Late-rejection, a strategy to perform an overflow policy

Benjamin Legros

Version of Record: https://www.sciencedirect.com/science/article/pii/S0377221719307076
Manuscript_f87a9a1cf9a05ca2a431ff0a0d1e5388

Late-rejection, a strategy to perform an overflow policy

Benjamin Legros

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

benjamin.legros@centraliens.net

Abstract

Motivated by overflow policies implemented in service systems, we consider a multi-server queue with customers' abandonment where rejection control is exercised on customers currently waiting in the queue. Our aim is to find a good balance between conflicting goals, namely, the rate of rejected customers and a cost function which may involve wait and abandonment metrics like percentiles of the waiting time or rate of abandonment. We develop a Markov decision process approach where the waiting time of the first customer in line is used in a discretized form to define the system state. We show that a time-based threshold policy is optimal, and develop a procedure to compute the optimal threshold. Our analysis explains some known behaviors in practice. For instance, if the cost function is constant in the system state like with wait percentiles, then the optimal threshold is one of the time limits defining the percentiles. Also, abandonment is shown to have beneficial or detrimental effect depending on the system manager's objective.

Keywords: Queueing systems; Markov decision process; performance evaluation; threshold policy; Erlang approximation; routing; rejection; abandonment.

1 Introduction

In many service systems, a timeout threshold is set by administrators in order to control the time spent in the system by customers. In particular, time-based-rejection policies can be found in service contact centers. Some call centers use CTI software to manage call rejection after a certain waiting time threshold. The threshold value is adjusted by the customers (the call centers) according to their business requirements. For instance, it is 6 minutes for the energy company Primagaz, 5 minutes for the pharmaceutical company Sanofi, 3 minutes for the telecom operator Keyyo' commercial call center and 15 minutes for its technical hotline. For Carglass, a callback option is also proposed at a waiting time threshold of 10 minutes. Another illustration of task rejection in contact centers is the use of automatic replies for emails. Based on a list of key words, an automatic email is sent in reply to an email that has not been answered within an acceptable response time. Another case where time-based policies are employed is in IVR systems. For instance, robot agents may be used if the

waiting time is too long, or live agents may take over a service if the service time with a robot is too long (or unsuccessful). Finally, in Emergency Departments, the patient-priority is often increased after reaching some wait-thresholds.

One objective for the controller in the aforementioned examples is to reduce the customer flow by rejecting or proposing a service alternative to some customers. Thus, the *late-rejection* policies presented above belong to the family of *overflow policies*. Late-rejection policies however differ from the rejection strategies studied in the academic literature. In most of these studies, customers are rejected or rerouted *upon arrival* (Akşin et al., 2008; Ren and Zhou, 2008; Koçağa and Ward, 2010; Schrieck et al., 2014). The decision to reject a newly arrived customer is based on the system state, i.e., on its expected waiting time. The alternative, proposed here, is to accept that the new customer joins the queue, but allow it to be rejected later if its experienced waiting time reaches a timeout threshold.

Intuitively, late-rejection seems worse than rejection at arrival. Rejecting customers after a wait (late-rejection) is particularly customer-unfriendly in the sense that rejected customers are doubly punished; they wait and then they are rejected. This should preclude the implementation of late-rejection policies. However, late-rejection is nonetheless of interest. Compared to rejection at arrival, its main operational value is to improve management of (i) service time variability, and (ii) non-linear objectives. Variability in service times may lead to a shorter realized wait than the expected wait evaluated at arrival. Therefore, rejecting customers at arrival could lead to wrong decisions if the idea is to serve customers with sufficiently short waits. Consider for instance an M/M/1 queue with an expected service duration of 4 minutes. If a customer arrives when 5 customers are already present in the system, then the expected waiting time is equal to 20 minutes. This might seem long, and a controller may choose to reject this customer. However, the customer has a 56% chance of waiting less than 20 minutes, a 32% chance of waiting less than 15 minutes, an 11% chance of waiting less than 10 minutes, and a 1% chance of waiting less than 5 minutes. So, it might worth trying to serve this customer.

Moreover, in many service systems, the quality of service is evaluated based on percentiles of waiting time. This metric is often preferred to the average speed of answer as the former was perceived to be more informative (Bailey and Sweeney (2003)). For instance, a minimum service level of 80% of customers served in less than 20 seconds is common in call centers (Aksin et al. (2007),Koole (2013)), while a minimum service level of 90% of patients served in less than four hours is used in emergency departments (Thompson et al. (1996)). Wait percentiles can be seen as

the Value-at-Risk of a service system (Lotfi and Zenios, 2018). For percentile objectives, time-based decisions seem valuable. For instance, with a time-based rejection threshold, 0% of customers would wait more than the timeout threshold. On the contrary, with rejection at arrival, the future wait is not perfectly controlled due to the variability in the service times.

We should also mention that although late-rejection seems useful for handling non-linear objectives and service time variability, it appears to have other advantages for contact center managers. First, it reduces the attractiveness of the competition. If a customer has already waited some time at a contact center and has paid a cost per minute of wait, then an investment has been made to receive a service, and callers will be less likely to contact the competition. So they will either try to call back the same call center or agree to be called back later. When a callback offer is proposed at a caller's arrival, no investment is made, and the announcement of a long waiting time and/or a long delay for being called back may cause callers to balk (and try to obtain the service elsewhere). Another aspect is related to the customer's learning experience. A recent article by Emadi and Swaminathan (2018) shows that new callers who do not have any prior experience with the call center tend to be optimistic about their delay in the system and underestimate its length. Dissatisfaction due to too long wait may then be stronger than that of accustomed callers. Making customers wait before being rejected is one way of giving them an experience of the congestion. The call center may then expect these customers to be less demanding in the future.

In this paper, our main aim is to analyze a queue in which late-rejection control is exercised. To understand the implementation of this policy, we formulate a time-based rejection control problem for a stationary M/M/s+G queue when there are costs associated with customer rejection, customer abandonment, and customer wait. The wait-cost can be a function of the expected waiting time, percentiles of the waiting time or may involve higher moments of the wait. Therefore, the late-rejection control problem can be formulated as one that tries to minimize the system cost by providing a good trade-off between the waiting or abandonment cost and the rejection cost. More precisely, the question we wish to analyze is *how long should a customer remain in the queue before being rejected?* The main difficulty in the analysis is the decision variable, namely, the customer wait. The system therefore cannot be modeled as a simple Markov decision process where a state of the system corresponds to the number of customers. To overcome this difficulty, we approximate the waiting time of the first customer in line in the queue by an Erlang distribution as in Koole et al. (2012) and Legros et al. (2017). This allows us to represent the system evolution by a Markov chain. As the parameters of the Erlang distribution tend to infinity, the approximated model converges to

the real one. Another difficulty is the non-linearity of the objective function which does not allow us to prove the form of the optimal policy using an induction step approach. We overcome this difficulty by studying the properties of the relative value function directly. We show that a waiting time-based threshold control policy is optimal for our problem. Therefore, our model is reduced to an M/M/s+min(G,D) queue where the deterministic rejection threshold is an endogenous controlled parameter. This result is a natural translation, with time-based decisions of the queue-size threshold policy for controlling rejections at arrival resulting in an M/M/s+G/s+n queue. In order to compute the optimal threshold, using the approximated model via an n-terminating formulation as in Koçağa and Ward (2010) and Adusumilli and Hasenbein (2010), we prove that the first local minimum found by increasing the time-threshold is necessarily a global one when the patience has the decreasing failure rate property. Other practical results are also derived. We show that if the cost function is constant, for instance if it is made up of percentiles of the waiting time, then the optimal threshold is one of the time limits defining the percentiles. The role of abandonment is also investigated. We show that although abandonment is perceived as a negative phenomenon, it may be beneficial when the system manager is focused on the service quality of served customers.

Structure of the article. The remainder of the article is structured as follows. Section 2 reviews the related literature, while Section 3 defines the optimization problem. Section 4 investigates the computation of the optimal policy, and Section 5 shows the applicability of these results. Finally, Section 6 concludes the article. The proofs are given in the Appendix.

2 Literature review

We distinguish two streams of literature related to this paper. The first considers queueing models where decisions are based on the customers' waiting experience, while the second deals with admission control problems. One particularity of our queueing model is that decisions are taken according to the experienced waiting time of the oldest customer in the queue. Although using the waiting time as a decision variable is common in practice, the academic literature almost always focuses on quantity-based policies, where the number of customers is the decision variable. One reason is the difficulty of providing a Markov chain analysis when the wait is the decision variable. To overcome this difficulty, Koole et al. (2012) created a tool to develop Markov decision processes analysis where the first-in-line waiting time is used as a decision variable. Later, Legros et al. (2017) extended this method to queueing models with rejection and abandonment. Yet, the complexity of the transition

structure provided in these references makes it complicated to prove the optimality of a policy using this method. Hence, for policy computation, only numerical applications (Koole et al., 2015; Legros, 2018) have so far been considered using this method. For performance evaluation, this method has been successfully employed with the restrictive assumption of a deterministic abandonment (Legros, 2016, 2019). In this article, we attempt to tackle the limitations of existing results and provide a theoretical method to compute the optimal rejection policy.

A famous routing problem in the queueing literature is the *admission control problem* (Ku and Jordan, 2003; Maglaras and Van Mieghem, 2005; Ward and Kumar, 2008; Xu, 2015; Niyirora and Zhuang, 2017; Bountali and Economou, 2017). For a given optimization problem where a trade-off between the wait and the rejection flow has to be determined, a controller has to decide whether or not to keep a customer in the system. In most of these studies, the decision is taken upon arrival based on the number of customers in the system. This differs from our setting where the control variable is the experienced waiting time. We refer to Hassin and Haviv (2003) for an overview of classic routing solutions for individual and social optimization with observable and non-observable simple queues. In a complex queueing network of service facilities, Cosyn and Sigman (2004) investigate the admission control problem with queueing and reneging from a revenue maximizer perspective. Using orbiting as an approximation of queueing, they show that a target tracking policy is close to optimal. Lin and Ross (2004) analyze a single server loss queueing system. A gatekeeper has to decide whether to admit a customer without knowing the status (idle or busy) of the server. They show that a threshold policy which blocks arrivals for a certain time period following each admission and then admits the next customer is optimal. The present paper shows the benefits of time control as found in our study. Bassamboo et al. (2005) consider a service system model with several customer classes, server pools and doubly stochastic arrivals. A double control is exercised: rejection control at arrival and routing control to a given pool after a certain wait. Under asymptotic assumptions on the system parameters, they show 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. They showed that the larger the system becomes, the more negligible the dependency between each in-house station and overflow station. 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 quantity-threshold for their respective models. This method is employed in this article to compute the optimal time-dependent threshold.

