Dimensioning a queue with state-dependent arrival rates

Benjamin Legros

Ecole de Management de Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France

benjamin.legros@centraliens.net

Abstract

In an observable queue, customers joining decisions may be influenced by wait-aversion and crowd-attraction. These opposing phenomena and the diversity of arriving customers lead to an arrival process that depends on the number of present customers. For the system manager, having more customers may be beneficial as it can increase future arrivals due to the attraction generated. It may also saturate the system, resulting in long waits. Rejection at arrival may then be employed as a way to obtain a trade-off between these conflicting objectives. With this in mind, we developed a Markov decision process approach to determine how to optimally reject customers in a queueing system with state-dependent arrivals.

When the arrival rate is bounded, we compute the optimal policy from a value iteration approach. When the arrival rate is decreasing and convex, we prove that it has a threshold form. When the arrival rate is increasing and potentially unbounded, uniformization may not apply. In dealing with this case, we restrict the analysis to stationary policies and prove the optimality of threshold policies from a computational approach. In addition, we show how to compute the optimal threshold within a finite number of iterations and prove that the long-run expected cost is decreasing and convex in the number of servers. We finally illustrate the applicability of our results through the analysis of a linearly increasing arrival rate, determining the main drivers of control decisions.

Keywords: Queueing; admission control; state-dependent arrivals; threshold policy.

1 Introduction

Understanding customer behavior is an important step in modelling a queueing system. In particular, the perception of congestion influences a customer's decision on whether to join a system. When the service is unknown, congestion may be seen as an indication of quality. The empty restaurant syndrome is the most famous example. A restaurant with mostly empty tables gives the idea of low quality. The underlying psychology of empty restaurant syndrome is simple: if so few people come to this restaurant, it cannot be good. The dynamic is based on uninformed consumers taking their cues from informed ones. This behavior of *congestion-attraction* is also called *herding behavior*. It can be partially explained by the rational lack of knowledge about the quality of a service, but it may also be driven by some aspects of human psychology.

Academic research has focused on understanding herding behavior. One major domain of application of this behavior is Finance (Bacry et al., 2013; Da Fonseca and Zaatour, 2014; Rambaldi et al., 2017). In

such studies, crowd-attraction is used to capture the contagious nature of financial activity. Hawkes (1971) developed a stochastic process to capture the effect of past events on present arrivals. From an operations management orientated perspective, Debo and Veeraraghavan (2009) provides an overview of crowd-attraction models, in which the idea of unknown service is prevalent. For instance, Veeraraghavan and Debo (2011) study customers' joining decisions when there are two competing congested services of unknown service quality. Debo and Veeraraghavan (2014) evaluate the equilibrium joining decision when the queue length is positively correlated with unknown service quality. The focus of the aforementioned studies, as well as that of empirical studies (e.g., see Simonsohn and Ariely (2008)), is to capture the drivers of customers' herding behavior.

Another and better known aspect of human psychology is the *wait-aversion*. A congested system leading to long waits discourages customers from joining. This negative perception of the congestion is an underlying assumption in most queueing studies where optimization is involved. When considering customers acting rationally, congestion is also viewed as reducing the attractiveness of a system. For instance, from queueing games approach, depending on whether the queue is observable of not, customers may adopt a threshold or a randomized joining policy to maximize their utility (Hassin and Haviv, 2003). Empirical studies confirm the phenomenon of wait-aversion, although agreeing that it may not be well captured by rational models (Bennett, 1998; Kumar and Krishnamurthy, 2008; Buell, 2020).

Motivated by the phenomena of crowd-attraction and wait-aversion, we consider a queueing model with a state-dependent arrival rate. Instead of focusing on the drivers of state-dependency, we investigate how a controller may regulate the customer flow by rejecting or proposing a service alternative to some customers. The need to exercise admission control is particularly relevant when congestion attracts more arrivals, as the uncontrolled growth of population in the system may lead to poor service quality. While the admission control problem has been widely studied in the literature, most of these studies assume that the customers' arrival rate is constant (Koçağa and Ward, 2010; Schrieck et al., 2014; Legros, 2020). The aim of this paper, is to determine optimal rejection decisions at arrival in the presence of state-dependent arrival rates.

First, we consider the case of a bounded arrival rate. We employ an iterative Markov decision process approach to compute the optimal policy. Our numerical experiments show the optimality of a single-threshold policy. When the arrival rate is decreasing and convex, we prove the convexity of the value function. This also proves the optimality of a threshold policy. This result is extended to the case of an increasing and potentially unbounded arrival rate. In this case uniformization may not apply and an alternative computational approach based on an $n-$terminating formulation is developed. An algorithm to compute the optimal threshold level is proposed. In addition, we prove that the long-run expected cost is decreasing and convex in the number of servers. These results indicate that the system capacity can be viewed as a decision variable by the system

manager since the question of dimensioning the system is equivalent to computing the optimal policy. Finally, we illustrate our results by considering a case where the arrival rate is linearly increasing. This model may correspond to a situation with informed and uninformed customers. The arrival process of informed customers is Poisson while the arrival process of uninformed customers is generated by the presence of customers in the system. This simple queueing model may capture some of the aspects of herding. For this queue, our numerical investigations show that fewer mistakes are made when dimensioning a crowded system, having systems that are too large may be detrimental to both rejection and congestion, and that crowd-attraction may only have a significant effect in low workload situations.

Structure of the paper. The rest of the paper is organized as follows. Section 2 presents a literature review. Section 3 defines the model and the optimization problem. Section 4 investigates the case of a bounded arrival rate. Section 5 develops an algorithmic approach for the computation of the system capacity in the increasing case. Finally, Section 6 concludes the paper and highlights avenues for future research. The proofs are given in the appendix at the end of the paper.

2 Literature review

The focus of our paper is on *admission control* in queueing systems. Admission control corresponds to short-term decisions as compared to system parameters' optimization considered for determining the system design (Stidham, 2009). Specifically, for a given optimization problem where a trade-off between the system congestion and the rejection flow has to be established, a controller has to decide whether or not to accept a customer in the system. This issue has been extensively addressed in various ways by many authors (Ku and Jordan, 2003; Maglaras and Van Mieghem, 2005; Ward and Kumar, 2008; Xu, 2015; Niyirora and Zhuang, 2017; Bountali and Economou, 2017). In what follows, we present the literature related to this paper. First, we detail admission problems in single queues and explain the methodological tools that are employed in this paper. Second, we present admission problems in queueing networks. Next, we show applications of the admission control in service facilities. Finally, we provide references in relation with queueing games as a way to explain the state-dependency of the arrival rate.

Methodological approach for single queues. Stidham and Weber (1989) developed a method to prove monotonicity results for arrival rate control. Their method inspired our present approach, although they did not directly consider the case of admission control. Koole (2007) presented an event-based dynamic programming framework to prove monotonicity properties in queueing control problems as considered in this

paper. Adopting this approach, they proved that the optimal admission control policy is of a threshold type when the arrival rate is constant. We extend their result in Section 4, using the same approach when the arrival rate is decreasing and convex in the number of customers in the system. This method however relies on uniformization and fails to provide the desired result when the arrival rate is unbounded as in Section 5 of this paper. Studies by Koçağa and Ward (2010) and Adusumilli and Hasenbein (2010) considered the problem of admission and rate control. They developed the so-called $n-$terminating problem to compute the optimal threshold for their respective models. The method is used in this paper to compute the optimal rejection threshold and prove the threshold form of the optimal policy while restricting the analysis to stationary policies in Section 5. Extensions of the result can be found with general interarrival times for a loss system (Örmeci and van der Wal, 2006), batch arrivals (Örmeci and Burnetas, 2005; Yildirim and Hasenbein, 2010), and state-dependent service rates (Xia et al., 2017), but to the best of our knowledge, the result has not yet been extended to the case with state-dependent arrival rates.

Admission control in queueing networks. Admission control has been investigated for complex queueing networks. With a dynamic programming approach close to ours, Hordijk and Koole (1991) investigated routing to parallel queues in which each queue has its own single server. They provided conditions on the cost function such that the optimal policy assigns customers to a faster queue when that server has a shorter queue. Ku and Jordan (2002) considered admission control in a multi-server loss queue fed by a set of upstream parallel multi-server loss queues and by an arrival of new customers. They proved that the policy that maximizes total discounted revenue is a multi-threshold policy. Chang and Chen (2003) considered a two-stage loss tandem queueing system and showed the value of admission control to ensure that customers will go through the second stage successfully. Ziedins (2007) considered a network of parallel finite tandem queues with two stages with loss customers. Surprisingly, the authors showed that as the service rate increases at the second stage, the optimal policy changes in such a way that the total expected cost due to loss increases. Also with two stages of service, Silva et al. (2013) showed that the optimal control policy has a threshold form. Their conjecture is proven in Kim and Kim (2014).

Admission control for service systems. Admission control is often implemented in service systems like hospitals or call centers. For instance, in a queueing network of service facilities, Cosyn and Sigman (2004) investigated the admission control issue with waiting and reneging from a revenue perspective. Using orbiting as an approximation of queueing, they showed that a particular tracking policy is close to be optimal. Lin and Ross (2004) analyzed a single server loss queueing system where a gatekeeper has to decide whether to admit a customer without knowing the status (idle or busy) of the server. They showed that a threshold policy which

rejects arrivals for a certain time interval after each admission and next accepts the next customer is optimal. Bassamboo et al. (2005) considered a multi-class service system with several server pools and doubly stochastic arrivals. A double control is exercised: rejection control at arrival and routing control to a given team after a certain wait. Under asymptotic assumptions on the system parameters, they showed how to implement the fluid model's optimal control in the original service system context. In the context of outsourcing, Gans and Zhou (2007) studied a call center with high and low values calls and evaluated routing schemes for outsourcing part of the low values calls, investing different priority queues. Gurvich and Perry (2012) considered a service network operated under a threshold-type overflow mechanism. If the waiting room is full, the call is overflowed to an outsourcer. Also in an outsourcing context, Legros et al. (2020) proved that rejecting customers after letting them experiment some wait leads to an improvement in the service quality for served customers as compared to rejection at arrival.

Queues with state-dependent arrival rates. The dynamic control of queues with state-dependent arrivals is complicated due to the break of monotonicity properties that this phenomenon can induce, queues with state-dependent parameters have been widely analyzed for performance analysis purposes (e.g., see Boxma et al. (2005); Bekker et al. (2011); Legros (2018)). Queueing games provide a methodological framework to determine how the arrival rates may evolve as functions of the system state. There is a large body of literature on this topic, including Gavirneni and Kulkarni (2016); Dimitrakopoulos and Burnetas (2016); Cui and Veeraraghavan (2016); Hassin and Roet-Green (2017). This literature stream focuses on the impact of customers acting rationally (i.e., deciding whether to join or to balk) when trying to obtain the best trade-off between the value of a service and the cost of waiting. The books of Hassin and Haviv (2003) and Hassin (2016) explain the main principles of decision making from the customers' perspective. In this paper, we consider the state-dependent arrival rate as exogenously determined and turn our focus on external control.

