CWI
Centrum voor Wiskunde en Informatica
REPORTRAPPORT

On the maximum number of customers simultaneously present in a queue

R.J. Boucherie

Department of Operations Research, Statistics, and System Theory

Report BS-R9423 July 1994

Copyright © Stichting Mathematisch Centrum
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

On the Maximum Number of Customers Simultaneously Present in a Queue

Richard J. Boucherie

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

This paper studies the behaviour of the maximum queue length both in queues with uniformly bounded service rates, such as multiserver queues, and in queues with unbounded service rates, such as the infinite server queue. It is shown that the behaviour of this maximum is substantially different for these two classes of queues. First the maximum queue length is characterised in one busy cycle. The relation between this maximum and blocking probabilities, and the Palm distribution at arrival instants is discussed. Then, bounds for the asymptotic distribution of normalised maxima are derived. For queues with uniformly bounded service rates it is shown that the sequence of partial maxima for uniformizable queues is stochastically compact, and that this sequence increases logarithmically with probability 1. For queues with unbounded service rates convergence in probability is obtained. Almost sure convergence of the sequence of normalised partial maxima to a constant is obtained for a subclass, which includes the infinite server queue.

AMS Subject Classification (1991): Primary: 60K25, Secondary: 90B22.
Keywords & Phrases: extremal values, limit theorems, loss formula, maximum number of customers.
Note: The main part of this work was carried out while the author was visiting CWI with the support of the ERCIM Fellowship Programme. This work is also partly supported by the European Grant BRA-QMIPS of CEC DG XIII, and by a Human Capital and Mobility Fellowship, Nr ERB4001GT930392.

1. Introduction

The maximum number of customers simultaneously present in a queueing system is an important performance measure directly related to the quality of service. It not only determines the maximum queue size found by customers arriving to the queueing system, but might also be used to determine optimal buffer sizes for queues with finite waiting room. This paper characterises the behaviour of the maximum number of customers simultaneously present in a single queue.

Let $\mathbf{Y}$ denote the maximum number of customers simultaneously present during a busy cycle. Then, for all queues studied in this paper,

$$\mathbf{P}(\mathbf{Y} \leq n) = 1 - \pi(0)^{-1} \Big/ \sum_{k=0}^{n} \pi(k)^{-1}, \quad n = 0, 1, 2, \ldots, \tag{1.1}$$

where $\pi(n)$ is the probability that an arriving customer finds n customers in the queue. The remarkable simplicity of this expression allows us to obtain not only the limiting

Report BS-R9423
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

behaviour of $\mathbf{Y}$, but also the limiting behaviour of $\mathbf{Y}^{(k)}$, the maximum number of customers simultaneously present in k subsequent busy cycles.

From (1.1) we obtain an interesting dual relation between the discrete failure rate $\mathbf{P}(\mathbf{Y} = n | \mathbf{Y} \geq n) = \pi(n)^{-1} / \sum_{k=0}^{n} \pi(k)^{-1}$ and the probability that an arriving customer in a queue with finite capacity n is lost (blocked and cleared upon arrival): $B(n) = \pi(n) / \sum_{k=0}^{n} \pi(k)$. This form of duality was first observed in Cohen [1971] for the Erlang model, and is of a form similar to the dual or inverse relation between the M/G/1 and GI/M/1 queue observed in Takács [1962], and further discussed and generalised in Niu and Cooper [1989], and Kimura [1993].

The limiting behaviour of $\mathbf{Y}$ can be obtained from (1.1). It will be shown that $\mathbf{P}(\mathbf{Y} > n) \sim \pi(n)$ $(n \to \infty)$, that is the probability that the maximum exceeds n is proportional to the probability that an arriving customer meets n customers. This result allows us to use the theory on extremal values to characterise the limiting behaviour of $\mathbf{Y}^{(k)}$ as $k \to \infty$. In particular, results from extremal value theory for discrete distributions such as developed by Anderson [1970] and Vervaat [1973] can be used to characterise $\mathbf{Y}^{(k)}$. Of main interest in this respect is the existence of norming constants $\{a_k > 0,\ b_k\}_{k \geq 1}$ such that $(\mathbf{Y}^{(k)} - b_k)/a_k$ converges weakly to a non-degenerate distribution $G(\cdot)$. However, for the discrete distributions $\pi(\cdot)$ typically occurring for stable queues, it is a consequence of standard results (e.g. Leadbetter *et al.* [1983], pp. 24-27) that norming constants such that $\lim_{k \to \infty} \mathbf{P}((\mathbf{Y}^{(k)} - b_k)/a_k \leq x) = G(x)$, where $G(\cdot)$ is non-degenerate, do not exist. Anderson [1970] computes lim inf's and lim sup's. We will show that norming constants $\{a_k > 0,\ b_k\}_{k \geq 1}$ exist such that the sequence of sample maxima $\mathbf{Y}^{(k)}$ for stable, uniformizable queues is *stochastically compact*: we apply results of De Haan and Resnick [1984] to show that every sequence $(\mathbf{Y}^{(n(k))} - b_{n(k)})/a_{n(k)}$ contains a subsequence that converges weakly to a non-degenerate distribution $G(x) = \exp[-\exp[-(x+\epsilon)]]$, $-\infty < x < \infty$. This adequately characterises the limiting behaviour of $\mathbf{Y}^{(k)}$ for queues with uniformly bounded service rates. In contrast, for queues with unbounded service rates such as the infinite server queue, we show that the sequence of sample maxima is not stochastically compact, but we do compute lim inf's and lim sup's of a form closely related to the results for queues with uniformly bounded service rates.

From the lim inf's and lim sup's for $\mathbf{P}((\mathbf{Y}^{(k)} - b_k)/a_k \leq x)$, following arguments such as developed in Cohen [1982], and Galambos [1987], we obtain that $Y^{(k)}/b_k$ converges in probability to 1. For queues with uniformly bounded service rates this can be seen as an application of known results for sample maxima. For such queues, under mild additional assumptions, this result is strengthened: we will show that $Y^{(k)}/b_k$ converges to 1 with probability 1. For queues with unbounded service rates we cannot obtain this result in the general framework, but from the specific form of the distribution $\pi(\cdot)$ a similar result is obtained for a subclass containing the infinite server queue.

This paper applies results from extremal value theory such as developed in Anderson [1970], Galambos [1987], De Haan and Resnick [1984], and Leadbetter *et al.* [1983] in a queueing framework. We will focus on positive recurrent processes. Null recurrent processes and processes that are close to null recurrent processes are discussed in Serfozo [1988]. Transient processes are the topic of heavy traffic theory, and results can for example be found in Iglehart and Whitt [1970], Iglehart [1972].

The paper is organised as follows. Section 2 introduces the model for a single queue and obtains several relations for the maximum number of customers simultaneously present in a single busy cycle. Although the queue studied in this paper is introduced via birth-and-death processes, the distribution for the maximum queue size is derived via the embedded chain for which the transition probabilities are of a general form. In section 3 we characterise the tail properties of $\mathbf{P}(\mathbf{Y} \leq n)$ for stable and unstable queues. These results are then applied to study weak convergence of sample maxima. In section 4 we investigate convergence in probability and almost sure convergence of the sample maxima. Finally, section 5 presents some examples from queueing theory, and section 6 presents conclusions and aims for further research.

2. The maximum in one busy cycle

Consider a single queue to which customers arrive singly and according to a state-dependent Poisson process with arrival rate $\lambda\widehat{\psi}(n)$ when n customers are present at the queue. The service times of the customers are independent stochastic variables, all negative exponentially distributed with rate μ. The service speed of the server at the queue is $\widehat{\phi}(n)$ when n customers are present.

Let $\mathbf{X} = (X(t),\ t \geq 0)$ record the number of customers in the queue at time t, then $\mathbf{X}$ is a birth-and-death process at state space $S = \mathbb{N}_0$, with birth and death rates $q(n, n+1) = \lambda\widehat{\psi}(n)$, $q(n+1, n) = \mu\widehat{\phi}(n+1)$, $n = 0, 1, 2, \ldots$. Define $\phi(n) = \prod_{k=1}^{n}(\widehat{\psi}(k-1)/\widehat{\phi}(k))$, and $\psi(n) = \widehat{\psi}(n)\prod_{k=1}^{n}(\widehat{\psi}(k-1)/\widehat{\phi}(k))$, $\psi(n) = 0$ if $n < 0$, and $\psi(0) = 1$, then the birth and death rates are

$$q(n, n+1) = \frac{\psi(n)}{\phi(n)}\lambda, \quad n \geq 0, \qquad q(n, n-1) = \frac{\psi(n-1)}{\phi(n)}\mu, \quad n > 0. \tag{2.1}$$

This form of the birth and death rates is chosen in agreement with the form used in the literature on product form queueing networks (cf. Boucherie and van Dijk [1991], Henderson and Taylor [1990]). For example, the single server queue corresponds to $\psi(n) = \phi(n) = 1$, $n \geq 0$, and the infinite server queue is obtained by setting $\psi(n) = \phi(n) = 1/n!$, $n \geq 0$. The form (2.1) for the rates enables us to obtain simple expressions for the equilibrium distribution, and the Palm distribution at arrival instants, as will be illustrated below.

Queues with finite waiting room are included in the model. If $N = \min\{n > 0 : \psi(n) = 0\}$, then $q(N, N+1) = 0$. As a consequence, when the queue starts off empty the number of customers in the queue cannot exceed N. Without loss of generality we will also set $\psi(n) = 0$, $n > N$. In this case, if $X(0) \leq N$, then $S = \{0, 1, \ldots, N\}$. For the birth and death rates to be properly defined, it must be assumed that $\phi(n) > 0$, $n = 0, \ldots, N$.

A measure $m = (m(n),\ n \in S)$ is an invariant measure for $\mathbf{X}$ if m satisfies the *detailed balance equations* (Kelly [1979], p. 5), for all $n, n' \in S$,

$$m(n)q(n,n') - m(n)q(n',n) = 0. \tag{2.2}$$

From the birth and death rates (2.1) we obtain that, with $\rho \equiv \lambda/\mu$,

$$m(n) = \phi(n)\rho^n, \quad n = 0, 1, \ldots, N \leq \infty \tag{2.3}$$

is an invariant measure for $\mathbf{X}$. Observe that the form of m is independent of N, i.e., independent of the size of the waiting room (finite or infinite).