3 System modeling and the optimization problem

We consider a multi-server queue with s servers, infinite capacity, and a first-come-first-served (FCFS) discipline. The arrival process of customers is Poisson with rate λ, and the service times are assumed to be independent and exponentially distributed with rate μ. Moreover, we assume that a customer accepts to wait with probability b $(0 \leq b \leq 1)$ or balks with probability $1 - b$. Finally, a waiting customer has finite patience and will abandon if the waiting time exceeds a random time that is *generally distributed*. Based on the wait experienced by a customer in the queue, a controller may decide at any instant to reject this customer from the system. The objective is to develop a waiting time-based stationary rejection control policy that minimizes the long-run average expected cost. The idea is to use rejection to control the system congestion and subsequently reduce the wait-cost function and the abandonment from the queue.

The particularity of this optimization problem is that decisions are taken based on time-related information rather than quantity-based information as in most control problems. Therefore, using the number of customers in the system as a state variable may not allow us to reach the optimal policy. Moreover, a quantity-based Markov decision process formulation fails to capture percentiles of the wait when those are involved in the wait cost function. To overcome this problem, we choose to explicitly model the waiting time of the First Customer in Line (FIL) in a dicretized form in the system state of a continuous time Markov chain as in Koole et al. (2012) and Legros et al. (2017). This approximation model is referred to as the *Erlang Approximation* (EA).

Let us denote by x a state of the corresponding Markov chain, where $x \geq -s$. States with $-s \leq x \leq 0$ correspond to an empty queue and $s + x$ busy servers. States with $x > 0$ correspond to a situation where the FIL is waiting at phase x and all servers are busy. In states $x \leq 0$, the Markov chain is identical to the one of a classical M/M/s queue where λ-transitions (respectively $(s + x)\mu$-transitions) correspond to an increment (respectively, to a decrement) of the number of customers in the system. In state $x = 0$, after a λ-transition, a customer enters the queue and the entity FIL is created starting at phase $x = 1$. For $x > 0$, further increase of x corresponds to phase-increase of the FIL. We approximate the time spent in each waiting phase by an exponential distribution with rate γ. Therefore, a γ-transition from phase $x > 0$ corresponds to an increase

of the waiting phase of the FIL. Note that with $x > 0$, the number of customers in the system is ignored.

Having large values of γ improves the approximation as it better represents the continuously elapsing time. As γ tends to infinity, this approximate setup converges to the original one, which leads to an exact analysis. This result comes from the convergence in distribution of the Erlang distribution to a deterministic one. The Laplace transform in the variable y of the Erlang distribution with parameters x and γ such that $x/\gamma = t$ is $\left(\frac{\gamma}{\gamma+y}\right)^x$. We thus have

$$\left(\frac{\gamma}{\gamma+y}\right)^x = e^{x\ln((1+y/\gamma)^{-1})} \underset{\gamma\to\infty}{\sim} e^{x\ln(1-y/\gamma)} \underset{\gamma\to\infty}{\sim} e^{-xy/\gamma} = e^{-yt},$$

where we write $f(a) \underset{a\to a_0}{\sim} g(a)$ to express that $\lim_{a\to a_0} \frac{f(a)}{g(a)} = 1$, for $a_0 \in \mathbb{R}$. Applying the Levy continuity theorem for Laplace transforms, this result ensures that as x and γ go to infinity, the considered Erlang random variable converges in distribution to the deterministic duration t.

After a service completion or an abandonment, the FIL leaves the queue. Either the queue becomes empty or another customer is present in the queue and the second customer in line becomes the new FIL. The difficulty with abandonment is to determine the transition probability, because the next customer first in line, if any, is no longer necessarily the customer that arrived after the FIL who just left. The former might actually have abandoned. To overcome this difficulty, as in Legros et al. (2017), we approximate times before abandonment by an homogeneous Coxian distribution evolving with the same rate γ as the elapsing of time of the FIL. This Coxian distribution is defined by the parameters r_x for $x \geq 1$ and $r_x \in [0, 1]$. Specifically, after a γ-transition from waiting phase $x > 0$, a customer abandons with probability $1 - r_x$, or stays in the queue with probability r_x. As proven in Theorem 1 of Legros et al. (2017), any non-negative random variable representing customers' patience can be approximated as close as wanted by an homogeneous Coxian distribution.

This approximation allowed expressing the transition probabilities, $\overline{p}_{x,k}$, to move from state $x > 0$ to state k, with $0 \leq k \leq x$. From Theorem 2 of Legros et al. (2017), we have

$$\overline{p}_{x,0} = \prod_{i=1}^{x} q_i, \text{ and, } \overline{p}_{x,k} = (1-q_k) \prod_{i=k+1}^{x} q_i, \text{ for } 0 < k \leq x, \text{ where } q_i = \left[1 + \frac{b\lambda}{\gamma} \prod_{j=1}^{i} r_j\right]^{-1}.$$

In summary, the five possible transitions in the approximation model are as follows:

1. An arrival with rate λ while the queue is empty ($-s \leq x \leq 0$), which changes the state to $x+1$. If $-s \leq x < 0$, then the number of busy servers is incremented by 1. If $x = 0$, then the

FIL entity is created.

2. A service completion with rate $(s+x)\mu$ while the queue is empty ($-s < x \leq 0$), which changes the state to $x-1$. The number of busy servers is decremented by 1.

3. A service completion with rates $s\mu\overline{p}_{x,k}$ while the queue is not empty ($x > 0$), which changes the state to k, that is, the new FIL is in waiting phase k if $k > 0$ or the queue is empty if $k = 0$.

4. A phase increase which does not lead to an abandonment with rate γr_x while the queue is not empty ($x > 0$), which changes the state to $x+1$. The waiting phase of the FIL is incremented by 1.

5. A phase increase which leads to an abandonment with rate $\gamma(1-r_x)\overline{p}_{x,k}$ while the queue is not empty ($x > 0$), which changes the state to k, that is, the new FIL is in waiting phase k if $k > 0$ or the queue is empty if $k = 0$.

We propose a cost model that is flexible enough to accommodate the variety of performance metrics encountered in service systems. Among those are the rate of abandonment, the expected waiting time, percentiles of the wait, the average excess, or higher moments of the wait. For wait related metrics, we may also distinguish between customers who are served, rejected or who have abandoned the queue. In what follows, we give examples of cost function, $c(x)$, that may capture some of these metrics.

- **Abandonment.** Abandonment occurs after a $(1-b)\lambda$-transition from state $x = 0$ and a $(1-r_x)\gamma$-transition from states $x > 0$. Therefore, by defining $c(0) = (1-b)\lambda$, and $c(x) = (1-r_x)\gamma$, for $x > 0$, a cost of 1 is counted per customer who abandons the queue. With this definition of $c(x)$, we capture the rate of customers who abandon the system. With the same definition, by diving $c(x)$ by λ, we instead consider the proportion of customers who abandon the system.

- **Expected wait.** The expected time spent in each waiting phase is $1/\gamma$. Therefore, the expected wait of a customer who leaves the queue from state $x > 0$ is $\frac{x}{\gamma}$. A service occurs after a $s\mu$-transition from state $x \geq 0$. Therefore, with $c(x) = s\mu\frac{x}{\gamma}$, for $x \geq 0$, a cost of 1 is counted per time unit spent in the queue for served customers only. With this definition, we capture the product of the rate of served customers multiplied by their expected wait. By changing $s\mu$ in this definition by $\gamma(1-r_x)$ (respectively, by γr_x if a rejection is decided), we

instead consider the expected wait of customers who abandon the queue (respectively, the expected wait of customers who are rejected).

- **Wait percentile.** For a wait percentile, a penalty is paid when the wait is longer than a threshold, t^*. In our model, the wait is discretized. We then choose an integer n^*, such that $\frac{n^*}{\gamma} = t^*$. In this way, by letting n^* and γ tend to infinity, the waiting phase n^* corresponds to the time threshold t^*. Therefore, with $c(x) = s\mu \mathbb{1}_{x \geq n^*}$, for $x \geq 0$, where $\mathbb{1}_{x \in A}$ is the indicator function of a given subset A, a cost of 1 is counted per served customer from a waiting phase higher than n^*. As for the expected wait, the coefficient $s\mu$ in $c(x)$ can be changed into $\gamma(1 - r_x)$ or γr_x to consider customers who abandon the queue or who are rejected in case of rejection.

- **Average excess.** The average excess is the expected waiting time in excess of a threshold, t^*. As for wait percentiles, we choose n^* such that $\frac{n^*}{\gamma} = t^*$. For instance, with $c(x) = s\mu \frac{(x-n^*)^+}{\gamma}$, where $z^+ = \max(z, 0)$, a cost of 1 is counted per time unit spent in the queue in excess of t^* for served customers only.

- **Higher moments.** The moment-generating function of an Erlang random variable with x phases and rate γ per phase as a function of t is $\left(1 - \frac{t}{\gamma}\right)^{-x} = \sum_{i=0}^{\infty} \frac{(x-1+i)!}{i!(x-1)!} \left(\frac{t}{\gamma}\right)^i$. Therefore, by defining $c(x) = s\mu \frac{(x-1+i)!}{(x-1)!\gamma^i}$, we capture the i^{th} moment of the wait in the cost function for served customers.

