The transient blended queue

Benjamin Legros

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

benjamin.legros@centraliens.net

Abstract

This study aims to determine the transient behavior of the blended queue. Priority customers arrive over time and benefit from a threshold reservation policy, while non-priority ones can be contacted at any time. We show how to compute the Laplace transforms of the transient probabilities. Using the uniformization technique, we prove some monotonicty properties of the expected number of customers in the queue, explaining why the optimal transient reservation threshold should be lower than the stationary one.

Keywords: Blended queue; reservation; threshold; Laplace transform; uniformization

1 Introduction

For reasons of stability, service capacity should be greater than demand in service systems modeled by queues. Furthermore, difficulty making accurate forecasts may incentivize decision-makers to over-staff their services to achieve a high quality of service. One consequence is an agent's utilization being too low, leading to long period of idle time. One way to reduce idle capacity is to allow agents to initiate services by contacting customers. This combination of inbound and outbound customer service provided by a single group of agents is referred to as the *blended queue* and is often used to model call centers with inbound and outbound callers.

The implementation of blended operations comes with operational challenges. Since the number of customers to be contacted can be considerable, agents may be continually occupied, either serving inbound customers or contacting outbound ones. In such conditions, when agents can initiate outbound services, a high degradation of service level for inbound customers will be observed. Initiating outbound services should therefore be restricted. One routing solution proposed in the queueing literature for this type of problem is the employment of a *threshold reservation policy* [1, 4], whereby a certain number of agents, denoting the reservation level, are not allowed to initiate outbound services.

Various studies have considered the blended queue under a threshold reservation policy to provide practical insights for the selection of an optimal threshold level. Unlike the previous contributions, we

analyze the blended queue in transient regime. The transient analysis is motivated by the frequency of changes to the reservation threshold in call centers. [11] explained that the threshold level is reevaluated every 15 minutes by an automatic call distributor at a call center, while the stationary regime is achieved within half-hourly or hourly intervals with constant parameters [5, 3]. This means that stationary results fail to capture the real evolution of the system state and may not allow optimal decisions to be taken for the selection of the reservation threshold. In particular, the initial state of the system should be considered to determine the optimal threshold level for a finite time interval. As for stationary studies, we focus on the optimization problem, which consists of maximizing the rate of served outbound customers, while maintaining congestion in the queue below a threshold. We measure the congestion by the expected number of inbound customers waiting in the queue.

First, we determine the Laplace transforms of the transient probabilities in the case with equal service rates between inbound and outbound customers and when there is no reservation. We express the transient probabilities in terms of complex integrals and show how these integrals can be computed. The case with no reservation can serve as a basis for computing the Laplace transforms of the transient probabilities in the general case, provided that a finite number of equations remains to be solved. The complexity of the Laplace transforms precludes their use for purposes other than numerical computations.

To overcome this limitation, we consider a truncated approximation of the system and employ a uniformization technique to compute the transient probabilities in the equivalent discrete time Markov chain. Using this technique, we prove by iteration on the elapse of time that the expected number of customers in the queue is increasing with the arrival rate and the initial number of customers present and is decreasing and convex in the reservation threshold. These results support our numerical investigations revealing that in many cases, even when the initial number of customers present is above the manager's objective, it is advisable to select a reservation threshold that is lower than or equal to the stationary reservation threshold. This means that stationary results lead to decisions being taken that are too safe for the service quality of inbound customers while under-using the service capacity.

2 Literature review

There is a large body of literature on the blended queue, primarily for applications in call centers. The first formal proof that a threshold-type reservation policy is optimal for maximizing the rate of served outbound customers with a service level constraint on the inbound customers' waiting time was by [4] and [1]. Later, [2], showed the value of this policy in a more general setting with a piecewise-constant

doubly-stochastic arrival rate. With unequal service rates, the optimal policy is also of a threshold type, but the optimal threshold level should be state-dependent. This result was proven in [9] in a system wherein inbound customers would balk if they had to wait. However, the complexity of the state-dependent threshold policy and close proximity in performance between a state-dependent and fixed threshold policy have led most studies to restrict their investigation to the fixed threshold policy. This was the case for [14], who investigated a large call blending model and proposed a logarithmic safety staffing rule combined with a threshold reservation policy that would simultaneously manage having agents' utilization close to one, with idle agents almost always present. Extensions of the blended queue model with reservation have been investigated with retrials [15], reservation for arriving customers, where delayed ones are viewed as outbound ones [8], or in combination with outsourcing decisions in a sales environment [12]. In the above references, the analyses are made in the stationary regime. We should also mention the paper of [11], who considered transient experiments through simulations. They showed empirically that changes in the reservation threshold should be made more slowly than the stationary results suggest from the evolution of the arrival rate. This paper aims to further investigate this question and provide structural results on the effect of the reservation threshold on the expected queue length in the transient regime.

3 Model description

We consider a blended queue in the transient regime with a single pool of s homogeneous agents and two types of customer, referred to as class-1 and class-2 customers. Class-1 customers arrive in the system according to a Poisson process with rate λ. If class-1 customers are not routed to the service immediately upon arrival, then they will wait in an infinite capacity queue for their turn to be served, with customers being served in order of arrival. Unlike class-1 customers, we assume that there is an infinite supply of class-2 customers, so an available agent can always initiate service with them. The service times of all class-i customers are assumed to be exponential random variables with rate μ_i for $i = 1, 2$.

The routing of customers to service is controlled by a *priority-reservation threshold policy* with reservation parameter c for $c = 0, 1, \cdots s$. The priority is provided to class-1 customers, which means that agents serve class-1 customers in priority until their queue is empty. Furthermore, class-1 customers benefit from reservation that determines an agent's decision at service completion time when the queue is empty. Specifically, if the number of idle agents (excluding the idle agent considered) is at least c, then this agent initiates the service of a class-2 customer. Otherwise, this agent remains idle. In other words, there are c agents that are reserved for class-1 customers, thus there are at least $s - c$ agents working at any time. We

consider the optimization problem that consists of maximizing the rate of served class-2 customers while maintaining the expected number of class-1 customers in the queue below a threshold. In the stationary case, due to Little's Law [13], this is equivalent to a constraint on the expected waiting time.

In the transient case, we distinguish between a performance measure observed at time t and its average value on an interval of length t^* given an initial condition. In practice, both can be interesting to study. The former indicates the state of the system to determine the distance with the stationary system, while the latter informs the experimented performance of past customers. We consider a finite interval of time $I = [0, t^*]$ and denote by $E(Q)_t$ and $E(N_2)_t$ the expected number of customers in the queue and number of agents serving class-2 customers at time t for $0 \leq t \leq t^*$. The expected rate of served class-2 customers is estimated at time t as $E(T)_t = \mu_2 E(N_2)_t$. It should be noted that although $E(Q)_t$ and $E(N_2)_t$ can be determined by the transient probabilities, $E(T)_t$ is only an estimation of the potential rate of served class-2 from a situation whereby $E(N_2)_t$ are serving class-2 customers. The average values of $E(Q)_t$ and $E(T)_t$ are given by $\overline{E(Q)_I} = \dfrac{1}{t^*} \displaystyle\int_{t=0}^{t=t^*} E(Q)_t \mathrm{d}t$ and $\overline{E(T)_I} = \dfrac{\mu_2}{t^*} \displaystyle\int_{t=0}^{t=t^*} E(N_2)_t \mathrm{d}t$, respectively.

A state of the system in the blended queue is defined by the couple (x, y), where x is the number of customers (class-1 + class-2) present in the system and y is the number of class-2 customers present in service, with $0 \leq y \leq s - c \leq x$. We assume that at the beginning of the interval (i.e., at $t = 0$), the system is in state (x_0, y_0), and denote by $p^t_{x,y}$ the transient probability to be in state (x, y) at time $t \geq 0$. The

evolution of the system state is governed by the following set of equations:

$$\frac{\partial p^t_{s-c,s-c}}{\partial t} = -\lambda p^t_{s-c,s-c} + \mu_1 p^t_{s-c+1,s-c} + \mu_1 p^t_{s-c,s-c-1}, \tag{1}$$

$$\frac{\partial p^t_{s-c+x,s-c}}{\partial t} = -(\lambda + x\mu_1 + (s-c)\mu_2)p^t_{s-c+x,s-c} + \lambda p^t_{s-c+x-1,s-c} + (x+1)\mu_1 p^t_{s-c+x+1,s-c}$$
$$\text{for } 1 \leq x \leq c-1,$$

$$\frac{\partial p^t_{s+x,s-c}}{\partial t} = -(\lambda + c\mu_1 + (s-c)\mu_2)p^t_{s+x,s-c} + \lambda p^t_{s+x-1,s-c} + c\mu_1 p^t_{s+x+1,s-c} \text{ for } x \geq 0,$$

$$\frac{\partial p^t_{s-c,y}}{\partial t} = -(\lambda + (s-c-y)\mu_1)p^t_{s-c,y} + \mu_1(s-c+1-y)p^t_{s-c+1,y} + (s-c-(y-1))\mu_1 p^t_{s-c,y-1}$$
$$+ (y+1)\mu_2 p^t_{s-c+1,y+1} \text{ for } 1 \leq y \leq s-c-1,$$

$$\frac{\partial p^t_{s-c+x,y}}{\partial t} = -(\lambda + (s-c+x-y)\mu_1 + y\mu_2)p^t_{s-c+x,y} + \lambda p^t_{s-c+x-1,y}$$
$$+ (s-c+x+1-y)\mu_1 p^t_{s-c+x+1,y} + (y+1)\mu_2 p^t_{s-c+x+1,y+1} \text{ for } 1 \leq x \leq c-1, 0 \leq y \leq s-c-1,$$

$$\frac{\partial p^t_{s+x,y}}{\partial t} = -(\lambda + (s-y)\mu_1 + y\mu_2)p^t_{s+x,y} + \lambda p^t_{s+x-1,y} + (s-y)\mu_1 p^t_{s+x+1,y}$$
$$+ (y+1)\mu_2 p^t_{s+x+1,y+1} \text{ for } x \geq 0, 0 \leq y \leq s-c-1, \text{and}$$

$$\frac{\partial p^t_{s-c,0}}{\partial t} = -(\lambda + (s-c)\mu_1)p^t_{s-c,0} + \mu_1(s-c+1)p^t_{s-c+1,0} + \mu_2 p^t_{s-c+1,1}.$$

These equations are complex to analyze and do not lead to simple solutions due to the two-dimensional problem. In the particular cases where $\mu_1 = \mu_2$ or when $c = 0$, we develop a method to obtain the Laplace transforms of the transient probabilities in Section 4. Having Laplace transforms is interesting as it allows one particular transient probability to be determined without knowing the complete distribution of the system state. However, the involved expressions cannot be inverted explicitly. In Section 5, we instead focus on the equivalent discrete time formulation of Equation (1) in a truncated system. This in turn allows for a better understanding of the impact of the system parameters.

