Early Reservation for Follow-up Appointments in a Slotted-Service Queue

Yichuan Ding
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada, daniel.ding@mcgill.ca

Diwakar Gupta
McCombs School of Business, University of Texas, Austin, TX 78712, diwakar.gupta@mccombs.utexas.edu

Xiaoxu Tang
Corporate Model Risk, Wells Fargo Bank, Minneapolis, MN 55414,xiaoxu.tang@wellsfargo.com

We study an appointment-based slotted-service queue with the goal of maximizing service volume. Returning customers prefer to be served by the same service agent that they visited in their previous visit. Applications of this model include a whole host of medical clinics, lawyers, councillors, tutors, and government officials who deal with the public. We consider a simple strategy that a service provider may use to reduce balking among returning customers – designate some returning customers as high-priority customers. These customers are placed at the head of the queue when they call for a follow-up appointment. In an appointment-based system, this policy can be implemented by booking a high-priority returning customer's appointment right before she leaves the service facility. We focus on a need-based policy in which the decision to prioritize some customers depends on their return probability. We analyze three systems, an open-access system, a traditional appointment system, and a carve-out system. We show that in an open-access system, the service provider should never prioritize returning customers in order to maximize the throughput rate. However, it is always optimal to prioritize some customers in a traditional appointment system. In the carve-out system, which may be modeled as a system with two parallel queues, the optimal prioritized follow-up appointments booking policy varies depending on which queue is more congested. In the traditional systems, we prove that the throughput rate is a quasi-concave function of the threshold under the assumption that returning customers see time averages (RTA). This allows service systems to determine optimal operating policies that are both easy to implement and provably optimal.

Key words: Appointment Scheduling, Re-entrant Queue, Returning Customers See-Time-Average, Infinitesimal Perturbation Analysis

1. Introduction

Many service providers divide their consult time into slots and require customers to book appointments only in these slots, giving rise to slotted service systems. Examples include lawyers, councillors, tutors, government officials who deal with the public, and health professionals such as doctors, dentists, psychiatrists, and physical therapists. In several such settings, returning customers have a preference for consulting the same service provider that they met in their earlier visit. For example, patients may need to consult with their physical therapists several times before they feel well

enough to continue their treatments, exercises, and/or lifestyle changes on their own. These types of systems can be modeled as appointment-based queues with slotted service and returning customers. Because the motivation for this study comes from outpatient clinics, we use the terms "customers" and "patients" interchangeably throughout the paper. In non-revenue-oriented systems, the service provider may wish to complete as many service requests as possible to maximize social welfare. In revenue-oriented systems, if the service content is standard and each appointment generates approximately the same fee, then the service provider maximizes revenue by maximizing throughput. Therefore, our research objective is to characterize policies that maximize throughput in single-server appointment-based queues with slotted service and returning customers.

We investigate a strategy that involves reserving a potential follow-up appointment right after the customer's previous visit. Appointments booked using this strategy are hereafter referred to as the *prioritized follow-up* (PFU) appointments. Alternatively, one may interpret PFU appointments as being equivalent to designating some returning customers as high-priority customers, who are placed at the head of the backlog queue when they need an appointment. PFU is commonly used in the practice. For example, such a strategy was used in many outpatient clinics in a large health system in Minnesota (see Section 3 for details). Viewed in this light, our research objective is to determine which customers should have high priority to maximize throughput rate. Characterizing an optimal policy is challenging and involves trading off two opposing effects of booking PFU appointments: (1) a *holding effect*, through which the match between returning customers and service agents is improved, and (2) a *blocking effect*, through which spoilage is increased, leading to lower throughput.

The holding effect: Usually, patients with the potential need for a follow-up visit are advised to observe their health conditions at home for a period, which we refer to as the *observation period*. After the observation period, they decide based on whether the reason for their original visit is resolved or not, that is, if they need a follow-up appointment or not. Unfortunately, by the time a returning customer calls to book a follow-up appointment with his or her preferred service agent, there may exist a backlog of appointments for that server. The customer may balk upon anticipating a long wait, which results in revenue loss. If the customer visits a substitute provider, then the latter has to spend time learning the customer's information, which also increases cost. Such service disruption is more costly in the primary care setting in which interpersonal continuity of care (COC) is a core element of high-quality service (Shin et al. 2014). Booking a PFU appointment in advance, however, allows some returning customers to secure their slots in the queue, which prevents them from balking and reduces the potential costs associated with the disruption of service continuity.

The blocking effect: Whereas PFU appointments reduce service disruption costs, they can also lead to greater spoilage. This happens because at the time a slot is held for the high-priority returning customer, he or she does not know if follow-up appointment will be needed. That becomes known only at the end of the observation period. If returning customers cancel their appointments late, or in some cases forget to cancel and no-show, the appointment slot intended for them will go unused, causing spoilage. Late cancellations and no-shows are common in outpatient appointment systems, see for example, Dixon et al. (2010) and Liu et al. (2015). Furthermore, we present empirical evidence in Section 3 that spoilage rate is higher among PFU customers. Thus, holding slots for PFU customers increases backlog, resulting in a higher balking rate of regular returning customers and episode-initiating customers, and potentially a lower average throughput rate.

We develop mathematical models of slotted-service queues with returning customers under three different appointment regimes to determine if and when PFU should be booked. The first system, has limited buffer capacity, and customers balk if the buffer is full. The second system has no size limit on the waiting room and customers balk according to a queue-length dependent balking probability. The third system consists of two parallel queues whose properties are the same as those of the queues in the first and the second system, respectively. The three systems represent an open-access system, a traditional appointment system that allows patients to book in advance, and a carve-out system with certain appointments reserved for walk-in appointments (late-arriving requests) and others reserved for patients who book in advance (Dobson et al. 2011).

We assume that the service provider can estimate each customer's revisit probability and uses that as a criterion to determine whether a potential returning customer should be designated high-priority or not. We show that the effective throughput rate is always maximized by a threshold-type policy, i.e., a returning customer gets prioritized only if his revisit probability is above a threshold. We restrict our study to the class of *fixed-threshold PFU policy*, i.e. the service provider does not use the size of the current queue or anticipated future arrivals to adjust this threshold. The rationale is that although a state-dependent threshold policy may achieve a higher effective throughput rate (net of cancellations and no-shows), non-need-based priority schemes are difficult to implement in service systems involving humans, because of ethical and fairness concerns. In contrast, people do not consider need-based criterion as being unfair (Larson 1987).

Slots may go unused primarily because of two reasons: (a) customers may realize very close to the appointment date that they do not need the appointment (e.g., when patients get better) resulting in an unusable late cancellation or no-show, and (b) customers may encounter a conflict or forget the appointment, resulting in spoilage. In this paper, we do not differentiate between these two situations but refer to those unusable slots as spoilage. Scenario (a) is more likely to happen among appointments booked using the PFU policy, but Scenario (b) may occur equally among all types

of appointments. The key tradeoff in this paper comes from the greater spoilage rate caused by higher late cancellation rate among the PFUs on the one hand, and the higher balking rates among the remaining customers on the other. We assume that the spoilage rates among regular-returning and episode-initiating appointments are the same. This assumption is supported by our data, as shown in Section 3.

We show that booking PFUs is never a good option in a limited waiting buffer (i.e., an open-access system) because the holding effect does not pay off. In contrast, in the unlimited waiting-buffer (traditional) system with state-dependent balking, prioritizing a certain proportion of customers maximizes the effective throughput rate. Furthermore, we prove that under some mild conditions, the effective throughput rate is a quasi-concave function of the proportion of returning customers being prioritized, which leads to a simple method for identifying the optimal threshold needed to implement the PFU policy.

We call a queue a slotted-service queue if (1) each customer's service time equals a unit of time (referred to as a "slot"), and (2) start times of appointments are integers. The analysis of slotted-service queues with re-entrant customers is a known difficult problem. Closed-form expressions for the steady-state distribution of the number in queue are generally not available. To overcome this difficulty, we develop an analytical framework that builds on the *returning customer see time average* (RTA) approximation, which was first proposed by Greenberg and Wolff (1987) for studying an M/M/c orbits' queue, and later adopted for analyzing other orbits' queueing models (Greenberg 1989, Artalejo 1995, Greenberg and Wolff 1987). An equivalent interpretation of RTA is that the re-entrant customers are assumed to arrive according to a time-homogeneous Poisson process, with the mean arrival rate equal to the inverse of the mean orbit time. Yang and Templeton (1987) and Wolff (1989) presented detailed discussions of the RTA approximation. Under this framework, we are able to derive key performance metrics for the queueing system of interest and find the optimal probability threshold for high-priority designation to maximize the system throughput rate.

A summary of the paper's contributions is as follows.

- Utilizing data from a large number of outpatient clinics, we identify outcomes associated with PFU appointments. For example, PFU appointments in the data have a lower likelihood of balking and seeking service from a different service provider, but have a higher spoilage rate. This motivates us to study the optimal strategy for booking PFU appointments to balance the holding and the blocking effects, which is a new topic in the appointment scheduling literature.
- We develop stylized models for an open-access system, a traditional appointment booking system, and a carve-out system, respectively, and characterize the optimal policy for PFU capacity control in each system. In an open access system, we show PFU should not be booked; whereas in a traditional system with advanced scheduling, under mild conditions, we

show that the throughput rate first increases, and then decreases with the amount of PFUs being booked. Finally, we analyze the optimal PFU booking policy in a carve-out system in which some no-show or late-cancellation slots can be avoided by utilizing those slots for walk-in patients. This feature reduces the blocking effect, which suggests at first glance that PFU booking should be a dominating strategy. Contrary to this intuition, our analysis reveals that booking all follow-up appointments (FUAs) as PFUs is not optimal because the optimal control policy needs to take into account load balancing between the two patient groups as well as the reduced blocking effect.

- There is no closed-form characterization for the steady-state queue length of a slotted service queue with returning customers. Thus, our proof technique, which is built on the RTA approximation and utilizes sample-path comparisons, provides a novel approach for analyzing such queueing systems.
- We used infinitesimal perturbation analysis (IPA) to characterize the second-order effect of the control parameter on the steady-state performance. The analysis may be generalized and applied to other similar stochastic problems.

The paper is organized as follows. Section 2 reviews the related literature. Section 3 presents empirical evidence from a large set of outpatient clinics in Minnesota. Section 4 presents the model formulation and shows that the optimal need-based PFU booking policy possesses a threshold structure. Sections 5, 6, and 7 characterize the optimal threshold under an open-access system, a traditional appointment system, and a carve-out system, respectively. Section 8 presents a numerical study that complements the theoretical findings and validates the robustness of key assumptions. Section 9 concludes the paper and discusses potential future research topics.

2. Literature Review

There is a rich body of literature that studies different aspects of scheduling appointment, including such issues as intra-day or inter-day scheduling, capacity planning, panel sizing, and customer priority and capacity reservation policies. We will discuss papers that are closely related to our problem, which come predominantly from the healthcare domain. Comprehensive reviews can be found in Cayirli and Veral (2003), Gupta and Denton (2008) and Erdogan and Denton (2010).

A stream of papers have studied appointment scheduling in the presence of patient no-show and late cancellations. A common strategy used in these situations is overbooking, which, however, may lead to longer patient waiting times and clinic overtime. LaGanga and Lawrence (2007) developed a scheduling model for a single server with deterministic service times and common no-show probability for all patients. They show that overbooking increases with no-show probability and overbooking is effective in mitigating the negative impact of no-shows. Liu (2010) formulated a

dynamic programming model that takes into account future demand, and state dependent cancellation and no-show rates. This model obtains the number of appointments to schedule on each day in order to optimize the long-run average of the expected net reward, when cancellation and no-show probability distributions are known. They show that a two-day booking window outperforms same-day booking. Our paper focuses on achieving the balance between the holding effect and the blocking effect by carefully allocating capacity to PFU patients.

A few papers have studied appointment scheduling with heterogeneous patient types, with different criteria for classifying patients, e.g., patients may be grouped by arrival pattern and/or cost-structure. Patrick et al. (2008) developed an MDP model to dynamically schedule patients with different priority classes based on different waiting costs. Saure et al. (2012) considered wait-time dependent no-show rates and their simulation results demonstrated that a short booking window with overbooking can provide greater benefit to a clinic than open access. Schuetz and Kolisch (2013) considered the scheduling problem of two CT-scanners in a hospital's radiology department to provide medical service to three patient groups: scheduled outpatients, non-scheduled inpatients, and emergency patients.

A number of papers have used slotted service times to model appointment booking systems. Gupta and Wang (2008) studied the revenue maximization problem using a slotted-service framework with patient choice. Gupta and Denton (2008) formulated the problem of determining optimal appointment lengths as a two-stage stochastic linear program and used a sequential bounding approach to determine upper bounds. Zhou et al. (2021) showed that under mild conditions, a schedule consisting of equal-sized slots achieves near optimal performance. Green and Savin (2008) modeled the appointment booking as an M/D/1 queue and studied the optimal panel sizing under both an open access system and a traditional appointment booking system. Wang et al. (2020) studied appointment scheduling with potential walk-in patients using a slotted queue and derived properties of the optimal schedule. Robinson and Chen (2010) compared traditional and open-access appointment scheduling policies using a slotted queue model.

In our model, patients who finish a consultation may re-enter the queue after the observation period, and can go through infinitely many such loops. This model is referred to as queues with re-entrant customers in the literature and has many applications in service operations; e.g., Armony and Maglaras (2004) and Kostami and Ward (2009). Some papers use fluid or diffusion approximations to derive asymptotic characteristics for such systems (Huang et al. 2015, Chan et al. 2014, Dobson et al. 2013). Exact analysis, however, is only available when the service time and observation period both follow exponential distributions; see e.g., Campello et al. (2017) and Yom-Tov and Mandelbaum (2014). For appointment-based queue, however, exponential service time is usually

not a valid assumption. Our paper develops a framework based on the returning-customer-see-time-average approximation to analyze the steady-state performance of a non-Markovian queue with re-entrant customers, which may be viewed as a contribution to the literature on re-entrant queues.

3. Empirical Evidence

To motivate our model, we report empirical evidence from 75 outpatient clinics in urban, sub-urban, and rural areas of Minnesota. The data consists of 623,592 appointment records over a 12-month study period. Variables include encrypted patient and physician IDs, time stamps for the appointment request, the appointment booking, and the actual appointment, appointment status (scheduled, completed, canceled, no show, etc.), patient demographics (age category, insurance carrier status, and zip code), and clinic location.

We examine each patient's consecutive appointments and find that some of those are booked on the same date as the patient's preceding appointment (called Type-1 appointments), while others are booked on a later date (called Type-2 appointments). We also find that 92% of the Type-1 appointments are booked for a date within 60 days of the preceding appointment. Furthermore, when the consecutive appointments are more than 60 days apart, only 2.16% of them are Type-1 and the rest are Type-2. Thus, the data reveals that Type-1 appointments are mostly booked for consecutive appointments within a short time interval.

Type-1 appointments are most likely related to the previous appointment, whereas Type-2 appointments may be either related or independent. Because our data does not indicate which appointments are related to which previous appointments, we used an inter-appointment time threshold of 45 days to define FUAs. The choice of this threshold is guided by two considerations. First, the objective of the paper is to study when doctors prioritize FUAs by booking them on the same day, i.e., as Type-1 appointments, which are typically booked within 60 days. Second, we want to focus on consecutive appointments that can be reliably classified as follow-ups rather than independent new visits. By choosing the 45-day threshold, we ensure that Type-2 appointments included in our study cohort are more likely to be FUAs.

Correspondingly, we define PFUs and regular follow-ups (RFUs) as the Type-1 and Type-2 FUAs that fall within the 45-day cutoff, although this definition may count certain episode-initiating (labeled as NEW) visits as RFUs. According to this definition, among a total of 623,592 appointments, there are 466,111 NEW visits, 128,130 RFUs, and 29,351 PFUs, which gives an estimated mean revisit probability of 25.25%. This number is an upper bound on the true probability as it may have overcounted the RFUs. By analyzing the appointments data, we further identified the following patterns.

1. PFUs on average have a higher spoilage rate and a lower probability of same doctor balking rate (i.e. the need to switch to another doctor).

Recall that late cancellations and no-shows are collectively referred to as spoilage because they result in wasted slots (Gallucci et al. 2014). We refer to the probability of switching to a different doctor or clinic as the (same-doctor) balking rate. We calculate the spoilage rates and same-doctor balking rates for PFU, RFU, and NEW visits when FUAs are defined using a threshold of 30, 45, 60 days, respectively. We report the results in Table 1.

We find that the spoilage rate of PFU appointments is on average 8% higher than that of RFU appointments while the chance that an RFU will not be matched with the same doctor is 30% higher than that of a PFU. A plausible explanation is that by securing a patient's access to her doctor well in advance, there is increased probability that the patient can be seen by the same doctor as her preceding appointment. In contrast, RFU patients may not be able to book with the same doctor at a later date when they call and thus have a higher chance of balking and switching to a different doctor.

Table 1 Comparison of Spoilage Rate and Same-Doctor Balking Rate for Different Types of Appointments

Cutoff		PFU	RFU	New
30 days	Spoilage Rate	18.2%	10.5%	12.9%
	Same-Doctor Balking Rate	15.0%	42.4%	N/A
45 days	Spoilage Rate	19.1%	10.7%	12.9%
	Same-Doctor Balking Rate	13.6%	42.3%	N/A
60 days	Spoilage Rate	19.3%	10.8%	13.0%
	Same-Doctor Balking Rate	13.3%	42.5%	N/A

2. Patients are likely to book NEW and RFU appointments into earlier times.

We test this claim by examining a total of 594,241 NEW and RFU appointment requests booked in our study period. We check how many of those requests booked the earliest slot, or a slot on the first day or the first week that has an available slot (but might be not the first slot on that day or week). Because the data does not indicate a doctor's available times (but only the booked slots), We determine whether an appointment was available or not by checking whether that appointment was eventually booked by some other patient. This method might slightly overestimate the probability of "booking the first slot/day/week", because a small number of slots earlier than the booked one might be available but remain unused in the end, in which case the booked appointments would still be regarded as the earliest available one according to this method. We summarize the results in Table 2. We find that the majority NEW and RFU appointments had booked a slot in the first available week, while more than half had booked a slot in the first available day. Thus, we conclude that patients are inclined to book a slot which is chronologically close to the earliest available one.

Table 2 Percent of NEW and RFU requests being booked in the earliest slot/day/week

	First Slot	First Day	First Week
Percent	35%	59%	74%

In our model, for analytic tractability, we need to make an even stronger assumption that the NEW and RFU appointments always book the earliest available slot. Although the data shows that it is not always the case, the literature (Green and Savin 2008) has shown that a slotted-service queue will exhibit similar behavior as long as the booked slot is not far from the earliest available one.

3. PFUs usually can be booked into the desired slots without further delay.

To identify the above pattern, we calculated the empirical distribution of the lead time of PFU appointments as well as the NEW and RFU appointments. The lead time of an appointment is defined as the time between the appointment request date and the appointment completion date. The empirical distribution of PFU lead times is plotted in Figure 1, and the empirical cumulative probability distribution of the lead times for RFU and NEW appointments is plotted in Figure 2.

From Figure 1, we find that 65.6% and 41.8% PFU appointments have a lead time more than one week and two weeks, respectively; whereas 80.5% and 89.3% of the NEW and RFU appointments have lead time less than one week and two weeks, respectively. This implies that when a patient attempts to book a PFU appointment in a slot one or two weeks from that day, NEW and RFU appointment request for that slot likely have not arrived at the appointment system yet. Therefore, PFU appointments usually can book the desired slot without being delayed by the backlogs of NEW and RFU appointments. We thus assume that PFUs have head-of-line priority over the NEW and RFU requests. Under this assumption, the system would behave similar to what we observed from the data.

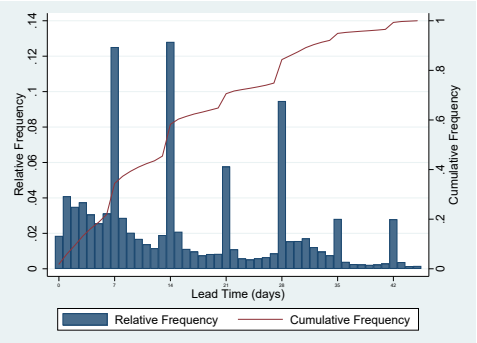


Figure 1 Histogram of Lead Times for PFU

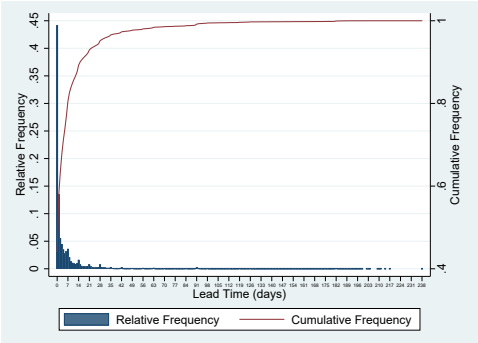


Figure 2 Histogram of Lead Times for NEW and RFU

4. Each patient's revisit probability can be predicted with standard error for the log-odds being no more than 0.537.

We fit a logistic regression model to our administrative data to predict a patient's revisit probability, i.e., the probability of needing an FUA. We have limited data elements, none of which contain clinical information. The response in the logistic regression model is defined as 1 if the patient had an FUA appointment in the next 45 days and 0 otherwise, and explanatory variables consist of the cumulative time since the start of observation period, the number of previous appointments by the same patient (which serves as a proxy for the patient's health status), the lead time (difference between appointment and book date converted to quantiles), and the insurance category. The first and the last 45 days of data were not considered due to possible censoring. This left 9 months of data, of which we used 6 months for training (269,133 observations) and 3 months for testing (135,807 observations) our model. Upon fitting the model to the data, the AUC-ROC were 0.661 and 0.651 on the training and test sets respectively. In practice, the doctor would have access to the entire medical history and clinical diagnoses for each patient, which were not included in our data. Thus, the doctor should be able to predict each patient's revisit probability with even greater accuracy.

The revisit probability $\hat{p}$ predicted by the above logistic regression has the following performance guarantee,

$$\Pr(|\ln(\frac{\hat{p}}{1-\hat{p}}) - \ln(\frac{p}{1-p})| > 0.537 z_{1-\eta/2}) \leq \eta, \tag{1}$$

where 0.537 is the maximal standard error of the log-odds among all observations in our data, and $z_{1-\eta/2}$ denotes the z-score at the $1-\eta/2$ cutoff. For example, $z_{1-\eta/2} = 1.96$ when $\eta = 0.05$. Later in the simulation experiments, we will show that even if the doctor predicts the revisit probability no more accurately than $\hat{p}$, the system behaves almost the same as the one in which the doctor can determine p accurately.

4. Model Formulation

The main decision studied in this paper is whether to book PFUs, and if yes, for which patients. The booking system used in practice is both dynamic and stochastic, making it difficult to identify optimal policies. Models with re-entrant patients are more complicated than the typical appointment scheduling models because the demand for FUAs depends on the existing bookings on each day as well as the patients that have received care but are likely to request an FUA in the future. Because of the complicated relationship between each appointment and the potential FUA demand, the optimal policy may not be of threshold type in the presence of FUAs and therefore may not permit a simple characterization, which also raises obstacles for implementation in practice. For this reason, we restrict our attention to an easy-to-implement policy that we call the *need-based FUA capacity-control policy*. Under this policy, a doctor decides whether to recommend a PFU or

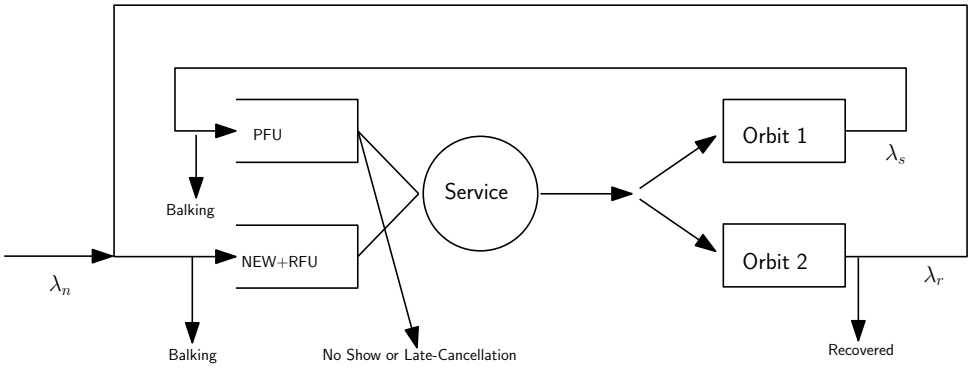
not based solely on the probability that the patient will need an FUA. In particular, the doctor does not consider the current backlog and anticipated bookings in the future.

