m_f \left(\frac{N_f}{m_f} - \sum_{\ell=0}^{\lfloor j/m_f \rfloor} \bar{x}_{\pi^*}(j - \ell m_f) \right) + \sum_{j=1}^{2m_f} m_s \left(\frac{N_s}{m_s} - \sum_{\ell=0}^{\lfloor j/m_s \rfloor} \bar{y}_{\pi^*}(j - \ell m_s) \right),$$

which follows using Lemma 7.3, i.e., $\Delta(\pi^*) = 1$. Lastly, the inequality (d) follows from the fact that $2m_f > m_s$ and $\epsilon_2 \geq \epsilon > 0$.

We have established that for a policy $\hat{\pi}$ to be optimal, we must have $\hat{\pi} \in \bar{\Pi}_{\theta^*}$. Thus, we can restate the optimization problem (26) as $\min_{\pi \in \bar{\Pi}_{\theta^*}} \bar{V}(\pi)$. Noting that for two policies $\pi_1, \pi_2 \in \bar{\Pi}_\theta$ if $\Delta(\pi_1) = 1$ and $\Delta(\pi_2) \neq 1$, we must have $\bar{V}(\pi_1) < \bar{V}(\pi_2)$, it follows that $\hat{\pi}$ is a solution to this problem if and only if $\Delta(\hat{\pi}) = 1$, and thus part 1 follows.

We now prove part 2 by contradiction. Suppose there exists π such that $\Theta(\pi) > \theta^*$ and $\Delta(\pi) = 1$. Using Lemma 2.1 we have that $\bar{V}(\pi) = m_f(N_f - C) + N_s m_f - \Theta(\pi)(m_s - m_f) < m_f(N_f - C) + N_s m_f - \theta^*(m_s - m_f) = \bar{V}(\pi^*)$ which contradicts the fact that π^* is a solution to (26). □

Proof of Proposition 5. The following observations follow immediately by the construction of π^*.

Lemma 8. *The policy π^* has the following properties:*

1. *π^* is an admissible policy, i.e., $\pi^* \in \bar{\Pi}_{\theta^*}$.*
2. *$\Delta(\pi^*) = 1$.*
3. *$k_{\pi^*}(m_f) = C$.*
4. *Either $k_{\pi^*}(j) = C$ for $m_f \leq j \leq m_s - m_f$ and $j = m_s$ or $\bar{y}_{\pi^*}(j) = \frac{N_s}{m_s}$ for $j = 1, \ldots, m_s$.*

We first consider the case when $S = 0$ as obtained after completing Step 2 of the construction of π^*. For this case, one can verify that $k_{\pi^*}(j) = C$ for $m_f \leq j \leq m_s - m_f$ and $j = m_s$. We first prove part 1 of the result. We begin by proving that $\pi' \in \bar{\Pi}_\theta$ where $\theta \neq \theta^*$ cannot be optimal.

Pick any $\pi' \in \bar{\Pi}_\theta$ for $\theta < \theta^*$. Applying Lemmas 2.1 and 8, we have $\bar{V}(\pi') \geq m_f(N_f - C) + N_s m_f - \theta^*(m_s - m_f) = \bar{V}(\pi^*)$.

Pick any $\pi' \in \bar{\Pi}_\theta$ for $\theta > \theta^*$. Define $\epsilon \equiv \theta - \theta^*$. Using Lemma 1, we can restrict our attention to the case $k_{\pi'}(m_f) = C$ without loss of generality. As $\pi' \in \bar{\Pi}_{\theta^* + \epsilon}$ and $\pi^* \in \bar{\Pi}_{\theta^*}$ with $k_{\pi^*}(m_f) = C$ by construction, we must have

$$\sum_{j=1}^{m_s - 2m_f} \bar{x}_{\pi'}(j) = \sum_{j=1}^{m_s - 2m_f} \bar{x}_{\pi^*}(j) - \epsilon_1 \tag{53}$$

$$\sum_{j=m_s-2m_f+1}^{m_f} \bar{x}_{\pi'}(j) = \sum_{j=m_s-2m_f+1}^{m_f} \bar{x}_{\pi^*}(j) - \epsilon_2 \tag{54}$$

$$\sum_{j=1}^{m_f} \bar{y}_{\pi'}(j) = \sum_{j=1}^{m_f} \bar{y}_{\pi^*}(j) + \epsilon, \tag{55}$$

where $\epsilon_1 + \epsilon_2 = \epsilon$.

