HAL
open science

A discrete-time single-server queue with balking: economic applications

Macarena Lozano, Pilar Moreno

To cite this version:

Macarena Lozano, Pilar Moreno. A discrete-time single-server queue with balking: economic applications. Applied Economics, 2008, 40 (06), pp.735-748. 10.1080/00036840600749607 . hal-00581999

HAL Id: hal-00581999

https://hal.science/hal-00581999v1

Submitted on 1 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

For Peer Review

A discrete-time single-server queue with balking: economic applications

Journal:	*Applied Economics*
Manuscript ID:	APE-05-0539.R1
Journal Selection:	Applied Economics
JEL Code:	C60 - General < C6 - Mathematical Methods and Programming < C - Mathematical and Quantitative Methods, C69 - Other < C6 - Mathematical Methods and Programming < C - Mathematical and Quantitative Methods, C65 - Miscellaneous Mathematical Tools < C6 - Mathematical Methods and Programming < C - Mathematical and Quantitative Methods
Keywords:	Balking, Discrete-time queue, Optimal control, Markov chain, Practical applications

powered by ScholarOne
Manuscript Central™

Submitted Manuscript

A discrete-time single-server queue with balking: economic applications

Macarena Lozano Pilar Moreno*

Departamento de Economía, Métodos Cuantitativos
e Historia Económica
Facultad de Ciencias Empresariales
Universidad Pablo de Olavide
Ctra. de Utrera, km. 1, 41013 Sevilla, Spain
{mlozoyo,mpmornav}@upo.es

For Peer Review

Abstract

This paper studies a discrete-time single-server queue with finite and infinite buffer where the users have the option to leave the queue upon arrival (balking). We consider two variants of the model in accordance with the balking policies. Firstly, all the arriving customers balk with a constant probability. Secondly, arriving customers increase their balking probabilities as more customers join the system. Specifically, we find the ergodicity condition and closed form expressions for the stationary distribution of the system size, of the waiting/spending time in the FCFS system and of the unfinished work. The mathematical model is applied in order to resolve several real-life problems in the economic field; in this sense, practical applications in the secondary and tertiary sector are shown. We also develop a cost model to determine the buffer capacity that minimizes certain cost function and give some numerical examples.

Keywords: Balking; Discrete-time queue; Ergodicity; Markov chain; Optimal control; Practical applications; Unfinished work; Waiting/spending time.

1 Introduction

Recently, different continuous-time queueing systems have been applied to a wide class of economic activities, including dynamic market process [22], labor market [19], holdups in markets with frictions [2], manufacturing and computer systems [17], multiproduct monopoly [21], hospitals [24], manufacturing context [26], job shop environment [12] and repair shops [23]. However, the discrete-time systems are more appropriate than their continuous-time counterparts for

*Her work is supported by the MEC through the project MTM2005-01248.

Editorial Office, Dept of Economics, Warwick University, Coventry CV4 7AL, UK

modelling diverse productive processes, since the basic units in these systems are digital such as machine cycle time. The analysis of discrete-time queueing models has received considerable attention in the scientific literature over the past years, in view of its applicability in the study of many computer and communication systems in which time is slotted, see [7, 14, 25, 28] and the references therein. Another important application stems from the secondary and tertiary sector since, for example, the current production systems of numerous factories operate on a discrete-time basis where events can only happen at regularly spaced epochs. Nonetheless, despite the wide applications in the aforementioned areas, no work seems to have been done concerning these last applications. That is why this paper makes efforts to fill this space and proposes a new approach modelling some economic activities in discrete-time.

On the other hand, in many queueing situations an arriving customer may balk, so there exists literature devoted to the design and applications of such models [1, 3, 15, 16, 18, 20, 27, 29]. Intuitively, the introduction of the balking assumption makes the system less congested than if it was not present, thus the existence of balking provides a mechanism to control an excessive congestion at the system. There exist another queueing models where the customers can leave the system before completing his service, for instance, queues with negative customers [4, 5], disasters [4] or impatient customers [8]. Unlike what happens in our model, in the case of negative arrivals or disasters, the abandonments are generated by external causes. Our system could be thought as a related model with impatient customers; however, the balking supposition involves an automatic system abandonment upon arrival at the system, whereas impatience means to take the abandonment decision after some random time.

Several continuous-time queueing models with balking have been discussed during the last years. Nevertheless, to the best of our knowledge, the only work about balking in discrete-time can be found in [16], where the authors regard a discrete-time multi-server queue with balking and reneging under arbitrary initial conditions and finite waiting space. This work studies a discrete-time single-server queue with finite and infinite buffer, in which arriving customers balk with a certain probability: in a case constant and in the another one depending on the number of customers in the system upon arrival. Moreover, certain economic activities can be modelled by this queueing system; specifically, two examples are shown in the secondary and tertiary sector. Besides, a cost model is developed to determine the optimal buffer size in order to minimize the steady-state expected cost per unit time.

The rest of the paper is organized as follows. The mathematical model is described in the next section. Section 3 provides some examples of real systems which can be modelled by our queue. Section 4 investigates the stationary distribution and the stability condition of the model under study. Waiting/spending time in the FCFS system is analysed in Section 5. Section 6 shows the stationary distribution of the unfinished work. Section 7 develops a cost model to determine the optimum buffer capacity. Finally, in Section 8, some numerical results are presented to illustrate the effect of the parameters on several performance characteristics.

For Peer Review

Submitted Manuscript

2 The mathematical model

We consider a discrete-time queueing system where the time axis is divided into intervals of equal length, called slots. It is assumed that all queueing activities (arrivals and departures) occur at the slot boundaries, and therefore they may occur at the same time. For mathematical clarity, we will suppose that the departures occur at the moment immediately before the slot boundaries and the arrivals occur at the moment immediately after the slot boundaries; that is, we will discuss the model only for the *early arrival system policy* (details on this and related concepts can be found in [11, 14]).

We assume that the system conforms to the following assumptions:

- Customers arrive according to a Bernoulli process with probability p, thus p is the probability that a customer arrives at a slot and $\bar{p} = 1 - p$ is the probability that an arrival does not take place in a slot.
- If the system is free at the instant of an arrival, the service of the arriving customer commences immediately and the customer leaves the system after service completion.
- If the arrival finds the system busy, the arriving customer either with a probability r_k, when the system size is k, joins the waiting line in order to be served (persistent customer), or with complementary probability abandons the system forever without service (non-persistent customer).
- The balking probability is $1 - r_k$ if there are k customers in the system upon arrival.
- We will study two cases separately: the case of a fixed $r_k = r$ (constant rate policy) and the case of $r_k = r^k$ (discouraged rate policy) where $0 \leq r \leq 1$.
- It is always assumed that services can begin only at slot boundaries and their durations are integer multiples of a slot duration.
- Service times are independent and geometrically distributed with probability $\bar{s} = 1 - s$, where s is the probability that a customer does not conclude his service in a slot.
- It is supposed that the buffer capacity is N. We consider two cases singly: an infinite buffer $N = \infty$ and a finite buffer $2 \leq N < \infty$. For a finite buffer, an arriving customer who finds the system completely full is lost.

