Perfect sampling of a single-server queue with periodic Poisson arrivals

Yaofei Xiong · Duncan J. Murdoch ·
David A. Stanford

Received: 10 September 2014 / Revised: 8 December 2014 / Published online: 24 December 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract In this paper we present algorithms for the perfect sampling of single-server time-varying queues with periodic Poisson arrivals under the first come first served (FCFS) discipline. The service durations have periodically time-dependent exponential ($M_t/M_t/1$) or homogeneous general ($M_t/G/1$) distributions. Assuming a cycle length of 1, we construct discrete dominating processes at the integer instants $n \in \{0, \pm 1, \dots\}$. Perfect sampling of the $M_t/M_t/1$ queue is obtained using dominated CFTP (Kendall and Møller 2000) when the system is relatively lightly loaded or with the regenerative method (Sigman 2012) in the general case. For the $M_t/G/1$ queue, perfect sampling is achieved with dominated CFTP.

Keywords Time-varying queues · Periodic Poisson process · Perfect sampling · Regenerative method · Dominated CFTP

Mathematics Subject Classification 60J22 · 65C05 · 60K25 · 68U20

1 Introduction

Time-varying queueing models are more realistic than time-homogeneous queues, but they are not usually mathematically tractable [17, p. 697]. As noted by [14], computational methods and approximation techniques involved in time-varying queueing problems have long been regarded as challenging.

In this paper we consider cases in which the time-dependent stochastic processes follow periodic patterns. Hasofer [10] showed that the Laplace-Stieltjes Transform (LST) of virtual waiting time in the $M_t/G/1$ queue is asymptotically periodic in time.

Y. Xiong · D. J. Murdoch (✉) · D. A. Stanford
Department of Statistical and Actuarial Science, University of Western Ontario, London, ON, Canada
e-mail: murdoch@stats.uwo.ca

Harrison and Lemoine [9] proved that the virtual waiting time at any given time has its own limiting distribution, and there is one such distribution for each point within the period of the queue. Asmussen and Thorisson [4] extended the context to more general cases, where the inter-arrival times and service durations both depend on the arrival instant within the time period. They proved that with more conditions (such as Harris ergodicity [1, p. 202] of the phase parameter which the inter-arrival time and service duration depend on), the virtual waiting time and queue length also have time-dependent limiting distributions in periodic patterns.

Due to the complexity of the time-varying systems, only asymptotic solutions have been developed, and this has happened gradually over recent decades. By assuming some state at time 0, [22] and [13] found the transient distributions of the number of customers in the system (Q_t) in $M_t/M_t/1$ and $M_t/M_t/c$ queues, respectively, using generating functions and Volterra integral equations. Zeifman et al. [21] approximated the limiting mean value ($\mathbb{E}(Q_t)$) of the $M_t/M_t/1$ queue with the transient distribution of the truncated time-varying birth and death processes by restricting their difference to some controllable extent. The asymptotic periodic solutions for $M_t/M_t/1$ and $M_t/M_t/c$ systems were achieved by Margolius [14], where distributions and moments are given in terms of integral equations.

If we could draw samples directly from the steady-state distribution, statistical inference would become straightforward. Perfect sampling is an approach to draw a sample directly from the steady-state distribution without explicitly solving for it. The first well-known perfect sampling algorithm is commonly referred to as coupling from the past (CFTP), introduced by [15]. “Dominated CFTP” [11] is an important extension of CFTP. It enables the coupling of Markov chains with unbounded state spaces by reducing the number of past chains that need to be simulated.

Recently, [18] applied dominated CFTP to achieve perfect sampling of an $M/G/c$ queue with a “super stable” (i.e., $\rho < 1/c$) condition. The backward simulation of an $M/G/1$ queue was implemented by running the coupled $M/G/1$ Processor-sharing (PS) model, which is time reversible. Connor and Kendall [7] showed how to generalize the dominated CFTP idea used by [18] to relax the super stable restriction to $\rho < 1$. They further proposed a sandwiching dominated CFTP algorithm for the perfect sampling of the $M/G/c$ queue. It significantly reduces the expected runtime. However, in the time-varying circumstance ($M_t/G/1$), it is hard to construct a dominating process which empties, or whose upper and lower envelopes coalesce.