Ergodicity of $\mathbf{X}$ can be determined from the birth and death rates. If $\psi(n) > 0$, $n = 0, 1, 2, \ldots$, we will assume that $\sum_n \{q(n,n+1) + q(n,n-1)\}^{-1} = \infty$. This guarantees that $\mathbf{X}$ can make at most a finite number of steps in any finite interval ($\mathbf{X}$ is regular). From (2.3) we obtain that $\mathbf{X}$ is ergodic if and only if

$$B_\phi^{-1} = \sum_{n=0}^{N} \phi(n)\rho^n < \infty. \tag{2.4}$$

If this is the case, then $\mathbf{X}$ has a unique stationary distribution

$$\pi_q(n) = \lim_{t\to\infty} \mathrm{P}(X(t) = n) = B_\phi \phi(n)\rho^n, \quad n = 0, 1, \ldots, N \leq \infty.$$

Ergodicity of $\mathbf{X}$ is determined by $\phi(\cdot)$ via (2.4). The function $\psi(\cdot)$ characterises the behaviour of $\mathbf{X}$ when $\sum_n m(n) = \infty$. To this end, from the assumption that $\sum_n \{q(n,n+1) + q(n,n-1)\}^{-1} = \infty$ we first obtain that for $B_\phi^{-1} < \infty$ it must be that $\sum_n \{\psi(n)\rho^n\}^{-1} = \infty$. From Foster's criterions (Foster [1953]) we see that $\mathbf{X}$ is null recurrent iff $\sum_n m(n) = \infty$ and $\sum_n \{\psi(n)\rho^n\}^{-1} = \infty$, and transient iff $\sum_n \{\psi(n)\rho^n\}^{-1} < \infty$. Therefore, if

$$B_\psi^{-1} = \sum_{n=0}^{\infty} \psi(n)\rho^n < \infty,$$

then $\mathbf{X}$ is recurrent (ergodic or null recurrent). Furthermore, if $B_\psi^{-1} < \infty$, the Palm probability at arrival instants is well defined, and is given by (Brumelle [1978])

$$\pi(n) = B_\psi \psi(n)\rho^n, \quad n \geq 0,$$

the probability that an arriving customer finds n customers in the queue. Note that, in general, $\pi(n) \neq \pi_q(n)$. Equality holds iff the birth rates are state independent, which corresponds to a Poisson arrival process.

Denote by $Z(k)$ the state $X(t)$ immediately after the k-th jump of $\mathbf{X}$ after $t = 0$. The sequence $\mathbf{Z} = (Z(k),\ k = 1, 2, \ldots)$ is a discrete-time Markov chain with state space S and with stationary one-step transition probabilities $p(n, n') \overset{\text{def}}{=} \mathbf{P}(Z(k+1) = n' | Z(k) = n)$. If $N = \infty$ these probabilities are

$$p(n-1, n) = \lambda \frac{\psi(n-1)}{\varphi(n-1)}, \quad p(n, n-1) = \mu \frac{\psi(n-1)}{\varphi(n)}, \quad n \geq 1, \tag{2.5}$$

and if $N < \infty$

$$p(n-1, n) = \lambda \frac{\psi(n-1)}{\varphi(n-1)}, \quad p(n, n-1) = \mu \frac{\psi(n-1)}{\varphi(n)}, \quad n = 1, \ldots, N, \quad p(N, N) = \lambda \frac{\psi(N)}{\varphi(N)},$$

where $\varphi(n) \equiv \lambda\psi(n) + \mu\psi(n-1)$, $n \geq 0$. By insertion into the detailed balance equations (2.2), with $p(\cdot, \cdot)$ replacing $q(\cdot, \cdot)$, we obtain that the jump-chain $\mathbf{Z}$ has an invariant measure

$$m_p(n) = \varphi(n)(\lambda/\mu)^n, \quad n = 0, \ldots, N.$$

Ergodicity of $\mathbf{Z}$ is completely determined by ψ.

For the jump-chain we denote for $0 \leq h < N$, the taboo probability that the chain, when starting in state i reaches state j without reaching state $h + 1$ (taboo state), by

$$f_{h;i,j} = \mathbf{P}\Big(\bigcup_{k=2}^{\infty} \{Z(k) = j,\ j < Z(r) \leq h,\ r = 1, \ldots, k-1\} | Z(1) = i\Big), \quad j \leq i \leq h.$$

For the model with finite waiting room, define $f_{N;i,j}$ as the probability that the chain, when starting in state i reaches state j without the occurence of a transition from state N to state N, i.e., without the occurence of an arrival for $\mathbf{X}$ when the waiting room is full.

The taboo probability $f_{h;i,j}$ is an entrance probability. For a proper definition of $f_{h;i,j}$, stability of $\mathbf{X}$ or $\mathbf{Z}$ is not required. Therefore, in the following result $f_{h;i,j}$ is expressed in the invariant measure at arrival instants.

Proposition 2.1 (Chung [1967], p. 73) *For finite and infinite queues, for $h \leq N$*

$$f_{h;i,j} = \sum_{k=i}^{h} \{\psi(k)\rho^k\}^{-1} \Big/ \sum_{k=j}^{h} \{\psi(k)\rho^k\}^{-1}, \quad i = j, \ldots, h;\ j = 0, 1, \ldots. \tag{2.6}$$

Proof Since $|Z(m+1)-Z(m)|=1$ with probability one for $m<N$, it is easily seen that for the model with finite and infinite waiting room, for $h \leq N$, $\{f_{h;i,j}\}_{h\geq 0,\ 0\leq j\leq i}$ is the solution of

$$\begin{aligned} f_{h;j,j} &= 1, \quad j=0,1,\ldots, \\ f_{h;i,j} &= p(i,i-1)f_{h;i-1,j}+p(i,i+1)f_{h;i+1,j}, \ i=j+1,\ldots,h; \ j=0,1,\ldots, \\ f_{h;i,j} &= 0, \quad i=h+1,h+2,\ldots, \end{aligned}$$

where for simplicity of notation we have introduced $p(N,N+1)=0$. □

Observe that each discrete time Markov chain with transition probabilities $p(n,n+1)=\alpha_n$, $p(n,n-1)=\beta_n$, where $\alpha_n+\beta_n=1$, can be modelled using $\psi(\cdot)$ and $\varphi(\cdot)$ as specified in (2.5). Therefore, the result of Proposition 2.1 is not restricted to processes $\mathbf{X}$ with negative exponentially distributed interarrival and service times.

The result of Proposition 2.1 allows us to determine the probability that customers are lost during a busy cycle for a queue with finite waiting room. This loss probability can be derived from the distribution of the maximum number of customers present in the queue during a busy cycle, as is illustrated below.

From $f_{h;i,j}$ the maximum number of customers in the queue can be determined. Denote by $\mathbf{Y}$ the maximum number of customers simultaneously present during a busy period. It follows that $\mathbf{P}(\mathbf{Y}\leq n)=f_{n;1,0}$, and thus

$$\mathbf{P}(\mathbf{Y}\leq n)=1-\left[\sum_{k=0}^{n}\{\psi(k)\rho^k\}^{-1}\right]^{-1}, \quad n=1,\ldots,N\leq\infty. \tag{2.7}$$

For $N<\infty$ the probability $\mathbf{P}(\mathbf{Y}\leq N)$ must be interpreted as the probability that during a busy cycle no customers are lost. For $N=\infty$ and $\mathbf{X}$ recurrent, from Foster's criterion we have that $\sum_{k=0}^{\infty}\{\psi(k)\rho^k\}^{-1}=\infty$. As a consequence, $\mathbf{P}(\mathbf{Y}\leq n)$ is an honest distribution. If $\mathbf{X}$ is transient, then $\lim_{n\to\infty}\mathbf{P}(\mathbf{Y}\leq n)<1$. This is obvious, since for transient $\mathbf{X}$ the number times the process returns to state n is finite with probability 1, for all n.

The probability that the maximum number of customers during a busy cycle is n given that the maximum number of customers in a busy cycle is at least n is given by

$$\mathbf{P}(\mathbf{Y}=n|\mathbf{Y}\geq n)=\{\psi(n)\rho^n\}^{-1}\Big/\sum_{k=0}^{n}\{\psi(k)\rho^k\}^{-1}. \tag{2.8}$$

Distributions of the form (2.8) are known as discrete failure rates in reliability theory. For Markov chain $\mathbf{X}$ the probability that an arriving customer meets n customers upon its arrival given that a customer meets at most n customers is defined when $\mathbf{X}$ is recurrent and is given by

$$\mathbf{P}_0(\mathbf{X} = n | \mathbf{X} \leq n) = \pi(n) \Big/ \sum_{k=0}^{n} \pi(k) = \psi(n)\rho^n \Big/ \sum_{k=0}^{n} \psi(k)\rho^k. \tag{2.9}$$

Comparison of (2.8) and (2.9) shows an interesting duality: we find $\psi(k)\rho^k$ in (2.9) where we have found $\{\psi(k)\rho^k\}^{-1}$ in (2.8). This form of duality was first observed for the Erlang model in Cohen [1971], and is of a form very similar to the dual or inverse relation between the M/G/1 queue and the GI/M/1 queue (Niu and Cooper [1989]), which is extended to a similar dual relation between the M/G/s/s queue and the GI/M/1/s/s queue in Kimura [1993]: The queue obtained from a queue by interchanging the interarrival and service-time distributions is called the *dual* or *inverse*. Therefore, if the equilibrium distribution of a queue contains the factor λ/μ then the equilibrium distribution of its dual contains the factor μ/λ. Note that in our case the inversion is not restricted to the factor λ/μ, but takes up the whole invariant measure.

From (2.8) and (2.9) we obtain that

$$\mathbf{P}_{\mathrm{M}(\lambda)/\mathrm{M}(\mu)/1}(\mathbf{Y} = n | \mathbf{Y} \geq n) = \mathbf{P}_{\mathrm{M}(\mu)/\mathrm{M}(\lambda)/1}(\mathbf{X} = n | \mathbf{X} \leq n),$$

where the non-stable M(λ)/M(μ)/1 queue is the dual of the stable M(μ)/M(λ)/1 queue. This shows that the dual relation between the M/G/1 and GI/M/1 queue goes beyond performance measures related to the forward Kolmogorov equations and equilibrium distribution.

Consider the system with finite waiting room. From the interpretation of $\mathbf{P}(\mathbf{Y} \leq N)$ we see that $\mathbf{P}(\mathbf{Y} = N | \mathbf{Y} \geq N)$ is the probability that during a busy cycle no customers are lost given that during this busy cycle the waiting room is full at least once. Thus $\mathbf{P}(\mathbf{Y} = N | \mathbf{Y} \geq N)$ is the probability that no customer meets congestion during a busy cycle given that at least once no vacancies exist. For this finite system the probability that an arriving customer finds the queue full and is lost equals $B(N, \rho) = \mathbf{P}_0(\mathbf{X} = N | \mathbf{X} \leq N)$, which is also the probability that a customer meets congestion during a busy cycle. This gives an interesting interpretation of the dual relation between (2.8) and (2.9).