It appears reasonable to assume, in certain situations, that the balking probability is constant and is not dependent on the system length; such situations may arise when the system length is not observable by the customer and he has no knowledge of the system length. It is also logical to suppose that the balking probability increases as more customers enter the system; in fact, if the system size increases, then the waiting time is higher and the arrivals are discouraged to wait in order to receive their services. It is obvious that we introduce the

For Peer Review

balking probabilities as a mechanism to control the level of internal congestion in the system. It should be pointed out that the discouraged rate policy is more intelligent and rational than the constant one, since the first discipline has more information concerning the queueing system.

In Figures 1 and 2, the system is depicted for an infinite and finite buffer, respectively.

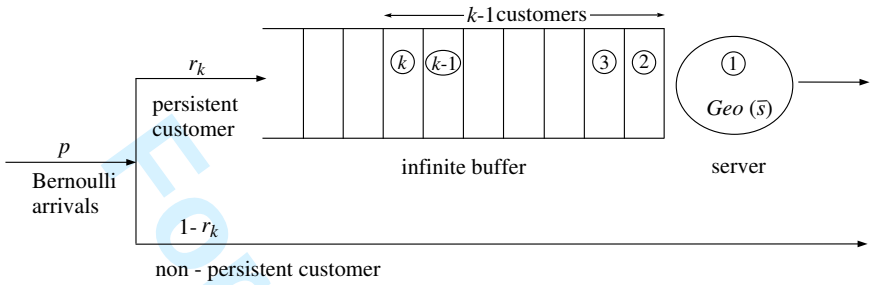


Figure 1: Mathematical model for the queueing system with infinite buffer.

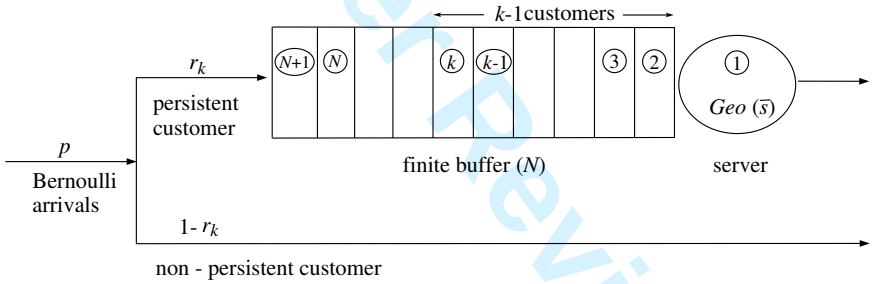


Figure 2: Mathematical model for the queueing system with finite buffer.

In order to avoid trivial cases, we assume $0 < p < 1$ and $0 < s < 1$. The traffic intensity is given by $\rho = p/\bar{s}$.

At time m^+ (the instant immediately after the m-th slot), the system can be described by the random variable X_m, which designates the number of customers in the system (including the one in service, if any).

To conclude this section it should be pointed out that if $r = 1$, arriving customers always decide to enter the system in order to receive their service, hence we obtain a well-known special case: the classical discrete-time single-server queue with geometrical arrivals and service times. In the particular case

For Peer Review

Submitted Manuscript

$r = 0$, arriving customers only enter the system when their waiting time in the buffer is zero, i.e., when the system is totally empty (it is an extreme case of impatience).

3 Practical applications

This section gives two examples of productive processes which can be modelled as a discrete-time single-server queue with balking.

3.1 Secondary sector

This subsection applies the model under consideration to the industrial sector. The production system operates in the following way. It is considered a manufacturing plant in which the production process is divided into several phases. We will concentrate our attention on a specific stage, as appears in Figure 3. Customers are components or products that pass by different jobs on consecutive machines.

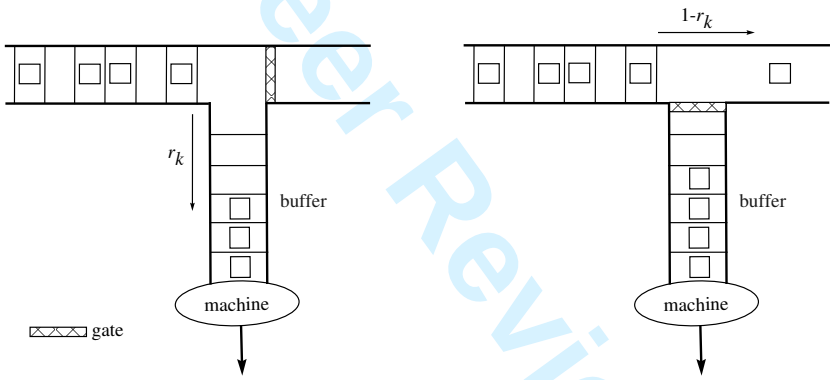


Figure 3: Industry.

Customers are transported in a conveyor belt with regularly spaced partitions (which, for example, move every thirty seconds). Each division of the conveyor belt contains a component/product with probability p and consequently the arrivals follow a Bernoulli process. In addition, a buffer is opposite the machine (server) in a lower level, which impedes empty positions in the waiting line (see Figure 3).

The components are processed on a first-come-first-served (FCFS) basis. If the machine is empty at the moment of an arrival, the service begins immediately

For Peer Review

and the corresponding component/product leaves the system after concluding his service. When the system is busy, customers with a probability r_k, if the system size is k, join the buffer (see the left-hand side of Figure 3). On the other hand, if the queue in front of the machine is "too occupied" (right-hand side of Figure 3), a gate prevents the passage to the waiting line; in this case, the customers with a probability $1 - r_k$ abandon the studied system forever without receiving service (non-persistent customers) and pass to another phase of the productive process.

Components/products may be distinct and consequently manufactured at different times, that is, service times are not deterministic. For instance, a service finishes with probability $\bar{s}$ each half minute (if the duration of a slot is thirty seconds).

3.2 Tertiary sector

The mathematical model described in section 2 allows to analyse problems that take place in activities of the services sector. In this subsection the previous queueing system has been applied to a particular attraction of an amusement park, as you can see in the enlargement of Figure 4.

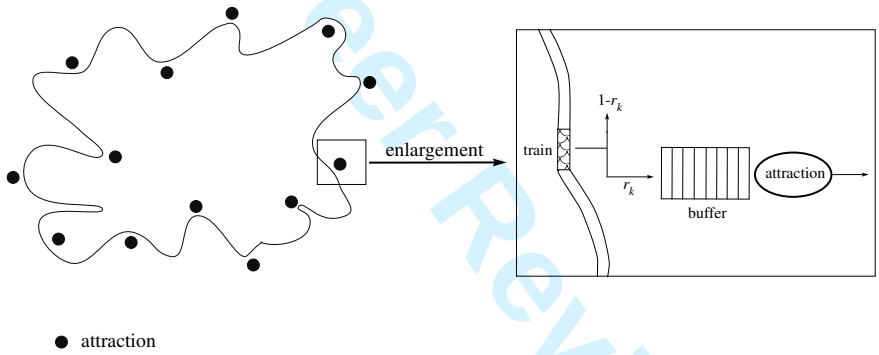


Figure 4: Amusement park.

A train with as many carriages as attractions covers all the amusement park so every five minutes, for example, an only carriage stops at each of the attractions. The group of people who get off the carriage in each attraction (for instance, one or twenty persons) will be called a visitor. In the attraction under study, a carriage arrives empty or busy each five minutes (the duration of a slot) with probabilities $\bar{p}$ and p, respectively. When the attraction is unoccupied, the visitor will join the attraction immediately. Otherwise, based on the characteristics of the queue, the visitor will choose either to stay in the carriage or to

For Peer Review

alight in order to enter the attraction. That is, when the attraction is occupied, there are two possible situations.