Using Lemma 8.3, we have $k_{\pi^*}(m_s) = C$, and thus noting $k_{\pi'}(m_s) \leq C = k_{\pi^*}(m_s)$, we obtain

$$\sum_{j=m_s-m_f+1}^{m_s} \bar{x}_{\pi'}(j) + \sum_{j=1}^{m_s} \bar{y}_{\pi'}(j) \leq \sum_{j=m_s-m_f+1}^{m_s} \bar{x}_{\pi^*}(j) + \sum_{j=1}^{m_s} \bar{y}_{\pi^*}(j), \tag{56}$$

which by using (55) gives us

$$\sum_{j=m_s-m_f+1}^{m_s} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi'}(j) \leq \sum_{j=m_s-m_f+1}^{m_s} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi^*}(j) - \epsilon. \tag{57}$$

We claim that

$$\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi'}(j) \leq \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j). \tag{58}$$

To see that this relation holds, note that for the case $\epsilon_1 \geq 0$, as $k_{\pi'}(m_f) = C$, the total allocation in periods $m_f + 1, \ldots, m_s - m_f$ must be less than the returns in periods $m_f + 1, \ldots, m_s - m_f$

$$\begin{aligned}
\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi'}(j) &\leq \sum_{j=1}^{m_s-2m_f} \bar{x}_{\pi'}(j) \\
&\overset{(a)}{=} \sum_{j=1}^{m_s-2m_f} \bar{x}_{\pi^*}(j) - \epsilon_1 \\
&\overset{(b)}{=} \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j) - \epsilon_1 \\
&\leq \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j),
\end{aligned}$$

where (a) follows from (53), and (b) follows from the definition of π^*.

For the case $\epsilon_1 < 0$, we note that the allocation to the fast customers in periods $m_f + 1, \ldots, m_f - m_s$, $\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi'}(j)$, must be less than the requests from fast customers in these periods $\sum_{j=1}^{m_s-2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j)\right)$. Thus, we obtain

$$\begin{aligned}
\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi'}(j) \leq \sum_{j=1}^{m_s-2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j)\right) &= \sum_{j=1}^{m_s-2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j)\right) + \epsilon_1 \\
&\overset{(a)}{=} \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \epsilon_1 \\
&\leq \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j),
\end{aligned}$$

where (a) follows by the definition of π^*. Further, using $\sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi'}(j) \leq \frac{N_s}{m_s}(m_s - 2m_f) = \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j)$, we obtain $\sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi'}(j) \leq \sum_{j=m_f+1}^{m_s-m_f} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j)$. Thus, the relation (58) holds.

Adding equations (53), (54), (57), (58) we obtain

$$\sum_{j=1}^{m_s} \bar{x}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi'}(j) + \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi'}(j) \leq \sum_{j=1}^{m_s} \bar{x}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s-m_f} \bar{y}_{\pi^*}(j) + \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi^*}(j) - 2\epsilon. \tag{59}$$

Using the fact that $\bar{y}_{\pi'}(j) \le \frac{N_s}{m_s} = \bar{y}_{\pi^*}(j)$ for $m_s - m_f + 1 \le j \le m_s$ and the above inequality, we have

$$\sum_{j=1}^{m_s} \bar{x}_{\pi'}(j) + 2 \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi'}(j) \le \sum_{j=1}^{m_s} \bar{x}_{\pi^*}(j) + 2 \sum_{j=m_f+1}^{m_s} \bar{y}_{\pi^*}(j) - 2\epsilon. \tag{60}$$

This immediately gives the following:

$$\begin{aligned}
&\sum_{j=m_s-m_f+1}^{2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) \right) + \sum_{j=2m_f+1}^{m_s} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) - \bar{x}_{\pi'}(j-2m_f) \right) \\
&\qquad + 2 \sum_{j=m_f+1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) \\
&\ge \sum_{j=m_s-m_f+1}^{2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j) - \bar{x}_{\pi^*}(j-m_f) \right) + \sum_{j=2m_f+1}^{m_s} \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j) - \bar{x}_{\pi^*}(j-m_f) - \bar{x}_{\pi^*}(j-2m_f) \right) \\
&\qquad + 2 \sum_{j=m_f+1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi^*}(j) \right) + 2\epsilon \\
&= 2\epsilon.
\end{aligned} \tag{61}$$