Other methods are also available for perfect sampling. In [2, p. 420] the perfect sampling of regenerative processes was described and then applied to the $M/G/c$ queue by [19]. A special case is perfect sampling of the $GI/G/1$ queue [2, p. 437], assuming that the Exponential Change of Measure (ECM) [1, p. 352] exists for the underlying random walk.

In practice, dominating processes are key elements for perfect sampling. They are processes defined on the same probability space, which bound the states of the target process, reducing the range of unknown values to a bounded set. In this paper, dominating processes are constructed by modifying the arrival instants or potential departure instants on each periodic cycle of the time-varying queues. In the $M_t/M_t/1$ queue, we can simulate the steady-state draw of the dominating process using the ECM method mentioned above for the $GI/G/1$ queue. For the $M_t/G/1$ queue, we

estimate the upper bound of the dominating process based on a coupled homogeneous queue.

In Sect. 2, we present our assumptions and notation. Section 3 presents perfect sampling of the $M_t/M_t/1$ queue in both the relatively lightly loaded and more general cases. In Sect. 4, we use dominated CFTP to achieve perfect sampling for the $M_t/G/1$ queue. Section 5 concludes the paper.

2 Assumptions and notation

To ensure consistency and clarity, we present our assumptions and notation which we will require in the subsequent analyses.

- Define $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{R} = (-\infty, \infty)$ and $\mathbb{R}^- = (-\infty, 0]$. Let x^+ be the non-negative part of x ; i.e., if $x > 0$, then x^+ equals x ; otherwise it is 0.
- All queueing systems involved in this paper are work conserving and non preemptive. Queueing disciplines are generally presumed to be First Come First Served (FCFS) unless described otherwise.
- Arrivals to the queue form a time-varying Poisson process, with non-trivial periodic arrival rate $\lambda(t) \geq 0$. Without loss of generality, assume the length of the period is 1. We also assume that “potential service events” form a time-varying Poisson process with periodic rate $\mu(t) \geq 0$, also with period 1. These are the departures that would occur if the system were to remain busy; there may be fewer actual departures if the queue empties. Both $\lambda(t) > 0$ and $\mu(t) > 0$ except possibly at discrete points, so their integrals are strictly increasing. Time “0” is congruent with our time point of interest. For both $\lambda(t)$ and $\mu(t)$, t is measured in clock time:

$$\lambda(t) = \lambda(t + 1), \text{ and } \mu(t) = \mu(t + 1), \quad \forall t \in \mathbb{R}.$$

- Define

$$\begin{aligned} \bar{\lambda} &= \int_0^1 \lambda(t) dt, \quad \bar{\mu} = \int_0^1 \mu(t) dt, \text{ as well as} \\ F_{\lambda}(t) &= \frac{\int_0^t \lambda(s) ds}{\bar{\lambda}}, \text{ and } F_{\mu}(t) = \frac{\int_0^t \mu(s) ds}{\bar{\mu}} \end{aligned} \quad (2.1)$$

for $t \in (0, 1]$. These functions are strictly increasing on the defined interval, according to the definitions above of $\lambda(t)$ and $\mu(t)$. Therefore their inverse functions exist, and are denoted by $F_{\lambda}^{-1}(x)$ and $F_{\mu}^{-1}(x)$, $x \in (0, 1]$, respectively.

- To ensure the stability of the $M_t/M_t/1$ queue, the occupancy ρ must be less than unity; i.e.,

$$\rho = \frac{\bar{\lambda}}{\bar{\mu}} < 1.$$

See [4, Theorem 4.5] for more details. For the $M_t/G/1$ queue, we define $\mu = 1/\mathbb{E}(B)$ where B is the homogeneous service duration, and for stability,

$$\rho = \frac{\bar{\lambda}}{\mu} < 1.$$

- For the time-varying queues, let N_k^A be the number of arrivals on the interval $(k - 1, k]$, $k \in \mathbb{Z}$, so

$$N_k^A \sim \text{Poi}(\bar{\lambda}).$$

The N_k^A 's constitute an i.i.d. sequence of random variables.