(a) If the visitor has knowledge of the system length, he will join or leave the attraction (system) depending on the waiting time. In this case, the balking probability increases as more visitors enter the system, that is, it is an increasing function of the system size (discouraged rate policy). Two different possibilities are assumed: a finite queue if there is a limited space, or an infinite queue when the extension is so large that it may be considered infinite.

(b) If the visitor does not know the system size, the balking probability will be fixed (constant rate policy). For example, when there is an opaque fence with a gate and some instructions about the attraction. Like in the previous assumption, the buffer can be regarded finite or infinite.

We assume that visitors arrive at the attraction through a single waiting line and are served in order of arrival (FCFS). An only visitor can be served in each attraction and his service time is a multiple of five minutes. Every five minutes, the attraction concludes with probability $\bar{s}$. Once the visitor is served leaves the attraction (system).

4 System size

This section studies the stationary distribution of the system size and its ergodicity condition, and it also links our discrete-time model to the continuous-time one.

It can be readily shown that $\{X_m\,,\ m \in \mathbb{N}\}$ is the one-dimensional Markov chain of our queueing system, whose state space is $\{0, 1, 2, \ldots, N+1\}$. One of the main objectives is to find the stationary distribution

$$\pi_k = \lim_{m\to\infty} P[X_m = k]\,,\ k = 0, \ldots, N+1$$

of the Markov chain $\{X_m\,,\ m \in \mathbb{N}\}$. The state space and one-step transitions are illustrated in Figure 5.

The Kolmogorov equations for the distribution π_k are

$$\begin{aligned}
\pi_0 &= \bar{p}\,\pi_0 + \bar{s}\,\bar{p}\,\pi_1 \\
\pi_1 &= p\,\pi_0 + [\bar{s}\,p + s\,\bar{p} + s\,p\,(1-r_1)]\,\pi_1 + [\bar{s}\,\bar{p} + \bar{s}\,p\,(1-r_1)]\,\pi_2 \\
\pi_k &= s\,p\,r_{k-1}\,\pi_{k-1} + [\bar{s}\,p\,r_{k-1} + s\,\bar{p} + s\,p\,(1-r_k)]\,\pi_k + \\
&\quad + [\bar{s}\,\bar{p} + \bar{s}\,p\,(1-r_k)]\,\pi_{k+1}\,,\ k = 2, \ldots, N \\
\pi_{N+1} &= s\,p\,r_N\,\pi_N + (\bar{s}\,p\,r_N + s)\,\pi_{N+1} \text{ if } N < \infty
\end{aligned}$$

and the normalization condition is $\sum_{k=0}^{N+1} \pi_k = 1$. The solution of these equilibrium equations is presented in the following theorems, whose proofs are easily obtained proceeding by recurrence.

For Peer Review

Theorem 1 (Constant rate discipline) *The Markov chain* $\{X_m\,,\ m \in \mathbb{N}\}$ *is ergodic if and only if* $\rho\, r\, \delta_{N,\infty} < 1$ *and its stationary distribution is given by the formulae:*

$$\begin{aligned}
\pi_0 &= \frac{\bar{p}\,(1-\rho\, r)\,(1-p\, r)^N}{[\bar{p}+\rho\,(1-r)]\,(1-p\, r)^N - \rho\,(\rho\, s\, r)^{N+1}} \\
\pi_k &= \frac{\rho^k\,(s\, r)^{k-1}\,(1-\rho\, r)\,(1-p\, r)^{N-k+1}}{[\bar{p}+\rho\,(1-r)]\,(1-p\, r)^N - \rho\,(\rho\, s\, r)^{N+1}}\,,\ k = 1,\ldots,N+1\,.
\end{aligned}$$

Theorem 2 (Discouraged rate discipline) *The Markov chain* $\{X_m\,,\ m \in \mathbb{N}\}$ *is ergodic if and only if* $\rho\, \delta_{r,1}\, \delta_{N,\infty} < 1$ *and its stationary distribution is given by the formulae:*

$$\begin{aligned}
\pi_0 &= \left\{1 + \sum_{k=1}^{N+1} \frac{\rho^k\, s^{k-1}\, r^{\frac{(k-1)\,k}{2}}}{\prod_{j=0}^{k-1}(1-p\, r^j)}\right\}^{-1} \\
\pi_k &= \frac{\rho^k\, s^{k-1}\, r^{\frac{(k-1)\,k}{2}}}{\prod_{j=0}^{k-1}(1-p\, r^j)} \left\{1 + \sum_{l=1}^{N+1} \frac{\rho^l\, s^{l-1}\, r^{\frac{(l-1)\,l}{2}}}{\prod_{i=0}^{l-1}(1-p\, r^i)}\right\}^{-1}\,,\ k = 1,\ldots,N+1\,.
\end{aligned}$$

Let us remark that if $r < 1$, in case of infinite buffer and discouraged rate, the system is always ergodic independently of the parametric values. This fact was expected since as the system size increases, the effective rate of entrance diminishes, which constitutes a mechanism to impede an overloaded and unstable system.

We must emphasize the importance of finding the ergodicity hypothesis, which is related to the term "propagation of chaos" [6]. Although chaotic systems obey certain rules that can be described by mathematical equations, chaos theory shows the difficulty of predicting their long-range behaviour. This queueing system could be tackled in an alternative way using the chaos theory [6, 10, 13]; specifically, we should follow the two crucial steps given by Borovkov [6]. Of course, when the parametric values tend to the ergodicity condition, the system becomes unstable and tends to the disintegration. Furthermore, although the system is always ergodic under a finite buffer, it is seriously disturbed-unbalanced when the parametric values approach the ergodicity condition of the corresponding model under an infinite buffer (see the maxima of the graphics in Figure 8).

Remark 1 (Relation to the continuous-time system) *This remark analyses the relation between our discrete-time model and its continuous-time counterpart.*

We consider the continuous-time single-server balking queueing system where customers arrive according to a Poisson process with rate λ*. The balking operation is exactly the same than in discrete-time. Service times are independent and exponentially distributed with mean* μ^{-1}*. Interarrival and service times are assumed to be mutually independent.*

For Peer Review

If time is divided into intervals of length $\Delta \in \left(0, \min\left\{\frac{1}{\lambda}, \frac{1}{\mu}\right\}\right)$, *the previous continuous-time system can be approximated by our discrete-time model choosing the parameters as follows:* $p = \lambda\,\Delta$ *and* $s = 1 - \mu\,\Delta$.

- *For the constant rate policy, we have:*

$$\begin{aligned}
\lim_{\Delta\to 0} \pi_0 &= \frac{1 - \frac{\lambda}{\mu}\, r}{1 + \frac{\lambda}{\mu}\,(1-r) - \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\, r\right)^{N+1}} \\
\lim_{\Delta\to 0} \pi_k &= \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\, r\right)^{k-1} \lim_{\Delta\to 0} \pi_0\,, \;\; k = 1, \ldots, N+1\,.
\end{aligned}$$

- *For the discouraged rate policy, we obtain:*

$$\begin{aligned}
\lim_{\Delta\to 0} \pi_0 &= \left\{1 + \sum_{k=1}^{N+1} \left(\frac{\lambda}{\mu}\right)^k r^{\frac{(k-1)\,k}{2}}\right\}^{-1} \\
\lim_{\Delta\to 0} \pi_k &= \left(\frac{\lambda}{\mu}\right)^k r^{\frac{(k-1)\,k}{2}} \lim_{\Delta\to 0} \pi_0\,, \;\; k = 1, \ldots, N+1\,.
\end{aligned}$$

These stationary probabilities can be checked since this continuous-time system is a special case of a birth-death process with state-dependent arrivals and departures [9].