3. Limit theorems: convergence in distribution

From the results above we can obtain limit theorems for the maximum number of customers simultaneously present in a queue. In this section we will first discuss the tail behaviour of the distribution of $\mathbf{Y}$. Then we apply the obtained results to characterise convergence of $\mathbf{P}((\mathbf{Y}^{(k)} - a_k)/b_k \leq x)$ for suitable normalisation constants $\{a_k > 0,\ b_k\}_{k \geq 1}$. In section 4 we will discuss convergence in probability and convergence with probability 1 for the sequence of partial maxima $\mathbf{Y}^{(k)}$.

Let $\mathbf{Y}_n$ be the maximum number of customers simultaneously present in the $n+1$st busy cycle after $t = 0$, $n \geq 0$, where we assume that the 1st busy cycle starts at $t = 0$. By the properties of $\mathbf{X}$, the busy cycles are independent and statistically identical, and therefore the random variables $\mathbf{Y}_n$ are iid, and all have the same distribution as $\mathbf{Y}$. Let $\mathbf{Y}^{(k)} = \max\{\mathbf{Y}_0, \ldots, \mathbf{Y}_{k-1}\}$, the maximum of the number of customers simultaneously present in k subsequent busy cycles. From (2.7) we obtain

$$\mathbf{P}(\mathbf{Y}^{(k)} \leq n) = \mathbf{P}(\mathbf{Y} \leq n)^k = \left(1 - \left[\sum_{k=0}^{n} \{\psi(k)\rho^k\}^{-1}\right]^{-1}\right)^k .$$

We want to investigate the behaviour of $\mathbf{P}((\mathbf{Y}^{(k)} - b_k)/a_k \leq x)$ for suitable normalisation constants $\{a_k > 0,\ b_k\}_{k \geq 1}$. To this end we will first investigate the tail behaviour of $\mathbf{P}(\mathbf{Y} \leq n)$. Then we will use the theory on extremal distributions to investigate the distribution of the normalised sample maxima. By Theorem 1.7.13 of Leadbetter *et al.* [1983], a necessary and sufficient condition for the sequence of partial maxima $\mathbf{Y}^{(k)}$ to possess normalisation constants leading to a non-degenerate limit is that $\lim_{x \to x_F} \mathbf{P}(\mathbf{Y} > x)/\mathbf{P}(\mathbf{Y} \geq x) = 1$, where x_F is the right endpoint of the support. This condition is not satisfied for most discrete valued distributions, such as the Poisson distribution (Leadbetter *et al.* [1983], Example 1.7.14), and the geometric distribution (Leadbetter *et al.* [1983], Example 1.7.15), distributions typically occurring when queues are involved. Anderson [1970], and Cohen [1982] show that asymptotic bounds for the limit of $\mathbf{P}((\mathbf{Y}^{(k)} - b_k)/a_k \leq x)$ can be found, and compute lim inf's and lim sup's. These lim inf's and lim sup's are sufficient to characterise weak convergence of the partial maxima for recurrent processes.

3.1 The tail behaviour of $\mathbf{P}(\mathbf{Y} \leq n)$

In the sequel it will be assumed that all arriving customers are accepted ($N = \infty$), i.e., that that $\psi(n) > 0$ for all $n \geq 0$. We will first investigate $\mathbf{P}(\mathbf{Y} \leq n)$ when $\mathbf{X}$ is recurrent. In particular, we will first investigate the case

$$B_\psi^{-1} = \sum_{n=0}^{\infty} \psi(n)\rho^n < \infty. \tag{3.1}$$

Then we will investigate the case on the boundary between transient and recurrent $\mathbf{X}$

$$\sum_{n=0}^{\infty} \psi(n)\rho^n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} \{\psi(n)\rho^n\}^{-1} = \infty. \tag{3.2}$$

Finally, we will discuss results for transient $\mathbf{X}$

$$\sum_{n=0}^{\infty} \{\psi(n)\rho^n\}^{-1} < \infty. \tag{3.3}$$

We will use the following notation:

$$\bar{\beta} = \limsup_{n\to\infty}(\psi(n+1)/\psi(n)),$$
$$\underline{\beta} = \liminf_{n\to\infty}(\psi(n+1)/\psi(n)),$$
$$\beta = \lim_{n\to\infty}(\psi(n+1)/\psi(n)) \quad \text{(if it exists)}.$$

Assumption (3.1) suggests that $\bar{\beta} < 1/\rho$. The following lemma characterises the limiting behaviour of $\mathbf{P}(\mathbf{Y} \le n)$ for general ψ and for two interesting special cases corresponding to the single server queue, and the infinite server queue.

Lemma 3.1 *Assume that $\bar{\beta} < 1/\rho$. Then*

$$(1-\rho\bar{\beta})/B_\psi \le \liminf_{n\to\infty} \mathbf{P}(\mathbf{Y} > n)/\pi(n) \le \limsup_{n\to\infty} \mathbf{P}(\mathbf{Y} > n)/\pi(n) \le (1-\rho\underline{\beta})/B_\psi . \quad (3.4)$$

If $\beta < 1/\rho$ exists, then

$$\lim_{n\to\infty} \pi(n)^{-1}[1-\mathbf{P}(\mathbf{Y} \le n)] = (1-\rho\beta)/B_\psi. \tag{3.5}$$

If $\psi(n) = \beta^n$, then $\pi(n) = (1-\rho\beta)(\rho\beta)^n$, and $\lim_{n\to\infty} \pi(n)^{-1}[1-\mathbf{P}(\mathbf{Y} \le n)] = 1$.
If $\psi(n) = 1/n!$, then $\pi(n) = \mathrm{e}^{-\rho}\rho^n/n!$, and $\lim_{n\to\infty} \pi(n)^{-1}[1-\mathbf{P}(\mathbf{Y} \le n)] = \mathrm{e}^{\rho}$.

Proof Define $T(n) = (1/\rho)^n\psi(n)^{-1}\mathbf{P}(Y > n)$. We have $0 < T(n) < \infty$ for $n \ge 0$, and

$$\frac{1}{T(n+1)} = \rho\frac{\psi(n+1)}{\psi(n)}\frac{1}{T(n)} + 1.$$

Taking limits gives since $\frac{\psi(n+1)}{\psi(n)} > 0$ for all n, and $\frac{1}{T(n)} > 0$ for all n

$$\begin{aligned}
\limsup_{n\to\infty}\frac{1}{T(n+1)} &= \rho\limsup_{n\to\infty}\frac{\psi(n+1)}{\psi(n)}\frac{1}{T(n)} + 1 \le \rho\bar{\beta}\limsup_{n\to\infty}\frac{1}{T(n)} + 1\\
\liminf_{n\to\infty}\frac{1}{T(n+1)} &= \rho\liminf_{n\to\infty}\frac{\psi(n+1)}{\psi(n)}\frac{1}{T(n)} + 1 \ge \rho\underline{\beta}\liminf_{n\to\infty}\frac{1}{T(n)} + 1
\end{aligned}$$

or equivalently, since $\bar{\beta}\rho < 1$

$$\frac{1}{(1-\underline{\beta}\rho)} \le \liminf_{n\to\infty}\frac{1}{T(n+1)} \le \limsup_{n\to\infty}\frac{1}{T(n+1)} \le \frac{1}{(1-\bar{\beta}\rho)},$$

yielding the first statement.

In the two special cases presented in the lemma, $\lim_{n\to\infty}(\psi(n+1)/\psi(n))$ exists. If $\psi(n) = \beta^n$ then $(\psi(n+1)/\psi(n)) = \beta$ for all n. Furthermore, $B_\psi = (1-\rho\beta)$ yielding the first special case. If $\psi(n) = 1/n!$ then $\lim_{n\to\infty}(\psi(n+1)/\psi(n)) = 0$. Since $B_\psi = \mathrm{e}^{-\rho}$ this gives the second special case. □

The result of Lemma 3.1 is that $\mathbf{P}(Y > n) \sim \pi(n)$ $(n \to \infty)$: the probability that the maximum number of customers simultaneously present in a busy cycle exceeds n equals the probability that an arriving customer meets n customers in the queue. If $N < \infty$ but large, this suggests that $\mathbf{P}(Y > N)$ can be used to approximate the loss probability.

Next, we investigate the limiting behaviour of $\mathbf{P}(\mathbf{Y} \leq n)$ in the case $\beta = 1/\rho$, the case in which the ratio test does not give a conclusion about the convergence or divergence of $\sum_{k=0}^{\infty}\{\psi(k)\rho^k\}^{-1}$. This conclusion depends on the detailed behaviour of $\psi(k)$ for large values of k. The following lemma characterises $\mathbf{P}(\mathbf{Y} \leq n)$ for polynomial functions $\psi(k)\rho^k$. Depending on the dominant power of the polynomial, the behaviour of $\mathbf{P}(\mathbf{Y} \leq n)$ for $n \to \infty$ changes completely.

Lemma 3.2 *Assume that there exist constants α, $0 < \alpha < \infty$, and p, $-\infty < p < \infty$, such that $\lim_{k\to\infty} \psi(k)\rho^k/k^p = \alpha$. Then, for all p, $-\infty < p < \infty$, there exists a constant $\gamma(p)$, $0 < \gamma(p) < \infty$, such that*

$$\lim_{n\to\infty} \delta(n,p)[1 - \mathbf{P}(\mathbf{Y} \leq n)] = \gamma(p), \tag{3.6}$$

where

$$\delta(n,p) = \begin{cases} n^{1-p}, & \text{if } p < 1, \\ \log n, & \text{if } p = 1, \\ 1, & \text{if } p > 1, \end{cases} \qquad \gamma(p) = \begin{cases} \alpha(1-p), & \text{if } p < 1, \\ \alpha, & \text{if } p = 1, \\ \left[\sum_{k=0}^{\infty}[\psi(k)\rho^k]^{-1}\right]^{-1}, & \text{if } p > 1. \end{cases} \tag{3.7}$$

Proof From the assumptions, for all $\epsilon > 0$, $\exists\ n(\epsilon)$ such that $(\alpha - \epsilon)k^p < \psi(k)\rho^k < (\alpha + \epsilon)k^p$ for all $k \geq n(\epsilon)$. Therefore, for all $n > n(\epsilon)$

$$\sum_{k=0}^{n(\epsilon)} \frac{1}{\psi(k)\rho^k} + \frac{1}{\alpha+\epsilon}\sum_{k=n(\epsilon)+1}^{n} \frac{1}{k^p} \leq [1 - \mathbf{P}(\mathbf{Y} \leq n)]^{-1} \leq \sum_{k=0}^{n(\epsilon)} \frac{1}{\psi(k)\rho^k} + \frac{1}{\alpha-\epsilon}\sum_{k=n(\epsilon)+1}^{n} \frac{1}{k^p}.$$

For $p > 1$, $\sum_k k^{-p}$ converges, and (3.6) holds true with $\gamma(p)^{-1} = \sum_{k=0}^{\infty}[\psi(k)\rho^k]^{-1}$, and $\delta(n,p) = 1$. For $p = 1$, $\sum_{k=1}^{n} k^{-p}$ behaves like $\log n$, and for $p < 1$, $\sum_{k=1}^{n} k^{-p}$ behaves like n^{1-p}, which can easily be checked via the integral test. □

The parameter p figuring in the rates of convergence, $\delta(n,p)$, of $1 - \mathbf{P}(\mathbf{Y} \leq n)$ also determines recurrence properties of $\mathbf{X}$. In particular, $\mathbf{X}$ is transient if and only if $p > 1$. Observe that the distinction between positive recurrence and null recurrence can be made on the basis of $\phi(\cdot)$ (and not of $\psi(\cdot)$) only. For $\phi \equiv \psi$ we have that $\mathbf{X}$ is ergodic if and only if $p < 1$.