The cost function, $c(x)$, can also be a linear combination of the different aforementioned examples. Finally, if a rejection is decided, we count a penalty P per rejected customer.

As for the service discipline, we apply a first-come-first-rejected (FCFR) discipline for the rejection decision. Therefore, the rejection decision is taken in priority to the FIL. More precisely, the possible actions after an elapse of time of the FIL are either to maintain the customer in the queue or to reject the customer from the system. The dynamic programming optimality equations for the

relative value function, $V(x)$, for $x \geq -s$, and an average constant cost g, are given by

$$V(-s) + g = \lambda V(-s+1) + (1-\lambda)V(-s), \tag{1}$$

$$V(x) + g = \lambda V(x+1) + (s+x)\mu V(x-1) + (1-\lambda-(s+x)\mu)V(x), \text{ for } -s < x < 0, \tag{2}$$

$$V(0) + g = c(0) + b\lambda \min(V(1), V(0)+P) + s\mu V(-1) + (1-b\lambda - s\mu)V(0), \text{ for } x = 0, \text{ and,} \tag{3}$$

$$\begin{aligned} V(x) + g = c(x) &+ \gamma r_x \min(V(x+1), F(V(x))+P) + (s\mu + \gamma(1-r_x))F(V(x)) \\ &+ (1-\gamma - s\mu)V(x), \text{ for } x > 0, \end{aligned} \tag{4}$$

where $F(V(x)) = \sum_{h=0}^{x} \overline{p}_{x,x-h} V(x-h)$ for $x \geq 1$, and $F(V(x)) = V(x)$ for $-s \leq x \leq 0$. Note that we allow the controller to reject a customer at arrival (Equation (3)) when all servers are busy and the queue is empty. However, customers are not rejected at arrival if there is at least one idle server. This is without loss of generality since rejecting a customer at arrival when a server is idle is not optimal for our optimization problem.

The non-convexity of the cost function, $c(x)$, prohibits us from showing the form of the optimal policy using an induction step approach as developed in Puterman (1994). Instead, in Section 4, we develop a method to compute the optimal policy via an analysis of the relative value function.

4 Computation of the optimal policy

We introduce the relative cost difference defined as $y(x) = V(x) - F(V(x-1))$, for $x > -s$. We rewrite Equations (1)-(4) in terms of y as follows:

$$g = \lambda y(-s+1), \tag{5}$$

$$g = \lambda y(x+1) - (s+x)\mu y(x), \text{ for } -s < x < 0, \tag{6}$$

$$g - c(0) = b\lambda \min(y(1), P) - s\mu y(0), \text{ for } x = 0, \text{ and,} \tag{7}$$

$$\begin{aligned} g - c(x) = r_x\gamma \min(&V(x+1) - V(x), F(V(x)) + P - V(x)) \\ &- (s\mu + \gamma(1-r_x))(V(x) - F(V(x))), \text{ for } x > 0. \end{aligned} \tag{8}$$

For $x > 0$, we have

$$\begin{aligned} V(x) - F(V(x)) &= V(x) - \sum_{k=1}^{x}(1-q_k)\left(\prod_{j=k+1}^{x} q_j\right)V(k) - \left(\prod_{j=1}^{x} q_j\right)V(0) \\ &= q_x V(x) - q_x\left[\sum_{k=1}^{x-1}(1-q_k)\left(\prod_{j=k+1}^{x-1} q_j\right)V(k) - \left(\prod_{j=1}^{x-1} q_j\right)V(0)\right] \\ &= q_x(V(x) - F(V(x-1))). \end{aligned}$$

An equivalent form of Equation (8) is thus

$$g - c(x) = \gamma r_x \min(y(x+1), P) - q_x(s\mu + \gamma)y(x), \text{ for } x > 0. \tag{9}$$

Theorem 1 proves that there is no randomized policy that has a strictly lower infinite horizon average expected cost than g. Let us denote the set of stationary policies for customer rejection by Ω_S. A policy in this class allows customers to enter the system whenever a server is available, associate a probability $p(0)$ to reject a customer at arrival if the queue is empty and all servers are busy and a probability $p(x)$ to reject the FIL after its x waiting phase. The set Ω_S includes threshold policies.

Theorem 1 *If g^S is the associated cost with a policy in Ω_S, then $g^S \geq g$.*

The existence of a stationary solution to Equations (5)-(8) is due to the aperiodic irreducible Markov chain considered here (see Theorem 8.5.3 part c of Puterman (1994)). The stationary solution of these equations can only lead to a deterministic threshold policy. Either $y(1) > P$, and it is optimal to reject arriving customers if all servers are busy, or the first waiting phase x such that $y(x+1) > P$ is the optimal rejection threshold. Note that even if there exists $x' > x$ such that $y(x'+1) < P$, the optimal rejection policy remains a threshold policy with threshold level x since waiting phase x' is never reached by any customer.

Equations (5)-(9) are difficult to solve explicitly due to the minimizing operator. To overcome this difficulty, we consider the $n-$terminating problem as in Koçağa and Ward (2010) and Adusumilli and Hasenbein (2010). To this end, we consider a threshold policy with waiting threshold level n.

In other words, we want to find a constant g^n and a vector $(y(1)^n, y(2)^n, \cdots, y(n+1)^n)$ such that

$$g^n = \lambda y^n(-s+1), \tag{10}$$

$$g^n = \lambda y^n(x+1) - (s+x)\mu y^n(x), \text{ for } -s < x < 0, \tag{11}$$

$$g^n - c(0) = b\lambda y^n(1) - s\mu y^n(0), \text{ for } x = 0, \tag{12}$$

$$g^n - c(x) = \gamma r_x y^n(x+1) - q_x(s\mu+\gamma)y^n(x), \text{ for } 0 < x \leq n, \tag{13}$$

and $y^n(n+1) = P$, for $n \geq 0$. The n-terminating problem is a Markovian Reward Process with finite state space. Using an induction step, we can show after some algebra that the unique solution of Equations (10)-(13) is

$$y^n(-s+x) = \frac{g^n}{\lambda}\sum_{k=0}^{x-1}\frac{(x-1)!}{(x-1-k)!}\left(\frac{\mu}{\lambda}\right)^k, \text{ for } 1 \leq x \leq s, \text{ and,} \tag{14}$$

$$y^n(x) = g^n\left[\frac{\sum_{k=0}^{s}\frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k}{b\lambda}\left(\frac{s\mu+\gamma}{\gamma}\right)^{x-1}\prod_{j=1}^{x-1}\frac{q_j}{r_j} + \sum_{k=1}^{x-1}\frac{1}{\gamma r_k}\left(\frac{s\mu+\gamma}{\gamma}\right)^{x-k-1}\prod_{j=k+1}^{x-1}\frac{q_j}{r_j}\right]$$
$$- \frac{c(0)}{b\lambda}\left(\frac{s\mu+\gamma}{\gamma}\right)^{x-1}\prod_{j=1}^{x-1}\frac{q_j}{r_j} - \sum_{k=1}^{x-1}\frac{c(k)}{\gamma r_k}\left(\frac{s\mu+\gamma}{\gamma}\right)^{x-k-1}\prod_{j=k+1}^{x-1}\frac{q_j}{r_j}, \text{ for } 1 \leq x \leq n+1.$$

Using $y^n(n+1) = P$, we get

$$g^n = \frac{P\left(\frac{\gamma}{s\mu+\gamma}\right)^n + \frac{c(0)}{b\lambda}\prod_{j=1}^{n}\frac{q_j}{r_j} + \sum_{k=1}^{n}\frac{c(k)}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k\prod_{j=k+1}^{n}\frac{q_j}{r_j}}{\frac{\sum_{k=0}^{s}\frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k}{b\lambda}\prod_{j=1}^{n}\frac{q_j}{r_j} + \sum_{k=1}^{n}\frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k\prod_{j=k+1}^{n}\frac{q_j}{r_j}}. \tag{15}$$

Using the results of Lemma 1, in Theorem 2 we prove that the first local minimum of g^n found by increasing n is the optimal threshold.

Lemma 1 *For the $n-$terminating problem, if $g^{n_1} \geq g^{n_2}$ for $n_1, n_2 \in \mathbb{N}$, then*

$$y^{n_1}(x) \geq y^{n_2}(x), \text{ for } -s+1 \leq x \leq \min(n_1, n_2)+1. \tag{16}$$

Theorem 2 *Let r_x and $c(x) - q_x P(s\mu+\gamma)$ be increasing in x and suppose that there exists a solution to the $n-$terminating problem (Equations (10)-(13)) with $g^m < g^k$, for $0 \leq k \leq m-1$ and $g^{m+1} > g^m$, and if g^S is the average cost associated with a policy in Ω_S, then we have $g^S \geq g^m$.*

g^n may be strictly decreasing in n. In this case, it is optimal not to reject any customers. In what

follows, we prove that the $n-$terminating solution converges to the optimal one in the decreasing case. In Lemma 2, we prove that there exists a solution to Equations (5)-(9) in the decreasing case for a modified cost function.

Lemma 2 *Let r_x be increasing in x and $b\lambda < s\mu$ and suppose there exists a sequence of $n-$terminating solutions $\{g^n, y^n(-s+1), y^n(-s+2), \cdots, y^n(n)\}$ such that $g^{n+1} < g^n$ for $n \geq 0$. Then, for each $n \geq 0$, there exists a vector $(\overline{g}^n, \overline{y}^n(-s+1), \overline{y}^n(-s+2), \cdots, \overline{y}^n(n-1), \overline{y}^n(n), \overline{y}^n(n), \cdots)$ which solves Equations (5)-(9) with the modified cost function $\overline{c}(x) = c(x)$ for $0 \leq x \leq n$ and $\overline{c}(x) = c(n)$, for $x > n$. For this solution, we have*

$$\overline{g}^n = \lambda \overline{y}^n(-s+1), \tag{17}$$

$$\overline{g}^n = \lambda \overline{y}^n(x+1) - (s+x)\mu \overline{y}^n(x), \text{ for } -s < x < 0, \tag{18}$$

$$\overline{g}^n - c(0) = b\lambda \overline{y}^n(1) - s\mu \overline{y}^n(0), \text{ for } x = 0, \tag{19}$$

$$\overline{g}^n - c(x) = r_x \gamma \overline{y}^n(x+1) - q_x(s\mu + \gamma)\overline{y}^n(x), \text{ for } 0 < x < n, \tag{20}$$

$$\overline{g}^n - c(n) = r_x \gamma \overline{y}^n(n) - q_x(s\mu + \gamma)\overline{y}^n(n), \text{ for } x \geq n. \tag{21}$$

In Theorem 3 we show that if g^n is decreasing in n, then the sequence of solutions to the $n-$terminating problem converges to the optimal solution.