5 Waiting/spending time in the FCFS system

This section presents the stationary distribution of the waiting/spending time in the FCFS system of an arriving customer.

Firstly, we define the waiting time W (measured in slots) in the FCFS system in the $m+1$-th slot as the time that a customer would wait in the corresponding system with FCFS discipline if he arrived in the $m+1$-th slot.

Theorem 3 *The stationary distribution of the waiting time* W *in the FCFS system is given by:*

$$\begin{aligned}
P[W=0] &= \pi_0 + \bar{s}\,\pi_1 \\
P[W=j] &= \sum_{k=1}^{\min\{j,N+1\}} \binom{j-1}{k-1} \bar{s}^k\, s^{j-k+1}\, \pi_k + \\
&\quad + \sum_{k=2}^{\min\{j+1,N+1\}} \binom{j-1}{k-2} \bar{s}^k\, s^{j-k+1}\, \pi_k\,, \;\; j \geq 1\,.
\end{aligned}$$

Proof: The waiting time of a customer who arrives in the $m+1$-th slot will be zero if $X_m = 0$ or ($X_m = 1$ and the current service finishes in the $m+1$-th slot).

For Peer Review

Let us remember that the remaining service time follows a geometrical distribution with probability $\bar{s}$ due to the memoryless property of the geometrical law.

If $X_m = k \geq 1$ and the current service does not conclude in the $m+1$-th slot, the waiting time will be the sum of $k-1$ complete service times plus the remaining service time of the customer currently being served; then the waiting time will be the sum of k independent and geometrically distributed random variables with parameter $\bar{s}$.

If $X_m = k \geq 2$ and the current service ends in the $m+1$-th slot, the waiting time will be the sum of $k-1$ complete service times, that is, the waiting time will be the sum of $k-1$ independent geometrical distributions with parameter $\bar{s}$. □

We now define the spending time S (measured in slots) in the FCFS system in the $m+1$-th slot as the time that a customer would spend in the corresponding system with FCFS discipline if he arrived in the $m+1$-th slot. It is equal to the waiting time in the $m+1$-th slot plus the service time of the arriving customer in the $m+1$-th slot.

Theorem 4 *The stationary distribution of the spending time S in the FCFS system is given by:*

$$\begin{aligned} P[S=j] &= \bar{s}\, s^{j-1}\,(\pi_0 + \bar{s}\,\pi_1) + (1-\delta_{j,1}) \sum_{k=1}^{\min\{j-1,N+1\}} \binom{j-1}{k} \bar{s}^{k+1}\, s^{j-k}\, \pi_k + \\ &\quad +(1-\delta_{j,1}) \sum_{k=2}^{\min\{j,N+1\}} \binom{j-1}{k-1} \bar{s}^{k+1}\, s^{j-k}\, \pi_k\,, \ j \geq 1\,. \end{aligned}$$

The proof of this theorem follows the steps given in the proof of the previous theorem.

6 Unfinished work

This section finds the stationary distribution of the unfinished work, that is, the stationary distribution of the remaining number of slots needed to empty the system of all the customers present. Specifically, the unfinished work U (measured in slots) immediately after the $m+1$-th slot is the sum of the service times of all customers in the queue and the remaining service time of the customer being served at that time. It is equal to the waiting time in the $m+1$-th slot plus the service time of a possible arrival in the $m+1$-th slot.

For Peer Review

Theorem 5 *The stationary distribution of the unfinished work U is given by:*

$$
\begin{aligned}
P[U=0] &= \bar{p}\,(\pi_0 + \bar{s}\,\pi_1) \\
P[U=j] &= p\,\bar{s}\,s^{j-1}\,(\pi_0 + \bar{s}\,\pi_1) + \\
&\quad +(1-\delta_{j,1}) \sum_{k=1}^{\min\{j-1,N+1\}} (1-\delta_{k,N+1})\,p\,r_k \binom{j-1}{k} \bar{s}^{k+1}\,s^{j-k}\,\pi_k + \\
&\quad + \sum_{k=1}^{\min\{j,N+1\}} [1-(1-\delta_{k,N+1})\,p\,r_k] \binom{j-1}{k-1} \bar{s}^{k}\,s^{j-k+1}\,\pi_k + \\
&\quad +(1-\delta_{j,1}) \sum_{k=2}^{\min\{j,N+1\}} p\,r_{k-1} \binom{j-1}{k-1} \bar{s}^{k+1}\,s^{j-k}\,\pi_k + \\
&\quad + \sum_{k=2}^{\min\{j+1,N+1\}} (1-p\,r_{k-1}) \binom{j-1}{k-2} \bar{s}^{k}\,s^{j-k+1}\,\pi_k\,,\ j \geq 1\,.
\end{aligned}
$$

Proof: If $X_m = 0$ or ($X_m = 1$ and the current service finishes in the $m+1$-th slot):

- if an arrival occurs in the $m+1$-th slot (with probability p), the unfinished work will be the service time of the arriving customer.
- if an arrival does not occur in the $m+1$-th slot (with probability $\bar{p}$), the unfinished work will be zero.

If $X_m = k \geq 1$ and the current service does not conclude in the $m+1$-th slot:

- if a customer enters the system in the $m+1$-th slot (with probability $(1-\delta_{k,N+1})\,p\,r_k$), the unfinished work will be the sum of k complete service times plus the remaining service time of the customer currently being served; then the unfinished work will be the sum of $k+1$ independent and geometrically distributed random variables with parameter $\bar{s}$.
- if a customer does not enter the system in the $m+1$-th slot (with probability $1-(1-\delta_{k,N+1})\,p\,r_k$), the unfinished work will be the sum of $k-1$ complete service times plus the remaining service time of the customer currently being served; then the unfinished work will be the sum of k independent and geometrically distributed random variables with parameter $\bar{s}$.

If $X_m = k \geq 2$ and the current service ends in the $m+1$-th slot:

- if a customer joins the system in the $m+1$-th slot (with probability $p\,r_{k-1}$), the unfinished work will be the sum of k complete service times, that is, the unfinished work will be the sum of k independent geometrical distributions with parameter $\bar{s}$.

For Peer Review

- if a customer does not join the system in the $m+1$-th slot (with probability $1 - p\, r_{k-1}$), the unfinished work will be the sum of $k-1$ complete service times, that is, the unfinished work will be the sum of $k-1$ independent geometrical distributions with parameter $\bar{s}$.

□

7 Optimal control

In practice of designing concrete systems (computer and communication networks, manufacturing systems,...), a question arises to the manager: *How is the buffer size chosen to minimize the cost function?* This type of matters motivates that the optimal control of the buffer size is investigated in this section.