- In the $M_t/M_t/1$ queue, let N_k^D be the number of potential departures on the interval $(k - 1, k]$, $k \in \mathbb{Z}$. Then

$$N_k^D \sim \text{Poi}(\bar{\mu}).$$

The N_k^D 's also constitute an i.i.d. sequence of r.v.'s, and they are independent of the N_k^A 's. When the server is idle, potential departure events have no effect.

- In the algorithms that follow we will couple homogeneous queues to the time-varying queues that we are studying. Denote by Q_t^N and Q_t^H the numbers of customers at time t in the time-varying and homogeneous queues, respectively. Similarly, let V_t^N and V_t^H be the unfinished workloads in these two systems, respectively. These processes are all right continuous in t .

3 Perfect sampling of the $M_t/M_t/1$ queue

In this section, we present perfect sampling for the $M_t/M_t/1$ queue using one of two methods, depending upon the occupancy level in the queue. For the lightly loaded case where the minimum service rate is greater than the maximum arrival rate, dominated CFTP works as a straightforward solution. In the general setting, where we only have $\bar{\lambda} < \bar{\mu}$, we achieve the perfect sampling using an ECM to sample from the GI/G/1 queue and the regenerative method to extend this to the time-varying queue.

3.1 Perfect sampling of the $M_t/M_t/1$ queue with $\inf \mu(t) > \sup \lambda(s)$

Let $\mu_l = \inf \mu(t)$ and $\lambda_u = \sup \lambda(t)$. Assume

$$\mu_l > \lambda_u.$$

A stable M/M/1 queue can be generated with arrival and service rates of λ_u and μ_l , respectively, since $\mu_l > \lambda_u$. Based on its homogeneous arrival and potential departure events, the time-varying inputs of the coupled $M_t/M_t/1$ queue are simulated as follows:

- The time-varying arrival events are filtered from the homogeneous arrivals using the thinning method [17, p. 697, Method 1], thinning arrivals to rate $\lambda(t) \leq \lambda_u$.

- Time-varying potential departure events are reproduced based on the homogeneous ones by supplementing the homogenous events with events from a time-varying Poisson process at rate $\mu(t) - \mu_l$. These extra events are generated with the inter-event time method [17, p. 702, Method 3]. Thus, the composite potential departure process is a superposition of the homogeneous and the time-varying parts, due to the aggregation property of independent Poisson processes.

Under the coupling scheme described above, it is easy to see that the coupled M/M/1 queue dominates the time-varying one in Q_t (the number of customers in the system), because the arrivals in the time-varying queue are a subset, and the potential departures a superset, of those in the coupled M/M/1 queue.

Conceptually, we start the dominating homogeneous queue and the coupled time-varying queue infinitely long ago. At time 0, both of them are in steady state. In a past time $\tau \in \mathbb{R}^-$, if $Q_\tau^H = 0$, then the coupled time-varying queue must also be empty at this time. By running it forward with the time-varying events generated as above, we get a steady-state draw of the time-varying queue at time 0.

In practice, only a finite number of values are needed. The algorithm is described as follows:

1. We simulate the M/M/1 dominating queue's stationary value at time 0. In stationarity $Q_0^H \sim \text{Geom}(1 - \rho_0)$, where $\rho_0 = \lambda_u/\mu_l$ [12, p. 96].
2. Simulate the M/M/1 queue backwards with parameters λ_u and μ_l until it becomes idle at time $\tau \in \mathbb{R}^-$, where $\tau = \sup\{t : t \in \mathbb{R}^-, Q_t^H = 0\}$. This step is implemented based on the time reversibility of the M/M/1 queue [17, p. 399, Proposition 6.5]. (See Algorithm 1 for the pseudocode for steps 1 and 2.) Record the times of the arrival and departure instants of the time-homogeneous queue on the interval $[\tau, 0)$. If $\tau = 0$, return $Q_0^N = 0$. Otherwise, continue.
3. Filter the arrival events generated by step 2 above by thinning as follows: an arrival event at time ζ is retained with probability $\lambda(\zeta)/\lambda_u$, otherwise it is deleted from the collection of arrival events.
4. Supplement the potential departures according to a Poisson process with rate $\mu(t) - \mu_l$. Based upon our assumptions in Sect. 2, since $\int_0^1 \mu(t)dt = \bar{\mu} < \infty$, we can proceed as follows: Let $t_0 = \tau$, and repeat the two sub-steps below for $n = 0, 1, \dots, N$, where $t_N = \max\{t_k : t_k < 0, k = 0, 1, \dots\}$.
 - (1) Simulate X_{t_n} from the distribution with c.d.f.