Theorem 3 *Let r_x be increasing in x and $b\lambda < s\mu$. If $g^n > g^{n+1}$ for $n \geq 0$, then*

$$\lim_{n \longrightarrow \infty} g^n = g^*,$$

where g^* is the solution of Equations (5)-(9).

Finally, in Proposition 1, we provide a stopping criterion in the decreasing case.

Proposition 1 *If g^n is decreasing in n, then*

$$g^n - g^* \leq \gamma r_n (P - y_n(n)).$$

We are now in a position to establish an algorithm to compute the optimal threshold using the solution of the $n-$terminating problem. The result of Theorem 2 indicates that the first local minimum for the expected cost is also the optimal one. In the decreasing case, Proposition 1 provides a stopping criterion for the algorithm. The algorithm is as follows:

Algorithm 1: Computation of the optimal rejection threshold.

1. *Initialisation.* Set $n = 0$ and compute g^n, g^∞, and $y^n(n)$ using Equations (14) and (15).

2. *Iteration step:* Increase n by one and compute g^n and $y^n(n)$ using Equations (14) and (15).
   If $g^n > g^{n-1}$, then the rejection threshold $n-1$ is optimal.
   If $g^n \leq g^{n-1}$, then

   - If $g^\infty - g^n \leq \gamma r_n(P - y^n(n))$, then it is optimal not to reject any customer (i.e., $n = \infty$ is optimal).
   - Otherwise, go back to the Iteration step.

5 Applicability of the algorithm and discussions

In this section, we show the applicability of our algorithm. First, we comment the conditions needed to implement Algorithm 1. Next, we provide a numerical illustration of our algorithm. Finally, we discuss the effect of the patience on the expected cost.

Comments. Although the threshold nature of the optimal policy does not depend on additional conditions on the system parameters, the applicability of Algorithm 1 is restricted to the assumptions made to prove the related results of Section 4. Even if numerically our results hold if these conditions are not satisfied, it is interesting to explain the meaning of these assumptions.

- **"Let r_x be increasing in x."** This condition is needed to prove Theorem 2, Lemma 2, and Theorem 3. Having r_x increasing in x means that the longer a customer waits, the more patient she/he is. This characterizes patience distribution with a *decreasing failure rate* (DFR) property. As an example from practice, Jouini et al. (2013) showed that the patience distribution in a call center fits well with a hyperexponential distribution. This distribution has the DFR property. Note that r_x does not need to be *strictly* increasing in x. Thus, if customers are infinitely patient (i.e., $r_x = 1$, for $x \geq 1$) or if the patience is exponentially distributed with rate $\beta > 0$ (i.e., $r_x = \frac{\gamma}{\gamma+\beta}$, for $x \geq 1$ as in Table 1 of Legros et al. (2017)), then the results of Section 4 are still valid.

  The assumption for r_x can be understood through the system manager's objective. It consists of rejecting some customers so as to reduce the system congestion and to improve the service quality. When r_x is increasing in x, the FIL is the customer who has the most negative impact

on future congestion in the sense that the FIL has the smallest probability to abandon the queue in the future. Therefore, if one customer should be rejected, it should be in priority the FIL. This is actually what is being done while applying a FCFS and a FCFR discipline. With patience distributions which do not have the DFR property, the optimal discipline for rejection or service may not necessarily be FCFS or FCFR.

Note that for patience distributions without the DFR property, we cannot prove that the first minimum found by increasing the rejection threshold is the optimal threshold as proven in Theorem 2. However, from our numerical investigations, we couldn't find a counterexample where the result of Theorem 2 wouldn't apply.

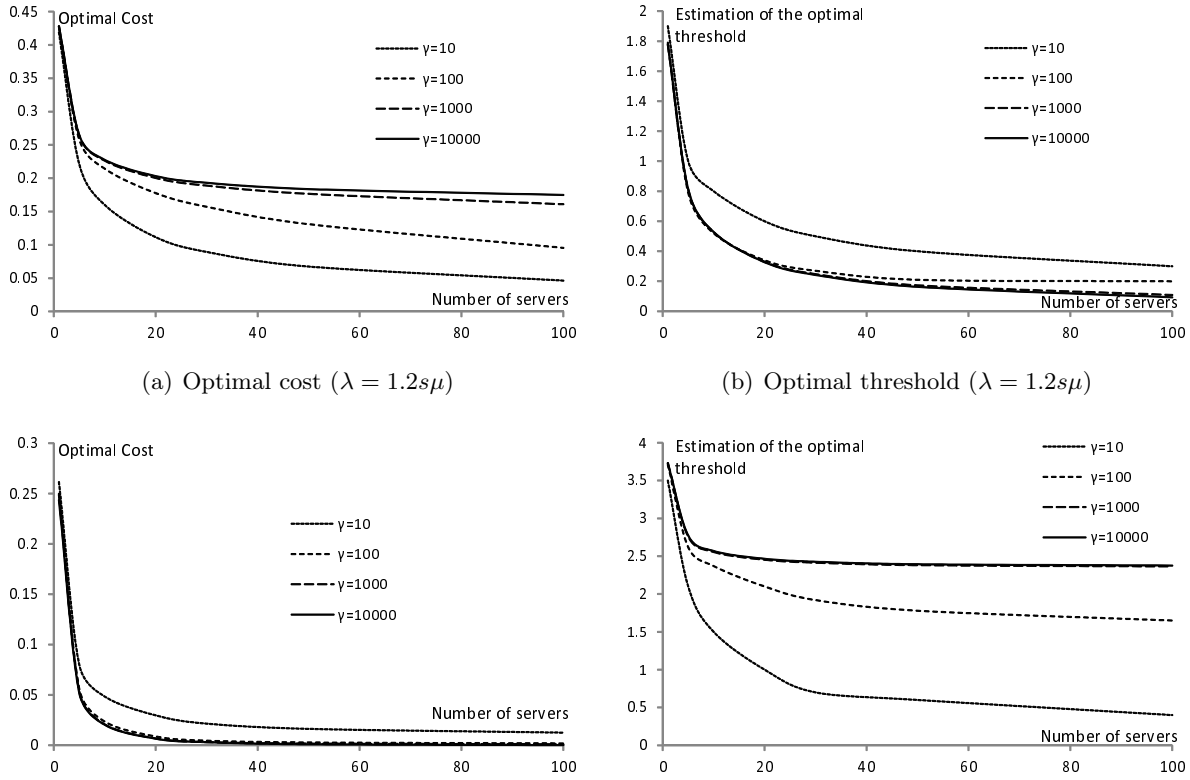
- **"Let $c(x) - q_x P(s\mu + \gamma)$ be increasing in x."** This assumption is needed to prove Theorem 2. Having $c(x) - q_x P(s\mu+\gamma)$ increasing in x is equivalent to having $c(x) + (1-q_x)P(s\mu+\gamma) = c(x) + \overline{p}_{x,x} P(s\mu+\gamma)$ increasing in x. In the latter expression, $\overline{p}_{x,x}$ is the probability that after a service completion, a rejection, or an abandonment, the system state remains unchanged. Therefore, $\overline{p}_{x,x}$ can be seen as the risk that a rejection from state $x > 0$ is paid with cost P without having any positive effect on the system congestion. With this interpretation, the term $\overline{p}_{x,x} P(s\mu+\gamma)$ should be added to $c(x)$ to account for the overall cost of state x. In other words, having $c(x) - q_x P(s\mu + \gamma)$ increasing in x means that the cost of state x is increasing in x.

  Note that we have $q_x = \left[1 + \frac{b\lambda}{\gamma} \prod_{i=1}^{x} r_i\right]^{-1}$. This function is increasing in x. Moreover, we have $q_x \leq 1$. Therefore, q_x has a finite limit as x tends to infinity. This means that for x sufficiently large, the variation of q_x in x can be neglected and the condition of having $c(x) - q_x P(s\mu+\gamma)$ increasing in x is equivalent to simply having $c(x)$ increasing in x, for x sufficiently high.

- **"Let $b\lambda < s\mu$."** This condition is needed to prove Lemma 2 and Theorem 3 when g^n is decreasing in n. This condition means that the system is stable without abandonment and subsequently ensures that the stationary probabilities can be computed in all cases. Note that for most cases of $c(k)$, if $b\lambda \geq s\mu$ then g^n cannot be decreasing in n. It means that the decreasing case for g^n is incompatible with $b\lambda \geq s\mu$.

