Adv. Appl. Prob. **29**, 1039–1059 (1997)
Printed in N. Ireland

APPROXIMATIONS OF GENERAL DISCRETE TIME QUEUES BY DISCRETE TIME QUEUES WITH ARRIVALS MODULATED BY FINITE CHAINS

VINOD SHARMA,* *Indian Institute of Science*

Abstract

Recently, Asmussen and Koole (*Journal of Applied Probability* **30**, pp. 365–372) showed that any discrete or continuous time marked point process can be approximated by a sequence of arrival streams modulated by finite state continuous time Markov chains. If the original process is customer (time) stationary then so are the approximating processes. Also, the moments in the stationary case converge. For discrete marked point processes we construct a sequence of *discrete* processes modulated by discrete time finite state Markov chains. All the above features of approximating sequences of Asmussen and Koole continue to hold. For discrete arrival sequences (to a queue) which are modulated by a countable state Markov chain we form a different sequence of approximating arrival streams by which, unlike in the Asmussen and Koole case, even the stationary moments of waiting times can be approximated. Explicit constructions for the output process of a queue and the total input process of a discrete time Jackson network with these characteristics are obtained.

MARKED POINT PROCESSES; DISCRETE TIME QUEUES; APPROXIMATIONS OF POINT PROCESSES; CONVERGENCE OF STATIONARY MOMENTS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60K25
SECONDARY 60G55; 60J10; 90B22

1. Introduction

Queues with Markov modulated arrivals in continuous as well as discrete time have been extensively studied recently (see, e.g., [10], [9], [17]). The main attraction in using this arrival model is modeling flexibility (against i.i.d. arrivals) and analytical tractability (against more general processes). In fact, several explicit algorithms are available to calculate different performance indices (see [10] for continuous time, and [17], [16] and the references therein for discrete time). If the modulating Markov chain has infinite state space then, although a certain amount of analytical tractability is retained, exact algorithms are no longer available. However, some processes can be modeled exactly only by a process modulated by an infinite state chain, e.g. the output process of a discrete time queue with infinite buffer. Despite its importance, a discrete time queue with arrivals modulated by an infinite

Received 25 September 1995; revision received 4 June 1996.
* Postal address: Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India.

state chain has been studied only in [15]. However, to compute any performance indices one would still need an approximation by a process modulated by a finite state chain.

Recently, Asmussen and Koole [2] have shown that any marked point process can be approximated by a sequence of Markovian arrival streams, where the modulation is by a finite state chain. Hence, if we have a discrete time queue with a general arrival stream, one can form an approximating Markovian arrival stream using [2] and compute, for example, the mean waiting time for a queue with that process. However, such a solution is not completely satisfactory. First, the approximating process we obtain is no longer discrete—the interarrival times and the marks (service times) are approximated by sums of exponentially distributed random variables. It would be aesthetically more pleasing to have an approximation by a discrete time process (queue). Further, discrete queues have also been extensively studied and some of the results of the discrete queues are not available in continuous time (e.g. [14]). Second, if the original stream is stationary, then considering a particular approximating process (e.g. for simulation) in [2], as the time of approximation increases, the state space of the modulating chain also increases. Third, creating the approximating process can be very cumbersome. Fourth, Asmussen and Koole [2] only ensure that the stationary distribution of the waiting time for a queue fed with such a process will be properly approximated but not any of the moments.

In this paper, we try to obtain more satisfactory approximations for discrete time queues. In Section 2 we provide two different ways to obtain approximate Markov modulated (with finite state space) discrete time processes for a given general discrete process.

The first approximation is straightforward, and obtained by replacing the exponential distributions construction in [2] by geometric distributions. We then have the desired features of [2]: a stationary (customer or time) process has a stationary (respectively customer or time) approximation, and all the moments of the interarrival times and the mark distributions converge. The second approximation also has all of these features but is much easier to construct (in fact, it is more intuitive than the first construction and hence solves the third problem mentioned above). This exploits the feature of discrete time. We also suggest a way to solve the second problem, i.e. the state space of the modulating Markov chain does not have to increase with time. The fourth problem remains in this generality, but in Sections 3 and 4 we provide various approximations in certain more specific cases. In Section 3 we provide conditions (not the most natural ones but no others seem to exist [2]) for the continuity of the stationarity distributions and the moments of waiting times for queues with infinite state Markov modulated arrivals. We use these conditions to provide an explicit approximation of an infinite state modulated process by a sequence of finite state modulated arrival streams such that not only the stationary distribution of waiting times (of a queue fed with these processes) but also their moments converge. Then, in Section 4, we use these results to explicitly provide approximations for the output process of an infinite buffer queue when the input is modulated by an infinite state chain. Finally, we provide approximations for the total input process to a queue in a discrete time generalized Jackson network studied in [14]. All our constructions are intuitive and simple, exploit the discrete time property and provide stronger results (convergence of moments) than one would obtain if the general results of [2] are applied.

2. Approximations for general arrival streams

Consider a general marked point process $\{(a_n, y_n), n \geqq 0\}$, where a_n and y_n have discrete values taken for convenience to be $\{1, 2, 3, \cdots\}$. We interpret (a_n) as interarrival times of points while y_n is the mark associated with the nth point. We assume $a_n \geqq 1$ because any batch sizes could also be made part of the marks. We will form a sequence of finite state Markov chains $\{X_k^{(m)}\}$ such that, for a fixed m, an arrival at time k will be decided by the state $X_k^{(m)}$ of the chain and, if an arrival occurs at time k, the corresponding mark y_k will also depend upon $X_k^{(m)}$. To be more precise, given $X_0^{(m)}, \cdots, X_k^{(m)}$ and any arrivals and their marks till time $k-1$ the probability of an arrival at time k and the distribution of its mark will depend only on $X_k^{(m)}$. We will denote the arrival stream formed from $\{X_k^{(m)}\}$ by $\{(a_n^{(m)}, y_n^{(m)})\}$ and it will be such that if $\{(a_n, y_n)\}$ is customer (time) stationary then so will be $\{(a_n^{(m)}, y_n^{(m)}), n \geqq 0\}$ (the time stationary version will be denoted by the arrival epochs and the marks). Further, we will ensure that $E[(a_n^{(m)})^s] \to E[(a_n)^s]$ and $E[(y_n^{(m)})^s] \to E[(y_n)^s]$ as $m \to \infty$ also hold for all $s \geqq 0$. The general approach considered is that of Asmussen and Koole [2], although the actual constructions will be different.

First we obtain an approximation of the (a_1, y_1). This is in the spirit of Neut's theory of phase type distribution [12]. Take $y_1^{(m)} = \min(m, y_1)$ and approximate a_1 by $a_1^{(m)} = \min(m, a_1)$. Of course $(a_1^{(m)}, y_1^{(m)}) \to (a_1, y_1)$ as $m \to \infty$. Let $p_1^{(i)} = P[a_1 = i]$, $i \geqq 1$. Then the distribution of $a_1^{(m)}$ is

$$\sum_{i=1}^{m-1} p_1^{(i)} \delta_i + \left(1 - \sum_{i=1}^{m-1} p_1^{(i)}\right) \delta_m,$$

where δ_i denotes a probability measure with the entire mass at i. Fix constants p_m such that $0 < p_m < 1$ and $p_m \uparrow 1$ as $m \to \infty$.

Define an r.v. $Z^{(m)}$ with distribution

$$P[Z^{(m)} = n] = (1 - p_m)^{n-1} p_m \qquad \text{for } n \geqq 1.$$

Let $\{Z_i^{(m)}\}$ and $\{Z_{ij}^{(m)}\}$ be sequences of i.i.d. r.v.s. with the distribution of $Z^{(m)}$. We approximate δ_i at the mth step by the distribution of $\sum_{k=1}^{i} Z_k^{(m)}$. It is easy to see that $\sum_{k=1}^{i} Z_k^{(m)} \xrightarrow{w} \delta_i$ (where $\xrightarrow{w}$ denotes weak convergence) as $m \to \infty$ and $E[(\sum_{k=1}^{i} Z_k^{(m)})^s] \to i^s$ for $s \geqq 0$. We approximate $a_1^{(m)}$ by

$$a_1^{(m,n)} \triangleq \sum_{i=1}^{m} 1(t^{(m)} = i) \sum_{k=1}^{i} Z_{ik}^{(n)},$$

where $t^{(m)}$ is a r.v. independent of the other r.v.s. and $P(t^{(m)} = i) = p_1^{(i)}$ for $0 \leqq i < m$ and $P(t^{(m)} = m) = 1 - \sum_{i=0}^{m-1} p_1^{(i)}$. We can show that $a_1^{(m,n)} \xrightarrow{w} a_1^{(m)}$ and $E[(a_1^{(m,n)})^s] \to E[(a_1^{(m)})^s]$ for all $s \geqq 0$, as $n \to \infty$. Thus we have an approximation of a_1 by a mixture of sums of geometric distributions, while y is approximated by a simple truncation.