Using the policy described above, we develop a model to calculate the throughput rate as a function of the threshold for the doctor's PFU recommendation. The model assumes that doctors know each patient's revisit probability at the end of an appointment with that patient. To test the robustness of this assumption, we show by simulation that the system behaves almost the same if the doctor can predict p with reasonable accuracy, which is supported by the empirical evidence presented in Section 3. Therefore, a need-based policy in which a doctor books a PFU only when the predicted revisit probability is above a certain threshold can be implemented.

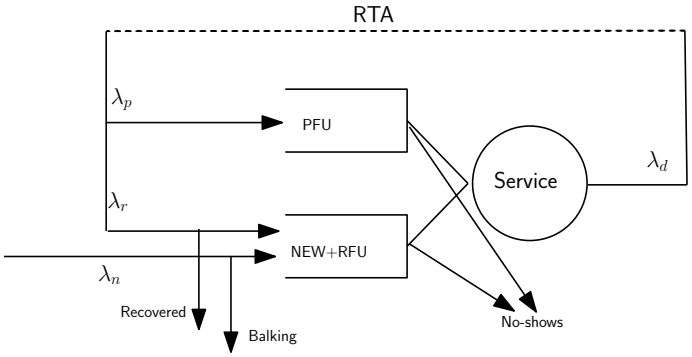
Mathematical models of service systems need to balance model fidelity and tractability. We make this tradeoff by approximating the appointment-based service system for a single doctor as a single-server priority queue with re-entrant customers, which can be either PFUs or RFUs. Each service, regardless of customer type, can be completed in a single slot. Time periods when the service system is closed (e.g. nights, weekends, and holidays) are not considered, which results in a model of the system in which a single server works continuously without vacation. Customers whose requests cannot be served immediately are placed in a queue of backlogs. For modeling purposes, there are two queues, one for PFU customers and the other for RFU and NEW customers. As demonstrated in Table 2, the FCFS service priority applies to customers in the same queue.

Figure 3(a) provides a graphic illustration of the queueing system described above. FUAs are modeled by entering each served customer into an orbit, which lasts for a random length of time that we call the *observation period*. After the observation period, one of the following four outcomes may occur: the service provider recommends a PFU and either (1) the customer needs an FUA, or (2) the customer does not need an FUA; or the service provider does not recommend a PFU and either (3) the customer needs an FUA, or (4) the customer does not need an FUA. The customer joins the higher priority PFU queue in case (1) and the lower priority queue with RFUs and NEW requests in case (3). In case (2), we further assume that with probability γ the slot may be rescued, i.e. canceled sufficiently early to avoid spoilage, hence with probability $p+(1-p)(1-\gamma)$ that PFU is actually booked; whereas in case (4) with probability one the customer leaves the system without causing spoilage. According to the above discussion, if we assume that a regular appointment has a spoilage rate $\eta \geq 0$ (because of occasional no-shows), then the PFUs with revisit probability p have spoilage rate $\frac{p\eta+(1-\gamma)(1-p)}{p+(1-\gamma)(1-p)}$, which is strictly greater than the spoilage rate η of a regular appointment as long as $\gamma < 1$, i.e., not all unneeded PFUs could be rescued.

To summarize, after completion of a service, a customer's health condition may improve, ending the episode of care, or generate an FUA in the future with a random revisit probability p. The generation of an FUA is memoryless and does not depend on how many visits the patient has



(a) The Original Model



(b) An Approximate Model Using RTA

Figure 3 The RTA Approximation for a Priority Queue with Orbits

already made. After each visit, the revisit probability p is randomly drawn from a distribution with cumulative distribution function (cdf) $F(\cdot)$, and is also independent of the system state and the number of previous visits made by that patient. We define $G(x) := \int_0^x p dF(p)$. Thus, $\overline{p} = G(1)$ denotes the average revisit probability of a randomly drawn patient from the population. We further assume that $F(\cdot)$ has a continuous density function $f(\cdot)$ on $[0, 1]$. No-shows or late-cancellations do not generate a follow-up visit.

The RTA Approximation The queueing system shown in Figure 3(a) includes multiple streams of arrivals: the NEW visits as well as those returning from different orbits. Such a system is difficult to analyze because of the complicated correlation between the number in the queue and in orbits. Analytical results are only available in a few special cases, such as the M/M/1/1 queue with geometric orbits (Cohen 1957), the M/M/1/2 queue with infinite orbits (Keilson et al. 1968), and the M/G/1/1 queue with infinite orbits (Keilson et al. 1968, Aleksandrov 1974). The model we consider is a discrete-time priority queue with heterogeneous customers and geometric orbits,

which is not included among the above-mentioned cases and happens to be analytically intractable. Therefore, we utilize the RTA approximation from the literature on retrial queues.

According to Artalejo (1995), the RTA approximation "is equivalent to keeping constant the expected number of customers in orbit when there are i customers at the waiting line", that is, putting $L_i = L$ for all i, where "L_i is the expected number of customers in orbit when there are i customers in the service facility". A direct implication of RTA is that the returning customers (in our paper, the RFUs and PFUs) arrive at the queue according to an independent Poisson process with constant rate $L\nu$, where ν is the rate for an in-orbit customer to return to the queue (so ν^{-1} denotes the mean in-orbit time). Furthermore, extensive numerical experiments in Greenberg (1989), Artalejo (1995), Greenberg and Wolff (1987) have shown that the RTA approximation yields accurate estimation of mean throughput rate, particularly when the in-orbit time is much larger than the service time. This condition is easily satisfied in our setting because the in-orbit time is several days or weeks compared to the service time of one slot (typically less than $1/8^{th}$ of a day).

Under the RTA assumption, the appointment queueing system is equivalent to a single-server priority queue with three independent Poisson arrival streams: PFUs, RFUs, and NEW visits, with mean arrival rates λ_p, λ_r, and λ_n, respectively. An extra constraint resulting from RTA is that the mean effective throughput rate λ_d should equal the mean arrival rate of the returning patients from the orbits, which is captured by expressing λ_p and λ_r as functions of λ_d. Figure 3(b) illustrates the simpler model that results from the RTA approximation.

In our model, the PFU customers have non-preemptive strict priority over other types of back-logged appointments. The rationale behind this assumption is that PFUs are usually booked many days before the appointment day, at which time the doctor's appointment book is largely open as shown in Section 3. A robustness test in Section 8 shows that imposing this assumption has minor impact on system performance.

Let λ_d denote the *average effective departure (or throughput or service) rate* which counts the average number of patients that receive service per time unit (1 unit of time = 1 slot), excluding late cancellations and no-shows. The RTA approximation requires that the departure rate must equal the arrival rate of returning patients, which leads to the following rate-balance equations,

$$\begin{aligned} \lambda_p &= \lambda_d \int_w^1 f(p)(p+(1-\gamma)(1-p))dp = \lambda_d((1-\gamma)(1-F(w))+\gamma(\overline{p}-G(w))) \\ \lambda_r &= \lambda_d \int_0^w f(p)pdp = \lambda_d G(w). \end{aligned} \tag{2}$$

where w is a threshold value chosen by the doctor. That is, a PFU is booked if and only if the probability that the patient needs an FUA is larger than w.

We define the *virtual arrival rate* λ_v as the total booking rate of all types of appointment, including PFU, RFU, and NEW visits. Note that some of the PFUs may be booked by recovered customers and turn out to be no-shows. The virtual arrival rate can be computed as

$$\lambda_v = \lambda_n + \lambda_p + \lambda_r = \lambda_n + \lambda_d((1-\gamma)(1+G(w)-F(w)) + \gamma\overline{p}), \tag{3}$$

where the second equality is obtained after plugging in the expression of λ_p and λ_r from (2).

Using a sample path argument, it can be shown that optimal need-based FUA capacity control policies are always of the threshold type. That is, every need-based FUA capacity control policy is characterized by a *PFU control threshold* $w \in [0,1]$ such that a PFU is booked if and only if the observed revisit probability $p > w$. In the rest of the paper, we will discuss how to find an optimal threshold w that maximizes the average effect service rate λ_d. We investigate this question in three different models: (1) an open access system with a limited buffer size for the total number of NEW, RFU, and PFU patients; (2) a traditional appointment system in which the arrival rate of NEW and RFU patients decreases with queue length due to state-dependent balking; and (3) a carve-out system that consists of two parallel queues as described in (1) and (2), respectively. We will next study how to characterize the steady-state effective throughput rate, and how to maximize it by choosing an optimal PFU control threshold w in these models.

5. The Open-Access System

5.1. Steady State Characterization using RTA

An open-access service system, or advance-access service system (Murray and Tantau 2000), attempts to "do all today's work today". The open-access system was proposed for outpatient care, but the idea can be generalized to many other appointment-based service systems. An open-access service system strives to keep a short waitlist by allowing customer to book their appointments only one or two days in advance. Green and Savin (2008) modeled such an open-access system using an $M/D/1/K$ queue with K standing for limit of the booking horizon. We will use a similar model to study the use of PFU strategy in an open-access system. We assume that service must start at integer times $t = 0, 1, \ldots$, which differs from that in an $M/D/1/K$ queue in which service starts either upon the completion of the previous service or upon the arrival of the first customer to an empty queue.

To analyze the open access system with a need-based PFU capacity control policy, we use $X(t)$ and $Y(t)$ to denote the total number of appointments (regardless of their types) and the number of PFU appointments waiting in the queue at time t, respectively, including the one currently being served. We assume that the system has a finite capacity $K > 0$, such that the NEW and RFU patients would balk when they arrive and see K patients in the system ($X(t) = K$). Since PFU

patients enjoy head-of-line priority, whenever they arrive and see a full buffer, they would bump out a non-PFU slot, if there is any; or balk if $Y(t) = K$ (the buffer is full of PFU slots). The assumption that a low-priority appointment (NEW or RFUs) can be bumped by PFUs builds on the fact that the latter were actually booked much earlier.

Since $X(t)$ has an upper bound K, $\{X(t) \,|\, t \geq 0\}$ is an ergodic continuous-time Markov chain (CTMC). However, because the service completes only at integer times, this CTMC is time-inhomogeneous and may not have a steady-state probability distribution in the conventional sense. In particular, the distribution of $X(t)$ always depends on $t - \lfloor t \rfloor$, the specific time in the period (a slot). However, if we define the limiting distribution as the proportion of time that $X(t)$ resides in each state, then it can be represented by a single vector $\boldsymbol{\pi}^*$, with $\pi_i^* := \int_0^1 \pi_i(t)dt$. We use a random variable X^* to represent the long-run average of $X(t)$, which has a probability distribution $\boldsymbol{\pi}^*$. Throughout this paper, we refer to $\boldsymbol{\pi}^*$ and X^* as the steady-state distribution and steady-state queue length, respectively, which may be viewed as a slight abuse of terminology. Following the definition, given a reward function $f(\cdot)$, the average reward in the long run is defined as $\mathbb{E}f(X^*) := \lim_{T\to\infty} \frac{1}{T} \int_0^T f(X(t))dt$. Its value, if finite, can be computed as

$$\mathbb{E}f(X^*) = \sum_{i\geq 0} \pi_i^* f(i) = \boldsymbol{\pi}^{*T} \boldsymbol{f}, \tag{4}$$

where $\boldsymbol{\pi}^{*T}$ denotes transpose of the vector $\boldsymbol{\pi}^*$. Note that $\pi_0(1^-) := \lim_{t\nearrow 1} \pi_0(t)$ gives the probability that the server will be idled in the next service slot, and thus $1 - \pi_0(1^-)$ gives the average (virtual) departure rate of the queue, including the no-show and late-cancelled slots. In a queue with buffer size K and arrival rate λ, we use $\rho_K(\lambda) = 1 - \pi_0(1^-)$ to denote the corresponding average throughput (departure) rate. Thus, the average departure rate of the stochastic process described by $X(t)$ is given by $\rho_K(\lambda_v)$.

Since the PFU appointments have head-of-line priority, $Y(t)$ has the same stochastic behavior as the number of jobs in a queue with buffer limit K and mean arrival rate λ_p. Therefore, the average departure rate of the stochastic process $Y(t)$ can also be described by the function $\rho_K(\lambda_p)$. The effective throughput rate can be computed by the conservation law as

$$F(\lambda_d, w) = (\rho_K(\lambda_v) - \rho_K(\lambda_p))(1-\eta) + \rho_K(\lambda_p) \frac{\int_w^1 pf(p)dp}{\int_w^1 (p + (1-p)(1-\gamma))f(p)dp}(1-\eta) \tag{5}$$

$$= \rho_K(\lambda_v)(1-\eta) - \rho_K(\lambda_p) \frac{\int_w^1 (1-p)(1-\gamma)f(p)dp}{\int_w^1 (p + (1-p)(1-\gamma))f(p)dp}(1-\eta). \tag{6}$$

In the RHS of (5), $(\rho_K(\lambda_v) - \rho_K(\lambda_p))(1-\eta)$ computes the effective throughput rate of NEW and PFUs, and $\rho_K(\lambda_p) \frac{\int_w^1 pf(p)dp}{\int_w^1 (p+(1-p)(1-\gamma))f(p)dp}(1-\eta)$ computes the effective throughput rate of PFUs. Note

that λ_v and λ_p are both functions of λ_d and w as defined in (2) and (36), so the RHS of (6) can be represented as a function of λ_d and w.

The next Lemma shows that for any fixed $w \in [0,1]$, the function $F(\cdot, w)$ has a unique fixed point. Its proof is provided in Appendix EC.1.

LEMMA 1. *Under the RTA approximation, for any given $w \in [0,1]$, there is a unique $\lambda_d \in [0,1)$ such that*

$$F(\lambda_d, w) = \lambda_d. \tag{7}$$

Suppose λ_d is the fixed point described in Lemma 1. If we use λ_d as the input rate of the returning orbit to compute the arrival rates λ_p, λ_r, and λ_v according to (2) and (3), then the effective throughput rate calculated from equation (5) is exactly λ_d. In other words, λ_d is the unique value in $[0,1]$ under which the returning orbit falls within the bounds of the conservation law. We thus call the λ_d (and the associated λ_p, λ_r, λ_v) as the *mean-preserving effective throughput rate (PFU arrival rate, RFU arrival rate, total arrival rate)*. As discussed in the literature and validated in our numerical study (Section 8), the mean-preserving effective throughput rate closely approximates the true value of the effective throughput rate, which justifies our analytical framework built on RTA.

5.2. Optimal PFU Control Threshold

Let $\lambda_d(w)$ denote the effective throughput rate under RTA for a given PFU-control threshold w. The next theorem shows that in an open access system, it is always optimal to never book any PFU if the system designer's objective is to maximize the effective throughput rate. The proof of Theorem 1 is provided in Section EC.2.

THEOREM 1. *In an open-access system with buffer capacity $K > 0$, $\lambda_d(w)$ has a unique maximizer at $w^o = 1$ (i.e., booking no PFUs).*

Theorem 1 delivers a clear message – there is no need to book any PFU appointments, even when the doctor is 100% sure that the patients will need an FUA. The rationale is that if a patient needs an FUA, she can always book one if the buffer is not full. In case the buffer is full, then some appointment requests have to be turned away anyway. Then, it makes no difference to turn away a PFU request or another appointment request (NEW or RFU). This result is based on our assumption that the service provider's objective is to maximize long-run average effective throughput rate, which also maximizes revenue. However, in some cases the service provider may care about quality-related performance metrics. An important measure of quality is the same-doctor matching rate (Hennen 1975, Rogers and Curtis 1980). In that case, the optimal policy will choose some threshold $w^o < 1$ and book some PFUs. See Appendix EC.3 for a rigorous proof.

6. Traditional Appointment Booking System

6.1. Steady State Characterization using RTA

In a traditional appointment booking system, patients can book appointments well in advance so we may assume that the buffer has infinite capacity. However, a patient is likely to balk and choose a different service provider if there are no open slots in the near future. To count the demand loss due to congestion, we assume that NEW and RFU patients have a state-dependent balking rate $b(i)$ when the current backlog $X(t)$, including the one being served, equals i. The balking rates are assumed to satisfy the following conditions:

(i) $b(i)$ is strictly increasing for $1 \le i \le L$ for some constant $L \in [1, \infty]$ and flattens after L;

(ii) $\lambda_n(\frac{1-b(\infty)}{1-\overline{p}}) < 1$.

Condition (i) covers linear or concavely increasing functions, such as $b(i) = \min\{ci, cL\}$ or $b(i) = 1-\exp(-ci)$ for some constant $c > 0$. These functional forms of balking rates are widely used in the queueing literature (Ancker Jr and Gafarian 1963a,b, Armony et al. 2009). Condition (ii) ensures that the queue is stable, i.e., the backlogs will not converge to infinity. This condition is necessary to establish positive recurrence of the Markov chain as shown in Proposition 1.

Similar to what we did in Section 5, the arrival rate of PFU and RFU (λ_p and λ_r) can be computed according to equations (2) for a given input rate for the returning orbit, λ_d. The virtual arrival rate λ_v depends on the state $X(t)$ and has the following expression,

$$\begin{aligned} \lambda_v(X(t), w, \lambda_d) &:= (\lambda_n + \lambda_r)(1 - b(X(t))) + \lambda_p \\ &= (\lambda_n + \lambda_d G(w))(1 - b(X(t))) + \lambda_d((1-\gamma)(1-F(w)) + \gamma(\overline{p} - G(w))), \end{aligned} \tag{8}$$

where the first term in the RHS counts the total arrival rates of NEW and RFUs, and the second term counts the arrival rate of PFUs. Proposition 1 next summarizes some basic properties of the time inhomogeneous CTMC $\{X(t)|t \ge 0\}$, with the proof attached in Appendix EC.4.

Proposition 1. *Suppose the balking rate $b(i)$ satisfies Condition (ii), then for any $w \in [0,1]$, $\{X(t)|t \ge 0\}$ is a positively recurrent and irreducible Markov process with period one. Its steady-state distributions can be represented by unique probability distributions $\{\boldsymbol{\pi}(t)|t \in [0,1)\}$, such that $X(s) \stackrel{d}{=} \boldsymbol{\pi}(s - \lfloor s \rfloor)$ for all $s > 0$ provided that $X(0) \stackrel{d}{=} \boldsymbol{\pi}(0)$.*

We define the steady-state distribution $\boldsymbol{\pi}^*$ and the steady-state queue length X^* as we did in Section 5. According to the RTA approximation, the effective throughput rate must satisfy the following rate-balance equation,

$$\lambda_d = (1-\eta)\mathbb{E}_{\boldsymbol{\pi}^*}\left((1 - b(X^*))(\lambda_n + G(w)\lambda_d) + (1-\eta)(\overline{p} - G(w))\lambda_d\right), \tag{9}$$

where the first term on the RHS represents the arrival rate of NEW and RFU requests, excluding those that either balk, or no-show, or late-cancel, and the second term represents the arrival rate of PFU requests, excluding no-shows and late-cancellations.

Analogous to Lemma 1, we prove in Lemma 2 that Equation (9) admits a unique fixed point.

LEMMA 2. *Given any fixed $w \in [0,1]$, there exists a unique feasible performance vector $(\lambda_d,\ \lambda_r,\ \lambda_p,\ \lambda_v,\ \boldsymbol{\pi}^*)$ with $\lambda_d \in (0,1)$.*

The proof of Lemma 2 requires Lemma 3, stated below. A proof of Lemma 3 is presented in EC.5.

LEMMA 3. *Suppose λ_d is the expected effective throughput rate corresponding to a steady-state distribution $\boldsymbol{\pi}^*$. The following conditions hold for any non-decreasing sequence $\boldsymbol{x} := (x_i)$.*

$$\nabla_w(\boldsymbol{\pi}^*)^T\boldsymbol{x} \le 0, \quad \textit{and} \quad \nabla_{\lambda_d}(\boldsymbol{\pi}^*)^T\boldsymbol{x} \ge 0. \tag{10}$$

Proof of Lemma 2: It suffices to prove that there exists a unique λ_d which solves Equation (9). Because the remaining variables in the feasible performance vector, i.e., λ_p, λ_r, and λ_v, can be calculated from λ_d by (2), (3), and (9) and the existence and uniqueness of $\boldsymbol{\pi}^*$ follow from Proposition 1. Equation (9) holds if and only if λ_d solves the following equation,

$$(V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d)) :=)\mathbb{E}_{\boldsymbol{\pi}^*}\ (1-b(X^*))(\lambda_n+G(w)\lambda_d)+(\overline{p}-G(w))\lambda_d-\frac{1}{1-\eta}\lambda_d = 0. \tag{11}$$

Therefore, it suffices to show that the above equation has a unique solution $\lambda_d \in (0,1)$. To that end, we prove that (i) function $V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d))$ is positive at $\lambda_d = 0$ and is negative when $\lambda_d \to 1$, and (ii) $V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d))$ strictly decreases in λ_d in $(0,1)$. The two claims imply that Equation (11) has exactly one solution in $(0,1)$.

We first show (i). It is straightforward to see that $V(0,\boldsymbol{\pi}^*(0)) > 0$. To show $\lim_{\lambda_d\to 1} V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d)) < 0$, we first upper bound the first two terms of $V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d))$ for all $\lambda_d \in (0,1)$ as follows,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\pi}^*(\lambda_d)}\ (1-b(X^*))(\lambda_n+G(w)\lambda_d)+(\overline{p}-G(w))\lambda_d \\ \le\ &\mathbb{E}_{\boldsymbol{\pi}^*(\lambda_d)}\ (1-b(X^*))(\lambda_n+G(w)\lambda_d)+\lambda_p \\ =\ &\mathbb{E}_{\boldsymbol{\pi}^*(\lambda_d)}\lambda_v^*(X^*,w,\lambda_d) < 1, \end{aligned} \tag{12}$$

where the first inequality follows from $(\overline{p}-G(w))\lambda_d \le \lambda_p$ because the latter also includes patients that end up not needing an FUA but forget to cancel their bookings. The equality follows from (8), and the fact that $\lambda_v^*(X^*,w,\lambda_d) < 1$ because $\{X(t)\}$ has to be positive recurrent according to Proposition 1. Then Inequality (12) implies that

$$\begin{aligned} \lim_{\lambda_d\to 1} V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d)) &= \lim_{\lambda_d\to 1}\left(\mathbb{E}_{\boldsymbol{\pi}^*(\lambda_d)}\ (1-b(X^*))(\lambda_n+G(w)\lambda_d)+(\overline{p}-G(w))\lambda_d-\tfrac{1}{1-\eta}\lambda_d\right) \\ &\le 1-\tfrac{1}{1-\eta} < 0. \end{aligned} \tag{13}$$

We next prove Claim (ii). The derivative of $V(\lambda_d,\boldsymbol{\pi}^*(\lambda_d))$ with respect to λ_d can be calculated as

$$\frac{dV}{d\lambda_d} = \frac{\partial V}{\partial \lambda_d} + (\nabla_{\boldsymbol{\pi}^*}V)^T\nabla_{\lambda_d}\boldsymbol{\pi}^*. \tag{14}$$

For the first term, we have

$$\begin{aligned}\frac{\partial V}{\partial \lambda_d} &= \mathbb{E}_{\boldsymbol{\pi}^*}\ (1-b(X^*))G(w)+\overline{p}-G(w)-\tfrac{1}{1-\eta}\\ &\le G(w)+\overline{p}-G(w)-\tfrac{1}{1-\eta}\\ &= \overline{p}-\tfrac{1}{1-\eta}<0.\end{aligned} \tag{15}$$

For the second term in the RHS of (14), since the i-th entry of the vector $\nabla_{\boldsymbol{\pi}^*}V$ is $(1-b(i))(\lambda_n+G(w)\lambda_d)$, the entries of $-\nabla_{\boldsymbol{\pi}^*}V$ must form a non-decreasing sequence. Then by invoking Lemma 3 (using $\boldsymbol{x}=-\nabla_{\boldsymbol{\pi}^*}V$), we deduce that $-(\nabla_{\boldsymbol{\pi}^*}V)^T\nabla_{\lambda_d}\boldsymbol{\pi}^*\ge 0$. Thus, the RHS of (14) is negative and V is strictly decreasing in λ_d. ■

6.2. Optimal PFU Control Threshold

Similar to Section 5, we refer to the unique fixed point λ_d of the rate-balanced condition (9) as the mean-preserving effective throughput rate. According to the RTA model, this λ_d approximates the actual effective throughput rate of the system. Let $\lambda_d(w)$ denote the mean-preserving effective throughput rate in the feasible performance vector corresponding to a PFU control-threshold $w\in[0,1]$. The main result of this paper, presented in Theorem 2, characterizes the function $\lambda_d(w)$. We let $Beta(\alpha,\beta)$ denote the Beta distribution with parameters α and β, with mean $\overline{p}=\alpha/(\alpha+\beta)$, and let $U[a,b]$ denote a uniform distribution over an interval $[a,b]\subseteq[0,1]$.