$$F_{t_n}(x) = 1 - e^{-\int_0^x [\mu(t_n+s) - \mu_l] ds},$$

where the subscript t_n indicates that $F_{t_n}(x)$ depends on t_n .

- (2) Assign $t_{n+1} = t_n + X_{t_n}$.

If $N > 0$, then append $\{t_1, \dots, t_N\}$ to the collection of potential departure events, and sort these in ascending order, yielding the aggregate collection of potential departure events for the time-varying queue.

5. Starting from the empty state at time $\tau -$ with the set of arrival and potential departure instants generated in the previous two steps, run the time-varying system forward until time 0 and output the state Q_0^N as a steady-state draw at an integral time for the M_t/M_t/1 queue.

Algorithm 1 M/M/1 queue backward simulation

```

1: Initialize vectors ARRIVALS and PDEPARTURES to empty.
2:  $t \leftarrow 0$  # event instant, initialized as zero
3: Simulate  $Q \sim \text{Geom}(1 - \rho_0)$ , where  $\rho_0 = \lambda_u / \mu_l$ 
4: while  $Q > 0$  do
5:   Simulate  $X \sim \text{Exp}(\lambda_u + \mu_l)$ 
6:    $t \leftarrow t + X$ 
7:   Simulate  $E$  from  $\{1, -1\}$  with  $\Pr(E = 1) = \lambda_u / (\lambda_u + \mu_l)$ 
8:   if  $E = 1$  then
9:     Append  $t$  to ARRIVALS
10:  else
11:    Append  $t$  to PDEPARTURES
12:  end if
13:   $Q \leftarrow (Q + E)^+$ 
14: end while
15:  $\tau \leftarrow -t$ 
16: Change the signs of ARRIVALS and PDEPARTURES then reverse their order.
17: Rename ARRIVALS as PDEPARTURES' and PDEPARTURES as ARRIVALS'
18: return  $\tau$ , ARRIVALS' and PDEPARTURES'.

```

3.2 Perfect sampling of the $M_t/M_t/1$ queue: general case

In this section, we use the regenerative method to perform perfect sampling of the $M_t/M_t/1$ queue with the general stationary condition, i.e., $\bar{\lambda} < \bar{\mu}$.

We start with a general description of the regenerative method to obtain a stationary draw from our distribution of interest, which we adapt from [19] in what follows. Let X_n , $n = 0, 1, \dots$ denote the number of customers in a stable queue just before the $(n + 1)^{\text{st}}$ arrival, with $X_0 = 0$. Then $\{X_n\}_{n \geq 0}$ is a positive recurrent non-delayed discrete-time regenerative process with $X_n = 0$ as the regenerative setting. Assume its cycle length is $T \in \mathbb{N}$, with $\mathbb{E}(T) < \infty$, where the cycle length is defined as

$$T = \min\{n : n \geq 1, X_n = 0\}$$

with $X_0 = 0$.

Explicitly, a generic cycle with length T can be defined as

$$C = \{X_n : 0 \leq n < T\}.$$

It is easy to simulate i.i.d. cycles and the sequentially generated ones are denoted by

$$C^{(j)} = \{X_n^{(j)} : 0 \leq n < T^{(j)}\}, j \geq 1,$$

with corresponding cycle lengths $T^{(j)}$.