Just as a mixture of Erlang distributions can be realized by a jump of a Markov jump process, a mixture of sums of geometric distributions can be realized by a jump of a discrete time finite state Markov chain. We show it for $a_1^{(m,n)}$. The chain has state space $\{0, (i, i_1, i_2, \cdots, i_{i-1}), 1 \leqq i \leqq m\}$ and it evolves as follows. At $k = 0$ let the chain be in state 0 and an arrival has just occurred. At $k = 1$ the chain changes states from 0 to i with probability $p_1^{(i)}$. We denote the transition probabilities of the Markov chain by P_{ij}, where $P_{ii} = p_m$ and $P_{ii_1} = 1 - p_m$. Also, $P_{i_k, i_k} = p_m$, $P_{i_k, i_{k+1}} = 1 - p_m$ for $1 \leqq k \leqq i - 2$, $P_{i_{i-1}, i_{i-1}} = p_m$ and $P_{i_{i-1}, 0} = 1 - p_m$. When the chain next reaches state 0, again there is an arrival and we associate a mark $y_1^{(m)}$ with it.

To approximate an arbitrary sequence $\{(a_n, y_n)\}$, first approximate it with a sequence $\{(a_n^{(m)}, y_n^{(m)})\}$, $a_n^{(m)} \triangleq \min(a_n, m)$, $y_n^{(m)} \triangleq \min(y_n, m)$ and then approximate $\{(a_n^{(m)}, y_n^{(m)})\}$ by $\{(a_n^{(m,k)}, y_n^{(m)})\}$ where $(a_n^{(m,k)}, y_n^{(m)})$ is obtained as above for the point (a_1, y_1).

We realize $\{(a_1^{(m,k)}, y_1^{(m)}), (a_2^{(m,k)}, y_2^{(m)}), \cdots, (a_n^{(m,k)}, y_n^{(m)})\}$ by a finite state Markov modulated process by retaining the whole past (which is finite because it is a finite sequence with finitely valued r.v.s.) with the required information about $a_l^{(m,k)}$ being $(p_l^{(i)}, i = 1, \cdots, m)$. The formal details can be carried out as in [2]. Using a realization similar to that of $(a_1^{(m,n)}, y_1^{(m)})$ given above (see also the details of the next construction), it can be shown, as in [2], that if $\{(a_n, y_n)\}$ is a customer stationary sequence then so is $\{(a_1^{(m,k)}, y_1^{(m)}), \cdots, (a_n^{(m,k)}, y_n^{(m)})\}$. A time stationary version will provide a time stationary approximation. We have already shown that if $E[a_l^s] < \infty$ then $E[(a_l^{(m,k)})^s] \to E[a_l^s]$ and $E[(y_l^{(m)})^s] \to E[y_l^s]$.

The second construction is simpler and more intuitive. Again we approximate, $\{(a_1^{(m)}, y_1^{(m)}), \cdots, (a_n^{(m)}, y_n^{(m)})\}$. Now we provide an explicit construction. Consider a Markov chain with state space

$$\{(S_1^{(m)}, \cdots, S_n^{(m)}): S_k^{(m)} = (s_{11}^{(m)}, \cdots, s_{1m}^{(m)}, y),\ s_{ij}^{(m)} = 0 \text{ or } 1 \text{ and } y \in \{1, 2, \cdots, m\}\}.$$

At time 0, the Markov chain is in state

$$\{(1,\cdots,\underset{\substack{\uparrow\\ a_1^{(m)}\text{th position}}}{1},0,\cdots,0,y_1^{(m)}),\ (1,\cdots,\underset{\substack{\uparrow\\ a_2^{(m)}\text{th position}}}{1},0,\cdots,0,y_2^{(m)}),\cdots,(1,\cdots 1,0\cdots,0,y_n^{(m)})\}$$

and jumps to state (the first 1 becomes 0, everything else stays the same)

$$\{(0,1,\cdots,1,0,\cdots 0,y_1^{(m)}),\ (1,\cdots 1,0,\cdots 0,y_2^{(m)}),\cdots,(1,\cdots,1,0\cdots,0,y_n^{(m)})\}.$$

Until time $a_1^{(m)}$ the first one is made zero and then, at $a_1^{(m)}$, an arrival occurs with mark $y_1^{(m)}$ and the state becomes $(S_1^{(m)},\cdots,S_n^{(m)})$ where $S_2^{(m)},\cdots,S_n^{(m)}$ are the same as at time 0 and $S_1^{(m)}$ becomes $(0,0,\cdots,0)$. Then the same procedure applies to $S_2^{(m)}$ and, at time $a_1^{(m)}+a_2^{(m)}$, a second arrival comes with mark $y_2^{(m)}$ and $(S_1^{(m)},S_2^{(m)})$ becomes $(0,0,\cdots,0)(0,\cdots,0)$. This repeats, until the nth arrival occurs. Again if the given sequence is customer (time) stationary, we obtain a customer (time) stationary approximation. Also, the moments of interarrival times and marks converge.

We notice that both the above approximations have the second drawback mentioned in the introduction—as the length of the sequence n increases, the state space increases. We now suggest a way to alleviate this problem. When $\{(a_1,y_1),(a_2,y_2),\cdots,\}$ is a stationary sequence, approximate it by a stationary m-dependent sequence $\{(\tilde{a}_1^{(m)},\tilde{y}_1^{(m)}),(\tilde{a}_2^{(m)},\tilde{y}_2^{(m)}),\cdots\}$ which has the m-dimensional distributions of $\{(a_1,y_1),(a_2,y_2),\cdots\}$. Thus as $m\to\infty$, all the finite-dimensional distributions of $\{(\tilde{a}_1^{(m)},\tilde{y}_1^{(m)}),(\tilde{a}_2^{(m)},\tilde{y}_2^{(m)}),\cdots\}$ converge to that of $\{(a_1y_1),(a_2,y_2),\cdots\}$ and $E[(\tilde{a}_1^{(m)})^s]=E[a_1^s]$ and $E[(\tilde{y}_1^{(m)})^s]=E[y_1^s]$. Now, instead of approximating $\{(a_1,y_1),(a_2,y_2)\cdots\}$ we approximate $\{(\tilde{a}_1^{(m)},\tilde{y}_1^{(m)}),\{(\tilde{a}_2^{(m)},\tilde{y}_2^{(m)})\cdots\}$ by the Markov modulated sequences presented above. The state space will no longer explode as time increases.

3. Approximations for discrete queues with countable state modulated arrivals

In this section we concentrate on discrete arrival streams which are obtained from a modulating countable state Markov chain. Instead of using a customer stationary representation, we use a time stationary representation (when the process is stationary). Specifically, let $\{X_k\}$ be a countable state, aperiodic, irreducible, ergodic Markov chain (we take its state space to be positive integers). Let Z_k be the number of arrivals at time k. Given $(X_0,\cdots.X_k,Z_0,\cdots,Z_{k-1})$, the distribution of Z_k

depends only on X_k. For simplicity, we do not attach any marks with the arrivals (we will comment on the general case at the end of the section). We are interested in approximating the process $\{Z_k\}$ by a process that will be modulated by a finite state Markov chain. The approximations should be such that if these processes are used as input to a discrete time queue then the waiting time and the work load process of the corresponding queues should have the stationary distributions and moments close to each other. The approximations obtained in the last section can ensure the convergence of stationary distributions but not of the moments. Moments are certainly of considerable practical interest.

To show that our approximations satisfy the desired properties, we will consider a single server discrete time queue. The service times are of one unit length and the service can start only at integer points. We take Z_k defined above as the number of customers arriving at the queue in time interval $(k, k+1)$. Let W_k be the total number of customers (and hence also the total work load) in the queue at time k. Then $W_{k+1} = (W_k - 1)^+ + Z_k$ and $\{(W_k, X_k)\}$ is a countable state Markov chain. This queue has been extensively studied in [15], providing rates of convergence to stationary distributions and functional limit theorems in light and heavy traffic etc. Let us also define the process $\tilde{W}_k = (W_k - 1)^+$. Then $\tilde{W}_{k+1} = (\tilde{W}_k + Z_k - 1)^+$ and $W_{k+1} = \tilde{W}_k + Z_k$. Let π be the stationary distribution of $\{X_k\}$. Then if $E_\pi[Z_1] < 1$, from [1], p. 237, $\{(\tilde{W}_k, X_k, Z_K)\}$ has a unique stationary distribution (again denoted by π) and starting from any initial distribution, convergence to π occurs in total variation. This implies that $W_{k+1} = \tilde{W}_k + Z_k$ also converges to a unique stationary distribution and so do (W_{k+1}, X_k, Z_k).