4 Laplace transforms of the transient probabilities

In this section, we determine the Laplace transforms (LT) of the transient probabilities for particular cases of the blended queue. We introduce the LT of $p^t_{x,y}$, $q^\theta_{x,y}$, defined by $q^\theta_{x,y} = \int_{t=0}^{\infty} e^{-\theta t} p^t_{x,y} \mathrm{d}t$. Since the LT of $\frac{\partial p^t_{x,y}}{\partial t}$ is $\theta q^\theta_{x,y} - p^0_{x,y}$, the set of equations governing the evolution of the system state can be transformed into a linear system. However, this set of equations is complex and can be solved only through a numerical procedure. We focus here on two special cases where $q^\theta_{x,y}$ can be determined. First, we consider the case

with equal service rates between class-1 and class-2 customers, $\mu_1 = \mu_2 = \mu$, and next we consider the case without reservation, $c = 0$. Other cases could be considered. For instance, the case $c = s$ (i.e., full reservation) corresponds to an M/M/s queue because class-2 customers are not served.

In the general case, the analysis for $c = 0$ can be employed to express the LT of the transient probabilities when all agents are busy. However, a finite set of equations remains to be solved, corresponding to situations where at least one agent is idling and the queue is empty. With the approach developed for the case $\mu_1 = \mu_2$ when at least one agent is idling, we can generate functions, close to $A_x(\theta)$ and $B_x(\theta)$ (see Equations (4) and (5)), such that the LT of the transient probabilities are expressed as a linear combination of these functions. There remains to determine the coefficients of this linear combination. This could be done using the equations that $q^\theta_{x,y}$ satisfy. However, this does not lead to simple relations, nor explicit solutions. The complexity can already be seen in the computation of the coefficients $c_{k,k}$ for $k = 0, 1, \cdots, s$ in the proof of Proposition 2 in the appendix for the simpler case $c = 0$.

Analysis with equal service rates. We analyze the case $\mu = \mu_1 = \mu_2$, where we do not need to distinguish between class-1 and class-2 customers in service. Therefore, the Markov chain representing the evolution of the system state becomes one-dimensional, as y does not need to be specified. We then omit the index y in $q^\theta_{x,y}$ such that q^θ_x expresses the LT of the transient probability of having x customers in the system. We determine q^θ_x by solving the following set of equations:

$$-(\lambda + \theta + x\mu)q^\theta_x + \lambda q^\theta_{x-1} + (x+1)\mu q^\theta_{x+1} = -\delta_{x,x_0} \text{ for } s - c \leq x \leq s - 1 \text{ and} \tag{2}$$

$$-(\lambda + \theta + s\mu)q^\theta_x + \lambda q^\theta_{x-1} + s\mu q^\theta_{x+1} = -\delta_{x,x_0} \text{ for } x \geq s, \tag{3}$$

with $q^\theta_{s-c-1} = 0$ and where the system is at state x_0 at time $t = 0$ and δ_{x,x_0} is the Kronecker symbol defined by $\delta_{x,x_0} = 1$ if $x = x_0$ and $\delta_{x,x_0} = 0$ if $x \neq x_0$. We express the solutions of (2) and (3) in Proposition 1 in terms of z_a, z_b, and $W_{m,n}$, where

$$z_a = \frac{\lambda + \theta + s\mu - \sqrt{(\lambda + \theta + s\mu)^2 - 4\lambda s\mu}}{2s\mu}, z_b = \frac{\lambda + \theta + s\mu + \sqrt{(\lambda + \theta + s\mu)^2 - 4\lambda s\mu}}{2s\mu}, \text{ and}$$

$W_{m,n} = A_m(\theta)B_{n-1}(\theta) - A_{m-1}(\theta)B_n(\theta)$, with

$$A_x(\theta) = \frac{1}{2i\pi}\int_{C_1} z^{-(x+1)}e^{\frac{\lambda}{\mu}z}(1-z)^{-\theta/\mu}\mathrm{d}z, \text{ and} \tag{4}$$

$$B_x(\theta) = \frac{1}{2i\pi}\int_{C_2} z^{-(x+1)}e^{\frac{\lambda}{\mu}z}(z-1)^{-\theta/\mu}\mathrm{d}z, \tag{5}$$

where the contour C_1 is defined as a small circle in the z-plane on which $|z| < 1$ and C_2 is a contour which goes from $-\infty - i\epsilon$ to $-\infty + i\epsilon$ for $\epsilon > 0$, encircling $z = 1$ in the counterclockwise sense.

Proposition 1. *The solutions of (2) and (3) are given by*

$$q_x^\theta = \frac{W_{x,s-c-1}(W_{x_0,s-1} - W_{x_0,s}z_a^{-1})}{\lambda\frac{\left(\frac{\lambda}{\mu}\right)^{x_0}}{x_0!}\frac{e^{\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^{\theta-1}}{\Gamma(\theta)}\left(z_a^{-1}W_{s,s-c-1} - W_{s-1,s-c-1}\right)} \text{ for } s-c \le x \le x_0,$$

$$q_x^\theta = \frac{W_{x_0,s-c-1}(W_{x,s-1} - W_{x,s}z_a^{-1})}{\lambda\frac{\left(\frac{\lambda}{\mu}\right)^{x_0}}{x_0!}\frac{e^{\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^{\theta-1}}{\Gamma(\theta)}\left(z_a^{-1}W_{s,s-c-1} - W_{s-1,s-c-1}\right)} \text{ for } x_0 \le x \le s, \text{ and}$$

$$q_x^\theta = \frac{z_a^{x-s}}{\mu}\frac{x_0!}{s!}\frac{\left(\frac{\lambda}{\mu}\right)^{s-1-x_0}W_{x_0,s-c-1}}{z_a^{-1}W_{s,s-c-1} - W_{s-1,s-c-1}} \text{ for } x \ge s,$$

if $x_0 \le s$ *and*

$$q_x^\theta = \frac{\left(\frac{s\mu}{\lambda z_a}\right)^{x_0-s}W_{x,s-c-1}}{s\mu z_b W_{s,s-c-1} - \lambda W_{s-1,s-c-1}} \text{ for } s-c \le x \le s,$$

$$q_x^\theta = \frac{\left(\frac{s\mu}{\lambda z_a}\right)^{x_0-s}}{s\mu(z_b - z_a)}\left[z_b^{x-s} + z_a^{x-s}\frac{\frac{\lambda}{s\mu}\frac{W_{s-1,s-c-1}}{W_{s,s-c-1}} - z_a}{z_b - \frac{\lambda}{s\mu}\frac{W_{s-1,s-c-1}}{W_{s,s-c-1}}}\right] \text{ for } s \le x \le x_0, \text{ and}$$

$$q_x^\theta = \frac{\left(\frac{s\mu}{\lambda z_a}\right)^{x_0-s}}{s\mu(z_b - z_a)}z_a^{x-s}\left[\left(\frac{z_b}{z_a}\right)^{x_0-s} + \frac{\frac{\lambda}{s\mu}\frac{W_{s-1,s-c-1}}{W_{s,s-c-1}} - z_a}{z_b - \frac{\lambda}{s\mu}\frac{W_{s-1,s-c-1}}{W_{s,s-c-1}}}\right] \text{ for } x \ge x_0,$$

if $x_0 \ge s$.

The expression of $A_x(\theta)$ in (4) can be obtained explicitly. By expanding $(1-z)^{-\frac{\theta}{\mu}}$, we obtain $(1-z)^{-\frac{\theta}{\mu}} = 1 + \frac{\theta}{\mu}z + \frac{\theta}{\mu}\left(\frac{\theta}{\mu}+1\right)\frac{z^2}{2!} + \cdots$. Therefore, we deduce that

$$A_x(\theta) = \sum_{k=0}^{x}\frac{\left(\frac{\lambda}{\mu}\right)^{x-k}}{k!(x-k)!}\frac{\Gamma\left(\frac{\theta}{\mu}+k\right)}{\Gamma\left(\frac{\theta}{\mu}\right)},$$

where $\Gamma(z)$ is the Gamma function defined by $\Gamma(z) = \int_{t=0}^{\infty} t^{z-1}e^{-t}\mathrm{d}t$. The computation of $B_x(\theta)$ is more

complex. In the online appendix, we explain how this complex integral can be derived.

Analysis without reservation. We now consider the case without reservation when the service rates are not necessarily equal. It means that whenever a service is completed, a new service starts either with a class-1 or class-2 customer. In this case, we need to determine the solutions of the following system:

$$-(\lambda+\theta+(s-y)\mu_1+y\mu_2)q^\theta_{x,y}+\lambda q^\theta_{x-1,y}+(s-y)\mu_1 q^\theta_{x+1,y} \tag{6}$$
$$+(y+1)\mu_2 q^\theta_{x+1,y+1}=-\delta_{(x,y),(x_0,y_0)} \text{ for } x>s, 0\le y\le s \text{ and}$$
$$-(\lambda+\theta+(s-y)\mu_1)q^\theta_{s,y}+(s-y)\mu_1 q^\theta_{s+1,y}+(y+1)\mu_2 q^\theta_{s+1,y+1} \tag{7}$$
$$+(s-y-1)\mu_1 q^\theta_{s,y-1}=-\delta_{(x,y),(x_0,y_0)} \text{ for } x=s, 0\le y\le s,$$

where $\delta_{(x,y),(x_0,y_0)}=1$ if $(x,y)=(x_0,y_0)$ and $\delta_{(x,y),(x_0,y_0)}=0$ if $(x,y)\neq(x_0,y_0)$, with the convention $q^\theta_{x,s+1}=0$. In Proposition 2, we express the LT of the transient probabilities, in terms of $z_{a,y}$, $z_{b,y}$, α_k and β_k, where

$$z_{a,y}=\frac{\lambda+(s-y)\mu_1+y\mu_2+\theta-\sqrt{(\lambda+(s-y)\mu_1+y\mu_2+\theta)^2-4\lambda\mu_1(s-y)}}{2(s-y)\mu_1},$$
$$z_{b,y}=\frac{\lambda+(s-y)\mu_1+y\mu_2+\theta+\sqrt{(\lambda+(s-y)\mu_1+y\mu_2+\theta)^2-4\lambda\mu_1(s-y)}}{2(s-y)\mu_1},$$
$$\alpha_k=\frac{\mu_2 z_{a,k}}{\mu_1(1-z_{a,k})-\mu_2}, \text{ and } \beta_k=\frac{\mu_2 z_{b,k}}{\mu_1(1-z_{b,k})-\mu_2}.$$

Proposition 2. *The solutions of (6) and (7) are given by*

$$q^\theta_{x,y}=\sum_{k=y}^{s}\binom{k}{y}\alpha_k^{k-y}c_{k,k}z_{a,k}^{x-s},$$

for $y_0<y\le s$ *or* $0\le y\le y_0$ *and* $x>x_0-y_0+y$, *and*

$$q^\theta_{x,y}=\sum_{k=y,k\neq y_0}^{s}\binom{k}{y}\alpha_k^{k-y}c_{k,k}z_{a,k}^{x-s}+\binom{y_0}{y}\alpha_{y_0}^{y_0-y}d_{y_0,y_0}z_{a,y_0}^{x-s}+\binom{y_0}{y}\beta_{y_0}^{y_0-y}e_{y_0,y_0}z_{b,y_0}^{x-s},$$

for $0\le y\le y_0$ *and* $x\le x_0-(y_0-y)$.