3 Model description

Below, we provide the model description and the routing problem. Our assumptions are partly driven by the actual problem that motivates the analysis, and partly by our concern to keep the model as simple as possible. The idea is to obtain an easy-to-implement dimensioning procedure, and to gain insights into the effect of crowd-attraction.

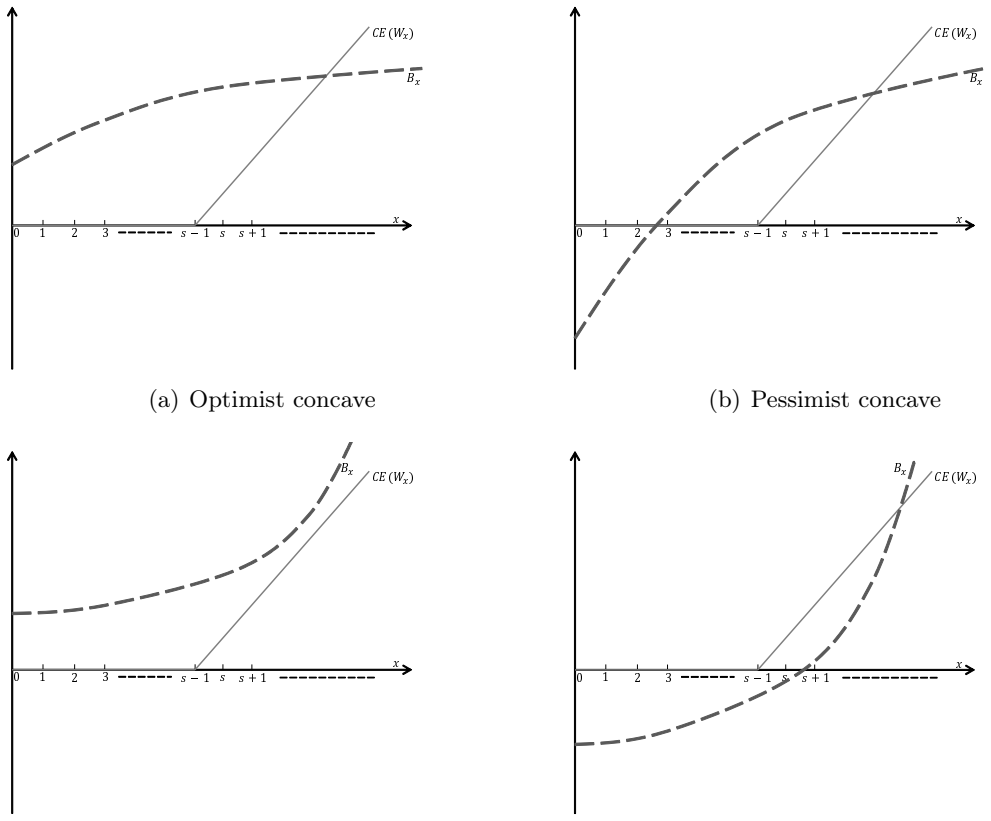
The queueing model. We consider a system with infinite capacity with a single pool of s homogeneous agents. Customers arrive at the system according to a Poisson process with state-dependent arrival rate $\lambda(x)$,

where x is the number of customers present in the system. If a customer is not routed to service immediately upon arrival, then she/he waits in a queue for her/his turn to be served, with customers being served in order of arrival. The service times of all customers are assumed to be exponential random variables with rate 1. As such, our queueing model is called an $M_x/M/s$ queue, under Kendall's notation.

State-dependent arrival rates as an effect of crowd-attraction and wait-aversion. Here, we discuss how state-dependent arrivals can result from the diversity of *rational* customers with regard to their perception of the wait and quality of service. Assume that we have $M+1$ classes of customers and that the arrival process of class-i customers is Poisson with parameter λ_i, for $i = 0, 1, 2, \cdots, M$. On arrival, customers observe the system state and decide whether to join the system or not. This decision is based on a trade-off between the wait and the service quality. Due to the first-come-first-served discipline, the expected wait at arrival when x customers are in the system is $E(W_x) = \frac{\max(x-s+1,0)}{s}$. However, each customer class has a different sensitivity to the wait. We denote by C_i, the wait sensitivity of class-i customers, for $i = 0, 1, 2, \cdots, M$. The particularity of this queue is the congestion-attraction feature. As customers lack information regarding the real quality of service, they tend to associate the number of customers present in the system with the service quality. Yet, this perception differs per customer class. Class-i customers consider that the benefit of the service is $B_{i,x}$, when x customers are in the system. We assume that $B_{i,x}$ is increasing in x, in order to capture the crowd-attraction feature. Note that we may have $B_{i,x} < 0$ when the price of the service is higher than the estimated reward for being served.

The joining decision for class-i customers depends on the difference $B_{i,x} - C_i E(W_x)$. This difference represents the utility for joining the system. Given that the utility of balking is zero, a class-i customer joins the system in state x if $B_{i,x} - C_i E(W_x) \geq 0$, for $x \geq 0$. A joining function, $\phi_i(x)$, can be defined for class-i customers such that $\phi_i(x) = 1$ when customers join at state x, and $\phi_i(x) = 0$ otherwise, for $i = 0, 1, 2, \cdots, M$. The overall arrival rate at state x, $\lambda(x)$, can then be written as $\lambda(x) = \sum_{i=0}^{M} \phi_i(x)\lambda_i$, for $x \geq 0$. We further assume that customers do not regret their choice after joining. This means that customers cannot abandon the system. This assumption is mainly made for tractability reason. Note that when the revenue for being served is fixed and known by customers, rational customers should not abandon as their wait can only reduce over time if the first-come-first-served discipline is applied (Hassin and Haviv, 2003).

Depending on the function $B_{i,x}$, different profiles of customer can be encountered. Figure 1 illustrates four of them. The four profiles in this figure are determined by the evaluation of the service quality when the system is empty and by the convexity property of $B_{i,x}$. Figure 1(a) illustrates a case where $B_{i,x}$ is concave with $B_{i,0} > 0$. This results in classical threshold behavior similar to that encountered with informed customers



(a) Optimist concave

(b) Pessimist concave

(c) Optimist convex

(d) Pessimist convex

Figure 1: Customer profiles

as in Hassin and Haviv (2003). Customers join the system only if the number of customers present is below a certain level. In this case, the wait-aversion is dominant compared to the crowd-attraction. This phenomenon is commonly observed in amusement parks, where customers join an activity if the wait is below a certain threshold level. For the pessimist ($B_{i,0} < 0$) concave profile of Figure 1(b), customers join within a certain interval. As the number of customers increases, the attraction of the system also increases if there is no wait. The attraction decreases with the number of customers present when the system is congested. This phenomenon is observed in restaurants where a sufficient number of customers should be present to attest that the service quality is good. Yet, a too crowded restaurant indicates that the wait will be too long which deters from entering. With optimist and convex customers, either customers always join as presented in Figure 1(c), or they always join except within a certain interval. The former case corresponds to situations where the crowd-attraction is stronger than the wait-aversion or when customers are not sensitive to the wait. This may happen in situations where customers do not actively wait, for instance in call centers where customers ask to be called. The latter case is less likely to happen in real systems. Pessimist and convex customers join only above a certain threshold as shown in Figure 1(d). This phenomenon can be observed in order books on financial markets when a threshold in the number of sent orders can be perceived as accrediting a preference for a price. Several other customer profiles may exist involving a more complex function $B_{i,x}$. This results in

a state-dependent arrival rate, $\lambda(x)$, capturing customers rational decisions.

Optimization problem. To ensure a sufficiently low waiting time and limit system congestion, the system manager should not accept all arriving customers. Therefore, customers' rejection at arrival is permitted. We consider a cost model where a cost of $c_R > 0$ is counted per rejected customer, and each customer in the system costs c_N per time unit. This cost model allows us to account for the conflicting objectives of customer rejection and system congestion. Our aim is to determine optimal rejection control to minimize the long-run system cost per customer.

Our aim is to determine the optimal rejection policy in the case with state-dependent arrival rates. The difficulty lies in the potential irregular behavior of $\lambda(x)$ which may break the monotonicity properties of the performance measures and subsequently preclude proving the optimal policy form. We put forward the following approach to tackle the problem:

- In Section 4, we develop an iterative Markov decision process approach to derive the optimal policy. This method can be employed for any bounded function $\lambda(x)$. In the case where $\lambda(x)$ is decreasing and convex in x, we prove that the optimal policy is of the threshold type.
- In Section 5, we restrict the analysis to stationary policies. In the case where $\lambda(x)$ is increasing in x, we show that the optimal policy is of the threshold type and we propose an algorithm to derive the optimal threshold after a finite number of iterations.

4 Admission control with bounded arrival rates

In this section, we tackle the admission control problem by employing an iterative one-dimensional Markov decision process approach when the arrival rate is bounded. Having a bounded arrival rate allows us to employ the uniformization technique in order to obtain a discrete time Markov chain. In Section 4.1, we formulate the system value function and derive the optimal policy numerically. We show that optimal policies are equivalent to threshold type policies. Next in Section 4.2, we prove the threshold form of the optimal policy in the decreasing and convex case for $\lambda(x)$.

4.1 Computation of the optimal policy

We first formulate the system value function. The two possible transitions of the Markov chain are as follows:

1. Arrival at the queue with rate $\lambda(x)$. The number of customers is increased by 1, which changes the state to $x + 1$, for $x \geq 0$.

2. Service completion with rate $\min(x, s)$. The number of customers is reduced by 1, which changes the state to $x - 1$, for $x > 0$.

Therefore, the Markov chain under consideration is a specific *birth and death process* with birth rate $\lambda(x)$ and death rate $\min(x, s)$.

Given that the arrival rate is bounded, there exists $\overline{\lambda} > 0$ such that $\overline{\lambda} = \sup_{x \geq 0} \lambda(x)$. Consequently, the event rate is bounded by $\overline{\lambda} + s$ and our system is uniformizable. By replacing the rates $\lambda(x)$ and $\min(x, s)$ by $\frac{\lambda(x)}{\overline{\lambda}+s}$ and $\frac{\min(x,s)}{\overline{\lambda}+s}$, we obtain the transition probabilities in the equivalent discrete time Markov chain. Thus, we define the value function, $V_k(x)$, over k steps as follows, with $V_0(x) = 0$, and

$$V_{k+1}(x) = c_N x + \frac{\lambda(x)}{\overline{\lambda} + s} \min(c_R + V_k(x), V_k(x+1)) + \frac{\min(x, s)}{\overline{\lambda} + s} V_k(x-1) + \left(1 - \frac{\lambda(x)}{\overline{\lambda} + s} - \frac{\min(x, s)}{\overline{\lambda} + s}\right) V_k(x), \tag{1}$$

for $x \geq 0$. Note that for $x = 0$, we have $\min(x, s) = 0$, which means that there is no transition from $x = 0$ to $x = -1$ and $V_k(-1)$ does not need to be specified.