Numerical illustration. In Figure 1, we consider a case without balking and reneging (i.e., $b = 1$, and $r_i = 1$, for $i \geq 1$). We want to minimize the expected cost per customer and per time unit given that a cost of 0.1 is counted per customer and per time unit spent in the queue and a cost of 1 is counted per rejected customer. In order to translate this objective, we define $c(x)$ as

$c(x) = 0.1\left(\frac{s\mu}{\lambda}\frac{x}{\gamma} + \mathbb{1}_{x=n}\frac{\gamma}{\lambda}\frac{n}{\gamma}\right)$, for $x > 0$, and $P = 1/\lambda$. The time-threshold, τ, is estimated via the relation $\tau = \frac{n}{\gamma}$. Therefore, the higher γ is the longer it takes to estimate the optimal threshold but the more accurate is the estimation of τ. The speed of convergence of the optimal cost and the optimal threshold as functions of γ depends on the system parameters. We observe that in cases with a large number of servers and $\lambda > s\mu$, high values of γ are needed to obtain an accurate estimation. As expected, in large systems, the expected cost per customer is reduced. Moreover,



(a) Optimal cost ($\lambda = 1.2s\mu$)

(b) Optimal threshold ($\lambda = 1.2s\mu$)

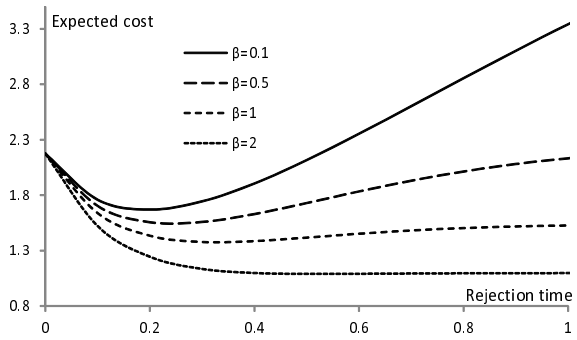
(c) Optimal cost ($\lambda = 0.8s\mu$)

(d) Optimal threshold ($\lambda = 0.8s\mu$)

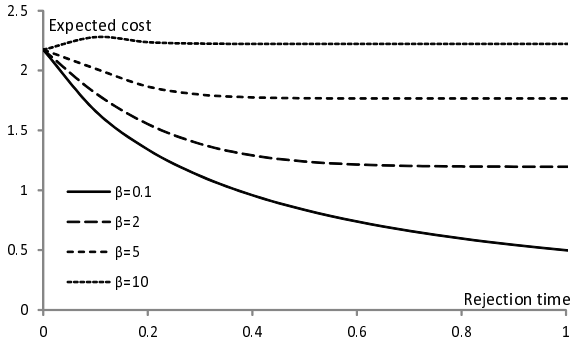
Figure 1: Numerical results ($\mu = 1$, $c(x) = 0.1\left(\frac{s\mu}{\lambda}\frac{x}{\gamma} + \mathbb{1}_{x=n}\frac{\gamma}{\lambda}\frac{n}{\gamma}\right)$, $P = 1/\lambda$, $b = 1$, and $r_i = 1$, for $i \geq 1$)

the optimal rejection threshold decreases with the system size. This translates that the optimal trade-off between rejection and wait is obtained with lower waits in larger systems.

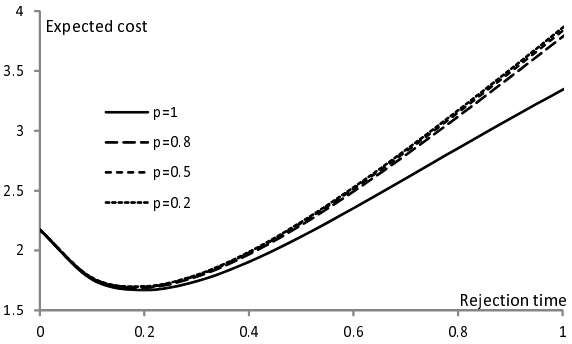
Effect of the impatience. In Figure 2, we represent the expected cost as a function of the rejection time in different situations of impatience. Figure 2(a) illustrates the effect of the expected patience time on the expected cost. We considered an exponential distribution with rate $\beta > 0$ for the patience time, and a cost function, $c(x)$, which captures the expected wait of served customers. We observe that while increasing customers' impatience (i.e., while increasing β), (i) the expected cost reduces, and (ii) the optimal rejection threshold increases. This can be understood by the



(a) Exponential distribution with $c(x) = \frac{s\mu x}{\gamma}$, and $r_x = \frac{\gamma}{\gamma+\beta}$, for $x \geq 1$



(b) Exponential distribution with $c(0) = (1-b)\lambda$, $c(x) = (1-r_x)\gamma$, and $r_x = \frac{\gamma}{\gamma+\beta}$, for $x \geq 1$



(c) Hyper-exponential distribution with $c(x) = \frac{s\mu x}{\gamma}$, and $r_x = \frac{p\left(\frac{\gamma}{\gamma+\beta}\right)^x + 1 - p}{p\left(\frac{\gamma}{\gamma+\beta}\right)^{x-1} + 1 - p}$, for $x \geq 1$

Figure 2: Optimal cost ($\mu = 1$, $\lambda = 12$, $s = 10$, $b = 0.8$, $P = 0.5$, $\gamma = 10000$)

objective function. In this illustration, the system manager only cares about the service quality of served customers. From this perspective, a strong impatience has an appreciable effect for reducing the system congestion. In addition, customers' rejection is used as a tool to further reduce the system congestion. The more customers abandon the queue, the less it is needed to reject customers and the higher is the rejection threshold.

Figure 2(b) considers the same distribution as in Figure 2(a). For this illustration, the definition of the cost function, $c(x)$, allows us to evaluate the rate of abandonment. As expected, contrary to Figure 2(a), the expected cost increases with customers' impatience. Another interesting result is that the optimal rejection threshold is either 0 (with $\beta = 10$) or ∞ (with $\beta = 0.1, 2, 5$). This means that either all delayed customers should be rejected or they should all be kept in the queue. This preference for extreme decisions is related to the constant value of the cost function, $c(x)$, for $x \geq 1$. In our case, this function is constant due to the exponential distribution of the patience. Due to the memoryless property of the patience, if it is optimal to keep a customer in the queue, then the same decision is optimal at a later time as neither the probability of abandonment nor the

cost of rejection will be modified. Note that a wait percentile objective is another example where the cost function is partially constant. Thus, with percentiles objective, it is either optimal to reject all customers at the instant just before the penalty for exceeding the time-threshold is paid or it is optimal to not reject any customer.

In Figure 2(b) the preference between rejecting customers or letting them wait is driven by the competition between the cost of abandonment and the cost of rejection. We chose a cost of rejection of 0.5 per customer, while the cost of abandonment is 1 per customer. Therefore, it is cheaper to reject a customer than to let this customer abandon the queue. With highly impatient customers, the probability that a delayed customer abandons the queue is high, therefore by rejecting all delayed customers, the abandonment cost is avoided. With more patient customers, the chance to be served is higher, and both the rejection and the abandonment cost could be avoided.

Finally, in Figure 2(c), we investigate the impact of the patience variability with a cost function representing the expected wait of served customers. For this purpose, we consider a particular hyperexponential distribution for which either a customer abandons the queue after an exponential time with parameter β or abandons directly without waiting. The former event occurs with probability p, while the latter occurs with probability $1-p$. The parameters of this distribution are chosen such that the expected time before abandonment, $\frac{p}{\beta}$, is equal to 10 time units. The coefficient of variation of this distribution, defined as the ratio of the standard deviation to the mean, is $\sqrt{\frac{2-p}{p}}$. Therefore, by decreasing p, we increase the patience variability. As expected, the system cost increases and the optimal rejection threshold decreases with the patience variability. However, by comparing Figure 2(c) with Figure 2(a), we conclude that the mean value of the patience has significantly more impact than its standard deviation on system cost and optimal threshold.

6 Conclusion

We considered a multi-server queue with general abandonment where rejection control could be exercised after allowing customers to wait. This problem may be seen as an alternative to the classic admission control problem where control is exercised at customers' arrival. Using an approximated model where the waiting time of the first customer in line is discretized via an Erlang distribution, we showed that the optimal control for rejection is a time-based threshold policy. Under this policy, all customers are admitted in the system. However, if a customer's wait reaches a timeout threshold, then the customer is rejected. To compute the optimal threshold, we showed, in the case where patience has the decreasing failure rate property, that a local minimum was necessarily a global

one for the approximated system. By extension, this also proved the result for the real system. Other interesting results were explained. For instance, if the cost function is constant on some intervals, then the search for the optimal threshold is limited to a finite number of time thresholds. Another interesting aspect is the role of customers' abandonment which may either be detrimental or beneficial to the system cost depending whether the system manager is interested by the rate of abandonment or by the service quality of served customers.

In future research, going one step beyond the current framework, we could investigate the potential of time-based control in other queueing systems involving customer retrial, other service time distributions, or agents' heterogeneity. Another challenging avenue would be to determine a framework for comparing time-based and quantity-based policies. More specifically, it would be interesting to associate the appropriate decision variable (the time or the quantity) to each metric used in practice for evaluating a system's performance.

References

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

Akşin, O., De Véricourt, F., and Karaesmen, F. (2008). Call center outsourcing contract analysis and choice. *Management Science*, 54(2):354–368.

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

Bailey, D. and Sweeney, T. (2003). Considerations in establishing emergency medical services response time goals. *Prehospital Emergency Care*, 7(3):397–399.

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.

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

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

Emadi, S. M. and Swaminathan, J. M. (2018). Customer learning in call centers from previous waiting experiences. *Operations Research*, 66(5):1433–1456.

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

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

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

Jouini, O., Koole, G., and Roubos, A. (2013). Performance indicators for call centers with impatient customers. *IIE Transactions*, 45(3):341–354.

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. (2013). *Call Center Optimization*. MG Books.

Koole, G., Nielsen, B., and Nielsen, T. (2012). First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60(5):1258–1266.

Koole, G., Nielsen, B., and Nielsen, T. (2015). Optimization of overflow policies in call centers. *Probability in the Engineering and Informational Sciences*, 29(3):461–471.

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.

Legros, B. (2016). Unintended consequences of optimizing a queue discipline for a service level defined by a percentile of the waiting time. *Operations Research Letters*, 44(6):839–845.

Legros, B. (2018). Waiting time based routing policies to parallel queues with percentiles objectives. *Operations Research Letters*, 46(3):356–361.

Legros, B. (2019). Transient analysis of a markovian queue with deterministic rejection. *Operations Research Letters*.

Legros, B., Jouini, O., and Koole, G. (2017). A uniformization approach for the dynamic control of queueing systems with abandonments. *Operations Research*, 66(1):200–209.