We develop a steady-state expected cost function per unit time, in which N is a decision variable. Our aim is to determine the optimum value of the control parameter N, say N^*, so that this function is minimized. The following costs are considered:

$$\begin{aligned} C_1 &\equiv \text{holding cost per unit time per customer in the waiting line,} \\ C_2 &\equiv \text{cost incurred per unit time for keeping the server idle,} \\ C_3 &\equiv \text{cost per unit time per balked customer,} \\ C_4 &\equiv \text{cost per unit time per lost customer.} \end{aligned}$$

The performance measures, such as the expected number of customers in the system μ_s and the expected number of customers in the queue μ_q, can be obtained from the steady-state probabilities and are given by:

$$\mu_s = \sum_{k=1}^{N+1} k\, \pi_k \ , \quad \mu_q = \sum_{k=2}^{N+1} (k-1)\, \pi_k \, .$$

Since the probability that a customer balks in the system is $1 - r_k$ if the system size is $k = 1, \ldots, N$, the balking probability is given by:

$$\alpha = \sum_{k=1}^{N} (1 - r_k)\, \frac{\pi_k}{1 - \pi_0 - \pi_{N+1}} \, .$$

Using the definitions of each cost element listed above, the total expected cost function per unit time is given by:

$$C(N) = C_1\, \mu_q + C_2\, \pi_0 + C_3\, \alpha + C_4\, \pi_{N+1} \, .$$

We now carry out a sensitivity analysis on the optimum value N^* based on changes in specific values of the system parameter r under the two balking disciplines. The following cost elements are used:

$$C_1 = 2 \text{ euros/slot} \ , \quad C_2 = 1 \text{ euro/slot} \ , \quad C_3 = 3 \text{ euros/slot} \ , \quad C_4 = 4 \text{ euros/slot.}$$

For Peer Review

The lightest cases (C_1 and C_2) are due to the costs induced per buffer size and per maintaining the server inactive, respectively; albeit holding customers in the buffer is costlier than having an unoccupied server ($C_2 < C_1$). The highest costs (C_3 and C_4) are presented when a customer does not receive his service, though it is less serious when the omission of the service is a decision of the customer ($C_3 < C_4$). Obviously, the values of the costs can change as function of the decisor priorities.

The analytical study of the behaviour of the expected cost function is an arduous task since the decision variable N is discrete and appears in a complex and non-linear expression. Hence, the optimum value N^* will be the first integer by satisfying the inequality $C(N^*-1) \geq C(N^*) \leq C(N^*+1)$.

	$r = 0.05$	$r = 0.1$	$r = 0.2$	$r = 0.4$
N^*	5	6	6	2
$C(N^*)$	3.16927	3.13	3.15438	3.68382
μ_q	0.0496361	0.110831	0.285909	0.568395
π_0	0.22	0.208333	0.18183	0.134916
α	0.95	0.9	0.8	0.6
π_{N^*+1}	$5.62163 \cdot 10^{-7}$	$2.38219 \cdot 10^{-6}$	0.000184056	0.153029

	$r = 0.6$	$r = 0.8$	$r = 0.9$	$r = 0.95$
N^*	2	2	2	2
$C(N^*)$	4.00389	4.91617	5.15506	5.25639
μ_q	0.816667	1.21226	1.3421	1.40122
π_0	0.0816667	0.0568687	0.0437719	0.0381296
α	0.4	0.2	0.1	0.05
π_{N^*+1}	0.272222	0.458694	0.531774	0.566452

Table 1: Constant rate policy.

Tables 1–2 present results for constant and discouraged rate policy, respectively. The results are rounded to six significant digits. We first fix $(p, \bar{s}) = (0.5, 0.3)$ and choose several values of r. The minimum expected cost $C(N^*)$ and the values of various system performance measures μ_q, π_0, α and π_{N^*+1} at the optimum values N^* are shown in Tables 1–2 for different values of r. One sees that N^*, as function of r, starts increasing to a maximum and then decreases until 2. From the last five columns of Table 1, N^* does not change at all when r varies from 0.4 to 0.95. From the last three columns of Table 2, the optimum values N^* are the same even though r changes from 0.8 to 0.95. As the intuition tells us, the mean buffer size μ_q increases with increasing values of r, while π_0 and α reduce as r increases.

Let us briefly compare the data in Table 1 with those in Table 2. It is observed that $C(N^*)$ is slightly superior with a constant rate policy and the balking probability is moderately higher with a discouraged rate discipline. As

For Peer Review

	$r = 0.05$	$r = 0.1$	$r = 0.2$	$r = 0.4$
N^*	3	4	4	6
$C(N^*)$	3.16715	3.11701	3.05614	3.11441
μ_q	0.0442488	0.0882514	0.180884	0.421937
π_0	0.220588	0.210637	0.190874	0.149172
α	0.952689	0.909958	0.8345	0.707121
π_{N^*+1}	$1.87364 \cdot 10^{-8}$	$1.37687 \cdot 10^{-10}$	$1.37509 \cdot 10^{-7}$	$7.89711 \cdot 10^{-9}$

	$r = 0.6$	$r = 0.8$	$r = 0.9$	$r = 0.95$
N^*	7	2	2	2
$C(N^*)$	3.59937	4.63386	5.02358	5.19561
μ_q	0.859605	1.08942	1.27753	1.36864
π_0	0.100583	0.0657373	0.0480679	0.0402047
α	0.593192	0.297391	0.159062	0.0822321
π_{N^*+1}	$1.36671 \cdot 10^{-6}$	0.374278	0.48582	0.542858

Table 2: Discouraged rate policy.

For Peer Review

was expected, the costs are minimized with the discouraged policy since, under this rule, the balking probability increases as function of the system size. Moreover, the probability that the system is empty (respectively, the loss probability) is smaller (respectively, bigger) in the constant case than in the discouraged one.

8 Numerical work

This section shows some numerical results regarding the stationary distribution of our systems. Besides, as performance measures, we choose the mean system size μ_s, the balking probability α, the variance σ_s^2 of the random variable "system size", the mean waiting/spending time and the mean unfinished work. The arrival and departure probabilities p and $\bar{s}$ in a slot are assumed to be equal to 0.5 and 0.3, respectively. The following graphics and tables corroborate what the analytical results and intuition tell us.

In Figure 6, the stationary probabilities π_k are plotted versus r when the buffer size is 3. The curves corresponding to the constant and discouraged rate policies show a similar behaviour. We can observe that π_0 and π_1 decrease with increasing values of r, which also agrees with the intuition. On the other hand, we see that π_2 and π_3 are increasing and then decreasing according to the rises of r. As is to be expected, the loss probability π_4 is increasing as function of r.

Figure 7 considers the effect of the parameter r on the mean system size and the balking probability.

Figure 7(a) illustrates the development of μ_s as function of r. We present six curves which correspond to $N = 10, 20, 30$ subject to the two disciplines. As was expected, the mean system size μ_s grows with increasing values of r

and N. Under the constant rate rule, this increase is moderate when r varies from 0 to 0.5, while the increment is considerable when r is superior to 0.5. An analogous behaviour is observed for the discouraged rate policy, since the growth is controlled if $r < 0.9$ and is heightened if r approaches 1. Moreover, assuming r inferior to 0.5, μ_s is scarcely affected by the discipline and the buffer capacity. We also note that, for any value of r, the mean system size μ_s is bigger in the constant case than in the discouraged one; although the difference between both rules, which increases with N, is intensified when r exceeds 0.5.

In Figure 7(b), the balking probability α is plotted against r for the two disciplines. The curves are decreasing as function of r, in accordance with the analytical expression of α. As is to be expected, the highest curves correspond to the discouraged rate policy. Under the discouraged case, the balking probability α rises with increasing values of N, but these differences are only appreciated when r is close to 1.