Since we are interested in showing the continuity of stationary distributions the following result of Borovkov [3], p. 53, will be repeatedly used.

Consider a sequence of $G/G/1$ queues with stationary ergodic arrival sequences $\{(a_k^{(n)}, s_k^{(n)})\}$, $\{(a_k, s_k)\}$ and let $W^{(n)}$, W denote the corresponding stationary waiting times. Then $W^{(n)} \xrightarrow{w} W$ if the finite-dimensional distributions of $\{(a_k^{(n)}, s_k^{(n)}), k \geqq 0\}$ converge to that of $\{(a_k, s_k)\}$ and $E[(s_1^{(n)} - a_1^{(n)}); s_1^{(n)} > a_1^{(n)}] \to E[s_1 - a_1; s_1 > a_1]$.

Since we are interested in the Markov modulated discrete queues defined above, we will translate the result of Borovkov into more specific terms. Consider a sequence of inputs $\{(X_k^{(m)}, Z_k^{(m)})\}$ and $\{(X_k, Z_k)\}$. If Borovkov's conditions are satisfied (when these sequences are in stationarity) then we actually get $\tilde{W}^{(m)} \xrightarrow{w} \tilde{W}$. However, $(W^{(m)} - 1)^+ \stackrel{\text{dist}}{=} \tilde{W}^{(m)}$ and hence $(W^{(m)} - 1)^+ \xrightarrow{w} (W - 1)^+$. Thus for any $k > 0$, $P[(W^{(m)} - 1)^+ = k] = P[W^{(m)} = k + 1] \to P[W = k + 1]$. Therefore, $P[W^{(m)} \leqq 1] \to P[W \leqq 1]$. But $P[W^{(m)} = 0] = 1 - E_\pi[Z_1^{(m)}]$ and we will require that $E_\pi[Z_1^{(m)}] \to E_\pi[Z_1]$. Then we get $W^{(m)} \xrightarrow{w} W$. Under the same conditions we will show below the convergence of stationary distributions of waiting times also.

We will show in the appendix that for our system the following conditions are sufficient for that of Borovkov. As $m \to \infty$,

$$
\begin{aligned}
&P[X_1^{(m)} = i,\ Z_o^{(m)} = k \mid X_0^{(m)} = j] \to P[X_1 = i,\ Z_o = k \mid X_0 = j], \quad \forall i, j, k \\
(1) \quad &E_\pi[Z_1^{(m)}] \to E_\pi[Z_1], \\
&P_\pi[X_1^{(m)} = i] \to P_\pi[X_1 = i], \quad \forall i.
\end{aligned}
$$

The first condition in (1) implies that

$$P[X_1^{(m)} = i \mid X_0^{(m)} = j] \to P[X_1 = i \mid X_0 = j]$$

and

$$P[Z_0^{(m)} = k \mid X_0^{(m)} = j] \to P[Z_0 = k \mid X_0 = j] \qquad \text{for all } i, j, k.$$

Next we show that, under the above conditions, the waiting time distributions also converge. For this we need an expression for the stationary queue lengths seen by an arriving batch of customers. Consider the Markov chain $\{((W_k, X_k), (W_{k+1}, X_{k+1}))\}$. An arrival occurs whenever this two-step Markov chain enters the set

$$S = \{((n, i), (m, j)) : m \geqq n \text{ when } n > 0 \text{ and } m > n \text{ when } n = 0\}.$$

Consider this chain only at the epochs when it enters S. This modified chain has the stationary distribution

$$
\begin{aligned}
(2) \qquad &P_\pi[((W_k, X_k), (W_{k+1}, X_{k+1})) = ((i, j), (i', j'))] \\
&\qquad = \frac{\pi(i, j)P((i, j), (i', j'))}{\sum_S \pi(k, l)P((k, l), (k', l'))},
\end{aligned}
$$

where $P(\cdot, \cdot)$ on the right-hand side denotes the transition probability matrix of $\{(W_k, X_k)\}$. If we sum over (i', j') then we obtain the stationary distribution of W_k as seen by an arriving batch. Thus we obtain the stationary waiting time of the first customer in a batch. We will show below that this distribution for the corresponding queues converges under the conditions (1). Next observe that, since the time stationary distributions of $\{Z_k^{(m)}\}$ converge to that of $\{Z_k\}$, we also have convergence of batch stationary distributions of $\{Z_k^{(m)}\}$ to that of $\{Z_k\}$ (for terminology see [5], ch. 7). Indeed, taking batch size as marks, the batch stationary process can be considered a customer stationary process. Thus from Theorem 3.7.1, p. 114 of [5], we get convergence of distributions of this process. Further, from [18] (see also [5], p. 224) we get that in the customer stationary version of $\{Z_k\}$ the position of the zeroth customer being $n > 0$ has probability

$$P_\pi[Z_1 \geqq n \mid Z_1 > 0] / E_\pi[Z_1 \mid Z_1 > 0].$$

Hence from (1) we get that the distribution of position of a typical customer (customer zero) in its batch for $\{Z_k^{(m)}\}$ (assuming the customers in a batch are ordered and are served in that order) also converges to that of $\{Z_k\}$. These results together provide the convergence of the stationary distributions of the waiting times.

Now for convergence of stationary waiting times we only have to show the convergence of waiting times of the first customer in a batch, obtained from (2). We first show that under conditions stated in (1), $\pi^{(m)}(i,j) \to \pi(i,j)$ and

$$P_\pi^{(m)}((i,j),(i',j')) \to P_\pi((i,j),(i',j')).$$

Borovkov [3] only shows that $W^{(m)} \xrightarrow{w} W$. But under the same conditions, Brandt *et al.* [5] show that $(Z^{(m)}, W^{(m)}) \xrightarrow{w} (Z, W)$. However, Borovkov [4] shows that under these conditions 'renovating' events exist, i.e. the conditions of Theorem 1.7.1 in [5] are satisfied and hence we get $(Z^{(m)}, W^{(m)}, X^{(m)}) \xrightarrow{w} (Z, W, X)$, where (Z, W, X) represent r.v.s with the stationary distribution of (Z_k, W_k, X_k) (similarly for the others). Thus, by a general convergence theorem in Royden [13], p. 270,

$$\sum_S \pi^{(m)}(k,l) P^{(m)}((k,l),(k',l')) \to \sum_S \pi(k,l) P((k,l),(k',l')).$$

Hence under stationarity, when the two-step chain is watched in the set S,

$$((W_k^{(m)}, X_k^{(m)}), (W_{k+1}^{(m)}, X_{k+1}^{(m)})) \to ((W_k, X_k)), (W_{k+1}, X_{k+1}))$$

in total variation. This provides the convergence of the stationarity waiting time distribution of the first customer in a batch.

We will also be interested in showing the convergence of stationary moments of $\{W_k\}$ and waiting time. Such results do not seem to be available except for $GI/GI/1$ queues (see [2]) although recently the finiteness of moments for a $G/GI/1$ queue has been obtained [6]. Hence the following results, although somewhat restricted (for countable state modulated chains at least; the finite state case seems more satisfactory) are still worthwhile. (Our recent result removes these conditions—in particular, (5) and the inequality following it are not required.) Let $\tau^{(m)}$ be the regeneration length for the mth system where regeneration epochs will be the visit times of $\{(W_k^{(m)}, X_k^{(m)})\}$ to state $(0, i)$, i being some fixed state of $\{X_k^{(m)}\}$. Then, for any $\alpha > 0$, taking $k = 0$ as a regeneration epoch,

$$E[\tau^{(m)}] E_\pi[(W^{(m)})^\alpha] = E\left[\sum_{k=0}^{\tau^{(m)}-1} (W_k^{(m)})^\alpha\right] \leqq E[(\tau^{(m)})^{\alpha+1}],$$

where the inequality holds because $W_k^{(m)} \leqq \tau^{(m)}$ for all $k = 0, \cdots, \tau^{(m)} - 1$. Since $\tau^{(m)} \geqq 1$, we now get

$$\sup_m E_\pi[(W^{(m)})^\alpha] \leqq \sup_m E[(\tau^{(m)})^{\alpha+1}]. \tag{3}$$

Thus, if we show that the right-hand side in (3) is finite, then under (1), by uniform integrability, we will get $E[(W^{(m)})^{\alpha'}] \to E[W^{\alpha'}]$ for any $\alpha' < \alpha$. Since inequality (3) is also satisfied by the moments of the stationary waiting times, we will also obtain convergence of moments of the waiting times.