In the expressions of $q^\theta_{x,y}$ in Proposition 2, the coefficients $c_{0,0}, c_{1,1}, \cdots, c_{s,s}$, d_{y_0,y_0} and e_{y_0,y_0} remain to be determined. This can be done by the initial condition and the boundary equations at state $x=s$ and $0\le y\le s$. The details are provided after the proof of Proposition 2 in the appendix.

5 Discrete time approximation using the uniformization technique

The LT technique presented in Section 4 is interesting because it leads to the derivation of the performance measures without computing each transient probability [10]. Furthermore, the case $\mu_1 = \mu_2$ can be connected to the analysis of the M/M/s queue, as the M/M/s queue is a special case of the blended queue with $c = s$. However, the involved expressions in the LT are difficult to analyze and can only be used for numerical purposes. In this section, we instead use the uniformization technique, which consists of discretizing the elapse of time using a parameter $\epsilon > 0$, such that $\frac{\partial p^t_{x,y}}{\partial t} \approx \frac{p^{t+\epsilon}_{x,y} - p^t_{x,y}}{\epsilon}$. By letting ϵ tend to zero, this approximation tends to the exact set of equations governing the evolution of the system state. This approach is efficient to compute the transient distribution of finite states Markov chain [19, 17, 16, 18]. For numerical experiments, to apply this approach for the blended queue, the number of customers in the queue needs to be bounded with parameter N. The selection of the parameters ϵ and N determines the quality of the approximation.

After uniformization, the system (1) becomes

$$
\begin{aligned}
p^{t+\epsilon}_{s-c,s-c} &= (1-\lambda\epsilon)p^t_{s-c,s-c} + \mu_1\epsilon p^t_{s-c+1,s-c} + \mu_1\epsilon p^t_{s-c,s-c-1}, \qquad (8)\\
p^{t+\epsilon}_{s-c+x,s-c} &= (1-(\lambda + x\mu_1 + (s-c)\mu_2)\epsilon)p^t_{s-c+x,s-c} + \lambda\epsilon p^t_{s-c+x-1,s-c} + (x+1)\mu_1\epsilon p^t_{s-c+x+1,s-c}\\
&\qquad \text{for } 1 \le x \le c-1,\\
p^{t+\epsilon}_{s+x,s-c} &= (1-(\lambda + c\mu_1 + (s-c)\mu_2)\epsilon)p^t_{s+x,s-c} + \lambda\epsilon p^t_{s+x-1,s-c} + c\mu_1\epsilon p^t_{s+x+1,s-c} \text{ for } 0 \le x < N,\\
p^{t+\epsilon}_{s-c,y} &= (1-(\lambda + (s-c-y)\mu_1)\epsilon)p^t_{s-c,y} + \mu_1(s-c+1-y)\epsilon p^t_{s-c+1,y} + (s-c-(y-1))\mu_1\epsilon p^t_{s-c,y-1}\\
&\qquad + (y+1)\mu_2\epsilon p^t_{s-c+1,y+1} \text{ for } 1 \le y \le s-c-1,\\
p^{t+\epsilon}_{s-c+x,y} &= (1-(\lambda + (s-c+x-y)\mu_1 + y\mu_2)\epsilon)p^t_{s-c+x,y} + \lambda\epsilon p^t_{s-c+x-1,y}\\
&\qquad + (s-c+x+1-y)\mu_1\epsilon p^t_{s-c+x+1,y} + (y+1)\mu_2\epsilon p^t_{s-c+x+1,y+1} \text{ for } 1 \le x \le c-1, 0 \le y \le s-c-1,\\
p^{t+\epsilon}_{s+x,y} &= (1-(\lambda + (s-y)\mu_1 + y\mu_2)\epsilon)p^t_{s+x,y} + \lambda\epsilon p^t_{s+x-1,y} + (s-y)\mu_1\epsilon p^t_{s+x+1,y}\\
&\qquad + (y+1)\mu_2\epsilon p^t_{s+x+1,y+1} \text{ for } 0 \le x < N, 0 \le y \le s-c-1, \text{and}\\
p^{t+\epsilon}_{s-c,0} &= (1-(\lambda + (s-c)\mu_1)\epsilon)p^t_{s-c,0} + \mu_1(s-c+1)\epsilon p^t_{s-c+1,0} + \mu_2\epsilon p^t_{s-c+1,1}.
\end{aligned}
$$

At states $x = s + N$ and $0 \le y \le s - c$, we need to modify the equations as follows:

$$
p^{t+\epsilon}_{s+N,y} = (1-((s-y)\mu_1 + y\mu_2)\epsilon)p^t_{s+N,y} + \lambda\epsilon p^t_{s+N-1,y} \text{ for } 0 \le y \le s-c.
$$

We estimate $E(Q)_t = \sum_{x=0}^{N}\sum_{y=0}^{s-c} x p_{s+x,y}^t$ and $E(T)_t = \mu_2 \sum_{x=s-c}^{s+N}\sum_{y=0}^{s-c} y p_{x,y}^t$ by selecting $\epsilon = \frac{t}{n}$ for a sufficiently high value of n. For the average values of these measures on the interval $I = [0, t^*]$, we select $\epsilon = \frac{t^*}{n}$ and obtain

$$\overline{E(Q)_I} = \frac{\epsilon}{t^*}\sum_{k=1}^{n}\sum_{x=0}^{N}\sum_{y=0}^{s-c} x p_{s+x,y}^{k\epsilon}, \text{ and } \overline{E(T)_I} = \frac{\mu_2\epsilon}{t^*}\sum_{k=1}^{n}\sum_{x=s-c}^{s+N}\sum_{y=0}^{s-c} y p_{x,y}^{k\epsilon}.$$

In Section 5.1, we employ this method to investigate the behavior of the transient blended queue and provide insights on the selection of the optimal threshold level. Next, in Section 5.2, we provide monotonicity results in the case $\mu_1 = \mu_2$ to support the observations.

5.1 Numerical observations

In Table 1, we provide the average transient performance measures for different values of n during an interval of time of one time unit of observation and for the stationary case (i.e., $t^* = \infty$). We present a case where, initially, 4 agents are idling with $x_0 = 6$ and one where the queue is congested with $x_0 = 24$. In both cases, 2 agents are initially busy with serving class-2 customers, $y_0 = 2$. We observe the convergence of the

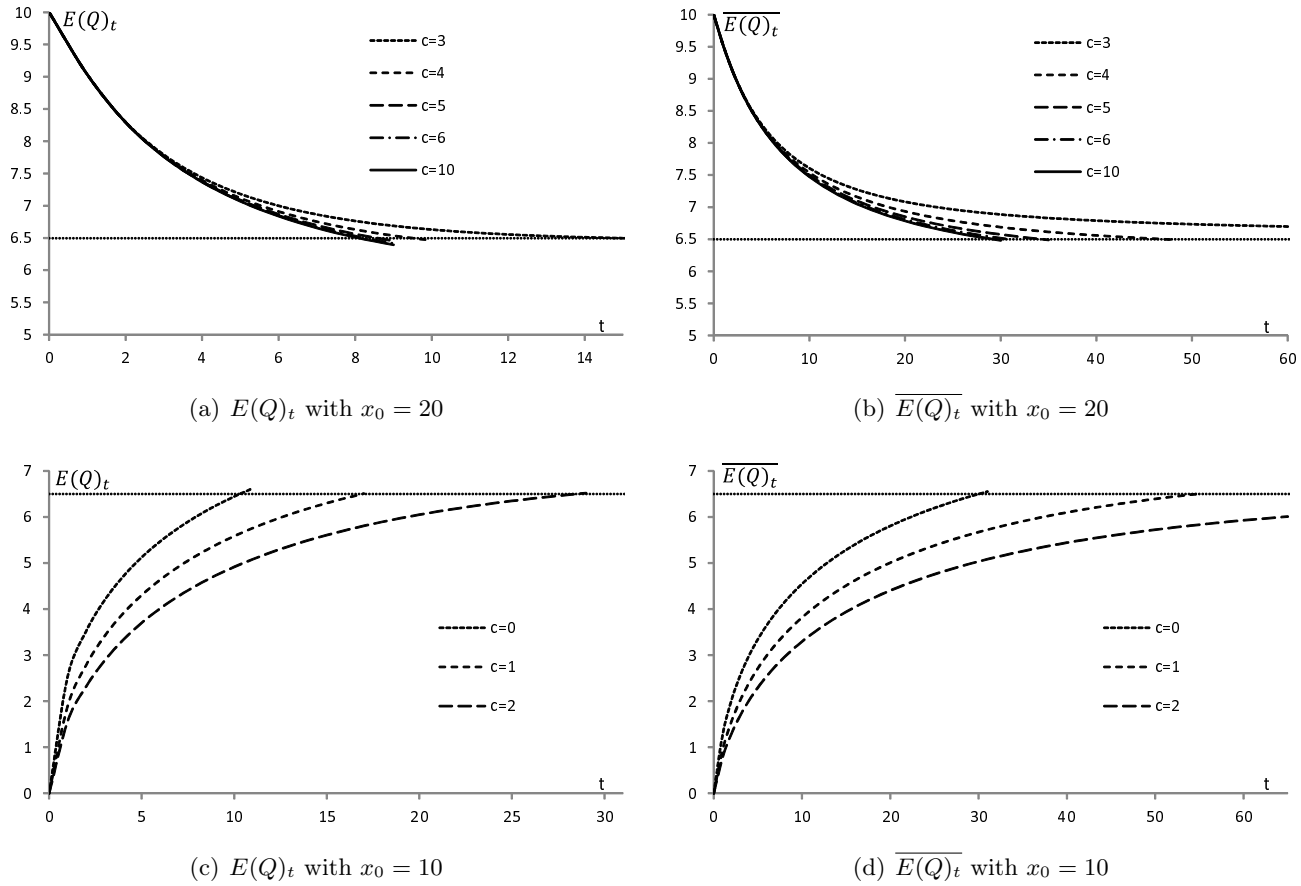
Table 1: Computation of the transient performance measures ($\mu_1 = 3, \mu_2 = 4, s = 10, c = 6, t^* = 1$, $N = 20$)

		$\overline{E(T)_I}$				$\overline{E(Q)_I}$			
	λ	$n = 100$	$n = 500$	$n = 1000$	$t^* = \infty$	$n = 100$	$n = 500$	$n = 1000$	$t^* = \infty$
	0.01	11.627	11.579	11.557	15.992	0.000	0.000	0.000	0.000
(x_0, y_0)	10	6.095	6.110	6.105	6.959	0.002	0.002	0.002	0.001
$= (6, 2)$	20	3.312	3.358	3.362	1.752	0.142	0.140	0.140	0.240
	25	2.676	2.729	2.735	0.581	0.483	0.478	0.476	1.716
	0.01	3.536	3.613	3.614	15.992	3.226	3.338	3.352	0.000
(x_0, y_0)	10	2.120	2.192	2.200	6.959	5.159	5.259	5.271	0.001
$= (24, 2)$	20	1.906	1.968	1.975	1.752	8.824	8.875	8.876	0.240
	25	1.892	1.953	1.960	0.581	10.990	11.018	11.013	1.716

performance measures as n increases. Since the horizon of observation is short with average performance measures, the initial condition strongly impacts $\overline{E(T)_I}$ and $\overline{E(Q)_I}$, which is observed by the high distance between the transient and stationary results. As for the stationary case, $\overline{E(T)_I}$ decreases and $\overline{E(Q)_I}$ increases with the arrival rate. Furthermore, the impact of λ depends on the initial condition. When the system is initially with idle capacity, the sensitivity of $\overline{E(T)_I}$ in λ is high while the sensitivity of $\overline{E(Q)_I}$ in λ is low. The opposite is true when starting from a congested situation. This suggests that if the system needs to reset the reservation threshold due to the anticipation of an increase in the arrival rate, the new reservation threshold should be (not be) highly reduced if the system is observed without (with) idle agents. In the following, we further investigate this question.