Multiplying both sides by m_f and using the fact that $2m_f \le m_s$ and $\sum_{j=m_f+1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) \ge 0$, we obtain

$$\begin{aligned}
&m_f \sum_{j=m_s-m_f+1}^{2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) \right) + m_f \sum_{j=2m_f+1}^{m_s} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) - \bar{x}_{\pi'}(j-2m_f) \right) \\
&\qquad + m_s \sum_{j=m_f+1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) \\
&\ge 2m_f\epsilon.
\end{aligned} \tag{62}$$

Arguing as in (52), we can write

$$\begin{aligned}
\bar{V}(\pi') &= \sum_{j=1}^{\infty} m_f \left(\frac{N_f}{m_f} - \sum_{\ell=0}^{\lfloor j/m_f \rfloor} \bar{x}_{\pi'}(j - \ell m_f) \right) + \sum_{j=1}^{\infty} m_s \left(\frac{N_s}{m_s} - \sum_{\ell=0}^{\lfloor j/m_s \rfloor} \bar{y}_{\pi'}(j - \ell m_s) \right) \\
&\ge m_f \sum_{j=1}^{m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) \right) + m_f \sum_{j=m_f+1}^{2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) \right) \\
&\quad + m_f \sum_{j=2m_f+1}^{m_s} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) - \bar{x}_{\pi'}(j-2m_f) \right) + m_s \sum_{j=1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) \\
&\overset{(a)}{\ge} m_f \sum_{j=1}^{m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) \right) + m_f \sum_{j=m_s-m_f+1}^{2m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) \right) \\
&\quad + m_f \sum_{j=2m_f+1}^{m_s} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) - \bar{x}_{\pi'}(j-m_f) - \bar{x}_{\pi'}(j-2m_f) \right) + m_s \sum_{j=1}^{m_f} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) \\
&\quad + m_s \sum_{j=m_f+1}^{m_s} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right)
\end{aligned} \tag{63}$$

$$
\begin{aligned}
&\overset{(b)}{\geq} m_f \sum_{j=1}^{m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi'}(j) \right) + m_s \sum_{j=1}^{m_f} \left(\frac{N_s}{m_s} - \bar{y}_{\pi'}(j) \right) + 2\epsilon m_f \\
&\overset{(c)}{=} m_f \sum_{j=1}^{m_f} \left(\frac{N_f}{m_f} - \bar{x}_{\pi^*}(j) \right) + \epsilon m_f + m_s \sum_{j=1}^{m_f} \left(\frac{N_s}{m_s} - \bar{y}_{\pi^*}(j) \right) - \epsilon m_s + 2\epsilon m_f \\
&= \bar{V}(\pi^*) + (3m_f - m_s)\epsilon \\
&\overset{(d)}{\geq} \bar{V}(\pi^*),
\end{aligned}
\tag{64}
$$

where (a) uses the fact that $\bar{x}_{\pi'}(j - m_f) + \bar{x}_{\pi'}(j) \leq \frac{N_f}{m_f}$ for $j = m_f + 1, \ldots, m_s - m_f$, (b) follows by applying (62), and (c) follows from (53-55), and (d) follows as $3m_f > m_s$.

This establishes that for a policy $\hat{\pi}$ to be optimal, we must have $\hat{\pi} \in \bar{\Pi}_{\theta^*}$. Thus, we can restate the optimization problem (26) as $\min_{\pi \in \bar{\Pi}_{\theta^*}} \bar{V}(\pi)$. It follows that $\hat{\pi}$ is a solution to this problem if and only if $\Delta(\hat{\pi}) = 1$, and thus the part 1 of the result follows.

Proof of part 2 proceeds as in the proof of Proposition 4.

