Rational Abandonment from Priority Queues: Equilibrium Strategy and Pricing Implications

Philipp Afèche
Joseph L. Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada, afeche@rotman.utoronto.ca

Vahid Sarhangian
Graduate School of Business, Columbia University, New York, New York 10027, vs2573@columbia.edu

Observable priority queues are prevalent in practice and create incentives for utility-maximizing customers to abandon after joining. However, the literature has so far ignored this behavior and the resulting system control issues. This paper studies the rational abandonment behavior of utility-maximizing customers in the context of an observable two-class priority queue, and identifies novel implications. We characterize the equilibrium abandonment strategy of low-priority customers and show that it has a threshold structure that depends on the fee structure. We then consider pricing as a means to control the balking and abandonment behavior, both under welfare maximization and revenue maximization. Our pricing results highlight the importance of the *timing* of payments. We show that welfare-maximization requires charging only a service fee and no entrance fee. In contrast, revenue maximization generally requires a combination of both an entrance and a service fee. This two-fee structure is equivalent to charging only upon entrance but offering a partial cancellation refund. Moreover, charging only an entrance fee may generate more or less revenue than charging only a service fee, but the performance of the latter policy is more robust. This appears to be the first paper that (i) gives an analytical characterization of equilibrium abandonment behavior in observable priority queues, and (ii) studies pricing for any queueing system in presence of rational customer abandonment.

Key words: service operations; rational abandonment; pricing; priority queues; optimal stopping

1. Introduction

Priority queues are prevalent in the delivery of congestion-prone services, ranging from hospital emergency departments to cultural institutions and amusement parks. Such priority systems create a natural incentive for low-priority customers to abandon the queue after joining. Specifically, in observable priority queues, the focus of this paper, low-priority customers may abandon in response to observing "too many" higher-priority customers arrive and overtake them while they wait for service. For example, Batt et al. (2015) report this phenomenon in an empirical study of the factors affecting abandonment from a hospital emergency department, where upon arrival, patients are assigned a priority level based on the severity of their conditions. The authors find

that the observable aspect of the queue affects the abandonment behavior of patients, in particular, observing additional arrivals of sicker (and higher-priority) patients increases the abandonment rate of lower-priority patients. Priority queues are also common in the entertainment industry. For example, the Van Gogh Museum in Amsterdam prioritizes visitors based on the type of ticket they hold. Customers who paid for admission in advance of their visit (holders of e-tickets, nationwide museum cards, or discount tickets purchased through a tour) get priority admission over customers who intend to purchase their ticket on site, at the time of their visit. Therefore, these low-priority customers have an incentive to abandon if they observe "too many" high-priority customers arrive at the queue. Similarly, visitors to The London Eye (a Ferris wheel in London) can purchase a Fast Track ticket online and avoid the long queue at the location by entering a separate line. Again, ordinary customers who observe the arrival of Fast Track customers may decide to abandon the system if they feel that the wait is going to be longer than they expected when they joined the queue.

For services with price-sensitive demand, such as museums and tourist attractions (but perhaps not emergency departments), this abandonment behavior may also be affected by the pricing policy. Interestingly, pricing policies in practice differ in terms of the *timing* of charges. In some cases, including the Van Gogh Museum, low-priority customers pay the admission fee *after* waiting in line and before entering the museum, whereas in others, including the London Eye, they must purchase their ticket *before* waiting in line.

These priority systems point to two fundamental questions that have remained unanswered so far: (1) How do rational customers make abandonment decisions in observable priority queues? (2) Given this abandonment behavior, how should a service provider structure its pricing policy to maximize its revenue or the system welfare?

We answer these questions for an observable $M/M/1$ queue with two priority classes. We assume that customers are forward-looking upon arrival and make join/balk/abandon decisions to maximize their expected utility, which is the difference of expected service reward minus linear delay cost minus possible fees for entrance (joining the queue) and/or service delivery. We characterize customers' equilibrium join/balk/abandon strategy and identify its implications for the service provider's pricing decisions. Our results make the following contributions to the analysis and optimization of queueing systems with strategic customers:

1. *Rational abandonment in observable priority queues.* Whereas high-priority customers have no incentive to abandon, we show that the equilibrium join/balk/abandon strategy of low-priority customers has a threshold structure that depends on the fee structure. We provide an explicit, recursive characterization of the equilibrium thresholds and the customer utility. We develop these results by solving inter-related optimal stopping problems.

2. *Pricing in the presence of rational abandonment.* This paper is the first to study pricing for any queueing model that accounts for customers' rational abandonment decisions. We study pricing as a means to control the balking and abandonment behavior of low-priority customers. This analysis yields pricing guidelines that demonstrate, both analytically and numerically, the importance of the timing of payments for low-priority customers: (i) Welfare maximization requires charging only a service fee, and no entrance fee. This pricing policy yields equal balking and abandonment thresholds. (ii) In contrast, revenue maximization generally requires charging not only a service fee, but also an entrance fee. Under this policy the balking threshold is smaller than the abandonment threshold. This policy corresponds to charging only for entrance but offering a partial refund for order cancellation. (iii) The revenue-maximizing single-fee policy may be to charge only for entrance or only for service, depending on whether the high-priority utilization is below or above a threshold.

More broadly, our results imply that priority pricing policies based on demand models that mistakenly ignore customer abandonment, may potentially reduce system performance significantly. As such this paper stimulate more research, both on further models of rational abandonment in priority queues, and on pricing and operational controls of such systems. As detailed in Section 2, the literature on priority queues has so far focused on exogenous abandonment models or outright ignored customers' abandonment behavior.

The rest of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the model. Section 4 characterizes the equilibrium behavior of low-priority customers in the absence of pricing. As a building block for the pricing analysis, Section 5 characterizes how the customers' equilibrium behavior and the system's steady-state performance measures depend on the fee structure. Section 6 discusses pricing for welfare maximization, and Section 7 discusses pricing for revenue maximization. Section 8 offers concluding remarks. All proofs are in the Appendix.

2. Related Literature

Our study relates to several branches of the literature, briefly summarized below.

Modeling customer abandonment through exogenous deadlines. Going back to Barrer (1957), the classical approach in modeling customer abandonment in queueing systems is to assume that customers arrive at the system with i.i.d. patience thresholds, and abandon once their waiting time exceeds their threshold. There are numerous examples in the performance evaluation and control literature, especially with applications to design and control of contact centers. See for instance Bachelli and Hebuterne (1981), Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Baron and Milner (2009), Bassamboo and Randhawa (2010), Down et al. (2011), Ward (2011), Bassamboo and Randhawa (2015) for examples of single-class queues and Whitt (2006), Iravani

and Balcıoğlu (2008), Atar et al. (2010), Jennings and Reed (2012), Sarhangian and Balcıoğlu (2013), Jouini and Roubos (2013) for multi-class queues.

Pricing is generally ignored in this stream of literature. One exception is Lee and Ward (2014) who study the joint pricing and capacity decision for a $GI/GI/1+GI$ queue using a diffusion approximation. In contrast to this paper, they model abandonment through exogenous deadlines and the system arrival rate as independent of delay; as a result, the timing of charges is irrelevant in their model - they consider only a service fee, which is equivalent to a fully refundable entrance fee.

Models of rational abandonment. A smaller set of papers endogenize the abandonment behavior of customers. However, they consider service disciplines that differ from the priority policy considered in this paper, and do not consider pricing (see Hassin and Haviv 2003, Chapter 5).

In the unobservable setting, i.e., when the queue is hidden from the customers such as in contact centers, this is usually done by assuming a nonlinear waiting cost or service reward; see, e.g., Hassin and Haviv (1995), Haviv and Ritov (2001), Shimkin and Mandelbaum (2004). Mandelbaum and Shimkin (2000) assume that the reward is constant and the waiting cost is linear, but with a certain probability and without knowing so, some customers may enter a "fault position" and never get served. Akşin et al. (2013) and Ata et al. (2015) propose a discrete-time dynamic abandonment model. A customer's utility is comprised of a waiting cost, service reward, and exogenous random shocks. In each period, after observing the realizations of the random shocks, a customer decides whether to abandon or stay in the system given that she knows the probability of getting served in that period. The random shocks play a key role in this model in that customers have no incentive to abandon without these shocks. Akşin et al. (2013) use a structural estimation approach to estimate callers' parameters using call center data and implement a simulation-based approach to compute the equilibrium abandonment distribution. Ata et al. (2015) prove the existence and uniqueness of the equilibrium abandonment distribution by studying the corresponding virtual waiting time distribution, using exact analysis for a system with a single customer class, and approximate analysis in the heavy traffic regime for a multi-class system.

There are also a few papers that study the rational abandonment of customers in the observable setting. For example, Assaf and Haviv (1990) study equilibrium abandonment strategies in a processor sharing queue and Hassin (1985) considers a Last-Come, First-Served (LCFS) preemptive-resume discipline. Maglaras et al. (2015) present a model of customer abandonment for an observable First-Come, First-Served (FCFS) queue assuming that customers do not know the service rate but learn it through observing the service times of other customers while they wait.

Empirical studies of the factors driving the abandonment behavior in a hospital emergency department include Batt et al. (2015), and in a parallel study Bolandifar et al. (2014). Both papers

find that beside waiting time, the number of patients in the system and the service rate are also significant drivers of abandonment, and they discuss the modeling implications of their findings. Yu et al. (2015) empirically study for a medium sized call center the impact of delay announcements on customer behavior, including abandonment.

Pricing for queues with rational customers. There is an extensive literature on pricing for queues with rational customers, cf. Hassin and Haviv (2003) for an overview. Nevertheless, to the best of our knowledge, this paper is the first to investigate pricing for such systems in the presence of rational customer abandonment. Studies of optimal pricing and scheduling decisions for priority queues consider customers' rational join/balk and class choice decisions but preclude abandonment. For examples in unobservable queues we refer to Mandelson and Whang (1990) for welfare maximization, and Afèche (2004, 2013), Ata and Olsen (2013), Maglaras et al. (2015b), and Nazerzadeh and Randhawa (2015) for revenue maximization. For observable queues, Adiri and Yechiali (1974) and Hassin and Havi (1997) study the case of a fixed number of priority classes where upon arrival a customer can decide whether to purchase any of the priority levels or balk from the system, and Alperstein (1988) characterizes the revenue-maximizing number of priority classes. Guo and Zhang (2013) study alternative information scenarios for a system with congestion-based staffing and pricing for faster service. In contrast to the existing priority pricing studies, and as discussed in Section 3, we assume that customers arrive at the system with preassigned priority levels but are strategic with respect to their join/balk/abandon decisions.

Expulsion/termination control. When considered from the perspective of the social planner, our problem relates to the literature on expulsion and termination control in queueing systems, where customers can be removed from the system by the central planner after joining. Xu and Shanthikumar (1993) study the optimal admission control policy for a FCFS $M/M/m$ system with a single customer class and identical servers by considering the corresponding optimal expulsion control policy for the dual system where all customers are admitted but served with the preemptive-resume LCFS discipline. Xu (1994) applies the same duality approach to obtain the optimal admission and scheduling policy of an $M/M/2$ queue with nonidentical servers, and Righter (2000) extends the analysis to multiple classes of customers. Brouns and van der Wal (2006) consider admission and termination control for a preemptive priority queue with two customer classes. They show that the optimal policy for both decisions has a threshold structure. Our welfare maximization analysis (see Section 6) differs from that of Brouns and van der Wal (2006), in that, (a) we focus on controlling the low-priority segment, (b) our results relate the optimal admission and expulsion thresholds to each other (theirs do not), and (c) we also investigate pricing as a means to achieve the socially optimal policy.

Pricing for service cancellation without queueing. Finally, our work also relates to a few papers in the marketing literature that study service cancellation without queueing considerations. Xie and Gerstner (2007) consider services with capacity constraints and explore the benefits of allowing customers to cancel their tickets in return for a refund. They show that the provider can profit by offering refunds to canceling customers and reselling the capacity to new customers. Guo (2009) investigates the same problem in a competitive setting with multiple providers. In our setting, customer cancellation corresponds to abandoning the queue and, the revenue-maximizing two-fee structure for entrance and service is equivalent to charging only an entrance fee, and offering a partial refund for cancellation.

3. The Model

We consider an observable $M/M/1$ queue with two customer types; high- and low-priority. Customers are served according to a preemptive priority discipline in favor of the high-priority type. Within each priority class the service discipline is FCFS. We assume that the firm can distinguish between low- and high-priority customers. This implies that customers do not get to choose their priority class. This holds in settings where the provider assigns priority levels to customers based on attributes that she can readily observe (e.g., business versus residential customers) or determine (e.g., based on triage in the emergency department). This assumption also makes sense in cases where customers can purchase the high-priority level only well in advance of observing the system, whereas the low-priority purchase is only available on site, at the time of the visit. For example, this applies both to the Van Gogh Museum in Amsterdam and to the London Eye.

The service rate is μ for both customer types. Low-priority (high-priority) customers arrive to the system with rate λ_l (λ_h), incur a delay cost c_l (c_h) per unit of time in the system (including service) and have service valuation R_l (R_h). We assume that $c_l < R_l\mu$ and $c_h < R_h\mu$, which implies that a customer of either type prefers receiving service to balking or abandoning if there is no other customer ahead of her in the system. All parameters are common knowledge. Customers can observe the number and priority level of customers in the system.

The arrival process to the system is exogenous. However, once they arrive to the system, customers are forward-looking and maximize their expected utility with respect to their join/balk/abandonment decisions. Upon arriving to and observing the state of the system, each customer decides whether to join or balk. A joining customer may decide to later abandon the queue if that maximizes her expected utility. In other words, each customer, starting from the time of her arrival to the system until the end of service, has the *option* of abandoning the queue. Accordingly, abandoning the queue upon arrival corresponds to balking. We assume that customers join (stay in) the system if they are indifferent between joining and balking (staying and abandoning).

4. Equilibrium Behavior of Low-Priority Customers in the Absence of Pricing

From the viewpoint of high-priority customers the system operates as a FCFS queue with only high-priority customers. Therefore, high-priority customers who join the system have no incentive to later abandon, and their join/balk strategy is readily available from Naor (1969). A high-priority customer joins the system if and only if the number of high-priority customers (including herself) does not exceed $\bar{n}_h \equiv \lfloor R_h \mu / c_h \rfloor$.

In contrast, the waiting time of a low-priority arrival depends both on the number of customers she finds in system upon arrival, and on future arrivals of high-priority customers. Therefore, a low-priority customer has an incentive to abandon if "too many" high-priority customers arrive during her sojourn time. Further, the waiting time of a low-priority customer is not affected by future low-priority arrivals, but it depends on the stay/abandon decisions of low-priority customers in front of her. In equilibrium, we require each customer to take actions that maximize her expected utility, given that all other customers also behave as rational utility maximizers.

Since the low-priority customers are homogeneous and the system is Markovian, we focus without loss of generality on symmetric, Markovian, and stationary strategies that depend only on the number of low- and high-priority customers in the system. To specify the equilibrium strategy of low-priority customers, we need the following definition.

Definition 1. We say a low-priority customer is in *position* (m, n) if m is her service order among low-priority customers and n is her service order among all customers. As long as the customer is in the system we have $1 \leq m \leq n \leq \bar{n}_h + m$. Upon finishing service the customer moves to position $(0, 0)$. We also refer to m as her *low-priority position* and to n as her *system position*.

4.1. Structure of the Equilibrium Strategy

Let the value function $v(m, n)$ denote the maximum expected utility of the customer at position (m, n) under the equilibrium strategy. As detailed in the Appendix, these value functions are obtained as the solutions of inter-related optimal stopping problems: Specifically, the value function of the first low-priority customer, $v(1, n)$, is the solution of her stopping problem that only accounts for the behavior of high-priority customers. The value function of the customer in low-priority position $m > 1$ is the solution of her stopping problem that accounts both for the behavior of high-priority customers and for the behavior of low-priority customers in positions $i < m$. The following result characterizes the equilibrium strategy for low-priority customers.

Proposition 1. *There is a unique equilibrium join/balk/abandon strategy for low-priority customers which has the following threshold structure.*

1. A low-priority customer joins/stays in the system if and only if her position (m,n) satisfies $m \leq L < \infty$ and $n \leq \bar{n}(m)$, where the thresholds L and $\{\bar{n}(1),\ldots,\bar{n}(L)\}$ are functions of the problem parameters.

2. There is a threshold K satisfying $0 \leq K < L$ such that (i) $\bar{n}(m) = \bar{n}_h + m$ for $m \in \{1,...,K\}$ and (ii) $\bar{n}(m) = L < \bar{n}_h + m$ for $m \in \{K+1,...L\}$. Furthermore, if $K > 0$, then $L = \bar{n}_h + K$.

The first part of Proposition 1 implies that there exists a set of positions $\mathcal{S} \equiv \{(m,n); m \leq L, n \leq \bar{n}(m)\}$ that characterizes the equilibrium behavior of low-priority customers as follows: A customer balks from the system if the position that she would occupy upon joining the system does not belong to the set $\mathcal{S}$. Otherwise, the customer joins the system and abandons if her position falls outside of the set $\mathcal{S}$ due to new high-priority arrivals.

The second part of Proposition 1 characterizes the properties of the thresholds $\bar{n}(m)$ that determine the set $\mathcal{S}$. Recall that $\bar{n}_h$ is the maximum number of high-priority customers in the system. Therefore, for low-priority positions $m \leq K$, property (i), $\bar{n}(m) = \bar{n}_h + m$, means that a low-priority customer does not abandon as long as she is in low-priority position m. In this case, low-priority customers ahead of her do not abandon either, that is, $\bar{n}(i) = \bar{n}_h + i$ for $i < m$. However, for low-priority positions $m > K$, property (ii), $\bar{n}(m) < \bar{n}_h + m$, means that a customer at such positions may abandon, because her position can fall out of the set $\mathcal{S}$ if enough high-priority customers arrive during her sojourn time (so that $n = \bar{n}(m) + 1$). In this case, low-priority customers behind her abandon from the same system position, that is, $\bar{n}(i) = \bar{n}(m) = L$ for $i \geq m$. It follows that the next customer to abandon (if any) is always the customer at the end of the line. In other words, the customers abandon in a last-come, first-abandon order.

Large high-priority balking threshold. Proposition 1 yields the following corollary for the case where the high-priority balking threshold $\bar{n}_h$ is sufficiently large.

Corollary 1. *If $\bar{n}_h \geq R_l\mu/c_l$ then the equilibrium join/balk/abandon strategy of a low-priority customer only depends on her system position: There is a single finite threshold $\bar{n}$ such that $\bar{n} = L = \bar{n}(1) = \bar{n}(2) = \cdots = \bar{n}(L)$.*

The result is intuitive. If $\bar{n}_h \geq R_l\mu/c_l$ then a customer in system position $\bar{n}_h + 1$ would get strictly negative utility even in a FCFS system. That is, the high-priority balking threshold $\bar{n}_h$ is so large that all low-priority customers abandon before the number of high-priority customers in the system reaches this threshold. (The condition $\bar{n}_h \geq R_l\mu/c_l$ is sufficient; a necessary condition can be computed using Proposition 2.) As a result, the expected number of high-priority customers that will overtake a low-priority customer only depends on the number, but not the types, of customers ahead of her, so that she abandons once that number exceeds a certain threshold.

4.2. Explicit Characterization of the Equilibrium Thresholds and Customer Utility

Given the structure of the equilibrium strategy we can compute the threshold values and the maximal expected utility of a customer at a given position, namely, the value function $v(m,n)$. We express the value function in terms of the following performance metrics of the birth-death processes associated with service completions and high-priority arrivals.