Results similar to (3.6) can be obtained for other choices of the tail behaviour of $\psi(k)$, $k \to \infty$. Obvious choices are $\psi(k)\rho^k \sim k(\log k)^p$ and $\psi(k)\rho^k \sim k\log k(\log\log k)^p$ for which $\delta(n,p)$ is obtained from (3.7) by changing n into $\log n$ and n into $\log\log n$

respectively. The important observation made in the lemma above is that (3.6) is sensitive to the specific choice for ψ. This behaviour is completely different from the result of Lemma 3.1, where we found that $1 - \mathbf{P}(\mathbf{Y} \leq n) \sim \psi(n)\rho^n$.

If $\mathbf{X}$ is transient, then the tail behaviour of $\mathbf{P}(\mathbf{Y} > n)$ is identical to the result for $p > 1$ in Lemma 3.2. If we condition on $\mathbf{Y} < \infty$, that is if we only consider asymptotic relations for busy cycles in which the maximum number of customers present remains finite, then asymptotic relations similar to the results of Lemma 3.1 can be obtained. This is the purpose of the following lemma.

Lemma 3.3 *Assume that $\psi(n) > 0$ for all $n \geq 0$, and that $\beta > 1/\rho$ exists. Then*

$$\lim_{n\to\infty} \psi(n)\rho^n[1 - \mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} < \infty)] = [(\rho\beta - 1)B_*(B_* - 1)]^{-1}, \tag{3.8}$$

where $B_* = \sum_{k=0}^{\infty}\{\psi(k)\rho^k\}^{-1}$.

Proof Define $T(n) = \rho^n\psi(n)[1 - \mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} < \infty)]$. Then, $0 < T(n) < \infty$ for $n \geq 0$, and

$$T(n+1) = \rho\frac{\psi(n+1)}{\psi(n)}\frac{\sum_{k=0}^{n}\{\psi(k)\rho^k\}^{-1}}{\sum_{k=0}^{n+1}\{\psi(k)\rho^k\}^{-1}}T(n) - \frac{1}{(B_* - 1)\sum_{k=0}^{n+1}\{\psi(k)\rho^k\}^{-1}}.$$

Taking limits, and using that $\lim_{n\to\infty}\sum_{k=0}^{n}\{\psi(k)\rho^k\}^{-1}$ and $\lim_{n\to\infty}(\psi(n+1)/\psi(n))$ exist and are finite, completes the proof. □

Remark 3.4 *It is interesting to compare the result (3.5) for $\rho\beta < 1$ with the result (3.8) for $\rho\beta > 1$:*

$$\lim_{n\to\infty}[1 - \mathbf{P}(\mathbf{Y} \leq n)]/[B_\psi\psi(n)\rho^n] = (1 - \rho\beta)/B_\psi, \quad \rho\beta < 1,$$

$$\lim_{n\to\infty}[1 - \mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} < \infty)]/[1/(B_*\psi(n)\rho^n)] = 1/[(\rho\beta - 1)\mathbf{P}(\mathbf{Y} < \infty)B_*], \quad \rho\beta > 1.$$

For $\rho\beta < 1$, $B_\psi\psi(n)\rho^n$ is a probability distribution, and for $\rho\beta > 1$, $1/(B_*\psi(n)\rho^n)$ is a probability distribution. This shows that the behaviour of $\mathbf{Y}$ conditioned on $\mathbf{Y} < \infty$ for transient $\mathbf{X}$ is similar to the behaviour of $\mathbf{Y}$ for recurrent $\mathbf{X}$.

3.2 Sample maxima

Bounds for $\mathbf{P}(\mathbf{Y}^{(k)} \leq n)$ $(k \to \infty)$ in the case that $0 \leq \bar{\beta} < 1/\rho$ can be obtained from Lemma 3.1.

Theorem 3.5 *Assume that $0 \leq \bar{\beta} < 1/\rho$. Then there exists a continuous and decreasing function $f : [0,\infty) \to [0,\infty)$ such that for $-\infty < x < \infty$*

$$\begin{aligned}
\liminf_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\underline{\beta})k}) + 1) &\geq \mathrm{e}^{-\mathrm{e}^{-x}}, \\
\limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\bar{\beta})k})) &\leq \mathrm{e}^{-\mathrm{e}^{-x}}.
\end{aligned}$$

If in addition $\underline{\beta} > 0$, then for $-\infty < x < \infty$

$$\begin{aligned}
\liminf_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\underline{\beta})k})) &\geq \mathrm{e}^{-(\underline{\beta}\rho)^{-1}\,\mathrm{e}^{-x}} \\
\limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\bar{\beta})k})) &\leq \mathrm{e}^{-\mathrm{e}^{-x}}.
\end{aligned}$$

Proof Let $g : [0,\infty) \to [0,\infty)$ the function obtained from $\rho^n\psi(n)$ by linear interpolation, that is, for $y = \alpha n+(1-\alpha)(n+1)$ define $g(y) = \alpha g(n)+(1-\alpha)g(n+1)$. It can easily be shown that $\liminf_{y\to\infty}(g(y+1)/g(y)) = \underline{\beta}$, and $\limsup_{y\to\infty}(g(y+1)/g(y)) = \bar{\beta}$. This implies also that, for $\epsilon < 1-\rho\bar{\beta}$ there exists an y_0 such that for all $y > y_0$ $g(y+1) < (\bar{\beta}\rho+\epsilon)g(y)$, which implies that $g(y)$ is decreasing for $y > y_0$. If $y_0 = 0$ set $f \equiv g$. If $y_0 > 0$ set $f(y) = g(y)$ for $y \geq y_0$, and $f(y) = y_0 - y + g(y_0)$ for $0 \leq y \leq y_0$. Then f is a continuous and decreasing function, and f^{-1} is well-defined.

From (3.4) we obtain that for all $\epsilon > 0$, $\exists$ $n(\epsilon)$ such that for all $n \geq n(\epsilon)$

$$1-((1-\rho\underline{\beta})+\epsilon)\rho^n\psi(n) \leq \mathbf{P}(\mathbf{Y} \leq n) \leq 1-((1-\rho\bar{\beta})-\epsilon)\rho^n\psi(n). \tag{3.9}$$

For $\max\{y_0, n(\epsilon)\} < y-1 < n \leq y$ we have $\mathbf{P}(\mathbf{Y} \leq n) = \mathbf{P}(\mathbf{Y} \leq y)$, and

$$\begin{aligned}
1-((1-\rho\underline{\beta})+\epsilon)f(y-1) &\leq 1-((1-\rho\underline{\beta})+\epsilon)f(n) \leq \mathbf{P}(\mathbf{Y} \leq y) \\
&\leq 1-((1-\rho\bar{\beta})-\epsilon)f(n) \leq 1-((1-\rho\bar{\beta})-\epsilon)f(y).
\end{aligned}$$

For $\bar{\beta} = 0$, insertion of

$$y = f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\underline{\beta})k}), \quad \text{and} \quad y = f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\bar{\beta})k}), \tag{3.10}$$

in the left and right inequalities respectively, yields the first set of results.

If $\underline{\beta} > 0$ then for all $\underline{\beta}\rho > \epsilon > 0$, $\exists$ $y(\epsilon)$ such that

$$(\underline{\beta}\rho-\epsilon)g(y) < g(y+1) < (\bar{\beta}\rho+\epsilon)g(y), \tag{3.11}$$

for all $y \geq y(\epsilon)$, and the same relation holds for f for all $y \geq \max\{y_0, y(\epsilon)\}$. For $\max\{y_0, n(\epsilon), y(\epsilon)\} < y-1 \leq n < y$ we now obtain

$$1 - \{((1-\rho\underline{\beta}) + \epsilon)/(\underline{\beta}\rho - \epsilon)\}f(y) < \mathbf{P}(\mathbf{Y} \leq y) < 1 - ((1-\rho\bar{\beta}) - \epsilon)f(y)$$

Taking y as defined in (3.10) in the left and right inequalities, yields the second set of results. □

If β exists, and $0 < \beta < 1/\rho$, then from Anderson [1970], or the result above (via the observation that we may apply results for distributions with a continuous support in the proof above) we obtain that there exist norming constants $\{a_k > 0,\ b_k\}_{k\geq 1}$, a_k bounded, $b_k \to \infty$ as $k \to \infty$, such that for $-\infty < x < \infty$

$$\begin{aligned} \mathrm{e}^{-(\beta\rho)^{-1}\mathrm{e}^{-x}} &\leq \liminf_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq a_k x + b_k) \\ &\leq \limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq a_k x + b_k) \leq \mathrm{e}^{-\mathrm{e}^{-x}}. \end{aligned} \tag{3.12}$$

For the single server queue this result is also obtained in Cohen [1982], p. 623.

When $\bar{\beta} \neq \underline{\beta}$ (3.12) is not valid. This is the main reason for the fact that we have used $f^{-1}(\cdot)$ in the formulation of the theorem above. Norming constants exist, but the sequence of norming constants corresponding to $f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\underline{\beta})k})$ will be different from the sequence corresponding to $f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\bar{\beta})k})$. Furthermore, the lim inf's above are not necessarily smaller than the lim sup's due to the difference in the arguments.

The result (3.12) cannot be strengthened, that is $\lim_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq a_k x + b_k)$ does not exist. This can, for example, be obtained from Serfozo [1988], Theorem 2.3. (3.12) shows that every sequence has a subsequence with limit between the upper and lower bounds presented in (3.12) for fixed x. From the results on stochastic compactness of sample extremes (De Haan and Resnick [1984]) we can show that this limit in fact produces a non-degenerate distribution of the form $\mathrm{e}^{-\mathrm{e}^{-x}}$. This is a powerful result that completely characterises the limiting behaviour of $\mathbf{P}(\mathbf{Y}^{(k)} \leq a_k x + b_k)$ for $0 < \beta < 1/\rho$.