We now consider the case $S > 0$ as obtained after completing Step 2 of the construction of π^*. In this case, the value of S is given by $S = \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4m_f m_s} - \left[\frac{N_f}{m_f} - \frac{1}{2} \left(\frac{N_f}{m_f} + \frac{N_s}{m_s} \right) \right] (3m_f - m_s) = \frac{(N_s - 4N_f)m_f + N_f m_s}{4m_f m_s}$. Using (30), we can calculate $\theta^* = \frac{N_s}{m_s} m_f$. Thus, noting that $k_{\pi^*}(m_f) = C$, the result immediately follows. □

Proof of Proposition 8. We can write the fraction of slow customers' requests rejected as $\rho_s = \frac{m_f}{m_s} - \frac{\theta^*}{N_s}$. We now compute an upper bound for ρ_s based on each case described in Section 4.2. Clearly, we only need to focus on the cases $\frac{N_s}{m_s} < \frac{N_f}{m_f}$ and $\frac{m_f}{m_s} > 1/3$. First, consider the case $\frac{m_f}{m_s} > 1/2$. In this case, using θ^* as given in (28), we can compute

$$
\begin{aligned}
\rho_s &= \frac{m_f}{m_s} - \min \left[\frac{C}{N_s} - \left(\frac{N_f}{2N_s} + \frac{1}{m_s} \frac{(2m_s - 3m_f)}{2} \right), \frac{m_f}{m_s} \right] \\
&= \max \left[-\frac{C}{N_s} + \frac{N_f}{2N_s} + 1 - \frac{m_f}{2m_s}, 0 \right].
\end{aligned}
\tag{65}
$$

Thus, noting that the right hand side in the above equation is maximized at $N_f = N_s = C$, and $\frac{m_f}{m_s} = 1/2$, we obtain $\rho_s \leq \frac{1}{4}$.

We now consider the case $1/3 < \frac{m_f}{m_s} \leq 1/2$. Using the argument in the last paragraph of the proof of Proposition 5, we can further consider two cases $N_f < N_s \frac{m_f}{4m_f - m_s}$, for which $\theta^* = \frac{N_s}{m_s} m_f$, and thus $\rho_s = 0$, and $N_f \geq N_s \frac{m_f}{4m_f - m_s}$, in which case the variable S has a value of zero at the end of Step 2 in the construction of π^*. In this case, we can write $\theta^* = \min \left[C - \frac{N_f}{2} - \frac{N_s}{m_s} \frac{m_f}{2} - \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4m_f m_s}, \frac{N_s}{m_s} m_f \right]$. Thus, we have

$$
\begin{aligned}
\rho_s &= \frac{m_f}{m_s} - \min \left[\frac{C}{N_s} - \frac{N_f}{2N_s} - \frac{1}{m_s} \frac{m_f}{2} - \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4m_f m_s N_s}, \frac{m_f}{m_s} \right] \\
&= \max \left[-\frac{C}{N_s} + \frac{N_f}{2N_s} + \frac{3m_f}{2m_s} + \frac{(m_s - 2m_f)(3N_s m_f - N_f m_s)}{4m_f m_s N_s}, 0 \right].
\end{aligned}
\tag{66}
$$

The right hand side of the above equation is again maximized at $N_f = N_s = C$, and $\frac{m_f}{m_s} = 1/2$, and we obtain the upper bound $\rho_s \leq \frac{1}{4}$, and the result follows. □

B.1 Proof of additional results

Proof of Lemma 4. The proof follows by contradiction. Pick an $\omega \in A \subseteq \Omega$ such that $\mathbb{P}(A) > 0$ and there exists a sequence denoted by n_m such that $\tau^{n_m}(\omega) \to \hat{\tau}(\omega) < \infty$ as $m \to \infty$. Noting that $\tau^{n_m} \in \mathbb{Z}_+$, the convergence can be equivalently stated as $\tau^{n_m}(\omega) = \hat{\tau}(\omega)$ for all $n_m \geq \hat{n}(\omega)$ sufficiently large. For ease of notation we drop the explicit reference to ω. We can write

$$\sum_{k=1}^{\tau^{n_m}} \frac{X_{\pi^{n_m}}(k)}{n_m} = \sum_{k=1}^{\hat{\tau}} \frac{X_{\pi^{n_m}}(k)}{n_m}, \text{ for } n_m \geq \hat{n}. \tag{67}$$