We obtain the long-run average optimal actions by applying the value iteration technique introduced by Bellman (1957) and Howard (1960), by recursively evaluating V_k using Equation (1), for $k \geq 0$. As k tends to infinity, the optimal policy converges to the unique average optimal policy and the difference $V_{k+1} - V_k$ converges to expected average long-run costs (see Puterman (1994), Section 8.1 and Koole (2007), Section 4.1). Moreover, the optimal long-run policy is independent of the choice of V_0.

Numerical illustration. For computational reason, we need to define an upper bound for the state space, M, such that $0 \leq x \leq M$. At state $x = M$, Equation (1) is modified in

$$V_{k+1}(M) = c_N M + \frac{\lambda(M)}{\overline{\lambda} + s}(c_R + V_k(M)) + \frac{\min(M, s)}{\overline{\lambda} + s} V_k(M-1) + \left(1 - \frac{\lambda(M)}{\overline{\lambda} + s} - \frac{\min(M, s)}{\overline{\lambda} + s}\right) V_k(M).$$

In what follows, we present examples of computed policies with $c_R = 1$, $c_N = 0.3$, $s = 3$ and where $\lambda(x)$ has a periodic form with $\lambda(3x) = \lambda_0$, $\lambda(3x+1) = \lambda_1$, and $\lambda(3x+2) = \lambda_2$, for $x \geq 0$. The first three columns of the table indicate the values of the arrival rates. The next 11 columns provide information regarding the optimal action at states $x = 0, 1, \cdots, 10$. We use the letter A (respectively, the letter R) when it is optimal to accept (respectively, to reject) an arriving customer. In the presented examples, it is optimal to reject arriving customers for $x \geq 10$. The 15^{th} column recalls the first state at which it is optimal to reject arriving customers. The last column gives the long-run expected cost under the optimal policy.

For the first two lines, with constant arrival rates, we obtain a classical threshold policy. For this policy, it

Table 1: Computation of the optimal policy ($c_R = 1$, $c_N = 0.3$, $s = 3$, $\lambda(3x) = \lambda_0$, $\lambda(3x+1) = \lambda_1$, and $\lambda(3x+2) = \lambda_2$, for $x \geq 0$)

Arrival rates			States											Threshold	Expected cost
λ_0	λ_1	λ_2	0	1	2	3	4	5	6	7	8	9	10		
1	1	1	A	A	A	A	A	A	A	R	R	R	R	7	0.3135
2	2	2	A	A	A	A	A	R	R	R	R	R	R	5	0.7858
0.1	1.5	0.1	A	A	A	A	R	A	R	R	R	R	R	4	0.0669
1	2	3	A	A	A	A	A	R	R	A	R	R	R	5	0.6909
0.1	2	5	R	R	A	A	R	R	R	A	R	R	R	0	0.1000
0.01	0.1	5	A	R	A	A	A	R	A	A	A	R	R	1	0.0040

is optimal to accept all arriving customers if and only the number of customers present in the system is below a certain threshold level. On the contrary, in the last four lines with non-constant arrival rates, we observe multiple switches between the "accept" and "reject" actions. This indicates that the optimal policy is defined by multiple thresholds. However, these complex policies result in the same single-threshold policy as in the case with a constant arrival rate. The reason is that the first rejection threshold has the role of the unique threshold since the system cannot have more customers than this first threshold level. As expected, we observe that the rejection threshold decreases with the arrival rates. Similarly, the threshold should increase with the number of servers. The effect of the number of servers is investigated in Section 5.

In conclusion, we observe from our numerical experiments that the optimal policy for our optimization problem always results in a deterministic single threshold policy. However, this result is not possible to prove in general due to the multiple switches observed between the "accept" and "reject" actions. We overcome this difficulty in the next section by considering the case where the arrival rate is decreasing and convex.

4.2 Analysis of the decreasing and convex case

We now assume that the arrival rate is decreasing and convex. In this case, we have $\overline{\lambda} = \lambda(0)$. We prove in this section that the optimal policy is a single threshold policy. A single threshold policy is characterized by the fact that if rejection is optimal in state x, then rejection is also optimal in state $x+1$. A sufficient condition for this is: if $V_k(x) + c_R \leq V_k(x+1)$ then $V_k(x+1) + c_R \leq V_k(x+2)$. This implication may happen if $V_k(x+2) - V_k(x+1) - c_R \geq V_k(x+1) - V_k(x) - c_R$, which is equivalent to the convexity of V_k: $V_k(x+2) + V_k(x) \geq 2V_k(x+1)$, for $x \geq 0$. Therefore, by showing the convexity of the value function, we prove the optimality of a single threshold policy. This convexity property is proven in Theorem 1 by induction on k. The proof consists of showing the induction step for every component of the value function. In addition to the convexity property, we need to prove that $V_k(x)$ is increasing in x. The induction step requires the decreasing and convex property of $\lambda(x)$.

Theorem 1. *If $\lambda(x)$ is decreasing and convex in x, then the optimal policy for the admission control problem is of threshold type.*

The result of Theorem 1 indicates that the optimal policy can be defined by a threshold, n, such that customers are allowed to join only if $x < n$. This threshold is the system capacity. Therefore, the optimization problem can be interpreted as dimensioning the system.

For computational purposes, we need to set the value of the upper bound M sufficiently high. If the optimal threshold, n, is such that $n < M$, then the switch from "accept" to "reject" will be observed. However if $n \geq M$, then in all states $x < M$ the optimal action is to accept an arriving customer. This does not allow us to conclude on the value of the optimal threshold. We may then increase M by one and check whether the switch from "accept" to "reject" can be observed. If the optimal threshold is finite, then after a finite number of iterations, the optimal threshold can be found. Yet, the optimal threshold could be infinite. In this case, this procedure will not enable us to determine the optimal policy after a finite number of iterations. We do not have a sharp condition to determine if the optimal threshold is infinite. However, to remove some particular cases from the analysis, in Proposition 1 we provide a necessary condition for the optimal threshold to be infinite.

Proposition 1. *If* $c_R \frac{\lambda(x)-\lambda(x+1)+\min(x+1,s)-\min(x,s)}{\lambda(0)+s} \geq c_N$ *for* $x \geq 0$*, then it is optimal to accept all arriving customers (i.e., the optimal threshold is infinite).*

In the following section, we restrict the analysis to stationary policies. In this set of policies, we prove that deterministic policies are optimal, and we show how the optimal capacity can be computed.

5 Analysis of the increasing case

In this section, we consider the case where $\lambda(x)$ is increasing in x and not necessarily bounded. Since the state space is countably infinite, this may result in an unbounded birth rate. Therefore, contrary to the assumption in Section 4, uniformization does not apply here. This means that the value iteration technique used in Section 4 in the equivalent discrete time Markov chain cannot be employed (Koole, 2007).

Instead, in Section 5.1, we develop an algorithmic approach to obtain the optimal policy within the set of stationary policies. First, we prove that a threshold policy is optimal within the stationary policy set. Next, we develop an algorithm to compute the optimal threshold level. In Section 5.2, we show the applicability of our results in the particular case of a linearly increasing arrival rate.

5.1 Computation of the optimal threshold

The optimality equation for the relative value function $V(x)$ for the system, and average constant cost g is

$$V(x) = \frac{c_N x - g}{\lambda(x) + \min(x,s)} + \frac{\lambda(x)}{\lambda(x) + \min(x,s)} \min\left(V(x+1), V(x) + c_R\right) + \frac{\min(x,s)}{\lambda(x) + \min(x,s)} V(x-1), \quad (2)$$

for $x \geq 0$. We define $\Delta(x) = V(x) - V(x-1)$, for $x > 0$. Therefore, Equation (2) can be rewritten as

$$c_N x - g = \min(x,s)\Delta(x) - \lambda(x)\min(\Delta(x+1), c_R), \quad (3)$$

for $x > 0$, and

$$g = \lambda(0)\min(\Delta(1), c_R), \quad (4)$$

for $x = 0$. Note that although the state space and the total event rate are unbounded, due to the minimizing operator, the system's stability is ensured and the approach in Puterman (1994), chapter 11, can apply to obtain Equation (2).

In Theorem 2, we confirm the validity of the optimality equations for bounded solutions of Equations (3)-(4). This theorem is also called a *verification theorem*. Let us denote the set of stationary stable policies by Ω_S. A policy $\pi_S \in \Omega_S$ is defined by the probabilities $\{p(0), p(1), \cdots, p(x), \cdots\}$, where $p(x)$ is the probability of accepting an arriving customer in state x. Let us denote by q_x the stationary probability of being in state x under policy π_S, for $x \geq 0$. The probabilities q_x are solutions of

$$\lambda(x)p(x)q_x = \min(s, x+1)q_{x+1}, \quad (5)$$

for $x \geq 0$. We denote the expected cost associated with a policy π_S by g^S.

Theorem 2. ***Verification Theorem.*** *Suppose we have $g < \infty$ and a bounded sequence ($\Delta(1)$, $\Delta(2)$,...,$\Delta(x)$,...) satisfying Equations (3)-(4). If g^S is the average cost associated with a policy in Ω_S, we then have $g^S \geq g$.*

The consequence of Theorem 2 is that a solution to Equations (3) and (4) exists and is unique. Although we cannot affirm that this policy is stationary, we propose restricting the analysis to this set of policies. In what follows, we prove that the optimal policy within the set of stationary policies is a threshold policy. In addition, we prove that the optimal threshold level can be computed after a finite number of iterations. To this end, we introduce the $n-$terminating problem approach as in Koçağa and Ward (2010) and Adusumilli and Hasenbein (2010). The objective is to determine a constant g^n and a vector $(\Delta^n(1), \Delta^n(2), \cdots, \Delta^n(n+1))$

such that

$$g^n = \lambda(0)\Delta^n(1), \tag{6}$$

and,

$$c_N x - g^n = \min(x, s)\Delta^n(x) - \lambda(x)\Delta^n(x+1), \tag{7}$$

for $0 < x \leq n$, with $\Delta^n(n+1) = c_R$. The cost g^n is the average cost associated with a threshold policy at level n. Equations (6) and (7) define a Markovian Reward Process with finite state space. For such processes, the existence of a unique solution is proven in Puterman (1994) (see Proposition 8.2.1).

In what follows, we prove in Theorem 3 that the optimal policy within the set of stationary policies is of the threshold type and that the first local minimum obtained by solving Equations (6)-(7) is the optimal threshold level. For this purpose, we prove an inequality in Lemma 1 relating the cost g^n and the relative difference $\Delta^n(x)$. Next, in Lemma 2, we prove that $\Delta^n(x) \geq 0$, for $1 \leq x \leq n+1$.

Lemma 1. *If $g^{n_1} \geq g^{n_2}$, for $n_1, n_2 \in \mathbb{N}$, then $\Delta^{n_1}(x) \geq \Delta^{n_2}(x)$, for $x \in \{1, 2, \cdots, \min(n_1, n_2) + 1\}$.*

Lemma 2. *We have $\Delta^n(x) \geq 0$, for $1 \leq x \leq n+1$.*

Theorem 3. *Assume that $\lambda(x)$ is increasing in x. Suppose there is a sequence of solutions to the $n-$terminating problem (6) and (7) such that $g^m < g^k$, for $k \in \{0, 1, 2, \cdots, m-1\}$ and $g^{m+1} \geq g^m$. Then if g^S is the cost associated with a policy in Ω_S, we have $g^S \geq g^m$.*

From Theorem 3, whenever a local minimum of g^n is found, this minimum is the optimal capacity. To reduce the set of possible threshold levels, in Proposition 2 we provide conditions under which either all customers should be rejected or the rejection threshold should be higher than or equal to s.