Define the birth-death processes $N(t)$ and $\overline{N}(t)$ on $\{0,1,\cdots,\mathcal{J}\}$. The processes differ only in their death rates in state $\mathcal{J}$: For both processes, $\lambda_0=\mu_0=0$, $\lambda_i=\lambda_h$ and $\mu_i=\mu$ for $i\in\{1,\cdots,\mathcal{J}-1\}$, whereas $\mu_{\mathcal{J}}=0$ for $N(t)$ and $\mu_{\mathcal{J}}=\mu$ for $\overline{N}(t)$. That is, 0 is an absorbing state in both processes, whereas $\mathcal{J}$ is absorbing in $N(t)$ but reflecting in $\overline{N}(t)$, corresponding to the cases where a given low-priority customer abandons and remains in the system, respectively.

Let $q(i,\mathcal{J})$ denote the probability that the process $N(t)$ reaches state 0 before $\mathcal{J}$ given the starting point i. Also denote by $w(i,\mathcal{J})$ the expected first passage time of the process $N(t)$ to state 0 or $\mathcal{J}$, and by $\bar{w}(i,\mathcal{J})$ the expected first passage time of the process $\overline{N}(t)$ to state 0, both given the starting point i. All metrics can be obtained by considering the embedded random walks associated with these processes. Accordingly, $q(i,\mathcal{J})$ and $w(i,\mathcal{J})$ can be viewed as the "ruin probability" and "expected length of the game" in a Gambler's Ruin problem (see, e.g., Rosenthal 2010 page 75), and $\bar{w}(i,\mathcal{J})$ can be viewed as the expected length of the busy period of an $M/M/1/\mathcal{J}$ queue initiated with i customers. Let $\rho_h\equiv\lambda_h/\mu$, we have

$$q(i,\mathcal{J})\equiv\frac{1-\rho_h^{\mathcal{J}-i}}{1-\rho_h^{\mathcal{J}}},\tag{1}$$

$$w(i,\mathcal{J})\equiv\frac{i(1-\rho_h^{\mathcal{J}})-\mathcal{J}\left(\rho_h^{\mathcal{J}-i}-\rho_h^{\mathcal{J}}\right)}{(\mu-\lambda_h)\left(1-\rho_h^{\mathcal{J}}\right)},\tag{2}$$

$$\bar{w}(i,\mathcal{J})\equiv\frac{i(\mu-\lambda_h)+\lambda_h\left(\rho_h^{\mathcal{J}}-\rho_h^{\mathcal{J}-i}\right)}{(\mu-\lambda_h)^2}.\tag{3}$$

Proposition 2 specifies the value function $v(m,n)$ for each low-priority position m in a recursive fashion, by relating it to the value of moving ahead by one low-priority position with no high-priority customers in the system, i.e., $v(m-1,m-1)$.

Proposition 2. *Let*

$$u(i,h,\mathcal{R})=\begin{cases}\mathcal{R}q(i,h+1)-c_l w(i,h+1), & h<\bar{n}_h+1,\\ \mathcal{R}-c_l\bar{w}(i,h), & h=\bar{n}_h+1,\end{cases}\tag{4}$$

for $i\le h$ with q,w and $\bar{w}$ defined in equations (1),(2) *and* (3)*, respectively. Then, letting $v(0,0)=R_l$, the equilibrium threshold and the value function for $1\le m\le L$ are given by the recursive relations*

$$\bar{n}(m)=(m-1)+\max\{h\in\{1,...,\bar{n}_h+1\};u(h;h,v(m-1,m-1))\ge 0\},\tag{5}$$

$$v(m,n)=u(n-(m-1);\bar{n}(m)-(m-1),v(m-1,m-1)),\ m\le n\le\bar{n}(m),\tag{6}$$

where $L \equiv max\{m \geq 1; v(m-1, m-1) \geq c_l/\mu\}$.

From Proposition 1 we know that in equilibrium the customer at low-priority position m abandons if her system position exceeds some threshold $\bar{n}(m)$. We obtain this threshold for each low-priority position m using the function $u(i, h, \mathcal{R})$ given in (4). The function returns the expected utility of staying in system for a low-priority customer who currently faces i service completions before advancing by one low-priority position and receiving the value $\mathcal{R}$, and stays in the system as long as the number of service completions that are required before advancing by one low-priority position does not exceed h.

As shown in (4) the structure of the utility function $u(i, h, \mathcal{R})$ depends on whether the system position of the tagged low-priority customer can or cannot exceed the threshold h. In the first case, that is, $h < \overline{n}_h + 1$, the low-priority customer abandons if the number of service completions required to advance in the queue reaches $h+1$. As a result, i, the number of service completions required for this low-priority customer to advance by one position, evolves according to the birth-death process $N(t)$ with absorbing state $\mathcal{J} = h+1$. Therefore, starting in state i, the probability of advancing by one low-priority position before abandoning is $q(i, h+1)$, given by (1), and the expected reward is $\mathcal{R}q(i, h+1)$. The expected time until the customer either advances in line or abandons is $w(i, h+1)$, given by (2), so the expected delay cost is $c_l w(i, h+1)$. In the second case of (4), that is, $h = \bar{n}_h + 1$, the low-priority customer does not abandon; she eventually advances by one low-priority position and receives the reward $\mathcal{R}$. In this case, i, the number of service completions required to advance by one position, evolves according to the birth-death process $\overline{N}(t)$ with reflecting state $\mathcal{J} = h$. It follows using (3) that the expected delay cost for the customer in state i is $c_l \bar{w}(i, h)$.

For a customer in low-priority position m, the threshold $h = \bar{n}(m) - m + 1$ as she is willing to wait for $\bar{n}(m) - m$ high-priority customers plus one low-priority customer ahead of her. (Note that $1 \leq h \leq \bar{n}_h + 1$ since $m \leq \bar{n}(m) \leq \bar{n}_h + m$). For each low-priority position m, the equilibrium threshold $\bar{n}(m)$ is the maximum system position such that the expected utility of staying in system is nonnegative. Hence, for each m, the equilibrium threshold $\bar{n}(m)$ and the value function $v(m, n)$ satisfy (5) and (6). To calculate the equilibrium thresholds we start from $m = 1$, for which moving ahead by one position in queue is equivalent to completing service and hence $\mathcal{R} = v(0, 0) = R_l$. This allows us to recursively compute the value function and equilibrium thresholds for higher positions.

Large high-priority balking threshold. If the high-priority balking threshold $\bar{n}_h$ is so large that low-priority customers may abandon even if at the head of their queue, the analysis simplifies since by Proposition 2 the strategy of low-priority customers only depends on their system position. Let $u(n, \bar{n})$ denote the expected utility of a low-priority customer at system position $n \leq \bar{n}$ if all

low-priority customers follow the threshold strategy whereby they join/stay in the system if their system position does not exceed $\bar{n}$ and they balk/abandon otherwise. We have

$$u(n,\bar{n}) = R_l q(n,\bar{n}+1) - c_l w(n,\bar{n}+1), \tag{7}$$

where $q(n,\bar{n}+1)$ is the probability that a low-priority customer in system position n reaches system position 0 before $\bar{n}+1$, and $w(n,\bar{n}+1)$ is the expected time until this customer reaches system position 0 or $\bar{n}+1$, whichever happens first. As in Proposition 3, these quantities can be viewed as the ruin probability and the expected length of the game in a Gambler's Ruin problem with initial wealth n, target wealth $\bar{n}+1$, and winning probability $\lambda_h/(\mu+\lambda_h)$. We have from (1) and (2):

$$q(n,\bar{n}+1) = \frac{1-\rho_h^{\bar{n}+1-n}}{1-\rho_h^{\bar{n}+1}}, \tag{8}$$

$$w(n,\bar{n}+1) = \frac{n(1-\rho_h^{\bar{n}+1}) - (\bar{n}+1)(\rho_h^{\bar{n}+1-n} - \rho_h^{\bar{n}+1})}{(\mu-\lambda_h)(1-\rho_h^{\bar{n}+1})}. \tag{9}$$

Note that $u(0,\bar{n}) = R_l$ and $u(\bar{n}+1,\bar{n}) = 0$ for any $\bar{n}$.

The equilibrium threshold $\bar{n}$ is uniquely determined by

$$\bar{n} = \max\{n \geq 0; u(n,n) \geq 0\}, \tag{10}$$

and the value function is given by

$$v(n) = u(n,\bar{n}), \quad n \in \{0,1,...,\bar{n}\}. \tag{11}$$

Example 1: equilibrium strategy of low-priority customers. We illustrate the two possible equilibrium structures with the following two cases for a system with parameters $R_l = 10, \mu = 1, \lambda_h = 0.4$, and $c_l = 0.5$: (i) $\bar{n}_h = 10$ and (ii) $\bar{n}_h = 15$. Figure 1 illustrates the equilibrium strategies. In case (i) the equilibrium thresholds $\bar{n}(m)$ depend on the low-priority position m. In particular, $\bar{n}(m) = \bar{n}_h + m = 10 + m$ for $m = 1,2$, $\bar{n}(m) = L = 12$ for $3 \leq m \leq 12$, and $K = 2$. That is, a low-priority customer does not abandon in low-priority positions $m = 1$ and $m = 2$ as she receives a positive expected utility in all possible system positions. In case (ii), however, all low-priority customers abandon once their system position exceeds the threshold $\bar{n} = 12$.

5. Pricing under Rational Abandonment: Preliminaries

We now turn to the analysis of pricing under rational abandonment. For simplicity and to highlight the interaction between abandonment and pricing implications, we focus on controlling the low-priority segment, and we assume a fixed arrival rate $\lambda_h < \mu$ of high-priority customers who do not balk ($\bar{n}_h = \infty$). This setting applies, for example, to cases with a predetermined high-priority price and where high-priority customers do not react to the queue length when they arrive to the service

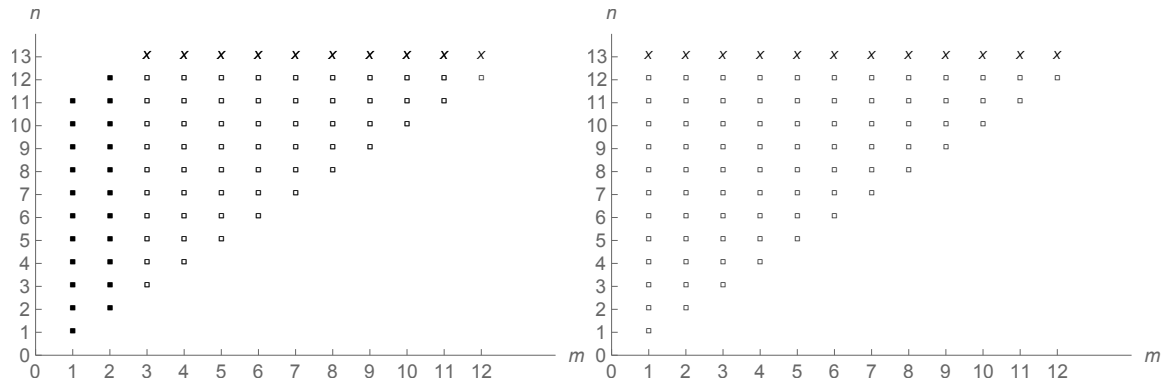

Figure 1 An illustration of the equilibrium strategy of low-priority customers in Example 1 with Case (i) on the left and Case (ii) on the right. A customer joins/stays in positions marked with □ and ■ and abandons if her position reaches a position marked with ×. Customers do not abandon once they reach the positions marked with ■.

facility. For example, the Van Gogh Museum in Amsterdam grants high priority to holders of nationwide museum cards that are purchased well before their visit. These high-priority customers are likely less sensitive to the length of the high-priority queue because it advances relatively quickly. Furthermore, our structural results on the pricing implications of rational abandonment extend to cases with high-priority balking ($\bar{n}_h < \infty$), and to the solution of the joint low- and high-priority pricing problem (where the high-priority price controls $\bar{n}_h$).

We study pricing for welfare maximization in Section 6 and for revenue maximization in Section 7. The remainder of this section presents two important preliminaries for these analyses; the equilibrium behavior of low-priority customers in the presence of pricing (Section 5.1), and steady-state performance measures (Section 5.2).

5.1. Equilibrium Behavior of Low-Priority Customers in the Presence of Pricing

In the presence of customer abandonment the service provider faces an important pricing question that has so far been ignored in the literature: How to structure the pricing policy to jointly control customers' balking and abandonment decisions? In line with the three key events that characterize each customer's path through the system, we allow for the following three fees that are potentially relevant and commonly observed in practice: An entrance fee P_e that is charged for joining the system (placing an order), a cancellation fee P_c that is charged for abandoning (canceling the order), and a service fee P_s that is charged upon service completion (order delivery). Note that these fees must satisfy $R_l - P_s > P_e$ (otherwise low-priority customers prefer balking to getting served) and $P_e + P_c \geq 0$ (otherwise low-priority customers can make money by placing and then immediately canceling an order). These fees play an important role for the equilibrium behavior

of low-priority customers. In particular, as formalized in Proposition 3 below, in presence of an entrance fee the balking and abandonment thresholds need not be, and typically are not, equal.

Let $\bar{n}_a$ and $\bar{n}_b$ denote the equilibrium balking and abandonment thresholds, respectively. Given a cancellation fee P_c and service fee P_s, let $u(n, \bar{n}_a; P_c, P_s)$ denote the expected utility of a low-priority customer of staying in the system at system position $n \leq \bar{n}_a$ if all low-priority customers follow the threshold strategy whereby they stay if their system position does not exceed $\bar{n}_a$ and they abandon otherwise. We have

$$u(n, \bar{n}_a; P_c, P_s) = (R_l - P_s)q(n, \bar{n}_a + 1) - P_c(1 - q(n, \bar{n}_a + 1)) - c_l w(n, \bar{n}_a + 1), \tag{12}$$

with q and w given in (8) and (9), respectively. In the absence of service and cancellation fees, $P_c = P_s = 0$ so that $u(n, \bar{n}_a; 0, 0) = u(n, \bar{n}_a)$, i.e., (12) specializes to (7).

Proposition 3. *Given an entrance fee $P_e \geq 0$, cancellation fee $P_c \geq 0$, and service fee $P_s \geq 0$, there is an unique join/balk/abandon equilibrium characterized by an abandonment threshold $\bar{n}_a < \infty$ and a balking threshold $\bar{n}_b \leq \bar{n}_a$, where $\bar{n}_b < \bar{n}_a$ only if $P_e > 0$. That is, a low-priority customer stays in the system if and only if her system position does not exceed*

$$\bar{n}_a(P_c, P_s) \equiv \max\{n \geq 0; u(n, n; P_c, P_s) \geq -P_c\}, \tag{13}$$

and joins the system if and only if her system position does not exceed

$$\bar{n}_b(P_e, P_c, P_s) \equiv \max\{n \geq 0; u(n, \bar{n}_a(P_c, P_s); P_c, P_s) \geq P_e\}. \tag{14}$$

We omit a formal proof of Proposition 3 (it follows from the results in Section 4) and discuss the underlying intuition. By (13) the equilibrium abandonment threshold is uniquely determined as the maximum system position at which a low-priority customer (weakly) prefers to stay in the system versus canceling her order at a cost of P_c. Note that (13) is equivalent to condition (10) with the reward R_l replaced by $R_l - (P_s - P_c)$:

$$\bar{n}_a(P_c, P_s) \equiv \max\{n \geq 0; (R_l - (P_s - P_c))q(n, n+1) - c_l w(n, n+1) \geq 0\}.$$

That is, the incentive to stay/abandon is the same, regardless of whether the cancellation fee P_c is paid upon abandoning the system, or this fee is paid upfront (when joining the system) and deducted from the service fee P_s when completing service. Therefore, in equilibrium the expected utility of low-priority customers who join the system is given by the value function

$$v(n; P_c, P_s) = u(n, \bar{n}_a(P_c, P_s); P_c, P_s), \quad n \in \{0, 1, ..., \bar{n}_a\},$$

which specializes to (11) in the absence of cancellation and service fees. Since joining the system requires an immediate payment of P_e, a low-priority customer joins at system position n iff $v(n; P_c, P_s) \geq P_e$. Therefore, by (14) the equilibrium balking threshold is uniquely determined as the maximum system position at which the customer joins. Notice that $P_e = 0$ implies equal balking and abandonment thresholds ($\bar{n}_b = \bar{n}_a$). A positive entrance fee implies $\bar{n}_b \leq \bar{n}_a$.

5.2. Steady-State System Performance Measures

We characterize three steady-state system performance measures that are required to evaluate the system welfare and the provider's revenue, the queue-length distribution (π_i), the joining probability (q_J), and the service probability (q_S). These three measures are functions of the abandonment threshold $\bar{n}_a$ and the balking threshold $\bar{n}_b$.

Queue-length Distribution. Let π_i be the steady-state probability of having i customers in the system (in this section we suppress the dependence of π_i on the thresholds $\bar{n}_a$ and $\bar{n}_b$). Define $\lambda \equiv \lambda_l + \lambda_h$, $\rho = \lambda/\mu$, and $\rho_h = \lambda_h/\mu$. Then π_i satisfies the balance equations

$$\begin{aligned} \lambda\pi_i &= \mu\pi_{i+1}, \quad i = 0, 1, ..., \bar{n}_b - 1, \\ \lambda_h\pi_i &= \mu\pi_{i+1}, \quad i = \bar{n}_b, ..., \bar{n}_a - 1. \end{aligned}$$

It follows that

$$\pi_i = \begin{cases} \pi_0\rho^i, & i = 0, 1, ..., \bar{n}_b, \\ \pi_0\rho_h^{i-\bar{n}_b}\rho^{\bar{n}_b}, & i = \bar{n}_b + 1, ..., \bar{n}_a. \end{cases} \tag{15}$$

For $i \geq \bar{n}_a + 1$ since all the customers in system are high-priority, we have

$$\pi_i = (1 - \rho_h)\rho_h^i, \quad i \geq \bar{n}_a + 1.$$

The normalization condition $\sum_{i=0}^{\infty} \pi_i = 1$ implies that

$$\pi_0 = \frac{1 - \rho_h^{\bar{n}_a+1}}{\frac{1-\rho^{\bar{n}_b+1}}{1-\rho} + \rho^{\bar{n}_b}\frac{\rho_h - \rho_h^{\bar{n}_a-\bar{n}_b+1}}{1-\rho_h}}. \tag{16}$$

Joining Probability. Let $q_J(\bar{n}_b, \bar{n}_a)$ denote the "joining probability", i.e., the probability that a low-priority arrival joins the system (does not balk) upon observing the queue, given the thresholds $\bar{n}_a$ and $\bar{n}_b$. We have

$$q_J(\bar{n}_b, \bar{n}_a) = \sum_{i=0}^{\bar{n}_b-1} \pi_i = \sum_{i=0}^{\bar{n}_b-1} \rho^i\pi_0 = \pi_0\frac{1 - \rho^{\bar{n}_b}}{1 - \rho}. \tag{17}$$

Service Probability. Let $q_S(\bar{n}_b, \bar{n}_a)$ denote the "service probability", i.e., the probability that a customer is eventually served; that is, she joins the system and does not abandon before completing service. Then

$$q_S(\bar{n}_b, \bar{n}_a) = \sum_{i=0}^{\bar{n}_b-1} \pi_i q(i+1, \bar{n}_a + 1),$$

where $q(i+1, \bar{n}_a + 1)$ is the probability that the customer joining at position $i+1$, reaches system position 0 before $\bar{n}_a + 1$. Substituting for π_i from (15) and for $q(i+1, \bar{n}_a + 1)$ from (8) yields after some algebra

$$q_S(\bar{n}_b, \bar{n}_a) = \pi_0\frac{1}{1 - {\rho_h}^{\bar{n}_a+1}}\left(\frac{1 - \rho^{\bar{n}_b}}{1 - \rho} - {\rho_h}^{\bar{n}_a}\frac{1 - \left(\frac{\rho}{\rho_h}\right)^{\bar{n}_b}}{1 - \left(\frac{\rho}{\rho_h}\right)}\right). \tag{18}$$

6. Pricing for Welfare Maximization

In this section we study the problem of maximizing the system welfare. Since we consider a fixed high-priority arrival stream and these customers are not affected by low-priority customers, this problem is equivalent to that of maximizing the surplus from low-priority customers.