Then, applying (10), we obtain $N_f - \sum_{k=1}^{\hat{\tau}} \bar{x}_\pi(k) \leq \epsilon$. Combining this with a similar argument for $\sum_{k=1}^{\tau^{n_m}} \frac{Y_{\pi^{n_m}}(k)}{n_m}$, we obtain $\bar{\tau} \leq \hat{\tau}$, which contradicts (39). □

Proof of Lemma 5. Fix an $\omega \in \Omega$ such that (10) holds. Define

$$\mathcal{Q}^j_{\pi^n}(\bar{\tau}) = \{i : i \in \mathcal{Q}_{\pi^n}(\bar{\tau}) \cap (\mathcal{R}_\pi(j) - \mathcal{A}_\pi(j)),\ i \notin \cup_{k \neq j, k \leq \bar{\tau}} (\mathcal{R}_\pi(k) - \mathcal{A}_\pi(k))\},$$

and $\epsilon^n_j = \frac{\#\mathcal{Q}^j_{\pi^n}(\bar{\tau})}{n}$ for $j = 1, \ldots, \bar{\tau}$. That is, ϵ^n_j denotes the fraction of customers who have not been allocated the new product by period $\bar{\tau}$ and whose last request for the new product came in period j. The convergences in (10) and (19) imply that there exists $K < \infty$ (that depends on ω) such that for all $n > K$, we have $\sum_{j=1}^{\bar{\tau}} \epsilon^n_j \leq 2\epsilon$. Thus, $\{(\epsilon^n_1, \epsilon^n_2, \ldots, \epsilon^n_{\bar{\tau}}) : n = 1, 2, \ldots\}$ lies in a compact set, which implies that there exists a further subsequence that converges, i.e.,

$$(\epsilon^{n_m}_1, \epsilon^{n_m}_2, \ldots, \epsilon^{n_m}_{\bar{\tau}}) \to (\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_{\bar{\tau}}).$$

Thus appealing to the strong law of large numbers, we have for $j = 1, \ldots, \bar{\tau}$

$$\lim_{m\to\infty} \sum_{i \in \mathcal{Q}^j_{\pi^{n_m}}(\bar{\tau})} \frac{R_i(\bar{\tau})}{n_m} = \lim_{m\to\infty} \frac{\sum_{i \in \mathcal{Q}^j_{\pi^{n_m}}(\bar{\tau})} R_i(\bar{\tau})}{\#\mathcal{Q}^j_{\pi^{n_m}}(\bar{\tau})} \frac{\#\mathcal{Q}^j_{\pi^{n_m}}(\bar{\tau})}{n_m} = \mathbb{E}[v_{i1} - (\bar{\tau} - j) | v_{i1} > (\bar{\tau} - j)]\bar{\epsilon}_j \leq M\bar{\epsilon}_j.$$

The result now follows by summing over all $j = 1, \ldots, \bar{\tau}$ and using the fact that $\sum_{j=1}^{\bar{\tau}} \bar{\epsilon}_j \leq 2\epsilon$. □

Proof of Lemma 6. Let r^n denote the time period of the last rejection, i.e., $r^n := \sup\{\ell : \mathcal{R}_{\pi^n}(\ell) \neq \mathcal{A}_{\pi^n}(\ell)\}$. Then, by the definition of $\bar{\tau}$ and the convergence (10) applied at period $\bar{\tau}$, we obtain $\limsup_{n\to\infty} r^n \leq \bar{\tau}$. Noting that for $j \geq r^n$, $X_{\pi^n}(j) = D_{\pi^n,f}(j)$ and $Y_{\pi^n}(j) = D_{\pi^n,s}(j)$, we obtain $\bar{x}_\pi(j) = \bar{d}_{\pi,f}(j)$ and $\bar{y}_\pi(j) = \bar{d}_{\pi,s}(j)$ for $j \geq \bar{\tau}$. Thus, we can write

$$\begin{aligned} \sum_{j=\bar{\tau}+1}^{\infty} (N_f - \sum_{k=1}^{j} \bar{x}_\pi(k)) + \sum_{j=\bar{\tau}+1}^{\infty} (N_s - \sum_{k=1}^{j} \bar{y}_\pi(k)) &= \sum_{j=\bar{\tau}+1}^{\infty} (j - \bar{\tau})(\bar{d}_{\pi,f}(j) + \bar{d}_{\pi,s}(j)) \\ &\leq 2\epsilon \sup_{\ell \leq \bar{\tau}, \alpha_i = s,f} \mathbb{E}[v_{i1} - \ell | v_{i1} > \ell], \end{aligned} \tag{68}$$