THEOREM 2. *Suppose $0\le\gamma<1$ and the revisit probability p satisfies either of the following conditions:*

- $p\sim Beta(\alpha,\beta)$ *with* $\beta>1$ *and* $\alpha/(\alpha+\beta)\le 0.25$.
- $p\sim U[a,b]$ *with* $2b^2\le b-a$.

Then $\lambda_d(w)$ is quasi-concave over $[0,1]$. There is a unique $w^*\in(0,1)$ such that $\lambda_d(w)$ strictly increases over $[0,w^*)$ and strictly decreases over $(w^*,1]$. Consequently, w^* is the unique PFU control threshold that maximizes $\lambda_d(w)$.

Before presenting its proof, we first make a few remarks concerning Theorem 2. Not only Theorem 2 implies the existence and uniqueness of an optimal PFU control threshold w^*, but also allows the system manager to search for such a w^* without knowing the parameters of the system. The manager may permit more PFUs to be booked, gradually decreasing w until the effective throughput rate $\lambda_d(w)$ stops increasing further. In this way, the manager can adopt the optimal need-based FUA capacity control policy that maximizes the effective throughput rate.

To prove Theorem 2, we need to study the second-order properties of the holding effect and the blocking effect of booking a PFU. When w increases, the marginal blocking effect has a diminishing trend because the probability that a PFU will recover during the observation period increases. A key step of the proof is to show that this trend dominates other factors, including the second

derivative of the holding effect. For that purpose, we want the holding effect to be relatively small. This can be achieved by imposing an upper bound on the average revisit probability, which is largely proportional to the holding effect. For this reason, conditions in Theorem 2 require that if p has a Beta distribution with $\beta > 1$, then $\overline{p} = \alpha/(\alpha+\beta) \le 0.25$; and if p has a uniform distribution over $[0, b]$, then $b \le 0.5$, or equivalently $\overline{p} \le 0.25$. Our empirical analysis in Section 3 provides an upper bound of 0.2525 for the mean revisit probability $\overline{p}$, which is consistent with the assumption of $\overline{p} \le 0.25$. When $\overline{p} > 0.25$, despite the lack of theoretical results, our numerical experiments show that the quasi-concavity property remains robust.

Before presenting a formal proof, we provide a sketch of our proof technique. We take the derivative of w at both sides of (9) to obtain an expression for $\lambda_d'(w)$. We show that $\lambda_d'(w) > 0$ at $w = 0$ and $\lambda_d'(w) < 0$ at $w = 1$. To show quasi-concavity, we need to study the second-order partial derivative $\lambda_d''(w)$, which, however, is difficult to analyze. In fact, we are unable to determine the sign of $\lambda_d''(w)$ for any w, and thus cannot deduce concavity of $\lambda_d(w)$. However, we can show that at any w^* such that $\lambda_d'(w^*) = 0$, the second derivative $\lambda_d''(w^*)$ must be negative. This observation, along with the fact that the continuous function $\lambda_d'(w)$ is positive at $w = 0$ and negative at $w = 1$, implies that once $\lambda_d(w)$ stops increasing in w, it must keep decreasing when w increases further. In other words, $\lambda_d(w)$ is quasi-concave in w and the w^* at which $\lambda_d'(w^*) = 0$ is the unique maximizer of $\lambda_d(w)$.

The challenge of the proof is to analyze the second derivative $\lambda_d''(w)$ and show that its value at w^* is upper bounded by zero. One of the term in the expansion of $\lambda_d''(w)$ is $(\lambda_n + G(w)\lambda_d)\sum_{i\ge 0}\frac{\partial \pi_i^*}{\partial w}(1 - b(i))$, which can be interpreted as the change rate of the virtual arrival rate with respect to w. To analyze this term, we use infinitesimal perturbation analysis (IPA) method, which has been used in the queuing literature for performance evaluation (e.g., Ho et al. 1983, Wardi et al. 2009). We use the IPA method to bound the second derivative. This proof technique may be of independent interest to researchers.

Proof of Theorem 2: We modify the definition of V function in Equation (11) by allowing it to change with w, i.e.,

$$V(\lambda_d, \boldsymbol{\pi}^*(\lambda_d, w), w) := \mathbb{E}_{\boldsymbol{\pi}^*(\lambda_d, w)}\ (1 - b(X^*))(\lambda_n + G(w)\lambda_d) + (\overline{p} - G(w))\lambda_d - \frac{1}{1-\eta}\lambda_d. \tag{16}$$

In Lemma 2, we have shown that given any w, there is a unique λ_d that solves the equation $V(\lambda_d, \boldsymbol{\pi}^*(\lambda_d, w), w) = 0$. Therefore, the function $\lambda_d(w)$ is well defined. Furthermore, as we have shown $\frac{\partial V}{\partial \lambda_d} < 0$ for all w in the proof of Lemma 2[1], the implicit function theorem states that $\lambda_d(w)$

[1] In equation (14), $\frac{\partial V}{\partial \lambda_d}$ was expressed in the form of complete derivative $\frac{dV}{d\lambda_d}$, because at that time we had not assumed that w is a variable of the function V.

has continuous derivative $\lambda_d'(w)$. The specific expression of $\lambda_d'(w)$ can be obtained by taking implicit derivative of $V(\lambda_d, \boldsymbol{\pi}^*(\lambda_d, w), w) = 0$, which gives

$$\begin{aligned} 0 &\equiv \tfrac{dV}{dw} \\ &= [\textstyle\sum_i \frac{\partial V}{\partial \pi_i^*}\frac{\partial \pi_i^*}{\partial w} + \frac{\partial V}{\partial w}] + [\frac{\partial V}{\partial \lambda_d} + \sum_i \frac{\partial V}{\partial \pi_i^*}\frac{\partial \pi_i^*}{\partial \lambda_d}]\lambda_d'(w) \\ &=: \Xi_1(\lambda_d, w) + \Xi_2(\lambda_d, w)\lambda_d'(w), \end{aligned} \tag{17}$$

where $\Xi_1(\lambda_d, w)$ and $\Xi_2(\lambda_d, w)$ denote the two terms enclosed in square brackets $[\cdot]$ in equation (17). We can then express $\lambda_d'(w)$ as

$$\lambda_d'(w) = \frac{\Xi_1(\lambda_d, w)}{-\Xi_2(\lambda_d, w)}. \tag{18}$$

Since $\Xi_2(\lambda_d, w) < 0$ for all $w \in [0,1]$ as we show in the proof of Lemma 2, $\lambda_d'(w)$ must possess the same sign as that of $\Xi_1(\lambda_d, w)$. The rest of the proof focuses on the function $\Xi_1(\lambda_d, w)$.

$$\begin{aligned} \Xi_1(\lambda_d, w) &= \textstyle\sum_i \frac{\partial V}{\partial \pi_i^*}\frac{\partial \pi_i^*}{\partial w} + \frac{\partial V}{\partial w} \\ &= \textstyle\sum_i (1 - b(i))(\lambda_n + G(w)\lambda_d)\frac{\partial \pi_i^*}{\partial w} - wf(w)\lambda_d \sum_i \pi_i^* b(i) \\ &= -(\lambda_n + G(w)\lambda_d)\frac{\partial \mathbb{E}_{\boldsymbol{\pi}^*} b(X^*)}{\partial w} - wf(w)\lambda_d \mathbb{E}_{\boldsymbol{\pi}^*} b(X^*), \end{aligned} \tag{19}$$

where the last equality follows from $\sum_i (\lambda_n + G(w)\lambda_d)\frac{\partial \pi_i^*}{\partial w} = 0$ and the fact that $\sum_i \pi_i^* \equiv 1$. We next provide some intuition behind the RHS of the above expression. Increasing w reduces the proportion of PFUs, which affects the effective throughput rate λ_d in two ways. First, if there are more RFUs and fewer PFUs, then the virtual arrival rate λ_v will decrease because RFUs might balk. As a result, the steady-state queue length may be shorter and the balking rate may decrease when w increases. In this scenario, increasing w may increase the effective arrival rate as there are fewer balked visits. This reduces *blocking effect* as captured by the first term in the RHS of (19). In contrast, when more follow-up visits are booked as RFUs rather than PFUs, some RFUs may balk. Therefore, increasing w (i.e., booking fewer PFUs) may have a negative effect on effective throughput by diminishing the *holding effect*, as captured by the second term in the RHS of Equation (19).

The first term in the RHS of (19) contains a partial derivative of the steady-state probability with respect to w. In order to analyze this term, we analyze the stochastic process $X(t)$ and derive an alternative expression for $\partial \mathbb{E}_{\boldsymbol{\pi}^*} b(X^*)/\partial w$ in the following lemma. The proof of Lemma 4, which uses the IPA method, is provided in Appendix EC.6.

Lemma 4. *At all* $w \in [0,1]$,

$$\begin{aligned} \tfrac{\partial \mathbb{E}_{\boldsymbol{\pi}^*} b(X^*)}{\partial w} &= \tfrac{1}{\lambda_d G(w) + \lambda_n} \textstyle\sum_i \frac{\partial \lambda_v(i, w, \lambda_d)}{\partial w} \left[\int_0^1 \pi_i(s) q_i(s, w) ds\right] \\ &= -\tfrac{\lambda_d f(w)}{\lambda_d G(w) + \lambda_n} \textstyle\sum_i (1 - w(1 - b(i)) - \gamma(1 - w)) \left[\int_0^1 \pi_i(s) q_i(s, w) ds\right], \end{aligned} \tag{20}$$

and

$$\begin{aligned}\frac{\partial \mathbb{E}_{\boldsymbol{\pi}^*} b(X^*)}{\partial \lambda_d} &= \frac{1}{\lambda_d G(w)+\lambda_n} \sum_i \frac{\partial \lambda_v(i,w,\lambda_d)}{\partial \lambda_d} \left[\int_0^1 \pi_i(s) q_i(s,w) ds\right] \\ &= \sum_i \frac{G(w)(1-b(i))+1-F(w)-\gamma(1-F(w)-\overline{p}+G(w))}{\lambda_d G(w)+\lambda_n} \left[\int_0^1 \pi_i(s) q_i(s,w) ds\right],\end{aligned} \tag{21}$$

where $q_i(s,w)$ is a continuous function that increases in i. Moreover, $q_i(s,w) \in (0,1)$ for all $s \in [0,1)$, and

$$\frac{\partial q_i(s,w)}{\partial w} \leq \frac{f(w) w \lambda_d}{\lambda_d G(w)+\lambda_n} q_i(s,w). \tag{22}$$

Lemma 4, particularly (20), allows us to reformulate $\Xi_1(\lambda_d, w)$ as

$$\Xi_1(\lambda_d,w) = \sum_i f(w)\lambda_d(1-w(1-b(i))-\gamma(1-w)) \left[\int_0^1 \pi_i(s) q_i(s,w) ds\right] - w f(w) \lambda_d \sum_i \pi_i^* b(i). \tag{23}$$

Since $q_i(s,w)$ is differentiable in w, $\Xi_1(\lambda_d, w)$ is also differentiable in w. Using the above expression for $\Xi_1(\lambda_d, w)$, we can determine the sign of $\Xi_1(\lambda_d(w), w)$ at the two end points, $w=0$ and $w=1$, respectively.

If we plug $w=0$ into the expression (23), the second term vanishes. As $\gamma < 1$, we get

$$\Xi_1(\lambda_d,0) = \sum_i f(0)\lambda_d(1-\gamma) \left[\int_0^1 \pi_i(s) q_i(s,0) ds\right] > 0. \tag{24}$$

By plugging $w=1$ into expression (23), we get

$$\begin{aligned}\Xi_1(\lambda_d,1) &= \sum_i f(1)\lambda_d b(i) \left[\int_0^1 \pi_i(s) q_i(s,1) ds\right] - f(1)\lambda_d \sum_i \pi_i^* b(i) \\ &= f(1)\lambda_d \sum_i b(i) \left[\int_0^1 \pi_i(s)(q_i(s,1)-1) ds\right] < 0.\end{aligned} \tag{25}$$

The above facts imply that $\lambda_d'(w)$ is positive at $w=0$, and negative at $w=1$. Since $\lambda_d'(w) = -\Xi_1(\lambda_d,w)/\Xi_2(\lambda_d,w)$ is continuous, there must be at least one $w^* \in (0,1)$ at which $\lambda_d'(w^*)=0$. Furthermore, since $\Xi_1(\lambda_d,w)$ is differentiable on $w \in [0,1]$, and $\Xi_2(\lambda_d,w)$, with its expression given in (14), is also differentiable in w, equation (18) then implies that the second derivative $\lambda_d''(w)$ exists everywhere on $[0,1]$. We next show that at any $\lambda_d''(w^*) < 0$ at any w^* at which $\lambda_d'(w^*)=0$. This fact implies that w^* is a maximizer on $[0,1]$ and there is no minimizer in the interior $(0,1)$. Furthermore, we deduce that such a minimizer must be unique, because a twice differentiable function cannot have two maximizers without a minimizer in between. This completes the proof of quasi-concavity of $\lambda_d(w)$.

To estimate $\lambda_d''(w^*)$, we first multiply both sides of (17) by $\frac{1}{f(w)}$ and obtain the following identity

$$0 \equiv \frac{1}{f(w)} \Xi_1(\lambda_d,w) + \frac{1}{f(w)} \Xi_2(\lambda_d,w) \lambda_d'(w). \tag{26}$$

At both sides of the above identity, take the derivative of w at w^* and get

$$\begin{aligned} 0 &= \tfrac{\partial}{\partial w}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right)+\tfrac{\partial}{\partial \lambda_d}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right)\lambda_d'(w^*) \\ &\quad+\tfrac{d}{dw}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_2(\lambda_d,w)\right)\lambda_d'(w^*)+\tfrac{1}{f(w^*)}\Xi_2(\lambda_d,w^*)\lambda_d''(w^*) \\ &= \tfrac{\partial}{\partial w}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right)+\tfrac{1}{f(w^*)}\Xi_2(\lambda_d,w^*)\lambda_d''(w^*), \end{aligned} \tag{27}$$

where the second equality follows from the assumption that $\lambda_d'(w^*)=0$. Since $\Xi_2(\lambda_d,w^*)<0$, equation (27) implies that $\lambda_d''(w^*)$ and $\tfrac{\partial}{\partial w}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right)$ share the same sign. Thus, to show $\lambda_d''(w^*)<0$, it suffices to show that $\tfrac{\partial}{\partial w}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right)<0$. To do that, we plug in the expression (23) of Ξ_1 and get

$$\begin{aligned} &\tfrac{\partial}{\partial w}|_{w=w^*}\left(\tfrac{1}{f(w)}\Xi_1(\lambda_d,w)\right) \\ &= \tfrac{\partial}{\partial w}|_{w=w^*}\left(\textstyle\sum_{i\geq 0}\lambda_d(1-w(1-b(i))-\gamma(1-w))\left[\int_0^1\pi_i(s)q_i(s,w)ds\right]-w\lambda_d\sum_i\pi_i^*b(i)\right) \\ &= \underbrace{\sum_i\lambda_d(1-w^*(1-b(i))-\gamma(1-w^*))(\int_0^1\frac{\partial q_i(s,w)}{\partial w}|_{w=w^*}\pi_i(s)ds)}_{C_1} \\ &\quad\underbrace{-\left(\sum_i(1-b(i)-\gamma)\lambda_d(\int_0^1\pi_i(s)q_i(s,w^*)ds)+\sum_i\pi_i^*b(i)\lambda_d\right)}_{C_2} \\ &\quad\underbrace{+\sum_i(\int_0^1\frac{\partial\pi_i(s)}{\partial w}|_{w=w^*}q_i(s,w^*)(1-w^*(1-b(i))-\gamma(1-w^*))\lambda_d)ds)}_{C_3}\underbrace{-\sum_i\frac{\partial\pi_i^*}{\partial w}|_{w=w^*}b(i)\lambda_d w^*}_{C_4}, \end{aligned} \tag{28}$$

where C_1, C_2, C_3, and C_4 are expressions involving w^* and λ_d. Because of Lemma 3 and the fact that $\{q_i(s,w^*)(1-w^*(1-b(i))-\gamma(1-w^*))\}$ is a non-decreasing sequence of i for all $s\in[0,1]$, C_3 is nonnegative. Consequently,

$$\frac{\partial}{\partial w}|_{w=w^*}\left(\frac{1}{f(w)}\Xi_1(\lambda_d,w)\right)\leq C_1-C_2+C_4. \tag{29}$$

To analyze the rest terms at the RHS of (28), we use the fact that $\lambda_d'(w^*)=0$, which implies $\Xi_1(\lambda_d,w^*)=0$. Therefore,

$$\begin{aligned} 0 &= \Xi_1(\lambda_d,w^*) \\ &= \textstyle\sum_i f(w^*)\lambda_d(1-w^*(1-b(i))+\gamma(1-w))\left[\int_0^1\pi_i(s)q_i(s,w^*)ds\right]-w^*f(w^*)\lambda_d\sum_i\pi_i^*b(i) \\ &= \textstyle\sum_i f(w^*)\lambda_d(1+\gamma)\left[\int_0^1\pi_i(s)q_i(s,w^*)ds\right] \\ &\quad -w^*f(w^*)\lambda_d\left(\sum_i(1-b(i)-\gamma)\left[\int_0^1\pi_i(s)q_i(s,w^*)ds\right]+\sum_i\pi_i^*b(i)\right) \\ &= \textstyle\sum_i f(w^*)\lambda_d(1+\gamma)\left[\int_0^1\pi_i(s)q_i(s,w^*)ds\right]-w^*f(w^*)C_2, \end{aligned} \tag{30}$$

which leads to an important equality

$$C_2=\frac{\lambda_d(1+\gamma)}{w^*}\sum_i\left[\int_0^1\pi_i(s)q_i(s,w^*)ds\right]>0. \tag{31}$$

Using the upper bound for $\frac{\partial q_i(s,w)}{\partial w}$ derived in Lemma 4, i.e., Equation (22), we can upper bound C_1 at w^* as

$$\begin{aligned} C_1 &= \textstyle\sum_i (\int_0^1 \frac{\partial q_i(s,w^*)}{\partial w}|_{w=w^*}\pi_i(s)ds)(1-w^*(1-b(i))-\gamma(1-w^*))\lambda_d \\ &\le \tfrac{\lambda_d w^* f(w^*)}{\lambda_n+G(w^*)\lambda_d}\lambda_d\left[\textstyle\int_0^1 \sum_i \pi_i(s)(1-w^*(1-b(i))-\gamma(1-w^*))q_i(s,w^*)ds\right] \\ &< \tfrac{\lambda_d w^* f(w^*)}{\lambda_n+G(w^*)\lambda_d}\lambda_d\left[\textstyle\int_0^1 \sum_i \pi_i(s)q_i(s,w^*)ds\right] \\ &= \tfrac{\lambda_d (w^*)^2 f(w^*)}{(1+\gamma)(\lambda_n+G(w^*)\lambda_d)}C_2, \end{aligned} \tag{32}$$

where the first inequality follows from the upper bound for $\frac{\partial q_i(s,w)}{\partial w}$ derived in Lemma (4), and the last equality follows from equation (31).

We can also obtain an upper bound for C_4 at w^* as

$$\begin{aligned} C_4 &= -\textstyle\sum_i \frac{\partial \pi_i^*}{\partial w}|_{w=w^*} b(i)\lambda_d w^* \\ &= \tfrac{w^*\lambda_d}{G(w^*)\lambda_d+\lambda_n}\textstyle\sum_i f(w^*)\lambda_d(1-w^*(1-b(i))-\gamma(1-w^*))\left[\int_0^1 \pi_i(s)q_i(s,w^*)ds\right] \\ &< \tfrac{\lambda_d (w^*)^2 f(w^*)}{(1+\gamma)(\lambda_n+G(w^*)\lambda_d)}C_2, \end{aligned} \tag{33}$$

where the second equality follows from equation (20) of Lemma 4, and the last equality follows the same logic as that we used in equation (32).

Plugging the upper bound for C_1 and C_4 into inequality (29) leads to

$$\frac{\partial}{\partial w}|_{w=w^*}\left(\frac{1}{f(w)}\Xi_1(\lambda_d,w)\right) < (-1+\frac{2\lambda_d (w^*)^2 f(w^*)}{(1+\gamma)(G(w^*)\lambda_d+\lambda_n)})C_2 \le 0, \tag{34}$$

where the last inequality follows from the fact that C_2 is positive at w^* and Lemma 5. Inequality (34), according to our previous argument, implies that $\lambda_d''(w^*)<0$ and that concludes the proof. ■

Lemma 5. *If the distribution of p satisfies either one of the two conditions specified in Theorem 2, then for all $w\in[0,1]$,*

$$\frac{2\lambda_d w^2 f(w)}{(1+\gamma)(G(w)\lambda_d+\lambda_n)} \le 1. \tag{35}$$

The proof of Lemma 5 is presented in Appendix EC.7.

Finally, we provide an algorithm to compute the stationary distribution $\boldsymbol{\pi}^*(w)$ for any fixed w and prove convergence of this algorithm in EC.8.

7. A Carve-Out Appointment Booking System

7.1. The Setting

Some doctors use a "carve-out" scheduling approach to manage their service capacity (Robinson and Chen 2010). A carve-out system operates similarly to a traditional appointment system except that it reserves a certain portion of slots for walk-in patients. The queueing model developed earlier in this paper can be adapted to fit a carve-out system by allowing two parallel queues that share the same server (doctor), with one queue (Queue A) modeling the backlog of appointment bookings

and the other queue (Queue W) representing walk-in patients who have arrived at the clinic for same-day service; see Figure 4 for a graphical illustration.

Because both queues share the same server (doctor), the care provider has to split the service capacity between the two queues. Let $R \in (0,1)$ denote the portion of service capacity that is allocated to Queue A. For example, if $R = 3/5$, then three out of every five slots are allocated to Queue A and the remaining two slots are reserved for the walk-in queue. For analytical tractability, we do not consider how these service slots are sequenced (e.g., whether 3 consecutive slots for Queue A and 2 consecutive ones for Queue W, or A-W-A-W-A) because different appointment sequences will result in similar patient wait times. Instead, we simply assume that the service slots for each queue are distributed uniformly over time. Consequently, the service times for each appointment in Queue A and Queue W are $1/R$ slots and $1/(1-R)$ slots, respectively.

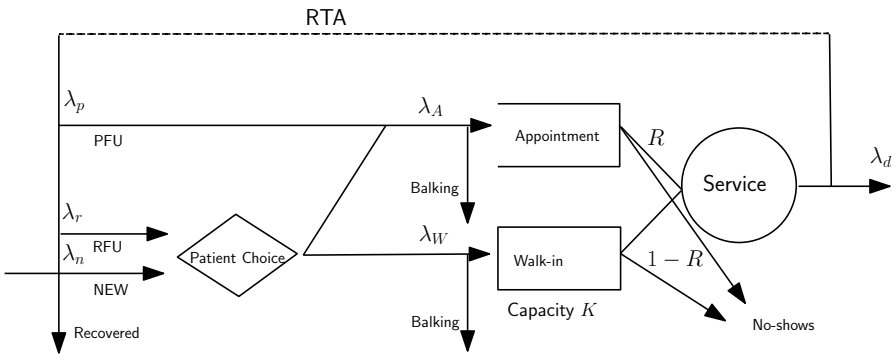


Figure 4 A Hybrid System under RTA

For the arrival process, we assume that among the doctor's NEW and RFU patients, a fixed portion r choose to book an appointment in advance by joining Queue A, and the rest walk in. i.e., join Queue W. Since our focus is the optimal PFU capacity management policy, we do not consider the patients' strategic behavior. Therefore, we assume that r is exogenous and independent of system states as well as control parameters such as w and R. A careful study of the carve-out system considering patient strategic behavior would require extensive empirical evidence, which is beyond the scope of this paper.

We assume that the arrival of new patients follows a time-stationary Poisson process with mean rate λ_n. Furthermore, because of the RTA assumption, the RFUs also arrive according to a time-stationary Poisson process with mean rate λ_r. Then, by Poisson thinning theorem, the arrivals for Queue A and Queue W also follow a time-stationary Poisson process with mean rate $r(\lambda_n + \lambda_r)$ and $(1-r)(\lambda_n + \lambda_r)$, respectively. Nevertheless, not all of these patients will join the queue as some

will balk. We make the same set of assumptions on patient balking behavior as in the previous sections. That is, patients who choose to join Queue A have a state-dependent balking rate $b(i)$, which satisfies Conditions (i) and (ii) given in Section 6; whereas walk-in patients balk only if the number of people in Queue W has reached its upper limit K. As a result, the walk-in queue is modeled as a single-server slotted service queue with finite capacity.