6.1. Welfare Maximization Requires Equal Balking and Abandonment Thresholds

We start by considering how a welfare-maximizing service provider would control the system through balking and abandonment thresholds. Suppose the social planner accepts a new low-priority arrival up to system position $\bar{m}_b$ and keeps a low-priority customer in system up to position $\bar{m}_a \geq \bar{m}_b$. Let $S(\bar{m}_b, \bar{m}_a)$ denote the social welfare as a function of these thresholds. We have

$$S(\bar{m}_b, \bar{m}_a) = \lambda_l \sum_{i=0}^{\bar{m}_b - 1} \pi_i(\bar{m}_b, \bar{m}_a) u(i+1, \bar{m}_a), \tag{19}$$

where $\pi_i(\bar{m}_b, \bar{m}_a)$, given in (15), is the steady-state probability of having i customers in system, and $u(i+1, \bar{m}_a)$, given in (7), is the expected utility of a low-priority customer who joins at position $i+1$ and is removed if she reaches position $\bar{m}_a + 1$.

Proposition 4 establishes that such a threshold policy, with equal thresholds, is socially optimal.

Proposition 4. *The socially optimal policy has a threshold structure: Accept a new low-priority arrival up to system position $\bar{m}_b^*$ and keep a low-priority customer in system up to system position $\bar{m}_a^*$. Furthermore, the socially optimal thresholds are (i) equal and (ii) not greater than the self-optimization threshold $\bar{n}$ given in* (10)*; that is, $\bar{m}_a^* = \bar{m}_b^* \leq \bar{n}$.*

By property (i), the welfare-maximizing balking and abandonment thresholds are equal. This follows because the expected utility of a low-priority customer in a given position, and her externality on other low-priority customers, are independent of how she reached this position, whether by joining this position upon arrival, or by joining further ahead and "falling behind" later due to high-priority arrivals.

By property (ii), the welfare-maximizing threshold is smaller than or equal to the equilibrium threshold under "self-optimization", that is, if the provider does not control the queue. The intuition for this result extends the one for the FCFS case analyzed by Naor (1969) to settings with abandonment: Under self-optimization low-priority customers ignore the externality they impose on customers that join after they do, and therefore join, and remain in the system, up to a higher position.

6.2. Welfare-Maximizing Pricing Requires a Service Fee and No Entrance Fee

The pricing that induces the welfare-maximizing operation consists of a single fee that is charged upon completing service, that is, $P_e = P_c = 0$. Let P_s^* be the socially optimal service fee and $\bar{m}^* \equiv \bar{m}_a^* = \bar{m}_b^*$. Then

$$P_s^* = R_l - c_l \frac{w(\bar{m}^*, \bar{m}^* + 1)}{q(\bar{m}^*, \bar{m}^* + 1)}, \tag{20}$$

induces the equilibrium balking/abandonment threshold $\bar{m}^*$, as is evident from (12)–(13). The service fee P_s^* reflects the system externality of a customer in position $\bar{m}^*$.

This pricing result is in line with the classic FCFS analysis of Naor (1969), *except* for one important distinction: If customers have an incentive to abandon after joining the system, as in our model with priorities and unlike the FCFS case, the *timing* of payments plays a key role for queueing control. Specifically, in our setting the welfare-maximizing operation typically *cannot* be induced if the provider charges an entrance fee (upon joining the queue) instead of a service fee (upon service completion). With an entrance fee the provider can only control the balking threshold but not the abandonment threshold, and the two thresholds will typically differ, which is suboptimal by Proposition 4. In contrast, by charging only a service fee the provider can ensure equal balking and abandonment thresholds at the desired level.

7. Pricing for Revenue Maximization

In this section we study the problem of maximizing the provider's revenue. Since we consider a fixed high-priority arrival stream and these customers are not affected by low-priority customers, this problem is equivalent to that of maximizing the revenue from low-priority customers.

As shown in Section 6, welfare-maximization requires charging only a service fee, and no entrance fee. In contrast, as we show in this section, revenue-maximization typically requires *both* an entrance and a service fee. Furthermore, charging an entrance fee only may generate more or less revenue than charging a service fee only.

7.1. Problem Formulation

Consider a revenue-maximizing service provider charging three different fees; P_e for entrance, P_c for cancellation, and P_s for service. The provider solves

$$\max_{\{P_e \geq 0, P_c \geq 0, P_s \geq 0\}} \quad \lambda_l (P_e + P_c) q_J(\bar{n}_a, \bar{n}_b) + \lambda_l (P_s - P_c) q_S(\bar{n}_a, \bar{n}_b) \tag{21}$$

$$s.t. \quad \bar{n}_a = \max\{n \geq 0; u(n, n; P_c, P_s) \geq -P_c\}, \tag{22}$$

$$\bar{n}_b = \max\{n \geq 0; u(n, \bar{n}_a; P_c, P_s) \geq P_e\}. \tag{23}$$

The revenue rate in (21) has two components. The first derives from all customers who join the system, regardless of whether they get served; they each pay $P_e + P_c$, the sum of entrance plus

cancellation fees. The second derives from all customers who get served; they each pay $P_s - P_c$, the service fee net of cancellation fee. The joining and service probabilities, $q_J(\bar{n}_a, \bar{n}_b)$ and $q_S(\bar{n}_a, \bar{n}_b)$, respectively, are given in (17) and (18). The constraints (22) and (23) specify, respectively, the abandonment threshold $\bar{n}_a$ and the balking threshold $\bar{n}_b$ that are implied in equilibrium by the fee triple (P_e, P_c, P_s). Recall that $u(n, \bar{n}_a; P_c, P_s)$, defined in (12), is the expected utility of a low-priority customer of staying in the system at system position $n \leq \bar{n}_a$ if all low-priority customers stay in the system if their system position does not exceed $\overline{n}_a$ and they abandon otherwise.

We simplify and reformulate the revenue-maximization problem (21)–(23) into an equivalent optimization problem over the thresholds $\bar{n}_a$ and $\bar{n}_b$. Let $\Pi(\bar{n}_a, \bar{n}_b)$ denote the revenue rate from low-priority customers as a function of these thresholds.

Proposition 5. *The problem* (21)–(23) *is equivalent to*

$$\max_{\{\bar{n}_b \leq \bar{n}_a\}} \quad \Pi(\bar{n}_a, \bar{n}_b) \equiv \lambda_l \bar{P}_e q_J(\bar{n}_a, \bar{n}_b) + \lambda_l \bar{P}_s q_S(\bar{n}_a, \bar{n}_b) \tag{24}$$

$$s.t. \quad \bar{P}_s = R_l - c_l \frac{w(\bar{n}_a, \bar{n}_a + 1)}{q(\bar{n}_a, \bar{n}_a + 1)}, \tag{25}$$

$$\bar{P}_e = \left(R_l - \bar{P}_s\right) q(\bar{n}_b, \bar{n}_a + 1) - c_l w(\bar{n}_b, \bar{n}_a + 1), \tag{26}$$

where $\bar{P}_e \equiv P_e + P_c$ denotes the net entrance fee and $\bar{P}_s \equiv P_s - P_c$ the net service fee.

The net fees $\bar{P}_e$ and $\bar{P}_s$ fees have the following intuitive revenue equivalence interpretation alluded to in Section 5.1: Charging a cancellation fee P_c upon abandonment is equivalent to charging P_c upon entrance and then refunding it if the customer completes service. Therefore, the provider can achieve maximum revenues using only an entrance and a service fee. For simplicity, we henceforth use the terms "entrance fee" and "service fee" to refer to the amounts $\bar{P}_e$ and $\bar{P}_s$, respectively. However, in terms of the timing of the fees, the following two implementations are clearly revenue-equivalent: (i) charge $\bar{P}_e$ when the customer places her order and $\bar{P}_s$ if and when the customer gets served; or (ii) charge $\bar{P}_e + \bar{P}_s$ when the customer places her order and refund $\bar{P}_s$ if and when the customer abandons. By (25) the service fee $\bar{P}_s$ directly controls the abandonment threshold, by extracting all surplus of the "marginal staying" customer, that is, the one in the system position $\bar{n}_a$, the highest position without abandonment. By (26) the entrance fee $\bar{P}_e$ controls the balking threshold, given the abandonment threshold induced by the service fee, by extracting all surplus of the "marginal joining" customer, that is, the one joining into position $\bar{n}_b$, the highest position without balking. It follows from (25)–(26) that the balking threshold is strictly smaller than the abandonment threshold if the entrance fee $\bar{P}_e$ is positive, whereas the thresholds are equal if the entrance fee is zero.

7.2. Revenue-Maximizing Pricing Requires a Service Fee and an Entrance Fee

The provider can typically achieve a higher revenue rate by charging *both* an entrance and a service fee, because the two fees give the provider separate control over the balking and abandonment thresholds. Proposition 7 formalizes this result.

PROPOSITION 6. *It is generally not revenue-maximizing to charge only for entrance (with $\bar{P}_s = 0$), or only for service (with $\bar{P}_e = 0$).*

To pinpoint the revenue effects of charging an entrance fee on top of a service fee, rewrite the revenue function (24) as

$$\Pi(\bar{n}_a, \bar{n}_b) \equiv \lambda_l q_J(\bar{n}_a, \bar{n}_b)\Big(\bar{P}_e + \bar{P}_s \frac{q_S(\bar{n}_a, \bar{n}_b)}{q_J(\bar{n}_a, \bar{n}_b)}\Big),$$

the product of quantity (the effective arrival or joining rate $\lambda_l q_J(\bar{n}_a, \bar{n}_b)$) by the expected payment per joining customer $\bar{P}_e + \bar{P}_s q_S(\bar{n}_a, \bar{n}_b)/q_J(\bar{n}_a, \bar{n}_b)$. Without entrance fee the provider is restricted to a *fixed* relationship between quantity and expected payment in trading off these two factors, because the service fee determines both through a single threshold ($\bar{n}_a = \bar{n}_b$). The ability to charge both for entrance and for service relaxes the trade-off between quantity and expected payment, giving the provider the flexibility to optimize one factor given the other. Charging an entrance fee, in addition to a service fee, has two countervailing revenue effects, a negative quantity effect as it reduces the joining probability, and a positive price effect as it increases the expected payment per joining customer. By lowering the service fee to offset the entrance fee, the provider can further optimize this trade–off. Charging both fees is optimal whenever the positive price effect dominates the negative quantity effect. This holds unless the high-priority utilization exceeds a threshold: In this case, the low-priority joining probability is (i) relatively low even without entrance fee, which reduces the significance of the price effect, and (ii) experiences a significant percentage drop when charging for entrance, which increases the magnitude of the quantity effect.

As alluded to in Section 7.1, a natural implementation of the two-fee structure is to charge $\bar{P}_e + \bar{P}_s$ when the customer places her order and offer the partial refund $\bar{P}_s$ if and when the customer cancels her order. From this perspective, offering the partial refund generates more revenue because it allows the provider to attract more low-priority customers by reducing their abandonment risk.

The following example illustrates these effects and proves Proposition 6.

Example 2: optimal pricing versus optimal service fee. Consider a system with parameters $R_l = 60$, $\mu = 1$, $\lambda_l = 0.4$, and $c_l = 1$. We compare the optimal pricing scheme with the optimal service fee if the provider does not charge for entrance (see Section 7.3). Figure 2 shows the optimal fees under these two pricing schemes, as well as the resulting balking and abandonment thresholds, joining probabilities, and expected per-customer payments, as functions of the high-priority arrival

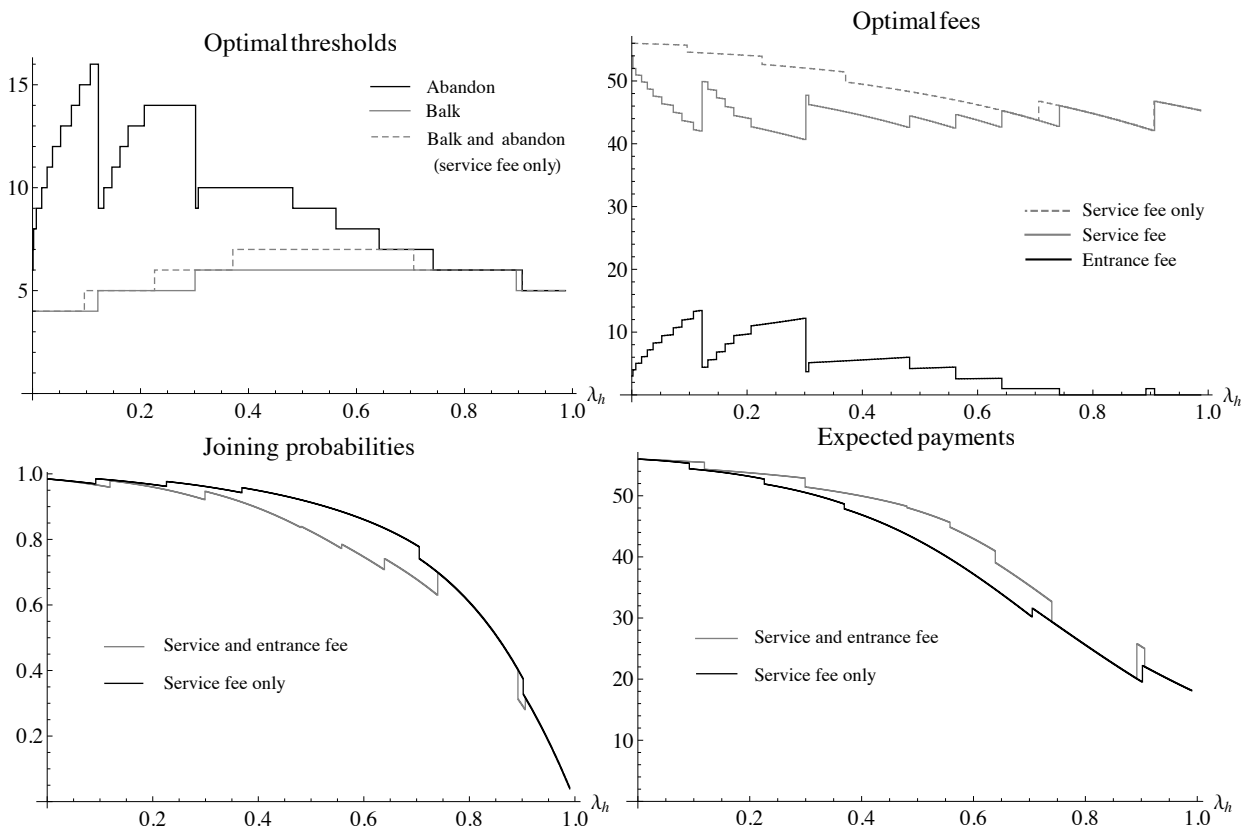

Figure 2 Comparison of optimal pricing (service and entrance fees) vs. optimal service fee only. System parameters: $R_l = 60, \mu = 1, \lambda_l = 0.4, c_l = 1$.

rate λ_h. Observe that it is optimal to charge both an entrance and a service fee as long as the high-priority utilization is not too high, below around 0.73. Furthermore, the service fee is significantly higher than the (nonrefundable) entrance fee, because the former can be viewed as a refundable entrance fee. As the high-priority utilization increases to this level, the optimal entrance fee declines to zero, and the optimal strategy is to charge a service fee only (the exception around $\lambda_h = 0.9$ is due to the discrete nature of the problem). Further, cases where it is optimal to charge both entrance and service fees result in a lower service fee and typically a lower balking threshold but a higher abandonment threshold, compared to charging only for service. As a result, customers are less likely to join the system, but their expected payment is higher because conditional on joining, they are more likely to complete service instead of abandoning.

7.3. Maximizing Revenues with a Single Fee: Service Fee or Nonrefundable Entrance Fee?

As shown in Section 7.2, it is typically optimal to charge customers both an entrance and a service fee, which is equivalent to charging only for entrance and offering a partial cancellation refund. However, charging only a single fee is practically appealing due to its simplicity. We investigate

whether the firm can generate more revenue by charging customers only a service fee or only a nonrefundable entrance fee. The answer lies in the following intuitive trade-off captured by the model: On one hand, if customers are only charged for service, they are more willing to enter the system but only those who do not abandon contribute to the firm's revenue. On the other hand, if customers are only charged a nonrefundable fee for entering the system they are more reluctant to join the system yet the firm collects revenue from all entering customers regardless of whether they receive service or not.

Service fee. Suppose the service provider charges a single fee $\bar{P}_s$ for service, that is, $\bar{P}_e = 0$. It follows from (25) and (26) that a zero entrance fee induces equal balking and abandonment thresholds. Let Π_s denote the revenue rate under a service fee only. The service provider maximizes revenues by solving

$$\begin{aligned} \max_{\{\bar{n}_b = \bar{n}_a\}} \quad & \Pi_s(\bar{n}_b, \bar{n}_a) \equiv \lambda_l \bar{P}_s q_S(\bar{n}_b, \bar{n}_a) \\ s.t. \quad & \bar{P}_s = R_l - c_l \frac{w(\bar{n}_a, \bar{n}_a + 1)}{q(\bar{n}_a, \bar{n}_a + 1)}. \end{aligned} \tag{27}$$

The service fee $\bar{P}_s$ is set such that the customer joining at system position $\bar{n}_a$ is indifferent between joining or balking.

Nonrefundable entrance fee. Assume that the service provider charges a single fee $\bar{P}_e$ upon entering the system, that is, $\bar{P}_s = 0$. In this case the firm can control only the balking threshold, whereas the abandonment threshold is independent of the entrance fee. Let Π_e denote the revenue rate under an entrance fee only. The service provider maximizes revenues by solving

$$\begin{aligned} \max_{\{\bar{n}_b \le \bar{n}_a\}} \quad & \Pi_e(\bar{n}_a, \bar{n}_b) \equiv \lambda_l \bar{P}_e q_J(\bar{n}_a, \bar{n}_b) \\ s.t. \quad & \bar{n}_a = \max\{n \ge 0; R_l q(n, n+1) - c_l w(n, n+1) \ge 0\}, \\ & \bar{P}_e = R_l q(\bar{n}_b, \bar{n}_a + 1) - c_l w(\bar{n}_b, \bar{n}_a + 1). \end{aligned} \tag{28}$$

With a positive entrance fee the balking threshold is typically smaller than the abandonment threshold.