Proposition 2. *Special cases: If $c_N < (1 + \lambda(m) - \lambda(m+1))c_R$, for $0 \leq m \leq s$, then $g^0 > g^1 > \cdots > g^s$. If $c_N \geq (1 + \lambda(m) - \lambda(m+1))c_R$ for $m \geq 0$, then $g^0 < g^1 < \cdots < g^s < g^{s+1} < \cdots < g^n < \cdots$.*

We now provide the algorithm to compute the optimal system capacity, n. To facilitate the computation, we specify the system cost, g^n, obtained from a standard Markov chain analysis, and $\Delta^n(n)$, obtained from

Equation (7). So

$$g^n = c_R\lambda(n)\frac{\prod_{i=0}^{n-1}\frac{\lambda(i)}{\min(i+1,s)}}{\sum_{x=0}^{n}\prod_{i=0}^{x-1}\frac{\lambda(i)}{\min(i+1,s)}} + c_N\frac{\sum_{x=0}^{n}x\cdot\prod_{i=0}^{x-1}\frac{\lambda(i)}{\min(i+1,s)}}{\sum_{x=0}^{n}\prod_{i=0}^{x-1}\frac{\lambda(i)}{\min(i+1,s)}}, \text{ and,} \tag{8}$$

$$\Delta^n(n) = \frac{c_N n - g^n + \lambda(n)c_R}{\min(n,s)}. \tag{9}$$

The algorithm is as follows:

Algorithm 1: Computation of the optimal rejection threshold.

1. *Initialisation.* If $c_N \geq (1+\lambda(m)-\lambda(m+1))c_R$ for $m \geq 0$, then $n=0$ is optimal. If $c_N < (1+\lambda(m)-\lambda(m+1))c_R$, for $0 \leq m \leq s$, then set $n=s$. Otherwise, set $n=0$. Compute g^n Equations (8) and (9).

2. *Iteration step:* Increase n by one and compute g^n using Equations (8) and (9).

   If $g^n > g^{n-1}$, then the rejection threshold $n-1$ is optimal.

   If $g^n \leq g^{n-1}$, then go back to the iteration step.

We do not know if Algorithm 1 will stop. It could be that the optimal capacity is infinite. Unfortunately, as in Section 4, we cannot obtain a condition which determines if the threshold is infinite. From Theorem 3, the condition to have an infinite optimal capacity is to have g^n decreasing in n. In this case, since we also have $g^n \geq 0$, we may define the limit of g^n as $\lim_{n\longrightarrow\infty} g^n = \overline{g}$. In Proposition 3, we provide an inequality which could be used as a stopping criterion for Algorithm 1.

Proposition 3. *Stopping criterion: If g^n is decreasing in n, then $\lim_{n\longrightarrow\infty} g^n = \overline{g}$ and $g^n - \overline{g} \leq \lambda(n)(c_R - \Delta^n(n))$.*

Note however that this proposition does not provide a necessary and sufficient condition for having an infinite optimal threshold.

Impact of the number of servers. We end this section by determining the effect of the number of servers on the expected cost for a given rejection threshold, n. Using Equations (6) and (7), we prove in Proposition 4 that the system expected cost, g^n, is strictly decreasing and convex in the number of servers.

Proposition 4. *For a system with a given rejection threshold, n, the expected cost, g^n, is strictly decreasing and convex in s.*

5.2 Applicability of the results for a linearly increasing arrival rate

In this section, we show the applicability of our results to the case $\lambda(x) = \lambda + x\gamma$. From a modeling perspective, this queueing model can be viewed as an off-shoot of Hawkes processes. Hawkes (1971) introduced the idea of

self-excitement, a model in which the current intensity of events is determined by events in the past. The rate of new event occurrences increases as each event occurs. This concept has been extended to queueing systems via the definition of the Queue-Hawkes process in Daw and Pender (2020). As such, our model with a linearly increasing arrival rate is an *affine Queue-Hawkes Process* as studied in Section 3 of Daw and Pender (2020).

First, we provide here the stationary performance measures of this queue for a given rejection threshold, n. The expected number of customers in the system, $E(N)$, and the expected rate of rejected customers, $E(R)$, are derived from the stationary probabilities. We denote by q_x the stationary probability to have x customers in the system. We have $(\lambda + \gamma x)q_x = \min(x+1, s)q_{x+1}$, for $0 \leq x \leq n-1$. This leads to

$$q_x = q_0 \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma} + x\right)}{x! \Gamma\left(\frac{\lambda}{\gamma}\right)}, \text{ for } 0 \leq x \leq s, \text{ and, } q_x = q_0 \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma} + x\right)}{s! s^{x-s} \Gamma\left(\frac{\lambda}{\gamma}\right)}, \text{ for } s \leq x \leq n, \text{ with,} \tag{10}$$

$$q_0 = \left[\sum_{x=0}^{s-1} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma} + x\right)}{x! \Gamma\left(\frac{\lambda}{\gamma}\right)} + \sum_{x=s}^{n} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma} + x\right)}{s! s^{x-s} \Gamma\left(\frac{\lambda}{\gamma}\right)} \right]^{-1}, \tag{11}$$

where $\Gamma(z)$ is the Gamma-function defined for all complex numbers except the non-positive integers as $\Gamma(z) = \int_{t=0}^{\infty} t^{z-1} e^{-t} \, \mathrm{d}t$. Recall that this queueing model as been studied in Daw and Pender (2020) with an infinite number of servers. As such, the stationary probabilities in Daw and Pender (2020), given in Theorem 3.3 can be deduced from our results in Equations (10) and (11) by letting s tend to infinity. The expected number of customers in the system is given by $E(N) = \sum_{x=0}^{n} x \cdot q_x$. The expected rate of rejected customers is the rate of customers arriving at state $x = n$; $E(R) = (\lambda + n\gamma)q_n$. Therefore, we deduce that

$$E(N) = \frac{\sum_{x=0}^{s-1} x \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{x!} + \sum_{x=s}^{n} x \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{s! s^{x-s}}}{\sum_{x=0}^{s-1} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{x!} + \sum_{x=s}^{n} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{s! s^{x-s}}}, \text{ and, } E(R) = (\lambda + \gamma n) \frac{\frac{\gamma^n \Gamma\left(\frac{\lambda}{\gamma}+n\right)}{s! s^{n-s}}}{\sum_{x=0}^{s-1} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{x!} + \sum_{x=s}^{n} \frac{\gamma^x \Gamma\left(\frac{\lambda}{\gamma}+x\right)}{s! s^{x-s}}}. \tag{12}$$

We now illustrate the applicability of our algorithm and investigate the effect of the congestion-attraction feature. In Table 2, we give the optimal threshold and optimal cost computed using the results of Section 5.1.

We observe that the rejection threshold, n^*, decreases with λ and γ and increases with s. When customer arrivals increase, either due to an increase in λ and or in γ, congestion in the system increases and the rejection threshold needs to be decremented to avoid excessive congestion. Large systems are known to handle a given workload better than small systems. Therefore, the rejection threshold should be incremented as s increases in order to accept more customers. This result is a consequence of Proposition 4. While these results are intuitive and expected, it should be noted that they are only significant for low arrival rates (or a high number

Table 2: Numerical Illustration ($c_R = 1$)

Exogenous parameters			Optimal results				Approximations	
	λ	γ	$E(N)$	$E(R)$	n^*	Optimal Cost	Approx1-Cost	Approx2-Cost
$s = 10$, $c_N = 0$	6	0.1	7.469	0.003	40	0.003	-3	0
	6	0.5	8.840	2.076	12	2.076	1	0
	6	1	8.571	6.000	10	6.000	6	0
	12	0.1	12.558	3.548	15	3.548	3	0
	12	0.5	9.818	7.521	11	7.521	7	0
	12	1	9.231	12.000	10	12.000	12	0
$s = 10$, $c_N = 0.1$	6	0.1	7.251	0.013	23	0.738	-2	0.667
	6	0.5	8.179	2.092	11	2.910	2	1.2
	6	1	0.000	6.000	0	6.000	7	-
	12	0.1	10.849	3.624	13	4.709	4	1.333
	12	0.5	8.927	7.536	10	8.429	8	2.4
	12	1	0.000	12.000	0	12.000	13	-
$s = 50$, $c_N = 0$	30	0.1	33.353	0.000	186	0.000	-15	0
	30	0.5	47.456	6.960	53	6.960	5	0
	30	1	48.387	30.000	50	30.000	30	0
	60	0.1	53.974	15.736	57	15.736	15	0
	60	0.5	49.643	35.595	51	35.595	35	0
	60	1	49.180	60.000	50	60.000	60	0
$s = 50$, $c_N = 0.1$	30	0.1	33.353	0.000	186	3.335	-10	3.333
	30	0.5	46.678	7.025	52	11.693	10	6
	30	1	0.000	30.000	0	30.000	35	-
	60	0.1	52.080	15.795	55	21.003	20	6.667
	60	0.5	48.675	35.663	50	40.530	40	12
	60	1	0.000	60.000	0	60.000	65	-

of servers). This means that fewer mistakes can be made when dimensioning a crowded system when the arrival rate is hard to forecast.

We also observe that the rate of rejected customers is not necessarily decreasing in n. This explains why we do not have $n^* = \infty$ when $c_N = 0$. With $\gamma > 0$, two phenomena are in competition as the rejection threshold increases. On the one hand increasing n allows the system to accept more customers and should reduce the rate of rejected customers. On the other hand increasing n increases the generation of new customers as the system is more congested. For small values of n, the former phenomenon is dominant while the latter becomes dominant for higher values of n. From a managerial viewpoint, this means that by accepting more customers, we may create an attraction phenomemon which could result in rejecting more customers at the end. This may be an unintended consequence of increasing the threshold level. For our optimization problem, the threshold level was understood as a way to provide a trade-off between customers rejection and the number of customers in the system. The intuition was that increasing the threshold would be detrimental to the number of customers in the system while being beneficial to the rate of rejected customers. Our study instead reveals that increasing the threshold can be detrimental for both the two aforementioned metrics when the system is large. This precludes creating systems that are too large where over-congestion and high rejection rates would be encountered simultaneously.

Finally, we observe that the effect of γ is reduced with high arrival rates while it increases in large systems. With high arrival rates, the rejection threshold should be reduced so as to avoid excessive congestion. As the number of customers in the system cannot grow large, the rate of arrivals generated by these customers cannot be high either. With symmetrical arguments, this also explains why the effect of γ increases with the number

of servers. Consequently, in small and busy systems, like in restaurants at peak hours, it is not possible for crowd-attraction to operate. In the long-run, it also means that such systems do not have much opportunity to attract new (uninformed) customers. Extending opening hours to less busy periods may help to increase the benefit of attraction created by present customers.