Although the walk-in queue in the carve-out system is modeled in the same way as the open-access queue in Section 5, they represent different entities in reality. The backlogs in the open-access queue represent appointments that have been booked in advance, with the extra constraint that patients by themselves can only book appointments within a short lead time (e.g., at most 24-48 hours in advance), except PFUs who are allowed to book well in advance. Whereas, the backlogs in Queue W in the carve-out system represent walk-in patients physically present at the clinic. Therefore, in a carve-out system, the PFUs can only join Queue A as they represent appointments being booked in advance.

We still consider need-based PFU capacity management policies. That is, the doctor observes the patient's probability p of needing an FUA and decides whether to book a PFU or not. As we have shown earlier, such a policy is characterized by a control threshold w such that PFUs are booked only for patients with $p \geq w$. In this section, we characterize the optimal control threshold w in a carve-out system in certain cases. Our study complements the existing literature on carve-out systems (Dobson et al. 2011, Qu et al. 2007) by studying the management of FUAs.

One important feature relevant to the management of FUA is that no-show and late cancellation causes less damage in a carve-out system, because the care provider can redirect some of those slots to walk-in patients, if there are any walk-in patients in the queue. To underscore this important feature, we assume that $\gamma = 1$ so that all recovered PFUs can be rescued[2]. We do not model in detail the processes by which those slots are rescued and redirected to the walk-in queue because that will introduce complicated interactions between the two queues. Instead, in our aggregate-level model, these slots can be rescued by the system, and then $1 - R$ portion of the total slots are allocated to walk-in queue. This setting accommodates the possibility that some rescued slots will be wasted because the walk-in queue may be empty.

When $\gamma = 1$, there is no cost to book a PFU because if the patient does not need an appointment, then the care provider can safely rescue the slot. It might be plausible to book all potential FUAs as PFUs by setting $w = 0$. In fact, in a traditional appointment system, Equation (23) in the proof of Theorem 2 implies that $w = 0$ is a dominant strategy when $\gamma = 1$. Contrary to this intuition, a rigorous analysis shows that letting $w = 0$ is not optimal in a carve-out system. The optimal value

[2] We do not impose any assumption on the spoilage rate η for regular slots because it has no significant impact on our conclusion, but it would be fair to assume $\eta = 0$ in a carve-out system.

of w depends mostly on the arrival rates of Queue A and Queue W. To enable a rigorous analysis, we first introduce the RTA framework which allows us to derive an approximate characterization for the steady-state of the carve-out system.

7.2. Steady-State Characterization using RTA

Under the RTA assumption, the arrival of RFUs and PFUs both follow independent Poisson processes. Let $X_A(t)$ and $X_W(t)$ denote the number of patients in Queue A and Queue W at the beginning of slot t, respectively, including the one being served. Since both arrival and service processes for the two queues are independent under the RTA assumption, we consider each process as an independent process and characterize its steady state distribution. The next proposition shows that the steady-state distributions[3] of $X_A(\cdot)$ and $X_W(\cdot)$ both exist regardless of r, R, and K, and are unique. The intuition is that since Queue A has state-dependent balking and Queue W has a finite buffer, both queues are stable regardless of the arrival rates.

PROPOSITION 2. *Suppose the balking rate $b(i)$ satisfies conditions (i) and (ii) in Section 6.1. Then for all $w \in [0,1]$, $R \in (0,1)$, $X_A(\cdot)$ and $X_W(\cdot)$ are positively recurrent and irreducible Markov chains with periods $1/R$ and $1/(1-R)$, respectively. Consequently, there exist unique probability distributions $\{\pi^A(t)|t \in [0,1/R)\}$ and $\{\pi^W(t)|t \in [0,1/(1-R))\}$ such that $X^A(s) \stackrel{d}{=} \pi^A(s-\lceil s \rceil)$ and $X^W(s) \stackrel{d}{=} \pi^W(s-\lceil s \rceil)$ for all $s>0$ provided that $X^A(0) \stackrel{d}{=} \pi^A(0)$ and $X^W(0) \stackrel{d}{=} \pi^W(0)$. The steady state distribution $\pi^{A,*} = \int_0^{1/R} \pi^A(t)dt$ and $\pi^{W,*} = \int_0^{1/(1-R)} \pi^W(t)dt$ both exist and are unique.*

The proof of Proposition 2 is similar to that of Proposition 1 and is provided in Appendix EC.9.

We use X_A^* and X_W^* to denote the steady-state queue lengths in the two queues (with some abuse of notation), which have probability distributions $\pi^{A,*}$ and $\pi^{W,*}$, respectively. Then, we can express the virtual arrival rates of the two queues λ_A and λ_W as functions of the total effective throughput rate λ_d and queue lengths. Considering $\gamma = 1$, the expressions are given by

$$\begin{aligned} \lambda_A(w,\lambda_d) &= r(\lambda_n+\lambda_r)\mathbb{E}_{\pi^{A,*}}(1-b(X_A^*)) + \lambda_p \\ &= r(\lambda_n+\lambda_d G(w))\mathbb{E}_{\pi^{A,*}}(1-b(X_A^*)) + (\overline{p}-G(w))\lambda_d, \\ \lambda_W(w,\lambda_d) &= (1-r)(\lambda_n+\lambda_r)(1-\pi_K^{W,*}) = (1-r)(\lambda_n+\lambda_d G(w))(1-\pi_K^{W,*}), \end{aligned} \tag{36}$$

where the expressions of λ_r and λ_p follow Equation (2) in the case of $\gamma = 1$, and $\pi_K^{W,*}$ denotes the steady-state probability that an arrived patient finds a full buffer (i.e., queue length $= K$) and balks.

[3] As before, we define steady-state distribution as the long-run average of the stochastic process by abuse of terminology.

As before, we define a performance vector for the carve-out system $(\lambda_d, \lambda_r, \lambda_p, \lambda_A, \lambda_W)$ and any feasible performance vector must satisfy Equation (36) as well as the following equation, which is analogous to Equation (11),

$$\begin{aligned} & V^r(\lambda_d, \pi^{A,*}(\lambda_d), \pi^{W,*}(\lambda_d)) \\ := & \, r\mathbb{E}_{\boldsymbol{\pi}^{A,*}}(1-b(X_A^*))(\lambda_n + G(w)\lambda_d) + (\overline{p} - G(w))\lambda_d + (1-r)(\lambda_n + \lambda_d G(w))(1-\pi_K^{W,*}) - \tfrac{\lambda_d}{1-\eta} \\ = & \, 0. \end{aligned} \tag{37}$$

Our next result, which is analogous to Proposition 2, shows that there always exists a unique solution to Equation (37) for all w, r, and R.

LEMMA 6. *Given any $w \in [0,1]$, R, $r \in (0,1)$ and $K > 0$, there exists a unique feasible performance vector $(\lambda_d, \lambda_r, \lambda_p, \lambda_A, \lambda_W)$ with $\lambda_d \in (0,1)$.*

Proof of Lemma 6: The proof is similar to that of Lemma 2. To show that Equation (37) has a unique solution $\lambda_d \in (0,1)$, it suffices to prove that

$$V^r(0, \pi^{A,*}(0), \pi^{W,*}(0)) > 0 > V^r(1, \pi^{A,*}(1), \pi^{W,*}(1)), \tag{38}$$

and that $V^r(\lambda_d, \pi^{A,*}(\lambda_d), \pi^{W,*}(\lambda_d))$ is strictly decreasing in λ_d. It is straightforward to show that $V^r(0, \pi^{A,*}(0), \pi^{W,*}(0)) > 0$. To show $V^r(1, \pi^{A,*}(1), \pi^{W,*}(1)) < 0$, because Proposition 2 ensures stability of both queues, we have $r\mathbb{E}_{\boldsymbol{\pi}^{A,*}}(1-b(X_A^*))(\lambda_n + G(w)\lambda_d) + (\overline{p} - G(w))\lambda_d = \lambda_A(w, \lambda_d) < R$, and $(1-r)(\lambda_n + \lambda_d G(w))(1-\pi_K^{W,*}) = \lambda_W(w, \lambda_d) \leq 1 - R$. Finally, $V^r(\lambda_d, \pi^{A,*}(\lambda_d), \pi^{W,*}(\lambda_d))$ is strictly decreasing in λ_d because

$$\begin{aligned} \tfrac{dV^r}{d\lambda_d} &= -r\textstyle\sum_i (\pi_i^{A,*}(\lambda_d))' b(i)(\lambda_n + G(w)\lambda_d) + r\mathbb{E}_{\boldsymbol{\pi}^*}\,(1-b(X^*))G(w) + \overline{p} - G(w) \\ &\quad + (1-r)G(w)(1-\pi_K^{W,*}) - (1-r)(\lambda_n + \lambda_d G(w))(\pi_K^{W,*}(\lambda_d))' - \tfrac{1}{1-\eta} \\ &\leq rG(w) + \overline{p} - G(w) + (1-r)G(w) - \tfrac{1}{1-\eta} \\ &= \overline{p} - \tfrac{1}{1-\eta} < 0, \end{aligned} \tag{39}$$

where the inequality follows from $r\sum_i (\pi_i^{A,*}(\lambda_d))' b(i) G(w)\lambda_d \geq 0$ as a result of Lemma 3 and the fact that $(\pi_K^{W,*}(\lambda_d))' \geq 0$. ■

7.3. Optimal PFU Control Threshold

For a given set of parameters including λ_n, r, R, $b(\cdot)$, and $f(\cdot)$, we let $\lambda_d(w)$ denote the unique solution to Equation (37), which represents the steady-state effective throughput rate. We are interested in solving the throughput maximization problem, $\max_{w\in[0,1]} \lambda_d(w)$, and derive some structural properties of the optimal threshold w^C. However, this is challenging even under the RTA assumption. We can characterize w^C only in two extreme cases, that is, when $R \to 0$ and when $R \to 1$. In the first case, almost all service capacity is allocated to Queue W. As a result, Queue A becomes extremely crowded, in which case we show that $w^C = 1$ if Queue W is not too

crowded. In the second case when $R \to 1$, Queue W becomes extremely crowded, in which case we prove $w^C = 0$, so the care provider should use the PFU strategy whenever possible. Although the analytical results are limited to the two special cases, they show that the choice of the optimal threshold varies by system parameters, particularly the demand-supply ratio in each queue. We formally state the result in the following theorem and attach its proof in Appendix EC.10.

THEOREM 3. *In a carve-out system with $r \in (0,1)$, we have*

$$\begin{array}{l} \lambda_d(w) \text{ increases in } w \text{ and } w^C \to 1 \quad \text{when } R \to 0 \text{ and } \frac{(1-r)\lambda_n}{1-\overline{p}} \le 1, \\ \lambda_d(w) \text{ decreases in } w \text{ and } w^C \to 0 \quad \text{when } R \to 1. \end{array} \tag{40}$$

Despite the complex algebra involved in the proof of Theorem 3, the underlying intuition is not difficult to explain. When Queue A is crowded but Queue W is less crowded, the holding effect of booking a PFU diminishes because even if a PFU is not booked, $1-r$ portion of the RFU will join Queue W without balking because Queue W is less congested. Meanwhile, even though $\gamma = 1$, the blocking effect remains in effect because by booking PFUs, the care provider directs more patients to Queue A and exacerbates the congestion there. On the other hand, if Queue W is crowded, the holding effect becomes dominant. Then, if the care provider does not book a PFU and that patient joins the walk-in queue, that patient will likely see a full buffer and balk. Thus, PFU appointments not only help secure slots for FUAs (without the cost of increasing the number of no-shows as much as in the traditional system), they also play a role in helping to balance the supply and demand in each queue.

In certain scenarios, the care provider is able to select a capacity split R to balance the workload in the two queues. In this case, our RTA framework provides a way to numerically search for a pair (R, w) that jointly maximizes the effective throughput rate. We conjecture that the optimal policy will book a portion of FUAs as PFUs, similar to the traditional system, though we were not successful in proving quasi-concavity in this case. We explore the selection of the optimal pair in our simulation study in Section 8.

8. Numerical Study

We present a numerical study that serves several purposes: (1) Validate the RTA approximation by comparing the fixed point $\lambda_d(w)$ calculated from the RTA algorithm with the simulation result. (2) Check robustness of our claims upon relaxing the assumption "PFUs have strict head-of-line priority". (3) Check robustness of our claims upon relaxing the assumption "doctors know exactly each patient's revisit probability p". (4) Compute $\lambda_d(w)$ and check robustness when the distribution of p does not satisfy the assumptions in Theorem 2. (5) Investigate how $\lambda_d(w)$ varies in w and R in the carve-out system and to identify an optimal (R, w) pair. We focus on the traditional appointment booking system and the carve-out system because in the open-access system, the optimal policy is simple and has been fully characterized in Theorem 1.

8.1. Test of the RTA Assumption

We develop a simulation model for an appointment booking system and compute its effective throughput rate $\lambda_d^{Sim}(w)$ for different values of w. The simulation model is coded within the *Arena* platform Version 16.10.00001 (Kelton 2002). Instead of simulating the original appointment system, we simulate an FCFS queue with slotted service time. The difference between the two systems is that patients in the former system do not always take the first available slot. Despite this difference, Green and Savin (2008) have shown that the latter provides a close approximation to the former system when a majority of the patients prefer an earlier slot, which is consistent with our data; see Section 3. We also compute $\lambda_d^{RTA}(w)$, the fixed point to Equation (9), using the RTA algorithm presented in Appendix EC.8. The computation of $\lambda_d^{RTA}(w)$ is performed using Matlab (MATLAB 2020).

In the base case of the slotted queuing model, we set the length of each appointment equal to 30 minutes. New appointment booking requests arrive according to a Poisson process and the mean inter-arrival time is 50 minutes, corresponding to $\lambda_n = 30/50 = 0.6$ per slot. Other base case parameters are $w = 0.6$, $f(\cdot) = \text{Beta}(0.5, 0.5)$, and $b(i) = 1 - e^{0.1i}$. In subsequent experiments, one or more of these parameters is varied and we compute $\lambda_d^{RTA}(w)$ and $\lambda_d^{Sim}(w)$. A complete list of input parameters can be found in Table 3, Columns 1–4. If $p \geq w$, then the patient is routed to the PFU orbit and stays there for an observation period L_{o1}, which is simulated according to the empirical distribution of the PFU lead times as plotted in Figure 1, assuming 1 day = 6 hours = 12 slots. After the observation period, with probability p that PFU needs another visit and returns to the queue as prioritized patient, and with probability $(1-p)$ the PFU patient recovers. If the patient recovers, then with probability $(1-\gamma)$ that a patient is a no-show and the slot is wasted and with probability γ the slot is rescued (equivalent in our model to not booking that slot). The latter can happen, for example, when the patient cancels early realizing that he or she is on the path to recovery. If $p < w$, then the patient is routed to the RFU orbit and stays there for an observation period L_{o2}, which is also simulated according to the empirical distribution of inter-appointment times for RFUs. After the observation period, with probability p the RFU will return the queue, and with probability $(1-p)$ she recovers and leaves the system permanently. Each NEW or RFU patient balks with probability $b(i)$ upon arrival. Each booked slot has a probability η to be a spoilage.

In each parametric setting, we run the simulation for 50 replications to obtain a robust evaluation. The choice of 50 replications is based on the observation that the standard error (half-width of the 95% confidence interval) is within 1% of the simulated value. This is a recommended method for determining the number of replications of simulation runs; see for example Law (2007). In each replication, we run the simulation for 2000 slots, and record the average effective throughput rate

in the last 1000 slots. The first 1000 slots are used as a warm-up period to allow the stochastic process to reach a steady state.

Table 3 Numerical Results

Parameters*				Effective Throughput Rate			
w	λ_n	$f(\cdot)$	$b(i)$	$\lambda_d^{RTA}(w)$	$\lambda_d^{SIM}(w)$	$\lambda_d^{NoPr}(w)$	$\lambda_d^{NoAc}(w)$
0.6	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5892	0.5918 (0.0045)**	0.5888 (0.0063)	0.5838 (0.0054)
0	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5084	0.5129 (0.0051)	0.5132 (0.0045)	0.5146 (0.0040)
0.1	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5486	0.5503 (0.0043)	0.545 (0.0053)	0.5541 (0.0045)
0.2	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5634	0.565 (0.0047)	0.566 (0.0055)	0.5658 (0.0060)
0.3	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5734	0.5766 (0.0057)	0.5772 (0.0061)	0.5746 (0.0053)
0.4	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5805	0.5823 (0.0054)	0.5857 (0.0057)	0.5779 (0.0070)
0.5	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5857	0.5844 (0.0063)	0.5913 (0.0064)	0.5877 (0.0050)
0.7	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5913	0.5949 (0.0062)	0.5925 (0.0061)	0.5896 (0.0052)
0.8	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5920	0.5979 (0.0074)	0.5965 (0.0071)	0.5893 (0.0064)
0.9	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5910	0.5916 (0.0067)	0.5939 (0.0066)	0.5888 (0.0063)
1	0.6	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5843	0.5858 (0.0078)	0.5901 (0.0068)	0.5887 (0.0059)
0.6	1	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.6972	0.6952 (0.0044)	0.6956 (0.0049)	0.6850 (0.0037)
0.6	0.95	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.6931	0.6937(0.0044)	0.6926 (0.0040)	0.6845 (0.0042)
0.6	0.9	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.6874	0.686 (0.0044)	0.6889 (0.0047)	0.6765 (0.0035)
0.6	0.8	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.6692	0.6721 (0.0044)	0.6705 (0.0054)	0.6646 (0.0050)
0.6	0.75	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.6555	0.6583(0.0053)	0.6578 (0.0051)	0.6521 (0.0053)
0.6	0.5	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.5210	0.5227 (0.0069)	0.5243 (0.0066)	0.5143 (0.0075)
0.6	0.4	Beta(0.5,0.5)	$1-e^{-0.1i}$	0.4353	0.4384 (0.0061)	0.4357 (0.0066)	0.4355 (0.0071)
0.6	0.6	Beta(5,1)	$1-e^{-0.1i}$	0.6627	0.6648(0.0046)	0.6656 (0.0043)	0.6705 (0.0045)
0.6	0.6	Beta(1,3)	$1-e^{-0.1i}$	0.4951	0.4918(0.0051)	0.4923 (0.0048)	0.4941 (0.0064)
0.6	0.6	Beta(2,2)	$1-e^{-0.1i}$	0.5793	0.5800(0.0064)	0.5823 (0.0063)	0.5739 (0.0064)
0.6	0.6	Unif(0,1)	$1-e^{-0.1i}$	0.5835	0.5879 (0.0066)	0.5901 (0.0057)	0.5788 (0.0058)
0.3	0.6	Unif(0,0.5)	$1-e^{-0.1i}$	0.4806	0.4862 (0.0057)	0.4857 (0.0055)	0.4907 (0.0058)
0.6	0.6	Beta(0.5,0.5)	$1-e^{-i}$	0.464	0.4757 (0.0050)	0.4760 (0.0045)	0.4674 (0.0057)
0.6	0.6	Beta(0.5,0.5)	$\min\{1, 0.1i\}$	0.5823	0.5856 (0.0066)	0.5849 (0.0064)	0.5803 (0.0052)
0.6	0.6	Beta(0.5,0.5)	$\min\{1, 0.2i\}$	0.5416	0.5465 (0.0056)	0.5409 (0.0071)	0.5450 (0.0055)

*Other Parameters: $\epsilon = 10^{-5}$ (tolerance of Matlab code), $\eta = 0.26$, $\gamma = 0$, L_{o1} and L_{o2} follow empirical distribution.
**The number in () denotes the standard error of the simulation results in 50 replications.

The outputs of the RTA algorithm and the simulation experiments are summarized in Table 3 (Column 5 and 6). For most instances, the absolute error $|\lambda_d^{RTA}(w) - \lambda_d^{Sim}(w)|$ is within 0.006, and the relative error $|\lambda_d^{RTA}(w) - \lambda_d^{Sim}(w)|/\lambda_d^{Sim}(w)$ is within 1.2%, except for the instance in which the balking rate is determined by the expression $(1 - \exp(-i))$. In that case, the error is 2.5%, which is still reasonable. Overall, the results support the accuracy of our analytical framework based on the RTA assumption.

8.2. Test of Robustness of PFUs' Head-of-Line Priority

We assume that the PFU slots have head-of-line priory over NEW and RFUs. The empirical evidence supporting this assumption has been provided in Section 3, showing that the request for PFU bookings usually arrive earlier than those of the NEW and RFU appointments for the same slot. To further check if a violation of this assumption would have a significant impact on system throughput, we performed a simulation experiment in *Arena*. In the simulation, which closely matches the real system, when the PFU patient returns to the queue, then that and other patients already in the queue will be served in an FCFS fashion. Because the PFU's slot is typically requested early, the PFU patient will most likely be placed at the head-of-queue just as our theoretical model assumes. However, with a small probability, there can be NEW or RFU patients in the queue who have requested their appointment even before the PFU patient, and then the PFU patient will not have the head-of-line priority.

We ran the above simulation model with 50 replications for each parameter setting and report the mean effective throughput rate $\lambda_d^{NoPr}(w)$ in Column 7 of Table 3. We compare the reported values to the values of $\lambda_d^{SIM}(w)$ in Column 6, the output of the simulation model which assumes head-of-line priority for PFUs. The comparison shows that the difference between the $\lambda_d^{NoPr}(w)$ and $\lambda_d^{SIM}(w)$ is always less then 0.01 or 2% across all instances. Therefore, the assumption that PFU patients have head-of-line priority over RFU and NEW patients has a relatively minor impact on the system performance.

8.3. Test of Robustness of the Assumption that Revisit Probability is Observable

Our model assumes that a doctor knows each patient's precise probability of needing an FUA later. We tested the robustness of this assumption by simulating a system in which a doctor's estimate is imprecise. We let p denote a patient's true probability of needing an FUA and let $\overline{p}$ denote an estimation of p that will be used in the simulation. We assume that $\ln(\frac{\overline{p}}{1-\overline{p}})$ follows a normal distribution with mean $\ln(\frac{p}{1-p})$ and standard deviation 0.537, where the latter is the largest standard error observed upon fitting a logistic regression model to our data; see Section 3. Then, if we were to use $\overline{p}$ in Inequality (1), it would be an equality.

From the resulting equation and the monotonicity of the log-odds function, we deduce that $|\overline{p}-p|$ stochatically dominates $|\hat{p}-p|$, where $\hat{p}$ is the revisit probability predicted by the logistic regression for our data. Therefore, applying a logistic regression to our data will lead to a more accurate prediction than $\overline{p}$. In reality, the doctor observes much more information than that recorded in the data, can make a more accurate prediction than $\hat{p}$, and thus $\overline{p}$.

Therefore, is suffices to show $\lambda_d^{NoAc}(w)$, the effective throughput rate in a system that uses $\overline{p}$, is close to $\lambda_d^{SIM}(w)$, the effective throughput rate in the original system. If that were true, then the

effective throughput in the real system in which a doctor had a more accurate prediction would be even closer to $\lambda_d^{SIM}(w)$. To estimate $\lambda_d^{NoAc}(w)$, we run 50 replications of the simulation model for each of the parametric setting in Table 3, and report the mean effective throughput rate $\lambda_d^{NoAc}(w)$ in the last column of Table 3. The comparison shows that the difference between $\lambda_d^{NoAc}(w)$ and $\lambda_d^{SIM}(w)$ is small and the relative error is less than 2% across all instances.

8.4. $\lambda_d(w)$ in a Traditional Appointment System for Large $\overline{p}$

In Theorem 2, we proved that $\lambda_d(w)$ is a quasi-concave function when the revisit probability p follows a Beta(α,β) distribution with mean $\overline{p}=\alpha/(\alpha+\beta)\leq 1/4$ and $\beta>1$, or uniformly distributed over $[a,b]$ with $2b^2\leq b-a$. Next, we plot $\lambda_d(w)$ for parameters that violate the assumption of Theorem 2 with beta distributed p to check whether the conclusion still holds. We use the RTA algorithm to compute and plot $\lambda_d(w)$ under different parameters of α, β, c, where c is a coefficient in the balking rate function $b(i)=1-\exp(-ci)$. The plots are presented in Figure 5.