Optimality conditions for service fee vs. nonrefundable entrance. Our numerical experiments show that, depending on the problem parameters, either pricing strategy may dominate the other. Figure 3 illustrates this fact, by showing for $c_l = \mu = 1$ and two different values of the low-priority reward R_l, the combinations of arrival rates λ_h and λ_l for which the service/entrance fee is optimal. As Figure 3 demonstrates the optimal pricing scheme does not have a monotone structure with respect to either arrival rate. This can be explained by the non-monotone behavior of the optimal prices, as well as the joining and service probabilities (see Example 3 below). The following proposition provides sufficient conditions under which either an entrance or a service fee

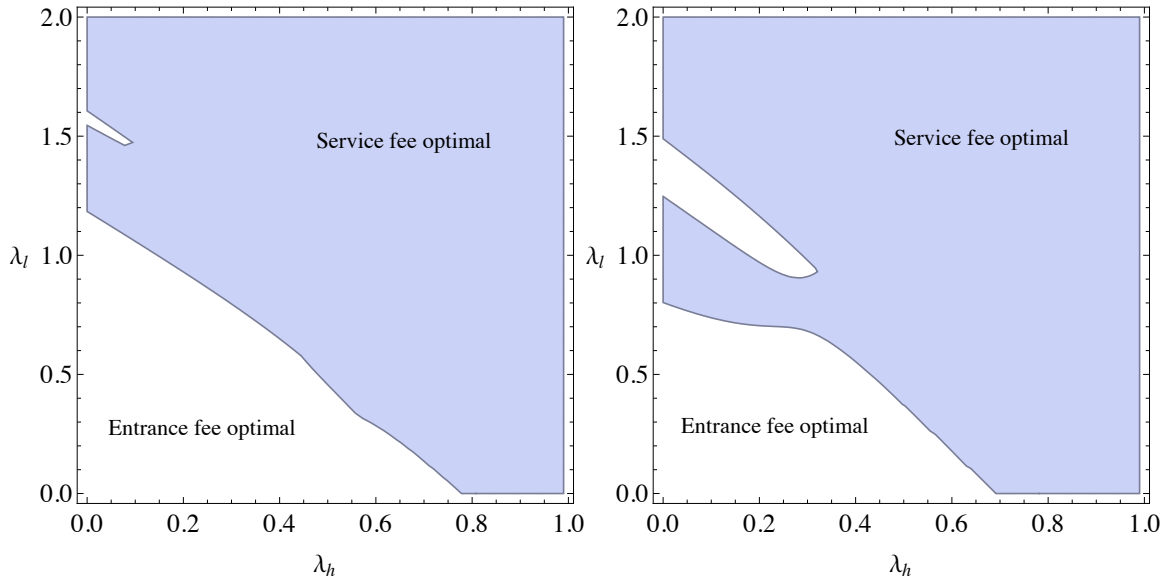

Figure 3 Regions where a service/entrance fee is optimal for $R_l = 40$ (left) and $R_l = 20$ (right) and $c_l = \mu = 1$.

is optimal. (It seems infeasible to fully characterize analytically which pricing strategy dominates as a function of the problem parameters.)

PROPOSITION 7. *Assume λ_h is strictly positive but sufficiently small, i.e., let $\lambda_h \to 0^+$.*

(i) If $R_l\mu/c_l > 2$ and $\lambda_l \to 0^+$, then the optimal nonrefundable entrance fee is more profitable than the optimal service fee.

(ii) If $R_l\mu/c_l > (\hat{n}^* + 1)^2 - 1$ where $\hat{n}^* = \left\lfloor \frac{1}{2}(\sqrt{5 + 4R_l\mu/c_l} - 1) \right\rfloor$, and $\rho_l \to 1^+$, then the optimal nonrefundable entrance fee is more profitable than the optimal service fee.

(iii) If $R_l\mu/c_l \geq 2$ and $\lambda_l \to +\infty$, then the optimal service fee is more profitable than the optimal nonrefundable entrance fee.

Proposition 7 reinforces the observation made in Sections 6.2 and 7.2 by showing that in the presence of (even a little) abandonment, the *timing* of the payment matters. Specifically, it identifies three parameter regimes where charging a nonrefundable entrance fee or a service fee dominates. Regime (i) corresponds to a very lightly loaded system with both low- and high-priority arrival rates in the neighborhood of zero. In this case, charging an entrance fee dominates if the low-priority reward exceeds the expected delay cost of two service completions. This condition is met in both examples presented in Figure 3. Our numerical experiments show that an entrance fee is also the optimal choice if the load conditions of regime (i) are somewhat relaxed, specifically, when the low- and high-priority arrival rates are not infinitesimal but still sufficiently low. Regime (ii) corresponds to a system with low-priority load approaching one, and high-priority arrival rate in the neighborhood of zero. In this case, a non-trivial condition on the system parameters is required

for the entrance fee to dominate. This condition holds for the first example (left) presented in Figure 3, but not for the second one (right). Regime (*iii*) corresponds to a system with high-priority arrival rate in the neighborhood of zero and low-priority arrival rate sufficiently high. In this case, a service fee generates more revenue than an entrance fee if $R_l\mu/c_l \geq 2$. This condition holds in both examples in Figure 3.

In general, we observe that for sufficiently high low- and high-priority loads the service fee is the optimal choice. Intuitively, under these load conditions the added revenue due to the higher joining rate when charging for service vs. for entrance dominates the revenue loss associated with abandoning customers.

Revenue gain/loss of service vs. nonrefundable entrance fee. We are also interested in the revenue loss that the provider incurs when choosing the suboptimal fee structure. Based on extensive numerical experiments, we observe that the percentage loss due to charging a service fee when an entrance fee is optimal is generally small. However, especially when the high-priority load is high, charging a service fee yields a significant revenue gain. This is because in this case, low-priority customers anticipate a high probability of abandonment, so that charging a nonrefundable entrance fee would discourage a large proportion of these customers from joining the system.

We illustrate these observations with the following representative example.

Example 3: optimal service fee versus optimal nonrefundable entrance fee. Consider a system with parameters $R_l = 60$, $\mu = 1$, $\lambda_l = 0.4$, $c_l = 1$. Figure 4 shows the optimal service fee and the optimal nonrefundable entrance fee, as well as the resulting performance measures, as functions of the high-priority utilization. Figure 4 illustrates three key points. First, the revenue under the optimal service fee is smaller than under the optimal entrance fee for high-priority utilization levels below approximately 0.55, and vice versa for higher levels. Second, when charging a service fee is suboptimal, that is, for $\rho_h < 0.55$, the revenue loss versus an entrance fee is relatively small, less than 1%. Third, when charging an entrance fee is suboptimal, that is, for $\rho_h > 0.55$, the revenue loss versus a service fee increases in the high-priority utilization, to very significant levels – in this example to over 20% as $\rho_h \to 1^+$. This significant discrepancy between the pricing schemes can be explained as follows. As illustrated in Figure 4, regardless of the high-priority load, the joining probability under the optimal entrance fee is close to the service probability under the optimal service fee. Moreover, in order to encourage customers to purchase, the optimal nonrefundable entrance fee is set such that the conditional service probability for customers who join is close to 100% (this is not shown in Figure 4). However, when the high-priority utilization is high, the optimal service fee is significantly larger than the optimal entrance fee. Therefore, by charging customers for service versus for entrance, the firm can collect a significantly higher fee from approximately the same proportion of customers and thereby boost revenues by a large margin.

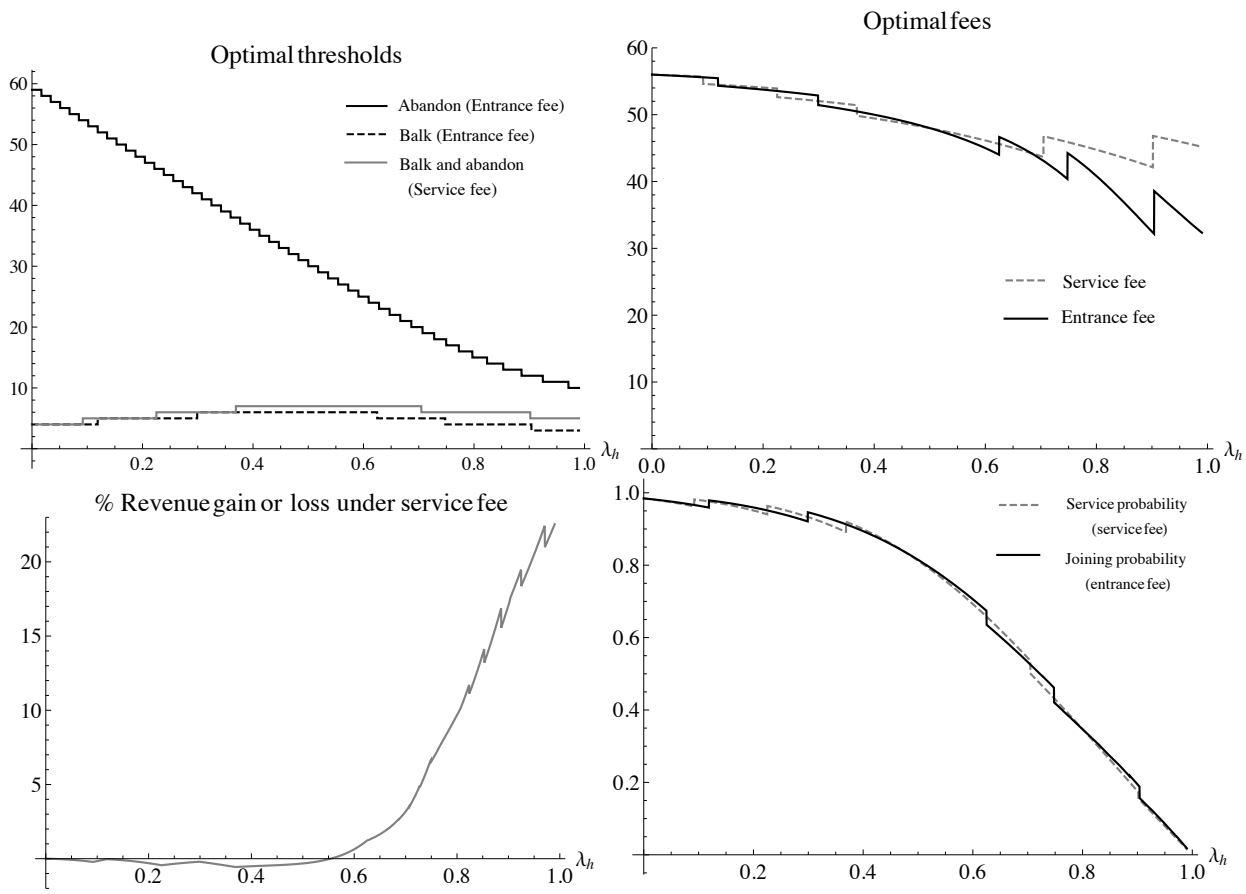

Figure 4 Comparison of service fee vs. entrance fee. System parameters: $R_l = 60, \mu = 1, \lambda_l = 0.4, c_l = 1$.

8. Concluding Remarks

Summary and discussion of results. Observable priority queues are prevalent in practice and create incentives for utility-maximizing low-priority customers to abandon after joining the system. The literature has so far ignored this behavior and the resulting system control issues. This paper contributes to closing these gaps by studying the abandonment strategy of rational customers and its pricing implications in an observable priority queue with two customer classes. Our main contributions are to characterize the equilibrium abandonment strategy of low-priority customers, and to provide guidelines on the optimal design of pricing policies that account for this behavior.

Our pricing results highlight the importance of the timing of charges for low-priority customers in the presence of rational abandonment, whereas this dimension of the fee structure is irrelevant in existing pricing models that ignore abandonment. In particular, welfare-maximization requires charging only a service fee and no entrance fee, whereas revenue-maximization typically requires both an entrance and a service fee. Charging low-priority customers both for entrance and service generates more revenue by reducing their abandonment risk: The two-fee structure is equivalent to charging only an entrance fee and offering a partial cancellation refund. In contrast, charging only

a nonrefundable fee for entrance (upon order placement) is common, and certainly more plausible, in models that ignore abandonment. Interestingly, this is how the Van Gogh Museum and the London Eye charge high-priority customers, who have no incentive to abandon.

These results also imply that a revenue-maximizing service provider can further increase revenues by capturing the entire system surplus through queue-length-dependent pricing, charging customers entrance fees that decrease in their system position upon joining, and offering a (common) partial cancellation refund that equals the socially optimal service fee. Intuitively, the longer the queue length, the higher the abandonment risk of a joining customer, and so the larger the (common) cancellation refund as a fraction of her upfront fee.

Whereas revenue-maximization typically requires a two-fee structure, charging a single fee is simple and practically appealing. We find that when the high-priority load is not too high, charging low-priority customers only for entrance and offering no cancellation refund typically generates only a little more revenue, compared to charging only for service. However, for sufficiently large high-priority loads, charging for service only may generate a lot more revenue than a nonrefundable entrance fee, because the latter structure requires dropping the fee in order to offset the significant abandonment risk. Therefore, charging for service could be a more robust strategy, especially when the arrival rate of high-priority customers varies over time and can be potentially high.

In light of these results, the London Eye's policy of charging low-priority customers a nonrefundable entrance fee, before standing in line, is justified as long as the utilization of high-priority customers is not too high, in relation to low-priority customers' reward-to-delay cost ratio. Otherwise, the London Eye may be better off switching to a pay-for-service model or even to a two-fee model with a cancellation refund. Compared to a service-fee-only model, the entrance-fee model may potentially make it operationally easier to keep the Ferris wheel fully utilized as it keeps a significant queue of "all-paid" low-priority customers ready to load. However, the deployment of wireless e-payment technologies would make it easy to quickly charge customers only if and when they get on the Eye. The Van Gogh Museum's policy of charging only for service is consistent with our results, both under revenue-maximization and under welfare-maximization; it is operated by a foundation as a national museum and may therefore moderate revenue generation by other objectives.

More generally, our results suggest that priority pricing policies based on demand models that mistakenly ignore customer abandonment, may potentially reduce system performance significantly.

Future research. Recent empirical studies (Batt et al. 2015, Bolandifar et al. 2014) highlight the need for models of customer abandonment in observable queues that also account for factors other than the elapsed time in the system, which is at the core of standard abandonment models based on exogenous patience time distributions. Our study is a step in this direction and points

to numerous opportunities for further research on modeling and controlling priority queues with customer abandonment. We conclude by outlining a few of these.

One important problem is to consider multiple classes of customers with different cost and reward parameters. The approach developed in this work can also be applied to characterize the equilibrium strategy of customers in a multi-class system, as the solution to inter-related optimal stopping problems. The analysis is more complicated than in the two-class case, as for each class, one needs to keep track of the number of customers in all lower priority classes. Interestingly, the last-come, first-abandon structure does no longer hold in general. For example, in a three-class system, second-priority customers may abandon before the lowest-priority customers. Ongoing work considers how the rational abandonment behavior depends on the reward and delay cost parameters of the different customer classes.

We assume that all customers are fully rational forward-looking utility maximizers. One can relax this assumption and consider settings where all or some of the customers are myopic. That is, they ignore future arrivals of high-priority customers and join/stay in system if, and only if, their expected utility based on their *current* position is nonnegative. It is not hard to see that the abandonment strategy of such customers has a threshold structure with the threshold being equal to the balking threshold of Naor (1969). It follows that such customers join and stay in higher positions compared to forward-looking ones and hence do not affect their abandonment strategy. Another way of relaxing the rationality assumption is to assume that customers are boundedly rational. That is, they do not always make the choice that maximizes their expected utility due to lack of information or inability in estimating their expected waiting times; see Huang et al. (2013) who study how bounded rationality affects join/balk decisions in $M/M/1$ FCFS systems.

Our model can also be modified to allow for customers to purchase priority. In this case, a customer needs to decide which class to join, given that she may abandon the queue later if she joins the low-priority queue. Note that the waiting time of a low-priority customer also depends on the priority purchasing decisions of future customers. Adiri and Yechiali (1974) show that, when abandonment is *precluded*, the priority purchasing decision of customers is characterized by a single threshold, such that if the number of low-priority customers in system is below the threshold, a new arrival joins the low-priority queue and purchases priority otherwise. In the absence of abandonment, this implies that the low-priority line is chosen only when the high-priority line is empty. However, this is no longer the case when customers *can* abandon the queue after joining: In that case, as new arrivals start to purchase priority and join the high-priority line, customers in the low-priority line may start to abandon the queue, which may make the low-priority line the optimal choice for future arrivals.

Another direction is to model and analyze customers' retrial decisions in addition to their join/balk/abandon decisions. Whereas retrial decisions have been recognized as an important phenomenon in the call center literature (for example, see Mandelbaum et al. 2002, Gans et al. 2003, de Véricourt and Zhou 2005, Akşin et al. 2007, Aguir et al. 2008), these papers typically model retrials with a constant probability. The study of rational retrial models appears to be recent. Focusing on a single-class FIFO system, Cui et al. (2014) study the equilibrium and the welfare implications in a model of rational retrial in queues.

This paper also points to the need for more research on the control of queues with rational abandonment. This work appears to be the first to investigate pricing as a means to control the abandonment behavior of customers. Our findings show that considering abandonment has significant implications for pricing prescriptions, which suggests the need for more research on the interplay between pricing and abandonment, for example for unobservable queues or for customers with nonlinear delay costs. Furthermore, there is a need to investigate operational controls, such as scheduling, for systems in the presence of rational customer abandonment. We hope that this research stimulates more work on the design and control of queuing systems in the presence of rational customer abandonment.

References

Adiri I, Yechiali U (1974), Optimal priority purchasing and pricing decisions in nonmonopoly and monopoly queues, Operations Research 22:1051–1066.

Afèche P (2004) Incentive-compatible revenue management in queueing systems: optimal strategic idleness and other delay tactics. Working paper, University of Toronto.

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. Manufacturing & Service Operations Management 15(3):423–443.

Aguir S, Akşin Z, Karaesmen F, Dallery Y (2008) On the interaction between retrials and sizing of call centers. European Journal of Operational Research 191(2):398–408.

Akşin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. Production and Operations Management 16(6):665–688.

Akşin Z, Ata B, Emadi S, Su C (2013) Structural estimation of callers' delay sensitivity in call centers, Management Science 59(12):2727–2746.

Alperstein H (1988) Optimal pricing policy for the service facility offering a set of priority prices, Management Science, 34:666–671.

Assaf D, Haviv M (1990) Reneging from processor sharing systems and random queues, Mathematics of Operations Research 15(1):129–138.

Ata B, Glynn PW, Peng X (2015) An equilibrium analysis of a multiclass queue with endogenous abandonments. Working paper. University of Chicago, Chicago, IL.

Ata B, Olsen T (2013) Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. Queueing Systems 73(1):35–78.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. Operations Research 58(5):1427–1439.

Baccelli F, Hebuterne G (1981) On queues with impatient customers. Performance '81 (North-Holland Publishing Company, Amsterdam) 159–179.

Baron O, Milner J (2009) Staffing to maximize profit for call centers with alternate service level agreements. Operations Research 57(3):685–700.

Barrer DY (1957) Queuing with impatient customers and indifferent clerks. Operations Research 5:644–649.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity planning in queueing systems with impatient customers. Operations Research 58(5):1398–1413.

Bassamboo A, Randhawa RS (2015) Scheduling homogeneous impatient customers. Management Science. Forthcoming.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. Management Science 61(1):39–59.

Bolandifar E, Dehoratius N, Olsen T, Wiler JL (2014) Modeling the behavior of patients who Leave the emergency department without being seen by a physician. Working paper. University of Chicago, Chicago, IL.

Brandt A and Brandt M (2004). On the two-class M/M/1 system under preemptive resume and impatience of the prioritized customers. Queueing Systems 47(1/2): 147–168.