Approximations in extreme workload situations. We end this section by providing approximations for the optimal cost in high and low workload situations for large systems. These approximations are given in Proposition 5. They correspond to the limits of $E(N)$ and $E(R)$ when the system parameters n, s, and λ tend to infinity with the relations $n = rs$, with $r \geq 1$ and $\lambda = qs$, with $q > 0$. We introduce the $\sim$ symbol and write $a_n \sim b_n$ if $\lim_{n \longrightarrow \infty} \frac{a_n}{b_n} = 1$.

Proposition 5. *We let $n = rs$, with $r \geq 1$ and $\lambda = qs$, with $q > 0$. The following holds:*

- ***Approximation 1:*** *If $q + r\gamma - 1 > 0$ (High workload), then $E(R) \sim s\,(q - 1 + r\gamma)$, and $E(N) \sim rs$.*
- ***Approximation 2:*** *If $q + r\gamma - 1 \leq 0$ (Low workload), then $E(R)$ tends to zero as n tends to infinity, and $E(N) \sim \frac{\lambda}{1-\gamma}$.*

The simplicity of the expressions in Proposition 5 allows the system manager to take simple combined decisions for the system capacity and the staffing level. However, the number of servers should be very large to have a good approximation of the real system. In the last two columns of Table 2, the approximated costs are also computed. We observe an important gap between the approximations and the real values. This is due to the number of servers being less than 50. With a small number of servers, the approximations are good only in over-staffed (Approximation 1) under-staffed (Approximation 2) situations.

6 Conclusion

This paper analyzed a queue with state-dependent arrival rates accounting for wait-aversion and crowd-attraction. We investigated the admission control problem for this queue and proved that the optimal policy for this problem is of the threshold type when the arrival rate is decreasing and convex. Next, by restricting the analysis to stationary policies in the increasing case, we proved that the optimal policy remains of a threshold nature. Moreover, we showed that the first local minimum of the system cost obtained by increasing the threshold level is the optimal threshold level. We illustrated our results for a linearly increasing arrival rate of the form $\lambda(x) = \lambda + \gamma x$, for $x \geq 0$. Our numerical investigations showed that dimensioning a crowded system is easier as mistakes in predicting the arrival rate have fewer consequences, that very large systems

accumulate the negative aspect of high rejection, and that over-congestion, and and the crowd-attraction only has a small impact in highly congested situations.

In methodological terms, it would be interesting to extend the proof of optimality of a threshold policy to any form of arrival rate. We only determined a necessary condition for the optimal threshold to be infinite. It could be interesting for computational purposes to obtain a necessary and sufficient condition for this property. Rather than having all agents present at opening time, it would be less costly to give a later appointment to each server in order to avoid wasted capacity. From a modeling perspective, extensions could be considered to better model the operational complexity of service systems, as well as that of customer behavior. One important extension would be to allow for non-stationary, batch arrivals or customers' abandonment. While non-stationarity requires different methodological tools from ours, batch arrivals only involve higher jumps in the Markov chain which would not affect most of the proven properties in this paper if the batches are of equal size. The analysis might be more complicated with batches of unequal size. Moreover, the service process may include multiple pools of agents, as well as more complex service requirements. Including this complexity in our model could, however, lead to an increase in the dimensionality of the problem, which may make it difficult to obtain optimal policies.

References

Adusumilli, K. and Hasenbein, J. (2010). Dynamic admission and service rate control of a queue. *Queueing Systems*, 66(2):131–154.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J. (2013). Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499.

Bassamboo, A., Harrison, M., and Zeevi, A. (2005). Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3-4):249–285.

Bekker, R., Koole, G., Nielsen, B., and Nielsen, T. (2011). Queues with waiting time dependent service. *Queueing systems*, 68(1):61–78.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

Bennett, R. (1998). Queues, customer characteristics and policies for managing waiting-lines in supermarkets. *International Journal of Retail & Distribution Management*.

Bountali, O. and Economou, A. (2017). Equilibrium joining strategies in batch service queueing systems. *European Journal of Operational Research*, 260(3):1142–1151.

Boxma, O., Kaspi, H., Kella, O., and Perry, D. (2005). On/off storage systems with state-dependent input, output, and switching rates. *Probability in the Engineering and Informational Sciences*, 19(1):1–14.

Buell, R. W. (2020). Last-place aversion in queues. *Management Science*.

Chang, K. and Chen, W. (2003). Admission control policies for two-stage tandem queues with no waiting spaces. *Computers & Operations Research*, 30(4):589–601.

Cosyn, J. and Sigman, K. (2004). Stochastic networks: Admission and routing using penalty functions. *Queueing Systems*, 48(3-4):237–262.

Cui, S. and Veeraraghavan, S. (2016). Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Science*, 62(12):3656–3672.

Da Fonseca, J. and Zaatour, R. (2014). Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579.

Daw, A. and Pender, J. (2020). The Queue-Hawkes process: Ephemeral self-excitement. *arXiv preprint arXiv:1811.04282*.

Debo, L. and Veeraraghavan, S. (2009). Models of herding behavior in operations management. In *Consumer-Driven Demand and Operations Management Models*, pages 81–112. Springer.

Debo, L. and Veeraraghavan, S. (2014). Equilibrium in queues under unknown service times and service value. *Operations Research*, 62(1):38–57.

Dimitrakopoulos, Y. and Burnetas, A. (2016). Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control. *European Journal of Operational Research*, 252(2):477–486.

Gans, N. and Zhou, Y. (2007). Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management*, 9(1):33–50.

Gavirneni, S. and Kulkarni, V. (2016). Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, 25(6):979–992.

Gurvich, I. and Perry, O. (2012). Overflow networks: Approximations and implications to call center outsourcing. *Operations research*, 60(4):996–1009.

Hassin, R. (2016). *Rational queueing*. CRC press.

Hassin, R. and Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media.

Hassin, R. and Roet-Green, R. (2017). The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research*, 65(3):804–820.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hordijk, A. and Koole, G. (1991). *On the assignment of customers to parallel queues*. Rijksuniversiteit Leiden. Instituut voor Toegepaste Wiskunde en Informatica.

Howard, R. (1960). *Dynamic Programming and Markov Processes*. Massachusetts Institute of Technology Press, Cambridge.

Kim, B. and Kim, J. (2014). Optimal admission control for two station tandem queues with loss. *Operations Research Letters*, 42(4):257–262.

Koçağa, Y. L. and Ward, A. R. (2010). Admission control for a multi-server queue with abandonment. *Queueing Systems*, 65(3):275–323.

Koole, G. (2007). Monotonicity in markov reward and decision chains: Theory and applications. *Foundations and Trends in Stochastic Systems*, 1(1):1–76.

Ku, C. and Jordan, S. (2003). Near optimal admission control for multiserver loss queues in series. *European Journal of Operational Research*, 144(1):166–178.

Ku, C.-Y. and Jordan, S. (2002). Access control of parallel multiserver loss queues. *Performance Evaluation*, 50(4):219–231.

Kumar, P. and Krishnamurthy, P. (2008). The impact of service-time uncertainty and anticipated congestion on customers' waiting-time decisions. *Journal of Service Research*, 10(3):282–292.

Legros, B. (2018). M/G/1 queue with event-dependent arrival rates. *Queueing Systems*, 89(3-4):269–301.

Legros, B. (2020). Late-rejection, a strategy to perform an overflow policy. *European Journal of Operational Research*, 281(1):66–76.

Legros, B., Jouini, O., and Koole, G. (2020). Should we wait before outsourcing? analysis of a revenue-generating blended contact center. *Manufacturing & Service Operations Management*.

Lin, K. and Ross, S. (2004). Optimal admission control for a single-server loss queue. *Journal of Applied Probability*, 41(2):535–546.

Maglaras, C. and Van Mieghem, J. (2005). Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. *European journal of Operational Research*, 167(1):179–207.

Niyirora, J. and Zhuang, J. (2017). Fluid approximations and control of queues in emergency departments. *European Journal of Operational Research*, 261(3):1110–1124.

Örmeci, L. and Burnetas, A. (2005). Dynamic admission control for loss systems with batch arrivals. *Advances in Applied Probability*, 37(4):915–937.

Örmeci, L. and van der Wal, J. (2006). Admission policies for a two class loss system with general interarrival times. *Stochastic models*, 22(1):37–53.

Puterman, M. (1994). *Markov Decision Processes*. John Wiley and Sons.

Rambaldi, M., Bacry, E., and Lillo, F. (2017). The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, 17(7):999–1020.

Schrieck, J., Akşin, Z., and Chevalier, P. (2014). Peakedness-based staffing for call center outsourcing. *Production and Operations Management*, 23(3):504–524.

Silva, D., Zhang, B., and Ayhan, H. (2013). Optimal admission control for tandem loss systems with two stations. *Operations Research Letters*, 41(4):351–356.

Simonsohn, U. and Ariely, D. (2008). When rational sellers face nonrational buyers: evidence from herding on ebay. *Management science*, 54(9):1624–1637.

Stidham, S. (2009). *Optimal design of queueing systems*. CRC press.

Stidham, S. and Weber, R. (1989). Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations research*, 37(4):611–625.

Veeraraghavan, S. and Debo, L. (2011). Herding in queues with waiting costs: Rationality and regret. *Manufacturing & Service Operations Management*, 13(3):329–346.

Ward, A. and Kumar, S. (2008). Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202.

Xia, L., He, Q., and Alfa, A. (2017). Optimal control of state-dependent service rates in a MAP/M/1 queue. *IEEE Transactions on Automatic Control*, 62(10):4965–4979.

Xu, K. (2015). Necessity of future information in admission control. *Operations Research*, 63(5):1213–1226.

Yildirim, U. and Hasenbein, J. (2010). Admission control and pricing in a queue with batch arrivals. *Operations Research Letters*, 38(5):427–431.

Ziedins, I. (2007). A paradox in a queueing network with state-dependent routing and loss. *Advances in Decision Sciences*, 2007.

A Proof of Theorem 1

Proof. We show by induction on k that $V_k(x)$ is increasing and convex in x, for $x \geq 0$. Since $V_0(x) = 0$, $V_0(x)$ is increasing and convex. Assume that $V_k(x)$ is increasing and convex. We want to show that the same property holds for $V_{k+1}(x)$.

We rewrite Equation (1) as

$$(\lambda(0) + s)V_{k+1}(x) = c_N(\lambda(0) + s)x + A(V_k(x)) + S(V_k(x)), \tag{13}$$

where $A(V_k(x)) = \lambda(x)\min(c_R + V_k(x), V_k(x+1)) + (\lambda(0) - \lambda(x))V_k(x)$, and $S(V_k(x)) = \min(x, s)V_k(x-1) + (s - \min(x, s))V_k(x)$. The function $c_N(\lambda(0) + s)x$ is increasing and convex. There remains to prove that $A(V_k(x))$ and $S(V_k(x))$ are also increasing and convex.