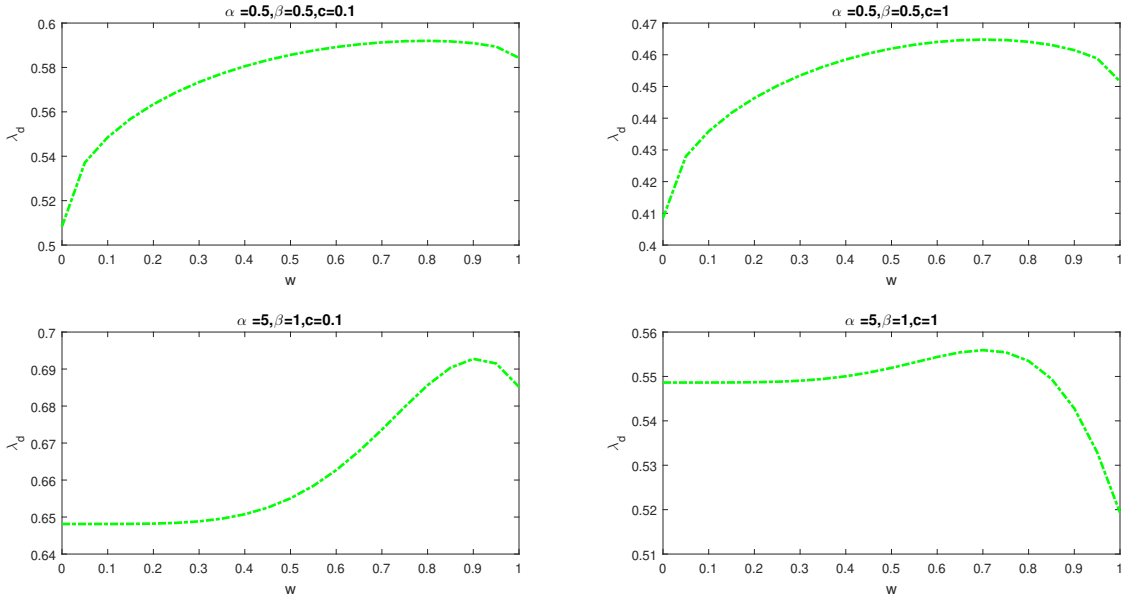


Figure 5 $\lambda_d(w)$ as a function of w

Note that the two figures in the first row correspond to $\overline{p}=\alpha/(\alpha+\beta)=0.5>0.25$. So these plots do not satisfy the assumption of Theorem 2, but $\lambda_d(w)$ exhibits quasi-concave property. We then increase $\overline{p}$ to $\alpha/(\alpha+\beta)=5/6$ and plot $\lambda_d(w)$ in the second row of Figure 5. Over the interval $w\in[0,0.4]$, $\lambda_d(w)$ is nearly flat, so quasi-concavity is not evident. Across all our experiments, we did not find a counter-example in which quasi-concavity was violated.

8.5. $\lambda_d(w)$ in a Carve-Out System

In Theorem 3, we show that in some special cases the total effective throughput rate in a carve-out system is monotonically increasing or decreasing in w. As a result, the optimal w^C is attained at 0 (no PFU) or at 1 (no RFU). However, these special cases require $R \to 0$ or $R \to 1$, which are not typical scenarios in practice. In this section, we run simulation experiments to compute the function $\lambda_d(w)$ under different values of R. These experiments serve three purposes. First, they show that the results of Theorem 3, i.e., the monotonicity of $\lambda_d(w)$, are robust when R is reasonably close to 0 or 1. Second, the simulation results characterize the optimal w^C for intermediate values of R, which is not covered by our analytical results. Third, the experiments provide intuition regarding how clinic managers should select R when R is a decision variable.

We choose the same parameters as in Table 3, except that we set $\gamma = 1$ and $\eta = 0$ to be consistent with the carve-out setting. We chose $r = 0.5$ and tested five different values of R: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each R, we computed $\lambda_d(w)$ for different values of w and plotted their relationship in Figure 6.

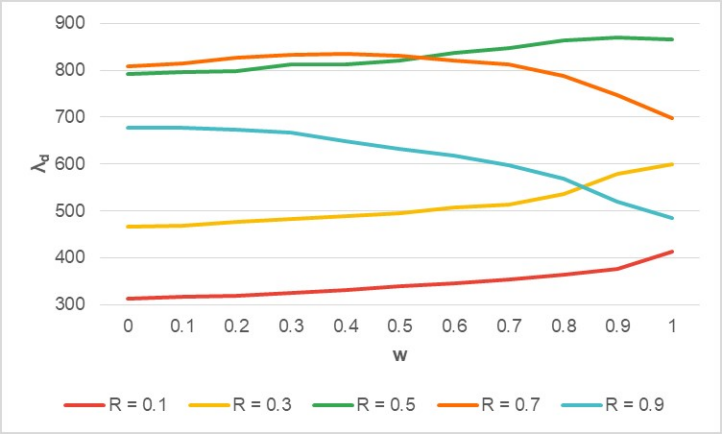


Figure 6 Total Effective Throughput Rate in a Carve-out System $\lambda_d(w)$

Note that for $R = 0.1, 0.3$, $\lambda_d(w)$ is increasing in w and the optimal threshold for booking PFUs is $w^C = 0$; whereas for $R = 0.9$, $\lambda_d(w)$ is decreasing and its maximum value is attained at $w^C = 1$, i.e., by not booking any PFUs. These observations show that the results of Theorem 3 are valid even for somewhat larger (or smaller) values of R. Second, Figure 6 shows that for $R = 0.5, 0.7$, $\lambda_d(w)$ is non-monotone and quasi-concave. This is what one would expect although it is difficult to prove this observation via formal arguments. Finally, the figure shows that for the problem parameters chosen for these experiments, the overall highest $\lambda_d(w)$ is attained when R is 0.5 (approximate load balance holds) and w is high but less than 1.

9. Concluding Remarks

Appointment systems are ubiquitous, especially in health care. Inspired by the empirically observed practice in outpatient clinics of giving priority to some follow-up appointments, we analyze three representative appointment systems. The objective of our investigation is both to establish the structure of a need-based priority rule and to develop a method to compute the optimal policy parameters when the health system wants to maximize throughput rate. We show that the optimal policy is to not have PFUs in an open access system, and that the optimal policy is of threshold-type with a single parameter in the other two systems. This parameter is such that if the patient's need exceeds a critical probability threshold, then he or she will be designated a priority follow-up, otherwise not.

On the methodological front, the paper presents an analysis of slotted-service queues with orbits under the RTA assumption. In this way, it adds to both the literature on queueing systems with feedback, and to the literature on appointment scheduling with follow-up visits. On the practitioner front, the contribution of this paper is that it provides implementable operating guidelines for appointment systems that aim to maximize throughout, and a tractable method for calculating the optimal probability threshold. We show that the optimality of this threshold is robust against estimation errors associated with the revisit probability. Therefore, even if the doctor cannot know exactly each patient's revisit probability, the suggested threshold policy is still implementable and provides near optimal performance.

In practice, the appointment booking system in place may not allow doctors to book follow-up appointments for the patients. An alternative in those cases would be for doctors to tell some patients that FUAs likely would be needed based on their diagnoses and typical courses of treatment, and to encourage those patients to book FUAs immediately after their earlier appointments. This would be tantamount to prioritizing some FUAs.

The paper focuses on throughput rate maximization, which is a realistic objective for many revenue-oriented care providers. Other prevalent objectives may incorporate concern for quality of care and continuity of care. We hope that these objectives would be incorporated into the RTA framework and that their analyses would provide promising topics for future work.

References

Aleksandrov, A. Queuing system with repeated orders. *Engineering Cybernetics*, 12(3):1–4, 1974.

Ancker Jr, C. and Gafarian, A. Some queuing problems with balking and reneging. i. *Operations Research*, 11(1):88–100, 1963a.

Ancker Jr, C. and Gafarian, A. Some queuing problems with balking and reneging-ii. *Operations Research*, 11(6):928–937, 1963b.

Armony, M. and Maglaras, C. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

Armony, M., Plambeck, E., and Seshadri, S. Sensitivity of optimal capacity to customer impatience in an unobservable M/M/s queue (why you shouldn't shout at the DMV). *Manufacturing & Service Operations Management*, 11(1):19–32, 2009.

Artalejo, J. R. A queueing system with returning customers and waiting line. *Operations Research Letters*, 17(4):191–199, 1995.

Campello, F., Ingolfsson, A., and Shumsky, R. A. Queueing models of case managers. *Management Science*, 63(3):882–900, 2017.

Cayirli, T. and Veral, E. Outpatient scheduling in health care: a review of literature. *Production and operations management*, 12(4):519–549, 2003.

Chan, C. W., Yom-Tov, G., and Escobar, G. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482, 2014.

Cohen, J. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18(2):49–100, 1957.

Dixon, A., Robertson, R., Appleby, J., Burge, P., and Devlin, N. J. *Patient choice: how patients choose and how providers respond*. 2010.

Dobson, G., Hasija, S., and Pinker, E. J. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

Dobson, G., Tezcan, T., and Tilson, V. Optimal workflow decisions for investigators in systems with interruptions. *Management Science*, 59(5):1125–1141, 2013.

Erdogan, S. A. and Denton, B. T. Surgery planning and scheduling. *Wiley Encyclopedia of operations research and management science*, 2010.

Gallucci, G., Swartz, W., and Hackerman, F. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. 2014.

Green, L. V. and Savin, S. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

Greenberg, B. S. M/G/1 queueing systems with returning customers. *Journal of Applied Probability*, 26: 152–163, 1989.

Greenberg, B. S. and Wolff, R. W. An upper bound on the performance of queues with returning customers. *Journal of Applied Probability*, 24:466–475, 1987.

Gupta, D. and Denton, B. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

Gupta, D. and Wang, L. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56(3):576–592, 2008.

Hennen, B. Continuity of care in family practice. part 1: dimensions of continuity. *The Journal of family practice*, 2(5):371–372, 1975.

Ho, Y.-C., Cao, X., and Cassandras, C. Infinitesimal and finite perturbation analysis for queueing networks. *Automatica*, 19(4):439–445, 1983.

Huang, J., Carmeli, B., and Mandelbaum, A. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

Keilson, J., Cozzolino, J., and Young, H. A service system with unfilled requests repeated. *Operations Research*, 16(6):1126–1137, 1968.

Kelton, W. D. *Simulation with ARENA*. McGraw-hill, 2002.

Kostami, V. and Ward, A. R. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4):644–656, 2009.

LaGanga, L. R. and Lawrence, S. R. An appointment overbooking model to improve client access and provider productivity. *Proceedings of the New Challenges in Service Operations, POMS College of Service Operations and EurOMA*, 2007.

Larson, R. C. Or forumâperspectives on queues: Social justice and the psychology of queueing. *Operations research*, 35(6):895–905, 1987.

Law, A. M. *Simulation modeling and analysis*. McGraw-Hill, 4th Edition, 2007.

Liu, N., Finkelstein, S. R., Kruk, M. E., and Rosenthal, D. Understanding patient preferences and choice behavior in appointment scheduling. *working paper*, 2015.

Liu, Y. New insights into epithelial-mesenchymal transition in kidney fibrosis. *Journal of the American Society of Nephrology*, 21(2):212–222, 2010.

MATLAB. *version 7.10.0 (R2020b)*. The MathWorks Inc., `www.mathworks.com`, Natick, Massachusetts, 2020.

Murray, M. and Tantau, C. Same-Day appointments: Exploding the Access Paradigm. *Family Practice Management*, 7(8):45–50, 2000.

Patrick, J., Puterman, M. L., and Queyranne, M. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525, 2008.

Qu, X., Rardin, R. L., Williams, J. A. S., and Willis, D. R. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2): 812–826, 2007.

Robinson, L. W. and Chen, R. R. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, 12(2):330–346, 2010.

Rogers, J. and Curtis, P. The concept and measurement of continuity in primary care. *American journal of public health*, 70(2):122–127, 1980.

Ross, S. M. *Introduction to probability models, 10th Edition*. Academic press, 2014.

Saure, A., Patrick, J., Tyldesley, S., and Puterman, M. L. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 223(2):573–584, 2012.

Schuetz, H.-J. and Kolisch, R. Capacity allocation for demand of different customer-product-combinations with cancellations, no-shows, and overbooking when there is a sequential delivery of service. *Annals of operations research*, 206(1):401–423, 2013.

Shin, D. W., Cho, J., Yang, H. K., Park, J. H., Lee, H., Kim, H., Oh, J., Hwang, S., Cho, B., and Guallar, E. Impact of continuity of care on mortality and health care costs: a nationwide cohort study in korea. *The Annals of Family Medicine*, 12(6):534–541, 2014.

Wang, S., Liu, N., and Wan, G. Managing appointment-based services in the presence of walk-in customers. *Management Science*, 66(2):667–686, 2020.

Wardi, Y., Adams, R., and Melamed, B. A unified approach to infinitesimal perturbation analysis in stochastic flow models: the single-stage case. *IEEE Transactions on Automatic Control*, 55(1):89–103, 2009.

Wolff, R. W. Stochastic modelling and the theory of queues. *Englewood Cliffs, NJ*, 96, 1989.

Yang, T. and Templeton, J. G. C. A survey on retrial queues. *Queueing systems*, 2(3):201–233, 1987.

Yom-Tov, G. B. and Mandelbaum, A. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

Zhou, S., Ding, Y., Huh, W. T., and Wan, G. Constant job-allowance policies for appointment scheduling: Performance bounds and numerical analysis. *Production and Operations Management*, 2021.

EC.1. Proof of Lemma 1

Proof. Given $w \in [0,1]$, $F(\cdot, w)$ maps a given $\lambda_d \in [0,1]$ to an effective throughput rate. Since $F(\cdot, w)$ is continuous and maps the closed convex set $[0,1]$ to itself, the Brouwer fixed-point theorem guarantees the existence of a fixed point λ_d in $[0,1]$ such that $F(\lambda_d, w) = \lambda_d$. Moreover, since the image of $F(\cdot, w)$ must be strictly smaller than 1, the fixed point must lie in $[0,1)$.

To show that the fixed point is unique, it suffices to show that $0 \le \partial F(\lambda_d, w)/\partial \lambda_d < 1$ so that $F(\cdot, w)$ is a contracting map. If the arrival rate of a queue with finite buffer size K has increased by ϵ, we know that the increment to the average departure rate of that queue will be between 0 and ϵ, because some would have balked. Therefore, a sample path argument leads to the inequality $0 \le \rho'_K(\cdot) \le 1$.

By expressing λ_v and λ_p as functions of λ_d as in (2) and (3), Equation (5) implies that

$$\begin{aligned} \tfrac{\partial F(\lambda_d, w)}{\partial \lambda_d} &= \rho'_K(\lambda_v)((1-\gamma)(1+G(w)-F(w))+\gamma\overline{p})(1-\eta) \\ &\quad -\rho'_K(\lambda_p)((1-\gamma)(1-F(w))+\gamma(\overline{p}-G(w)))\tfrac{\int_w^1 (1-p)(1-\gamma)f(p)dp}{\int_w^1 (p+(1-p)(1-\gamma))f(p)dp}(1-\eta) \\ &\le \rho'_K(\lambda_v)((1-\gamma)(1+G(w)-F(w))+\gamma\overline{p})(1-\eta) \\ &< 1, \end{aligned} \tag{EC.1}$$

where the second inequality follows from $G(w) - F(w) \le 0$ and $\rho'_K(\cdot) \le 1$. It is straightforward to see that $F(\lambda_d, w)$ is non-deceasing in λ_d, so we have $0 \le \partial F(\lambda_d, w)/\partial \lambda_d < 1$. ■

EC.2. Proof of Theorem 1

Proof. We consider three systems: (1) a system with a control threshold $w \in (0,1)$ and $\gamma \in (0,1)$; (2) a system with a control threshold $w \in (0,1)$ and $\gamma = 1$ (all recovered PFUs can be rescued); (3) a system with $w = 1$ (no PFUs so the parameter γ is moot). Let $\lambda_d^\gamma(w)$, $\lambda_d^1(w)$, and $\lambda_d(1)$ denote the mean-preserving effective throughput rates corresponding to the three systems, respectively. We will prove that

$$\lambda_d^\gamma(w) < \lambda_d^1(w) = \lambda_d(1). \tag{EC.2}$$

That is, the effective throughput rate is always larger when the recovered PFUs can be fully rescued, while the latter has the same mean-preserving service rate as a system with no PFUs.

To prove $\lambda_d^\gamma(w) < \lambda_d^1(w)$, we couple the arrival process of the effective appointments (not including the recovered PFUs) in system (1) and (2). Let $Z^1(t)$ ($X^1(t)$) and $Z^2(t)$ ($X^2(t)$) denote the number of effective backlogged slots (virtual slots) in the coupled systems. We next prove $Z^1(t) \le Z^2(t)$ at all t by contradiction. Let $\tau := \inf\{t \ge 0 \mid Z^1(t) > Z^2(t)\}$. If $\tau < \infty$, then because we have coupled the arrival processes, either of the following must happen at τ^- (an infinitesimal time period right before τ): (a) A customer balks in system (2) but not in system (1) at τ^-; (b) a customer was

bumped in system (2) by a PFU but not in system (1) at τ^-. We next show that neither (a) nor (b) could happen. For (a), if a customer balks in system (2) at τ^-, then $X^2(\tau^-) = K$. Since system (2) has fewer virtual slots as more PFUs have been rescued, we always have $X^1(t) \geq X^2(t)$ for all t, so $X^1(\tau^-) \geq K$. Thus, the customer must also balk in system (1). Therefore, (a) is impossible. For (b), by the definition of τ, we have $Z^1(\tau^-) = Z^2(\tau^-)$. Then if a PFU arrives at τ, it has an equal chance to bump out an effective slot (non-effective slot must be booked also as PFUs so cannot be bumped out). Thus, by coupling the two systems, we can ensure that an effective slot is bumped out in (2) only if it has also be bumped in (1). By ruling out the two cases (a) and (b), we conclude that such a τ does not exist and $Z^1(t) \leq Z^2(t)$ for all t. This implies that

$$\lambda_d^\gamma(w) = \Pr(Z^1(t) > 0) < \Pr(Z^2(t) > 0) = \lambda_d^1(w). \tag{EC.3}$$

To show $\lambda_d^1(w) = \lambda_d(1)$, we note that in both systems all appointments in the queue have the same no-show rate. The only difference is that system (3) is FCFS, and customers balk upon seeing a full buffer, whereas in system (2), the PFUs can bump out a non-PFU slot when it arrives and sees a full buffer. Whichever the case, the system loses one regular slot. If we consider all customers as a homogeneous class, the two systems are equivalent and thus have the same effective throughput rate. ■

EC.3. Incorporating Same-Doctor Matching into the Objective Function

We consider the case when the care provider tries to maximize same-doctor matching as well as effective throughput rate. We analyze the case of an open-access system and a traditional appointment system and reach similar conclusions.

EC.3.1. Open Access Systems

To improve same-doctor matching, we try to avoid cases when an FUA has to balk and look for service at other places. This leads to a multi-objective problem of maximizing throughput and simultaneously minimizing the number of FUAs that balk, which can be scalarized by utilizing a cost $c > 0$ associated with each FUA balking. Mathematically, it leads to the following objective function,

$$\begin{aligned} &\max_{w\in[0,1]} \lambda_d(w) - c(\text{RFU balking} + \text{PFU balking}) \\ =&\max_{w\in[0,1]} \lambda_d(w) - c(\pi_K^X(\lambda_r + \lambda_p \tfrac{\lambda_r}{\lambda_n+\lambda_r}) + \pi_K^Y \lambda_p (\tfrac{\int_w^1 pf(p)dp}{\int_w^1 (p+(1-p)(1-\gamma))f(p)dp} - \tfrac{\lambda_r}{\lambda_n+\lambda_r})). \end{aligned} \tag{EC.4}$$

In the above formulation, $\pi_K^X := \Pr(X(\infty) = K)$ and $\pi_K^Y := \Pr(Y(\infty) = K)$ denote the steady-state probability that the buffer is fully occupied by appointments of all types, and by PFU appointments, respectively. Thus, the term $\pi_K^X(\lambda_r + \lambda_p \frac{\lambda_r}{\lambda_n+\lambda_r})$ computes the expected number of FUA balkings

when $X(\infty)=K$. If a PFU arrives and finds a full buffer and if $Y(\infty)<K$, then she can bump out the last non-PFU patient in the queue. Since the arrival time is independent of patient type, the probability that the last non-PFU patient is an RFU (versus a NEW visit) equals $\lambda_r/(\lambda_n+\lambda_r)$. However, if $Y(\infty)=K$, then the PFU cannot find a non-PFU in the queue and has to balk. The probability that PFU will be an effective slot is $\frac{\int_w^1 pf(p)dp}{\int_w^1(p+(1-p)(1-\gamma))f(p)dp}$. Thus, the last term corrects the balking rate of FUAs when $Y(\infty)=K$.

The following proposition shows that the care provider always books a few PFUs when same-doctor matching has been incorporated into the objective function. The main idea of the proof is to show that the derivative of the multi-objective function (EC.4) is always negative at $w=1$ regardless of the value of c. Unfortunately, we cannot easily determine the sign of the derivative when $w<1$ nor the monotonicity of the objective function with respect to w. Therefore, we do not know whether the optimal threshold w^O could be further decreased if a larger weight c was chosen.

PROPOSITION EC.1. *If $c>0$ and $f(1)>0$, then the optimal solution to* (EC.4)*, denoted by w^O, is strictly less than* 1.

Proof. The derivative of (EC.4) is given by

$$\begin{aligned}&\lambda_d'(w)-c\tfrac{d\pi_K^X}{dw}(\lambda_r+\lambda_p\tfrac{\lambda_r}{\lambda_n+\lambda_r})+c\pi_K^X\tfrac{d}{dw}(\lambda_r+\lambda_p\tfrac{\lambda_r}{\lambda_n+\lambda_r})\\+\;&c\tfrac{d\pi_K^Y}{dw}(\tfrac{\lambda_p\int_w^1 pf(p)dp}{\int_w^1(p+(1-p)(1-\gamma))f(p)dp}-\tfrac{\lambda_p\lambda_r}{\lambda_n+\lambda_r})+c\pi_K^Y\tfrac{d}{dw}(\tfrac{\lambda_p\int_w^1 pf(p)dp}{\int_w^1(p+(1-p)(1-\gamma))f(p)dp}-\tfrac{\lambda_p\lambda_r}{\lambda_n+\lambda_r}).\end{aligned}\tag{EC.5}$$

We now evaluate each term in the above equation when $w=1$. The first term $\lambda_d'(1)=0$ because there is neither blocking nor holding effect to book a PFU when $w=1$ by the proof of Theorem 1. Since $\lambda_p=0$ and $\pi_K^Y=0$ when $w=1$, the last two terms vanish. So only the second and the third term remain non-zero and can be evaluated as

$$\begin{aligned}&-c\tfrac{d\pi_K^X}{dw}|_{w=1}(\lambda_r+\lambda_p\tfrac{\lambda_r}{\lambda_n+\lambda_r})+c\pi_K^X\tfrac{d}{dw}(\lambda_r+\lambda_p\tfrac{\lambda_r}{\lambda_n+\lambda_r})|_{w=1}\\=\;&-c\tfrac{d\pi_K^X}{dw}|_{w=1}\lambda_r-c\pi_K^X\lambda_d f(1)\tfrac{2\lambda_d\bar{p}+\lambda_n}{\lambda_d\bar{p}+\lambda_n}\\=\;&-c\pi_K^X\lambda_d f(1)\tfrac{2\lambda_d\bar{p}+\lambda_n}{\lambda_d\bar{p}+\lambda_n}.\end{aligned}\tag{EC.6}$$

The last equality follows from $\frac{d\pi_K^X}{dw}|_{w=1}=0$ as $\frac{d\lambda_v}{dw}|_{w=1}=0$ (there is no blocking effect by booking a patient with $w=1$ as PFU). Therefore, as long as $c>0$ and $f(1)>0$, the objective function has a strictly negative derivative at $w=1$, which implies the optimal PFU threshold $w^o<1$. ■

EC.3.2. Traditional Appointment System

In a traditional appointment system, the PFU will never balk. So we only need to penalize the number of RFUs that balk in order to improve same-doctor matching. That leads to the following objective function,

$$\begin{aligned}&\max_{w\in[0,1]}\ \lambda_d(w)-c(\text{RFU balking })\\=&\max_{w\in[0,1]}\ \lambda_d(w)-c\textstyle\sum_i\pi_i^*(w)b(i)\lambda_d(w)G(w).\end{aligned}\tag{EC.7}$$

We can prove a conclusion similar to that in an open access system – if one wants to maximize the effective throughput rate as well as same-doctor matching, then one should use a control threshold smaller than w^*, the optimal threshold for solely maximizing the effective throughput rate.