Figure 8 depicts the variance σ_s^2 of the random variable "system size" against r in relation to three values of N and the two considered schemes. Obviously, σ_s^2 increases with N, i.e., the concentration of the stationary distribution of the system size reduces as the buffer capacity grows. It can be seen that all the curves, as function of r, increase to a maximum and then decrease. In the constant rate discipline, observe Figure 8(a), the maxima are got when r takes the value of the ergodicity condition in the case of infinite buffer $(r = 1/\rho)$. In the discouraged rule, see Figure 8(b), the values of r where the maxima are reached tend to 1 as N increases. Moreover, the maxima of σ_s^2 are smaller in the discouraged case than in the constant one, what confirms that the discouraged rate policy continually receives and processes information in order to decide the entrance in the system or not of the arriving customers. In both policies, when r belongs to certain intervals, the variance σ_s^2 is only lightly affected by the buffer size, whereas the parameter N influences considerably in the complementary of such intervals.

Tables 3–5 present the mean waiting/spending time and the mean unfinished work for different values of N and r. The results are rounded to four decimal digits. As was expected, these performance measures increase with increasing values of r. In the constant case, for $r = 0.05, 0.2$ the values do not change when N varies from 10 to 30, whereas for $r = 0.4, 0.6, 0.8, 0.95$ the values are increasing as function of N. In the discouraged case, for $r = 0.05, 0.2, 0.4, 0.6$ the values remain constant when N varies from 10 to 30, while for $r = 0.8, 0.95$ the values grow with N. As is to be expected, these performance characteristics are bigger for the constant policy than for the discouraged one; though the differences, which increase with N, are very small when $r = 0.05, 0.2$ and are intensified when $r = 0.4, 0.6, 0.8, 0.95$.

9 Conclusions

It should be pointed out that the economic importance of this paper resides in the multiple applications to productive processes, since most of them operate on

For Peer Review

	$r = 0.05$	$r = 0.2$	$r = 0.4$	$r = 0.6$	$r = 0.8$	$r = 0.95$
	Constant rate policy					
$N = 10$	1.9855	2.8636	6.1334	19.6791	29.9141	32.6736
$N = 20$	1.9855	2.8636	6.2215	36.4278	63.0056	66.0000
$N = 30$	1.9855	2.8636	6.2222	52.5529	96.3334	99.3333
	Discouraged rate policy					
$N = 10$	1.9661	2.4909	3.3917	4.9640	9.2520	26.4841
$N = 20$	1.9661	2.4909	3.3917	4.9640	9.2521	34.3299
$N = 30$	1.9661	2.4909	3.3917	4.9640	9.2521	34.4109

Table 3: The mean waiting time.

	$r = 0.05$	$r = 0.2$	$r = 0.4$	$r = 0.6$	$r = 0.8$	$r = 0.95$
	Constant rate policy					
$N = 10$	5.3188	6.1970	9.4668	23.2257	33.2474	36.0069
$N = 20$	5.3188	6.1970	9.5548	39.8809	66.3389	69.3333
$N = 30$	5.3188	6.1970	9.5556	55.9326	99.6668	102.6667
	Discouraged rate policy					
$N = 10$	5.2995	5.8242	6.7251	8.2973	12.5853	29.8175
$N = 20$	5.2995	5.8242	6.7251	8.2973	12.5854	37.6633
$N = 30$	5.2995	5.8242	6.7251	8.2973	12.5854	37.7442

Table 4: The mean spending time.

a discrete-time basis. In fact, this paper presents two examples applied to the industrial and services sectors. Furthermore, the optimal control of the buffer size has been investigated in order to minimize a cost model, what represents a main objective from the enterprise point of view.

On the other hand, the mathematical importance is in the direct resolution of the systems of equations, since we find explicit formulae for the stationary probabilities of the system size, of the waiting/spending time in the FCFS system and of the unfinished work. From the solutions we can obtain, among others things, all the moments for the stationary distributions of the number of customers in the system and in the waiting line, the measures of effectiveness,... Moreover, the formulae of the finite buffer can be used for the approximate calculus of the stationary features of the respective queueing systems with infinite buffer capacity (for values of N large enough).

References

[1] Abou-El-Ata, M.O. and Hariri, A.M.A. Point estimation and confidence intervals of the $M/M/2/N$ queue with balking and heterogeneity. American

For Peer Review

	$r = 0.05$	$r = 0.2$	$r = 0.4$	$r = 0.6$	$r = 0.8$	$r = 0.95$
	Constant rate policy					
$N = 10$	2.7655	3.6818	7.0221	20.7149	30.9128	33.6735
$N = 20$	2.7655	3.6818	7.1104	37.4492	64.0055	67.0000
$N = 30$	2.7655	3.6818	7.1111	53.5571	97.3334	100.3333
	Discouraged rate policy					
$N = 10$	2.7455	3.3000	4.2426	5.8634	10.2110	27.4826
$N = 20$	2.7455	3.3000	4.2426	5.8634	10.2111	35.3290
$N = 30$	2.7455	3.3000	4.2426	5.8634	10.2111	35.4100

Table 5: The mean unfinished work.

Journal of Mathematical and Management Sciences 15, 35-55 (1995).

[2] Acemoglu, D. and Shimer, R. Holdups and efficiency with search frictions. International Economic Review 40, 827-849 (1999).

[3] Artalejo, J.R. and López-Herrero, M.J. On the single server retrial queue with balking. Infor 38, 33-50 (2000).

[4] Atencia, I. and Moreno, P. The discrete-time $Geo/Geo/1$ queue with negative customers and disasters. Computers and Operations Research 31, 1537-1548 (2004).

[5] Atencia, I. and Moreno, P. A single-server G-queue in discrete-time with geometrical arrival and service process. Performance Evaluation 59, 85-97 (2005).

[6] Borovkov, K.A. Propagation of chaos for queueing networks. Theory of Probability and its Applications 42, 385-394 (1998).

[7] Bruneel, H. and Kim, B.G. Discrete-time models for communication systems including ATM. Kluwer Academic Publishers, Boston (1993).

[8] Choi, B.D., Kim, B. and Chung, J. $M/M/1$ queue with impatient customers of higher priority. Queueing Systems 38, 49-66 (2001).

[9] Cooper, R.B. Introduction to queueing theory. Elsevier North Holland, New York (1981).

[10] Friedman, E.J. and Landsberg, A.S. Long-run dynamics of queues: stability and chaos. Operations Research Letters 18, 185-191 (1996).

[11] Gravey, A. and Hébuterne, G. Simultaneity in discrete-time single server queues with Bernoulli inputs. Performance Evaluation 14, 123-131 (1992).

[12] Haskose, A., Kingsman, B.G. and Worthington, D. Modelling flow and jobbing shops as a queueing network for workload control. International Journal of Production Economics 78, 271-285 (2002).

For Peer Review

[13] Haxholdt, C., Larsen, E.R. and Van Ackere, A. Mode locking and chaos in a deterministic queueing model with feedback. Management Science 49, 816-830 (2003).

[14] Hunter, J.J. Mathematical techniques of applied probability. Vol. 2. Discrete-time models: techniques and applications. Academic Press, New York (1983).

[15] Ikeda, Z. and Nishida, T. $M/G/1$ queue with balking. Mathematica Japonica 33, 707-711 (1988).