Denote by T^e a random variable which has the equilibrium distribution of the cycle length. Suppose we can sample T^e , and let $J = \min\{j \geq 1 : T^{(j)} \geq T^e\}$, then we have a steady-state draw of $\{X_n\}_{n \geq 0}$ as

$$X_{T^e}^{(J)}.$$

(Note that, if $T^{(J)} = T^e$, then $X = 0$.) The interested reader will find a proof of the correctness of the regenerative method to generate a stationary draw in [19]. More details on regenerative methods can be found in [2, p. 420] and [3].

The key to our algorithm is to construct a dominating process which can be simulated in steady-state. If we start with a realization of the time-varying system, and concentrate all arrivals to the end of the interval $(k - 1, k]$, $k \in \mathbb{N}$, and all potential departures to the beginning of it, the modified process would dominate the original one in the sense that at whole integer times it will have at least as many customers waiting. Intuitively, since more potential departure events might be lost to an empty queue due to the postponing of the arrival events, there would tend to be more customers remaining in the system after the arrivals than at the same instant in the original process. This idea is explicitly stated by the following proposition.

Proposition 1 *Construct a process by modifying a simulation of a stable $M_t/M_t/1$ queue as follows. On each interval $(k - 1, k]$, $k \in \mathbb{Z}$, let the number of arrivals in the $M_t/M_t/1$ queue be N_k^A , and the number of potential departures be N_k^D . In the modified queue let N_k^A customers arrive as a batch just before time k , and let N_{k+1}^D potential departures occur just after time k . Denote by L_k the number of customers counted at time k of the modified process, and by Q_k^N that in the corresponding $M_t/M_t/1$ queue. If $L_{k_0} = Q_{k_0}^N = 0$ for some k_0 , then the modified system dominates the original one in the number of customers at all integer times after k_0 :*

$$L_k \geq Q_k^N, \forall k \geq k_0.$$

Proof At the non-integer points, we define

$$L_t = \left(L_{k-1} - N_k^D \right)^+, \forall t \in (k - 1, k). \quad (3.1)$$

It is obvious that

$$L_k \geq N_k^A, \forall k \in \mathbb{Z},$$

since no matter what the system's state is, the N_k^A arrivals provide a lower bound.

It is clear that when $k = k_0$ the inequality is true. Assume that when $k = m$, $m \geq k_0$, $m \in \mathbb{Z}$, the inequality holds, then when $k = m + 1$ we have one of the following situations:

1. If $\exists t \in (m, m + 1)$ such that $Q_t^N = 0$, then $Q_{m+1}^N \leq N_{m+1}^A \leq L_{m+1}$.
2. Otherwise, $Q_t^N > 0$, $\forall t \in (m, m + 1)$ (i.e., the time-varying queue keeps busy on this interval), so that

$$Q_{m+1}^N = Q_m^N - N_{m+1}^D + N_{m+1}^A.$$

- (1) If $L_t > 0$, $\forall t \in (m, m + 1)$, then

$$L_{m+1} = L_m - N_{m+1}^D + N_{m+1}^A,$$

and it follows that

$$L_{m+1} - Q_{m+1}^N = L_m - Q_m^N \geq 0.$$

(2) Otherwise $\exists t \in (m, m+1)$ such that $L_t = 0$, then it must be the case that

$$L_m \leq N_{m+1}^D \Rightarrow Q_m^N \leq N_{m+1}^D.$$

So

$$Q_{m+1}^N = Q_m^N - N_{m+1}^D + N_{m+1}^A \leq N_{m+1}^A \leq L_{m+1}.$$

Thus in both cases, the inductive step is established, and the result follows. $\square$

It is clear that $\{L_k\}_{k \geq 0}$ is a non-delayed regenerative process with $L_k = 0$ as the regenerative setting. So its cycle length can be defined as

$$T = \min\{k : k \geq 1, L_k = 0\}, \quad (3.2)$$

with $L_0 = 0$. The generic cycle is defined as

$$C = \{L_k : k = 0, \dots, T-1\}. \quad (3.3)$$

The cycle starts with value zero and lasts through the subsequent positive values of L_k . For example, the L_k sequence $\{0, 3, 1, 0, \dots\}$ yields $C = \{0, 3, 1\}$ with length 3 and $\{0, 0, 3, \dots\}$ yields $C = \{0\}$ with unit-length.