Definition 3.6 (De Haan and Resnick [1984]) *The sequence of sample maxima $Y^{(k)}$ is stochastically compact if there exist $\{a_n > 0,\ b_n\}_{n\geq 1}$ such that every sequence $\{(\mathbf{Y}^{n(k)} - b_{n(k)})/a_{n(k)}\}_{k\geq 1}$ contains a subsequence whose distributions converge weakly to a non-degenerate probability distribution. Such a limit distribution is called a partial limit distribution. We will also call the distribution function of* **Y** *stochastically compact if the above holds.*

Theorem 3.7 *For $0 < \beta\rho < 1$, $\mathbf{P}(\mathbf{Y} \leq y)$ is stochastically compact. The possible partial limit distributions are $G(x) = \exp[-\exp[-(x+\epsilon)]]$, $-\infty < x < \infty$, with $\log[\beta\rho] \leq \epsilon \leq 0$.*

Proof For $\beta > 0$ we will show that $\mathbf{P}(\mathbf{Y} \leq y)$ satisfies the conditions of Theorem 4 in De Haan and Resnick [1984]: for some $\delta > 1$,

$$\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1}dy<\infty, \tag{3.13}$$

$$\liminf_{x\to\infty}\frac{\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta}dy}{(1-\mathbf{P}(\mathbf{Y}\leq x))\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1}dy}>0, \tag{3.14}$$

$$\limsup_{x\to\infty}\frac{\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta}dy}{(1-\mathbf{P}(\mathbf{Y}\leq x))\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1}dy}<1. \tag{3.15}$$

Let $0<\epsilon<\min\{\rho\beta,(1-\rho\beta)\}$. From (3.5) there exists an $n(\epsilon)$ such that for all $n+1>y\geq n>n(\epsilon)$ we have $((1-\rho\beta)-\epsilon)\psi(n)\rho^n\leq 1-\mathbf{P}(\mathbf{Y}\leq n)\leq((1-\rho\beta)+\epsilon)\psi(n)\rho^n$, and $(\beta\rho-\epsilon)\psi(n)\rho^n<\psi(n+1)\rho^{n+1}<(\beta\rho+\epsilon)\psi(n)\rho^n$. As a consequence, for $\delta>2$

$$\begin{aligned}\int_{n(\epsilon)}^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1}dy &\leq \sum_{n=n(\epsilon)}^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq n))^{\delta-1}\\ &\leq \sum_{n=n(\epsilon)}^{\infty}\left[((1-\rho\beta)+\epsilon)(\beta\rho+\epsilon)^{n-n(\epsilon)}\psi(n(\epsilon))\rho^{n(\epsilon)}\right]^{\delta-1}<\infty.\end{aligned}$$

By using the lower bounds on $\psi(n)\rho^n$ and $1-\mathbf{P}(\mathbf{Y}\leq y)$ it can also be shown that the above bound holds for $0<\delta<1$, which shows that (3.13) is satisfied for all $\delta>1$.

For $x>n(\epsilon)$ we have, with $\lfloor x\rfloor$ the integer part of x, and $\lceil x\rceil$ the smallest integer not less than x (note that $\lceil x\rceil=\lfloor x\rfloor$ iff $x=\lfloor x\rfloor$), for $\delta>2$

$$\begin{aligned}&\frac{\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta}dy}{(1-\mathbf{P}(\mathbf{Y}\leq x))\int_x^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1}dy}=\\ &\quad=\frac{(x-\lfloor x\rfloor)(1-\mathbf{P}(\mathbf{Y}\leq\lfloor x\rfloor))^{\delta}+\sum_{n=\lceil x\rceil}^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq n))^{\delta}}{(1-\mathbf{P}(\mathbf{Y}\leq\lfloor x\rfloor))\left\{(x-\lfloor x\rfloor)(1-\mathbf{P}(\mathbf{Y}\leq\lfloor x\rfloor))^{\delta-1}+\sum_{n=\lceil x\rceil}^{\infty}(1-\mathbf{P}(\mathbf{Y}\leq n))^{\delta-1}\right\}}\\ &\quad\leq\frac{\left[((1-\rho\beta)+\epsilon)(\beta\rho+\epsilon)^{\lceil x\rceil-\lfloor x\rfloor}\psi(\lfloor x\rfloor)\rho^{\lfloor x\rfloor}\right]^{\delta}\left(\frac{(x-\lfloor x\rfloor)}{(\beta\rho+\epsilon)^{\delta}}+\frac{1}{1-(\beta\rho+\epsilon)^{\delta}}\right)}{[((1-\rho\beta)-\epsilon)(\beta\rho-\epsilon)^{\lceil x\rceil-\lfloor x\rfloor}\psi(\lfloor x\rfloor)\rho^{\lfloor x\rfloor}]^{\delta}\left(\frac{(x-\lfloor x\rfloor)}{(\beta\rho-\epsilon)^{\delta}}+\frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^{\delta}}\right)}\\ &\quad\leq\frac{[((1-\rho\beta)+\epsilon)(\beta\rho+\epsilon)]^{\delta}\left(\frac{(x-\lfloor x\rfloor)}{(\beta\rho+\epsilon)^{\delta}}+\frac{1}{1-(\beta\rho+\epsilon)^{\delta}}\right)}{[((1-\rho\beta)-\epsilon)(\beta\rho-\epsilon)]^{\delta}\left(\frac{(x-\lfloor x\rfloor)}{(\beta\rho-\epsilon)^{\delta}}+\frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^{\delta}}\right)}.\end{aligned}$$

Observe that the right hand side is a periodic function of x. It therefore suffices to bound the expression for $0\leq x\leq 1$.

For $0\leq x\leq 1$ let $r(x)$ denote the x-dependent part of the right hand side above, that is

$$r(x)=\frac{\frac{x}{(\beta\rho+\epsilon)^{\delta}}+\frac{1}{1-(\beta\rho+\epsilon)^{\delta}}}{\frac{x}{(\beta\rho-\epsilon)^{\delta}}+\frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^{\delta}}}=\left(\frac{\beta\rho-\epsilon}{\beta\rho+\epsilon}\right)^{\delta}+\frac{\frac{1}{1-(\beta\rho+\epsilon)^{\delta}}-\left(\frac{\beta\rho-\epsilon}{\beta\rho+\epsilon}\right)^{\delta}\frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^{\delta}}}{\frac{x}{(\beta\rho-\epsilon)^{\delta}}+\frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^{\delta}}}.$$

For $2 < \delta < 1 + \log[1/(\beta\rho+\epsilon)] / \log[(\beta\rho+\epsilon)/(\beta\rho-\epsilon)]$, which is possible for ϵ sufficiently small (for $\epsilon \to 0$ the right hand side goes to ∞), the denominator of the second term, $\frac{1}{1-(\beta\rho+\epsilon)^\delta} - \left(\frac{\beta\rho-\epsilon}{\beta\rho+\epsilon}\right)^\delta \frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^\delta}$ is negative, and $r(x)$ is an increasing function on $[0,1]$. As a consequence, $\max_{0\leq x\leq 1} r(x) = r(1)$.

From the results above,

$$\limsup_{x\to\infty} \frac{\int_x^\infty (1-\mathbf{P}(\mathbf{Y}\leq y))^\delta dy}{(1-\mathbf{P}(\mathbf{Y}\leq x))\int_x^\infty (1-\mathbf{P}(\mathbf{Y}\leq y))^{\delta-1} dy} \leq \limsup_{x\to\infty} \frac{[((1-\rho\beta)+\epsilon)(\beta\rho+\epsilon)]^\delta}{[((1-\rho\beta)\epsilon)(\beta\rho-\epsilon)]^\delta} r(x)$$

$$= \frac{[((1-\rho\beta)+\epsilon)(\beta\rho+\epsilon)]^\delta}{[((1-\rho\beta)-\epsilon)(\beta\rho-\epsilon)]^\delta} \left\{ \left(\frac{\beta\rho-\epsilon}{\beta\rho+\epsilon}\right)^\delta + \frac{\frac{1}{1-(\beta\rho+\epsilon)^\delta} - \left(\frac{\beta\rho-\epsilon}{\beta\rho+\epsilon}\right)^\delta \frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^\delta}}{\frac{1}{(\beta\rho-\epsilon)^\delta} + \frac{1}{(\beta\rho-\epsilon)-(\beta\rho-\epsilon)^\delta}} \right\}$$

For ϵ sufficiently small the right hand side is strictly smaller than 1, which can easily be checked by taking the limit $\epsilon \to 0$.

Using similar arguments, it can also be shown that (3.15) is satisfied for all $\delta > 1$, and that (3.14) is satisfied for all $\delta > 1$, which completes the proof. □

For $\beta = 0$, corresponding to a non-uniformizable Markov chain $\mathbf{X}$, the situation is completely different. From Lemma 3.1 we obtain

$$\lim_{n\to\infty} \frac{1-\mathbf{P}(\mathbf{Y}\leq n-1)}{1-\mathbf{P}(\mathbf{Y}\leq n)} = \lim_{n\to\infty} \frac{\pi(n-1)}{\pi(n)} = \infty,$$

and Corollary 4 of De Haan and Resnick [1984] shows that $\mathbf{P}(\mathbf{Y}\leq y)$ is *not* stochastically compact. As a consequence, although we have obtained bounds on the limiting distributions, these bounds do not allow us to characterise the limiting behaviour of $\mathbf{P}(\mathbf{Y}^{(k)} \leq a_k x + b_k)$ for $\beta = 0$.

For $\beta = 1/\rho$ the distribution of $\mathbf{Y}^{(k)}$ can be obtained from Lemma 3.2.

Theorem 3.8 *Under the assumptions of Lemma 3.2, for $p < 1$,*

$$\lim_{k\to\infty} \mathbf{P}\left(\frac{\mathbf{Y}^{(k)}}{(k\alpha(1-p))^{1-p}} \leq x\right) = \begin{cases} \mathrm{e}^{-x^{-(1-p)}}, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{3.16}$$

and for $p = 1$

$$\lim_{k\to\infty} \mathbf{P}\left(\frac{\log \mathbf{Y}^{(k)}}{\alpha k} \leq x\right) = \begin{cases} \mathrm{e}^{-x^{-1}}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{3.17}$$

Proof For $p < 1$, from (3.6) we obtain that for all $0 < \epsilon < \alpha(1-p)$, $\exists\, y(\epsilon)$ such that for all $y \geq y(\epsilon)$

$$1 - \frac{\alpha(1-p)+\epsilon}{(y-1)^{1-p}} < \mathbf{P}(\mathbf{Y} \le y) < 1 - \frac{\alpha(1-p)-\epsilon}{y^{1-p}}.$$

We may set $y = (k\alpha(1-p))^{1/(1-p)}x$ for $x > 0$, $k = 1, 2, \ldots$, which proves (3.16). The proof of (3.17) follows the same lines. □

For the single server queue, corresponding to $p = 0$, the result of Theorem 3.8 is obtained by Cohen [1982]. Extensions of this result for $p < 1$ are presented in Serfozo [1988]. Here limit theorems are obtained for null recurrent, and almost null recurrent processes. In this reference the birth and death rates depend not only on the number of customers in the queue, but also on the number of busy cycles that have past, that is in busy cycle k we have $q(n, n+1) = \lambda_{n,k}$, and $q(n, n-1) = \mu_{n,k}$. Serfozo [1988] defines parameters ρ_k, related to the traffic intensity, and obtains limit results when $\rho_k \to 1$ along a suitable sequence.