[16] Kapur, P.K., Garg, R.B., Sehgal, V.K. and Mishra, G.D. Numerical computations of discrete-time solutions for a multi-server queue with balking and reneging. Asia-Pacific Journal of Operational Research 13, 1-15 (1996).

[17] Köchel, P. Finite queueing systems-structural investigations and optimal design. International Journal of Production Economics 88, 157-171 (2004).

[18] Krishna Kumar, B., Parthasarathy, P.R. and Sharafali, M. Transient solution of an $M/M/1$ queue with balking. Queueing Systems 13, 441-448 (1993).

[19] Masters, A.M. Wage posting in two-sided search and the minimum wage. International Economic Review 40, 809-826 (1999).

[20] Montazer-Haghighi, A., Medhi, J. and Mohanty, S.G. On a multiserver Markovian queueing system with balking and reneging. Computers and Operations Research 13, 421-425 (1986).

[21] Parra, I. and Aranda, J. Multiproduct monopoly: a queueing approach. Applied Economics 31, 565-576 (1999).

[22] Sattinger, M. A queuing model of the market for access to trading partners. International Economic Review 43, 533-547 (2002).

[23] Sleptchenko, A., Van Der Heijden, M.C. and Van Harten, A. Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems. International Journal of Production Economics 79, 209-230 (2002).

[24] Smet, M. Multi-product costs and standby capacity derived from queuing theory: the case of Belgian hospitals. Applied Economics 36, 1475-1487 (2004).

[25] Takagi, H. Queueing analysis: a foundation of performance evaluation. Vol. 3. Discrete-time systems. North-Holland, Amsterdam (1993).

[26] Van Nyen, P.L.M., Van Ooijen, H.P.G. and Bertrand, J.W.M. Simulation results on the performance of Albin and Whitt's estimation method for waiting times in integrated production-inventory systems. International Journal of Production Economics 90, 237-249 (2004).

For Peer Review

[27] Wang, K.-H. and Chang, Y.-C. Cost analysis of a finite $M/M/R$ queueing system with balking, reneging, and server breakdowns. Mathematical Methods of Operations Research 56, 169-180 (2002).

[28] Woodward, M.E. Communication and computer networks: modelling with discrete-time queues. IEEE Computer Society Press, Los Alamitos, California (1994).

[29] Yang, T., Posner, M.J.M. and Templeton, J.G.C. The $M/G/1$ retrial queue with nonpersistent customers. Queueing Systems 7, 209-218 (1990).

For Peer Review

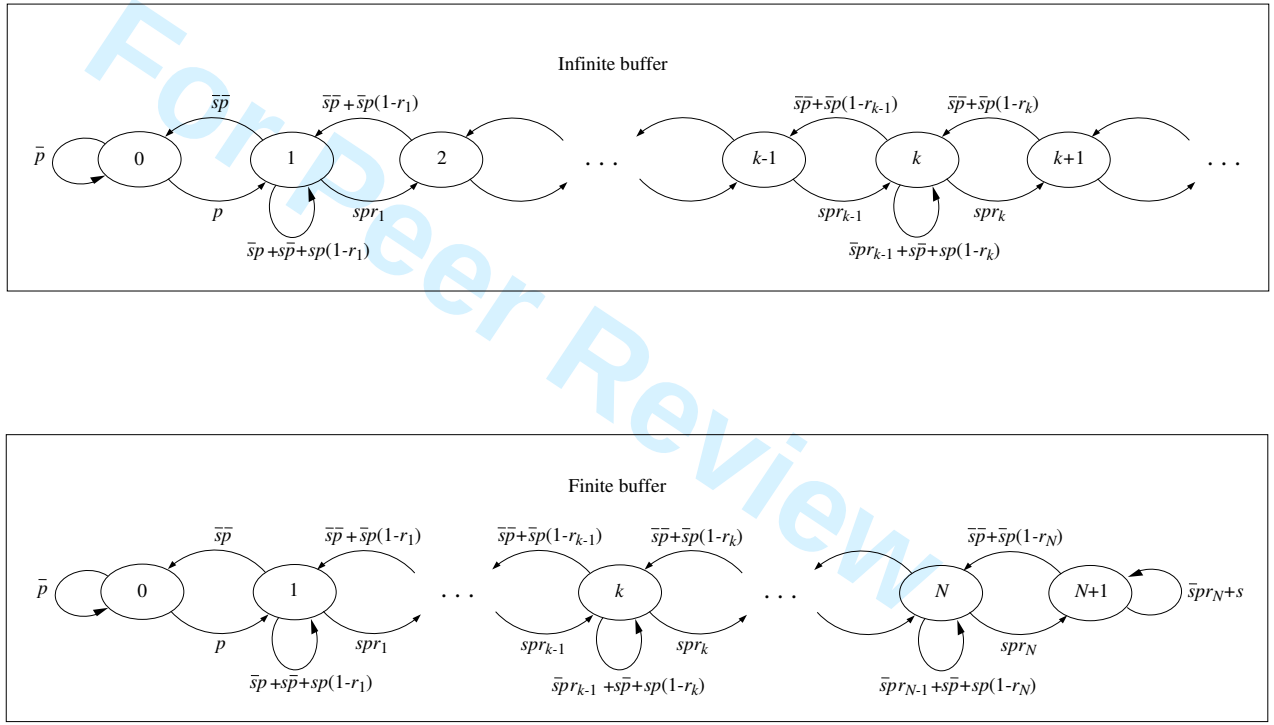
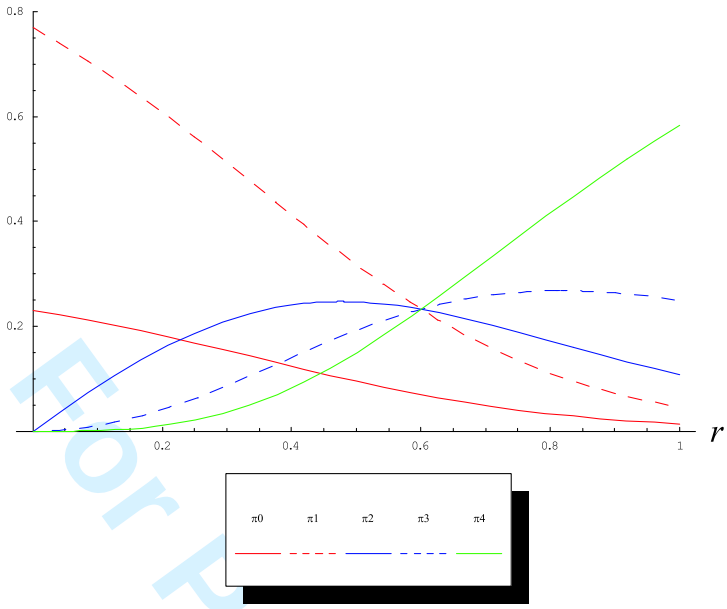
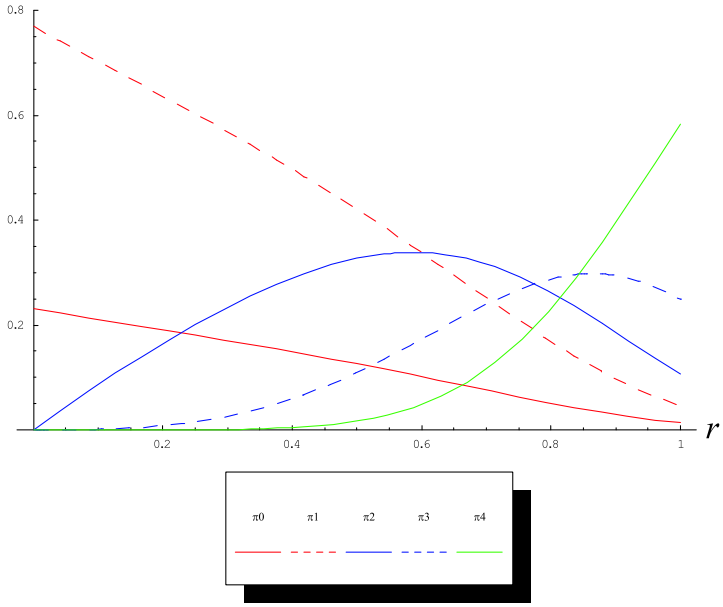


Figure 5: One-step transition probabilities diagrams.