$A(V_k(x))$ is increasing in x. For $x \geq 0$, we have

$$\begin{aligned} A(V_k(x+1)) - A(V_k(x)) &= \lambda(x+1)(\min(c_R + V_k(x+1), V_k(x+2)) - \min(c_R + V_k(x), V_k(x+1))) \\ &\quad + (\lambda(x+1) - \lambda(x))\min(c_R + V_k(x), V_k(x+1)) \\ &\quad + (\lambda(0) - \lambda(x))(V_k(x+1) - V_k(x)) - (\lambda(x+1) - \lambda(x))V_k(x+1). \end{aligned} \tag{14}$$

Three cases can be encountered.

Case 1: $\min(c_R + V_k(x+1), V_k(x+2)) = V_k(x+2)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$. In this case, Equation (14) can be rewritten as

$$A(V_k(x+1)) - A(V_k(x)) = \lambda(x+1)(V_k(x+2) - V_k(x+1)) + (\lambda(0) - \lambda(x))(V_k(x+1) - V_k(x)) \geq 0,$$

since $V_k(x)$ is increasing in x.

Case 2: $\min(c_R + V_k(x+1), V_k(x+2)) = c_R + V_k(x+1)$ and $\min(c_R + V_k(x), V_k(x+1)) = c_R + V_k(x)$. Equation (14) then becomes

$$\begin{aligned} A(V_k(x+1)) - A(V_k(x)) = \lambda(x+1)(V_k(x+1) - V_k(x)) + (\lambda(x+1) - \lambda(x))(c_R + V_k(x) - V_k(x+1)) \\ + (\lambda(0) - \lambda(x))(V_k(x+1) - V_k(x)). \end{aligned}$$

The terms on the right hand side of the equations are all positive, since $V_k(x)$ is increasing. The term proportional with $\lambda(x+1) - \lambda(x)$ is negative since we assumed that $c_R + V_k(x) \leq V_k(x+1)$. Yet, since $\lambda(x)$ is decreasing in x, the product $(\lambda(x+1) - \lambda(x))(c_R + V_k(x) - V_k(x+1))$ is positive.

Case 3: $\min(c_R + V_k(x+1), V_k(x+2)) = c_R + V_k(x+1)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$. Equation (14) is then

$$A(V_k(x+1)) - A(V_k(x)) = \lambda(x+1)c_R + (\lambda(0) - \lambda(x))(V_k(x+1) - V_k(x)) \geq 0,$$

since $c_R \geq 0$ and $V_k(x)$ is increasing in x. This proves that $A(V_k(x))$ is increasing in x.

$A(V_k(x))$ is convex in x. For $x \geq 0$, we may write

$$\begin{aligned} &A(V_k(x+2)) + A(V_k(x)) - 2A(V_k(x+1)) \qquad (15) \\ &= \lambda(x+2)(\min(c_R + V_k(x+2), V_k(x+3)) + \min(c_R + V_k(x), V_k(x+1)) - 2\min(c_R + V_k(x+1), V_k(x+2))) \\ &+ (\lambda(x) - \lambda(x+2))\min(c_R + V_k(x), V_k(x+1)) - 2(\lambda(x+1) - \lambda(x+2))\min(c_R + V_k(x+1), V_k(x+2)) \\ &+ (\lambda(0) - \lambda(x))(V_k(x+2) + V_k(x) - 2V_k(x+1)) \\ &- (\lambda(x+2) - \lambda(x))V_k(x+2) + 2(\lambda(x+1) - \lambda(x))V_k(x+1). \end{aligned}$$

We show that the right hand side of Equation (15) is positive. let us consider the first line after the equality. We call this line $L1$. We have

$$2\min(c_R + V_k(x+1), V_k(x+2)) \leq 2V_k(x+2), \qquad (16)$$

$$2\min(c_R + V_k(x+1), V_k(x+2)) \leq c_R + V_k(x+1) + V_k(x+2), \qquad (17)$$

$$2\min(c_R + V_k(x+1), V_k(x+2)) \leq 2c_R + 2V_k(x+1). \qquad (18)$$

Case 1: $\min(c_R + V_k(x+2), V_k(x+3)) = V_k(x+3)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$. In this case, Inequality (16) combined with the convexity of V_k proves that $L1$ is positive.

Case 2: $\min(c_R + V_k(x+2), V_k(x+3)) = c_R + V_k(x+2)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$. In this case, Inequality (17) proves that $L1$ is positive.

Case 3: $\min(c_R + V_k(x+2), V_k(x+3)) = c_R + V_k(x+2)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x) + c_R$. In this case, Inequality (18) combined with the convexity of V_k proves that $L1$ is positive.

Note that the case $\min(c_R + V_k(x+2), V_k(x+3)) = V_k(x+3)$ and $\min(c_R + V_k(x), V_k(x+1)) = V_k(x) + c_R$ cannot happen because it is in contradiction with the convexity property of V_k.

The third line of the right hand side of Equation (15) is also positive since V_k is convex in x.

Let us now consider the sum of the second and fourth line on the right hand side of Equation (15). This sum is called $L2 + L4$.

Case 1: $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$ and $\min(c_R + V_k(x+1), V_k(x+2)) = V_k(x+2)$. In this case, we have

$$L2 + L4 = (\lambda(x+2) + \lambda(x) - 2\lambda(x+1))(V_k(x+2) - V_k(x+1)) \geq 0,$$

since $\lambda(x)$ is convex in x and V_k is increasing in x.

Case 2: $\min(c_R + V_k(x), V_k(x+1)) = V_k(x+1)$ and $\min(c_R + V_k(x+1), V_k(x+2)) = c_R + V_k(x+1)$. We thus have

$$L2 + L4 = (\lambda(x+2) + \lambda(x) - 2\lambda(x+1))c_R + (\lambda(x) - \lambda(x+2))(V_k(x+2) - V_k(x+1) - c_R) \geq 0,$$

since $\lambda(x)$ is convex in x, $c_R \geq 0$, and since we assumed $V_k(x+1) + c_R \leq V_k(x+2)$.

Case 3: $\min(c_R + V_k(x), V_k(x+1)) = c_R + V_k(x)$ and $\min(c_R + V_k(x+1), V_k(x+2)) = c_R + V_k(x+1)$. We thus have

$$L2 + L4 = (\lambda(x+2) + \lambda(x) - 2\lambda(x+1))c_R + (\lambda(x) - \lambda(x+2))(V_k(x+2) + V_k(x) - 2V_k(x+1)) \geq 0,$$

since $\lambda(x)$ is convex in x, $c_R \geq 0$, $\lambda(x)$ is decreasing in x, and $V_k(x)$ is convex in x. This finally proves that $A(V_k(x))$ is convex in x.

$S(V_k(x))$ **is increasing in** x**.** For $x \geq 0$, we have

$$S(V_k(x+1)) - S(V_k(x)) = \min(x,s)(V_k(x) - V_k(x-1)) + (s - \min(x+1,s))(V_k(x+1) - V_k(x)) \geq 0, \quad (19)$$

since V_k is increasing in x.

$S(V_k(x))$ **is convex in** x**.** For $x \geq 0$, we have

$$\begin{aligned} S(V_k(x+2)) + S(V_k(x)) - 2S(V_k(x+1)) &= \min(x,s)(V_k(x+1) + V_k(x-1) - 2V_k(x)) \\ &+ (\min(x+2,s) - \min(x,s))V_k(x+1) - 2(\min(x+1,s) - \min(x,s))V_k(x) \\ &+ (s - \min(x+2,s))(V_k(x+2) + V_k(x) - 2V_k(x+1)) + (\min(x+2,s) - \min(x,s))V_k(x) \\ &- 2(\min(x+2,s) - \min(x+1,s))V_k(x+1) \\ &\geq (2\min(x+1,s) - \min(x,s) - \min(x+2,s))(V_k(x+1) - V_k(x)). \end{aligned} \quad (20)$$

If $x \leq s-2$, then $2\min(x+1,s) - \min(x,s) - \min(x+2,s) = 2(x+1) - x - (x+2) = 0$.

If $x = s-1$, then $2\min(x+1,s) - \min(x,s) - \min(x+2,s) = 2s - (s-1) - s = 1$.

If $x \geq s$, then $2\min(x+1,s) - \min(x,s) - \min(x+2,s) = 2s - s - s = 0$.

In all cases $2\min(x+1,s) - \min(x,s) - \min(x+2,s) \geq 0$. Therefore, since V_k is increasing in x, $S(V_k)$ is convex in x. □

B Proof of Proposition 1

Proof. Assume that $c_R \frac{\lambda(x) - \lambda(x+1) + \min(x+1,s) - \min(x,s)}{\lambda(0)+s} \geq c_N$ for $x \geq 0$. We prove by induction on k that $V_k(x) + c_R \geq V_k(x+1)$, for $x \geq 0$. For $k = 0$, we have $V_0(x) = V_0(x+1)$, for $x \geq 0$. Hence, $V_0(x) + c_R \geq V_0(x+1)$. Assume now that for a given $k \geq 0$, we have $V_k(x) + c_R \geq V_k(x+1)$ for $x \geq 0$. We deduce that

$$\begin{aligned} V_{k+1}(x) + c_R - V_{k+1}(x+1) &= -c_N + \frac{\lambda(x+1)}{\lambda(0)+s}(V_k(x+1) + c_R - V_k(x+2)) + \frac{\lambda(x) - \lambda(x+1)}{\lambda(0)+s}V_k(x+1) \\ &+ \frac{\min(x,s)}{\lambda(0)+s}(V_k(x-1) + c_R - V_k(x)) + \frac{\min(x,s) - \min(x+1,s)}{\lambda(0)+s}V_k(x) \\ &+ \left(1 - \frac{\lambda(x)}{\lambda(0)+s} - \frac{\min(x+1,s)}{\lambda(0)+s}\right)(V_k(x) + c_R - V_k(x+1)) - \frac{\lambda(x) - \lambda(x+1)}{\lambda(0)+s}V_k(x+1) \\ &- \frac{\min(x,s) - \min(x+1,s)}{\lambda(0)+s}V_k(x) + c_R\frac{\lambda(x) - \lambda(x+1) + \min(x+1,s) - \min(x,s)}{\lambda(0)+s} \\ &\geq c_R\frac{\lambda(x) - \lambda(x+1) + \min(x+1,s) - \min(x,s)}{\lambda(0)+s} - c_N \geq 0. \end{aligned}$$