Now we provide conditions for (3). For this we use a result from [11], p. 10. We provide the result for easy reference.

Let there be a function $V \geqq 0$ on the state space of the Markov chain (W_k, X_k). Also, let there be positive numbers Δ, b and $s > 1$ and an r.v. $\Delta((w, j))$ defined for each (w, j) in the state space such that:

(i) $v_Q \triangleq \sup_{(w,j) \in Q} V((w, j)) < \infty$ where Q is a subset of the state space;

(ii) $P_{(w,j)}[V(W_1, X_1) - V((w, j)) \leqq \Delta((w, j))] = 1, \quad \forall (w, j)$;

(iii) $\sup_{(w,j) \notin Q} E[\Delta((w, j))] \leqq -\Delta$;

(iv) $\sup_{(w,j)} E\, |\Delta((w, j))|^s \leq b < \infty$.

Then

$$E_{(w,j)}[\tau_Q^s] \leqq \begin{cases} \left(\alpha + \dfrac{2V((w, j))}{\Delta}\right)^s, & (w, j) \notin Q, \\ C(\Delta, b, s, v_Q), & (w, j) \in Q, \end{cases}$$

where τ_Q is the first time that the Markov chain $\{(W_k, X_k)\}$ enters the set Q and $C(\Delta, b, s, v_Q)$ is a constant, explicitly provided in [11], that depends only on Δ, b, s, v_Q.

For our purposes we take $Q = \{(w, j) : w \leqq M \text{ and } j \leqq M\}$ for some constant M such that $i \leqq M$. Also, let $\Delta((w, j)) = V(W_1, X_1) - V((w, j))$, $(w, j) \in$ state space of (W_k, X_k) and $V(w, j) \triangleq w + f(j)$ where f is a non-negative function on the state space of $\{X_k\}$. Then (i) and (ii) are immediately satisfied. For (iv),

$$E_{(w,j)}[|\Delta(w, j)|^s] = E[|V(W_1, X_1) - V(w, j)|^s \mid (W_0, X_0) = (w, j)] < \infty,$$

using the c_r-inequality, it is sufficient to have

$$\sup_j E[|Z_0|^s \mid X_0 = j] < b_1, \qquad \sup_j E[|f(X_1) - f(j)|^s \mid X_0 = j] < b_2, \tag{4}$$

where $b = c_s(b_1 + b_2)$, c_s being a constant dependent only on s. The unnatural restriction comes because of (iii), if the state space of $\{X_k\}$ is countably infinite. It can be easily shown that (iii) is satisfied for some $\Delta > 0$ if there exist $\varepsilon > 0$, $\delta > 0$ and $\varepsilon_1 > 0$ with $\delta < \varepsilon$ such that, for $j > M$,

$$E[f(X_1) \mid X_0 = j] < f(j) - \varepsilon, \qquad E[Z_0 \mid X_0 = j] < \delta, \tag{5}$$

and for all j,

$$E[Z_0 \mid X_0 = j] \leqq 1 - E[f(X_1) \mid X_0 = j] + f(j) - \varepsilon_1.$$

If the state space of $\{X_k\}$ is finite (taking $f(x) = 0$ and $Q = \{(0, j)$, for all j in state space of $\{x_k\}\}$) then instead of all these conditions (5) we only need $E[Z_0 \mid X_0 = j] < 1$ for all j.

As pointed out by the referee, it is important to generalize (5) and (for the finite state space of $\{X_k\}$) $E[Z_0 \mid X_0 = j] < 1$ for all j. Now we generalize these conditions such that, at least for the most important case (in practice) of a finite state space of

$\{X_k\}$ and Z_k having a finite support, the condition becomes completely satisfactory. The idea is to apply Kalashnikov's conditions (i)–(iv) on the Markov chain $\{(W_{Nk}, X_{Nk}), k \geqq 0\}$ for some non-negative integer N. Choosing all the constants and functions V, Δ as above, conditions (i), (ii) and (iv) are satisfied under the same conditions as for the chain $\{(X_k, Z_k)\}$ while for (iii) in (5) $f(X_1)$ is replaced by $f(X_N)$. When $\{X_k\}$ has finite state space, then for (5) we now need

(5a) $$E[Z_0 + Z_1 + \cdots + Z_{N-1} \mid X_0 = j] < N$$

for all j. Suppose $E_\pi[Z_0] < 1$. By the ergodic theorem (which holds because $\{Z_k\}$ is regenerative with finite mean cycle length) as $n \to \infty$

$$\frac{1}{n} \sum_{k=0}^{n-1} E[Z_k \mid X_0 = j] \to E_\pi[Z_0]$$

for any j. Hence we obtain an N such that $E[\sum_{k=0}^{N-1} Z_k \mid X_0 = j] < N$. If this N can be chosen independently of (Z_0, X_0) (which happens if Z_k and X_k have compact support) then (5a) will be satisfied. Now, if $\sup_x E_x[\tau_Q^s] < b_3$ for $\{(W_{Nk}, X_{Nk})\}$, then $\sup_x E_x[\tau_Q^s] < Nb_3$ for $\{(W_k, X_k)\}$. Therefore, to get $\sup_{x,m} E_x[(\tau_Q^{(m)})^s] < \infty$, we need a common N for all chains $\{(X_k^{(m)}, Z_k^{(m)})\}$. This will hold if (5a) holds and Z_k, X_k have compact support. If $\{Z_k\}$ does not have compact support, then in addition to (1) we also need $E[Z_0^{(m)} \mid X_0 = j] \to E[Z_0 \mid X_0 = j]$ for all j.

If $\{(Z_k, X_k)\}$ satisfies (4) and (5) for $s = 5$ then $E[\tau^5] < \infty$ and from (3), $E_\pi[W^4] < \infty$. Thus, from [15], the process $\{W_k\}$ satisfies a functional CLT. Also a certain rate of convergence to stationarity can be ensured. If (5) is satisfied by all $\{(W_k^{(m)}, X_k^{(m)})\}$ and $\{(W_k, X_k)\}$ for the same ε_1, ε, δ and (4) is satisfied for the same b_1, b_2 and s then we get $\sup_{m,x} E_x[(\tau_Q^{(m)})^s] < \infty$ and $E_x[(\tau_Q)^s] < \infty$ for all x. However, we need $\sup_m E[(\tau^{(m)})^s] < \infty$, which we show now.

Let $B = \sup_{m, x \in Q} \{E_x[(\tau_Q^{(m)})^s]\}$. First we show that $E[(\tau)^s] < \infty$. Since Q is a finite set and $\{(W_k, X_k)\}$ is irreducible, there is a constant $T < \infty$ and a probability $p > 0$ such that

(6) $$\inf_{(n,j) \in Q} P_{(n,j)} \{\text{first time state } (0, i) \text{ is hit} \leqq T\} > p.$$

Now let $(W_0, X_0) = (0, i)$ and let the subsequent epochs when the chain hits set Q be $\tau_{Q,1}$, $\tau_{Q,2}, \cdots$. Also, let N be such that the chain hits $(0, i)$ within time T of $\tau_{Q,N}$. Then by definition

$$P[N > n] \leqq (1 - \tfrac{1}{2}p)^n.$$

Define an r.v. $\tilde{\tau}_{Q,1}$ by $\tilde{\tau}_{Q,1} = \sup_{(n,j) \in Q} [\tau_{Q,1} \mid (W_0, X_0) = (n, j)]$. Then

$$E[(\tilde{\tau}_{Q,1})^s] \leqq E\Big[\Big(\sum_{(n,j) \in Q} [\tau_{Q,1} \mid W_0, X_0 = n, j]\Big)^s\Big] \leqq B_1 \sup_{x \in Q} E_x[(\tau_Q)^s]$$

by the c_r-inequality, where B_1 is a constant dependent only upon the cardinality of Q.