For $\beta > 1/\rho$ limit theorems can be obtained for $\mathbf{Y}^{(k)}$ conditioned on $\mathbf{Y}^{(k)} < \infty$. In view of Remark 3.4, it is not surprising that a result similar to Theorem 3.7 can be derived for $\beta\rho > 1$.

Theorem 3.9 *For* $1 < \beta\rho < \infty$, $\mathbf{P}(\mathbf{Y} \le y|\mathbf{Y} < \infty)$ *is stochastically compact. The possible partial limit distributions are* $G(x) = \exp[-\exp[-(x+\epsilon)]]$, $-\infty < x < \infty$, *with* $-\log[\beta\rho] \le \epsilon \le 0$.

Proof Follows the lines of the proof of Theorem 3.7, and is based on the inequalities $(\frac{1}{\rho\beta} - \epsilon)\frac{1}{\psi(n)\rho^n} < \frac{1}{\psi(n+1)\rho^{n+1}} < (\frac{1}{\rho\beta} + \epsilon)\frac{1}{\psi(n)\rho^n}$, and $([(\rho\beta - 1)B_*(B_* - 1)]^{-1} - \epsilon)\frac{1}{\psi(n)\rho^n} < 1 - \mathbf{P}(\mathbf{Y} \le y|\mathbf{Y} < \infty) < ([(\rho\beta-1)B_*(B_*-1)]^{-1} + \epsilon)\frac{1}{\psi(n)\rho^n}$. □

For $\beta = \infty$ we obtain from Corollary 4 of De Haan and Resnick [1984] that $\mathbf{P}(\mathbf{Y} \le y|\mathbf{Y} < \infty)$ is not stochastically compact, similar to the result for $\beta = 0$.

4. Limit theorems: almost sure convergence

This section records some results related to convergence in probability for the sample extremes. These results can be obtained as corollaries to the results of the previous section. Furthermore, almost sure convergence of the normalised maxima is investigated. It is shown that for $0 < \beta < 1/\rho$, $\mathbf{Y}^{(k)}/b_k \to 1$ with probability 1.

For $\beta > 0$ we obtain the asymptotic relation $\psi(n+k)/\psi(n) \sim \beta^k$ $(n \to \infty)$, which shows that the dominant behaviour of $\psi(n)$ for large n is geometrical. However, for ψ to have a geometric tail, that is for $\psi(n) \sim \beta^n$ $(n \to \infty)$ additional assumptions on ψ are required. From Theorem 3.5 and the discussion following that theorem, we obtain that the tail behaviour of f^{-1} is of interest ($f(n) = \psi(n)\rho^n$). Suggested by this discussion we now assume that for all y, $-\infty < y < \infty$, there exist norming constants $\{a_k > 0,\ b_k\}_{k \ge 1}$, a_k bounded, and $\lim_{k\to\infty} b_k = \infty$, such that

$$\lim_{k\to\infty} f^{-1}(\frac{e^{-y}}{k})/(a_k y + b_k) = 1. \tag{4.1}$$

The relation (4.1) can be seen as the asymptotic characterisation of the inverse of f. The assumption (4.1) holds for all functions f for which $F(y) = 1 - f(y)$ is in the domain of attraction of $\exp[-\mathrm{e}^{-x}]$. In particular, (4.1) is satisfied if $\beta > 0$, but is also satisfied for the infinite server queue, as is shown in the examples. The increasing sequence $\{b_k\}_{k=1}^{\infty}$ can be chosen as

$$b_k = \inf\{x : f(x) \leq \frac{1}{k}\} = f^{-1}(\frac{1}{k}), \tag{4.2}$$

where the first equality gives the definition of b_k, and the last equality is obtained by the continuity of f.

The following result is a corollary to Theorem 3.5. Its proof follows standard lines such as presented in Galambos [1987], section 4.1, but deviates from these lines because the arguments in the upper and lower bounds presented in Theorem 3.5 are not identical.

Corollary 4.1 *Assume that $0 \leq \bar{\beta} < 1/\rho$. Under the additional assumption (4.1), the variable $\mathbf{Y}^{(k)}/b_k$ converges for $k \to \infty$ in probability to 1.*

Proof From (4.1) for all y, $-\infty < y < \infty$, for all $\delta > 0$, $\exists\, k(\delta)$ such that $(1-\delta)(a_k y + b_k) < f^{-1}(\frac{\mathrm{e}^{-x}}{k}) < (1+\delta)(a_k y + b_k)$ for all $k \geq k(\delta)$. From Theorem 3.5 we obtain, for $-\infty < x < \infty$

$$\begin{aligned}
&\limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq (1-\delta)(a_k x + a_k \log[1-\rho\bar{\beta}] + b_k)) \\
\leq\ &\limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq f^{-1}(\frac{\mathrm{e}^{-x}}{(1-\rho\bar{\beta})k})) \leq \mathrm{e}^{-\mathrm{e}^{-x}},
\end{aligned}$$

and similarly, for $-\infty < x < \infty$

$$\liminf_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq (1+\delta)(a_k x + a_k \log[1-\rho\underline{\beta}] + b_k) + 1) \geq \mathrm{e}^{-\mathrm{e}^{-x}}.$$

This gives, for $\delta < 1$, $-\infty < x < \infty$, since $b_k > 0$

$$\begin{aligned}
&\limsup_{k\to\infty} \mathbf{P}(\left|\frac{\mathbf{Y}^{(k)}}{b_k} - 1\right| > (1+\delta)\frac{a_k}{b_k}x + (1+\delta)\frac{a_k}{b_k}\log[1-\rho\underline{\beta}] + \frac{1}{b_k} + \delta) \\
\leq\ &\limsup_{k\to\infty} \mathbf{P}(\frac{\mathbf{Y}^{(k)}}{b_k} - 1 \leq -(1+\delta)\frac{a_k}{b_k}x - (1+\delta)\frac{a_k}{b_k}\log[1-\rho\underline{\beta}] - \frac{1}{b_k} - \delta) \\
&\quad + 1 - \liminf_{k\to\infty} \mathbf{P}(\frac{\mathbf{Y}^{(k)}}{b_k} - 1 \leq (1+\delta)\frac{a_k}{b_k}x + (1+\delta)\frac{a_k}{b_k}\log[1-\rho\underline{\beta}] + \frac{1}{b_k} + \delta) \\
\leq\ &\limsup_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq -(1+\delta)a_k x - (1+\delta)a_k\log[1-\rho\bar{\beta}] + (1-\delta)b_k) \\
&\quad + 1 - \liminf_{k\to\infty} \mathbf{P}(\mathbf{Y}^{(k)} \leq (1+\delta)a_k x + (1+\delta)a_k\log[1-\rho\underline{\beta}] + (1+\delta)b_k + 1) \\
\leq\ &\mathrm{e}^{-\mathrm{e}^{\frac{(1+\delta)}{(1-\delta)}x+c}} + 1 - \mathrm{e}^{-\mathrm{e}^{-x}},
\end{aligned}$$

where $c = 2\log[1-\rho\bar{\beta}]/(1-\delta)$.

For fixed $0 < \epsilon < 1$, $\exists$ $K(\epsilon)$ such that $1/b_k < \epsilon/4$, for all $k \geq K(\epsilon)$. Let $\delta < \epsilon/4$. Since $a_k/b_k \to 0$ for all K there exists a $K_0 > K$ such that for all $k > K_0$ we have $a_k/b_k < a_K/b_K$. With $x = [(1-\delta)a_K/b_K]^{-1}\epsilon/2$,

$$\begin{aligned}
&\limsup_{k\to\infty} \mathbf{P}\left(\left|\frac{\mathbf{Y}^{(k)}}{b_k} - 1\right| > \epsilon\right) \\
&\quad \leq \limsup_{k\to\infty} \mathbf{P}\left(\left|\frac{\mathbf{Y}^{(k)}}{b_k} - 1\right| > (1+\delta)\frac{a_k}{b_k}[(1+\delta)a_K/b_K]^{-1}\epsilon/2 + (1+\delta)\frac{a_k}{b_k}\log[1-\rho\underline{\beta}] + \frac{1}{b_k} + \delta\right) \\
&\quad \leq \mathrm{e}^{-\mathrm{e}^{[(1-\delta)a_K/b_K]^{-1}\epsilon/2+c}} + 1 - \mathrm{e}^{-\mathrm{e}^{-[(1+\delta)a_K/b_K]^{-1}\epsilon/2}}.
\end{aligned}$$

For $K \to \infty$ the right-hand side converges to 0, which completes the proof. □

For $\beta\rho = 1$ and $\beta\rho > 1$ limit theorems for $\mathbf{Y}^{(k)}$ cannot be obtained from the results of section 3 based on extremal values. In these cases limit theorems agree with ordinary limit theorems since $\mathbf{X}$ is growing. Such results can be found in Iglehart and Whitt [1970]. For $\beta\rho > 1$ limit theorems for $\mathbf{Y}^{(k)}$ conditional on $\mathbf{Y}^{(k)} < \infty$ can be obtained from Theorem 3.9 by analogy with the result of Corollary 4.1.

If we assume that $0 < \beta < 1/\rho$ the result of Corollary 4.1 can be strengthened. The following theorem shows that $\mathbf{Y}^{(k)}/b_k$ converges for $k \to \infty$ to 1 with probability 1. The proof of this result is based on Theorem 4.4.4 on page 268 of Galambos [1987]. Observe that our definition of b_k differs from the definition used by Galambos. Therefore, we cannot directly apply Theorem 4.4.4, however, the proof of the result below follows the same lines.