PROPOSITION EC.2. *Suppose the revisit probability p satisfies either of the two conditions in Theorem 2. If $1 \geq c > 0$, then the optimal solution to* (EC.7) *is strictly less than w^*, the unique maximizer of $\lambda_d(w)$.*

Proof. At any $w \geq w^*$, the derivative of (EC.7) can be bounded as follows,

$$\begin{aligned}
&\text{Derivative of (EC.7)} \\
&= \lambda_d'(w) - \lambda_d'(w) c \textstyle\sum_i \pi_i^* b(i) G(w) - c\lambda_d(w) G(w) \sum_i \frac{\partial \pi_i^*}{\partial w} b(i) - cwf(w)\lambda_d \sum_i \pi_i^* b(i) \\
&< -c \textstyle\sum_i \frac{\partial \pi_i^*}{\partial w} b(i) \lambda_d(w) G(w) - cwf(w)\lambda_d \sum_i \pi_i^* b(i) \\
&= \tfrac{c\lambda_d G(w)}{\lambda_n + \lambda_d G(w)} \textstyle\sum_i \lambda_d f(w)(1 - w(1-b(i)) - \gamma(1-w)) \left[\int_0^1 \pi_i(s) q_i(s,w) ds\right] - cwf(w)\lambda_d \sum_i \pi_i^* b(i) \\
&< c\Xi_1(\lambda_d, w) \\
&\leq 0,
\end{aligned} \tag{EC.8}$$

where the first inequality follows from $c\sum_i \pi_i^* b(i) G(w) < 1$ for all w and $\lambda_d'(w) \leq 0$ for $w \in [w^*, 1]$, the second equality follows from Equation (20) in Lemma 4, the second inequality follows from $\frac{\lambda_d G(w)}{\lambda_n + \lambda_d G(w)} < 1$ and the expression of $\Xi_1(\lambda_d, w)$, i.e., Equation (23), and the last inequality follows the fact that $\Xi_1(\lambda_d, w) \leq 0$ for all $w \geq w^*$, which is implied by the proof of Theorem 2. We thus deduce that the multi-objective function (EC.7) is strictly decreasing over $[w^*, 1]$. Therefore, the maximizer of (EC.7) must be smaller than w^*. ■

EC.4. Proof of Proposition 1

Proof. $\{X(t) | t \geq 0\}$ is a time-inhomogeneous Markov process with period one, as the service completes at integer times $t = 1, 2, \ldots$. Also, $X(t)$ is irreducible because any pair of states communicate with each other. The key to proving positive recurrence of $X(t)$ is to show that the virtual arrival rate is less than 1 for all w and sufficiently large $X(t)$. To that end, we note that the virtual arrival rate increases when more FUAs are designated as PFUs. Thus, it suffices to bound arrival rate when $w = 0$, i.e., all FUAs are booked as PFUs. When the queue length approaches infinity, the average balking rate converges to $b(\infty)$, which gives virtual arrival rate as

$$\lambda_n(1 - b(\infty)) + \lambda_n(1 - b(\infty))\overline{p} + \lambda_n(1 - b(\infty))\overline{p}^2 + \ldots = \frac{\lambda_n(1 - b(\infty))}{1 - \overline{p}} < 1, \tag{EC.9}$$

where $\lambda_n(1 - b(\infty))$ counts the average arrival rate of the new arrivals excluding those that balk, $\lambda_n(1 - b(\infty))\overline{p}$ denotes the PFUs generated by those new visits, and $\lambda_n(1 - b(\infty))\overline{p}^2$ counts the next generation of PFUs, etc. The last inequality follows from Condition (ii). We have thus proved that the virtual arrival rate is upper bounded by 1 when $X(t) \to \infty$ when $w = 0$. This implies that the virtual arrival rate is bounded away from 1 for all w when $X(t)$ is sufficiently large. This proves that the queue is stable and $X(t)$ is positive recurrent. Therefore, $X(t)$ must possess a steady-state distribution (in the periodic sense) $\{\boldsymbol{\pi}(t) | t \in [0, 1)\}$, which is unique by irreducibility. ■

EC.5. Proof of Lemma 3

Proof. Let $X^w(t)$, $A^w(t)$ denote the number of jobs in the system at time t and the cumulative number of arrived jobs up to time t, respectively, given a PFU control threshold w. Since $\frac{\partial \lambda_v(i,w,\lambda_d)}{\partial w} = f(w)\lambda_d(-b(i)w-(1-w)(1-\gamma)) < 0$, for sufficiently small $\Delta w > 0$ and all $i \geq 0$, the virtual arrival rate in the system with threshold $w-\Delta w$ must be larger than that with threshold w by an amount equal to $|\frac{\partial \lambda_v(i,w,\lambda_d)}{\partial w}|\Delta w + o(\Delta w)$. Thus, the process $A^{w-\Delta w}$ stochastically dominates A^w, which implies that $X^{w-\Delta w}$ stochastically dominates X^w. The stochastic dominance implies that $\mathbb{E}X^{w-\Delta w}(t) \geq \mathbb{E}X^w(t)$ for all t and consequently $\nabla_w(\boldsymbol{\pi}^*)^T\boldsymbol{x} \leq 0$. A similar argument can be applied to prove that $\nabla_{\lambda_d}(\boldsymbol{\pi}^*)^T\boldsymbol{x} \geq 0$ by noting that $\frac{\partial \lambda_v(i,w,\lambda_d)}{\partial \lambda_d} = G(w)(1-b(i)) + (1-\gamma)(1-F(w)) \geq 0$ for all i. ■

EC.6. Proof of Lemma 4

In order to streamline the proof of Lemma 4, we present some intermediate results in the following propositions.

Proposition EC.3. *For all $i = 1, 2, \ldots$, and for all w and Δw such that $0 \leq w - \Delta w \leq w \leq 1$,*

$$\mathbb{E}\ \omega(i,s) = -\frac{q_i(s,w)}{\lambda_n + G(w)\lambda_d}. \tag{EC.10}$$

where $q_i(s,w)$ is a continuous function that increases in i. Moreover, $q_i(s,w) \in (0,1)$ for all $s \in [0,1)$, and $\frac{\partial q_i(s,w)}{\partial w} \leq \frac{f(w)w\lambda_d}{\lambda_d G(w)+\lambda_n} q_i(s,w)$ for all $s \in [0,1)$.

Proposition EC.4. *If we randomly pick $T_{k-1}, T_k \in \mathcal{T}$, then with probability $1-o(1)$, (T_{k-1}, T_k) is a regular interval.*

Proposition EC.5. *For all sufficiently small $\Delta w > 0$ and $ds > 0$,*

$$\Pr([T_{k-1}, T_k] \in \Gamma_{i,ds}) = \frac{\Delta\lambda_v(i)\pi_i(s)ds}{\Delta\lambda_v^*} + o(1), \tag{EC.11}$$

where $o(1) \to 0$ when $\Delta w \to 0$.

Proofs for Propositions EC.3, EC.4, and EC.5 are provided in EC.6.1, EC.6.2, and EC.6.3, respectively. We next prove Lemma 4 using these propositions.

Proof of Lemma 4: To provide an overview of the proof, we will use the IPA method by considering two queue backlog processes X^w and $X^{w-\Delta w}$, and their corresponding stationary distributions $\boldsymbol{\pi}^*(w)$ and $\boldsymbol{\pi}^*(w-\Delta w)$, respectively. By the definition of $\boldsymbol{\pi}^*$, we have

$$\frac{\partial \mathbb{E}_{\boldsymbol{\pi}^*(w)} b(X^*)}{\partial w} = \lim_{\Delta w \to 0} \frac{1}{\Delta w} \lim_{T\to\infty} \frac{1}{T} \int_0^T (b(X^w(t)) - b(X^{w-\Delta w}(t)))dt. \tag{EC.12}$$

Since the virtual arrival rate decreases in w, we may couple X^w and $X^{w-\Delta w}$ such that $X^{w-\Delta w} \geq X^w$ for almost all the sample paths. We can also show that most of the time, $X^{w-\Delta w} - X^w = 0$;

but occasionally, the difference process $\Delta X(w,\Delta w) := X^{w-\Delta w} - X^w$ may jump upward by one, which is caused by the difference in the virtual arrival rates of the two processes, say, $\Delta\lambda_v(t)$. The difference $\Delta X(w,\Delta w)$ follows a nonhomogeneous Poisson process with time-varying arrival intensity $\Delta\lambda_v(t)$. Let $T_1, T_2, \ldots,$ denote the arrival epochs of $\Delta X(w,\Delta w)$. Then the long-run average difference in the balking rates, $\frac{1}{T}\int_0^T \frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt$, can be expressed as the sum of $E[\frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))]$ in all intervals $[T_{k-1}, T_k]$ $(k = 1, 2, \ldots,)$. Fortunately, in most of the intervals, the expected difference can be efficiently characterized.

Formally, let X^w and $X^{w-\Delta w}$ denote two backlog queue-length processes with PFU control-thresholds w and $w-\Delta w$, respectively. Because of equation (EC.12), it suffices to characterize the limit of $\frac{1}{\Delta w}\lim_{T\to\infty}\frac{1}{T}\int_0^T (b(X^w(t)) - b(X^{w-\Delta w}(t)))dt$ when $\Delta w \to 0$. As we argued in Lemma 3, $X^{w-\Delta w}$ stochastically dominates X^w, thus by coupling[4] the sample paths of X^w and $X^{w-\Delta w}$, it can be argued that $\Delta X(w,\Delta w,t) := X^{w-\Delta w}(t) - X^w(t) \geq 0$ for all $t \geq 0$ almost surely. Furthermore, since the long-run average does not depend on the initial distribution of the stochastic processes, we can assume both the initial distribution of X^w and $X^{w-\Delta w}$ to be exactly their steady-state distribution at time zero, i.e., $\boldsymbol{\pi}^w(0)$ and $\boldsymbol{\pi}^{w-\Delta w}(0)$, respectively, so that the two processes are periodic stationary processes with period one (having the same distribution at time points $k+s$ for all $s \in [0,1)$ and $k = 0, 1, \ldots$). Figure EC.1 plots an example of coupled sample paths of X^w and $X^{w-\Delta w}$.

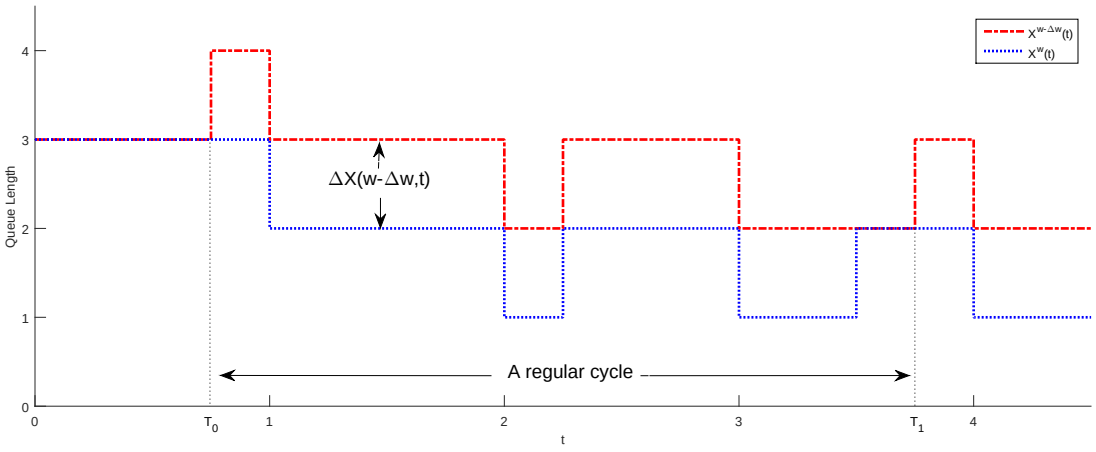


Figure EC.1 Sample paths for X^w, $X^{w-\Delta w}$. Their difference gives $\Delta X(w,\Delta w,t)$. According to the plot, $[T_0, T_1]$ is a regular interval.

[4] Coupling means that we redefine the mapping from the probability space to the space of sample paths, $\omega \mapsto X^w(\omega)(\cdot)$ and $\omega \mapsto X^{w-\Delta w}(\omega)(\cdot)$. Since the result of lemma only depends on the expected values of $\mathbb{E}\int_0^T b(X^w(s))ds$ and $\mathbb{E}\int_0^T b(X^{w-\Delta w}(s))ds$, coupling will not change the conclusion of the lemma.

The RHS of equation (EC.12) represents the long-run average value of $\frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))$, which, however, is difficult to calculate. Instead of calculating the long-run average directly, we first identify certain time intervals during which the average value of $\frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))$ is easier to calculate and can approximate the long-run average.

We consider possible time points t at which the gap $\Delta X(w, \Delta w, t)$ could possibly increase by one, that is, when $X^{w-\Delta w}$ has extra arrival compared to queue X^w. This happens only when

1. $\lambda_v(X^{w-\Delta w}(t), w-\Delta w, \lambda_d) - \lambda_v(X^w(t), w, \lambda_d) > 0$;
2. the above difference in their arrival rates leads to an actual arrival.

According to our coupling, $X^{w-\Delta w}(t) \geq X^w(t)$ at all t, so

$$\begin{aligned}\lambda_v(X^{w-\Delta w}(t), w-\Delta w, \lambda_d) - \lambda_v(X^w(t), w, \lambda_d) &< \lambda_v(X^w(t), w-\Delta w, \lambda_d) - \lambda_v(X^w(t), w, \lambda_d) \\ &=: \Delta\lambda_v(X^w(t)).\end{aligned} \tag{EC.13}$$

Therefore, the arrival process of $\Delta X(w, \Delta w, t)$ can be dominated by a non-homogeneous Poisson process $\mathcal{N}^{\Delta w}$ with time-varying arrival rate given by $\Delta\lambda_v(X^w(t))$, which is the difference in the arrival rates of the two processes upon ignoring the possible difference in their backlogs and balking rates. Let $\mathcal{T} := \{T_0, T_1, \ldots\}$ denote the sequence of arrival epochs of $\mathcal{N}^{\Delta w}$. Then the arrival epochs of $\Delta X(w, \Delta w, t)$ must be a subset of the arrival epochs of $\mathcal{N}^{\Delta w}$. Nevertheless, $\Delta X(w, \Delta w, t)$ does not have to increase at every time point in $\mathcal{T}$, because at some $T_k \in \mathcal{T}$, $X^{w-\Delta w}(T_k)$ may be strictly larger than X^w, in which case the $\lambda_v(X^{w-\Delta w}(t), w-\Delta w, \lambda_d) - \lambda_v(X^w(t), w, \lambda_d)$ may be negative due to the extra balking in queue $X^{w-\Delta w}$.

The time points in $\mathcal{T}$ partition any horizon $[0, T]$ into countably many intervals in the form of $[T_{k-1}, T_k)$. Within each interval, Δw cannot contribute to any extra arrival to the queue $X^{w-\Delta w}$ compared to queue X^w. We then reformulate the long-term average of $\frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))$ according to this partition,

$$\begin{aligned}&\lim_{T\to\infty} \tfrac{1}{T}\int_0^T \tfrac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\ =& \lim_{T\to\infty} \tfrac{1}{\Delta w T}\sum_{k=1}^{\mathcal{N}^{\Delta w}(T)} \int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\ =& \lim_{T\to\infty} [\tfrac{\mathcal{N}^{\Delta w}(T)}{\Delta w T}][\tfrac{1}{\mathcal{N}^{\Delta w}(T)}\sum_{k=1}^{\mathcal{N}^{\Delta w}(T)} \int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt].\end{aligned} \tag{EC.14}$$

Since $\mathcal{N}^{\Delta w}(T)$ is a Poisson random variable with intensity $\int_0^T \Delta\lambda_v(X^w(s))ds$, and $X^w(s)$ is a stationary process with distribution $\boldsymbol{\pi}(s - \lfloor s \rfloor)$ at any s. When $T \to \infty$, by ergodicity of $X^w(t)$, we have

$$\frac{\mathcal{N}^{\Delta w}(\Delta w T)}{T} = \frac{\int_0^T \Delta\lambda_v(X^w(s))ds + o(T)}{T} \to \int_0^1 \mathbb{E}(\Delta\lambda_v(X^w(s)))ds := \Delta\lambda_v^*, \tag{EC.15}$$

where $\Delta\lambda_v^*$ denotes the average virtual arrival rate for queue $X^w(\cdot)$, similar to the definition of $\boldsymbol{\pi}^*$.

We then continue to express the long-term average of $\frac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))$ as

$$\begin{aligned} &\lim_{T\to\infty} \tfrac{1}{T}\int_0^T \tfrac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\ &= [\tfrac{\Delta\lambda_v^*}{\Delta w}] \lim_{T\to\infty}[\tfrac{1}{\mathcal{N}^{\Delta w}(T)}\sum_{k=1}^{\mathcal{N}^{\Delta w}(T)}\int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt]. \end{aligned} \tag{EC.16}$$

We call a time point $s \in \mathcal{T}$ a *regular point* if

$$\Delta X(w, \Delta w, s^-) := \lim_{t\uparrow s} \Delta X(w, \Delta w, t) = 0. \tag{EC.17}$$

Intuitively, a regular point refers to a time right before which the sample paths of X^w and $X^{w-\Delta w}$ stick together. As a result, at time s, $\Delta X(w, \Delta w, s)$ must increase by one due to the extra arrival to queue $X^{w-\Delta w}$ led by the arrival-rate difference $\Delta\lambda_v(X^w(s))$. An interval $[T_{k-1}, T_k]$ is called a *regular interval* by having both its end points T_{k-1} and T_k as regular points; if at least one of the two endpoints of interval $[T_{k-1}, T_k]$ is non-regular, then we call $[T_{k-1}, T_k)$ a *non-regular interval*. A non-regular means that the sample paths of X^w and $X^{w-\Delta w}$ already have a gap before T_{k-1}, or the gap has not been closed before T_k. One may refer to Figure EC.1 for a graphic illustration of a regular interval.

In general, the term $\int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt$ is difficult to calculate. However, we can calculate this term if both T_{k-1} and T_k are both *regular points*. We define $\tau_= := \inf\{t > T_{k-1} | X^w(t) = X^{w-\Delta w}(t)\}$. Since T_{k-1} is a regular point, the two processes must be coupled before T_{k-1}, in which case an actual arrival would be triggered at T_{k-1} due to the difference in their arrival rates. So we have $X^{w-\Delta w}(t) \equiv X^w(t) + 1$ at $t \in [T_{k-1}, \tau_=)$, and the virtual arrival rate of $X^{w-\Delta w}$ is given by $\lambda_v(X^w + 1, w, \lambda_d)$, which does not depend on Δw (because the next extra arrival caused by Δw would occur not earlier than T_k). On the other hand, since T_k is a regular point, we must have $\tau_= < T_k$. So $X^{w-\Delta w}(t) \equiv X^w(t)$ at $t \in [\tau_=, T_k)$. Thus, we can define the following random variable, which is the difference in the balking rates during a regular interval,

$$\begin{aligned} \omega(i,s) &:= \textstyle\int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt | (T_{k-1}, T_k) \text{ regular, } X^w(T_{k-1}^-) = i, T_{k-1} - \lfloor T_{k-1}\rfloor = s \\ &= \textstyle\int_{T_{k-1}}^{\tau_=}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt | (T_{k-1}, T_k) \text{ regular, } X^w(T_{k-1}^-) = i, T_{k-1} - \lfloor T_{k-1}\rfloor = s \\ &= \textstyle\int_{T_{k-1}}^{\tau_=}(b(X^w(t)) - b((X^w + 1)(t)))dt |\ X^w(T_{k-1}^-) = i, T_{k-1} - \lfloor T_{k-1}\rfloor = s. \end{aligned} \tag{EC.18}$$

The expected value of $\omega(i,s)$, which represents the expected difference in the balking rates of the two processes in regular intervals, can be expressed as Equation (EC.10) according to Proposition EC.3. However, when $[T_{k-1}, T_k]$ are non-regular intervals, we are unable to characterize $\int_{T_{k-1}}^{T_k}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt$. But good news is that only a negligible amount of intervals $[T_{k-1}, T_k]$ are non-regular intervals, which is formally stated in Proposition EC.4.

We next reformulate the RHS at equation (EC.16) by separating the sum of regular intervals and non-regular intervals. For each non-regular interval $[T_{k-1}, T_k]$, we simply replace the integral

with $\omega(X^w(T_{k-1}), T_{k-1} - \lfloor T_{k-1} \rfloor)$. Such a replacement will only cause a difference of $o(1)$ to the average, because the non-regular intervals only take a proportion of $o(1)$.

$$\begin{aligned}
&\lim_{T\to\infty} \tfrac{1}{T}\int_0^T \tfrac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\
&= \tfrac{\Delta\lambda_v^*}{\Delta w} \lim_{T\to\infty} \tfrac{1}{\mathcal{N}^{\Delta w}(T)} (\textstyle\sum_{[T_{k-1},T_k] \text{ regular}} \int_{T_{k-1}}^{T_k} (b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\
&\quad + \textstyle\sum_{[T_{k-1},T_k] \text{ not regular}} \int_{T_{k-1}}^{T_k} (b(X^w(t)) - b(X^{w-\Delta w}(t)))dt) \\
&= [\tfrac{\Delta\lambda_v^*}{\Delta w}] \lim_{T\to\infty} [\tfrac{1}{\mathcal{N}^{\Delta w}(T)} (\textstyle\sum \text{ All } {}_{[T_{k-1},T_k]} \omega(X^w(T_{k-1}), T_{k-1}, \Delta w) \\
&\quad + \textstyle\sum_{[T_{k-1},T_k] \text{ not regular}} (\int_{T_{k-1}}^{T_k} (b(X^w(t)) - b(X^{w-\Delta w}(t)))dt - \omega(X^w(T_{k-1}), T_{k-1}, \Delta w))) \\
&= [\tfrac{\Delta\lambda_v^*}{\Delta w}] \lim_{T\to\infty} [\tfrac{1}{\mathcal{N}^{\Delta w}(T)} \left(\textstyle\sum \text{ All } {}_{[T_{k-1},T_k]} \omega(X^w(T_{k-1}), T_{k-1}, \Delta w)\right)] + o(1).
\end{aligned} \tag{EC.19}$$

Computing the RHS of the above equation is not straightforward, because the process is time inhomogeneous (but periodic) and $\mathbb{E}\omega(X^w(T_{k-1}), T_{k-1}, \Delta w)$ depends on $T_{k-1} - \lfloor T_{k-1} \rfloor$. We need to classify the intervals by the values of $T_{k-1} - \lfloor T_{k-1} \rfloor$ and count the proportion in each class. To facilitate the computation, we define the following partition of the intervals:

$$\Gamma_{i,ds} = \{[T_{k-1}, T_k] | T_{k-1} - \lfloor T_{k-1} \rfloor \in ds, X^w(T_{k-1}) = i\}, \text{ for all } i \geq 0, \ s \in [0, 1). \tag{EC.20}$$

where ds denotes a neighborhood centered at time $s \in (0, 1)$. For intervals in the same class $\Gamma_{i,ds}$, the corresponding random variables $\{\omega(X^w(T_{k-1}), T_{k-1} - \lfloor T_{k-1} \rfloor, \Delta w) | [T_{k-1}, T_k] \in \Gamma_{i,ds}\}$ are independently and identically distributed, with an expected value $\mathbb{E}\omega(i.s)$ as characterized in (EC.10). Therefore,

$$\begin{aligned}
&\lim_{T\to\infty} \tfrac{1}{\mathcal{N}^{\Delta w}(T)} \left(\textstyle\sum \text{ All } {}_{[T_{k-1},T_k]} \omega(X^w(T_{k-1}), T_{k-1}, \Delta w)\right) \\
&= \lim_{T\to\infty} \tfrac{1}{\mathcal{N}^{\Delta w}(T)} \int_0^1 \textstyle\sum_{i\geq 0} |\Gamma_{i,ds}| [\tfrac{1}{|\Gamma_{i,ds}|} \sum_{[T_{k-1},T_k] \in \Gamma_{i,ds}} \omega(X^w(T_{k-1}), T_{k-1} - \lfloor T_{k-1} \rfloor, \Delta w)] \\
&= \lim_{T\to\infty} \tfrac{1}{\mathcal{N}^{\Delta w}(T)} \int_0^1 \textstyle\sum_{i\geq 0} |\Gamma_{i,ds}| \mathbb{E}[\omega(X^w(T_{k-1}), T_{k-1} - \lfloor T_{k-1} \rfloor \in ds, \Delta w) | [T_{k-1}, T_k] \in \Gamma_{i,ds}] \\
&= \int_0^1 \textstyle\sum_{i\geq 0} \lim_{T\to\infty} \tfrac{|\Gamma_{i,ds}|}{\mathcal{N}^{\Delta w}(T)}] \mathbb{E}[\omega(X^w(T_{k-1}), T_{k-1} - \lfloor T_{k-1} \rfloor \in ds, \Delta w) | [T_{k-1}, T_k] \in \Gamma_{i,ds}] \\
&= \int_0^1 \textstyle\sum_{i\geq 0} \Pr([T_{k-1}, T_k] \in \Gamma_{i,ds}) \mathbb{E}[\omega(i, s, \Delta w)] \\
&= \int_0^1 \textstyle\sum_{i\geq 0} \tfrac{\Delta\lambda_v(s)}{\Delta\lambda_v^*} \pi_i(s) \mathbb{E}[\omega(i, s, \Delta w)] ds + o(1),
\end{aligned} \tag{EC.21}$$

where both the second and fourth equations follow from the strong law of large number, and the last equation follows from Proposition EC.5.