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

Lotfi, S. and Zenios, S. (2018). Robust VaR and CVaR optimization under joint ambiguity in distributions, means, and covariances. *European Journal of Operational Research*, 269(2):556–576.

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.

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

Ren, Z. and Zhou, Y. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383.

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

Thompson, D., Yarnold, P., Williams, D., and Adams, S. (1996). Effects of actual waiting time, perceived waiting time, information delivery, and expressive quality on patient satisfaction in the emergency department. *Annals of Emergency Medicine*, 28(6):657–665.

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

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

Appendix

Proof of Theorem 1

Proof. Consider a policy in Ω_S. Let π_x be the stationary probability to be in state x, $x \geq -s$. Consider the cut between $A = \{-s, -s+1, \cdots, x\}$ and $B = \{x+1, x+2, \cdots\}$, where $x \geq -s$. Observing that for $x, k > 0$, we have

$$\begin{aligned}\sum_{i=0}^{x} \overline{p}_{x+k,i} &= \sum_{i=1}^{x}(1-q_i)\prod_{j=i+1}^{x+k} q_j + \prod_{j=1}^{x+k} q_j = \sum_{i=0}^{x}\prod_{j=i+1}^{x+k} q_j - \sum_{i=1}^{x}\prod_{j=i}^{x+k} q_j = \sum_{i=0}^{x}\prod_{j=i+1}^{x+k} q_j - \sum_{i=0}^{x-1}\prod_{j=i+1}^{x+k} q_j \\ &= \prod_{j=x+1}^{x+k} q_j,\end{aligned}$$

we deduce that the cumulative transition rate from state $x+k$ to states $0, 1, \cdots, x$ is $(s\mu + \gamma(1 - r_{x+k}) + \gamma r_{x+k} p(x+k)) \cdot \prod_{j=x+1}^{x+k} q_j$. Therefore, by equating flows across the cut, one may write

$$\lambda \pi_x = (s+x+1)\mu\pi_{x+1}, \text{ for } -s \leq x < 0, \tag{22}$$

$$b\lambda(1-p(0))\pi_0 = \sum_{k=1}^{\infty}[s\mu + \gamma(1-r_k) + \gamma r_k p(k)]\pi_k \left(\prod_{j=1}^{k} q_j\right), \text{ for } x = 0, \tag{23}$$

$$\gamma r_x (1-p(x))\pi_x = \sum_{k=1}^{\infty}[s\mu + \gamma(1-r_{x+k}) + \gamma r_{x+k} p(x+k)]\pi_{x+k} \left(\prod_{j=x+1}^{x+k} q_j\right) \text{ for } x > 0. \tag{24}$$

Multiplying Equation (6) by π_x, we get $(s+x)\mu y(x)\pi_x - \lambda y(x+1)\pi_x = -g\pi_x$, for $-s < x < 0$. From Equation (22), we deduce that

$$(s+x)\mu y(x)\pi_x - (s+x+1)\mu y(x+1)\pi_{x+1} + g\pi_x = 0, \tag{25}$$

for $-s < x < 0$. We have $\min(y(1), P) \leq p(0)P + (1-p(0))y(1)$. Multiplying Equation (7) by π_0, we next deduce that $\pi_0(g - c(0) + s\mu y(0)) \leq b\lambda\pi_0[p(0)P + (1-p(0))y(1)]$. Finally, from Equation (23), we get

$$s\mu y(0)\pi_0 - y(1)\sum_{k=1}^{\infty}[s\mu + \gamma(1-r_k) + \gamma r_k p(k)]\pi_k \left(\prod_{j=1}^{k} q_j\right) + g\pi_0 \leq \pi_0(b\lambda p(0)P + c(0)). \tag{26}$$

Similarly, for $x > 0$, we have $\min(V(x+1) - V(x), F(V(x)) + P - V(x)) \leq (1 - p(x))(V(x+1) - V(x)) + p(x)(F(V(x)) + P - V(x))$. So, from Equation (8), we get

$$(s\mu + \gamma(1 - r_x))(V(x) - F(V(x))) - \gamma r_x(1 - p(x))(V(x+1) - V(x)) \\ - \gamma r_x p(x)(F(V(x)) + P - V(x)) \leq c(x) - g,$$

for $x > 0$. This equation can be rewritten as

$$(s\mu + \gamma(1 - r_x) + \gamma r_x p(x))(V(x) - F(V(x))) - \gamma r_x(1 - p(x))(V(x+1) - V(x)) - \gamma r_x p(x) P \leq c(x) - g,$$

for $x > 0$. Multiplying the above equation by π_x and combining it with Equation (24) yields

$$\begin{aligned} &(s\mu + \gamma(1 - r_x) + \gamma r_x p(x))(V(x) - F(V(x)))\pi_x \\ &\quad - (V(x+1) - V(x)) \sum_{k=1}^{\infty} [s\mu + \gamma(1 - r_{x+k}) + \gamma r_{x+k} p(x+k)] \pi_{x+k} \left(\prod_{j=x+1}^{x+k} q_j \right) + g\pi_x \\ &\quad \leq \pi_x (c(x) + \gamma r_x P p(x)), \end{aligned} \tag{27}$$

for $x > 0$.

Summing up Equations (25) for $-s < x < 0$, Equation (26) and Equations (27) for $x > 0$, we get

$$\begin{aligned} &\sum_{x=-s+1}^{-1} ((s+x)\mu y(x)\pi_x - (s+x+1)\mu y(x+1)\pi_{x+1}) + s\mu y(0)\pi_0 \\ &\quad - y(1) \sum_{k=1}^{\infty} [s\mu + \gamma(1 - r_k) + \gamma r_k p(k)] \pi_k \prod_{j=1}^{k} q_j \\ &\quad + \sum_{x=1}^{\infty} (s\mu + \gamma(1 - r_x) + \gamma r_x p(x))(V(x) - F(V(x)))\pi_x \\ &\qquad - (V(x+1) - V(x)) \sum_{k=1}^{\infty} [s\mu + \gamma(1 - r_{x+k}) + \gamma r_{x+k} p(x+k)] \pi_{x+k} \prod_{j=x+1}^{x+k} q_j \\ &\quad + g \sum_{x=-s+1}^{\infty} \pi_x \leq b\lambda \pi_0 p(0) P + c(0)\pi_0 + \sum_{x=1}^{\infty} \pi_x (c(x) + \gamma r_x P p(x)). \end{aligned}$$

The first line of the left hand side of the inequality can be simplified into $\sum_{x=-s+1}^{-1} ((s+x)\mu y(x)\pi_x - (s+x+1)\mu y(x+1)\pi_{x+1}) + s\mu y(0)\pi_0 = \mu y(-s+1)\pi_{-s+1}$. Consider now the second, third and

fourth line of the left hand side of the inequality. These three lines can be rewritten as

$$\sum_{x=1}^{\infty}\pi_x(s\mu+\gamma(1-r_x)+\gamma r_x p(x))\left(V(x)-F(V(x))-\sum_{k=1}^{x}(V(k)-V(k-1))\prod_{j=k}^{x}q_j\right)$$
$$=\sum_{x=1}^{\infty}\pi_x(s\mu+\gamma(1-r_x)+\gamma r_x p(x))\left(V(x)-\sum_{k=1}^{x}(1-q_k)\prod_{j=k+1}^{x}q_jV(k)-\prod_{j=1}^{x}q_jV(0)\right.$$
$$\left.-\sum_{k=1}^{x}(V(k)-V(k-1))\prod_{j=k}^{x}q_j\right)$$
$$\sum_{x=1}^{\infty}\pi_x(s\mu+\gamma(1-r_x)+\gamma r_x p(x))\left(V(x)-\sum_{k=0}^{x}\prod_{j=k+1}^{x}q_jV(k)+\sum_{k=1}^{x}\prod_{j=k}^{x}q_jV(k)-\sum_{k=1}^{x}\prod_{j=k}^{x}q_jV(k)\right.$$
$$\left.+\sum_{k=0}^{x-1}\prod_{j=k+1}^{x}q_jV(k)\right)=0.$$

Moreover, $g\sum_{x=-s+1}^{\infty}\pi_x=g(1-\pi_{-s})$. We identify the right hand side of the inequality with the average cost for the policy in Ω_S; $b\lambda\pi_0p(0)P+\pi_0c(0)+\sum_{x=1}^{\infty}\pi_x(c(x)+\gamma r_xPp(x))=g^S$. The inequality then becomes

$$\mu y(-s+1)\pi_{-s+1}+g(1-\pi_{-s})\leq g^S.$$

Since $\mu\pi_{-s+1}=\lambda\pi_{-s}$, this inequality can be rewritten as

$$\pi_{-s}\left(\lambda y(-s+1)-g\right)+g\leq g^S.$$

Finally, Equation (5) indicates that $\lambda y(-s+1)-g=0$. Therefore, we obtain $g\leq g^S$. This finishes the proof of the Theorem. □

Proofs of Lemma 1 and Theorem 2

Proof of Lemma 1. We prove Lemma 1 by induction. From Equation (10), we have $g^{n_1}=\lambda y^{n_1}(-s+1)$ and $g^{n_2}=\lambda y^{n_2}(-s+1)$, so clearly $y^{n_1}(-s+1)\geq y^{n_2}(-s+1)$. For $-s<x<0$, using Equation (11), one may write

$$\lambda(y^{n_1}(x+1)-y^{n_2}(x+1))=g^{n_1}-g^{n_2}+(s+x)\mu(y^{n_1}(x)-y^{n_2}(x)).$$

This relation proves the induction step for Equation (16) for $-s < x \leq 0$. For $x = 0$, we have

$$b\lambda(y^{n_1}(1) - y^{n_2}(1)) = g^{n_1} - g^{n_2} + s\mu(y^{n_1}(0) - y^{n_2}(0)) \geq 0.$$

For $x > 0$, we have

$$\gamma r_x(y^{n_1}(x+1) - y^{n_2}(x+1)) = g^{n_1} - g^{n_2} + q_x(s\mu + \gamma)(y^{n_1}(x) - y^{n_2}(x)),$$

for $0 < x \leq \min(n_1, n_2)$. This proves the induction step for Equation (16), for $0 < x \leq \min(n_1, n_2)+1$. □

Proof of Theorem 2. Consider $m \in \mathbb{N}$ such that $g^m < g^k$ for $0 \leq k \leq m-1$ and $g^{m+1} > g^m$ for the n-terminating problem. Let us consider a modified version of the original problem such that the transition rates are identical to those of the original problem for $-s \leq x \leq m$ and become constant and equal to their value at $x = m+1$, for $x \geq m+1$. In the modified problem, the cost function is changed into $c'(x) = c(x)$ for $0 \leq x \leq m$ and $c'(x) = c(m+1) - g^{m+1} + g^m$, for $x > m$.