In Figure 1, we present the evolution of $E(Q)_t$ and $\overline{E(Q)_t}$ over time for different values of c in a situation

with 10 agents, identical service rates with $\mu_1 = \mu_2 = 1$, an arrival rate of $\lambda = 9$, and either $x_0 = 20$ or $x_0 = 10$ (i.e., either 10 customers or no customer in the queue). We do not specify the number of class-2 customers in service since it does not impact the evolution of $E(Q)_t$ when $\mu_1 = \mu_2$. We want to maintain the expected number of customers in the queue at below 6.5. The curves are drawn until the objective of 6.5 is reached. For the optimization problem, it is optimal to have c as low as possible, provided that the expected number of customers in the queue remains below 6.5. In the stationary case, it is optimal to have $c = 4$, which leads to $E(Q)_\infty = 6.164$ and $E(T)_\infty = 0.332$.



(a) $E(Q)_t$ with $x_0 = 20$

(b) $\overline{E(Q)_t}$ with $x_0 = 20$

(c) $E(Q)_t$ with $x_0 = 10$

(d) $\overline{E(Q)_t}$ with $x_0 = 10$

Figure 1: Evolution of $E(Q)_t$ and $\overline{E(Q)_t}$ ($s = 10$, $\lambda = 9$, $\mu_1 = \mu_2 = 1$, $N = 45$)

Depending on the horizon of observation, $c = 4$ is not optimal in the transient regime. In the case $x_0 = 20$ (Figures 1(a) and 1(b)), the system is initially with too many customers, as $E(Q)_0 = 10 > 6.5$. Therefore, the reservation threshold c should be selected sufficiently high to reduce the expected number of customers in the queue. As expected, we observe that increasing c reduces the time before reaching the objective of 6.5. However, any further increase of c above 4 only has marginal effects. For $E(Q)_t$, the objective of 6.5 is reached between 8 and 9 time units for $c = 5, 6$, and 10 (Figure 1(a)), while for $\overline{E(Q)_t}$, the objective is reached between 30 and 35 time units (Figure 1(b)). This shows that having $c > 4$ can be

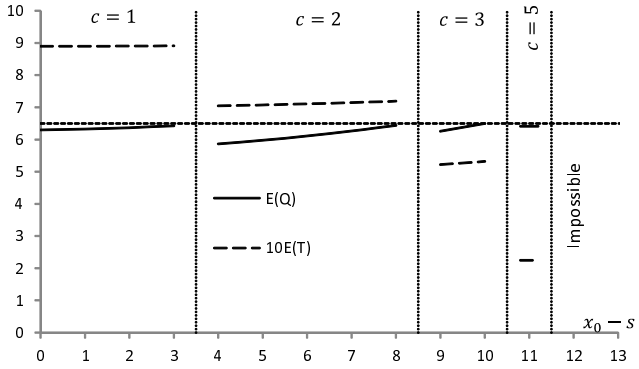
beneficial only during a restricted time interval. Surprisingly, we observe that $c = 3 < 4$ can be optimal for $E(Q)_t$ (Figure 1(a)). This can be explained by the behavior of $E(Q)_t$ starting from a congested situation. We observe that first $E(Q)_t$ decreases in t reaching a minimal value below its stationary value and later increases to reach the stationary value. The reason is that all agents are busy in a congested situation, which leads to the most efficient use of the system capacities and explains why we first observe a decrease of $E(Q)_t$ below its stationary value. This observation has also been made for the M/M/s queue (e.g., see Figure 1 in [7]). For decision making, it means that having the reservation threshold below its stationary value can be optimal, even if the system is highly congested at time $t = 0$.

In the case $x_0 = 10$ (Figures 1(c) and 1(d)), the system is initially with no customer in the queue and all agents busy. We select $c = 0, 1, 2$ to observe the time to reach the objective of 6.5. Contrary to the case $x_0 = 20$, the sensitivity to c is stronger in this case. This indicates that each threshold $c = 0, 1, 2$ has a large interval of optimality. It is then advisable to select c as low as possible, provided that the expected number in the queue does not reach 6.5. These observations are supported by the strong convexity of $E(Q)_t$ in c, as proven in Section 5.2.

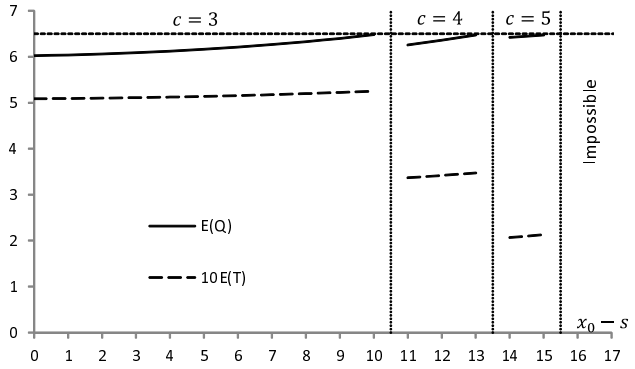
In Figure 2, we provide the optimal reservation threshold for different initial conditions on the number of customers in the queue $x_0 - s$ with the objective of $E(Q)_t \leq 6.5$ at time $t = 15, 30$, and 60 with the same system parameters as those of Figure 1. We also specify the corresponding value for $E(Q)_t$ and $10 \times E(T)_t$. The value of $E(T)_t$ is estimated through $\mu E(N_B)_t - \lambda$, where $E(N_B)_t$ is the expected number of busy agents at time $t \geq 0$. These figures confirm the observations made in Figure 1. Over a short time horizon the stationary threshold $c = 4$ may never be optimal, as in Figure 2(a). Furthermore, in many cases, even when $x_0 - s > 6.5$, we should select $c < 4$, while having $c > 4$ is almost never optimal. For decision-making, it means that the stationary threshold is in many cases "too safe" for the objective expected queue length. Given that we observe in Figure 2 that the rate of served class-2 customers is highly sensitive to the selection made for c, it is often advisable over a short time horizon to reserve less agents than what the stationary results would suggest.

5.2 Monotonicity properties of $E(Q)_t$ for the approximated model

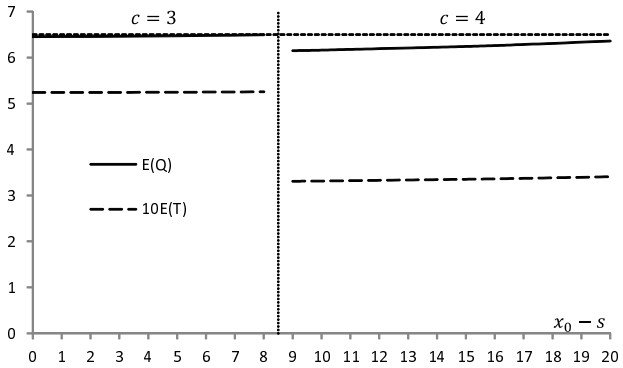
To support the observations made in Section 5.1, we provide the transient monotonicity properties of $E(Q)_t$ in Theorem 1 for the discrete time model. Although the approximated model can lead to a transient distribution of the queue length that is as close as wanted to the exact distribution (by selecting a high enough value for N and a low enough value for ϵ), Theorem 1 does not necessarily prove the same monotonicity



(a) $t = 15$



(b) $t = 30$



(c) $t = 60$

Figure 2: Optimal reservation threshold for $E(Q)_t$ ($s = 10$, $\lambda = 9$, $\mu_1 = \mu_2 = 1$, $N = 45$)

properties for the exact model. For future research, other techniques should be involved to prove the result for the exact model like sample path arguments [6] or stochastic ordering [20]. It should be noted that we could define the approximated discrete time model with $N = \infty$, as the maximal event rate is bounded. However in Theorem 1, we prove the monotonicty properties for $N < \infty$ in order to be consistent with the setting of the numerical experiments of Section 5.1.

This analysis is made for the case $\mu_1 = \mu_2$ since the complexity of the case $\mu_1 \neq \mu_2$ does not allow us to obtain the desired results. We prove the impact of the system parameters by showing the propagation of one property from the time instant t to the time instant $t + \epsilon$. Specifically, we show the monotonicity properties for the quantity $Q_x^t = \sum_{k=s-c}^{x} p_k^t$ and $Z_x^{c,t} = \sum_{k=x}^{s+N} p_k^t$, where x is the number of customers present in the system, c is the reservation threshold and p_k^t is transient probability to find k customers in the system. Recall that the index y does not need to be specified with equal service rates. The equations governing the

evolution of the state of the system for the approximated model are as follows:

$$
\begin{aligned}
p_{s-c}^{t+\epsilon} &= (1-\lambda\epsilon)p_{s-c}^{t} + (s-c+1)\mu\epsilon p_{s-c+1}^{t}, \\
p_{s-x}^{t+\epsilon} &= (1-(\lambda+(s-x)\mu)\epsilon)p_{s-x}^{t} + \lambda\epsilon p_{s-x-1}^{t} + (s-x+1)\mu\epsilon p_{s-x+1}^{t} \text{ for } 1 \leq x \leq c-1, \\
p_{x+s}^{t+\epsilon} &= (1-(\lambda+s\mu)\epsilon)p_{x+s}^{t} + \lambda\epsilon p_{s+x-1}^{t} + s\mu\epsilon p_{s+x+1}^{t} \text{ for } 0 \leq x \leq N-1, \text{ and} \\
p_{s+N}^{t+\epsilon} &= (1-s\mu\epsilon)p_{s+N}^{t} + \lambda\epsilon p_{s+N-1}^{t},
\end{aligned}
\tag{9}
$$

given an initial condition x_0 at time $t = 0$.