Define $\tilde{\tau}_{Q,1}$, $\tilde{\tau}_{Q,2}, \cdots$ i.i.d. with the distribution of $\tilde{\tau}_{Q,1}$ and independent of N. Then, with $(W_0, X_0) = (0, i)$, τ will be the next time the chain hits $(0, i)$ and

$$\begin{aligned} E[\tau^s] &\leqq E_{(0,i)}\left[\left(\sum_{k=1}^{N} \tau_{Q,k} + T\right)^s\right] \leqq E_{0,i}\left[\left(\sum_{k=1}^{N} \tilde{\tau}_{Q,k} + T\right)^s\right] \\ &\leqq E_{0,i}\left[\left(\sum_{k=1}^{\tilde{N}} \tilde{\tau}_{Q,k} + T\right)^s\right] \leqq B_2(E[\tilde{N}^s]E[\tilde{\tau}^s_{Q,k}] + T^s) \\ &\leqq BB_2E[\tilde{N}^s] + B_2T^s < \infty, \end{aligned}$$

where $\tilde{N}$ is an r.v. independent of all other r.v.s and has distribution $P[\tilde{N} = n] = (1 - \frac{1}{2}p)^{n-1}\frac{1}{2}p$ and B_2 is a constant dependent only on s (the third inequality follows from [8], p. 22).

We show that the upper bound given for $E[\tau^s]$ holds for the chains $\{(W_k^{(m)}, X_k^{(m)})\}$ for all m large enough. For this it is sufficient to show that we can have a lower bound $p' > 0$ in (6) for all chains with index m large enough. But observe that the left-hand side of (6) depends on only the first T-dimensional distributions of $\{(W_k, X_k)\}$. Under our conditions, we have

$$P_{(n,i)}\{((W_1^{(m)}, X_1^{(m)}), \cdots, (W_T^{(m)}, X_T^{(m)})) \in A\} \to P_{(n,i)}\{((W_1, X_1), \cdots, (W_T, X_T)) \in A\}$$

for all (n, i) and all sets A. Hence, by making the lower bound $p/2$ in (6) we can ensure that the inequality is satisfied by all the chains with indices m large enough. Thus, we get

$$\sup\left(\sup_m E[(\tau^{(m)})^s], E[\tau^s]\right) < \infty.$$

For easy reference we summarize the results obtained so far in the following theorem.

Theorem 1. Consider the countable state Markov modulated queues $\{(W_k^{(m)}, X_k^{(m)})\}$, $\{(W_k, X_k)\}$. If conditions (1) are satisfied then the transient as well as the stationary distributions of $\{(W_k^{(m)}, X_k^{(m)})\}$ converge to that of $\{(W_k, X_k)\}$. The stationary distributions of waiting times also converge. If in addition (4) and (5) hold for all the chains with the same positive constants b_1, b_2, ε_1, ε, δ and $s > 1$ with $\delta < \varepsilon$ then the α'th stationary moments of $\{W_k\}$ and waiting times also converge for any $\alpha' < s - 1$.

Although the above theorem is useful in practice, it does not provide any algorithms to compute the quantities of interest. For that, we now approximate an arrival stream modulated by a countable state Markov chain by a sequence of finite state modulated arrival streams in such a way that the stationary distributions as well as some moments of the work load and the waiting time processes converge. Now for simplicity we further assume that given X_k, X_{k+1} and Z_k are independent of each other. This is a natural assumption in the present context and is frequently made in the literature. Furthermore, with a little more effort, it should be possible to remove it.

Consider the arrival stream $\{(X_k, Z_k)\}$, where $\{X_k\}$ is the countable state (with state space, say $\{1, 2, \cdots\}$) modulating chain. We construct a sequence of arrival streams $\{(X_k^{(m)}, Z_k^{(m)})\}$ (where $\{(X_k^{(m)})\}$ has the state space $\{1, 2, \cdots, m\}$) by the truncation scheme provided in [7]. Fix some arbitrary state i of $\{X_k\}$. We construct $\{X_k^{(m)}\}$ for $m > i$ (asymptotically it does not matter) by

$$P[X_k^{(m)} = j \mid X_{k-1}^{(m)} = l| = P[X_k = j \mid X_{k-1} = l] \qquad \text{if } j, l \leqq m, j \neq i,$$

$$P[X_k^{(m)} = i \mid X_{k-1}^{(m)} = l] = \sum_{j=m+1}^{\infty} P[X_k = j \mid X_{k-1} = l] + P[X_k = i \mid X_{k-1} = l] \qquad \text{for } l \leqq m.$$

Thus effectively $\{X_k^{(m)}\}$ is formed from $\{X_k\}$ by making $X_k = i$, whenever the state of $\{X_k\}$ comes out of the set $\{1, 2, \cdots, m\}$.

One can easily check that, for any $j, l \in \{1, 2, \cdots\}$,

$$P[X_1^{(m)} = j \mid X_0^{(m)} = l] \to [X_1 = j \mid X_0 = l].$$

It is also shown in [7] that if $\{X_k\}$ is irreducible, aperiodic and positive recurrent then so are all $\{X_k^{(m)}\}$ chains, and their stationary distributions $\pi^{(m)}$ converge to π, the stationary distributions of $\{X_k\}$.

Define $P[Z_1^{(m)} = n \mid X_1^{(m)} = j]$ for $j \neq i$ by

$$P[Z_1^{(m)} = n \mid X_1^{(m)} = j] = P[Z_1 = n \mid X_1 = j] \times \min(\pi(j)/\pi^{(m)}(j), 1), \qquad n \neq 0$$

and

$$\begin{aligned} P[Z_1^{(m)} = 0 \mid X_1^{(m)} = j] = P[Z_1 = 0 \mid X_1 = j] \\ \times \min(\pi(j)/\pi^{(m)}(j), 1) + (1 - \pi(j)/\pi^{(m)}(j))^+. \end{aligned}$$

Also, let

$$\begin{aligned} P[Z_1^{(m)} = n \mid X_1^{(m)} = i] &= \frac{1}{a^{(m)}} \sum_{j=m+1}^{\infty} P[Z_1 = n \mid X_1 = j] \\ &\quad \times \frac{\pi(j)}{\pi^{(m)}(i)} + \frac{(\pi(i)/\pi^{(m)}(i))}{a^{(m)}} P(Z_1 = n \mid X_1 = i), \\ P[Z_1^{(m)} = 0 \mid X_1^{(m)} = i] &= 1 - \sum_{n=1}^{\infty} P[Z_1^{(m)} = n \mid X_1^{(m)} = i] \end{aligned} \tag{7}$$

where

$$a^{(m)} = \frac{[\sum_{j=m+1}^{\infty} \pi(j) + \pi(i)]}{\pi^{(m)}(i)}.$$

One observes that, for all n and j,

$$P[Z_1^{(m)} = n \mid X_1^{(m)} = j] \to P[Z_1 = n \mid X_1 = j].$$

We show that $P_\pi(Z_1^{(m)} = n) \to P_\pi[Z_1 = n]$ for all n and $E_\pi[(Z_1^{(m)})^\alpha] \to E_\pi[Z^\alpha]$ for $\alpha > 0$ whenever $E_\pi[Z^\alpha] < \infty$.

We have

$$E_\pi[(Z_1^{(m)})^\alpha] = \sum_{j=1}^{m} \sum_{n=1}^{\infty} n^\alpha P[Z_1^{(m)} = n \mid X_1^{(m)} = j]\pi^{(m)}(j)$$

$$(8)\qquad = \sum_{n=1}^{\infty} \left[\sum_{\substack{j=1\\ j\neq i}}^{m} n^\alpha P[Z_1 = n \mid X_1 = j] \min\left(1, \frac{\pi(j)}{\pi^{(m)}(j)}\right)\pi^{(m)}(j) + \sum_{j=m+1}^{\infty} n^\alpha P[Z_1 = n \mid X_1 = j]\frac{\pi(j)}{a^{(m)}} + n^\alpha P[Z_1 = n \mid X_1 = i]\frac{\pi(i)}{a^{(m)}}\right]$$

$$\leqq \sum_{n=1}^{\infty} \sum_{j} n^\alpha P(Z_1 = n \mid X_1 = j) \,.\, \pi(j) = E_\pi[Z^\alpha].$$

Therefore $\liminf E_\pi[(Z_1^{(m)})^\alpha] \leqq E_\pi[Z_1^\alpha]$. But since $\min(1, \pi(j)/\pi^{(m)}(j))\pi^{(m)}(j) \to \pi(j)$ and $\pi(i)/a^{(m)} \to \pi(i)$, by a general convergence theorem in [13], p. 231, we get $\liminf E_\pi[(Z_1^{(m)})^\alpha] \geqq E_\pi[Z_1^\alpha]$. Thus $E_\pi[(Z_1^{(m)})^\alpha] \to E_\pi[Z_1^\alpha]$. Similarly we can show the convergence of stationary distributions of $Z_1^{(m)}$ to that of Z_1.