By plugging equation (EC.21) into equation (EC.19), we obtain that

$$\begin{aligned}
&\lim_{T\to\infty} \tfrac{1}{T}\int_0^T \tfrac{1}{\Delta w}(b(X^w(t)) - b(X^{w-\Delta w}(t)))dt \\
&= \tfrac{\Delta\lambda_v^*}{\Delta w} \int_0^1 \textstyle\sum_{i\geq 0} \pi_i(s) \tfrac{\Delta\lambda_v(i)}{\Delta\lambda_v^*} \mathbb{E}[\omega(i, s, \Delta w)] ds + o(1) \\
&= \int_0^1 \textstyle\sum_{i\geq 0} \pi_i(s) \tfrac{\Delta\lambda_v(i)}{\Delta w} \mathbb{E}[\omega(i, s, \Delta w)] ds + o(1) \\
&\to \tfrac{-1}{\lambda_n + \lambda_d G(w)} \int_0^1 \textstyle\sum_{i\geq 0} \pi_i(s) \tfrac{\partial(-\lambda_v(i,s,w)}{\partial w}) q_i(s, w) ds,
\end{aligned} \tag{EC.22}$$

as $\Delta w \to 0$, which completes the proof of Equation (20). The proof of Equation (21) follows the same logic except for taking the derivative with respect to λ_d instead of w, and is therefore omitted.

■

EC.6.1. Proof of Proposition EC.3

Throughout this proof, we use a simplified notation (i,s) to represent the condition of $X^w(T_{k-1})=i, T_{k-1}-\lfloor T_{k-1}\rfloor=s$. By multiplying and dividing $-(\lambda_n+G(w)\lambda_d)$ at the RHS of (EC.18), we obtain the following alternative expression for $\mathbb{E}[\omega(i,s)]$,

$$\begin{aligned}&\mathbb{E}[\omega(i,s)]\\ =&-\tfrac{1}{\lambda_n+G(w)\lambda_d}\mathbb{E}[\textstyle\int_s^{\tau_=}(\lambda_n+G(w)\lambda_d)(b((X^w+1)(t))-b(X^w(t)))dt|(i,s)]\\ =&\tfrac{1}{\lambda_n+G(w)\lambda_d}\mathbb{E}[\textstyle\int_s^{\tau_=}(\lambda_v(X^w(t)+1,w,\lambda_d)-\lambda_v(X^w(t),w,\lambda_d))dt|(i,s)],\end{aligned}\tag{EC.23}$$

where the last equality follows from the fact that $(\lambda_n+G(w)\lambda_d)(b(X^w(t)+1)-b(X^w(t)))$ represents the difference in the virtual arrival rates of the two processes. Therefore, the expectation $\mathbb{E}[\int_s^{\tau_=}(\lambda_n+G(w)\lambda_d)(b(X^w(t)+1)-b(X^w(t)))dt]$ has an interesting interpretation: how many extra customers balk in queue X^w+1 compared to queue X^w during the time interval $[T_{k-1},\tau_=]$?

To answer this question, we notice that during interval $[T_{k-1},\tau_=]$, $\tilde{X}^w$ has lost an extra backlog compared to X^w. This happens only in either of the two cases:

Case 1 An extra customer is served in queue $\tilde{X}^w:=X^w+1$ during $[T_{k-1},\tau_=]$ compared to queue X^w. This could happen because of the extra backlog, so even when queue X^w is empty and its server is idled, the server in queue X^w+1 should still be working and serve an extra customer. If we define $\tau_S(w):=\inf\{t>T_{k-1}|(X^w+1)(t)=0\}$ as the first time when the backlog process X^w+1 hits zero, then we know that $\tau_S(w)\geq 1$. We can infer that in Case 1, it must be the case $\tau_==\tau_S$, and $X^w(\tau_=)=\tilde{X}^w(\tau_=)=0$;

Case 2 Queue $\tilde{X}^w$ has one more customer balk than queue X^w during $[T_{k-1},\tau_=]$. This could happen because the balking rate for the longer queue $\tilde{X}^w$ is larger. Let $\tilde{A}^w(U)$ and $A^w(U)$ denote the cumulative number of arrived jobs for queue $\tilde{X}^w$ and X^w during time interval $U\subseteq\mathbb{R}$, respectively. Define $\tau_A(w):=\inf\{t|A^w(T_{k-1},t)-\tilde{A}^w(T_{k-1},t)=1\}$, the first time when $\tilde{X}^w$ had one more balked customer than X^w. In Case 2, we have $\tau_==\tau_A$.

According to the above discussions, during interval $[T_{k-1},\tau_=]$, queue $\tilde{X}^w$ could have either zero or one more extra balked customer compared to X^w, which corresponds to Case 1 and Case 2, respectively. Once $\tilde{X}^w$ has one more balked customer than X^w, the two sample paths must coincide, so $\tilde{X}^w$ cannot have more than one balked customers than X^w. Therefore,

$$\begin{aligned}&\mathbb{E}[\textstyle\int_s^{\tau_=}(\lambda_n+G(w)\lambda_d)(b(\tilde{X}^w(t))-b(X^w(t)))dt|(i,s)])\\ =&\,\mathbb{E}[\text{number of extra balkings in queue }\tilde{X}^w\text{ compared to that in queue }X^w|(i,s)]\\ =&\,\Pr(\text{Case 1}|(i,s))*0+\Pr(\text{Case 2}|(i,s))*1\\ =&\,\Pr(\tau_A(w)<\tau_S(w)|(i,s))\\ =:&\,q_i(s,w).\end{aligned}\tag{EC.24}$$

where the last equation holds by the fact that Case 2 happens if and only if $\tau_A(w)<\tau_S(w)$. Plugging the above equation into (EC.23) leads to the expression (EC.10) for $\mathbb{E}[\omega(i,s,\Delta w)|(i,s)]$.

We next prove that the function $q_i(s,w)$ defined above satisfies the properties specified in Proposition EC.3. Since $q_i(s,w)$ could be represented as a conditional probability, it must range in $(0,1)$. Furthermore, the conditional probability of $\Pr(\tau_A(w) < \tau_S(w))$ must increase in the initial queue length i, because $\tau_S(w)$ has to be larger when the queue initially has more backlogs. To upper bound $\frac{\partial q_i(s,w)}{\partial w}$, we express $q_i(s,w)$ using an alternative expression as follows,

$$\begin{aligned} q_i(s,w) &= 1 - \Pr(\tau_A(w) > \tau_S(w)) \\ &= 1 - \Pr\left(A^w([T_{k-1}, \tau_S(w)]) = \tilde{A}^w([T_{k-1}, \tau_S(w)])\right), \end{aligned} \tag{EC.25}$$

where the second equation follows that $\tau_A(w) > \tau_S(w)$ if and only if the arrival processes of the two queues, A^w and $\tilde{A}^w$, are identical in $[T_{k-1}, \tau_S(w)]$. Because $\lambda_v(X^w+1, w, \lambda_d) \le \lambda_v(X^w, w, \lambda_d)$, the two arriving process $\tilde{A}^w$ and A^w can be coupled in the way that $\tilde{A}^w(s,t) \le A^w(s,t)$ on any interval $[s,t]$, and the process $A^w - \tilde{A}^w$ is a compound Poisson process $\mathcal{N}^{\delta w}$ which has arrival intensity

$$\Delta\lambda_b(X^w(t)) := \lambda_v(X^w+1, w, \lambda_d) - \lambda_v(X^w, w, \lambda_d) = (b(X^w(t)+1) - b(X^w))(\lambda_n + G(w)\lambda_d). \tag{EC.26}$$

Therefore, $A^w(T_{k-1}, \tau_S(w)) - \tilde{A}^w(T_{k-1}, \tau_S(w))$ is a Poisson random variable with mean $\int_{T_{k-1}}^{\tau_S(w)} \Delta\lambda_b(X^w(t))dt$. Note that $\Delta\lambda_b$ is led by the difference in balking rates, which differs from $\Delta\lambda_v$, which is led by Δw.

If we define $\tau_j^w(s,t) := \int_s^t I(X^w(x) = j)dx$ as the total time that X^w remains in state j within the time window $[s,t]$, then

$$\begin{aligned} \textstyle\int_{T_{k-1}}^{\tau_S(w)} \Delta\lambda_b(X^w(t))dt &= \textstyle\int_{T_{k-1}}^{\tau_S(w)} \Delta\lambda_b(X^w(t))dt \\ &= \textstyle\sum_j \int_{T_{k-1}}^{\tau_S(w)} \Delta\lambda_b(j) I(X^w(t) = j)dt \\ &= \textstyle\sum_j \tau_j^w(T_{k-1}, \tau_S(w))\Delta\lambda_b(j). \end{aligned} \tag{EC.27}$$

Thus, equation (EC.25) implies that

$$\begin{aligned} q_i(s,w) &= 1 - \mathbb{E}[\exp\left(-\textstyle\int_{T_{k-1}}^{\tau_S(w)} \Delta\lambda_b(X^w(t))(dt)\right)] \\ &= 1 - \mathbb{E}[\exp\left(-\textstyle\sum_j \tau_j^w(T_{k-1}, \tau_S(w))\Delta\lambda_b(j)\right)] \\ &= 1 - \mathbb{E}[\exp\left(-\textstyle\sum_j \tau_j^w(T_{k-1}, \tau_S(w))(b(j+1) - b(j))(\lambda_n + G(w)\lambda_d)\right)]. \end{aligned} \tag{EC.28}$$

Now suppose w has been increased by a sufficiently small amount $\delta w > 0$. We may express $\frac{\partial q_i(s,w)}{\partial w}$ as the following limit,

$$\begin{aligned} \tfrac{\partial q_i(s,w)}{\partial w} &= \lim_{\delta w \to 0} \tfrac{1}{\delta w}(q_i(s,w) - q_i(s, w-\delta w)) \\ &= \lim_{\delta w \to 0} \tfrac{1}{\delta w}\mathbb{E}[\exp\left(-\textstyle\sum_j \tau_j^{w-\delta w}(T_{k-1}, \tau_S(w-\delta w))(b(j+1) - b(j))(\lambda_n + G(w-\delta w)\lambda_d)\right)] \\ &\quad - \tfrac{1}{\delta w}\mathbb{E}[\exp\left(-\textstyle\sum_j \tau_j^w(T_{k-1}, \tau_S(w))(b(j+1) - b(j))(\lambda_n + G(w)\lambda_d)\right)], \end{aligned} \tag{EC.29}$$

where $X^{w-\delta w}$ is a queue backlog process corresponding to parameter $w - \delta w$ and with initial state $X^{w-\delta w}(T_{k-1}) = X^w(T_{k-1})$, and $\tau_j^{w-\delta w}(T_{k-1}, \tau_S(w-\delta w))$ denotes the total amount of time that a process $X^{w-\delta w}$ spends in state j.

We next show that with probability one, for all $j=0,1,\ldots,$

$$\tau_j^w(T_{k-1},\tau_S(w)) \le \tau_j^{w-\delta w}(T_{k-1},\tau_S(w-\delta w)). \tag{EC.30}$$

To prove the above inequality, we construct a sequence of intervals $\mathcal{A}:=\cup_{n=1}^N[a_n,b_n)$ over the time horizon $[T_{k-1},\tau_S(w-\delta w)]$, that is, before $X^{w-\delta w}$ hits zero. The specific construction of intervals $[a_n,b_n]$ proceeds as follows.

Let $\mathcal{S}$ denote the arrival epochs of a non-homogeneous Poisson process $\mathcal{N}^{\delta w}$. Let

$$a_n=\min\{v\in\mathcal{V}|v\ge b_{n-1}\}, \tag{EC.31}$$

that is, the next time at which queue $X^{w-\delta w}$ has an extra arrival compared to queue X^w after the previous interval (When $n=1$, we have $a_1=v_1$ by simply letting $b_0=T_{k-1}$.

Suppose $s=a_n-\lfloor a_n\rfloor$ and $j=X^{w-\delta w}(a_n^-)(:=\lim_{t\uparrow a_n}X^{w-\delta w}(t))$. Since $X^{w-\delta w}$ has increased to $j+1$ at time a_n, by looking into the sample paths of $X^{w-\delta w}$, we can deduce that $X^{w-\delta w}$ must visit state (j,s) at least once before hitting zero. We then define b_n as the first time at which $X^{w-\delta w}$ visits (j,s), that is,

$$b_n=\inf\{t>a_n|X^{w-\delta w}(t)=X^{w-\delta w}(a_n^-) \text{ and } s=t-\lfloor t\rfloor\}. \tag{EC.32}$$

According to the construction of $\mathcal{A}$, there are no points in $\mathcal{V}$ that lie in intervals $\cup_{n\ge1}^N[b_{n-1},a_n)\cup[b_N,\tau_S(w-\delta w))$, which is obtained by removing all intervals from $\mathcal{A}$ from time horizon $[T_{k-1},\tau_S(w+\delta w)]$. Therefore, δw causes no difference for the two sample paths during the time intervals $\cup_{n\ge0}^N[b_{n-1},a_n)\cup[b_N,\tau_S(w-\delta w))$. Moreover, we always have $X^{w-\delta w}(b_n)=X^w(a_n)$, and $b_n-\lfloor b_n\rfloor=a_n-\lfloor a_n\rfloor$ by our construction of b_n. Because the stochastic behavior of X^w and $X^{w-\delta w}$ depends on history only through state $(X^w(b_n),b_n-\lfloor b_n\rfloor)$, we can exactly couple the sample paths $\{X^w(t):t\in[T_{k-1},\tau_S(W)]\}$ and $\{X^{w-\delta w}(t):t\in\cup_{n\ge0}^N[b_{n-1},a_n)\cup[b_N,\tau_S(w-\delta w))\}$. As a result, for all $j=0,1,\ldots,$

$$\begin{aligned}\tau_j^w(T_{k-1},\tau_S(w)) &= \textstyle\int_{T_{k-1}}^{\tau_S(w)} I(X^w(t)=j)dt\\ &= \textstyle\int_{T_{k-1}}^{\tau_S(w-\delta w)} I\left(X^{w-\delta w}(t)=j, t\in\cup_{n\ge0}^N[b_{n-1},a_n)\cup[b_N,\tau_S(w-\delta w))\right)dt\\ &\le \textstyle\int_{T_{k-1}}^{\tau_S(w-\delta w)} I\left(X^{w-\delta w}(t)=j\right)dt\\ &= \tau_j^{w-\delta w}(T_{k-1},\tau_S(w-\delta w)),\end{aligned} \tag{EC.33}$$

which proves inequality (EC.30). Furthermore,

$$\begin{aligned}&\tfrac{1}{\delta w}\mathbb{E}[\exp\left(-\textstyle\sum_j\tau_j^{w-\delta w}(T_{k-1},\tau_S(w-\delta w))(b(j+1)-b(j))(\lambda_n+G(w-\delta w)\lambda_d)\right)\\ \le\ &\tfrac{1}{\delta w}\mathbb{E}[\exp\left(-\textstyle\sum_j\tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\right).\end{aligned} \tag{EC.34}$$

Plugging inequality (EC.34) into the RHS of equation (EC.29) leads to

$$\begin{aligned}
&\tfrac{\partial q_i(s,w)}{\partial w}\\
&= \lim_{\delta w\to 0} \tfrac{1}{\delta w}\mathbb{E}[\exp\Big(-\textstyle\sum_j \tau_j^{w-\delta w}(T_{k-1},\tau_S(w-\delta w))(b(j+1)-b(j))(\lambda_n+G(w-\delta w)\lambda_d)\Big)]\\
&\quad -\tfrac{1}{\delta w}\mathbb{E}[\exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&\le \lim_{\delta w\to 0} \tfrac{1}{\delta w}\mathbb{E}[\exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w-\delta w)\lambda_d)\Big)]\\
&\quad -\tfrac{1}{\delta w}\mathbb{E}[\exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&= -G'(w)\tfrac{\partial}{\partial G(w)}\mathbb{E}[\exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&= -f(w)(w)\mathbb{E}[\tfrac{\partial}{\partial G(w)}\exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&= -f(w)w\mathbb{E}[-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))\lambda_d\\
&\quad \exp\Big(-\textstyle\sum_{j\ge 0} \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&= \tfrac{f(w)w\lambda_d}{\lambda_n+G(w)\lambda_d}\mathbb{E}[\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\\
&\quad \exp\Big(-\textstyle\sum_j \tau_j^{w}(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)\Big)]\\
&=: \tfrac{f(w)w\lambda_d}{\lambda_n+G(w)\lambda_d}\mathbb{E}[z\exp(-z)],
\end{aligned} \tag{EC.35}$$

where the third equality follows from validity of swapping the derivative and expectation operators due to the bounded convergence theorem (the derivative is uniformly bounded), the last equation follows by defining $z := \sum_{j\ge 0}\tau_j^w(T_{k-1},\tau_S(w))(b(j+1)-b(j))(\lambda_n+G(w)\lambda_d)$. Because $z\exp(-z)\le 1-\exp(-z)$ for all $z\ge 0$, and $1-\exp(-z)=q_i(s,w)$ by equation (EC.28). We have

$$\frac{\partial q_i(s,w)}{\partial w} \le \frac{f(w)w}{\lambda_n+G(w)\lambda_d}(1-\exp(-z)) = \frac{f(w)w\lambda_d}{\lambda_n+G(w)\lambda_d}q_i(s,w). \tag{EC.36}$$

■

EC.6.2. Proof of Proposition EC.4

Since X^w is positive recurrent as we proved in Proposition 1, the sample path of X^w has to visit state 0 for infinitely many times. We use τ_0 and τ_0^C, respectively, to denote the random sojourn time when each time X^w visits state 0 and and all other states, respectively. By positive recurrence, we have $\mathbb{E}\tau_0^C<\infty$. Since the arrival rate of X^w at $X^w=0$ is upper bounded by $\lambda_d+\lambda_n$, it takes at least an exponentially distributed random period with mean $\frac{1}{\lambda_d+\lambda_n}$ for X^w to exit 0. Thus, we deduce that $\tau_0 \overset{st}{\ge} \mathrm{EXP}(\lambda_d+\lambda_n)$ [5]. When $X^{w-\Delta w}(t)>X^w(t)=0$, then the balking rate of $X^{w-\Delta w}(t)$ is strictly larger than that of $X^w(t)$ by at least $b(1)-b(0)$. Thus, the virtual arrival rate of $X^{w-\Delta w}$ is less than that of X^w by at least $(b(1)-b(0))(\lambda_n+G(w)\lambda_d)$ when the latter hits zero. By coupling $X^{w-\Delta w}$ and X^w, their difference $\Delta X(w,\Delta w,t) := X^{w-\Delta w}-X^w$ has a departure rate of at least $(b(1)-b(0))(\lambda_n+G(w)\lambda_d)$ after X^w visits zero.

[5] Note that $\overset{st}{\ge}$ represents the usual stochastic order. In particular, $X\overset{st}{\ge}Y$ if and only if $\Pr(X\le z)\le\Pr(Y\le z)$ for all $z\in\mathbb{R}$.

Let τ_0 denote the random sojourn time of X^w at state 0. Let τ_0^C denote the random period from X^w leaves state 0 till the next time when X^w visits 0 again. Since $\tau_0 \overset{st}{\geq} \text{EXP}(\lambda_d+\lambda_n)$, each time when $X^w(t)$ hits 0, a departure happens to $\Delta X(w,\Delta w,t)$ before X^w leaves state 0 with probability of at least $\frac{(b(1)-b(0))(\lambda_n+G(w)\lambda_d)}{(b(1)-b(0))(\lambda_n+G(w)\lambda_d)+\lambda_n+\lambda_d}$. Because X^w visits 0 once after a random period τ_0^C, the time period between two successive departures of $\Delta X(w,\Delta w,t)$ can be upper bounded by $\sum_{i=1}^{\kappa} Z_i$, where $Z_i \overset{d}{=} \tau_0^C+\tau_0$, and κ is a geometric distributed random variable with "success" rate $\frac{(b(1)-b(0))(\lambda_n+G(w)\lambda_d)}{(b(1)-b(0))(\lambda_n+G(w)\lambda_d)+\lambda_n+\lambda_d}$ and "success" corresponds to a departure event of $\Delta X(w,\Delta w,t)$.

By analyzing the expected inter-arrival and inter-departure time of $\Delta X(w,\Delta w,t)$, we deduce that $\Delta X(w,\Delta w,t)$ can be dominated by a stochastic process $\overline{\Delta X}(t)$, which is the length of an M/G/1 queue with arrival rate $o(1)$ and random service time $\sum_{i=1}^{\kappa} Z_i$, whose expectation $\mathbb{E}\kappa\mathbb{E}(\tau_0^C+\tau_0)$ is finite and independent to Δw. We know that at steady state $\Pr(\overline{\Delta X}(t)>0)=o(1)$, which also implies that a proportion of $1-o(1)$ time points in $\mathcal{T}$ qualify as the endpoints of regular intervals. ■

EC.6.3. Proof of Proposition EC.5

We first state two facts: (1) X^w is a periodic stationary process with distribution $\boldsymbol{\pi}(s)$ at any time t with $t-\lfloor t\rfloor = s$; (2) as a result of (1), $\mathcal{N}^{\Delta w}$ is a non-homogeneous Poisson process with time-varying arrival intensity $\sum_j \pi_j(s)\Delta\lambda_v(j)$ at any time t with such that $t-\lfloor t\rfloor = s$.

Using the Bayes' rule, we can formulate the probability in equation (EC.11) as follows,

$$\begin{aligned}\Pr([T_{k-1},T_k]\in\Gamma_{i,ds}) &= \Pr(T_{k-1}-\lfloor T_{k-1}\rfloor \in ds, X^w(T_{k-1})=i)\\ &= \Pr(T_{k-1}-\lfloor T_{k-1}\rfloor \in ds)\Pr(X^w(T_{k-1})=i|T_{k-1}-\lfloor T_{k-1}\rfloor\in ds)\\ &= \Pr(T_{k-1}-\lfloor T_{k-1}\rfloor \in ds)\tfrac{\pi_i(s)\Delta\lambda_v(i)}{\sum_j \pi_j(s)\Delta\lambda_v(j)}.\end{aligned} \tag{EC.37}$$

The arrival epochs of a non-homogeneous, periodic Poisson process are not mutually independent, so it is difficult to characterize the probability $\Pr(T_{k-1}-\lfloor T_{k-1}\rfloor\in ds)$ directly. We thus try to calculate the conditional probability $\Pr(T_{k-1}-\lfloor T_{k-1}\rfloor\in ds|T_{k-2}-\lfloor T_{k-2}\rfloor=h)$ for each $h\in[0,1)$. Given $T_{k-2}-\lfloor T_{k-2}\rfloor=h$, then by the properties of the non-homogeneous Poisson process, we have

$$\begin{aligned}&\Pr\left(T_{k-1}-\lfloor T_{k-1}\rfloor\in ds\,|\,T_{k-2}-\lfloor T_{k-2}\rfloor=h\right)\\ &=\textstyle\sum_j[\Pr(T_{k-1}<j+s+ds\,|\,T_{k-1}\geq j+s)][\Pr(T_{k-1}\geq j+s\,|\,T_{k-2}-\lfloor T_{k-2}\rfloor=h)]\\ &=\textstyle\sum_j[\sum_j \pi_j(s)\Delta\lambda_v(j)ds][\exp\left(-\int_h^{j+s}\Delta\lambda_v(t)dt\right)]\\ &=\textstyle[\sum_j \pi_j(s)\Delta\lambda_v(j)ds]\sum_{j\geq 0}[\exp\left(-\int_0^{j+s}\Delta\lambda_v(t)dt\right)]\exp(\int_0^h\Delta\lambda_v(t)dt)\\ &=\textstyle[\sum_j \pi_j(s)\Delta\lambda_v(j)ds]\frac{\exp\left(\int_0^h\Delta\lambda_v(t)dt-\int_0^s\Delta\lambda_v(t)dt\right)}{1-\exp(-\Delta\lambda_v^*)}\\ &=\textstyle[\sum_j \pi_j(s)\Delta\lambda_v(j)ds]\frac{1+o(1)}{\Delta\lambda_v^*+o(\Delta\lambda_v^*)}\\ &=\textstyle\frac{[\sum_j \pi_j(s)\Delta\lambda_v(j)ds](1+o(1))}{\Delta\lambda_v^*},\end{aligned} \tag{EC.38}$$

where the summation starts with $j=0$ when $s\geq h$, or with $j=1$ otherwise. The fourth equality follows from the fact that $\sum_{j\geq 0}[\exp\left(-\int_0^{j+s}\Delta\lambda_v(t)dt\right)]$ is a geometric series with ratio

$\exp(-\Delta\lambda_v^*)$ $\left(:=\exp(-\int_0^1 \Delta\lambda_v(t)dt)\right)$, and the fifth equality follows from the Taylor expansion of $\exp(-\Delta\lambda_v^*)$ at zero and the fact that $|\int_0^h \Delta\lambda_v(t)dt - \int_0^s \Delta\lambda_v(t)dt| < |\int_0^1 \Delta\lambda_v(t)dt| = o(1)$.