This proves the induction step and finishes the proof of the proposition. □

C Proof of Theorem 2

Proof. From Equation (3), we have

$$\begin{cases} \min(x,s)\Delta(x) - \lambda(x)\Delta(x+1) \leq c_N x - g, \text{ and,} \\ \min(x,s)\Delta(x) - \lambda(x)c_R \leq c_N x - g, \end{cases} \tag{21}$$

for $x > 0$. So,

$$\begin{cases} p(x)(\min(x,s)\Delta(x) - \lambda(x)\Delta(x+1)) \leq p(x)(c_N x - g), \text{ and,} \\ (1-p(x))(\min(x,s)\Delta(x) - \lambda(x)c_R) \leq (1-p(x))(c_N x - g), \end{cases} \tag{22}$$

for $x > 0$. Summing up the above equations and multiplying the result by q_x leads to

$$\min(x,s)\Delta(x)q_x - p(x)\lambda(x)\Delta(x+1)q_x - (1-p(x))\lambda(x)c_R q_x \leq (c_N x - g)q_x,$$

for $x > 0$. From Equation (5), we have $\lambda(x-1)p(x-1)q_{x-1} = \min(s,x)q_x$, for $x > 0$. Therefore, we get

$$\lambda(x-1)p(x-1)q_{x-1}\Delta(x) - p(x)\lambda(x)\Delta(x+1)q_x - (1-p(x))\lambda(x)c_R q_x \leq (c_N x - g)q_x, \tag{23}$$

for $x > 0$. We assumed that $\Delta(x+1)$ is bounded. Moreover, we restrict the analysis to *stable* stationary policies. Therefore, $p(x)\lambda(x)$ is also bounded and q_x tends to zero as x tends to infinity. Thus, we have $\lim_{x \longrightarrow \infty} p(x)\lambda(x)\Delta(x+1)q_x = 0$, and $\sum_{x=1}^{\infty} [\lambda(x-1)p(x-1)q_{x-1}\Delta(x) - p(x)\lambda(x)\Delta(x+1)q_x] = \lambda(0)p(0)q_0\Delta(1)$, $\sum_{x=1}^{\infty} q_x = 1 - q_0$, and $g^S = \sum_{x=0}^{\infty} (1-p(x))\lambda(x)c_R q_x + x c_N q_x$. Therefore, summing up Equation (23) for x from 1 to infinity, leads to

$$\lambda(0)p(0)q_0\Delta(1) + g(1-q_0) \leq g^S - \lambda(0)(1-p(0))c_R q_0. \tag{24}$$

Moreover, Equation (4) leads to

$$g q_0 \leq \lambda(0)q_0(p(0)\Delta(1) + (1-p(0))c_R). \tag{25}$$

Summing up Equations (24) and (25) finally proves that $g \leq g^S$. □

D Proof of Lemma 1

Proof. We prove the result by induction on x, for $x \in \{1, 2, \cdots, \min(n_1, n_2) + 1\}$. Assume that $g^{n_1} \geq g^{n_2}$, for $n_1, n_2 \in \mathbb{N}$. From Equation (6), we have $\Delta^{n_1}(1) = \frac{g^{n_1}}{\lambda(0)}$ and $\Delta^{n_2}(1) = \frac{g^{n_2}}{\lambda(0)}$. So, $\Delta^{n_1}(1) \geq \Delta^{n_2}(1)$. Assume now that for a given $x < \min(n_1, n_2) + 1$, we have $\Delta^{n_1}(x) \geq \Delta^{n_2}(x)$. Equation (7) leads to

$$\Delta^{n_1}(x+1) = -\frac{c_N x - g^{n_1}}{\lambda(x)} + \frac{\min(x,s)}{\lambda(x)}\Delta^{n_1}(x), \text{ and,}$$
$$\Delta^{n_2}(x+1) = -\frac{c_N x - g^{n_2}}{\lambda(x)} + \frac{\min(x,s)}{\lambda(x)}\Delta^{n_2}(x).$$

Subtracting these two equations leads to

$$\Delta^{n_1}(x+1) - \Delta^{n_2}(x+1) = \frac{g^{n_1} - g^{n_2}}{\lambda(x)} + \frac{\min(x,s)}{\lambda(x)}\left(\Delta^{n_1}(x) - \Delta^{n_2}(x)\right) \geq 0.$$

This finishes the proof of the lemma. □

E Proof of Lemma 2

Proof. Since the cost parameters are all positive, we have $g^n \geq 0$. Therefore, Equation (6) proves that $\Delta^n(1) \geq 0$. We also know that $\Delta^n(n+1) = c_R \geq 0$. Equation (7) can be rewritten as

$$\Delta^n(x+1) = \frac{\min(x,s)}{\lambda(x)}\Delta^n(x) + \frac{g^N - c_N x}{\lambda(x)}.$$

The above relation indicates that if $g^n - c_N x \geq 0$ and $\Delta^n(x) \geq 0$ then $\Delta^n(x+1) \geq 0$.

Now suppose $\Delta^n(2) < 0$. A necessary condition to have $\Delta^n(2) < 0$ given that $\Delta^n(1) \geq 0$ is to have $g^n - c_N \cdot 2 < 0$. This implies $g^n - c_N \cdot x < 0$ and $\Delta^n(x) < 0$ for $x \geq 2$. This contradicts $\Delta^n(n+1) = c_R \geq 0$. So, $\Delta^n(2) \geq 0$. Analogous arguments yield $\Delta^n(x) \geq 0$ for all $1 \leq x \leq n+1$. □

F Proof of Theorem 3

Proof. Assume that there exists $m > 0$ such that $g^m < g^k$, for $k \in \{0, 1, 2, \cdots, m-1\}$ and $g^{m+1} \geq g^m$. We assume that $m \geq s$. The case $m < s$ can be treated in an identical way.

Consider now the modified problem

$$c_N \min(x, m+1) - \mathbb{1}_{x>m}(g^{m+1} - g^m) - \tilde{g} = \min(\min(x, s), \min(m+1, s))\tilde{\Delta}(x) \\ - \min(\lambda(x), \lambda(m+1)) \min(\tilde{\Delta}(x+1), c_R), \tag{26}$$

for $x > 0$, and

$$\tilde{g} = \lambda(0) \min(\tilde{\Delta}(1), c_R), \tag{27}$$

for $x = 0$.

Since $g^m \leq g^{m+1}$, Lemma 1 indicates that $\Delta^m(x) \leq \Delta^{m+1}(x)$, for $x \in \{1, \cdots, m+1\}$. Since $\Delta^m(m+1) = c_R$, we deduce that $\Delta^{m+1}(m+1) \geq c_R$. Moreover, since $g^m \leq g^k$ for $k \in \{1, 2, \cdots, m\}$, Lemma 1 indicates that $\Delta^m(x) \leq \Delta^k(x)$, for $x \in \{1, \cdots, k+1\}$. Since $\Delta^k(k+1) = c_R$, we deduce that $\Delta^m(k) \leq c_R$, for $k \in \{1, 2, \cdots, m\}$. The inequalities $\Delta^m(x) \leq c_R$, for $x \in \{1, 2, \cdots, m\}$ and $\Delta^{m+1}(m+1) \geq c_R$ prove that $\tilde{g} = g^m$, $\tilde{\Delta}(x) = \Delta^m(x)$ for $1 \leq x \leq m$, and $\tilde{\Delta}(x) = \Delta^{m+1}(m+1)$ for $x > m$ is the solution of the modified problem (26)-(27). Therefore, the threshold policy with threshold m is optimal for the modified problem with associated optimal cost g^m.

Assume that m is not the optimal threshold for the original problem. Then, there exists $n \geq m+2$ such that $n = \inf\{k \geq m+2 : g^k < g^m\}$. We denote by $\tilde{g}^n$ the long-run average expected cost of a threshold policy with level n for the modified problem. Similarly, let $\tilde{\Delta}^n(.)$ be the relative cost difference for the associated modified problem. These values will coincide with those of the original problem for $k \leq m+1$. We want to prove that $\tilde{g}^n < g^n$ which would imply that such n cannot exist because $g^m < \tilde{g}^n$ (i.e., m is the optimal threshold level for the modified problem). Assume that $\tilde{g}^n - g^n \geq 0$. Therefore, we have

$$\tilde{g}^n = -(g^{m+1} - g^m) + c_N(m+1) - s\tilde{\Delta}^n(n) + \lambda(m+1)c_R, \text{and,} \tag{28}$$

$$g^n = c_N n - s\Delta^n(n) + \lambda(n)c_R \tag{29}$$

Subtracting Equation (29) from Equation (28) leads to

$$\tilde{g}^n - g^n = -(g^{m+1} - g^m) - c_N(n - (m+1)) + c_R(\lambda(m+1) - \lambda(n)) + s(\Delta^n(n) - \tilde{\Delta}^n(n)). \tag{30}$$

Let us now prove that $\Delta^n(n) - \tilde{\Delta}^n(n) < 0$. Under the assumption $\tilde{g}^n - g^n \geq 0$, Lemma 1 applies directly for states $x = 1, 2, \cdots, m+1$ giving $\Delta^n(x) \leq \tilde{\Delta}^n(x)$, for $1 \leq x \leq m+1$. For $x \geq m+2$, we show by induction

on x that $\Delta^n(x) - \tilde{\Delta}^n(x) \leq 0$. Since $m \geq s$, we may write

$$\begin{aligned}\lambda(m+1)(\Delta^n(x) - \tilde{\Delta}^n(x)) = s(\Delta^n(x-1) - \tilde{\Delta}^n(x-1)) + g^n - \tilde{g}^n + g^m - g^{m+1} \\ - c_N(x-2-m) + (\lambda(m+1) - \lambda(x-1))\Delta^n(x)\end{aligned} \tag{31}$$

We have $\Delta^n(x-1) - \tilde{\Delta}^n(x-1) \leq 0$, due to the induction assumption. We assumed that $g^n - \tilde{g}^n \leq 0$ and $g^m - g^{m+1} \leq 0$. Finally, Lemma 2 proves that $\Delta^n(x) \geq 0$ and since $x \geq m+2$, we have $\lambda(m+1) - \lambda(x-1) \leq 0$. Therefore, Equation (31) proves that $\Delta^n(x) - \tilde{\Delta}^n(x) \leq 0$, for $x \geq m+2$. Equation (30) then proves that $\tilde{g}^n - g^n < 0$ which is in contradiction with our initial assumption. This shows that n cannot exist and $g^m < g^k$, for $k \geq 0$. □

G Proof of Proposition 2

Proof. We prove the case $-c_N + (1 + \lambda(m) - \lambda(m+1))c_R > 0$, for $0 \leq m \leq s$. The case $-c_N + (1 + \lambda(m) - \lambda(m+1))c_R \leq 0$, for $m \geq 0$ can be proven with the same approach. Since $\Delta^0(1) = c_R$, Equation (6) leads to $g^0 = \lambda(0)c_R$. Using Equations (6) and (7), we may write $g^1 = \lambda(0)\Delta^1(1)$, and, $c_N - g^1 = \Delta^1(1) - \lambda(1)c_R$. This leads to $g^1 = \frac{\lambda(0)(c_N + \lambda(1)c_R)}{\lambda(0)+1}$. Therefore, we have $g^1 - g^0 = \frac{\lambda(0)(c_N + c_R(\lambda(1) - \lambda(0) - 1))}{\lambda(0)+1}$. We thus have, $g^1 < g^0$ if and only if $c_N < (1 + \lambda(0) - \lambda(1))c_R$. Assume now that $g^0 > g^1 > \cdots > g^{m-1} > g^m$, with $m < s$. We want to show that $g^m > g^{m+1}$. Let us assume that $g^{m+1} \geq g^m$. Since $g^{m-1} > g^m$, Lemma 1 indicates that $\Delta^m(x) < \Delta^{m-1}(x)$, for $x \in \{1, 2, \cdots, m\}$. Since $g^{m+1} \geq g^m$, Lemma 1 indicates that $\Delta^m(x) \leq \Delta^{m+1}(x)$, for $x \in \{1, 2, \cdots, m+1\}$. Moreover, $\Delta^{m-1}(m) = c_R$ and $\Delta^m(m+1) = c_R$. So, $\Delta^m(m) < c_R$ and $\Delta^{m+1}(m+1) \geq c_R$. Equation (7) leads to

$$\begin{aligned}g^m &= c_N m - m\Delta^m(m) + \lambda(m)c_R, \text{and,} \\ g^{m+1} &= c_N(m+1) - (m+1)\Delta^{m+1}(m+1) + \lambda(m+1)c_R.\end{aligned}$$