Using (3.1) for the definition of L_t , let

$$\mathcal{L}_k = L_{k-0.5}. \quad (3.4)$$

Then we have

$$\mathcal{L}_k = \left(\mathcal{L}_{k-1} + N_{k-1}^A - N_k^D \right)^+, \quad (3.5)$$

$$L_k = \mathcal{L}_k + N_k^A. \quad (3.6)$$

It is clear that N_k^A is independent of $\mathcal{L}_k$, since $\mathcal{L}_k$ is determined by $N_i^A (i < k)$ and $N_i^D (i \leq k)$, and N_k^A is independent of these r.v.'s as defined in Sect. 2. The limiting random variable L_∞ is defined by $\lim_{k \rightarrow \infty} L_k$. A segment of a sample path of the dominating process is shown in Fig. 1. It resembles the Late Arrival System discrete queue of [6], but the differences preclude us from using the LAS model directly. Instead, we exploit the form of (3.5) directly.

Fig. 1 Construction of the dominating process of the $M_t/M_t/1$ queue

3.2.1 Sampling from the steady-state of the dominating process

Equation (3.5) has the form of Lindley's equation of the waiting time in a GI/G/1 queue, and it leads to a special perfect sampling algorithm as shown by [2, p. 437] and [8]; we repeat the algorithm below, for reasons of completeness.

Let $Z_k = N_{k-1}^A - N_k^D$, $k \in \mathbb{N}$. The differences Z_k constitute an i.i.d. sequence, which we generically denote by $Z = N^A - N^D$. In light of (3.5) we find

$$\mathcal{L}_k = (\mathcal{L}_{k-1} + Z_k)^+.$$

Starting from $\mathcal{L}_0 = 0$, $S_0 = 0$, define $S_k = \sum_{i=1}^k Z_i$, $k \in \mathbb{N}$. Then $\{S_k\}_{k \geq 0}$ is a random walk with negative drift, since $\mathbb{E}(Z) = \mathbb{E}(N^A) - \mathbb{E}(N^D) = \bar{\lambda} - \bar{\mu} < 0$. It is shown in [2, p. 3] that

$$\mathcal{L}_k \stackrel{\mathcal{D}}{=} \max_{i=0,1,\dots,k} S_i.$$

So the limiting random variable $\mathcal{L}_\infty$, defined by $\lim_{k \rightarrow \infty} \mathcal{L}_k$, satisfies

$$\mathcal{L}_\infty \stackrel{\mathcal{D}}{=} \max_{k \geq 0} S_k.$$

To perform the Exponential Change of Measure (ECM) [2, p. 129], solve

$$M_Z(\gamma) = 1 \tag{3.7}$$

for $\gamma > 0$, where

$$M_Z(t) = \mathbb{E}(e^{tZ}) = e^{\bar{\lambda}e^t + \bar{\mu}e^{-t} - \bar{\lambda} - \bar{\mu}}.$$

The changed measure is given by

$$\mathbb{P}_\gamma(z) = e^{\gamma z} \mathbb{P}(z), z \in \mathbb{Z},$$

where $\mathbb{P}$ stands for the original measure.

Equation (3.7) has the equivalent form

$$\bar{\lambda}e^\gamma + \bar{\mu}e^{-\gamma} - \bar{\lambda} - \bar{\mu} = 0.$$

Let $g(\theta) = \bar{\lambda}e^\theta + \bar{\mu}e^{-\theta} - \bar{\lambda} - \bar{\mu}$. Since $g(0) = 0$, $g'(0) = \bar{\lambda} - \bar{\mu} < 0$, so $\exists \theta^* > 0$, such that $g(\theta^*) < 0$. Since $\lim_{\theta \rightarrow \infty} g(\theta) = \infty$, it follows that $g(\theta) = 0$ has a positive root on the interval (θ^*, ∞) . Furthermore, $g''(\theta) = \bar{\lambda}e^\theta + \bar{\mu}e^{-\theta} > 0$, so $g(\theta)$ is convex, and γ is the unique root of $g(\theta) = 0$ on (θ^*, ∞) .