By the property of non-homogeneous Poisson process, the probability of $T_{k-1} - \lfloor T_{k-1}\rfloor \in ds$ depends on the history only through the value of h. Equation (EC.38) further states that the choice of h only makes a difference $o(1)$ to that probability. We thus conclude that

$$\begin{aligned}\Pr\left(T_{k-1} - \lfloor T_{k-1}\rfloor \in ds\right) &= \mathbb{E}_h[\Pr\left(T_{k-1} - \lfloor T_{k-1}\rfloor \in ds | T_{k-2} - \lfloor T_{k-2}\rfloor = h\right)] \\ &= \frac{\sum_j \pi_j(s)\Delta\lambda_v(j)ds(1+o(1))}{\Delta\lambda_v^*},\end{aligned} \tag{EC.39}$$

where the second equation follows from equation (EC.38). Plugging equation (EC.39) into equation (EC.37) leads to the conclusion of the proposition. ■.

EC.7. Proof of Lemma 5

Proof. Because of balking and no-show, every new arrival only has probability $(1-\mathbb{E}b(X^*))(1-\eta)$ to be actually served and contribute to an effective departure. Each of those served patient has probability $G(w) = \int_0^w pf(p)dp$ and $\overline{p} - G(w) = \int_w^1 pf(p)dp$ to request an RFU or an effective PFU, respectively. By the RTA assumption, the backlog follows the stationary distribution upon each arrival of each PFU, so the average balking rate for each PFU is $\mathbb{E}b(X^*) := \sum \int_0^1 \pi_i(s)b(i)ds$. Therefore, each served patient results in a second effective slot with probability $(1-\eta)\left(G(w)(1-\mathbb{E}b(X^*)) + \overline{p} - G(w)\right)$. If we total up all the expected number of returning customers led by every new visit, we derive an expression for the effective departures λ_d as a sum of geometric series,

$$\lambda_d = \sum_{k=0}^{\infty}(1-\eta)(1-\mathbb{E}b(X^*))\lambda_n\left((1-\eta)(G(w)(1-\mathbb{E}b(X^*)) + \overline{p} - G(w))\right)^k. \tag{EC.40}$$

By ignoring the no-show rate, we can derive an upper bound for λ_d as

$$\begin{aligned}\lambda_d &\le \textstyle\sum_{k=0}^{\infty}(1-\mathbb{E}b(X^*))\lambda_n\left(G(w)(1-\mathbb{E}b(X^*)) + \overline{p} - G(w)\right)^k \\ &\le \tfrac{(1-\mathbb{E}b(X^*))\lambda_n}{1-(G(w)(1-\mathbb{E}b(X^*))+\overline{p}-G(w))}.\end{aligned} \tag{EC.41}$$

By $\mathbb{E}b(X^*) \ge 0$, we derive a lower bound for λ_n as

$$\lambda_n \ge \frac{1-\overline{p}+G(w)}{1-\mathbb{E}b(X^*)}\lambda_d - G(w)\lambda_d \ge (1-\overline{p})\lambda_d. \tag{EC.42}$$

Plugging the above inequality into inequality (35), we know that it suffices to show the following inequality

$$1 \ge \frac{2\lambda_d w^2 f(w)}{(1+\gamma)(1-\overline{p}+G(w))\lambda_d} = \frac{2w^2 f(w)}{(1+\gamma)(1-\overline{p}+G(w))}. \tag{EC.43}$$

We next prove the above inequality when p satisfies either condition in Theorem 2. If $p \sim Beta(\alpha,\beta)$ with $\beta > 1$, then $f(p)$ must decreases in p due to $\beta > 1$. Thus, $f(p) \ge f(w)$ for $p \in [0,w]$. Thus,

$$\int_0^w pf(p)dp \ge \int_0^w pf(w)dp = \frac{1}{2}w^2 f(w). \tag{EC.44}$$

In addition, since $\overline{p} \le 1/4$,

$$\begin{aligned} 2w^2 f(w) + (1+\gamma)(\overline{p} - G(w)) &= 4(\tfrac{1}{2} w^2 f(w)) + (1+\gamma)\int_w^1 f(p)pdp \\ &\le 4\int_0^w pf(p)dp + (1+\gamma)\int_w^1 f(p)pdp \\ &\le 4\int_0^1 p(fp)dp + \gamma(\overline{p} - G(w)) \\ &= 4\overline{p} + \gamma(\overline{p} - G(w)) \\ &\le 1 + \gamma(\overline{p} - G(w)) \\ &\le 1 + \gamma. \end{aligned} \tag{EC.45}$$

As a result, $2w^2 f(w) < (1+\gamma)(1-\overline{p}+G(w))$, which proves inequality (EC.43) and thus inequality (EC.41).

Alternatively, if p has a uniform distribution over $[a,b]$, then $f(w) = 1/(b-a)$ for all $w \in [a,b]$. Thus,

$$\begin{aligned} 2w^2 f(w) + (1+\gamma)(\overline{p} - G(w)) &= \tfrac{4w^2}{2(b-a)} + (1+\gamma)\tfrac{b^2-w^2}{2(b-a)} \\ &\le (1+\gamma)\tfrac{3w^2+b^2}{2(b-a)} \\ &\le (1+\gamma)\tfrac{4b^2}{2(b-a)} \\ &\le 1+\gamma. \end{aligned} \tag{EC.46}$$

where the second inequality follows from $w \le b$. ■

EC.8. Computation of the Effective Throughput Rate

The RTA-based analytical framework introduced in Sections 5 and 6 allows us to explicitly compute the stationary distribution of the queue lengths $\boldsymbol{\pi}^*(w)$ and the mean-preserving effective throughput rate $\lambda_d(w)$ under a given w in both an open-access system or a traditional appointment-booking system with state-dependent balking. Computing the values of $\lambda_d(w)$ allows us to evaluate the PFU policy being used as well as identify the optimal threshold w^*.

The main idea of computing $\lambda_d(w)$ is to iteratively solve the fixed-point equations (7) (for open-access system) and (11) (for traditional booking system). Because the state space for traditional booking system can be infinitely large, we truncate the state space by enforcing the queue length to stop increasing when it hits a large number M. Then we can compute the stationary distribution by solving $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P$. This algorithm, referred to as the RTA algorithm, is provided below. We only provide a version for the traditional appointment system, but one can easily adapt this version to the open-access system.

Initialize: Set the number of iteration $m \leftarrow 0$. Assign $\lambda_d^{(m)}$ with any number in $(0,1)$.

Step 1: In the m-the iteration, solve the probability transition matrix $\boldsymbol{P} := (p_{ij})$ for the embedded discrete-time Markov chain $\{X(k)|k = 1,2\ldots\}$. The stationary distribution of this discrete embedded Markov chain gives us $\boldsymbol{\pi}(0)$, the stationary distribution for the backlog size after the completion of each slot. The probability transition matrix $\boldsymbol{P}$ can be calculated as follows,

$$P_{0j} = \begin{cases} q_{0,0}(1) + q_{0,1}(1) & \text{if } j = 0 \\ q_{0,j+1}(1) & \text{if } j \ge 1, \end{cases} \qquad P_{ij} = \begin{cases} 0 & \text{if } j < i-1, i > 0 \\ q_{i,j+1}(1) & \text{if } j \ge i-1, i > 0. \end{cases} \tag{EC.47}$$

In the above equation, $\{q_{i,j}(s)|j=0,1,\ldots\}$ is the transition probability for a pure-birth process that stayed at initial state i at time 0 and had moved to state j at time $s\in[0,1]$. Since the pure-birth process has a state-dependent birth rate $\lambda_v(i,w,\lambda_d^{m-1})$ given by (8). The transition probability can be computed recursively using the following equation; see (Ross 2014).

$$q_{i,j}(s):=\begin{cases}0 & \text{if } j<i\\ \exp(-\lambda_v(i,w,\lambda_d^{(m-1)})s) & \text{if } j=i\\ \lambda_v(j-1,w,\lambda_d^{(m-1)})\exp(-\lambda_v(j,w,\lambda_d^{(m-1)})s)\int_0^s\exp(\lambda_v(j,w,\lambda_d^{(m-1)})t)q_{i,j-1}(t)dt & \text{if } j>i.\end{cases} \tag{EC.48}$$

Step 2: Truncate the transition matrix $\boldsymbol{P}$ and $\boldsymbol{Q}$ to only have state $0,1,2,\ldots,n$ for some large integer M (e.g.,$M=100$). Normalize each row to have sum of one. Calculate the stationary distribution for the embedded Markov process, as

$$(\boldsymbol{\pi}^{(m)}(0))^T\leftarrow\lim_{N\to\infty}(\frac{1}{M}\boldsymbol{e_M}^{(0)})^T\boldsymbol{P}^N, \tag{EC.49}$$

where e_M is an M-dimensional all-one vector.

Step 3: Calculate $\boldsymbol{\pi}^{(m)}(s)$ at other $s\in(0,1)$ using

$$(\boldsymbol{\pi}^{(m)}(s))^T\leftarrow(\boldsymbol{\pi}^{(m)}(0))^T\boldsymbol{Q}(s), \tag{EC.50}$$

where $\boldsymbol{Q}(s):=(q_{ij}(s))$ for $s\in[0,1)$. We then have update

$$(\pi_i^*)^{(m)}\leftarrow\int_0^1\pi_i^{(m)}(s)ds \text{ for } i=0,1,\ldots,n. \tag{EC.51}$$

Step 4: With $(\boldsymbol{\pi}^*)^{(m)}$, calculate the average balking rate as $\mathbb{E}b(X^*)^{(m)}=\sum_i(\pi_i^*)^{(m)}b(i)$. Update λ_d as

$$\lambda_d^{(m+1)}\leftarrow=(1-\eta)\mathbb{E}_{(\boldsymbol{\pi}^*)^{(m)}}\left((1-b(X^*))(\lambda_n+G(w)\lambda_d^{(m)})+(\overline{p}-G(w))\lambda_d^{(m)}\right). \tag{EC.52}$$

Whenever $|\lambda_d^{(m+1)}-\lambda_d^{(m)}|<\epsilon$ for some tolerance ϵ, terminate the algorithm; otherwise, update $m\leftarrow m+1$ and go back to Step 1.

The approximation mainly stems from Step 2 and 3, where we have to truncate the infinite-dimensional matrix $\boldsymbol{P}$ to a M-dimensional matrix and use the distribution the N-th transition to approximate $\boldsymbol{\pi}(0)$. If we ignore the approximation involved in these two steps, then the algorithm is guaranteed to converge by the following Proposition.

PROPOSITION EC.6. *If the RTA algorithm can calculate the $(\boldsymbol{\pi}^*)^{(m)}$ accurately for a given $\lambda_d^{(m)}$ in each iteration, then the algorithm generates a sequence of $\lambda_d^{(m)}$ that converges to $\lambda_d(w)$, the unique solution to Equation* (9).

Proof. Given any w, consider the function $V(\lambda_d)$ that we have defined in the proof of Proposition 1. According to Lemma 4, the derivative $\frac{\partial \mathbb{E}_{(\boldsymbol{\pi}^*)(\lambda_d)}b(X^*)}{\partial \lambda_d}$ exists. So we may calculate the derivative of $V(\lambda_d)$ as

$$\begin{aligned} V'(\lambda_d) &= -(1-\eta)\tfrac{\partial \mathbb{E}_{(\boldsymbol{\pi}^*)(x)}b(X^*)}{\partial \lambda_d}(\lambda_n + G(w)\lambda_d) + (1-\eta)\left((1-b(X^*))G(w) + (\overline{p} - G(w))\right) \\ &= -(1-\eta)\textstyle\sum_i \frac{\partial \lambda_v(i,w,\lambda_d)}{\partial \lambda_d}\left[\int_0^1 \pi_i(s)q_i(s,w)ds\right] + (1-\eta)\left((1-b(X^*))G(w) + (\overline{p} - G(w))\right) \\ &= -(1-\eta)\textstyle\sum_i \left((1-b(i))G(w) + 1 - F(w)\right)\left[\int_0^1 \pi_i(s)q_i(s,w)ds\right] \\ &\quad +(1-\eta)\left((1-b(X^*))G(w) + (\overline{p} - G(w))\right). \end{aligned} \tag{EC.53}$$

Since $\sum_i \left((1-b(i))G(w) + 1 - F(w)\right)\left[\int_0^1 \pi_i(s)q_i(s,w)ds\right] < 1$, and $(1-b(X^*))G(w) + (\overline{p} - G(w)) < 1$, we deduce that

$$-1 < V'(\lambda_d) < 1. \tag{EC.54}$$

Therefore, $V(\lambda_d)$ is a contracting mapping. As we proved in Proposition 1, the equation $V(\lambda_d) = \lambda_d$ always has a unique fixed point $\lambda_d \in [0,1)$. By contracting mapping theorem, $\lambda_d^{(m)}$, which is derived by applying the operator $V(\cdot)$ on $\lambda_d^{(0)}$ for m times, must converge to that fixed point. ■

EC.9. Proof of Proposition 2

Proof. The positive recurrence and irreducibility of $X^W(\cdot)$ follow from the finite buffer capacity of Queue W. To prove positive recurrence of $X^A(\cdot)$, we note that $X^A(\cdot)$ is stochastically larger when r is larger. So it suffices to prove positive recurrence of $\{X_A(t)\}$ when $r = 1$. In that case, X^A is exactly the queue-length process in a traditional appointment system (without the open-access queue) except that each slot has a length of $1/R$ instead of 1. The positive recurrence of X^A thus follows from Proposition 1. $X_A(\cdot)$ is clearly irreducible because all states communicate with the state 0. The existence of a (periodic) steady state distribution thus follows from positive recurrence, and the uniqueness follows from irreducibility. ■

EC.10. Proof of Theorem 3

Proof. To proceed with the proof, we consider the function V^r defined in (37) as a function of both λ_d and w, and express it as $V^r(\lambda_d, w, \pi^{A,*}(\lambda_d, w), \pi^{W,*}(\lambda_d, w))$. Then by taking the full derivative with respect to w at both sides of Equation (37), we get

$$\begin{aligned} 0 &\equiv \tfrac{dV^r}{dw} \\ &= \left[\textstyle\sum_i \frac{\partial V^r}{\partial \pi_i^{A,*}}\frac{\partial \pi_i^{A,*}}{\partial w} + \frac{\partial V^r}{\partial \pi_K^{W,*}}\frac{\partial \pi_K^{W,*}}{\partial w} + \frac{\partial V^r}{\partial w}\right] + \left[\textstyle\sum_i \frac{\partial V}{\partial \pi_i^{A,*}}\frac{\partial \pi_i^{A,*}}{\partial \lambda_d} + \frac{\partial V^r}{\partial \pi_K^{W,*}}\frac{\partial \pi_K^{W,*}}{\partial \lambda_W} + \frac{\partial V}{\partial \lambda_d}\right]\frac{\partial \lambda_d}{\partial w} \\ &=: \Xi_1^r(\lambda_d, w) + \Xi_2^r(\lambda_d, w)\lambda_d'(w). \end{aligned} \tag{EC.55}$$

where $\Xi_1^r(\lambda_d, w)$ and $\Xi_2^r(\lambda_d, w)$ are the partial derivatives of V^r with respect to w and λ_d (upon keeping the other variable constant when taking each derivative), respectively. We have shown in the proof of Lemma 6 that the function V^r is strictly decreasing in λ_d for all fixed values of

$w \in [0,1]$. Therefore, we have $\Xi_2^r(\lambda_d, w) < 0$ for all $w, \lambda_d, R \in [0,1]$, $\lambda_d'(w) = -\Xi_1^r(\lambda_d, w)/\Xi_2^r(\lambda_d, w)$ must share the same sign with $\Xi_1^r(\lambda_d, w)$, i.e., either both are positive or both are negative. So it suffices to analyze the sign of function $\Xi_1^r(\lambda_d, w)$. Similar to Equation (19), we express $\Xi_1^r(\lambda_d, w)$ as follows

$$\begin{aligned}\Xi_1^r(\lambda_d, w) &= \textstyle\sum_i \frac{\partial V^r}{\partial \pi_i^{A,*}} \frac{\partial \pi_i^{A,*}}{\partial w} + \frac{\partial V^r}{\partial \pi_K^{W,*}} \frac{\partial \pi_K^{W,*}}{\partial w} + \frac{\partial V^r}{\partial w} \\ &= -r \textstyle\sum_i (\lambda_n + G(w)\lambda_d) b(i) \frac{\partial \pi_i^{A,*}}{\partial w} - rwf(w)\lambda_d \sum_i \pi_i^{A,*} b(i) \\ &\quad -(1-r)\frac{\partial \pi_K^{W,*}}{\partial w}(\lambda_n + G(w)\lambda_d) - (1-r)\pi_K^{W,*} f(w) w \lambda_d.\end{aligned} \tag{EC.56}$$

Next we determine the sign of the right-hand-side (RHS) of equation (EC.56) by discussing the two extreme cases, namely $R \to 0$ and $R \to 1$.

When $R \to 0$, Queue A becomes extremely crowded. As a result, the change to the arrival rate equals the change to the total balking rate because all extra arrivals have to balk with probability one. That implies, $\partial(\sum_i r(\lambda_n + G(w)\lambda_d) b(i) \pi_i^{A,*})/\partial \lambda_A = \sum_i r(\lambda_n + G(w)\lambda_d) b(i) (\partial \pi_i^{A,*}/\partial \lambda_A) \to 1$, and thus

$$\begin{aligned}\textstyle\sum_i r(\lambda_n + G(w)\lambda_d) b(i) \frac{\partial \pi_i^{A,*}}{\partial w} &= \textstyle\sum_i r(\lambda_n + G(w)\lambda_d) b(i) \frac{\partial \pi_i^{A,*}}{\partial \lambda_A} \frac{\partial \lambda_A(w,\lambda_d)}{\partial w} \to \frac{\partial \lambda_A(w,\lambda_d)}{\partial w} \\ &= -rwf(w)\lambda_d \textstyle\sum_i \pi_i^{A,*} b(i) - (1-r) w f(w) \lambda_d.\end{aligned}$$

As a result, the RHS of Equation (EC.56) can be expressed as

$$\begin{aligned}&rwf(w)\lambda_d \textstyle\sum_i \pi_i^{A,*} b(i) + (1-r) w f(w) \lambda_d - rwf(w)\lambda_d \sum_i \pi_i^{A,*} b(i) \\ &-(1-r)\frac{\partial \pi_K^{W,*}}{\partial w}(\lambda_n + G(w)\lambda_d) - (1-r)\pi_K^{W,*} f(w) w \lambda_d \\ &= (1-r) w f(w) \lambda_d - (1-r)\frac{\partial \pi_K^{W,*}}{\partial w}(\lambda_n + G(w)\lambda_d) - (1-r)\pi_K^{W,*} f(w) w \lambda_d.\end{aligned} \tag{EC.57}$$

We apply the above logic to Queue W. Since not all patients in Queue W will balk, we have $(\lambda_n + G(w)\lambda_d)(\partial \pi_K^{W,*}/\partial \lambda_W) \le 1$. So

$$(1-r)(\lambda_n + G(w)\lambda_d)\frac{\partial \pi_K^{W,*}}{\partial w} = (1-r)(\lambda_n + G(w)\lambda_d)\frac{\partial \pi_K^{W,*}}{\partial \lambda_W}\frac{\partial \lambda_W}{\partial w} \le \frac{\partial \lambda_W}{\partial w} = (1-r)\pi_K^{W,*} f(w) w \lambda_d. \tag{EC.58}$$

We then can bound (EC.57) as

$$\text{(EC.57)} \ge (1-r) f(w) w \lambda_d - 2(1-r)\pi_K^{W,*} f(w) w \lambda_d. \tag{EC.59}$$

We want to show the above quantity is positive, or equivalently, $\pi_K^{W,*} < 1/2$. To show that, notice that the utilization of Queue W is given by $(\lambda_n + \lambda_d G(w))(1 - \pi_K^{W,*})/(1-R)$, which implies

$$\pi_K^{W,*} \le \frac{(1-\pi_K^{W,*})(1-r)(\lambda_n + \lambda_d G(w))}{1-R} \to (1-\pi_K^{W,*})(1-r)(\lambda_n + \lambda_d G(w)), \tag{EC.60}$$

Given $(1-r)\lambda_n/(1-\overline{p}) \le 1$, we have $(1-r)(\lambda_n + \lambda_d G(w)) \le (1-r)\lambda_n/(1-\overline{p}) \le 1$. This inequality and (EC.60) imply that $\pi_K^{W,*} \le 1 - \pi_K^{W,*}$. So $\pi_K^{W,*} \le 1/2$ and (EC.59) is positive for all w. Recall

that this quantity shares the same sign with $\lambda_d'(w)$. So $\lambda_d(w)$ is strictly increasing over $[0,1]$ and thus $w^C = 1$.

When $R \to 1$, still we look into the right-hand-side of Equation (EC.56). We want to derive an alternative expression for the first term $-r\sum_i(\lambda_n + G(w)\lambda_d)b(i)\frac{\partial \pi_i^{A,*}}{\partial w}$ using the handy expression (20) in Lemma 4. However, to adapt to the setting in Lemma 4, we need to scale the time horizon in Queue A by R so that its equivalent service time is 1 slot to keep consistency with the setting of Lemma 4. Then the virtual arrival rate in Queue A is actually λ_A/R. Therefore, the term $\frac{\partial \lambda_v(w,\lambda_d)}{\partial w}$ in Equation (20) should be expressed as

$$\frac{1}{R}\frac{\partial \lambda_A(w,\lambda_d)}{\partial w} = -\frac{rwf(w)\lambda_d}{R}\sum_i \pi_i^{A,*}b(i) - \frac{(1-r)f(w)w\lambda_d}{R}, \tag{EC.61}$$

and the term $\lambda_n + G(w)\lambda_d$ in Equation (20) should be replaced by $(\lambda_n + G(w)\lambda_d)/R$.

We thus derive the following alternative expression for the (EC.56)

$$\begin{aligned}
\text{(EC.56)} \;=\; & (rwf(w)\lambda_d\textstyle\sum_i \pi_i^{A,*}b(i) + (1-r)wf(w)\lambda_d)(\int_0^1 \pi_i^{A,*}(s)q(i,s)ds) - rwf(w)\lambda_d\textstyle\sum_i \pi_i^{A,*}b(i) \\
& -(1-r)\tfrac{\partial \pi_K^{W,*}}{\partial w}(\lambda_n + G(w)\lambda_d) - \pi_K^{W,*}(1-r)f(w)w\lambda_d \\
\le\; & (1-r)wf(w)\lambda_d(\textstyle\int_0^1 \pi_i^{A,*}(s)q(i,s)ds) - \pi_K^{W,*}(1-r)f(w)w\lambda_d \\
\to\; & (1-r)wf(w)\lambda_d(\textstyle\int_0^1 \pi_i^{A,*}(s)q(i,s)ds) - (1-r)f(w)w\lambda_d < 0,
\end{aligned} \tag{EC.62}$$

where the first inequality follows from $q(i,s) \le 1$ and $\frac{\partial \pi_K^{W,*}}{\partial w} > 0$. The convergence follows from $\pi_K^{W,*} \to 1$ when $R \to 1$.■