In this case, the threshold level m is *optimal* for this modified problem associated with the optimal cost g^m. To show the latter result, we give the explicit solution of Equations (5)-(9) for the modified problem. It is given by the vector $(g, y(-s+1), y(-s+2), \cdots, y(k), \cdots)$, where $g = g^m$ and $y(k) = y^m(k)$ for $-s+1 \leq k \leq m$ and $y(k) = y^{m+1}(m+1)$ if $k > m$. Since $g^{m+1} > g^m$, Lemma 1 proves that $y^{m+1}(m+1) > y^m(m+1) = P$. Moreover, since $g^k > g^m$ for $0 \leq k \leq m-1$, Lemma 1 shows that $y^m(k+1) < y^k(k+1) = P$. Therefore, we can easily check that $g, y(-s+1), y(-s+2), \cdots, y(m)$ are solutions of the optimality equations (5)-(9). For $x \geq m+1$, we have for the modified problem $g^m - c(m+1) + g^{m+1} - g^m = r_{m+1}\gamma P - (s\mu+\gamma)q_{m+1}y^{m+1}(m+1)$. This equation coincides with the equation at level $m+1$ for the threshold policy with the threshold level $m+1$ in the n-terminating problem.

For the modified problem, let g'^k be the average expected cost for the modified problem associated with the threshold level k and let $y'^k(x)$ be the associated solution of Equation (10)-(13). If m is not optimal for the original problem, then there exists $n > m+1$, such that $g^n < g^m$. First, we show that we cannot have $g'^n \geq g^n$. Let us assume that $g'^n \geq g^n$ and show that this leads to a contradiction. We have $g^n = \lambda y^n(-s+1)$, for the original problem, and, $g'^n = \lambda y'^n(-s+1)$, for the modified problem. So clearly, $y'^n(-s+1) \geq y^n(-s+1)$. For $-s < x < 0$,

$$\lambda(y'^n(x+1) - y^n(x+1)) = g'^n - g^n + (s+x)\mu(y'^n(x) - y^n(x)).$$

Therefore, by induction we have $y'^n(x) \geq y^n(x)$, for $-s < x \leq 0$. For $x = 0$, we have

$$b\lambda(y'^n(1) - y^n(1)) = g'^n - g^n + s\mu(y'^n(0) - y^n(0)).$$

Therefore, $y'^n(1) \geq y^n(1)$. For $0 < x \leq m$,

$$r_x\gamma(y'^n(x+1) - y^n(x+1)) = g'^n - g^n + (s\mu + \gamma)q_x(y'^n(x) - y^n(x)).$$

Therefore, by induction we have $y'^n(x) \geq y^n(x)$, for $1 < x \leq m+1$. For $x = m+1$, we have

$$r_{m+1}\gamma(y'^n(m+2) - y^n(m+2)) = g'^n - g^n + g^{m+1} - g^m + (s\mu + \gamma)q_{m+1}(y'^n(m+1) - y^n(m+1)).$$

So, $y'^n(m+2) \geq y^n(m+2)$. For $m+1 < x < n$,

$$\begin{aligned}
&\gamma(r_{m+1}y'^n(x+1) - r_x y^n(x+1)) \\
&\quad = g'^n - g^n + g^{m+1} - g^m + c(x) - c(m+1) + (s\mu + \gamma)(q_{m+1}y'^n(x) - q_x y^n(x)) \\
&\quad = g'^n - g^n + g^{m+1} - g^m + (s\mu + \gamma)q_{m+1}(y'^n(x) - y^n(x)) + (s\mu + \gamma)(q_{m+1} - q_x)y^n(x) \\
&\qquad\qquad + c(x) - c(m+1).
\end{aligned}$$

The definition of q_x indicates that q_x in increasing in x. Therefore, $q_{m+1} - q_x \leq 0$, for $x \geq m+1$. Moreover, using Lemma 1, we have $y^n(x) \leq P$, for $x \leq n$. Therefore,

$$(s\mu + \gamma)(q_{m+1} - q_x)y^n(x) + c(x) - c(m+1) \geq (s\mu + \gamma)(q_{m+1} - q_x)P + c(x) - c(m+1) \geq 0,$$

since $c(x) - q_x(s\mu + \gamma)P$ is increasing in x. This shows that if $y'^n(x) \geq y^n(x)$, then $r_{m+1}y'^n(x+1) - r_x y^n(x+1) \geq 0$. Using now the increasing property of r_x, we may write

$$\begin{aligned}
r_x(y'^n(x+1) - y^n(x+1)) &= r_{m+1}\frac{r_x}{r_{m+1}}y'^n(x+1) - r_x y^n(x+1) \\
&\geq r_{m+1}y'^n(x+1) - r_x y^n(x+1) \geq 0.
\end{aligned}$$

This proves by induction that $y'^n(x) \geq y^n(x)$, for $m+1 < x \leq n$. In summary, if $g'^n \geq g^n$ then

$y'^n(x) \geq y^n(x)$, for $-s+1 \leq x \leq n$. Consider now the case $x = n$. One may write

$$g^n - c(n) = \gamma r_n P - q_n(s\mu + \gamma)y^n(n), \text{ for the original problem, and,}$$
$$g'^n - c(m+1) + g^{m+1} - g^m = \gamma r_{m+1} P - q_{m+1}(s\mu + \gamma)y'^n(n), \text{ for the modified problem.}$$

Let us subtract these two equations. We get

$$\begin{aligned} g^n - g'^n = g^{m+1} - g^m + \gamma P(r_n - r_{m+1}) + q_{m+1}(s\mu + \gamma)(y'^n(n) - y^n(n)) \\ + c(n) - c(m+1) + (s\mu + \gamma)y^n(n)(q_{m+1} - q_n). \end{aligned}$$

We assumed that $g^{m+1} \geq g^m$ and that r_x is increasing in x and we showed that $y'^n(n) \geq y^n(n)$. Moreover, since $y^n(n) \leq P$ (using Lemma 1), $q_{m+1} \geq q_n$, and $c(x) - (s\mu + \gamma)q_x P$ is increasing in x, we have $c(n) - c(m+1) + (s\mu + \gamma)y^n(n)(q_{m+1} - q_n) \geq 0$. This proves that $g^n > g'^n$ and shows the contradiction. As a conclusion, we have $g^n > g'^n$. This leads to $g'^n < g^n < g^m = g'^m$ which is in contradiction with $g'^m < g'^n$ (optimality of the threshold level m for the modified problem). This proves that there cannot exist $n > m+1$, such that $g^n < g^m$ and shows that $g^m < g^k$ for $k \geq 0$, and $k \neq m$. □

Proof of Lemma 2

Consider an $n-$terminating solution $\{g^n, y^n(-s+1), y^n(-s+2), \cdots, y^n(n)\}$. Since g^n is decreasing in n, we have $g^n < g^k$ for $k < n$. So, Lemma 1 indicates that $y^n(x) < y^k(x)$ for $-s+1 \leq x \leq k+1$. So $y^n(k) < y^k(k+1) = P$ for $k \leq n$. Therefore, $\{g^n, y^n(-s+1), y^n(-s+2), \cdots, y^n(n)\}$ satisfies the optimality equations translated into Equations (17)-(20). However, Equation (21) is not necessarily satisfied by $\{g^n, y^n(-s+1), y^n(-s+2), \cdots, y^n(n)\}$.

The objective here is to build a new sequence $\{\overline{g}^n, \overline{y}^n(-s+1), \overline{y}^n(-s+2), \cdots, \overline{y}^n(n)\}$ which also satisfies Equation (21). Note that for any fixed g in Equations (5)-(9) there exists a unique sequence $y(x)$, for $x \geq -s+1$. Therefore, $y(n)$ can be seen as a function of g. We thus write $y(n) = y(n, g)$. We define

$$G(n, g) = -r_n \gamma y(n, g) + q_n(s\mu + \gamma)y(n, g) + g - c(n), \text{ for } n \geq 1, \text{ and}$$
$$G(0, g) = -b\lambda y(0, g) + s\mu y(0, g) + g - c(0).$$

We want to show that for each $n \geq 0$, there exists $\overline{g}^n$ such that $G(n, \overline{g}^n) = 0$. If for some $n \geq 1$, the original $n-$terminating problem is such that $G(n, g^n) < 0$, then $g^n - c(n) < r_n \gamma y^n(n) - q_n(s\mu + \gamma)y^n(n)$. However, from Equation (13), we have $g^n - c(n) = r_n \gamma P - q_n(s\mu + \gamma)y^n(n)$. So, $G(n, g^n) < 0$ indicates that $y^n(n) > P$. This is in contradiction with Lemma 1 and the assumption of decreasing values for g^n. The approach is identical to show that $G(0, g^0)$ cannot be strictly negative. Therefore, $G(n, g^n) \geq 0$, for $n \geq 0$.