Theorem 1. *The following holds for the approximated model:*

- $E(Q)_t$ *is increasing in* λ.
- $E(Q)_t$ *is decreasing and convex in* c.
- $E(Q)_t$ *is increasing in* x_0.

The monotonicty in λ and c are the same as those proven in the stationary case for the exact model (e.g., see Theorem 2 in [12]). The convexity of $E(Q)_t$ in c explains the observations made in Section 5.1, according to which the optimal transient reservation threshold should be less than or equal to the stationary one in most cases. We mention that the convexity in λ does not hold as we consider a truncated system with at most N customers in the queue. In our numerical illustrations in Figure 2, we observed that $E(Q)_t$ is convex in x_0 in addition to being increasing; this is, however, not the case at the beginning of the interval I as explained at the end of the online appendix.

Some of the results of Theorem 1 can be used for $E(T)_t$. With the one-dimensional Markov chain formulation, $E(T)_t$ can be estimated through $E(T)_t = \mu E(N_B)_t^c - \lambda$, where $E(N_B)_t^c$ is the expected number of busy servers at time t with reservation threshold c. If we express $E(N_B)_t^c$ as $E(N_B)_t^c = (s-c)Z_{s-c}^{c,t} + \sum_{k=s-c+1}^{s} Z_k^{c,t}$, we deduce that $E(N_B)_t^{c+1} - E(N_B)_t^c = Z_{s-c}^{c+1,t} - 1 + \sum_{k=s-c+1}^{s} (Z_k^{c+1,t} - Z_k^{c,t}) \leq 0$, which proves that $E(N_B)_t^c$ and $E(T)_t$ are decreasing in c. However, $E(N_B)_t^c$ and $E(T)_t$ are not convex in c. This was also not the case in the stationary regime (e.g., see Figure 3 in [1]).

References

[1] S. Bhulai and G. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48(8):1434–1438, 2003.

[2] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A.N. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

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

[4] N. Gans and Y. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271, 2003.

[5] L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

[6] O. Jouini and Y. Dallery. Monotonicity properties for multiserver queues with reneging and finite waiting lines. *Probability in the Engineering and Informational Sciences*, 21(3):335–360, 2007.

[7] W.D. Kelton and A.M. Law. The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

[8] B. Legros. Reservation, a tool to reduce the balking effect and the probability of delay. *Operations Research Letters*, 45(6):592–597, 2017.

[9] B. Legros. Agents' self-routing for blended operations to balance inbound and outbound services. *Production and Operations Management*, 30(10):3599–3614, 2021.

[10] B. Legros. Transient analysis of an affine Queue-Hawkes process. *Operations Research Letters*, 49(3):393–399, 2021.

[11] B. Legros, O. Jouini, and G. Koole. Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430, 2015.

[12] B. Legros, O. Jouini, and G. Koole. Should we wait before outsourcing? Analysis of a revenue-generating blended contact center. *Manufacturing & Service Operations Management*, 23(5):1118–1138, 2021.

[13] J. Little. A proof for the queuing formula: $L = \lambda W$. *Operations Research*, 9(3):383–387, 1961.

[14] G. Pang and O. Perry. A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91, 2014.

[15] T. Phung-Duc, W. Rogiest, Y. Takahashi, and H. Bruneel. Retrial queues with balanced call blending: Analysis of single-server and multiserver model. *Annals of Operations Research*, 239(2):429–449, 2016.

[16] R. Pulungan and H. Hermanns. Transient analysis of CTMCs: Uniformization or matrix exponential? *IAENG International Journal of Computer Science*, 45(2):267–274, 2018.

[17] R. Sidje, K. Burrage, and S. MacNamara. Inexact uniformization method for computing transient distributions of Markov chains. *SIAM Journal on Scientific Computing*, 29(6):2562–2580, 2007.

[18] N. van Dijk, S. van Brummelen, and R. Boucherie. Uniformization: Basics, extensions and applications. *Performance evaluation*, 118:8–32, 2018.

[19] A. Van Moorsel and W. Sanders. Adaptive uniformization. *Stochastic Models*, 10(3):619–647, 1994.

[20] W. Whitt. Comparing counting processes and queues. *Advances in Applied Probability*, 13(1):207–220, 1981.

A Proof of Proposition 1

Proof. We first consider Equation (2) in the case $x \neq x_0$. We introduce a function $F(z)$ for $z \in \mathbb{C}$ and we express q_x^θ as

$$q_x^\theta = \int_C z^{-(x+1)} F(z) \mathrm{d}z,$$

where C is a contour such that there are no boundary contributions arising in the integral from endpoints of C. This allows us to use the integration by part and show that $xq_x^\theta = \displaystyle\int_C z^{-(x+1)} z F'(z) \mathrm{d}z$. Equation (2) can then be rewritten as

$$\int_C z^{-(x+1)} \left(-F(z) \left[\lambda(1-z) + \theta \right] + \mu F'(z)(1-z) \right) \mathrm{d}z = 0.$$

Therefore, $F(z)$ is one solution of the differential equation

$$-F(z) \left[\lambda(1-z) + \theta \right] + \mu F'(z)(1-z) = 0.$$

Consequently, $F(z)$ is proportional with $e^{\frac{\lambda}{\mu} z}(1-z)^{-\theta/\mu}$. We thus determine two independent solutions of (2) by selecting two different contours encircling $z = 0$. We consider the contour C_1 defined as a small circle

in the z-plane, on which $|z| < 1$, and C_2 which goes from $-\infty - i\epsilon$ to $-\infty + i\epsilon$ for $\epsilon > 0$, encircling $z = 1$ in the counterclockwise sense. These contours define $A_x(\theta)$ and $B_x(\theta)$. Note that for (4) the integrand is analytic inside the unit circle, as we consider $(1-z)^{-\theta/\mu} = |1-z|^{-\theta/\mu} e^{-i\frac{\theta}{\mu}\arg(1-z)}$, with $|\arg(1-z)| < \pi$, such that for $z \in \mathbb{R}$ and $z < 1$, $\arg(1-z) = 0$. For (5), we use the branch $(z-1)^{-\frac{\theta}{\mu}} = |z-1|^{\frac{-\theta}{\mu}} e^{-i\frac{\theta}{\mu}\arg(z-1)}$, where $|\arg(z-1)| < \pi$, so the integrand is analytic in $\mathbb{C} - \{\mathrm{Im}(z) = 0, \mathrm{Re}(z) < 1\}$.

Consider now Equation (3). We need to determine two solutions of the equation in z, $-(\lambda+\theta+s\mu)z + \lambda + s\mu z^2 = 0$. These solutions are z_a and z_b. We can that $0 \leq z_a < 1 < z_b$. Consider the case where $s - c \leq x_0 \leq s - 1$. In this case, we express q_x^θ as

$$
\begin{aligned}
q_x^\theta &= c_1 A_x(\theta) + c_2 B_x(\theta) \text{ for } s - c \leq x \leq x_0, \\
q_x^\theta &= c_3 A_x(\theta) + c_4 B_x(\theta) \text{ for } x_0 \leq x \leq s, \\
q_x^\theta &= c_5 z_a^{x-s} \text{ for } x \geq s.
\end{aligned}
$$

Note that z_b is not part of the expression of q_x^θ since we cannot have $\lim_{x\to\infty} q_x^\theta = \infty$. We determine c_1, c_2, c_3, c_4 and c_5 from the boundary equations. These are given by

$$
\begin{aligned}
&c_1 A_{s-c-1}(\theta) + c_2 B_{s-c-1}(\theta) = 0, \\
&c_1 A_{x_0}(\theta) + c_2 B_{x_0}(\theta) = c_3 A_{x_0}(\theta) + c_4 B_{x_0}(\theta), \\
&\lambda(c_1 A_{x_0-1}(\theta) + c_2 B_{x_0-1}(\theta)) - (\lambda + x_0\mu + \theta)(c_1 A_{x_0}(\theta) + c_2 B_{x_0}(\theta)) + (x_0+1)\mu(c_3 A_{x_0+1}(\theta) + c_4 B_{x_0+1}(\theta)) = -1, \\
&\lambda(c_3 A_{s-1}(\theta) + c_4 B_{s-1}(\theta)) - (\lambda + s\mu + \theta)c_5 + s\mu c_5 z_a = 0, \text{ and} \\
&c_5 = c_3 A_s(\theta) + c_4 B_s(\theta).
\end{aligned}
$$

To express the solutions of this set of equations, we introduce the quantity $W_{m,n} = A_m(\theta)B_{n-1}(\theta) - A_{m-1}(\theta)B_n(\theta)$. In the case where $m = n+1$, $W_{n+1,n}$ is the Wronskian of $A_n(\theta)$ and $B_n(\theta)$. Using the equation defining $A_x(\theta)$ and $B_x(\theta)$, we find that $W_{x+1,x} = \frac{\frac{\lambda}{\mu}}{x+1} W(x, x-1)$. Therefore, $W_{x+1,x} = \frac{\left(\frac{\lambda}{\mu}\right)^{x+1}}{(x+1)!} W_{0,-1}$. Furthermore, $A_0(\theta) = 1$ and $A_{-1}(\theta) = 0$. With $x = -1$, the term in $z^{-(x+1)}$ is removed in the integral defining $B_{-1}(\theta)$. This allows us to obtain $B_{-1}(\theta) = \frac{e^{\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^{\theta-1}}{\Gamma(\theta)}$ and $W_{x+1,x} = \frac{\left(\frac{\lambda}{\mu}\right)^{x+1}}{(x+1)!} \frac{e^{\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^{\theta-1}}{\Gamma(\theta)}$. After some algebra, we then obtain the solution as in Proposition 1.