Theorem 4.2 *Assume that $0 < \beta < 1/\rho$. Let $b_k = \log[k]/\log[1/(\rho\beta)]$. Then*

$$\mathbf{P}\left(\lim_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1\right) = 1.$$

Proof For $n \leq 1$, and $\epsilon < \beta\rho - (\beta\rho)^{1/n}$, let $n(\epsilon)$ such that $1 - \mathbf{P}(\mathbf{Y} \leq nb_k) \geq ((1-\rho\beta)-\epsilon)f(\lfloor nb_k \rfloor)$ for all $k \geq n(\epsilon)$, and $(\beta\rho-\epsilon)f(\lfloor nb_k \rfloor) < f(\lfloor nb_k \rfloor + 1)$ for all $k \geq n(\epsilon)$. We obtain

$$\begin{aligned}
\sum_{k=n(\epsilon)}^{\infty} (1 - \mathbf{P}(\mathbf{Y} \leq nb_k)) &\geq \sum_{k=n(\epsilon)}^{\infty} ((1-\rho\beta)-\epsilon)f(\lfloor nb_k \rfloor) \\
&\geq ((1-\rho\beta)-\epsilon) \sum_{k=n(\epsilon)}^{\infty} (\beta\rho-\epsilon)^{\lfloor nb_k \rfloor - n(\epsilon)} f(n(\epsilon)) \\
&\geq \frac{f(n(\epsilon))}{(\beta\rho-\epsilon)^{n(\epsilon)}} \sum_{k=n(\epsilon)}^{\infty} \left(\frac{1}{k}\right)^{n\frac{\log[\beta\rho-\epsilon]}{\log[\beta\rho]}},
\end{aligned}$$

showing that $\sum_{k=1}^{\infty}(1-\mathbf{P}(\mathbf{Y} \leq nb_k))$ diverges for all $n \leq 1$.

For $n > 1$, and $\epsilon < (\beta\rho)^{1/n} - \beta\rho$, let $n(\epsilon)$ such that $1 - \mathbf{P}(\mathbf{Y} \leq nb_k) \leq ((1-\rho\beta)+\epsilon)f(\lfloor nb_k \rfloor)$ for all $k \geq n(\epsilon)$, and $(\beta\rho+\epsilon)f(\lfloor nb_k \rfloor) > f(\lfloor nb_k \rfloor + 1)$ for all $k \geq n(\epsilon)$. We obtain

$$\begin{aligned}\sum_{k=n(\epsilon)}^{\infty} (1-\mathbf{P}(\mathbf{Y} \leq nb_k)) &\leq \sum_{k=n(\epsilon)}^{\infty} ((1-\rho\beta)+\epsilon)f(\lfloor nb_k \rfloor) \\ &\leq \frac{f(n(\epsilon))}{(\beta\rho+\epsilon)^{n(\epsilon)+1}} \sum_{k=n(\epsilon)}^{\infty} \left(\frac{1}{k}\right)^{n\frac{\log[\beta\rho+\epsilon]}{\log[\beta\rho]}},\end{aligned}$$

which shows that $\sum_{k=1}^{\infty}(1-\mathbf{P}(\mathbf{Y} \leq nb_k))$ converges for all $n > 1$. As a consequence, Theorem 4.4.1 on p. 263 of Galambos [1987] implies that $\mathbf{P}\left(\limsup_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1\right) = 1$.

We now apply Lemma 4.3.3 on page 255 of Galambos [1987] with $u_k = nb_k$, $0 < n < 1$. For the application of this lemma, implying that $\mathbf{P}\left(\liminf_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1\right) = 1$, it is sufficient to prove that

$$\sum_{k=1}^{\infty}[1-\mathbf{P}(\mathbf{Y} \leq nb_k)]\exp\{-k[1-\mathbf{P}(\mathbf{Y} \leq nb_k)]\} < \infty$$

for each fixed $n < 1$. Application of the bounds on the distribution used above gives for $\epsilon < \beta\rho - (\beta\rho)^{1/n}$, and $n(\epsilon)$ sufficiently large (see above)

$$\begin{aligned}&\sum_{k=n(\epsilon)}^{\infty} [1-\mathbf{P}(\mathbf{Y} \leq nb_k)]\exp\{-k[1-\mathbf{P}(\mathbf{Y} \leq nb_k)]\} = \\ &\leq \sum_{k=n(\epsilon)}^{\infty} \frac{f(n(\epsilon))}{(\beta\rho+\epsilon)^{n(\epsilon)+1}} \left(\frac{1}{k}\right)^{n\frac{\log[\beta\rho+\epsilon]}{\log[\beta\rho]}} \exp\left\{-k\, \frac{f(n(\epsilon))}{(\beta\rho-\epsilon)^{n(\epsilon)}} \left(\frac{1}{k}\right)^{n\frac{\log[\beta\rho-\epsilon]}{\log[\beta\rho]}}\right\},\end{aligned}$$

which is finite, since $n\frac{\log[\beta\rho-\epsilon]}{\log[\beta\rho]} < 1$. □

The first part of the proof, showing that $\mathbf{P}\left(\limsup_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1\right) = 1$, cannot be extended to the case $0 < \underline{\beta} < \bar{\beta}$. If $\underline{\beta} < \bar{\beta}$ we obtain the following bounds

$$\mathbf{P}\left(\limsup_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{\underline{b}_k} \geq 1\right) = \mathbf{P}\left(\limsup_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{\bar{b}_k} \leq 1\right) = 1,$$

where $\underline{b}_k = \log[k]/\log[1/(\rho\underline{\beta})] < \log[k]/\log[1/(\rho\bar{\beta})] = \bar{b}_k$. If $0 = \underline{\beta} < \bar{\beta}$ only the second bound (involving $\bar{b}_k$) can be concluded from the proof above, and if $\bar{\beta} = 0$ we cannot apply any of the parts of the proof above. The problem involved is similar to the problem encountered in the proof of Theorem 3.5. To obtain results for the convergence of $\mathbf{Y}^{(k)}$ we need the specific form of $f(\cdot)$. The following result considers the logical extension of the geometric case $f(x) \sim \rho^x$ to a queue with unbounded service rates: $f(x) \sim (\rho/x)^x$, which approximates the Poisson distribution. This result includes the infinite server queue.

Theorem 4.3 *For $\alpha > 0$, $\gamma > 0$ define $f(x) = \alpha\gamma^x(1/x)^{x+\delta}$, $x > 0$, and assume that $\mathbf{P}(\mathbf{Y} > n) \sim f(n)$ $(n \to \infty)$, then*

$$\lim_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1, \quad \text{a.s.},$$

where $b_k = f^{-1}(1/k)$, $k \geq 1$.

Proof The proof follows the same lines as the proof of Theorem 4.2.

For $n < 1$, and $n(\epsilon)$ sufficiently large

$$\sum_{k=n(\epsilon)}^{\infty} (1 - \mathbf{P}(\mathbf{Y} \leq nb_k)) \geq (1-\epsilon) \sum_{k=n(\epsilon)}^{\infty} f(b_k) = \infty,$$

by the definition of b_k. For $n > 1$, $n(\epsilon)$ sufficiently large

$$\begin{aligned} \sum_{k=n(\epsilon)}^{\infty} (1 - \mathbf{P}(\mathbf{Y} \leq nb_k)) &\leq (1+\epsilon) \sum_{k=n(\epsilon)}^{\infty} f(nb_k - 1) \\ &= (1+\epsilon) \sum_{k=n(\epsilon)}^{\infty} \alpha\gamma \left(\frac{1}{k}\right)^n \exp\left[-(nb_k + \delta - 1)\log[n - (1/b_k)] + \log[b_k]\right] < \infty, \end{aligned}$$

which shows that $\limsup_{k\to\infty} \frac{\mathbf{Y}^{(k)}}{b_k} = 1$ almost surely.

For $0 < n < 1$ and $n(\epsilon)$ sufficiently large we have for $k \geq n(\epsilon)$ that $b_k \leq \log[k] \leq k$, $0 < \frac{1}{2}n < n - (1/b_k) < n < 1$. This gives

$$\begin{aligned} f(nb_k - 1) &= \alpha\gamma \left(\frac{1}{k}\right)^n \exp\left[-nb_k \log[n - (1/b_k)] + \log[b_k] + (1-\delta)\log[n - (1/b_k)]\right] \\ &\leq \alpha\gamma \left(\frac{1}{k}\right)^n \exp\left[-n\log[2/n]\log[k] + \log[k] + |(\delta - 1)\log[2/n]|\right], \end{aligned}$$

and

$$f(nb_k) = \alpha \left(\frac{1}{k}\right)^n \exp\left[-n(b_k + \delta)\log[b_k] - (nb_k + \delta)\log[nb_k]\right] \geq \alpha \left(\frac{1}{k}\right)^n.$$

As a consequence, with C_1 and C_2 positive constants incorporating the constants appearing above

$$[1 - \mathbf{P}(\mathbf{Y} \leq nb_k)] \exp\{-k[1 - \mathbf{P}(\mathbf{Y} \leq nb_k)]\} \leq C_1 \left(\frac{1}{k}\right)^{n + n\log[2/n] - 1} \exp[-C_2 k^{1-n}].$$

Lemma 4.3.3 on page 255 of Galambos [1987] completes the proof. □

5. Examples

Below we present some examples from queueing theory. In particular, in the first example we will discuss results for the single server queue, a model that combines a relatively simple structure with the property that the limiting behaviour is representative for uniformizable queues (queues with a uniform bound on the service speed). Results can be found in Cohen [1982], and the references therein. The second example presents some applications of the results of this paper to the M/M/s queue. Finally, the third example analyses the infinite server queue, the standard example of a non-uniformizable queue.

5.1 The single server queue

The single server queue with Poisson arrivals and negative exponential services is a birth-and-death process with birth rates $q(n, n+1) = \lambda$, and death rates $q(n+1, n) = \mu$, $n = 0, 1, \ldots, N-1 \leq \infty$, where N is the size of the waiting room. (We set $\beta = 1$.) The invariant measure $m(n) = \rho^n$, $n = 0, \ldots, N$, where $\rho = \lambda/\mu$, equals the invariant measure at arrival instants. We have

$$\pi(n) = \pi_q(n) = \begin{cases} (1-\rho)\rho^n, & n = 0,1,2,\ldots, \quad (N=\infty,\ \rho<1), \\ (1-\rho)\rho^n/(1-\rho^{N+1}), & n = 0,\ldots,N, \quad (N<\infty,\ \rho\neq 1), \\ 1/(N+1), & n = 0,\ldots,N, \quad (N<\infty,\ \rho=1), \end{cases}$$

where arrivals to a full queue are discarded.

The maximum number of customers simultaneously present in a busy cycle is obtained from (2.7)

$$\mathbf{P}(\mathbf{Y} \leq n) = \begin{cases} 1 - \frac{1-(1/\rho)}{1-(1/\rho)^{n+1}}, & \rho \neq 1, \\ 1 - \frac{1}{n+1}, & \rho = 1, \end{cases}$$

which can be expanded for $\rho < 1$

$$\mathbf{P}(\mathbf{Y} \leq n) = 1 - \frac{\pi(n)}{1-\rho^{n+1}} = 1 - \pi(n)(1 + O((\rho)^{n+1})). \tag{5.1}$$

This immediately establishes the result of Lemma 3.1. For $y - 1 \leq n < y$,

$$1 - (1-\rho)\rho^{y-1}(1 + O((\rho)^y)) \leq \mathbf{P}(\mathbf{Y} \leq y) < 1 - (1-\rho)\rho^y(1 + O((\rho)^y)).$$

Taking $y = -(x + \log[k(1-\rho)])/\log[\rho]$ reproduces the result of Theorem 3.5 with $f^{-1}(\frac{e^{-x}}{(1-\rho)k}) = -(x+\log[k(1-\rho)])/\log[\rho]$. Corollary 4.1 holds with $b_k = -\log[k]/\log[\rho]$, which indicates that $\mathbf{Y}^{(k)}$ increases logarithmically for the single server queue. Theorem 4.2 shows that $\mathbf{Y}^{(k)}/\log[k]$ converges almost surely to $\log[1/\rho]$.