Brouns GAJF, van der Wal J (2006) Optimal threshold policies in a two-class preemptive priority queue with admission and termination control. Queueing Syst. Theory Appl 54: 21–33.

Cui S, Su X, Veeraraghavan S (2014) A model of rational retrials in queues. Working paper, Georgetown University, Washington, DC.

de Véricourt F, Zhou Y (2005) Managing response time in a call-routing problem with service failure. Operations Research 53(6):968–981.

Down DG, Koole G, Lewis ME (2011) Dynamic control of a single-server system with abandonments. Queueing Systems 67(1): 63–90.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. Manufacturing & Service Operations Management 4(3):208-227.

Guo L (2009) Service cancellation and competitive refund policy. Marketing Science 28(5):901–917.

Guo P, Zhang ZG (2013) Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. Manufacturing & Service Operations Management 15(1):118–131.

Hassin R (1985) On the optimality of first come last served queues, Econometrica 53(1):201-202.

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. Oper. Res. Lett. 17(1):41–45.

Hassin R, Haviv M (1997) Equilibrium threshold strategies: The case of queues with priorities, Operations Research 45:966-973.

Hassin R, Haviv M (2003) To queue or not to queue: Equilibrium behavior in queueing systems (Kluwer Academic Publishers, Norwell, MA).

Haviv M, Ritov Y (2001) Homogeneous Customers Renege from invisible queues at Random Times under deteriorating waiting conditions. Queueing Systems 38: 495–508.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. Manufacturing Service Oper. Management 15: 263–279.

Iravani F, Balcıoğlu B (2008) On priority queues with impatient customers. Queueing Systems 58(4): 239–260.

Jennings OB, Reed JE (2012) An overloaded multiclass FIFO queue with abandonments. Operations Research, 60(5), 1282-1295.

Jouini O, Roubos A (2013) On multiple priority multi-server queues with impatience. Journal of the Operational Research Society, 65: 616-632.

Lee C, Ward AR (2014) Optimal pricing and capacity sizing when customers abandon. Working paper. Colorado State University, Fort Collins, CO.

Maglaras C, Yao J and Zeevi A (2015) Observational learning in queues with abandonments. Working paper. Columbia University, New York, NY.

Maglaras C, Yao J, Zeevi A (2015) Optimal price and delay differentiation in queueing systems. Forthcoming in Management Science.

Mandelbaum A, Shimkin N (2000) A model for rational abandonment from invisible queues, Queueing Systems: Theory and Applications, 36:141–173.

Mandelbaum A, Massey W, Reiman M, Stolyar A, Rider B (2002) Queue lengths and waiting times for multiserver queues with abandonment and retrials. Telecommunication Systems 21(2-4):149–171.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the $M/M/1$ queue. Oper. Res. 38(5):870–883.

Naor P (1969) On the regulation of queue size by levying tolls. Econometrica 37(1):15–24.

Nazerzadeh H, Randhawa R (2015) Asymptotic optimality of two service grades for customer differentiation in queueing systems. Working paper, University of Southern California, Los Angeles, CA.

Righter R (2000) Expulsion and scheduling control for multiclass queues with heterogeneous servers. Queueing Systems: Theory Appl. 34:289–335.

Rosenthal JS (2000) A first look at rigorous probability theory. World Scientific Publishing, Singapore.

Sarhangian V and Balcıoğlu B (2013) Waiting time analysis of multi-class queues with impatient customers. Probability in the Engineering and the Informaional Sciences 27(3): 333–352

Shimkin N, Mandelbaum A (2004) Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. Queueing Systems: Theory Appl. 47(1–2):117–146.

Yu Q, Allon G, and Bassamboo A (2015) How do delay announcements shape customer behavior? An empirical study. Management Science. Forthcoming.

Ward AR (2011) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. Surveys Oper. Res. Management Sci. 16(1):1–14.

Weber RR and Stidham S (1987) Optimal control of service rates in networks of queues. Advances in Applied Probability 19(1):202-218.

Whitt W (2006) Fluid models for multiserver queues with abandonments. Operations Research 54(1):37-54.

Xie J and Gerstner E (2007) Service Escape: Profiting from Customer Cancellations. Marketing Science, 26(1), 18-30.

Xu SH and Shanthikumar JG (1993) Optimal expulsion control–a dual approach to admission control of an ordered-entry system. Operations Research 41:1137–1152.

Xu SH (1994) A duality approach to admission and scheduling controls of queues. Queueing Systems: Theory and Applications, 18:273–300.

Zeltyn S, and Mandelbaum A (2005) Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. Queueing Systems 51(3/4), 361-402.

Appendix A: Proofs.

Proof of Proposition 1. The proof is by induction. In STEP 1 we start with the customer at the first low-priority position ($m=1$), whose strategy does not depend on that of other low-priority customers. We formulate her join/balk/abandon problem as a dynamic program and show that her strategy has the claimed threshold structure. In STEP 2 we show that, assuming the first $m \geq 1$ customers follow the claimed threshold strategy, the customer at low-priority position $m+1$ also follows a threshold strategy and the thresholds $\bar{n}(m)$ and $\bar{n}(m+1)$ satisfy properties (i) and (ii) in part 2 of the proposition. We do so by considering the solution properties of the dynamic programs for the customers in low-priority positions m and $m+1$. Finally, in STEP 3, we characterize the thresholds K and L.

STEP 1. Consider a tagged customer at low-priority position $m=1$, that is, there are no low-priority customers in front of her. Any waiting cost incurred so far is sunk, so given the customer is still in system, she maximizes her expected future utility. Since there can be a maximum of $\bar{n}_h$ high-priority customers in system, the position of the tagged customer is $(1,n)$ with $n \in \{1,...,\bar{n}_h+1\}$. For system positions $n \leq \bar{n}_h$ the next event is either a service completion, with probability $\bar{\mu} \equiv \mu/(\mu+\lambda_h)$, or a new-high priority arrival with probability $\bar{\lambda} \equiv \lambda_h/(\mu+\lambda_h)$. For position $(1,\bar{n}_h+1)$ since the maximum number of high-priority customers is reached, any new high-priority arrival balks and does not change the position of the tagged customer. Note that due to the memoryless property of the arrival and service processes, no new information becomes available between arrivals and service completions, so the customer would only abandon at these discrete epochs. At each decision epoch, the customer can either abandon and receive an immediate reward of 0 or stay until the next event. We consider the uniformized system with transition rate $\mu+\lambda_h$. Therefore, at each decision epoch, if the customer decides to stay in system she incurs an average waiting cost of $\bar{c} \equiv c_l/(\mu+\lambda_h)$ until the next epoch, where $\bar{c}/\bar{\mu} = c_l/\mu$. At position $(1,1)$ if the customer decides to wait and the next event is a service completion, she moves to $(0,0)$ and receives the reward R_l. Let $v(n,m)$ denote the value function of the customer at position (n,m), interpreted as the maximum future expected utility for a customer at that position, and set $v(0,0)=R_l$. Then, writing $g^+ \equiv \max(0,g)$, the value function for the customer at low-priority position $m=1$ satisfies

$$v(1,1) = [-\bar{c}+\bar{\lambda}v(1,2)+\bar{\mu}v(0,0)]^+, \tag{29}$$

$$v(1,n) = [-\bar{c}+\bar{\lambda}v(1,n+1)+\bar{\mu}v(1,n-1)]^+, \quad n \in \{2,...,\bar{n}_h\}, \tag{30}$$

$$v(1,\bar{n}_h+1) = [-\bar{c}+\bar{\lambda}v(1,\bar{n}_h+1)+\bar{\mu}v(1,\bar{n}_h)]^+, \tag{31}$$

Applying Lemma 1 (Appendix B) with $i=m$ and $T=\bar{n}_h$, we have that $v(1,n)$ strictly decreases in $n \leq \bar{n}(1)$, where by part (ii) of Lemma 1

$$\bar{n}(1) = \begin{cases} 1, & \text{if } v(1,1) < c_l/\mu, \\ \max\{n \in \{2,...,\bar{n}_h+1\}; v(1,n-1) \geq c_l/\mu\}, & \text{otherwise,} \end{cases} \tag{32}$$

because $v(0,0) = R_l > c_l/\mu$ by assumption. That is, a customer in low-priority position $m = 1$ joins/stays if and only if her system position $n \leq \bar{n}(1)$. Therefore, the solution of (29)–(31) satisfies

$$v(1,1) = -\bar{c} + \bar{\lambda} v(1,2) + \bar{\mu} v(0,0), \tag{33}$$

$$v(1,n) = -\bar{c} + \bar{\lambda} v(1,n+1) + \bar{\mu} v(1,n-1), \quad n \in \{2,...,\bar{n}(1)-1\}, \tag{34}$$

$$v(1,\bar{n}(1)) = \begin{cases} -\bar{c} + \bar{\lambda} v(1,\bar{n}(1)) + \bar{\mu} v(1,\bar{n}(1)-1), & \bar{n}(1) = \bar{n}_h + 1, \\ -\bar{c} + \bar{\mu} v(1,\bar{n}(1)-1), & 1 < \bar{n}(1) < \bar{n}_h + 1, \end{cases} \tag{35}$$

and $v(1,n) = 0$ for $n > \bar{n}(1)$, where

$$\bar{n}(1) > 1 \Leftrightarrow v(1,\bar{n}(1)-1) \geq c_l/\mu, \text{ and } \bar{n}(1) < \bar{n}_h + 1 \Rightarrow c_l/\mu > v(1,\bar{n}(1)). \tag{36}$$

STEP 2. We turn to the induction step. Assume that customers are willing to join in the first $m \geq 1$ low-priority positions and all customers in these positions follow a join/balk/abandon strategy with thresholds $\{\bar{n}(1),\ldots,\bar{n}(m)\}$ satisfying part 2 of Proposition 1. In this case, $\bar{n}(i) \leq \bar{n}(i-1)+1$ for $i \in \{2,3,...,m\}$, so the next customer to abandon (if any) is the customer at low-priority position m and from system position $\bar{n}(m)+1$, and her value function satisfies

$$v(m,m) = -\bar{c} + \bar{\lambda} v(m,m+1) + \bar{\mu} v(m-1,m-1), \tag{37}$$

$$v(m,n) = -\bar{c} + \bar{\lambda} v(m,n+1) + \bar{\mu} v(m,n-1), \quad n \in \{m+1,...,\bar{n}(m)-1\}, \tag{38}$$

$$v(m,\bar{n}(m)) = \begin{cases} -\bar{c} + \bar{\lambda} v(m,\bar{n}(m)) + \bar{\mu} v(m,\bar{n}(m)-1), & \bar{n}(m) = \bar{n}_h + m, \\ -\bar{c} + \bar{\mu} v(m,\bar{n}(m)-1), & m < \bar{n}(m) < \bar{n}_h + m, \end{cases} \tag{39}$$

$v(m,n) = 0$ for $n > \bar{n}(m)$, where

$$\bar{n}(m) > m \Leftrightarrow v(m,\bar{n}(m)-1) \geq c_l/\mu, \text{ and } \bar{n}(m) < \bar{n}_h + m \Rightarrow c_l/\mu > v(m,\bar{n}(m)). \tag{40}$$

Note that for $m = 1$, (37)–(40) agree with (33)–(36).

Next consider the problem of the customer at low-priority position $m+1$. Given the first m low-priority customers' strategy, there are at most $\bar{n}(m)$ customers (of either class) in front of her, so her maximum system position is $\bar{n}(m)+1$. We have two cases, 1) $\bar{n}(m) = m$, and 2) $\bar{n}(m) > m$.

Case 1) $\bar{n}(m) = m$. We show that customers prefer balking to joining in low-priority position $m+1$, so that $m = L$. Noting that $v(m+1,m+1) = v(m+1,\bar{n}(m)+1)$, a customer who considers joining low-priority position $m+1$ solves

$$v(m+1,\bar{n}(m)+1) = [-\bar{c} + \bar{\lambda} v(m,\bar{n}(m)+1) + \bar{\mu} v(m,m)]^+ = [-\bar{c} + \bar{\mu} v(m,m)]^+, \tag{41}$$

where the first equality follows because upon arrival of a high-priority customer, the customer at low-priority position m abandons as she moves to position $(m,\bar{n}(m)+1)$, which in turn moves the customer at low-priority position $m+1$ to position $(m,\bar{n}(m)+1)$. The second equality in (41)

follows because $v(m,\bar{n}(m)+1)=0$ by hypothesis. Noting from (40) that $c_l/\mu > v(m,m)$ because $\bar{n}(m)=m$, and $c_l/\mu=\bar{c}/\bar{\mu}$, it follows that $-\bar{c}+\bar{\mu}v(m,m)<0$ in (41), so the expected utility of joining low-priority position $m+1$ is strictly negative.

Case 2) $\bar{n}(m)>m$. We establish that the following holds in equilibrium for a customer at low-priority position $m+1$: First, she follows a threshold join/balk/abandon strategy with threshold $\bar{n}(m+1)$; second, the thresholds $\bar{n}(m)$ and $\bar{n}(m+1)$ satisfy the properties $(i)-(ii)$ in part 2 of Proposition 1 and her value function satisfies (37)–(40) with m replaced by $m+1$.

First, for the threshold structure, note that a customer in low-priority position $m+1$ solves:

$$v(m+1,m+1) = [-\bar{c}+\bar{\lambda}v(m+1,m+2)+\bar{\mu}v(m,m)]^+, \tag{42}$$

$$v(m+1,n) = [-\bar{c}+\bar{\lambda}v(m+1,n+1)+\bar{\mu}v(m+1,n-1)]^+, \quad n\in\{m+2,...,\bar{n}(m)\}, \tag{43}$$

$$v(m+1,\bar{n}(m)+1) = \begin{cases} [-\bar{c}+\bar{\lambda}v(m+1,\bar{n}(m)+1)+\bar{\mu}v(m+1,\bar{n}(m))]^+, & \bar{n}(m)=\bar{n}_h+m, \\ [-\bar{c}+\bar{\mu}v(m+1,\bar{n}(m))]^+, & \bar{n}(m)<\bar{n}_h+m. \end{cases} \tag{44}$$

Note that (44) captures two cases for position $(m+1,\bar{n}(m)+1)$ depending on the strategy of the customer at low-priority position m, who is then in position $(m,\bar{n}(m))$. If $\bar{n}(m)=\bar{n}_h+m$, then any new high-priority arrival would balk and hence does not affect the position of the customer in low-priority position $m+1$. However, for $\bar{n}(m)<\bar{n}_h+m$ we have, similar to (41), that

$$v(m+1,\bar{n}(m)+1)=[-\bar{c}+\bar{\lambda}v(m,\bar{n}(m)+1)+\bar{\mu}v(m+1,\bar{n}(m))]^+=[-\bar{c}+\bar{\mu}v(m+1,\bar{n}(m))]^+,$$

because a high-priority arrival triggers the customer at low-priority position m to abandon.

That the solution of (42)–(44) has a threshold structure follows because by Lemma 1 with $i=m+1$ and $T=\bar{n}(m)$, we have that $v(m+1,n)$ strictly decreases in $n\le\bar{n}(m+1)$, where by part (ii) of Lemma 1,

$$\bar{n}(m+1)=\begin{cases} m+1, & \text{if } v(m+1,m+1)<c_l/\mu, \\ \max\{n\in\{m+2,...,\bar{n}(m)+1\}; v(m+1,n-1)\ge c_l/\mu\}, & \text{otherwise,} \end{cases} \tag{45}$$

because $\bar{n}(m)>m$ and (40) imply that $v(m,m)\ge v(m,\bar{n}(m)-1)\ge c_l/\mu$. That is, a customer in low-priority position $m+1$ joins/stays in the system if and only if her system position $n\le\bar{n}(m+1)$.

Second, we establish that the thresholds $\bar{n}(m)$, defined in (40), and $\bar{n}(m+1)$, defined in (45), satisfy the properties $(i)-(ii)$ in part 2 of Proposition 1, and that the value function of a customer at low-priority position $m+1$ satisfies (37)–(40) with m replaced by $m+1$.

For property (i) we need to show that if $\bar{n}(m+1)=\bar{n}_h+m+1$ then $\bar{n}(m)=\bar{n}_h+m$. This follows because by (45) we have $v(m+1,\bar{n}(m+1)-1)\ge c_l/\mu$ and $m+1\le\bar{n}(m+1)-1\le\bar{n}(m)$, and by Lemma 2 we have $v(m,n)\ge v(m+1,n)$ for $n\in\{m+1,\ldots,\bar{n}(m)\}$. Applying the latter

inequality for $n = \bar{n}(m+1)-1$ and noting that $\bar{n}(m+1)-1 = \bar{n}_h + m$, we have that $v(m, \bar{n}_h + m) \geq v(m+1, \bar{n}_h + m) \geq c_l/\mu$, which implies in light of (40) that $\bar{n}(m) = \bar{n}_h + m$.

For property (ii) we need to show that if $\bar{n}(m) < \bar{n}_h + m$ then $\bar{n}(m+1) = \bar{n}(m)$. This follows because by (40) we have $v(m, \bar{n}(m)-1) \geq c_l/\mu > v(m, \bar{n}(m))$ and by part (ii) of Lemma 2 we have that if $\bar{n}(m) < \bar{n}_h + m$, then $v(m+1, n) = v(m, n)$ for $n \in \{m+1, \ldots, \bar{n}(m)+1\}$. Therefore, we have $v(m+1, \bar{n}(m)-1) \geq c_l/\mu$ if $\bar{n}(m) > m+1$ and also $c_l/\mu > v(m+1, \bar{n}(m))$ for $\bar{n}(m) \geq m+1$, which implies in light of (45) that $\bar{n}(m+1) = \bar{n}(m)$.

Furthermore, by substituting for $\bar{n}(m)$ into (42)–(45), using $\bar{n}(m+1) = \bar{n}(m)+1$ if $\bar{n}(m+1) = \bar{n}_h + m + 1$ and $\bar{n}(m+1) = \bar{n}(m)$ if $\bar{n}(m) < \bar{n}_h + m$, it is straightforward to verify that the value function of the customer in low-priority position $m+1$ satisfies (37)–(40) with m replaced by $m+1$.

STEP 3. We characterize the thresholds K and L, where $K \equiv 0$ if $\bar{n}(1) < \bar{n}_h + 1$ and $K \equiv \max\{m \geq 1; \bar{n}(m) = \bar{n}_h + m\}$ otherwise, and $L \equiv \min\{m \geq 1; \bar{n}(m) = m\}$. The fact that $L = \bar{n}_h + K$ if $K \geq 1$ follows because $\bar{n}(K) = \bar{n}_h + K$ and $\bar{n}(K+1) < \bar{n}_h + K + 1$ by definition, and $\bar{n}(K) \leq \bar{n}(K+1)$ by the properties (i) and (ii) established in STEP 2. Therefore, $\bar{n}(K) = \bar{n}(K+1)$. Furthermore $\bar{n}(K+1) = \bar{n}(L) = L$ as established above.