We next consider the case where $x_0 \geq s$. We may write

$$q_x^\theta = c_1 A_x(\theta) + c_2 B_x(\theta) \text{ for } s - c \leq x \leq s,$$
$$q_x^\theta = c_3 z_a^{x-s} + c_4 z_b^{x-s} \text{ for } s \leq x \leq x_0, \text{ and}$$
$$q_x^\theta = c_5 z_a^{x-s} \text{ for } x \geq x_0.$$

The constants c_1, c_2, c_3, c_4 and c_5 are solutions of

$$c_1 A_{s-c-1}(\theta) + c_2 B_{s-c-1}(\theta) = 0,$$
$$c_1 A_s(\theta) + c_2 B_s(\theta) = c_3 + c_4,$$
$$\lambda(c_1 A_{s-1}(\theta) + c_2 B_{s-1}(\theta)) - (\lambda + s\mu + \theta)(c_3 + c_4) + s\mu(c_3 z_a + c_4 z_b) = 0,$$
$$\lambda(c_3 z_a^{x_0-s-1} + c_4 z_b^{x_0-s-1}) - (\lambda + s\mu + \theta) c_5 z_a^{x_0-s} + s\mu c_5 z_a^{x_0-s+1} = -1, \text{ and}$$
$$c_3 z_a^{x_0-s} + c_4 z_b^{x_0-s} = c_5 z_a^{x_0-s}.$$

This system can also be solved and we find the solutions of Proposition 1. □

B Computation of $B_x(\theta)$

Since

$$(\lambda + \theta + x\mu) A_x(\theta) = \lambda A_{x-1}(\theta) + (x+1)\mu A_{x+1}(\theta), \text{ and}$$
$$(\lambda + \theta + x\mu) B_x(\theta) = \lambda B_{x-1}(\theta) + (x+1)\mu B_{x+1}(\theta),$$

by multiplying the first equation by $B_x(\theta)$ and the second one by $A_x(\theta)$ and next subtracting the two equations, we deduce a relation for the Wronskian of $A_x(\theta)$ and $B_x(\theta)$, defined by $W_{x,x-1} = A_x(\theta)B_{x-1}(\theta) - B_x(\theta)A_{x-1}(\theta)$. This relation is

$$(x+1)\mu W_{x+1,x} = \lambda W_{x,x-1} \text{ for } x \geq 0.$$

Therefore, we have $W_{x,x-1} = \frac{\left(\frac{\lambda}{\mu}\right)^x}{x!} W_{0,-1}$ for $x \geq 0$. Since $A_{-1}(\theta) = 0$ and $A_0(\theta) = 1$, we have $W_{0,-1} = B_{-1}(\theta)$. The expression of $B_{-1}(\theta)$ can be obtained explicitly. We find that $B_{-1}(\theta) = \dfrac{e^{\frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^{\frac{\theta}{\mu}-1}}{\Gamma\left(\frac{\theta}{\mu}\right)}$. This

leads to $W_{x,x-1} = \frac{e^{\frac{\lambda}{\mu}}\left(\frac{\lambda}{\mu}\right)^{x+\frac{\theta}{\mu}-1}}{x!\Gamma\left(\frac{\theta}{\mu}\right)}$ for $x \geq 0$.

Instead of $B_x(\theta)$, consider another solution of the equation defining $A_x(\theta)$ and $B_x(\theta)$, $\overline{B_x}(\theta)$, such that $\overline{B_{-1}}(\theta) = 1$ and $\overline{B_0}(\theta) = 0$. With these values, $A_x(\theta)$ and $\overline{B_x}(\theta)$ are independent. Therefore, we may write $B_x(\theta) = \alpha A_x(\theta) + \beta\overline{B_x}(\theta)$. Consider now the Wronskian of $A_x(\theta)$ and $\overline{B_x}(\theta)$ defined as $\overline{U_x} = A_x(\theta)\overline{B_{x-1}}(\theta) - A_{x-1}(\theta)\overline{B_x}(\theta)$ for $x \geq 0$. We show that $\overline{U_x} = \overline{U_0}\frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}$ for $x \geq 0$. Since $\overline{U_0} = 1$, we deduce that $A_x(\theta)\overline{B_{x-1}}(\theta) - A_{x-1}(\theta)\overline{B_x}(\theta) = \frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}$. We then deduce that

$$\overline{B_x}(\theta) = \frac{A_x(\theta)\overline{B_{x-1}}(\theta) - \frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}}{A_{x-1}(\theta)} \text{ for } x \geq 1.$$

From this expression, we conclude that

$$\overline{B_x}(\theta) = -A_x(\theta)\sum_{k=1}^{x}\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!A_k(\theta)A_{k-1}(\theta)} \text{ for } x \geq 0.$$

In the expression of $B_x(\theta)$, the coefficient β is found by replacing x by -1. We thus obtain $\beta = B_{-1}(\theta) = \frac{e^{\frac{\lambda}{\mu}}\left(\frac{\lambda}{\mu}\right)^{\frac{\theta}{\mu}-1}}{\Gamma\left(\frac{\theta}{\mu}\right)}$. The coefficient α can be expressed as

$$\alpha = \frac{B_x(\theta) - \beta\overline{B_x}(\theta)}{A_x(\theta)} \text{ for } x \geq 0.$$

We have $\frac{B_x(\theta)}{A_x(\theta)} \sim \frac{\Gamma\left(\frac{\theta}{\mu}\right)}{\sqrt{2\pi}e^{\frac{\lambda}{\mu}}\left(\frac{\theta}{\mu}+x\right)^{\frac{\theta}{\mu}+x}}$ as x grows large. Therefore, $\lim_{x\to\infty}\frac{B_x(\theta)}{A_x(\theta)} = 0$. We also have $\frac{\overline{B_x}(\theta)}{A_x(\theta)} = -\sum_{k=1}^{x}\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!A_k(\theta)A_{k-1}(\theta)}$. For a large value of k, we have $A_k(\theta) \sim \frac{e^{\frac{\lambda}{\mu}}k^{\frac{\theta}{\mu}-1}}{\Gamma\left(\frac{\theta}{\mu}\right)}$. As the sum $\sum_{k=1}^{m}\frac{\left(\Gamma\left(\frac{\theta}{\mu}\right)\right)^2\left(\frac{\theta}{\mu}\right)^k}{k!e^{2\frac{\lambda}{\mu}}k^{\frac{\theta}{\mu}-1}(k-1)^{\frac{\theta}{\mu}-1}}$ converges as m tends to infinity, the sum $\sum_{k=1}^{x}\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!A_k(\theta)A_{k-1}(\theta)}$ also converges as x tends to infinity. Thus we may write

$$\alpha = \frac{e^{\frac{\lambda}{\mu}}\left(\frac{\lambda}{\mu}\right)^{\frac{\theta}{\mu}-1}}{\Gamma\left(\frac{\theta}{\mu}\right)}\sum_{k=1}^{\infty}\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!A_k(\theta)A_{k-1}(\theta)}.$$

This leads to

$$B_x(\theta) = \frac{e^{\frac{\lambda}{\mu}}\left(\frac{\lambda}{\mu}\right)^{\frac{\theta}{\mu}-1}}{\Gamma\left(\frac{\theta}{\mu}\right)}A_x(\theta)\sum_{k=x+1}^{\infty}\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!A_k(\theta)A_{k-1}(\theta)},$$

which allows us to estimate $B_x(\theta)$.

C Proof of Proposition 2 and computation of the remaining coefficients

Proof. The homogeneous equation in z associated with Equation (6) is $-(\lambda+\theta+(s-y)\mu_1+y\mu_2)z+\lambda+(s-y)\mu_1 z^2=0$. For $y<s$, the solutions are

$$z_{a,y}=\frac{\lambda+(s-y)\mu_1+y\mu_2+\theta-\sqrt{(\lambda+(s-y)\mu_1+y\mu_2+\theta)^2-4\lambda\mu_1(s-y)}}{2(s-y)\mu_1}, \text{ and}$$
$$z_{b,y}=\frac{\lambda+(s-y)\mu_1+y\mu_2+\theta+\sqrt{(\lambda+(s-y)\mu_1+y\mu_2+\theta)^2-4\lambda\mu_1(s-y)}}{2(s-y)\mu_1}.$$

If $y=s$, the homogeneous equation associated with Equation (7) is different. We instead have

$$-(\lambda+s\mu_2+\theta)z+\lambda=0, \tag{10}$$

with solution $\frac{\lambda}{\lambda+s\mu_2+\theta}=\lim_{y\to s} z_{a,y}$. Note that $\lim_{y\to s} z_{b,y}=\infty$. Therefore, we can extend the definition of $z_{a,y}$ to $y=s$.

We then may write

$$q^\theta_{x,y}=\sum_{k=y}^{s} c_{k,y} z_{a,k}^{x-s} \text{ for } y>y_0 \text{ and } 0\le y\le y_0, x>x_0-(y_0-y) \text{ and}$$
$$q^\theta_{x,y}=\sum_{k=y,y\neq y_0}^{s} c_{k,y} z_{a,k}^{x-s}+d_{y_0,y} z_{a,y_0}^{x-s}+e_{y_0,y} z_{b,y_0}^{x-s} \text{ for } 0\le y\le y_0, x\le x_0-(y_0-y),$$

since $0<z_{a,y}<1$ and $z_{b,y}>1$. Therefore, for $y_0+1\le k\le s$ or $0\le y\le y_0, x>x_0-(y_0-y)$, we have

$$c_{k,y}=c_{k,y+1}\frac{\mu_2(y+1)z_{a,k}^2}{(\lambda+\theta+y\mu_2+(s-y)\mu_1)z_{a,k}-\lambda-(s-y)\mu_1 z_{a,k}^2}=c_{k,y+1}\frac{\mu_2(y+1)z_{a,k}}{(k-y)(\mu_1(1-z_{a,k})-\mu_2)}$$
$$=c_{k,y+1}\alpha_k\frac{y+1}{k-y},$$

where $\alpha_k=\frac{\mu_2 z_{a,k}}{\mu_1(1-z_{a,k})-\mu_2}$. This relation leads to

$$q^\theta_{x,y}=\sum_{k=y}^{s}\binom{k}{y}\alpha_k^{k-y}c_{k,k}z_{a,k}^{x-s},$$

for $y_0<y\le s$ or $0\le y\le y_0$ and $x>x_0-y_0+y$. Finally, for $0\le y\le y_0$ and $x\le x_0-(y_0-y)$, we obtain

in a similar way

$$q^{\theta}_{x,y} = \sum_{k=y,k\neq y_0}^{s} \binom{k}{y} \alpha_k^{k-y} c_{k,k} z_{a,k}^{x-s} + \binom{y_0}{y} \alpha_{y_0}^{y_0-y} d_{y_0,y_0} z_{a,y_0}^{x-s} + \binom{y_0}{y} \beta_{y_0}^{y_0-y} e_{y_0,y_0} z_{b,y_0}^{x-s},$$

where $\beta_{y_0} = \frac{\mu_2 z_{b,y_0}}{\mu_1(1-z_{b,y_0})-\mu_2}$. □

Computation of the remaining terms. There remains to determine the constants $c_{0,0}$, $c_{1,1}$, ..., $c_{s,s}$, d_{y_0,y_0}, and e_{y_0,y_0}. Using the initial condition, we obtain

$$\sum_{k=y_0+1}^{s} c_{k,y_0} z_{a,k}^{x_0-s} + d_{y_0,y_0} z_{a,y_0}^{x_0-s} + e_{y_0,y_0} z_{b,y_0}^{x_0-s} = \sum_{k=y_0+1}^{s} c_{k,y_0} z_{a,k}^{x_0-s} + c_{y_0,y_0} z_{a,y_0}^{x_0-s}, \text{ and}$$

$$\lambda \left(\sum_{k=y_0+1}^{s} c_{k,y_0} z_{a,k}^{x_0-1-s} + d_{y_0,y_0} z_{a,y_0}^{x_0-1-s} + e_{y_0,y_0} z_{b,y_0}^{x_0-1-s} \right)$$

$$+ (s-y_0)\mu_1 \left(\sum_{k=y_0+1}^{s} c_{k,y_0} z_{a,k}^{x_0+1-s} + c_{y_0,y_0} z_{a,y_0}^{x_0+1-s} \right)$$

$$+ (y_0+1)\mu_2 \sum_{k=y_0+1}^{s} c_{k,y_0+1} z_{a,k}^{x_0+1-s} - (\lambda + (s-y_0)\mu_1 + y_0\mu_2 + \theta) \left(\sum_{k=y_0+1}^{s} c_{k,y_0} z_{a,k}^{x_0-s} + c_{y_0,y_0} z_{a,y_0}^{x_0-s} \right) = -1.$$