Thus we have satisfied all the requirements in Theorem 1 for the convergence of stationary distributions of the workload and waiting time processes.

We apply Theorem 1 to obtain convergence of stationary moments of workload and waiting times. Now we assume that $\{(X_k, Z_k)\}$ satisfies (4) and (5) for some f, b_1, b_2, ε, δ, ε_1 and s with the required conditions on these constants. Also, let f be non-decreasing. We show that (4) and (5) will be satisfied by all $[(X_k^{(m)}, Z_k^{(m)})\}$ for m large enough and $\{X_k, Z_k\}$ by appropriately modifying the constants which will remain the same for all these chains.

Consider (4). For $j \neq i$,

$$(9)\qquad E[(Z_0^{(m)})^s \mid X_0^{(m)} = j] = \sum_{n=1}^{\infty} n^s P(Z_0 = n \mid X_0 = j) \min\left(1, \frac{\pi(j)}{\pi^{(m)}(j)}\right)$$

$$\leqq E[Z_0^s \mid X_0 = j] < b_1.$$

Also, for i,

$$E[(Z_0^{(m)})^s \mid X_0^{(m)} = i] = \sum_{j=m+1}^{\infty} \sum_{n=1}^{\infty} \left(n^s P[Z_0 = n \mid X_0 = j] \frac{\pi(j)}{\sum_{k=m+1}^{\infty} \pi(k) + \pi(i)}\right)$$

$$+ \sum_{n=1}^{\infty} n^s P[Z_0 = n \mid X_0 = i]\pi(i) \Big/ \left(\sum_{j=m+1}^{\infty} \pi(j) + \pi(i)\right)$$

$$\leqq \frac{\sum_{j=m+1}^{\infty} E[Z_0^s \mid X_0 = j]\pi(j) + \pi(i)E[Z_0^s \mid X_0 = i]}{\sum_{j=m+1}^{\infty} \pi(j) + \pi(i)} \leqq b_1.$$

Also, as $m \to \infty$,

$$E[|f(X_1^{(m)}) - f(j)|^s \, | X_0^{(m)} = j] = \sum_{k=1}^{m} |f(k) - f(j)|^s P[X_1^{(m)} = k \mid X_0^{(m)} = j]$$

$$(10) \qquad = \sum_{k=1}^{m} |f(k) - f(j)|^s P[X_1 = k \mid X_0 = j]$$

$$+ |f(i) - f(j)|^s P[X_1 \geqq m+1 \mid X_0 = j]$$

$$\to E[|f(X_1) - f(j)|^s \mid X_0 = j].$$

Thus the bound b_2 holds for all large m for any given j. For a bound on

$$\sup_j E[|f(X_1^{(m)}) - f(j)|^s \, | X_0^{(m)} = j]$$

for all large m, if f is bounded we obviously have a bound. If f is not bounded, then we further assume that

$$(11) \qquad f(m)^s P[X_1 \geqq m+1 \mid X_0 = j] < b_3$$

for all j and m large enough. A sufficient condition for (11) is $\sup_j E[(f(X_1))^s \mid X_0 = j] < \infty$. Then from (10), for all m large enough,

$$\sup_m \sup_j E[|f(X_1^{(m)} - f(j)|^s \, | X_0^{(m)} = j] < b_2 + 2b_3 c_s,$$

where c_s depends only on s.

Now consider (5). In (9) we have shown that, for $j \neq i$, $E[Z_0^{(m)} \mid X_0^{(m)} = j] \leqq E[Z_0 \mid X_0 = j]$ for all m. Also, $E[Z_0^{(m)} \mid X_0^{(m)} = i] \to E[Z_0 \mid X_0 = i]$. Thus, for all m large enough, for all $j > M$ (M specified in (5)), $E[Z_0^{(m)} \mid X_0^{(m)} = j] < \delta$. Further, since f is non-decreasing, for all $m \geqq i$,

$$E[f(X_1^{(m)}) \mid X_0^{(m)} = j] = \sum_{k=1, k \neq i}^{m} f(k) P[X_1 = k \mid X_0 = j]$$

$$+ f(i) P[X_1 = i \mid X_0 = j] + f(i) P[X_1 > m \mid X_0 = j]$$

$$\leqq \sum_{k=1}^{\infty} f(k) P[X_1 = k \mid X_0 = j] = E[f(X_1) \mid X_0 = j].$$

Hence the other conditions in (5) are satisfied for all $\{(X_k^{(m)}, Z_k^{(m)})\}$ whenever they are satisfied for $\{(X_k, Z_k)\}$. Thus we obtain the following result.

Theorem 2. *Let $\{(X_k, Z_k)\}$ be a Markov modulated arrival process with $\{X_k\}$ a countable state, irreducible, aperiodic ergodic chain and let $E_\pi[Z_1] < 1$. Also, let $\{(X_k^{(m)}, Z_k^{(m)})\}$ be a Markov modulated stream obtained from $\{(X_k, Z_k)\}$ as defined above. Then the stationary distributions of the workload and waiting time processes* *for* $\{(X_k^{(m)}, Z_k^{(m)})\}$ *converge to that of the queue with arrivals* $\{(X_k, Z_k)\}$. *Also, if* $\{(X_k, Z_k)\}$ *satisfies* (4), (5) *and* (11) *for a non-decreasing f, then the stationary moments of the workload and waiting time processes of all orders less than* $s-1$ *also converge.*

In Theorem 2 the only extra condition we need is (11). If the Markov chain $\{X_k\}$ has an upper Hessenberg structure, i.e. $P[X_1 = j \mid X_0 = k] = 0$ for $k > j+1$, then we need not fix state i for all Markov chains in Theorem 2. Hence instead of i for $\{(X_k^{(m)}, Z_k^{(m)})\}$ we could have made that state m and the stationary distributions of $\{X_k^{(m)}\}$ would still converge to those of $\{X_k\}$ (see [7]). One advantage of choosing state m is that in many practical queueing systems this naturally happens (e.g. the single server queue studied in the next section). Secondly, Theorem 2 then holds without the extra condition (11). Indeed, everything goes through in the same way except that $\sup_{j,m} E[|f(X_1^{(m)}) - f(j)|^s \,|X_0^{(m)} = j] < \infty$ requires verification.

But now, since f is non-decreasing,

$$E[|f(X_1^{(m)}) - f(j)|^s \,|X_0^{(m)} = j] = \sum_{n=1}^{m} |f(n) - f(j)|^s P[X_1^{(m)} = n \mid X_0^{(m)} = j]$$

$$= \sum_{n=1}^{m-1} |f(n) - f(j)|^s P[X_1 = n \mid X_0 = j] + [f(m) - f(j)]P[X_1 \geqq m \mid X_0 = j]$$

$$\leqq \sum_{n=1}^{\infty} |f(n) - f(j)|^s P[X_1 = n \mid X_0 = j] = E[|f(X_1) - f(j)|^s \mid X_0 = j]$$

and hence condition (4) is satisfied for all m and j.

The particular definition of $P[Z_1^{(m)} = n \mid X_1^{(m)} = j]$ given in (7) was just to make sure that $E_\pi[(Z_1^{(m)})^\alpha] \to E_\pi[Z_1^\alpha]$, $Z_1^{(m)} \xrightarrow{w} Z_1$ and $P[Z_1^{(m)} = n \mid X_1^{(m)} = j] \to P[Z_1 = n \mid X_1 = j]$. When $\{X_k\}$ has upper Hessenberg structure and we are taking state i to be m as above then we can show that $P_\pi[X_1^{(m)} = j] = P_\pi[X_1 = j]/P_\pi[X_1 \leqq m]$. Now taking $P[Z_1^{(m)} = n \mid X_1^{(m)} = j] = P[Z_1 = n \mid X_1 = j]$ for all n and j satisfies the above requirements and we might as well take these as probabilities.

Instead of upper Hessenberg structure, if $\{X_k\}$ has a state j_0 and an $\varepsilon > 0$ such that $P(X_1 = j_0 \mid X_0 = i) > \varepsilon$ for all i then again (from [7]) we can take state m instead of a particular state i and hence (11) is not necessary for Theorem 2.