Next we establish upper bounds on K and L by using the properties of the value function established in Lemma 1. (These bounds also follow by observing that the expected utility of a low-priority customer in a given position is lower than what it would be if no future high-priority arrivals were to overtake her; refer to the right-hand sides of (49) and (50).) We have from (92) that

$$\bar{n}(m) = \bar{n}_h + m \Leftrightarrow v(m, \bar{n}_h + m - 1) \geq c_l/\mu \text{ and } \bar{n}(m) = m \Leftrightarrow v(m-1, m-1) \geq c_l/\mu > v(m, m).$$

Therefore $K = 0$ if $v(1, \bar{n}_h) < c_l/\mu$ and $K = \max\{m \geq 1; v(m, \bar{n}_h + m - 1) \geq c_l/\mu\}$ otherwise, and $L \equiv \max\{m \geq 1; v(m-1, m-1) \geq c_l/\mu\}$ since $v(0,0) = R_l > c_l/\mu$. Furthermore, from (93) in part (iii) of Lemma 1 we have for any m with $v(m-1, m-1) \geq c_l/\mu$ that

$$v(m, m) \leq v(m-1, m-1) - c_l/\mu, \tag{46}$$

$$v(m, \bar{n}_h + m - 1) \leq v(m-1, m-1) - \bar{n}_h c_l/\mu, \text{ if } \bar{n}_h + m = \bar{n}(m). \tag{47}$$

Noting that $v(i-1, i-1) \geq c_l/\mu$ for every low-priority position $i \leq m$ that customers are willing to join, it follows from (46) that

$$v(m-1, m-1) = v(0,0) + \sum_{i=1}^{m-1} v(i,i) - v(i-1, i-1) \leq R_l - (m-1)c_l/\mu. \tag{48}$$

From the definition of K and by (46)–(48), the following must hold for $1 \leq m \leq K$:

$$c_l/\mu \leq v(m, \bar{n}_h + m - 1) \leq R_l - (\bar{n}_h + m - 1)c_l/\mu. \tag{49}$$

Therefore, if $\bar{n}_h \geq R_l\mu/c_l$ then this condition is violated for all m, so $K=0$, that is, $\bar{n}(1) < \bar{n}_h + 1$.

From the definition of L and by (48), the following must hold for $1 \leq m \leq L$:

$$c_l/\mu \leq v(m-1, m-1) \leq R_l - (m-1)c_l/\mu. \tag{50}$$

The proof is complete. □

Proof of Proposition 2. We show that the value function and the equilibrium thresholds satisfy (6) and (5) respectively for $1 \leq m \leq L$. From Proposition 1 (see also (37)–(39)) in equilibrium the value function satisfies the recursion

$$v(m,m) = -\bar{c} + \bar{\lambda} v(m, m+1) + \bar{\mu} v(m-1, m-1), \tag{51}$$

$$v(m,n) = -\bar{c} + \bar{\lambda} v(m, n+1) + \bar{\mu} v(m, n-1), \quad n \in \{m+1, ..., \bar{n}(m)-1\}, \tag{52}$$

$$v(m, \bar{n}(m)) = \begin{cases} -c_l/\mu + v(m, \bar{n}(m)-1), & \bar{n}(m) = \bar{n}_h + m, \\ -\bar{c} + \bar{\mu} v(m, \bar{n}(m)-1), & m < \bar{n}(m) < \bar{n}_h + m, \end{cases} \tag{53}$$

for $1 \leq m \leq L$ and with $v(0,0) = R_l$, $v(m, \bar{n}(m)+1) = 0$ and $v(m,n) \geq 0$ for $n \in \{m, ..., \bar{n}(m)\}$. Observe that for each m the value function depends on the value of other low-priority positions only through that of position $(m-1, m-1)$, i.e., the position at which the customer finds herself right after the next low-priority customer is served. Therefore, for each low-priority position m, given the "reward-to-go" value $\mathcal{R} \equiv v(m-1, m-1)$ one can characterize the equilibrium threshold by finding the maximum number of remaining service completions before advancing to the next low-priority position, such that the customer does not abandon, i.e., $\bar{n}(m) - (m-1)$.

Define $u(i, h, \mathcal{R})$ satisfying

$$\begin{aligned} u(1, h, \mathcal{R}) &= -\bar{c} + \bar{\lambda} u(2, h, \mathcal{R}) + \bar{\mu} u(0, h, \mathcal{R}), \\ u(i, h, \mathcal{R}) &= -\bar{c} + \bar{\lambda} u(i+1, h, \mathcal{R}) + \bar{\mu} u(i-1, h, \mathcal{R}), \quad i \in \{2, ..., h-1\}, \\ u(h, h, \mathcal{R}) &= \begin{cases} -c_l/\mu + u(h-1, h, \mathcal{R}), & h = \bar{n}_h + 1, \\ -\bar{c} + \bar{\mu} u(h-1, h, \mathcal{R}), & h < \bar{n}_h + 1, \end{cases} \end{aligned}$$

with $u(0, h, \mathcal{R}) = \mathcal{R} \geq c_l/\mu$, and $u(i, h, \mathcal{R}) \geq 0$. Note that $u(i, h, \mathcal{R})$ is the expected utility of staying in system for a tagged low-priority customer who currently faces $i \leq \bar{n}_h + 1$ remaining service completions before advancing one position in the queue and receiving reward $\mathcal{R}$, and stays in the system as long as the number of service completions that are required before advancing by one low-priority position does not exceed h. The value function thus satisfies

$$v(m,n) = u(n-(m-1); \bar{n}(m) - (m-1), v(m-1, m-1)),$$

for $1 \leq m \leq L$ and $m \leq n \leq \bar{n}(m)$ as claimed in (6). The explicit expression for $u(\cdot, h, \mathcal{R})$ given in (4) can be obtained, e.g., by directly solving the above difference equation for each case (i.e.,

$h = \bar{n}_h + 1$ and $h < \bar{n}_h + 1$). An intuitive derivation is provided in the main body of the paper and hence the detailed derivation is omitted here.

Finally, we show that the thresholds satisfy (5), i.e.,

$$\bar{h} \equiv \bar{n}(m) - (m-1) = \max\{h \in \{1, \ldots, \bar{n}_h + 1\}; u(h, h, v(m-1, m-1)) \geq 0\}. \tag{54}$$

First, consider the case where $u(\bar{n}_h + 1, \bar{n}_h + 1, v(m-1, m-1)) \geq 0$. In this case the tagged customer has nonnegative expected utility even if she faces the maximum possible number of service completions that can be required before advancing by one position, i.e., $\bar{n}_h + 1$. From (4) we have $u(i, \bar{n}_h + 1, v(m-1, m-1)) = v(m-1, m-1) - c_l \bar{w}(i, \bar{n}_h + 1)$, and since $\bar{w}(i, h)$ is increasing in i the customer receives nonnegative utility from staying in all system positions and hence has no incentive to abandon, i.e., $\bar{h} = \bar{n}_h + 1$. Otherwise from Proposition (1) we know that the customer abandons once her system position exceeds some threshold $\bar{n}(m) < m + \bar{n}_h$. To see that this threshold must satisfy (54) note that using (4),

$$u(i, h, v(m-1, m-1)) = v(m-1, m-1) q(i, h+1) - c_l w(i, h+1).$$

Hence, $u(i, h, v(m-1, m-1)) \geq 0$ iff $w(i, h+1)/q(i, h+1) \leq v(m-1, m-1)/c_l$. From part (ii) of Lemma (4) we know that $w(i, h+1)/q(i, h+1)$ is increasing in i, thus $u(\bar{h}, \bar{h}, v(m-1, m-1)) \geq 0$ implies that $u(i, \bar{h}, v(m-1, m-1)) \geq 0$ for all $i \leq \bar{h}$, i.e., the customer has no incentive to abandon in any of the lower positions. Further, from part (i) of Lemma (4) we know that $w(h, h+1)/q(h, h+1)$ is increasing in h, hence $u(\bar{h}+1, \bar{h}+1, v(m-1, m-1)) < 0$ implies that $u(h, h, v(m-1, m-1)) < 0$ for $\bar{h} + 1 \leq h \leq \bar{n}_h + 1$. Therefore, the customer has no incentive to stay in any higher positions either, so the unique threshold must satisfy (54). □

Proof of Proposition 4. We first show that the socially optimal policy has a threshold structure that satisfies property (i), that is, equal balking and abandonment thresholds. To this end, we reformulate the social planner's problem as an infinite horizon Markov decision process. We first consider the α-discounted problem and show that the claims are satisfied for any discount rate $\alpha > 0$. It is well-known that under certain conditions the long-run average optimal policy can be obtained from the discounted optimal policy by letting $\alpha \to 0$ (Weber and Stidham 1987). It is straightforward to verify that these conditions are satisfied for our problem. Thus, the average optimal policy must also satisfy the claimed properties.

Let $z(x_1, x_2)$ denote the maximum expected total discounted system welfare generated by low-priority customers when the initial state is (x_1, x_2), where $x_1 \geq 0$ and $x_2 \geq 0$ respectively denote the number of high- and low-priority customers in the system. Defining $\Lambda \equiv \lambda_l + \lambda_h + \mu$, and applying uniformization the optimality equations are given by:

$$z(x_1, x_2) = \frac{1}{\Lambda + \alpha} \left[-c_l x_2 + \lambda_l \mathcal{T}_1 z(x_1, x_2) + \lambda_h \mathcal{T}_2 z(x_1, x_2) + \mu \mathcal{T}_3 z(x_1, x_2) \right],$$

where

$$\mathcal{T}_1 z(x_1,x_2) = \max\left[z(x_1,x_2+1), z(x_1,x_2)\right], \tag{55}$$

$$\mathcal{T}_2 z(x_1,x_2) = \max_{0\le j\le x_2} z(x_1+1,j), \tag{56}$$

$$\mathcal{T}_3 z(x_1,x_2) = \begin{cases} z(x_1-1,x_2), & x_1>0, x_2\ge 0,\\ R_l + z(0,x_2-1), & x_1=0, x_2>0,\\ z(0,0), & x_1=x_2=0. \end{cases} \tag{57}$$

In the above formulation, (55) and (56) respectively correspond to arrival instances of low- and high-priority customers and (57) corresponds to a service completion. The decision epochs are arrival instances of customers to the system (note that it cannot be optimal to remove low-priority customers upon a service completion). Upon arrival of a low-priority customer, the controller can decide whether to admit or reject the customer. Upon arrival of a high-priority customer, the controller can remove low-priority customers (if any) from the system. Note that in (56) we allow "batch removals", i.e., the controller can remove more than one low-priority customer upon arrival of a high-priority customer. However, under the optimal policy this can only affect the transient states of the system (i.e., if the initial number of customers is above the socially optimal threshold $\bar{m}^*$) and hence does not change the steady-state probabilities.

We prove that $z(x_1,x_2)$ satisfies the following properties, which proves the result:

$$z(x_1,x_2+2) - z(x_1,x_2+1) \le z(x_1,x_2+1) - z(x_1,x_2), \tag{58}$$

$$z(x_1+1,x_2+1) - z(x_1+1,x_2) = z(x_1,x_2+2) - z(x_1,x_2+1). \tag{59}$$

First note that these conditions, together with (55), imply that the optimal admission policy of low-priority customers has a threshold structure that only depends on the total number of customers in the system. Inequality (58) states that $z(x_1,x_2)$ is concave in x_2. It follows from (55)–(58) that for each $x_1\ge 0$, it is optimal to reject new low-priority arrivals in state (x_1,j) if and only if $j\ge \bar{x}_2(x_1)\equiv \min\{x_2\ge 0; z(x_1,x_2+1)-z(x_1,x_2)<0\}$. Further, (59) states that the value of an additional low-priority customer only depends on the total number of customers in system. It follows that there exist a threshold on the total number of customers in system $\bar{m}_b^*$, such that $\bar{x}_2(x_1) = (\bar{m}_b^* - x_1)^+$ and it is socially optimal to accept new low-priority arrivals as long as the total number of customers in system is less than $\bar{m}_b^*$, and reject otherwise.

Next, (59) further implies that the optimal removal of low-priority customers matches the optimal admission policy. Specifically, (59) implies that if it is optimal to reject a new low-priority arrival in a given state, then it must also be optimal to remove a low-priority customer upon arrival of a high-priority customer to that state, and vice versa, i.e., $\bar{m}_a^* = \bar{m}_b^* \equiv \bar{m}^*$. To see this, assume that

it is optimal to reject a new low-priority arrival in state (x_1, x_2) with $x_2 \geq 1$, i.e., $z(x_1, x_2+1) - z(x_1, x_2) < 0$. By (59) this is equivalent to $z(x_1+1, x_2) - z(x_1+1, x_2-1) < 0$ and therefore by (56) it is also optimal to remove a low-priority customer upon arrival of a high-priority one. Note this also implies that bulk removals of low-priority customers cannot be optimal (unless the initial number of customers in the system exceeds the threshold).

To prove that z indeed satisfies properties (58) and (59), we use the value iteration algorithm. Let $z_0(x_1, x_2) = 0$ for all $x_1, x_2 \geq 0$, and

$$z_{n+1}(x_1, x_2) = \frac{1}{\Lambda + \alpha}\left[-cx_2 + \lambda_l \mathcal{T}_1 z_n(x_1, x_2) + \lambda_h \mathcal{T}_2 z_n(x_1, x_2) + \mu \mathcal{T}_3 z_n(x_1, x_2)\right],$$

for $n \geq 0$. Then, $z_n \to z$ as $n \to \infty$. The properties are clearly satisfied for $n = 0$. Lemma 3 in Appendix B shows that if z_n satisfies (58)–(59) then applying the operators $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ preserve these properties, and it is clear from (58)–(59) that these properties are also preserved under summation and multiplication by a constant. Thus, by induction the properties are satisfied for all $n \geq 0$ and hence the result follows.

To prove property (ii), i.e., that the socially optimal thresholds are not greater than the individually optimal threshold, we show that individual optimization leads to joining the queue at all positions where the social planner accepts new low-priority arrivals.

From property (i), we know that the socially optimal balking and abandonment thresholds are equal. Hence, we can express the social welfare function (19) as a function of a single threshold value $\bar{m} \equiv \bar{m}_a = \bar{m}_b$:

$$S(\bar{m}) = \lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) u(i+1, \bar{m}), \tag{60}$$

where $\pi_i(\bar{m})$ is the steady-state probability of having i customers in system when using threshold value $\bar{m}$ and $u(i+1, \bar{m})$ is the expected utility of a customer who joins at position $i+1$ and is removed when she reaches position $\bar{m}+1$. Recall from (10) that the self-optimization threshold satisfies $\bar{n} = \max\{n \geq 0; u(n,n) \geq 0\}$. Thus, to show the result it suffices to show that $S(\bar{m}) - S(\bar{m}-1) \geq 0$ implies $u(\bar{m}, \bar{m}) \geq 0$.

Consider $S(\bar{m})$ in (60) . Using the definition of u in (7) we have

$$\begin{aligned}
S(\bar{m}) - S(\bar{m}-1) &= \lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) \left(R_l q(i+1, \bar{m}+1) - c_l w(i+1, \bar{m}+1)\right) \\
&\quad - \lambda_l \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}-1) \left(R_l q(i+1, \bar{m}) - c_l w(i+1, \bar{m})\right) \\
&= R_l \Big(\lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) q(i+1, \bar{m}+1) - \lambda_l \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}-1) R_l q(i+1, \bar{m})\Big) \\
&\quad - c_l \Big(\lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) w(i+1, \bar{m}+1) - \lambda_l \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}-1) R_l w(i+1, \bar{m})\Big) \\
&= R_l (A_{\bar{m}} - A_{\bar{m}-1}) - c_l (I_{\bar{m}} - I_{\bar{m}-1}),
\end{aligned} \tag{61}$$

where,

$$A_{\bar{m}} \equiv \lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) q(i+1, \bar{m}+1), \tag{62}$$

$$I_{\bar{m}} \equiv \lambda_l \sum_{i=0}^{\bar{m}-1} \pi_i(\bar{m}) w(i+1, \bar{m}+1). \tag{63}$$

It is not hard to verify that $A_{\bar{m}} - A_{\bar{m}-1} > 0$. Therefore, $S(\bar{m}) - S(\bar{m}-1) \geq 0$ iff

$$\frac{R_l}{c_l} \geq \frac{I_{\bar{m}} - I_{\bar{m}-1}}{A_{\bar{m}} - A_{\bar{m}-1}}. \tag{64}$$

Also, note that $u(\bar{m}, \bar{m}) = R_l q(\bar{m}, \bar{m}+1) - c_l w(\bar{m}, \bar{m}+1) \geq 0$ iff

$$\frac{R_l}{c_l} \geq \frac{w(\bar{m}, \bar{m}+1)}{q(\bar{m}, \bar{m}+1)}. \tag{65}$$

To prove that (64) implies (65) we will show that

$$I_{\bar{m}} - I_{\bar{m}-1} \geq (A_{\bar{m}} - A_{\bar{m}-1}) \frac{w(\bar{m}, \bar{m}+1)}{q(\bar{m}, \bar{m}+1)}. \tag{66}$$

First, note that the following identities hold:

$$q(i+1, \bar{m}+1) = q(i+1, \bar{m}) + (1 - q(i+1, \bar{m})) q(\bar{m}, \bar{m}+1), \tag{67}$$

$$w(i+1, \bar{m}+1) = w(i+1, \bar{m}) + (1 - q(i+1, \bar{m})) w(\bar{m}, \bar{m}+1). \tag{68}$$

Note that the identities relate the probability of service and expected waiting time of a customer starting from position $i+1$ and abandoning at position $\bar{m}+1$, to those of a customer starting from the same position but abandoning at position $\bar{m}$. Both identities are easily obtained by conditioning on whether the customer starting from position $i+1$ and abandoning at position $\bar{m}+1$, first reaches position $\bar{m}$ or not. Using (62) and (67) we have

$$\begin{aligned} A_{\bar{m}} - A_{\bar{m}-1} &= \lambda_l \Big(\pi_{\bar{m}-1}(\bar{m}) q(\bar{m}, \bar{m}+1) + \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}) \left(q(i+1, \bar{m}) + (1 - q(i+1, \bar{m})) q(\bar{m}, \bar{m}+1) \right) \\ &\quad - \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}-1) q(i+1, \bar{m}) \Big) \\ &= \lambda_l \Big(\pi_{\bar{m}-1}(\bar{m}) q(\bar{m}, \bar{m}+1) + \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}) \left((1 - q(i+1, \bar{m})) q(\bar{m}, \bar{m}+1) \right) \\ &\quad + \sum_{i=0}^{\bar{m}-2} \left[\pi_i(\bar{m}) - \pi_i(\bar{m}-1) \right] q(i+1, \bar{m}) \Big), \end{aligned} \tag{69}$$

and similarly using (63) and (68) we have

$$\begin{aligned} I_{\bar{m}} - I_{\bar{m}-1} &= \lambda_l \Big(\pi_{\bar{m}-1}(\bar{m}) w(\bar{m}, \bar{m}+1) + \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}) \left(w(i+1, \bar{m}) + (1 - q(i+1, \bar{m})) w(\bar{m}, \bar{m}+1) \right) \\ &\quad - \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}-1) w(i+1, \bar{m}) \Big) \\ &= \lambda_l \Big(\pi_{\bar{m}-1}(\bar{m}) w(\bar{m}, \bar{m}+1) + \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}) \left((1 - q(i+1, \bar{m})) w(\bar{m}, \bar{m}+1) \right) \\ &\quad + \sum_{i=0}^{\bar{m}-2} \left[\pi_i(\bar{m}) - \pi_i(\bar{m}-1) \right] w(i+1, \bar{m}) \Big). \end{aligned} \tag{70}$$