After simplification, these two equations allow us to express d_{y_0,y_0} and e_{y_0,y_0} in c_{y_0,y_0}. We obtain

$$d_{y_0,y_0} = -\frac{z_{b,y_0}}{z_{b,y_0} - z_{a,y_0}} \frac{1}{\lambda z_{a,y_0}^{x_0-s-1}} + c_{y_0,y_0} \left(1 + \frac{z_{a,y_0}}{z_{b,y_0} - z_{a,y_0}} \left(\frac{z_{b,y_0}}{z_{a,y_0}} \right)^{x_0-s} \right), \text{ and}$$

$$e_{y_0,y_0} = \frac{z_{a,y_0}}{z_{b,y_0} - z_{a,y_0}} \left(\frac{1}{\lambda z_{b,y_0}^{x_0-s-1}} - c_{y_0,y_0} \right).$$

Next, using the boundary conditions at $x = s$, we can compute $c_{y,y}$ in terms of $c_{s,s}$. For instance, line $y = s$ leads to

$$(\lambda + \theta) c_{s,s} = \mu_1 (c_{s-1,s-1} + s\alpha_s c_{s,s}).$$

We thus obtain

$$c_{s-1,s-1} = c_{s,s} \frac{\lambda + \theta - s\alpha_s \mu_1}{\mu_1}.$$

We can compute the $c_{y,y}$'s for $y_0+2 \le y \le s-1$ using line y, we get

$$\begin{aligned}
c_{y-1,y-1} = \frac{1}{(s-(y-1))\mu_1} \sum_{k=y+1}^{s} \alpha_k^{k-(y+1)} c_{k,k} \Bigg(& -\mu_1(s-(y-1))\alpha_k^2 \binom{k}{y-1} \\
& + \alpha_k(\lambda+\theta+(s-y)\mu_1(1-z_{a,k}))\binom{k}{y} - \mu_2(y+1)z_{a,k}\binom{k}{y+1}\Bigg) \\
+ \frac{1}{(s-(y-1))\mu_1} c_{y,y} & \left(-\mu_1 y(s-(y-1))\alpha_y + (\lambda+\theta+(s-y)\mu_1(1-z_{a,y}))\right).
\end{aligned}$$

For $y \le y_0$, we proceed in the same way but the relations are more involved. We obtain for $1 \le y \le y_0+1$,

$$\begin{aligned}
c_{y-1,y-1} = & \frac{1}{(s-(y-1))\mu_1} \sum_{k=y+1, k\neq y_0}^{s} \alpha_k^{k-(y+1)} c_{k,k} \Bigg(-\mu_1(s-(y-1))\alpha_k^2 \binom{k}{y-1} \\
& + \alpha_k(\lambda+\theta+(s-y)\mu_1(1-z_{a,k}))\binom{k}{y} - \mu_2(y+1)z_{a,k}\binom{k}{y+1}\Bigg) \\
& + \frac{1}{(s-(y-1))\mu_1} c_{y,y} \mathbb{1}_{y\neq y_0} \left(-\mu_1 y(s-(y-1))\alpha_y + (\lambda+\theta+(s-y)\mu_1(1-z_{a,y}))\right) \\
& + \frac{\mathbb{1}_{y_0>y}}{(s-(y-1))\mu_1} \alpha_{y_0}^{y_0-(y+1)} d_{y_0,y_0} \Bigg(-\mu_1(s-(y-1))\alpha_{y_0}^2 \binom{y_0}{y-1} \\
& + \alpha_{y_0}(\lambda+\theta+(s-y)\mu_1(1-z_{a,y_0}))\binom{y_0}{y} - \mu_2(y+1)z_{a,y_0}\binom{y_0}{y+1}\Bigg) \\
& + \frac{\mathbb{1}_{y_0>y}}{(s-(y-1))\mu_1} \beta_{y_0}^{y_0-(y+1)} e_{y_0,y_0} \Bigg(-\mu_1(s-(y-1))\beta_{y_0}^2 \binom{y_0}{y-1} \\
& + \beta_{y_0}(\lambda+\theta+(s-y)\mu_1(1-z_{b,y_0}))\binom{y_0}{y} - \mu_2(y+1)z_{b,y_0}\binom{y_0}{y+1}\Bigg) \\
& + \frac{1}{(s-(y_0-1))\mu_1} d_{y_0,y_0} \mathbb{1}_{y=y_0} \left(-\mu_1 y(s-(y_0-1))\alpha_{y_0} + (\lambda+\theta+(s-y_0)\mu_1(1-z_{a,y_0}))\right) \\
& + \frac{1}{(s-(y_0-1))\mu_1} e_{y_0,y_0} \mathbb{1}_{y=y_0} \left(-\mu_1 y(s-(y_0-1))\beta_{y_0} + (\lambda+\theta+(s-y_0)\mu_1(1-z_{b,y_0}))\right).
\end{aligned}$$

With line $y=1$, we express $c_{0,0}$ as a function of $c_{s,s}$. Finally, line $y=0$ leads to $c_{s,s}$.

D Proof of Theorem 1

Proof. We want to prove that $E(Q)_t$ is increasing in λ. To this end, we introduce the notation $Q_x^t = \sum_{k=s-c}^{x} p_k^t$ for $s-c \leq x \leq s+N$. We then rewrite (9) in terms of Q_x^t as follows:

$$
\begin{aligned}
&Q_{s-c}^{t+\epsilon} = (1-(\lambda+(s-c+1)\mu)\epsilon)Q_{s-c}^t + (s-c+1)\mu\epsilon Q_{s-c+1}^t, \qquad (11)\\
&Q_{s-c+x}^{t+\epsilon} = (1-(\lambda+(s-c+x+1)\mu)\epsilon)Q_{s-c+x}^t + \lambda\epsilon Q_{s-c+x-1}^t + (s-c+x+1)\mu\epsilon Q_{s-c+x+1}^t\\
&\qquad \text{for } 1 \leq x \leq c-1,\\
&Q_{x+s}^{t+\epsilon} = (1-(\lambda+s\mu)\epsilon)Q_{x+s}^t + \lambda\epsilon Q_{s+x-1}^t + s\mu\epsilon Q_{s+x+1}^t \text{ for } 0 \leq x \leq N-1, \text{ and}\\
&Q_{s+N}^{t+\epsilon} = Q_{s+N}^t = 1.
\end{aligned}
$$

We now prove by induction on t that $\frac{\partial Q_x^t}{\partial \lambda} \leq 0$ for $s-c \leq x \leq s+N$. The initial condition is independent from λ. Therefore, we have $\frac{\partial Q_x^0}{\partial \lambda} = 0$ for $s-c \leq x \leq s+N$. We assume now that $\frac{\partial Q_x^t}{\partial \lambda} \leq 0$ and we prove that $\frac{\partial Q_x^{t+\epsilon}}{\partial \lambda} \leq 0$ for $s-c \leq x \leq s+N$. We obtain

$$
\begin{aligned}
&\frac{\partial Q_{s-c}^{t+\epsilon}}{\partial \lambda} = (1-(\lambda+(s-c+1)\mu)\epsilon)\frac{\partial Q_{s-c}^t}{\partial \lambda} + (s-c+1)\mu\epsilon\frac{\partial Q_{s-c+1}^t}{\partial \lambda} - \epsilon Q_{s-c}^t \leq 0, \qquad (12)\\
&\frac{\partial Q_{s-c+x}^{t+\epsilon}}{\partial \lambda} = (1-(\lambda+(s-c+x+1)\mu)\epsilon)\frac{\partial Q_{s-c+x}^t}{\partial \lambda} + \lambda\epsilon\frac{\partial Q_{s-c+x-1}^t}{\partial \lambda}\\
&\quad + (s-c+x+1)\mu\epsilon\frac{\partial Q_{s-c+x+1}^t}{\partial \lambda} - \epsilon(Q_{s-c+x}^t - Q_{s-c+x-1}^t) \leq 0, \text{ for } 1 \leq x \leq c-1,\\
&\frac{\partial Q_{x+s}^{t+\epsilon}}{\partial \lambda} = (1-(\lambda+s\mu)\epsilon)\frac{\partial Q_{x+s}^t}{\partial \lambda} + \lambda\epsilon\frac{\partial Q_{s+x-1}^t}{\partial \lambda} + s\mu\epsilon\frac{\partial Q_{s+x+1}^t}{\partial \lambda}\\
&\quad - \epsilon(Q_{x+s}^t - Q_{x-1+s}^t) \leq 0 \text{ for } 0 \leq x \leq N-1, \text{ and}\\
&\frac{\partial Q_{s+N}^{t+\epsilon}}{\partial \lambda} = \frac{\partial Q_{s+N}^t}{\partial \lambda} = 0,
\end{aligned}
$$

which proves the induction step. The expected number of customers in the queue at time t can be expressed as $E(Q)_t = \sum_{k=0}^{N} k p_{s+k}^t = N - \sum_{k=0}^{N-1} Q_{s+k}^t$. This proves that $\frac{\partial E(Q)_t}{\partial \lambda} \geq 0$. Therefore, the expected number of customers in the queue is increasing in λ. It should be noted that the same result holds for the expected number of customers in the system.