Assume non-negative integers a and d are the observations of N^A and N^D , respectively. Let $z = a - d$ represent the corresponding observation of Z . Since N^A and N^D are independent, it follows that

$$\begin{aligned} \mathbb{P}_\gamma(Z = z) &= \frac{e^{\gamma z} \mathbb{P}(Z = z)}{M_Z(\gamma)} \\ &= \frac{1}{M_Z(\gamma)} \sum_{a, d: a-d=z} e^{\gamma(a-d)} \mathbb{P}(N^A = a) \mathbb{P}(N^D = d) \\ &= \sum_{a, d: a-d=z} \frac{(\bar{\lambda}e^\gamma)^a e^{-\bar{\lambda}e^\gamma}}{a!} \frac{(\bar{\mu}e^{-\gamma})^d e^{-\bar{\mu}e^{-\gamma}}}{d!}. \end{aligned}$$

Hence under the measure $\mathbb{P}_\gamma$, Z can be treated as the difference of two Poisson r.v.'s: N^{A*} and N^{D*} , which satisfy

$$N^{A*} \sim \text{Poi}(\bar{\lambda}e^\gamma) \text{ and } N^{D*} \sim \text{Poi}(\bar{\mu}e^{-\gamma}).$$

Consequently

$$\begin{aligned} \mathbb{E}_\gamma(Z) &= \bar{\lambda}e^\gamma - \bar{\mu}e^{-\gamma} \\ &= g'(\gamma) > 0 \end{aligned}$$

due to the convexity of $g(\theta)$ as shown above. This implies that $\{S_k\}_{k \geq 0}$ becomes a random walk with positive drift under the measure of $\mathbb{P}_\gamma$.

So Z (under the measure $\mathbb{P}_\gamma$) can be simulated by generating N^{A*} and N^{D*} from their respective distributions, and then taking the difference.

Now using $\mathbb{P}_\gamma$, define a strictly increasing process with ladder heights $S_{\tau(n)}$, $n = 0, 1, \dots$, where

$$\tau(0) = 0, \tau(n+1) = \inf\{k > \tau(n) : S_k > S_{\tau(n)}\}.$$

Let $\mathcal{L} = \sup\{S_{\tau(n)} : S_{\tau(n)} \leq V\}$, where $V \sim \text{Exp}(\gamma)$. Then $\mathcal{L}$ is a stationary draw of $\mathcal{L}_\infty$. Thus the stationary draw of L_∞ given by

$$L_\infty = \mathcal{L} + N^A,$$

where $N^A \sim \text{Poi}(\bar{\lambda})$, and N^A is independent of $\mathcal{L}$.

3.2.2 Algorithm for perfect sampling of the $M_t/M_t/1$ queue

Based on the constructed dominating process, whose stationary state can be simulated, the perfect sampling of the $M_t/M_t/1$ queue is available using the regenerative method.

1. Simulate a random variable (denoted by T^e) from the equilibrium distribution of the cycle length (3.2) of the regenerative process of $\{L_k\}_{k \geq 0}$, which dominates the queue length process in the time-varying queue at the integer time points. We obtain T^e as follows. At time 0, sample a stationary draw of the dominating process, denoted by L_0 , using the method presented in the previous subsection. Continue simulating this process forward until it becomes 0. According to equations (3.5) and (3.6), we have $L_k = (L_{k-1} - N_k^D)^+ + N_k^A, k \in \mathbb{N}$. So

$$T^e = \min\{k \geq 1, L_k = 0\}.$$

2. Sequentially simulate generic cycles $C^{(j)} = \{L_k^{(j)} : 0 \leq k < T^{(j)}\}, j = 1, 2, \dots$, of the dominating process, where $T^{(j)}$ is the length of the j^{th} cycle. Record N_k^A and N_k^D ($1 \leq k \leq T^{(j)}$), until $T^{(j)} \geq T^e$, where