It follows that

$$\begin{aligned} (A_{\bar{m}} - A_{\bar{m}-1}) \frac{w(\bar{m}, \bar{m}+1)}{q(\bar{m}, \bar{m}+1)} &= \lambda_l \Big(\pi_{\bar{m}-1}(\bar{m}) w(\bar{m}, \bar{m}+1) + \sum_{i=0}^{\bar{m}-2} \pi_i(\bar{m}) \left((1 - q(i+1, \bar{m})) w(\bar{m}, \bar{m}+1) \right) \\ &\quad + \sum_{i=0}^{\bar{m}-2} \left[\pi_i(\bar{m}) - \pi_i(\bar{m}-1) \right] q(i+1, \bar{m}) \frac{w(\bar{m}, \bar{m}+1)}{q(\bar{m}, \bar{m}+1)} \Big). \end{aligned} \tag{71}$$

Comparing (71) and (70) and noting that $\pi_i(\bar{m}) - \pi_i(\bar{m}-1) < 0$, it follows that for (66) to hold, it is sufficient to have

$$\frac{w(i+1, \bar{m})}{q(i+1, \bar{m})} \leq \frac{w(\bar{m}, \bar{m}+1)}{q(\bar{m}, \bar{m}+1)},$$

for all $0 \leq i \leq \bar{m} - 2$. From Lemma (4) we know that $w(i+1, \bar{m})/q(i+1, \bar{m})$ is increasing in i and $w(\bar{m}, \bar{m}+1)/q(\bar{m}, \bar{m}+1)$ is increasing in $\bar{m}$. It follows that the above inequality holds and hence the proof is complete. □

Proof of Proposition 5. Consider the constraint in (22). Using the definition of u in (12) and noting that $w(\bar{n}_a, \bar{n}_a+1)/q(\bar{n}_a, \bar{n}_a+1)$ is increasing in $\bar{n}_a$ (see Lemma 4, part (i)), to induce the abandonment threshold $\bar{n}_a$ we must have

$$R_l - c_l \frac{w(\bar{n}_a+1, \bar{n}_a+2)}{q(\bar{n}_a+1, \bar{n}_a+2)} < P_s - P_c \leq R_l - c_l \frac{w(\bar{n}_a, \bar{n}_a+1)}{q(\bar{n}_a, \bar{n}_a+1)}. \tag{72}$$

Similarly, from the constraint (23), given $\bar{n}_a$ the balking threshold $\bar{n}_b$ must satisfy

$$\begin{aligned} (R_l - (P_s - P_c)) q(\bar{n}_b+1, \bar{n}_a+1) - c_l w(\bar{n}_b+1, \bar{n}_a+1) &< P_e + P_c \\ &\leq (R_l - (P_s - P_c)) q(\bar{n}_b, \bar{n}_a+1) - c_l w(\bar{n}_b, \bar{n}_a+1). \end{aligned} \tag{73}$$

Now let $\bar{P}_e \equiv P_e + P_c$ and $\bar{P}_s \equiv P_s - P_c$. Observe that for fixed thresholds the revenue (21) is increasing in $\bar{P}_e$ and (72) is independent of $\bar{P}_e$. Thus, at optimality, it must be that for any $\bar{n}_a$ and $\bar{P}_s$ satisfying (72), and $\bar{n}_b \leq \bar{n}_a$, the fee $\bar{P}_e$ must be set such the second inequality in (72) holds with equality. That is

$$\bar{P}_e = (R - \bar{P}_s) q(\bar{n}_b, \bar{n}_a+1) - c_l w(\bar{n}_b, \bar{n}_a+1). \tag{74}$$

Substituting from (74) into the revenue function (21) we get

$$\begin{aligned}\Pi(\bar{n}_b,\bar{n}_a) &= \lambda_l\left[(R_l-\bar{P}_s)q(\bar{n}_b,\bar{n}_a+1)-c_l w(\bar{n}_b,\bar{n}_a+1)\right]q_J(\bar{n}_b,\bar{n}_a)+\lambda_l\bar{P}_s q_S(\bar{n}_b,\bar{n}_a)\\ &= \lambda_l R_l q(\bar{n}_b,\bar{n}_a+1)q_J(\bar{n}_b,\bar{n}_a)+\lambda_l\bar{P}_s\left[q_S(\bar{n}_b,\bar{n}_a)-q(\bar{n}_b,\bar{n}_a+1)q_J(\bar{n}_b,\bar{n}_a)\right]\\ &\quad -\lambda_l c_l w(\bar{n}_b,\bar{n}_a+1)q_J(\bar{n}_b,\bar{n}_a). \end{aligned} \tag{75}$$

Now observe that

$$q_S(\bar{n}_b,\bar{n}_a)=\sum_{i=0}^{\bar{n}_b-1}\pi_i q(i+1,\bar{n}_a+1)\geq\sum_{i=0}^{\bar{n}_b-1}\pi_i q(\bar{n}_b,\bar{n}_a+1)=q(\bar{n}_b,\bar{n}_a+1)q_J(\bar{n}_b,\bar{n}_a),$$

where the inequality follows from the fact that $q(i,\bar{n}_a+1)$, defined in (1) is decreasing in i. It follows from (75) that the revenue increases in $\bar{P}_s$, so this fee is set such that the second inequality in (72) is binding, that is $\bar{P}_s=R_l-c_l w(\bar{n}_a,\bar{n}_a+1)/q(\bar{n}_a,\bar{n}_a+1)$. □

Proof of Proposition (7). We compare the optimal revenues under entrance fee and under service fee as functions of λ_h, when λ_h is in the neighborhood of zero. To do so we compare the right derivatives of these optimal revenue functions with respect to λ_h in the limit as $\lambda_h\to 0^+$. Then since both optimal revenue functions are decreasing in λ_h and have the same value at $\lambda_h=0$ (because in that case the system reduces to a single-class FCFS queue) the one with the larger derivative generates more revenue for sufficiently small λ_h.

Denote by $\bar{n}^*$ the balking and abandonment threshold under the optimal service fee, and by $\bar{n}_b^*$ and $\bar{n}_a^*$, respectively, the balking and abandonment thresholds under the optimal entrance fee. Further, denote the optimal revenue function under service fee by Π_s^*, and under entrance fee by Π_e^*. For clarity, here we state the optimal thresholds and the optimal revenues as functions of λ_h, i.e., we denote the thresholds by $\bar{n}^*(\lambda_h),\bar{n}_b^*(\lambda_h),\bar{n}_a^*(\lambda_h)$ and the optimal revenue functions by $\Pi_s^*(\lambda_h)=\Pi_s(\bar{n}^*(\lambda_h),\lambda_h)$ and $\Pi_e^*(\lambda_h)=\Pi_e(\bar{n}_a^*(\lambda_h),\bar{n}_b^*(\lambda_h),\lambda_h)$. It can be shown that the optimal thresholds are piecewise constant, right-continuous functions of λ_h. Hence, for sufficiently small λ_h, the optimal thresholds are constant and equal to their values at $\lambda_h=0$. It follows from (27) that

$$\begin{aligned}\lim_{\lambda_h\to 0^+}\frac{d\Pi_s^*(\lambda_h)}{d\lambda_h} &= \lim_{\lambda_h\to 0^+}\frac{\partial\Pi_s(\bar{n}^*(0),\lambda_h)}{\partial\lambda_h}\\ &= \lim_{\lambda_h\to 0^+}\frac{\partial}{\partial\lambda_h}\left[\lambda_l\left(R_l-c_l\frac{w(\bar{n}^*(0),\bar{n}^*(0)+1)}{q(\bar{n}^*(0),\bar{n}^*(0)+1)}\right)q_S(\bar{n}^*(0),\bar{n}^*(0))\right],\end{aligned} \tag{76}$$

and from (28) that

$$\begin{aligned}\lim_{\lambda_h\to 0^+}\frac{d\Pi_e^*(\lambda_h)}{d\lambda_h} &= \lim_{\lambda_h\to 0^+}\frac{\partial\Pi_e(\bar{n}_a^*(0),\bar{n}_b^*(0),\lambda_h)}{\partial\lambda_h}\\ &= \lim_{\lambda_h\to 0^+}\frac{\partial}{\partial\lambda_h}\left[\lambda_l\left(R_l q(\bar{n}_b^*(0),\bar{n}_a^*(0)+1)-c_l w(\bar{n}_b^*(0),\bar{n}_a^*(0)+1)\right)q_J(\bar{n}_b^*(0),\bar{n}_a^*(0))\right].\end{aligned} \tag{77}$$

Note that for $\lambda_h=0$, we have using (1) that $q(n,n+1)=1$, using (2) that $w(n,n+1)=n/\mu$, and using (17) and (18) that $q_S=q_J=(1-\rho_l^n)/(1-\rho_l^{n+1})$. Therefore, (27) and (28) yield

$$\hat{n}^*\equiv\bar{n}^*(0)=\bar{n}_b^*(0)=\text{argmax}_{n\geq 1}\lambda_l(R_l-c_l n/\mu)\frac{1-\rho_l^n}{1-\rho_l^{n+1}}, \tag{78}$$

$$\check{n}^*\equiv\bar{n}_a^*(0)=\max\{n\geq 0;R_l-c_l n/\mu\geq 0\}, \tag{79}$$

where $\rho_l \equiv \lambda_l/\mu$. Observe that (78) and (79) are respectively the revenue maximizing and self-optimization thresholds in the single class FCFS problem of Naor (1969) and hence $\hat{n}^* \leq \check{n}^*$.

We proceed by computing the derivatives in (76) and (77). With slight abuse of notation, we write $f' \equiv \partial f/\partial \lambda_h$ for the partial derivatives of these functions with respect to λ_h. As $\lambda_h \to 0^+$, the following limits hold:

$$w'(n, n+1) \to -1/\mu, \quad q'(n, n+1) \to -1, \tag{80}$$

$$q'_S(n, n) \to -\frac{\rho_l^{n-1}(\rho_l^{n+2} - (n+2)\rho_l + (n+1))}{(1-\rho_l^{n+1})^2}, \tag{81}$$

$$q'_J(n, n) \to -\frac{\rho_l^{n-1}(\rho_l^{n+1} - (n+1)\rho_l + n)}{(1-\rho_l^{n+1})^2}. \tag{82}$$

Moreover, for $n_b \leq n_a$, as $\lambda_h \to 0^+$,

$$w(n_b, n_a+1) \to n_b/\mu, \quad q(n_b, n_a+1) \to 1, \quad q_S(n_b, n_a) \to \frac{1-\rho_l^{n_b}}{1-\rho_l^{n_b+1}}, \quad q_J(n_b, n_a) \to \frac{1-\rho_l^{n_b}}{1-\rho_l^{n_b+1}}, \tag{83}$$

and for $n_b < n_a$, as $\lambda_h \to 0^+$,

$$w'(n_b, n_a+1) \to n_b/\mu, \quad q'(n_b, n_a+1) \to 0, \quad q'_J(n_b, n_a) \to -\frac{\rho_l^{n_b-1}(\rho_l^{n_b+2} - (n_b+1)\rho_l + n_b)}{(1-\rho_l^{n_b+1})^2}. \tag{84}$$

Computing the derivatives in (76) and substituting the limits from (80), (81), and (83) we have

$$\begin{aligned}
\lim_{\lambda_h \to 0^+} \frac{d\Pi_s^*(\lambda_h)}{d\lambda_h} &= \lim_{\lambda_h \to 0^+} \lambda_l \left[\left(-c_l \frac{w'(\hat{n}^*, \hat{n}^*+1) q(\hat{n}^*, \hat{n}^*+1) - w(\hat{n}^*, \hat{n}^*+1) q'(\hat{n}^*, \hat{n}^*+1)}{q(\hat{n}^*, \hat{n}^*+1)^2} \right) q_S(\hat{n}^*, \hat{n}^*) \right. \\
&\quad \left. + \left(R_l - c_l \frac{w(\hat{n}^*, \hat{n}^*+1)}{q(\hat{n}^*, \hat{n}^*+1)} \right) q'_S(\hat{n}^*, \hat{n}^*) \right] \\
&= \lambda_l \left[\left(-c_l \frac{\hat{n}^*-1}{\mu} \right) \frac{1-\rho_l^{\hat{n}^*}}{1-\rho_l^{\hat{n}^*+1}} - (R_l - c_l \hat{n}^*/\mu) \frac{\rho_l^{\hat{n}^*-1}(\rho_l^{\hat{n}^*+2} - (\hat{n}^*+2)\rho_l + (\hat{n}^*+1))}{(1-\rho_l^{\hat{n}^*+1})^2} \right].
\end{aligned} \tag{85}$$

For (77) we need to consider two cases depending on whether $\hat{n}^* < \check{n}^*$ or $\hat{n}^* = \check{n}^*$. Computing the derivatives in (77) and substituting the limits from (83) and (84) for $\hat{n}^* < \check{n}^*$, and from (80), (82) and (83) for $\hat{n}^* = \check{n}^*$, we get

$$\begin{aligned}
\lim_{\lambda_h \to 0^+} \frac{d\Pi_e^*(\lambda_h)}{d\lambda_h} &= \lim_{\lambda_h \to 0^+} \lambda_l \left[(R_l q'(\hat{n}^*, \check{n}^*+1) - c_l w'(\hat{n}^*, \check{n}^*+1)) \, q_J(\hat{n}^*, \check{n}^*) \right. \\
&\quad \left. + (R_l q(\hat{n}^*, \check{n}^*+1) - c_l w(\hat{n}^*, \check{n}^*+1)) \, q'_J(\hat{n}^*, \check{n}^*) \right] \\
&= \begin{cases} \lambda_l \left[(-c_l \hat{n}^*/\mu) \frac{1-\rho_l^{\hat{n}^*}}{1-\rho_l^{\hat{n}^*+1}} - (R_l - c_l \hat{n}^*/\mu) \frac{\rho_l^{\hat{n}^*-1}(\rho_l^{\hat{n}^*+2} - (\hat{n}^*+1)\rho_l + \hat{n}^*)}{(1-\rho_l^{\hat{n}^*+1})^2} \right], & \text{if } \hat{n}^* < \check{n}^*, \\ \lambda_l \left[(-R_l + c_l \hat{n}^*/\mu) \frac{1-\rho_l^{\hat{n}^*}}{1-\rho_l^{\hat{n}^*+1}} - (R_l - c_l \hat{n}^*/\mu) \frac{\rho_l^{\hat{n}^*-1}(\rho_l^{\hat{n}^*+1} - (\hat{n}^*+1)\rho_l + \hat{n}^*)}{(1-\rho_l^{\hat{n}^*+1})^2} \right], & \text{if } \hat{n}^* = \check{n}^*. \end{cases}
\end{aligned} \tag{86}$$

Using (85) and (86) after some algebra we can write

$$\lim_{\lambda_h \to 0^+} \left(\frac{d\Pi_e^*(\lambda_h)}{d\lambda_h} - \frac{d\Pi_s^*(\lambda_h)}{d\lambda_h} \right) = \begin{cases} \lambda_l \left[(R_l - c_l \hat{n}^*/\mu)\, \rho_l^{\hat{n}^*-1} \left(\frac{1-\rho_l}{1-\rho_l^{\hat{n}^*+1}} \right)^2 - (c_l/\mu) \frac{1-\rho_l^{\hat{n}^*}}{1-\rho_l^{\hat{n}^*+1}} \right], & \text{if } \hat{n}^* < \check{n}^*, \\ \lambda_l \left(R_l - c_l \hat{n}^*/\mu \right) \left(\frac{1-\rho_l^{\hat{n}^*} + \rho_l^{\hat{n}^*-1}(1-\rho_l)}{1-\rho_l^{\hat{n}^*+1}} \right), & \text{if } \hat{n}^* = \check{n}^*. \end{cases} \tag{87}$$

Case 1: For $\hat{n}^* < \check{n}^*$ (which is typically the case) the entrance fee is optimal iff

$$(R_l - c_l \hat{n}^*/\mu)\, \rho_l^{\hat{n}^*-1} \left(\frac{1-\rho_l}{1-\rho_l^{\hat{n}^*+1}} \right)^2 - (c_l/\mu) \frac{1-\rho_l^{\hat{n}^*}}{1-\rho_l^{\hat{n}^*+1}} > 0. \tag{88}$$

Case 2: For $\hat{n}^* = \check{n}^*$, it is easy to verify that for any ρ_l, the term $\left(1-\rho_l^{\hat{n}^*} + \rho_l^{\hat{n}^*-1}(1-\rho_l)\right)/(1-\rho_l^{\hat{n}^*+1})$ in (87) is positive. Thus, the entrance fee is optimal.

In the following we verify the statements $(i)-(iii)$.

(i) Let $\lambda_l \to 0^+$, then using (78) we find that $\hat{n}^* = 1$. Hence, (88) reduces to $(R_l - c_l/\mu) - c_l/\mu > 0$ or $R_l\mu/c_l > 2$. Also note that assuming $R_l\mu/c_l > 2$, we have $\check{n}^* \geq 2$. Hence, we are in Case 1 and the entrance fee is optimal.

(ii) Let $\rho_l \to 1^+$, then we have $(1-\rho_l)/(1-\rho_l^{\hat{n}^*+1}) \to 1/(1+\hat{n}^*)$ and $(1-\rho_l^{\hat{n}^*})/(1-\rho_l^{\hat{n}^*+1}) \to \hat{n}^*/(1+\hat{n}^*)$. Thus, (88) reduces to $(R_l - c_l\hat{n}^*)/(1+\hat{n}^*)^2 > c_l(\hat{n}^*/(1+\hat{n}^*))$ which simplifies to $R_l/c_l > (\hat{n}^*+1)^2 - 1$. In both cases the sufficient conditions for the entrance fee to be optimal are satisfied. Further, for $\rho_l \to 1^+$, $\hat{n}^*$ is the maximizer of $f(n) \equiv \lambda_l(R_l - c_l n/\mu)(n/(1+n))$. Therefore, it can be obtained by finding the positive root of $f(n) - f(n-1) = 0$ and applying the floor function to it, which yields $\hat{n}^* = \left\lfloor \frac{1}{2}(\sqrt{5 + 4R_l\mu/c_l} - 1) \right\rfloor$.

(iii) Let $\lambda_l \to \infty$, then it is easy to verify that the problem in (78) reduces to $\text{argmax}_{n\geq 1}\mu(R_l - c_l n/\mu)$ and thus $\hat{n}^* = 1$. Assuming that $R_l\mu/c_l \geq 2$, we have $\check{n}^* \geq 2$ and hence we are in Case 1. Setting $\hat{n}^* = 1$, the lhs of (88) simplifies to $(R_l - c_l/\mu)\left(\frac{1}{1+\rho_l}\right)^2 - (c_l/\mu)\frac{1}{1+\rho_l}$, which approaches zero from below as $\lambda_l \to \infty$, proving that the service fee is optimal. $\square$

Appendix B: Additional Lemmas

LEMMA 1. *Fix $i \geq 1$, $T \geq i$, and $v(i-1,i-1) \geq 0$. The value function*

$$v(i,i) = [-\bar{c} + \bar{\lambda} v(i,i+1) + \bar{\mu} v(i-1,i-1)]^+, \tag{89}$$

$$v(i,n) = [-\bar{c} + \bar{\lambda} v(i,n+1) + \bar{\mu} v(i,n-1)]^+, \quad n \in \{i+1,...,T\}, \tag{90}$$

$$v(i,T+1) = \begin{cases} [-\bar{c} + \bar{\lambda} v(i,T+1) + \bar{\mu} v(i,T)]^+, & T = \bar{n}_h + i - 1, \\ [-\bar{c} + \bar{\mu} v(i,T)]^+, & T < \bar{n}_h + i - 1, \end{cases} \tag{91}$$

satisfies the following properties:

(i) $v(i,n+1) \leq v(i,n) \leq v(i-1,i-1)$ for $n \in \{i,\ldots,T\}$.