We are now interested in the impact of the reservation threshold on the expected number of customers in the queue. We thus compare the expected number of customers in the queue with reservation thresholds c and $c+1$ for $0 \leq c < s$. We write $p_x^{c,t}$ instead of p_x^t to specify the reservation threshold under consideration. The initial condition is such that $x_0 \geq s-c$, otherwise the two systems cannot be compared. We define

$Z_x^{c,t} = \sum_{k=x}^{s+N} p_k^{c,t}$ and $\Delta_x^t = Z_x^{c+1,t} - Z_x^{c,t}$. We prove by induction on t that $\Delta_x^t \leq 0$ for $s-c \leq x \leq s+N$. At $t=0$, we have $\Delta_x^0 = 0$ for $s-c \leq x \leq s+N$. We now show the induction step. From (9), we deduce that

$$
\begin{aligned}
&Z_{s-c}^{c,t+\epsilon} = Z_{s-c}^{c,t} = 1, \qquad (13)\\
&Z_{s-c+x}^{c,t+\epsilon} = (1-(\lambda+(s-c+x)\mu)\epsilon)Z_{s-c+x}^t + \lambda\epsilon Z_{s-c+x-1}^t + (s-c+x)\mu\epsilon Z_{s-c+x+1}^t \text{ for } 1 \leq x \leq c-1,\\
&Z_{x+s}^{c,t+\epsilon} = (1-(\lambda+s\mu)\epsilon)Z_{x+s}^{c,t} + \lambda\epsilon Z_{s+x-1}^{c,t} + s\mu\epsilon Z_{s+x+1}^{c,t} \text{ for } 0 \leq x \leq N-1, \text{ and}\\
&Z_{s+N}^{c,t+\epsilon} = (1-(\lambda+s\mu)\epsilon)Z_{s+N}^{c,t} + \lambda\epsilon Z_{s+N-1}^{c,t}.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
&\Delta_{s-c+x}^{t+\epsilon} = (1-(\lambda+(s-c+x)\mu)\epsilon)\Delta_x^t + \lambda\epsilon\Delta_{x-1}^t + (s-c+x)\mu\epsilon\Delta_{x+1}^t \leq 0 \text{ for } 1 \leq x \leq c,\\
&\Delta_x^{t+\epsilon} = (1-(\lambda+s\mu)\epsilon)\Delta_x^t + \lambda\epsilon\Delta_{x-1}^t + s\mu\epsilon\Delta_{x+1}^t \leq 0 \text{ for } s \leq x \leq s+N-1, \text{ and}\\
&\Delta_{s+N}^{t+\epsilon} = (1-(\lambda+s\mu)\epsilon)\Delta_{s+N}^t + \lambda\epsilon\Delta_{s+N-1}^t \leq 0.
\end{aligned}
$$

Consider now the term at $x = s-c$. We have

$$
\begin{aligned}
&p_{s-c}^{c,t+\epsilon} = (1-\lambda\epsilon)p_{s-c}^{c,t} + (s-c+1)\mu\epsilon p_{s-c+1}^{c,t}, \text{ and}\\
&p_{s-c}^{c+1,t+\epsilon} = (1-(\lambda+(s-c)\mu)\epsilon)p_{s-c}^{c+1,t} + \lambda\epsilon p_{s-c-1}^{c+1,t} + (s-c+1)\mu\epsilon p_{s-c+1}^{c+1,t}.
\end{aligned}
$$

Therefore, we deduce that

$$
\Delta_{s-c}^{t+\epsilon} = (1-(\lambda+(s-c)\mu)\epsilon)\Delta_{s-c}^t + (s-c)\mu\epsilon\Delta_{s-c+1}^t + \lambda\epsilon\Delta_{s-c-1}^t - (s-c)\mu\epsilon p_{s-c}^{c,t} \leq 0.
$$

This proves the induction step. Therefore, $E(Q)_t = \sum_{k=s+1}^{s+N} Z_k^{c,t}$ is decreasing in c.

We now prove that $E(Q)_t$ is convex in c. We consider a reservation threshold c such that $0 \leq c \leq s-2$. We want to prove that $Z_x^{c+2,t} + Z_x^{c,t} - 2Z_x^{c+1,t} \geq 0$ for $t \geq 0$ and $s-c \leq x \leq s+N$ with an initial condition such that $x_0 \geq s-c$, otherwise the comparison cannot be made. At $t=0$ we have $Z_x^{c+2,0} + Z_x^{c,0} - 2Z_x^{c+1,0} = 0$.

Next, we show the induction with

$$
\begin{aligned}
&Z_{s+N}^{c+2,t+\epsilon} + Z_{s+N}^{c,t+\epsilon} - 2Z_{s+N}^{c+1,t+\epsilon} = (1-(\lambda+s\mu)\epsilon)\left(Z_{s+N}^{c+2,t+\epsilon} + Z_{s+N}^{c,t+\epsilon} - 2Z_{s+N}^{c+1,t+\epsilon}\right) \\
&\qquad + \lambda\epsilon(Z_{s+N-1}^{c+2,t+\epsilon} + Z_{s+N-1}^{c,t+\epsilon} - 2Z_{s+N-1}^{c+1,t+\epsilon}) \geq 0, \\
&Z_{x}^{c+2,t+\epsilon} + Z_{x}^{c,t+\epsilon} - 2Z_{x}^{c+1,t+\epsilon} = (1-(\lambda+s\mu)\epsilon)(Z_{x}^{c+2,t} + Z_{x}^{c,t} - 2Z_{x}^{c+1,t}) \\
&\qquad + \lambda\epsilon(Z_{x-1}^{c+2,t} + Z_{x-1}^{c,t} - 2Z_{x-1}^{c+1,t}) + s\mu\epsilon(Z_{x+1}^{c+2,t} + Z_{x+1}^{c,t} - 2Z_{x+1}^{c+1,t}) \geq 0 \text{ for } s \leq x \leq s+N-1, \\
&Z_{s-c+x}^{c+2,t+\epsilon} + Z_{s-c+x}^{c,t+\epsilon} - 2Z_{s-c+x}^{c+1,t+\epsilon} = (1-(\lambda+(s-c+x)\mu)\epsilon)(Z_{s-c+x}^{c+2,t} + Z_{s-c+x}^{c,t} - 2Z_{s-c+x}^{c+1,t}) \\
&\qquad + \lambda\epsilon(Z_{s-c+x-1}^{c+2,t} + Z_{s-c+x-1}^{c,t} - 2Z_{s-c+x-1}^{c+1,t}) + (s-c+x)\mu\epsilon(Z_{s-c+x+1}^{c+2,t} + Z_{s-c+x+1}^{c,t} - 2Z_{s-c+x+1}^{c+1,t}) \geq 0 \\
&\qquad \text{for } 1 \leq x \leq c-1.
\end{aligned}
$$

We now consider the term $x = s - c$. We have

$$
\begin{aligned}
p_{s-c}^{c+2,t+\epsilon} &= (1-(\lambda+(s-c)\mu)\epsilon)p_{s-c}^{c+2,t} + \lambda\epsilon p_{s-c-1}^{c+2,t} + (s-c+1)\mu\epsilon p_{s-c+1}^{c+2,t}, \\
p_{s-c}^{c,t+\epsilon} &= (1-\lambda\epsilon)p_{s-c}^{c,t} + (s-c+1)\mu\epsilon p_{s-c+1}^{c,t}, \text{ and} \\
p_{s-c}^{c+1,t+\epsilon} &= (1-(\lambda+(s-c)\mu)\epsilon)p_{s-c}^{c+1,t} + \lambda\epsilon p_{s-c-1}^{c+1,t} + (s-c+1)\mu\epsilon p_{s-c+1}^{c+1,t}.
\end{aligned}
$$

We then deduce that

$$
\begin{aligned}
Z_{s-c}^{c+2,t+\epsilon} + Z_{s-c}^{c,t+\epsilon} - 2Z_{s-c}^{c+1,t+\epsilon} &= (1-(\lambda+(s-c)\mu)\epsilon)(Z_{s-c}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c}^{c+1,t}) + (s-c)\mu\epsilon Z_{s-c}^{c,t} \\
&\quad + \lambda\epsilon(Z_{s-c-1}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c-1}^{c+1,t}) + (s-c)\mu\epsilon(Z_{s-c+1}^{c+2,t} + Z_{s-c+1}^{c,t} - 2Z_{s-c+1}^{c+1,t}).
\end{aligned}
$$

We have $(s-c)\mu\epsilon(Z_{s-c+1}^{c+2,t} + Z_{s-c+1}^{c,t} - 2Z_{s-c+1}^{c+1,t}) \geq 0$. Moreover, $Z_{s-c}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c}^{c+1,t} = 2p_{s-c-1}^{c+1,t} - p_{s-c-1}^{t,c+2} - p_{s-c-2}^{c+2,t}$, $Z_{s-c}^{c,t} = 1$, and $Z_{s-c-1}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c-1}^{c+1,t} = -p_{s-c-2}^{c+2,t}$. Therefore, we may write

$$
\begin{aligned}
&(1-(\lambda+(s-c)\mu)\epsilon)(Z_{s-c}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c}^{c+1,t}) + (s-c)\mu\epsilon Z_{s-c}^{c,t} + \lambda\epsilon(Z_{s-c-1}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c-1}^{c+1,t}) \\
&\geq (1-(2\lambda+(s-c)\mu)\epsilon)(2p_{s-c-1}^{c+1,t} - p_{s-c-1}^{t,c+2} - p_{s-c-2}^{c+2,t}) + \lambda\epsilon(2p_{s-c-1}^{c+1,t} - p_{s-c-1}^{t,c+2}) \\
&\geq (1-(2\lambda+(s-c)\mu)\epsilon)(Z_{s-c}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c}^{c+1,t}) + \lambda\epsilon(Z_{s-c}^{c+2,t} + Z_{s-c}^{c,t} - 2Z_{s-c}^{c+1,t}) \geq 0.
\end{aligned}
$$

This proves that $Z_{s-c}^{c+2,t+\epsilon} + Z_{s-c}^{c,t+\epsilon} - 2Z_{s-c}^{c+1,t+\epsilon} \geq 0$. The expression of $E(Q)_t$ indicates that $E(Q)_t$ has the same convexity property as the Z_x^t's.

The proof of the monotonicity property of Z_x^t in x_0 is direct from (13) by comparing two systems with

initial conditions x_0 and $x_0 + 1$. □

E The expected number of customers in the queue $E(Q)_t$ is not convex in x_0

We write $E(Q)_t^{x_0}$ instead of $E(Q)_t$ to specify the initial state x_0 for the expected number of customers in the queue at time $t \geq 0$. From the proof of Theorem 1, we may relate $E(Q)_{t+\epsilon}^{x_0}$ and $E(Q)_t^{x_0}$ via $E(Q)_{t+\epsilon}^{x_0} = E(Q)_t^{x_0} + \lambda\epsilon(Z_s^{c,t} - Z_{s+N}^{c,t}) - s\mu\epsilon Z_{s+1}^{c,t}$. At time $t = 0$, we have $E(Q)_0^{x_0} = \max(x_0 - s, 0)$ and $Z_k^{c,0} = \mathbb{1}_{k \geq x_0}$. Therefore, for the first iteration we have

$$
\begin{aligned}
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = 0 \text{ for } s - c \leq x_0 < s - 2,\\
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = \lambda\epsilon > 0 \text{ for } x_0 = s - 2,\\
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = 1 - (\lambda + s\mu)\epsilon > 0 \text{ for } x_0 = s - 1,\\
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = s\mu\epsilon > 0 \text{ for } x_0 = s,\\
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = 0 \text{ for } s < x_0 < s + N - 2, \text{ and}\\
&E(Q)_\epsilon^{x_0+2} + E(Q)_\epsilon^{x_0} - 2E(Q)_\epsilon^{x_0+1} = -\lambda\epsilon < 0 \text{ for } x_0 = s + N - 2,
\end{aligned}
$$

which shows that the convexity in x_0 does not hold due to the boundary state.