Consider the case $n = 0$, with $g = 0$. Equations (5)-(6) lead to $y(0, 0) = 0$. Therefore $G(0, 0) = (s\mu - b\lambda)y(0, 0) + g - c(0) = -c(0) < 0$. Moreover, we showed that $G(0, g^0) \geq 0$. Since $G(0, g)$ is continuous in g on $(0, \infty)$, there exists $0 \leq \overline{g}^0 \leq g^0$ such that $G(0, \overline{g}^0) = 0$.
Consider now the case $n = 1$, with $g = 0$. Equations (5)-(6) lead to $y(1, 0) = -\frac{c(0)}{b\lambda}$. Therefore, we have $G(1, 0) = -c(1) - \frac{c(0)}{b\lambda}(q_1(s\mu + \gamma) - r_1\gamma)$. Recall that $q_1 = \frac{1}{1+\frac{b\lambda}{\gamma}r_1}$. Thus, $q_1(s\mu + \gamma) - r_1\gamma = \frac{s\mu+\gamma(1-r_1)-b\lambda r_1^2}{1+\frac{b\lambda}{\gamma}r_1}$. The numerator of this expression is decreasing in r_1. Moreover, the highest possible value for r_1 is 1. Therefore, $s\mu + \gamma(1 - r_1) - b\lambda r_1^2 \geq s\mu - b\lambda > 0$. This shows that $G(1, 0) < 0$. Again, since $G(1, g)$ is continuous on $(0, \infty)$, there exists $0 \leq \overline{g}^1 \leq g^1$ such that $G(1, \overline{g}^1) = 0$.
Assume now that there exists $\overline{g}^n$ such that $G(n, \overline{g}^n) = 0$. We have

$$G(n + 1, \overline{g}^n) = -r_{n+1}\gamma y(n + 1, \overline{g}^n) + q_{n+1}(s\mu + \gamma)y(n + 1, \overline{g}^n) + \overline{g}^n - c(n + 1)$$

Moreover, we have $y(n + 1, \overline{g}^n) = y(n, \overline{g}^n)$ and $G(n, \overline{g}^n) = 0$. So,

$$G(n + 1, \overline{g}^n) = c(n) - c(n + 1) + y(n, \overline{g}^n)\left[(s\mu + \gamma)(q_{n+1} - q_n) + \gamma(r_n - r_{n+1})\right].$$

We assumed that $c(x)$ and r_x are increasing in x. Moreover q_x is decreasing in x. Therefore, $G(n + 1, \overline{g}^n) < 0$. Since $G(n + 1, g)$ is continuous on $(0, \infty)$, there exists $\overline{g}^n \leq \overline{g}^{n+1} \leq g^{n+1}$ such that $G(n + 1, \overline{g}^{n+1}) = 0$. This proves the induction step. □

Proof of Theorem 3

Consider the modified holding cost $\overline{c}(x) = c(x)$, for $0 \leq x \leq n$ and $\overline{c}(x) = c(n)$, for $x \geq n$. Lemma 2 indicates that there exists a vector $(\overline{g}^n, \overline{y}^n(-s + 1), \overline{y}^n(-s + 2), \cdots, \overline{y}^n(n - 1), \overline{y}^n(n), \overline{y}^n(n), \cdots\}$ which solves Equations (5)-(9) for the modified cost function and has $\overline{y}^n(k) < P$, for $-s+1 \leq k \leq n$. Therefore, it is never optimal to reject a customer. Let π_x^n and π_x^∞ be the steady state probabilities to be at state x in the $n-$terminating problem and in the policy that exercises no control. These

probabilities can be obtained using the expression of the long-run cost in Equation (15). Given that $g^n = P\gamma r_n \cdot \pi_n^n + \sum_{x=0}^{n} c(x)\pi_x^n$, we identify the stationary probabilities to be in state $x \geq 0$. For the policy with finite n, we obtain after some algebra

$$\pi_0^n = \left[\sum_{k=0}^{s} \frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k + b\lambda \sum_{k=1}^{n} \frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k \prod_{j=1}^{k} \frac{r_j}{q_j}\right]^{-1}, \text{ and,}$$

$$\pi_x^n = \frac{\frac{b\lambda}{\gamma r_x}\left(\frac{\gamma}{s\mu+\gamma}\right)^x \prod_{j=1}^{x} \frac{r_j}{q_j}}{\sum_{k=0}^{s} \frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k + b\lambda \sum_{k=1}^{n} \frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k \prod_{j=1}^{k} \frac{r_j}{q_j}}, \text{ for } 0 < x \leq n.$$

By letting n tend to infinity, we get

$$\pi_0^\infty = \left[\sum_{k=0}^{s} \frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k + b\lambda \sum_{k=1}^{\infty} \frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k \prod_{j=1}^{k} \frac{r_j}{q_j}\right]^{-1}, \text{ and,}$$

$$\pi_x^\infty = \frac{\frac{b\lambda}{\gamma r_x}\left(\frac{\gamma}{s\mu+\gamma}\right)^x \prod_{j=1}^{x} \frac{r_j}{q_j}}{\sum_{k=0}^{s} \frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k + b\lambda \sum_{k=1}^{\infty} \frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k \prod_{j=1}^{k} \frac{r_j}{q_j}}, \text{ for } x > 0.$$

We have $\overline{g}^n = \sum_{x=0}^{n} c(x)\pi_x^\infty + c(n) \sum_{x=n+1}^{\infty} \pi_x^\infty$. This leads to

$$g^n \leq \overline{g}^n + \gamma r_n P \cdot \pi_n^n + \sum_{x=0}^{n} c(x)(\pi_x^n - \pi_x^\infty)$$

Moreover, since $\overline{c}(x) \leq c(x)$, for $x \geq 0$, we have $g^* \geq \overline{g}^n$. Therefore,

$$g^n \leq g^* + \gamma r_n P \cdot \pi_n^n + \sum_{x=0}^{n} c(x)(\pi_x^n - \pi_x^\infty).$$

Note that

$$\gamma r_n \pi_n^n = \frac{b\lambda\left(\frac{\gamma}{s\mu+\gamma}\right)^n \prod_{j=1}^{n} \frac{r_j}{q_j}}{\sum_{k=0}^{s} \frac{s!}{(s-k)!}\left(\frac{\mu}{\lambda}\right)^k + b\lambda \sum_{k=1}^{n} \frac{1}{\gamma r_k}\left(\frac{\gamma}{s\mu+\gamma}\right)^k \prod_{j=1}^{k} \frac{r_j}{q_j}} \leq b\lambda\left(\frac{\gamma}{s\mu+\gamma}\right)^n \prod_{j=1}^{n} \frac{r_j}{q_j}.$$

For $1 \leq j \leq n$, we have $\frac{r_j}{q_j} = r_j\left(1 + \frac{b\lambda}{\gamma}\prod_{i=1}^{j} r_i\right) \leq 1 + \frac{b\lambda}{\gamma}$, since $r_i \leq 1$, for $1 \leq i \leq j$. Therefore, $\prod_{j=1}^{n} \frac{r_j}{q_j} \leq \left(1 + \frac{b\lambda}{\gamma}\right)^n$. Thus, we may write $\gamma r_n \pi_n^n \leq b\lambda\left(\frac{b\lambda+\gamma}{s\mu+\gamma}\right)^n$. Since we assumed $b\lambda < s\mu$, we have $\lim_{n\longrightarrow\infty}\left(\frac{b\lambda+\gamma}{s\mu+\gamma}\right)^n = 0$ and we deduce that $\lim_{n\longrightarrow\infty} \gamma r_n \pi_n^n = 0$. So, if we show that

$\lim_{n\longrightarrow\infty}\left(\sum_{x=1}^{n}c(x)(\pi_x^n-\pi_x^\infty)\right)=0$, then we have $\lim_{n\longrightarrow\infty}g^n\leq g^*$. Since $g^n\geq g^*$ for each $n\geq 0$, this proves that $\lim_{n\longrightarrow\infty}g^n=g^*$ and finishes the proof.

There remains to prove that $\lim_{n\longrightarrow\infty}\left(\sum_{x=0}^{n}c(x)(\pi_x^n-\pi_x^\infty)\right)=0$. By defining $r_0=\frac{b\lambda}{\gamma}$, we may write

$$\begin{aligned}\sum_{x=0}^{n}c(k)(\pi_x^n-\pi_x^\infty)&=(\pi_0^n-\pi_0^\infty)\sum_{x=0}^{n}c(x)\frac{b\lambda}{\gamma r_x}\left(\frac{\gamma}{s\mu+\gamma}\right)^x\prod_{j=1}^{x}\frac{r_j}{q_j}\\&\leq(\pi_0^n-\pi_0^\infty)\sum_{x=0}^{n}c(x)\frac{b\lambda}{\gamma r_x}\left(\frac{b\lambda+\gamma}{s\mu+\gamma}\right)^x\end{aligned}$$

If $\lim_{x\longrightarrow\infty}\frac{c(x+1)}{c(x)}\frac{r_x}{r_{x+1}}\frac{b\lambda+\gamma}{s\mu+\gamma}<1$, then $\sum_{x=0}^{n}c(x)\frac{b\lambda}{\gamma r_x}\left(\frac{b\lambda+\gamma}{s\mu+\gamma}\right)^x$ has a finite limit as n tend to infinity. With expressions of $c(x)$ translating the expected waiting time, a percentile of the waiting time or a combination of both, this condition is clearly satisfied. Since $\lim_{n\longrightarrow\infty}\pi_0^n=\pi_0^\infty$, then $\lim_{n\longrightarrow\infty}\left(\sum_{x=1}^{n}c(x)(\pi_x^n-\pi_x^\infty)\right)=0$. □

Proof of Proposition 1

Consider the modified problem with holding costs $c'(x)=c(x)-r_n\gamma(P-y_n(n))$, for $x<n$ and $c'(x)=c(n)$ otherwise. Therefore, $(g^n-r_n\gamma(P-y_n(n)),y^n(-s+1),y^n(-s+2),\cdots,y^n(n-1),y^n(n),y^n(n),\cdots)$ satisfies the optimality equations for the modified problem with lower holding costs implying that the optimal cost of the original problem cannot be any less (i.e., $g^n-r_n\gamma(P-y_n(n))\leq g^*$). □