(ii) If $v(i-1,i-1) < c_l/\mu$, customers prefer balking to joining in low-priority position i. If $v(i-1,i-1) \geq c_l/\mu$ then a customer joins/stays in low-priority position i if and only if her system position $n \leq \bar{n}(i)$, where

$$\bar{n}(i) = i \text{ if } v(i,i) < c_l/\mu \text{ and } \bar{n}(i) = max\{n \in \{i+1,...,T+1\}; v(i,n-1) \geq c_l/\mu\} \text{ otherwise.} \tag{92}$$

(iii) If $v(i-1,i-1) \geq c_l/\mu$ then

$$v(i,i) \leq v(i-1,i-1) - c_l/\mu \text{ and } v(i,n) \leq v(i,n-1) - c_l/\mu \text{ for } i < n \leq \bar{n}(i). \tag{93}$$

Proof of Lemma 1. (i) By induction on the steps of the value iteration algorithm. Let $v_0(i,n) = 0$ for $n \in \{i,\ldots,T\}$. For $k \geq 0$ set

$$v_{k+1}(i,i) = [-\bar{c} + \bar{\lambda} v_k(i,i+1) + \bar{\mu} v(i-1,i-1)]^+, \tag{94}$$

$$v_{k+1}(i,n) = [-\bar{c} + \bar{\lambda} v_k(i,n+1) + \bar{\mu} v_k(i,n-1)]^+, \quad n \in \{i+1,\ldots,T\}. \tag{95}$$

$$v_{k+1}(i,T+1) = \begin{cases} [-\bar{c} + \bar{\lambda} v_k(i,T+1) + \bar{\mu} v_k(i,T)]^+, & T = \bar{n}_h + i - 1, \\ [-\bar{c} + \bar{\mu} v_k(i,T)]^+, & T < \bar{n}_h + i - 1. \end{cases} \tag{96}$$

From (89)–(91) it follows that $v_k(i,n) \to v(i,n)$ as $k \to \infty$ by the convergence of the value iteration algorithm. The proof is complete if

$$v_k(i,n+1) \leq v_k(i,n) \leq v(i-1,i-1), \quad n \in \{i,\ldots,T\}, \tag{97}$$

for $k \geq 0$. This clearly holds for $k = 0$. Next, if (97) holds for some k then by using (94)–(97) and noting that $\bar{\lambda} + \bar{\mu} = 1$, it is straightforward to verify that $v_{k+1}(i,n+1) \leq v_{k+1}(i,n) \leq v(i-1,i-1)$ for $n \in \{i,\ldots,T\}$.

(ii) Recall that low-priority customers join/stay in position (i,n) if and only if the expected utility of doing so is nonnegative. The expressions inside the brackets on the right-hand sides of (89)–(91) represent these expected utilities, for system positions $n \in \{i,\ldots,T+1\}$. From (89) the

expected utility of joining/staying in position (i,i) is nonnegative if and only if $v(i,i) = \bar{\lambda} v(i,i+1) + \bar{\mu}\,(v(i-1,i-1) - \bar{c}/\bar{\mu}) \geq 0$. This holds if and only if $v\,(i-1,i-1) \geq \bar{c}/\bar{\mu}$ (where $\bar{c}/\bar{\mu} = c_l/\mu$), because $v(i,i)$ is a convex combination of $v(i,i+1)$ and $v\,(i-1,i-1) - \bar{c}/\bar{\mu}$, and we have $v(i,i+1) \leq v(i,i)$ from part (i). It follows similarly from (90)–(91) that for $n \in \{i+1,\ldots,T\}$ the expected utility of joining/staying in position (i,n) is nonnegative if and only if $v(i,n-1) \geq c_l/\mu$. These conditions, together with part (i), establish (92).

(iii) We have for position (i,i) that

$$v(i,i) = \bar{\lambda} v(i,i+1) + \bar{\mu}(v(i-1,i-1) - c_l/\mu) \leq v\,(i-1,i-1) - c_l/\mu, \tag{98}$$

where the equality follows from (89) and because by part (ii) customers have nonnegative expected utility of joining/staying in system position i. The inequality holds because $v(i,i)$ is a convex combination of $v(i,i+1)$ and $v(i-1,i-1) - c_l/\mu$, the latter term is nonnegative by hypothesis, and $v(i,i+1) \leq v(i,i)$ by part (i). Similarly, for positions (i,n) with $n \in \{i+1,\ldots,\bar{n}(i)\}$ we have

$$v(i,n) = \bar{\lambda} v(i,n+1) + \bar{\mu}(v(i,n-1) - c_l/\mu) \leq v(i,n-1) - c_l/\mu, \tag{99}$$

where the equality follows from (90)–(91) and because by part (ii) customers have nonnegative expected utility of joining/staying in system positions $n \in \{i,\ldots,\bar{n}(i)\}$. The inequality holds because $v(i,n)$ is a convex combination of $v(i,n+1)$ and $v(i,n-1) - c_l/\mu$, the latter term is nonnegative for $n \in \{i+1,\ldots,\bar{n}(i)\}$ by (92), and $v(i,n+1) \leq v(i,n)$ by part (ii). □

Lemma 2. *Consider the value functions $v(m,n)$ and $v(m+1,n)$ satisfying equations* (37)–(39) *and* (42)–(44)*, respectively. We have*

(i) *If $\bar{n}(m) = \bar{n}_h + m$, then $v(m+1,n) \leq v(m,n)$ for $n \in \{m+1,\ldots,\bar{n}_h + m\}$.*

(ii) *If $\bar{n}(m) < \bar{n}_h + m$, then $v(m+1,n) = v(m,n)$ for $n \in \{m+1,\ldots,\bar{n}(m)+1\}$.*

Proof of Lemma 2. (i) The proof is by induction and convergence of the value iteration algorithm. Note that since $\bar{n}(m) = \bar{n}_h + m$, the value of staying at all system positions is nonnegative for the customer at low-priority position m. Therefore, using (37)–(39), $v(m,n)$ satisfies

$$\begin{aligned} v(m,m) &= -\bar{c} + \bar{\lambda} v(m,m+1) + \bar{\mu} v(m-1,m-1), \\ v(m,n) &= -\bar{c} + \bar{\lambda} v(m,n+1) + \bar{\mu} v(m,n-1), \quad n \in \{m+1,...,\bar{n}_h + m - 1\}, \\ v(m,\bar{n}_h + m) &= -\bar{c} + \bar{\lambda} v(m,\bar{n}_h + m) + \bar{\mu} v(m,\bar{n}_h + m - 1). \end{aligned}$$

Next we approximate the value function for the customer at low-priority position $m+1$ given in (42)-(44) using the recursion

$$\begin{aligned} v_{k+1}(m+1,m+1) &= [-\bar{c} + \bar{\lambda} v_k(m+1,m+2) + \bar{\mu} v(m,m)]^+, \\ v_{k+1}(m+1,n) &= [-\bar{c} + \bar{\lambda} v_k(m+1,n+1) + \bar{\mu} v_k(m+1,n-1)]^+, \quad n \in \{m+2,...,\bar{n}_h + m\}, \\ v_{k+1}(m+1,\bar{n}_h + m + 1) &= [-\bar{c} + \bar{\lambda} v(m+1,\bar{n}_h + m + 1) + \bar{\mu} v(m+1,\bar{n}_h + m)]^+, \end{aligned}$$

for $k \geq 0$ and with $v_0(m+1,n) = 0$ for $n \in \{m+1, ..., \bar{n}_h + m + 1\}$. The claim is clearly satisfied for $k = 0$. Assuming

$$v_k(m+1,n) \leq v(m,n), \tag{100}$$

we show that $v_{k+1}(m+1,n) \leq v(m,n)$ for $n \in \{m+1, ..., \bar{n}_h + m\}$ and hence the result follows by induction. For $n = m+1$ the corresponding equations are

$$\begin{aligned} v(m,m+1) &= -\bar{c} + \bar{\lambda} v(m,m+2) + \bar{\mu} v(m,m), \\ v_{k+1}(m+1,m+1) &= [-\bar{c} + \bar{\lambda} v_k(m+1,m+2) + \bar{\mu} v(m,m)]^+, \end{aligned}$$

and for $n \in \{m+2, ..., \bar{n}_h + m - 1\}$, we have

$$\begin{aligned} v(m,n) &= -\bar{c} + \bar{\lambda} v(m,n+1) + \bar{\mu} v(m,n-1), \\ v_{k+1}(m+1,n) &= [-\bar{c} + \bar{\lambda} v_k(m+1,n+1) + \bar{\mu} v_k(m+1,n-1)]^+. \end{aligned}$$

In each case, comparing the two equations and using (100) the claim directly follows. For $n = \bar{n}_h + m$, we have

$$\begin{aligned} v(m,\bar{n}_h + m) &= -\bar{c} + \bar{\lambda} v(m,\bar{n}_h + m) + \bar{\mu} v(m,\bar{n}_h + m - 1), \\ v_{k+1}(m+1,\bar{n}_h + m) &= [-\bar{c} + \bar{\lambda} v_k(m+1,\bar{n}_h + m + 1) + \bar{\mu} v_k(m+1,\bar{n}_h + m - 1)]^+. \end{aligned}$$

Noting that $v_k(m+1,\bar{n}_h + m + 1) \leq v_k(m+1,\bar{n}_h + m)$ (see the proof of Lemma (1) part (i)) and using (100) the claim follows.

(ii) For $\bar{n}(m) < \bar{n}_h + m$, the result follows because the solution of (42)–(44) is unique and the function that satisfies (37)–(39) also satisfies (42)–(44) when setting $v(m+1,n) = v(m,n)$ for $n \in \{m+1, \ldots, \bar{n}(m) + 1\}$. □

Lemma 3. *Let $\mathcal{V}$ denote the set of functions satisfying* (58) *and* (59)*. If $z \in \mathcal{V}$, then* (i) *$\mathcal{T}_1 z \in \mathcal{V}$,* (ii) *$\mathcal{T}_2 z \in \mathcal{V}$, and* (iii) *$\mathcal{T}_3 z \in \mathcal{V}$.*

Proof of Lemma 3. (i) First, note that

$$\mathcal{T}_1 z(x_1,x_2) = z(x_1,x_2) + [z(x_1,x_2+1) - z(x_1,x_2)]^+. \tag{101}$$

Let

$$d(x_1,x_2) \equiv z(x_1,x_2+1) - z(x_1,x_2), \tag{102}$$

and note that $d(x_1,x_2)$ is decreasing in x_2 since z is concave in x_2. Using (101) we have

$$\begin{aligned} \mathcal{T}_1 z(x_1,x_2+1) - \mathcal{T}_1 z(x_1,x_2) &= z(x_1,x_2+1) - z(x_1,x_2) \\ &\quad + [z(x_1,x_2+2) - z(x_1,x_2+1)]^+ - [z(x_1,x_2+1) - z(x_1,x_2)]^+ \\ &= d(x_1,x_2) - [d(x_1,x_2)]^+ + [d(x_1,x_2+1)]^+, \end{aligned}$$

which is decreasing in x_2, establishing property (58) for $\mathcal{T}_1$. Next, to show that (59) holds for $\mathcal{T}_1$, using the definition in (102) we write (59) as

$$d(x_1+1,x_2) = d(x_1,x_2+1). \tag{103}$$

Now observe that

$$\begin{aligned}\mathcal{T}_1 z(x_1+1,x_2+1) - \mathcal{T}_1 z(x_1+1,x_2) &= d(x_1+1,x_2) - [d(x_1+1,x_2)]^+ + [d(x_1+1,x_2+1)]^+ \\ &= d(x_1,x_2+1) - [d(x_1,x_2+1)]^+ + [d(x_1,x_2+2)]^+ \\ &= \mathcal{T}_1 z(x_1,x_2+2) - \mathcal{T}_1 z(x_1,x_2+1),\end{aligned}$$

where the second equality follows from (103).

(*ii*) First, note that

$$\mathcal{T}_2 z(x_1,x_2) = x(x_1+1,0) + \sum_{j=0}^{x_2-1} [d(x_1+1,j)]^+,$$

and therefore,

$$\mathcal{T}_2 z(x_1,x_2+1) - \mathcal{T}_2 z(x_1,x_2) = [d(x_1+1,x_2)]^+, \tag{104}$$

which is decreasing in x_2, establishing property (58) for $\mathcal{T}_2$. Next, using (103) and (104) we have

$$\begin{aligned}\mathcal{T}_2 z(x_1+1,x_2+1) - \mathcal{T}_2 z(x_1+1,x_2) &= [d(x_1+2,x_2)]^+ \\ &= [d(x_1+1,x_2+1)]^+ \\ &= \mathcal{T}_2 z(x_1,x_2+2) - \mathcal{T}_2 z(x_1,x_2+1),\end{aligned}$$

which proves property (59) for $\mathcal{T}_2$.

(*iii*) We consider each of the three cases of (57) in turn.

(a) If $x_1 > 0$, then $\mathcal{T}_3 z(x_1,x_2) = z(x_1-1,x_2)$ by (57) and it is immediate that $\mathcal{T}_3$ preserves (58)–(59).

(b) If $x_1 = 0, x_2 > 0$, then it follows that $\mathcal{T}_3$ preserves (58) because by (57) we have

$$\begin{aligned}\mathcal{T}_3 z(0,x_2+2) - \mathcal{T}_3 z(0,x_2+1) &= [R_l + z(0,x_2+1)] - [R_l + z(0,x_2)] \\ &\le z(0,x_2) - z(0,x_2-1) \\ &= \mathcal{T}_3 z(0,x_2+1) - \mathcal{T}_3 z(0,x_2),\end{aligned}$$

where the inequality holds by concavity, and $\mathcal{T}_3$ preserves (59) because by (57) we have

$$\begin{aligned}\mathcal{T}_3 z(1,x_2+1) - \mathcal{T}_3 z(1,x_2) &= z(0,x_2+1) - z(0,x_2) \\ &= [R_l + z(0,x_2+1)] - [R_l + z(0,x_2)] \\ &= \mathcal{T}_3 z(0,x_2+2) - \mathcal{T}_3 z(0,x_2+1).\end{aligned}$$

(c) Finally, if $x_1 = x_2 = 0$, it follows that $\mathcal{T}_3$ preserves (58) because by (57) we have

$$\begin{aligned}\mathcal{T}_3 z(0,2) - \mathcal{T}_3 z(0,1) &= [R_l + z(0,1)] - [R_l + z(0,0)] \\ &\le R_l + z(0,0) - z(0,0) \\ &= \mathcal{T}_3 z(0,1) - \mathcal{T}_3 z(0,0),\end{aligned}$$

where the inequality follows because $z(0,1) - z(0,0) \le R_l$, that is, the value of an additional low-priority customer is bounded by R_l. Furthermore, $\mathcal{T}_3$ preserves (59) because by (57) we have

$$\begin{aligned}\mathcal{T}_3 z(1,1) - \mathcal{T}_3 z(1,0) &= z(0,1) - z(0,0) \\ &= [R_l + z(0,1)] - [R_l + z(0,0)] \\ &= \mathcal{T}_3 z(0,2) - \mathcal{T}_3 z(0,1).\end{aligned}$$

The proof is complete. □

Lemma 4. *Consider w and q given in* (1) *and* (2)*, respectively. We have*

(*i*) *$w(\bar{n}, \bar{n}+1)/q(\bar{n}, \bar{n}+1)$ is increasing in $\bar{n}$ for $\bar{n} \ge 1$.*

(*ii*) *$w(i, \bar{n}+1)/q(i, \bar{n}+1)$ is increasing in i for $0 \le i \le \bar{n}$.*

Proof of Lemma 4. (*i*) Using (1) and (2) we have

$$\begin{aligned}\frac{w(\bar{n}, \bar{n}+1)}{q(\bar{n}, \bar{n}+1)} &= \frac{\bar{n}(1-\rho_h^{\bar{n}+1}) - (\bar{n}+1)(\rho_h - \rho_h^{\bar{n}+1})}{\mu(1-\rho_h)^2}. \\ &= \frac{\bar{n}(1-\rho_h) - \rho_h(1-\rho_h^{\bar{n}})}{\mu(1-\rho_h)^2} = \frac{1}{\mu(1-\rho_h)} \sum_{i=1}^{\bar{n}} (1-\rho_h^i),\end{aligned}$$

which is clearly increasing in $\bar{n}$.

(*ii*) Using (1) and (2) after simplifying we have

$$\frac{w(i, \bar{n}+1)}{q(i, \bar{n}+1)} = \frac{i(1-\rho_h^{\bar{n}+1}) - (\bar{n}+1)(\rho_h^{\bar{n}+1-i} - \rho_h^{\bar{n}+1})}{\mu(1-\rho_h)(1-\rho_h^{\bar{n}+1-i})}.$$

Adding and subtracting $\bar{n}+1$ in the numerator and simplifying we get

$$\begin{aligned}\frac{w(i, \bar{n}+1)}{q(i, \bar{n}+1)} &= \frac{i(1-\rho_h^{\bar{n}+1}) - (\bar{n}+1)(\rho_h^{\bar{n}+1-i} - \rho_h^{\bar{n}+1}) + (\bar{n}+1) - (\bar{n}+1)}{\mu(1-\rho_h)(1-\rho_h^{\bar{n}+1-i})}. \\ &= \frac{i(1-\rho_h^{\bar{n}+1}) + (\bar{n}+1)(1-\rho_h^{\bar{n}+1-i}) - (\bar{n}+1)(1-\rho_h^{\bar{n}+1})}{\mu(1-\rho_h)(1-\rho_h^{\bar{n}+1-i})} \\ &= \frac{(\bar{n}+1)(1-\rho_h^{\bar{n}+1-i}) - (\bar{n}+1-i)(1-\rho_h^{\bar{n}+1})}{\mu(1-\rho_h)(1-\rho_h^{\bar{n}+1-i})} \\ &= \frac{\bar{n}+1}{\mu(1-\rho_h)} - \frac{(1-\rho_h^{\bar{n}+1})(\bar{n}+1-i)}{\mu(1-\rho_h)(1-\rho_h^{\bar{n}+1-i})}.\end{aligned}$$

Hence, noting that $\rho_h < 1$, it suffices to show that $(\bar{n}+1-i)/(1-\rho_h^{\bar{n}+1-i})$ is decreasing in i. To this end, let $\nu \equiv \bar{n}+1-i$ and note that ν decreases as i increases. Thus, it remains to show that $\nu/(1-\rho_h^{\nu})$ is increasing in ν. The first difference is

$$\frac{\nu+1}{1-\rho_h^{\nu+1}} - \frac{\nu}{1-\rho_h^{\nu}} = \frac{(1-\rho_h^{\nu}) - (1-\rho_h)\nu\rho_h^{\nu}}{(1-\rho_h^{\nu+1})(1-\rho_h^{\nu})}.$$

The denominator is clearly positive. To show the nominator is also positive, consider $r(x) \equiv x^{\nu}$. By the mean value theorem,

$$\frac{1-\rho_h^{\nu}}{1-\rho_h} = \frac{r(1)-r(\nu)}{1-\nu} = r'(c_0),$$

for some $\rho_h < c_0 < 1$. But $r'(c_0) = \nu c_0^{\nu-1} > \nu c_0^{\nu} > \nu\rho_h^{\nu}$, implying that

$$\frac{1-\rho_h^{\nu}}{1-\rho_h} > \nu\rho_h^{\nu},$$

and hence the nominator is indeed positive and the proof is complete. $\square$