For $\rho \neq 1$ we obtain the conditional probability (for $N < \infty$)

$$\mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} \leq N) = 1 - \frac{(\hat{\rho}^n - \hat{\rho}^N)(1-\hat{\rho})}{(1-\hat{\rho}^N)(1-\hat{\rho}^{n+1})}, \tag{5.2}$$

where $\hat{\rho} = 1/\rho$. If the queue is unstable ($\rho > 1$) we may take the limit $N \to \infty$, which gives

$$\mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} < \infty) = 1 - \frac{\hat{\pi}(n)}{1-\hat{\rho}^{n+1}} = 1 - \hat{\pi}(n)(1 + O((\hat{\rho})^{n+1})),$$

where $\hat{\pi}(n) = (1-\hat{\rho})\hat{\rho}^n$. Comparison with (5.1) shows that we may now draw similar conclusions as in the case $\rho < 1$.

A perhaps more interesting application of (5.2) is the following result that allows us to characterise the behaviour when the buffer is almost full:

$$\frac{1 - \mathbf{P}(\mathbf{Y} \leq n|\mathbf{Y} \leq N)}{\hat{\rho}^n - \hat{\rho}^N} = \frac{(1-\hat{\rho})}{(1-\hat{\rho}^N)(1-\hat{\rho}^{n+1})} \to (1-\hat{\rho})$$

when $N \to \infty$, and $n \to N$. For example, we have $\hat{\rho}^{N-1}\mathbf{P}(\mathbf{Y} = N|\mathbf{Y} \leq N)] \to 1$ for $N \to \infty$, and we may apply the theorems obtained in this paper to this distribution.

5.2 The multiserver queue

The M/M/s queue is a birth-and-death process with birth rates $q(n, n+1) = \lambda$ and death rates $q(n, n-1) = \mu \min\{n, s\}$, that is customers depart at rate $n\mu$ when less than s customers are present, and at rate $s\mu$ when s or more customers are present, where s is the number of servers of the queue. For the limit theorems derived in this paper, only the behaviour of the process for large values of the queue is of interest. For these values the multiserver queue behaves similar to the single server queue. The queue is positive recurrent when $\lambda/s\mu < 1$. In the formalism of this paper: $\rho = \lambda/\mu$, and $\beta = 1/s$. For $\beta\rho < 1$ we have

$$\pi(n) = \pi_q(n) = \begin{cases} B_\psi \frac{\rho^n}{n!}, & n \leq s, \\ B_\psi \frac{\rho^s}{s!} (\beta\rho)^{n-s}, & n > s. \end{cases}$$

Lemma 3.1 shows that

$$\lim_{n\to\infty} \mathbf{P}(\mathbf{Y} > n)/(\lambda/s\mu)^n = (1 - \frac{\lambda}{s\mu})\frac{s!}{s^s},$$

and Theorem 3.7 implies that the sequence of sample maxima $\mathbf{Y}^{(k)}$ is stochastically compact (with norming constants $a_k = [\log[s/\rho]]^{-1}$, $b_k = (\log[k] + \log[s^s/s!])/\log[s/\rho]$) with partial limit distributions $G(x) = \exp[-\exp[-(x+\epsilon)]]$, $-\infty < x < \infty$, with $\log[\lambda/s\mu] \leq \epsilon \leq 0$. Note that the single server queue corresponds to $s = 1$. Finally, we may apply Theorem 4.2 to show that $\mathbf{Y}^{(k)}/\log[k]$ converges to $\log[s/\rho]$ almost surely.

5.3 The infinite server queue

The infinite server queue is a birth-and-death process with birth rates $q(n, n+1) = \lambda$, and death rates $q(n, n-1) = n\mu$. It corresponds to a queue with Poisson arrivals in which each arriving customer is assigned its own server. The queue is positive recurrent for all values of $\rho = \lambda/\mu$, and the equilibrium distribution is Poisson with mean ρ:

$$\pi(n) = \pi_q(n) = \mathrm{e}^{-\rho}\rho^n/n!.$$

Lemma 3.1 presents this queue as a special case:

$$\lim_{n\to\infty} \mathbf{P}(\mathbf{Y} > n)/(\rho^n/n!) = 1.$$

Due to the explicit appearance of $n!$ in this expression, the analysis of the maximum number of customers simultaneously present in an infinite server queue is considerably more difficult than in a multiserver queue, although it seems that the former appears as the limit of the latter. Comparison of the multiserver queue with the infinite server queue shows that the multiserver queue is uniformizable for all choices of s, the number of servers in the queue (the service speed is uniformly bounded), whereas the infinite server queue is not uniformizable.

Theorem 3.5 presents upper and lower bounds for the distribution of the sample maxima, but as is shown after the proof of Theorem 3.7, these bounds do not characterise the sequence of sample maxima: $\mathbf{P}(\mathbf{Y} \leq n)$ is not stochastically compact.

For the characterisation of limit theorems the behaviour of $f(n) = \rho^n/n!$ for large n is of interest only. Stirling's formula for $n!$ shows that $f(x) = \frac{1}{\sqrt{(2\pi)}}(\rho \mathrm{e})^x/x^{x+\frac{1}{2}}$ is the function appearing in (4.1). This can be obtained from Von Mises conditions, e.g. Theorem 2.7.8 on page 115 of Galambos [1987]: The distribution function $F(x) = 1 - f(x)$ has a negative second derivative $F''(x)$ for all $x > 1$, and satisfies

$$\lim_{x\to\infty} \frac{F''(x)(1-F(x))}{[F'(x)]^2} = -1.$$

Therefore, there exist norming constants $\{a_k > 0,\ b_k\}_{k\geq 1}$ such that (4.1) is satisfied. We may take b_k as given in (4.2), and $a_k = (1 - F(b_k))/F'(b_k)$, that is b_k is the unique solution of

$$b_k = f^{-1}(1/k), \quad \text{and} \quad a_k = [\log[b_k] + (1/2b_k) - \log[\rho]]^{-1}, \quad k \geq 1.$$

Observe that $b_k \sim \log[k]/\log\log[k]$ $(k \to \infty)$. We may now apply Corollary 4.1 to show that $\mathbf{Y}^{(k)}/b_k$ converges to 1 in probability.

The infinite server queue is also covered by Theorem 4.3, which shows that $\mathbf{Y}^{(k)}/b_k$ converges to 1 with probability 1.

6. Conclusions and aims for further research

This paper has discussed the behaviour of the maximum number of customers that is simultaneously present in a queue. It has characterised the distribution of this maximum in a single busy cycle, and a (dual) relation with loss probabilities and Palm probabilities at arrival instants has been discussed. The tail of the distribution of the maximum has allowed us to study the limiting behaviour of normalised maxima over a large number of busy cycles. Applying results from the theory on extremal distributions, it is shown that for queues with bounded service rates, including multiserver queues, this maximum over k cycles, $\mathbf{Y}^{(k)}$, increases logarithmically with probability 1: $\mathbf{Y}^{(k)}/\log k \to 1$. For a class of queues with unbounded service rates, including the infinite server queue, this maximum grows as $\log k/\log\log k$, that is $\mathbf{Y}^{(k)}\log\log k/\log k \to 1$ almost surely.

Several interesting extensions remain to be investigated. Queues with general arrival and service processes will most likely behave similarly as can be seen from the results for the GI/GI/1-queue. The extension to queues with finite waiting room is of current interest as this allows the application of limit theorems to buffer-dimensioning problems. Furthermore, such results could be investigated in a queueing network. Then bottleneck analysis would be feasible using the maximum queue size in the bottleneck queue via methods from the theory on extremal distributions. The extension to batch queues (and batch queueing networks) is of current interest since many of the applications involve simultaneous transmission of customers.

Acknowledgements

The author would like to thank J.W. Cohen for the introduction to the problem discussed in this paper, and for many stimulating discussions, and L. de Haan for several suggestions which improved the final result.

References

1. Anderson, C.W. (1970) Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability* **7**, 99-113.
2. Boucherie, R.J. and van Dijk, N.M. (1991) Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability* **23**, 152-187.
3. Brumelle, S.L. (1978) A generalization of Erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research* **3**, 10-16.
4. Chung, K.L. (1967) *Markov chains with stationary transition probabilities*, 2nd edition, Springer-Verlag.
5. Cohen, J.W. (1971) On a formula dual to Erlang's loss formula. *Operations Research* **19**, 1759-1760.
6. Cohen, J.W. (1982) *The single server queue*, North-Holland.

7. Foster, F.G. (1953) On the stochastic matrices associated with certain queueing processes. *Annals of Mathematical Statistics* **24**, 355-360.

8. Galambos, J. (1987) *The asymptotic theory of extreme order statistics*, Robert E. Krieger Publishing Company.

9. Haan, L. de and Resnick, S.I. (1984) Asymptotically balanced functions and stochastic compactness of sample extremes. *The Annals of Probability* **12**, 588-608.

10. Henderson, W. and Taylor, P.G. (1990) Product forms in networks of queues with batch arrivals and batch services. *Queueing Systems* **6**, 71-88.

11. Iglehart, D.L. and Whitt, W. (1970) Multiple channel queues in heavy traffic. I. *Advances in Applied Probability* **2**, 150-177.

12. Iglehart, D.L. (1972) Extreme values in the GI/G/1 queue. *The Annals of Mathematical Statistics* **43**, 627-635.

13. Kelly, F.P. (1979) *Reversibility and stochastic networks*, Wiley.

14. Kimura, T. (1993) Duality between the Erlang loss system and a finite source queue. *Operations Research Letters* **13**, 169-173.

15. Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983) *Extremes and related properties of random sequences and processes*, Springer-Verlag.

16. Niu, S.-C. and Cooper, R.B. (1989) Duality and other results for M/G/1 and GI/M/1 queues, via a new ballot theorem. *Mathematics of Operations Research* **14**, 281-293.

17. Resnick, S.I. (1987) *Extreme values, regular variation, and point processes*, Springer-Verlag.

18. Serfozo, R.F. (1988) Extreme values of birth and death processes and queues. *Stochastic Processes and their Applications* **27**, 291-306.

19. Takács, L. (1962) *Introduction to the theory of queues*, Oxford University Press.

20. Vervaat, W. (1973) Limit theorems for records from discrete distributions. *Stochastic Processes and their Applications* **1**, 317-334.