where the inequality follows by noting that using (16), for the fast customers we can write $\sum_{j=\bar{\tau}+1}^{\infty}(j - \bar{\tau})\bar{d}_{\pi,f}(j) = \sum_{j=\bar{\tau}+1}^{\infty}(j-\bar{\tau})N_f p^e_{j,f} + \sum_{k=1}^{\bar{\tau}}(j-\bar{\tau})p_{j-k,f}(\bar{d}_{\pi,f}(k) - \bar{x}_\pi(k)) \leq \epsilon \sup_{\ell \leq \bar{\tau}} \mathbb{E}[v_{i1} - \ell | v_{i1} > \ell, \alpha_i = f]$. A similar relation holds for the slow customers using (17). □

References

Armony, M. (2005), 'Dynamic routing in large-scale service systems with heterogenous servers', *Queueing Systems* **51**, 287–329.

Ata, B. & Kumar, S. (2005), 'Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies.', *Annals of Applied Probability* **15**, 331–391.

Bassamboo, A., Harrison, J. & Zeevi, A. (2006), 'Dynamic routing and admission control in high volume service systems: Asymptotic analysis via multi-scale fluid limits.', *Queueing Systems Theory and Applications* **51**(2), 249–285.

Bassamboo, A., Kumar, S. & Randhawa, R. S. (2007), 'Dynamics of new product introduction in closed rental systems', *Submitted* .

Bell, S. L. & Williams, R. J. (2001), 'Dynamic scheduling of a system with two parallel servers in heavy trafc with resource pooling: Asymptotic optimality of a threshold policy.', *Annals of Applied Probability* **11**, 608–649.

Cox, D. R. & Smith, W. L. (1961), *Queues*, Methuen, London.

Dai, J. G. & Lin, W. (2005), 'Maximum pressure policies in stochastic processing networks', *Operations Research* **53**(2), 197–218.

Gurvich, I. & Whitt, W. (2007), 'Scheduling flexible servers with convex delay costs in many-server service systems', *Manufacturing and Service Operations Management, To appear* .

Halfin, S. & Whitt, W. (1981), 'Heavy-traffic limits for queues with many exponential servers', *Operations Research* **29**, 567–588.

Harrison, J. M. & Wein, L. M. (1990), 'Scheduling networks of queues: heavy traffic analysis of a two-station closed network', *Operations Research* **38**, 1052–1064.

Kumar, S. (2000), 'Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality', *Annals of Applied Probability* **10**(3), 930–961.

Maglaras, C. (2000), 'Discrete-review policies for scheduling stochastic networks: Trajectory tracking and uid-scale asymptotic optimality.', *Annals of Applied Probability* **10**, 897–929.

Mandelbaum, A. & Stolyar, A. (2004), 'Scheduling flexible servers with convex delay costs: Heavy traffic optimality of the generalized $c\mu$-rule', *Operations Research* **52**, 836–855.

Mortimer, J. (2004), 'Vertical contracts in the video rental industry', *Working paper* .

Nazarathy, Y. & Weiss, G. (2007), 'Near optimal control of queueing networks over a finite time horizon', *Submitted* .

Savin, S. V., Cohen, M. A., Gans, N. & Katalan, Z. (2005), 'Capacity management in rental businesses with two customer bases', *Operations Research* **53**(4), 617–631.

Smith, W. E. (1956), 'Various optimizers for single-stage production', *Naval Research Logistics Quart.* **3**, 59–66.

Tang, C. S. & Deo, S. (2007), 'Rental price and rental duration under retail competition', *European Journal of Operational Research, To appear* .

Tezcan, T. (2007), 'Asymptotically optimal control of many-server heterogeneous service systems with hyper-exponential service times', *Submitted* .

Tezcan, T. & Dai, J. (2006), 'Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic', *Submitted* .

van Mieghem, J. (1995), 'Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule', *Annals of Applied Probability* **5**(3), 808–833.

Weiss, G. (2007), 'A simplex based algorithm to solve separated continuous linear programs', *Submitted* .

Wolff, R. W. (1989), *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Inc.