Since $\Delta^m(m) < c_R$ and $\Delta^{m+1}(m+1) \geq c_R$, we have

$$\begin{aligned}g^m &> c_N m - mc_R + \lambda(m)c_R, \text{and,} \\ g^{m+1} &\leq c_N(m+1) - (m+1)c_R + \lambda(m+1)c_R.\end{aligned}$$

The two above inequalities lead to $g^m > g^{m+1} - c_N + (1 + \lambda(m) - \lambda(m+1))c_R > g^{m+1}$, since $-c_N + (1 + \lambda(m) - \lambda(m+1))c_R > 0$. This is in contradiction with $g^{m+1} \geq g^m$ and proves that $g^{m+1} < g^m$. □

H Proof of Proposition 3

Proof. Consider the following modified problem:

$$-\lambda(n)(c_R - \Delta^n(n)) - g' = -\lambda(0)\min(\Delta'(1), c_R), \tag{32}$$

$$c_N x - \lambda(n)(c_R - \Delta^n(n)) - g' = \min(x, s)\Delta'(x) - \lambda(x)\min(\Delta'(x+1), c_R), \text{ for } 1 \le x < n, \text{ and} \tag{33}$$

$$c_N n - (\lambda(x) - \lambda(n))\Delta^n(n) - g' = \min(x, s)\Delta'(x) - \lambda(x)\min(\Delta'(x+1), c_R), \text{ for } x \ge n. \tag{34}$$

For this modified problem, the cost per customer in the system of the original problem, $c_N x$, for $x \ge 0$, is replaced in the modified problem by $c_N x - \lambda(n)(c_R - \Delta^n(n))$, for $0 \le x \le n$ and by $c_N n - (\lambda(x) - \lambda(n))\Delta^n(n)$, for $x \ge n$. Since $\lambda(x)$ is increasing in x and $0 \le \Delta^n(n) \le c_R$ (Lemma 1 and Lemma 2), the cost per customer in the modified problem is reduced as compared to the cost per customer in the original problem. So, $g' \le g$.

For the modified problem, we observe that $g' = g^n - \lambda(n)\,(c_R - \Delta^n(n))$, $\Delta'(x) = \Delta^n(x)$, for $1 \le x \le n$ and $\Delta'(x) = \Delta^n(n)$, for $x \ge n$ is the solution of Equations (32)-(34). Therefore, we have $g^n - \lambda(n)\,(c_R - \Delta^n(n)) \le g$. We have $\overline{g} \ge g$ as stationary policies are a restriction of the set of admissible policies. Therefore, when g^n is decreasing in n, we have $g^n - \overline{g} \le g^n - g \le \lambda(n)\,(c_R - \Delta^n(n))$ which proves the result. □

I Proof of proposition 4

Proof. let us start with the decreasing property of the expected cost in s for a fixed rejection threshold, n. Since we are interested in the effect of s when n is kept constant, we denote the expected cost by g^s and the relative difference by Δ^s, when s servers are present in the system, for $1 \le s \le n$. Assume that for a given s, such that $1 \le s \le n-1$, we have $g^s \le g^{s+1}$. In what follows, we show that this leads to a contradiction. From Equation (6), we have $\lambda(0)(\Delta^s(1) - \Delta^{s+1}(1)) = g^s - g^{s+1}$. Therefore, since $g^s \le g^{s+1}$, we have $\Delta^s(1) \le \Delta^{s+1}(1)$. Assume that $\Delta^s(x) \le \Delta^{s+1}(x)$, for $x \ge 1$. Equation (7) can be rewritten as

$$\lambda(x)(\Delta^s(x+1) - \Delta^{s+1}(x+1)) = g^s - g^{s+1} + \min(x, s)\Delta^s(x) - \min(x, s+1)\Delta^{s+1}(x), \tag{35}$$

for $1 \le x \le n$. Since $\min(x, s+1) \ge \min(x, s)$, $\Delta^s(x) \le \Delta^{s+1}(x)$, and $g^s \le g^{s+1}$, Equation (35) proves that $\Delta^s(x+1) \le \Delta^{s+1}(x+1)$. Therefore, by induction on x, we proved that $\Delta^s(x) \le \Delta^{s+1}(x)$, for $1 \le x \le n$. By replacing x by n in Equation (35), given that $\Delta^s(n+1) = \Delta^{s+1}(n+1) = c_R$, we get

$$g^{s+1} - g^s = \min(n, s)\Delta^s(n) - \min(n, s+1)\Delta^{s+1}(n) \le 0.$$

This is in contradiction with the initial assumption. Therefore, we have $g^s > g^{s+1}$.

We employ the same approach to prove that g^s is convex in s. Assume that there exists s such that $2g^{s+1} - g^s - g^{s+2} \geq 0$. Equation (6) leads to $\lambda(0)(2\Delta^{s+1}(1) - \Delta^s(1) - \Delta^{s+2}(1)) = 2g^{s+1} - g^s - g^{s+2}$. Therefore, since $2g^{s+1} - g^s - g^{s+2} \geq 0$, we have $2\Delta^{s+1}(1) - \Delta^s(1) - \Delta^{s+2}(1) \geq 0$. Assume that $2\Delta^{s+1}(x) - \Delta^s(x) - \Delta^{s+2}(x) \geq 0$, for $x \geq 1$. Equation (7) can be rewritten as

$$\begin{aligned}\lambda(x)(2\Delta^{s+1}(x+1) - \Delta^s(x+1) - \Delta^{s+2}(x+1)) &= 2g^{s+1} - g^s - g^{s+2} \\ &+ 2\min(x, s+1)\Delta^{s+1}(x) - \min(x, s)\Delta^s(x) - \min(x, s+2)\Delta^{s+2}(x),\end{aligned} \tag{36}$$

for $1 \leq x \leq n$.

Case 1: $x \leq s$. In this case, we have

$$2\min(x, s+1)\Delta^{s+1}(x) - \min(x, s)\Delta^s(x) - \min(x, s+2)\Delta^{s+2}(x) = x(2\Delta^{s+1}(x) - \Delta^s(x) - \Delta^{s+2}(x)) \geq 0.$$

Case 2: $x \geq s+2$. For this case, we get

$$\begin{aligned}2\min(x, s+1)\Delta^{s+1}(x) - \min(x, s)\Delta^s(x) - \min(x, s+2)\Delta^{s+2}(x) &= s(2\Delta^{s+1}(x) - \Delta^s(x) - \Delta^{s+2}(x)) \\ &+ 2(\Delta^{s+1}(x) - \Delta^{s+2}(x)).\end{aligned}$$

Since g^s is decreasing in s, we can prove by induction that $\Delta^s(x)$ is also decreasing in s. This proves that $2\min(x, s+1)\Delta^{s+1}(x) - \min(x, s)\Delta^s(x) - \min(x, s+2)\Delta^{s+2}(x) \geq 0$.

Case 3: $x = s+1$. This case leads to

$$\begin{aligned}2\min(x, s+1)\Delta^{s+1}(x) - \min(x, s)\Delta^s(x) - \min(x, s+2)\Delta^{s+2}(x) &= s(2\Delta^{s+1}(x) - \Delta^s(x) - \Delta^{s+2}(x)) \\ &+ (2\Delta^{s+1}(x) - \Delta^{s+2}(x)) \geq 0.\end{aligned}$$

This proves by induction on x that $2\Delta^{s+1}(x) - \Delta^s(x) - \Delta^{s+2}(x) \geq 0$, for $x \geq 1$. As for the decreasing property, by replacing x by n in Equation (36), we obtain a contradiction which subsequently proves that g^s is convex in s. □

J Proof of Proposition 5

Proof. Let us denote by $E(R)_n$, the throughput of rejected customers for threshold level n ($n \geq s$). We want to obtain an equivalent expression of $E(R)_n$ as n tends to infinity with $n = rs$, $r \geq 1$ and $\lambda = qs$ with $q > 0$. Observe that $\frac{E(R)_n}{E(R)_{n+1}} = \frac{s}{\lambda+(n+1)\gamma}\left(1 + \frac{E(R)_n}{s}\right)$. This relation leads to $E(R)_{n+1} = E(R)_n \frac{\lambda+(n+1)\gamma}{s+E(R)_n} = E(R)_n \frac{n\left(\frac{q}{r}+\gamma\right)+\gamma}{n\frac{1}{r}+E(R)_n}$, for $n \geq s$. We define the sequence u_n by $E(R)_n = n\frac{1}{r}u_n$, for $n \geq s$. This leads to $u_{n+1} = \frac{u_n}{1+u_n}\left(q + r\gamma + \frac{-q}{(n+1)}\right)$, for $n \geq s$. We now consider the sequence v_n defined by $v_{n+1} = f(v_n)$, for $n \geq s$, where $f(x) = \frac{x}{1+x}(q + r\gamma)$, for $x > 0$. As n tends to infinity u_n and v_n are equivalent.

Case 1: $q + r\gamma - 1 > 0$**.**

We have $f(0) = 0$, and $f(q+r\gamma-1) = q+r\gamma-1$. Since f is increasing, we have $f((0, q+r\gamma-1]) \subset (0, q+r\gamma-1]$ and $f([q + r\gamma - 1, \infty)) \subset [q + r\gamma - 1, \infty)$. We have $v_{n+1} - v_n = \frac{v_n}{1+v_n}(q + r\gamma - 1 - v_n)$. So $v_{n+1} > v_n$ if and only if $v_n < q + r\gamma - 1$. Therefore, either $v_s < q + r\gamma - 1$ and v_n is increasing with $v_n < q + r\gamma - 1$ for $n > s$, or $v_s > q + r\gamma - 1$ and v_n is decreasing with $v_n > q + r\gamma - 1$ for $n > s$. This proves in both cases that v_n has a finite limit as n tends to infinity. This limit is $q + r\gamma - 1$. This proves that $u_n \sim q + r\gamma - 1$ as n tends to infinity and $E(R) \sim s\,(q - 1 + r\gamma)$, as n tends to infinity.

Case 2: $q + r\gamma - 1 \leq 0$**.**

In this case, $v_{n+1} - v_n = \frac{v_n}{1+v_n}(q + r\gamma - 1 - v_n) \leq 0$ since $v_n > 0$. So v_n is decreasing and $v_n > 0$. The only possible limit of v_n (and u_n) as n tends to infinity is therefore zero. In a similar way, using $E(N)_{n+1} = \frac{E(N)_n + (n+1)\frac{E(R)_n}{s}}{1+\frac{E(R)_n}{s}}$, we show the asymptotic expressions of $E(N)$ as n tends to infinity. □