It is possible to generalize the results of this section at the expense of more complex notation. For example, the service time (or the 'mark') of each customer can be more than one slot. For an i.i.d. service such an extension has been studied in [13] for some related results for this queue. One can further generalize by making the service time distribution also dependent upon the state of the modulating Markov chain $\{X_k\}$.

In applying Theorems 1 and 2 the main restrictions, which often come up, are in verifying conditions $E[Z_0 \mid X_0 = x] < \delta$ for all x large enough and (11). These are needed only for the convergence of moments. Sometimes $E[\tau^s] < \infty$ can be satisfied

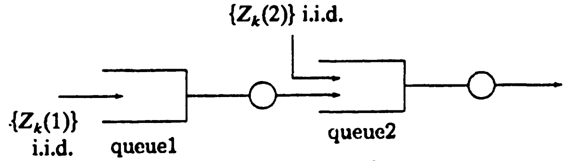

Figure 1. Example 1

under weaker conditions, and then we obtain our results under more realistic assumptions—as we see in the next section.

4. Some explicit examples

We provide a few examples of discrete time queues when the arrival process is modulated by a countable state Markov chain. We obtain a sequence of approximating queues with arrivals modulated by finite state queues such that the workload and waiting time stationary distributions and moments converge to that of the given queue. As mentioned in Section 1 in contrast to the general scheme of Asmussen and Koole [2] our approximating queues are discrete and we also obtain convergence of moments. In addition, the approximating queues are obtained in a natural, intuitive and easy way. The examples are useful in practice, illustrate a use of our results in Section 3, and the results appear to be new. Thus, we can obtain algorithms to compute the quantities of interest from these approximate systems.

Example 1. Consider a system of two discrete queues, each with infinite buffers (see Figure 1). Queue 1 is fed with an i.i.d. input $\{Z_k(1)\}$. After service the customers from queue 1 enter queue 2. There is another i.i.d. stream $\{Z_k(2)\}$ entering queue 2. At each queue there is one server (the generalization to multiserver queues is immediate).

The service time of each customer is one unit time (the generalization to i.i.d. service time is also immediate). The output process of queue 1 is a natural process which is modulated by the countable state Markov chain $\{W_k(1)\}$, the workload process at queue 1.

We assume that the customer who is served in slot k at queue 1 reaches just before time $k+1$ to queue 2 to be available for service at time $k+1$. Thus, if D_k are the arrivals to queue 2 at time $k+1$ from queue 1 then $D_k = 1\{W_k(1) > 0\}$, and denoting $Y_k = Z_k(2) + D_k$, we get

$$W_{k+1}(1) = (W_k(1) - 1)^+ + Z_k(1), \qquad W_{k+1}(2) = (W_k(2) - 1)^+ + Y_k,$$

where $\{W_k(2)\}$ is the workload process at queue 2. We assume that $E[Z_1(1) + Z_1(2)] < 1$. Then the system $\{(W_k(1), W_k(2))\}$ is stable and has a unique stationary distribution. Of course, computing stationary distributions and means will require the finite state approximations.

Since calculating the distributions and moments of $\{W_k(2)\}$ only cause a problem, we concentrate on it. The natural approximating arrival process for queue 2 is obtained if we replace the first queue by a finite buffer of length m. For such an approximation, we denote the various parameters by an additional superscript, e.g. $W_k^{(m)}(1)$ and $W_k^{(m)}(2)$. Now, $\{W_k^{(m)}(1)\}$ becomes a finite state modulating chain for queue 2. Since its transition probabilities satisfy the upper Hessenberg property mentioned at the end of the last section, its transition probability matrix satisfies the truncation requirements of [7]. Also we have $P[Y_1^{(m)}=n \mid W_1^{(m)}(1)=j]=P[Y_1=n \mid W_1(1)=j]$ for all m, n and j. Thus all the requirements for Theorem 2 (modified for upper Hessenberg at the end of the last section) for convergence of stationary distributions of waiting times and workload process at queue 2 are satisfied (of course, the time dependent probabilities of these processes also converge).

To obtain the convergence of stationary moments the only problem is in verifying assumption (5). However, to verify $\sup_m E[(\tau^{(m)})^s]<\infty$, $E[\tau^s]<\infty$, one can use a direct argument that will provide convergence of the stationary moments of waiting times and workload processes at queue 2 for all moments less than $s-1$. First let us specify the regeneration epochs for the mth system to be the times when $(W_k^{(m)}(1), W_k^{(m)}(2))=(0,0)$. Then comparing the mth system (where the first queue has buffer length m) with the original system we see that, if both are fed the same input $\{(X_k(1), X_k(2))\}$, then $W_k^{(m)}(1)\leqq W_k(1)$ and $W_k^{(m)}(2)\leqq W_k(2)$ for all $k\geqq 0$. Hence $\tau^{(m)}\leqq\tau$ for all m. Now we provide conditions for $E[\tau^s]<\infty$. First consider a discrete single server queue fed by both streams $\{Z_k(1)\}$ and $\{Z_k(2)\}$. Then it is known that the regeneration length has finite sth moment ($s\geqq 1$) if $E[Z_1(1)+Z_1(2)]<1$ and $E[(Z_1(i))^s]<\infty$ for $i=1,2$. Now compare this queue with the original system when both the systems are fed by the streams $\{Z_k(1), Z_k(2)\}$ and at time $k=0$ both systems are empty. When for the first time the single queue is empty, the original system will be in state $(0,0)$ or $(0,1)$. If in $(0,1)$ then, with probability $P[Z(1)=0]P[Z(2)=0]$, in the next step it will be in state $(0,0)$. Thus by the coin tossing argument we get $E[\tau^s]<\infty$ whenever $E[(Z_1(i))^s]<\infty$, $i=1,2$.

Example 2. Again we consider the system of two queues studied in Example 1, except that $\{(Z_k(1), Z_k(2))\}$ are now modulated by a countable state Markov chain $\{X_k\}$. If $\{Z_k(1)\}$ and $\{Z_k(2)\}$ are modulated by two separate countable state chains then this case can be converted to the case of the single chain $\{X_k\}$. Thus we consider a single chain $\{X_k\}$. Since queue 1 has already been studied, we concentrate on queue 2. The total input to queue 2 can be considered as Markov modulated by the countable state chain $\{(W_k(1), X_k)\}$. If $E_\pi[Z_k(1)]+E_\pi[Z_k(2)]<1$ and $\{X_k\}$ is aperiodic, irreducible and positive recurrent then we know that queue 2 is stable and has a unique stationary distribution. Now we obtain an approximation of queue 2 by queues modulated by finite state Markov chains. This can be done in two steps.

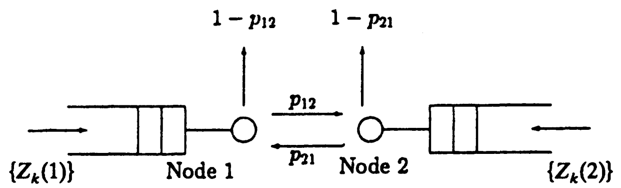

Figure 2. Discrete time Jackson network

First we approximate $\{X_k\}$ by $\{X_k^{(m)}\}$ obtained by truncating the state space as in Theorem 2. Then Theorem 2 can be used to show that the traffic entering the queues satisfies (as $m \to \infty$) the requirements in Borovkov's conditions and hence the stationary distribution of the work load and waiting times at queue 2 converge. If the chain $\{X_k\}$ also satisfies conditions (4), (5) and (11) for some $s > 1$ then again as in example 1 by comparing with a single server queue with input $\{(Z_k(1) + Z_k(2), X_k)\}$ we can get $E[\tau^s] < \infty$ and hence the convergence of all moments less than $s-1$. Now, as a next step one can replace queue 1 by a finite buffer queue with buffer size M. Then again one can show for any particular $\{X_k^{(m)}\}$ chain specified that, as $M \to \infty$, the stationary distributions and moments of workload and waiting time at queue 2 converge as $M \to \infty$ to the needed quantities. Thus choosing m and M appropriately will provide the desired approximation.

Example 3. We now study a discrete time Jackson network as described in [14] (see Figure 2). For simplicity we take a two node network each with infinite buffers.