For Peer Review

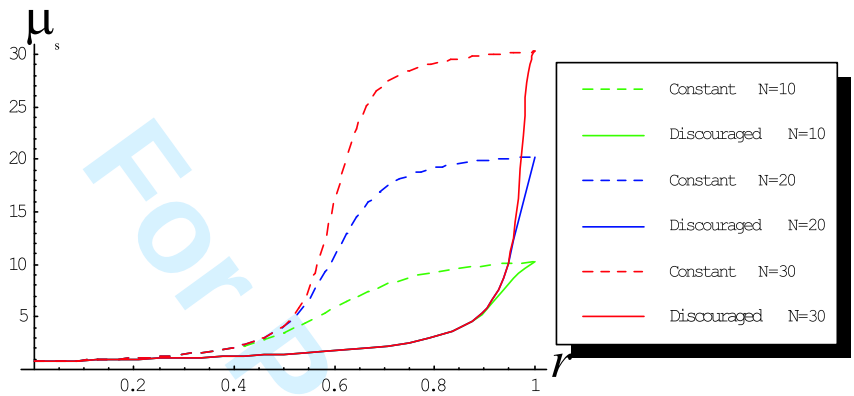


(a) Constant rate policy.

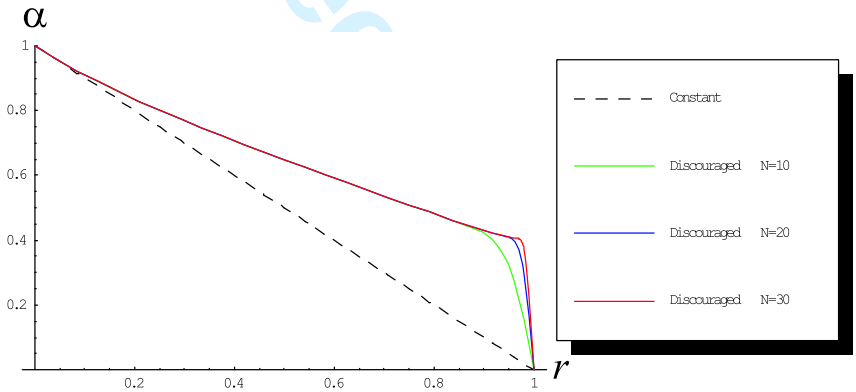


(b) Discouraged rate policy.

Figure 6: The buffer capacity is $N = 3$.

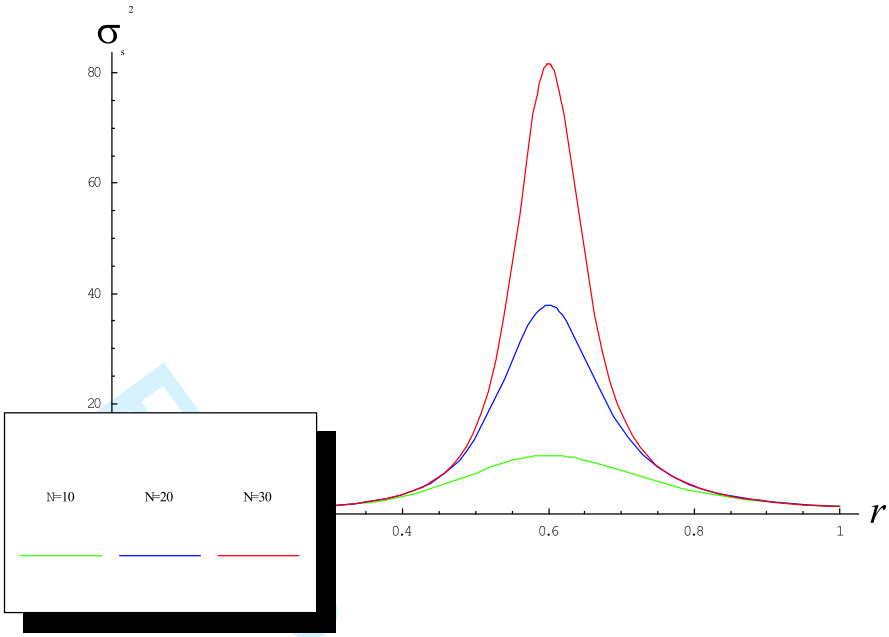


(a) The mean system size μ_s versus r .

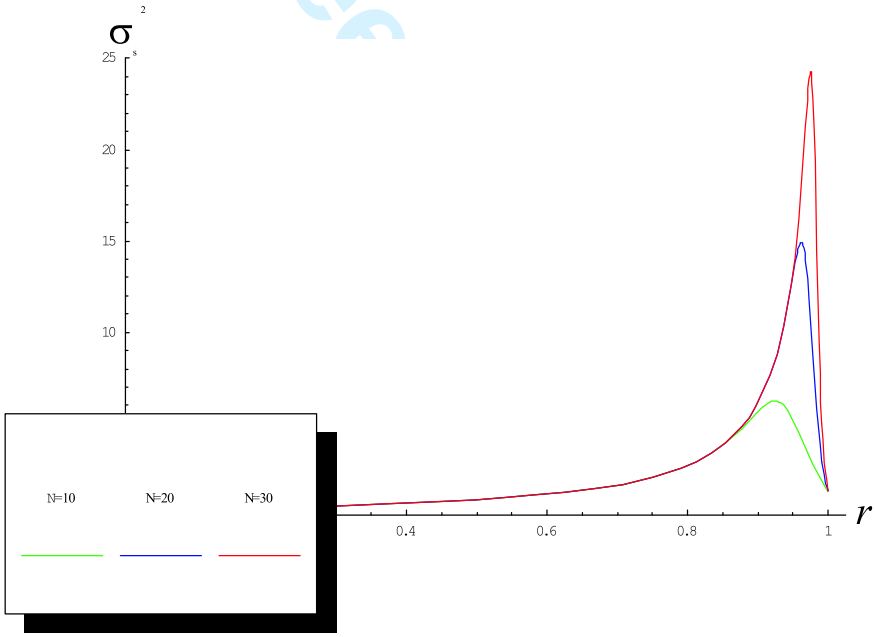


(b) The balking probability α versus r .

Figure 7: The effect of the parameter r .



(a) Constant rate policy.



(b) Discouraged rate policy.

Figure 8: The variance σ_s^2 of the random variable "system size" versus r .