Let $Z_k(i)$ be the number of external packets arriving in slot k at node i. We assume $\{Z_k(i)\}$ to be i.i.d. After service at node i, a packet goes to node j with probability p_{ij}, $j \neq i$ and leaves the system otherwise. Let $W_k(i)$ be the number of packets at time k at node i. The service times are one slot each. Let $R_k(i)$ be r.v.s independent of everything else, and $P(R_k(i) = 1) = p_{ij} = 1 - P(R_k(i) = 0)$. These r.v.s denote the routing of the packets which have been serviced in slot k. Define

$$Y_k(i) = 1\{W_k(i) > 0\}1\{R_k(i) = 1\}. \tag{12}$$

Then

$$W_{k+1}(i) = (W_k(i) - 1)^+ + Z_k(i) + Y_k(j), \qquad j \neq i. \tag{13}$$

The ergodicity of the process $\{(W_k(1), W_k(2))\}$ has been obtained in [14] when the total arrival intensity (as in the continuous Jackson network) at each node is less than one. Consider queue 1. Observe that the total arrival process at node 1 can be considered as Markov modulated by $\{W_k(2)\}$. Replacing queue 2 by a finite buffer queue will provide the necessary approximation. In the following we replace both queues with queues of buffer length m. The corresponding r.v.s will be denoted by $W_k^{(m)}(i)$, $Y_k^{(m)}(i)$ while $Z_k(i)$ and $R_k(i)$ will remain the same. Taking $W_0^{(m)}(i) = 0$, inductively (on k) from (12) and (13) we see that, as $m \to \infty$, $W_k^{(m)}(i) \to W_k(i)$ and $Y_k^{(m)}(i) \to Y_k(i)$, and the convergence is monotonically increasing. Denoting the corresponding stationary distributions with r.v.s $W^{(m)}(i)$, $Y^{(m)}(i)$, $W(i)$, and $Y(i)$, we can also easily show that

$$W^{(m)}(i) \leqq_{st} W^{(m+1)}(i) \leqq_{st} W(i), \qquad Y^{(m)}(i) \leqq_{st} Y^{(m+1)}(i) \leqq_{st} Y(i),$$

and $W^{(m)}(i) \xrightarrow{w} W, Y^{(m)}(i) \xrightarrow{w} Y(i)$ for $i = 1,2$ where $X \leqq_{st} Y$ means $P(X > a) \leqq P(Y > a)$, for all a. Because of monotonicity we also have convergence of all the moments. The conditions for $E[(W(i))^{\alpha}] < \infty$ for $\alpha > 0$ are provided in [14].

Various generalizations of Example 3 are possible. Instead of two nodes, we can have a finite number of nodes; the nodes can have multiple servers; and for these results the arrival processes and the routing variables could make a sequence of stationary ergodic r.v.s. (see [14]). Of course, for general stationary sequences, the finite buffer approximation will not provide a Markov modulated queue.

Acknowledgements

This research was partially supported by INRS-Telécom while the author was on leave from the Indian Institute of Science. The author thanks Professor Ravi Mazumdar and INRS-Télécom for their hospitality.

Appendix: Proof of sufficiency of (1) for Borovkov's conditions

We need to show that:

(i) $E_\pi[Z_1^{(m)} - 1; Z_1^{(m)} > 1] \to E_\pi[Z_1 - 1; Z_1 > 1]$, and

(ii) finite-dimensional distributions of $\{Z_k^{(m)}\}$ under stationarity, converge to that of $\{Z_k\}$.

Since

$$E_\pi[Z_1^{(m)}; Z_1^{(m)} > 1] - P_\pi[Z_1^{(m)} > 1] = E_\pi[Z_1^{(m)}] - P_\pi(Z_1^{(m)} = 1) - P_\pi[Z_1^{(m)} > 1]$$

we obtain (i) using (ii).

We will prove (ii) only for one- and two-dimensional distributions. The general case follows in the same way. Now,

$$\lim_{m\to\infty} P_\pi[Z_1^{(m)} = k] = \lim_{m\to\infty} \sum_j P[Z_1^{(m)} = k \mid X_1^{(m)} = j] P_\pi[X_1^{(m)} = j]$$

$$= \sum_j P[Z_1 = k \mid X_1 = j] P_\pi[X_1 = j],$$

where the second equality follows from [13], p. 270 once we realize that weak convergence for discrete r.v.s. implies the convergence in total variation.

Next we consider two-dimensional distributions. Observe that

$$
\begin{aligned}
P_\pi[Z_2^{(m)}=k_2,\, Z_1^{(m)}=k_1] &= \sum_{j_1} P[Z_2^{(m)}=k_2 \mid Z_1^{(m)}=k_1,\\
&\qquad X_1^{(m)}=j_1]P_\pi[X_1^{(m)}=j_1,\, Z_1^{(m)}=k_1]\\
&= \sum_{j_1}\sum_{j_2} P[Z_2^{(m)}=k_2 \mid X_2^{(m)}=j_2,\, Z_1^{(m)}=k_1,\, X_1^{(m)}=j_1]\\
&\qquad \times P[X_2^{(m)}=j_2,\, Z_1^{(m)}=k_1 \mid X_1^{(m)}=j_1]P_\pi[X_1^{(m)}=j_1]\\
&= \sum_{j_1}\Big(\sum_{j_2} P[Z_2^{(m)}=k_2 \mid X_2^{(m)}=j_2]P[X_2^{(m)}=j_2,\\
&\qquad Z_1^{(m)}=k_1 \mid X_1^{(m)}=j_1]\Big)P_\pi[X_1^{(m)}=j_1].
\end{aligned}
$$

Again by [13], p. 270,

$$
\begin{aligned}
&\sum_{j_2} P[Z_2^{(m)}=k_2 \mid X_2^{(m)}=j_2]P[X_2^{(m)}=j_2,\, Z_1^{(m)}=k_1 \mid X_1^{(m)}=j_1]\\
&\qquad\qquad \to \sum_{j_2} P[Z_2=k_2 \mid X_2=j_2]P[X_2=j_2,\, Z_1=k_1 \mid X_1=j_1].
\end{aligned}
$$

Now apply this to the last equality, and again use [13], p. 270, to obtain

$$
P_\pi[Z_2^{(m)}=k_2,\, Z_1^{(m)}=k_1] \to P_\pi[Z_2=k_2,\, Z_1=k_1].
$$

References

[1] ASMUSSEN, S. (1987) *Applied Probability and Queues.* Wiley, Chichester.

[2] ASMUSSEN, S. AND KOOLE, G. (1993) Marked point process as limits of Markovian arrival streams. *J. Appl. Prob.* **30,** 365–372.

[3] BOROVKOV, A. A. (1976) *Stochastic Processes in Queuing Theory.* Springer, Berlin.

[4] BOROVKOV, A. A. (1978) Ergodicity and stability theorems for a class of stochastic equations and their applications. *Theory Prob. Appl.* **23,** 227–247.

[5] BRANDT, A., FRANKEN, P. AND LISEK, B. (1990) *Stationary Stochastic Models.* Wiley, Chichester.

[6] DALEY, D. J. AND ROLSKI, T. (1992) Finiteness of waiting time moments in general stationary single server queue. *Ann. Appl. Prob.* **2,** 987–1008.

[7] GIBSON, D. AND SENETA, E. (1987) Augmented truncation of infinite stochastic matrices. *J. Appl. Prob.* **24,** 600–608.

[8] GUT, A. (1988) *Stopped Random Walks.* Springer, New York.

[9] HERMANN, C. (1993) Analysis of discrete time *SMAP/D/1/s* finite buffer queue with applications in ATM. In *INFOCOM '93.*

[10] LUCANTONI, D. (1991) New results on a single server queue with a batch Markovian arrival process. *Stoch. Models* **7,** 1–46.

[11] KALASHNIKOV, V. (1994) *Topics on Regenerative Processes.* CRC Press, Boca Raton, FL.

[12] NEUTS, M. F. (1981) *Matrix Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore, MD.

[13] ROYDEN, H. L. (1989) *Real Analysis.* Macmillan, Basingstoke.

[14] SHARMA, V. (1993) Open queuing networks in discrete time—some limit theorems. *Queueing Systems* **14**, 159–175.

[15] SHARMA, V. (1993) Markov modulated queues in discrete time—some limit theorems. *IISc-ISTC technical report.* Bangalore.

[16] SHARMA, V. (1995) Analysis of discrete time queues with applications to ATM based BISDNS. *IISc-ISTC technical report.*

[17] TAKINE, T., SUDA, T. AND HASEGAWA, T. (1993) Cell loss and output process analysis of a finite buffer discrete time ATM queuing system with correlated arrivals. In *INFOCOM'93.*

[18] WIRTH, K. D. (1982) On stationary queues with batch arrivals. *Inf. Kybernet.* **18**, 603–619.