$$J = \min\{j : T^{(j)} \geq T^e\}.$$

3. Use the order statistics method of simulating the time-varying Poisson process (see [17, p. 700, Method 2]) to construct time-varying events (arrival and potential departure instants) according to N_k^A and N_k^D ($1 \leq k \leq T^{(j)}$) generated in cycle $C^{(j)}$. The coupling scheme specified in Proposition 1 implies that the time-varying instant can be generated by shifting the coupled homogeneous one on each interval $(k-1, k]$. Let t^N and t^H be the instants in the time-varying and homogeneous systems, respectively, on this interval. For each such instant of a homogeneous arrival or service event on the interval, we know that independent of all other events, it will be uniformly distributed on $(k-1, k]$. The corresponding time-varying instant can be computed as

$$t^N = \lfloor t^H \rfloor + F^{-1}(t^H - \lfloor t^H \rfloor), \quad (3.8)$$

where F^{-1} corresponds to the inverse of functions $F_\lambda(t)$ or $F_\mu(t)$ defined in equation (2.1). From time 0, where the system is empty, simulate forward with these inputs to restore the time-varying queue. Output $Q_{T^e}^N$ as the stationary draw of the number of customers in the $M_t/M_t/1$ queue at integer time points.

- Remark 1* (1) According to the regenerative method, $L_{T^e}^{(J)}$ is a steady-state draw from the dominating process. Since it is coupled with the time-varying queue, at this time point (T^e), the corresponding sample of $Q_{T^e}^N$ is also stationary.
- (2) At time 0, even if the stationary draw of L_0 equals zero, we still continue simulating forward.
- (3) Since $\{L_k\}_{k \geq 0}$ is the dominating process, we can also directly output $Q_0^N = 0$ if $L_0 = 0$. But in this case, the condition of stopping the generic cycle simulation becomes

$$J = \min\{j : T^{(j)} > T^e\}.$$

- (4) Unfortunately the regenerative method has infinite expected runtime [5, 20]. Thus some runs may take so long that a practitioner would abort them, introducing a bias into the simulation.

3.2.3 An example

Let

$$\lambda(t) = 1 + \sin(2\pi t), \quad \mu(t) = 4 + 2 \cos(2\pi t).$$

These parameters are the same as those used by [14], and have $\bar{\lambda} = 1$ and $\bar{\mu} = 4$.

The regenerative method described in the previous subsection has been applied for the simulation. On a unit cycle, we chose 100 points at equal spacing from 0 to 1 and generated 10,000 samples for each point. Since only Q_0^N is generated in each trial, to get the samples at different points we changed the phases of the sinusoid functions in each run. For every point we repeated the algorithm 10,000 times. Although it is time consuming (around 100 times more than simulating the successive 99 points by continuing running the system after sampling the first point with the regenerative method), the simulation shows that our method works quite well. In this example, we did not need to abort any long runs. Since the occupancy (0.25) is quite small, they are very rare.

The idle probability and expected number of customers at time $t \in (0, 1)$ of the time-varying queue are illustrated in Fig. 2. The gray areas indicate point-wise 95 % confidence intervals. The time average of Q_t^N is around 0.383. The simulated values match very well with the analytical results derived by [14, Section 3], which involve solving a Volterra equation of the second kind numerically.

For a more efficient simulation, we could repeat the algorithm 10,000 times for just one phase, and continue simulating the process through a whole cycle using standard forward simulation methods. In this case, samples at different time points in the simulation would be correlated and could be studied as draws from their joint distribution.

Fig. 2 Idle probabilities and expected numbers in the $M_t/M_t/1$ queue for one period. 100 points were chosen on it with equal intervals. 10,000 samples were drawn for each point

4 Perfect sampling of the $M_t/G/1$ queue

In this section, we present a perfect simulation algorithm with a service time distribution that does not vary with time. Service durations (denoted by B) are drawn from some general distribution $G(\cdot)$. We require that $\mathbb{E}(B^2) < \infty$ in order to ensure that the algorithm for the backward simulation of the coupled $M/G/1$ queue has finite expected runtime. Since the service requirement of a customer can be considered to be known at an arrival instant, we can take the perspective of analyzing the unfinished workload to explore this time-varying system. As a result, this case is easier to handle than the $M_t/M_t/1$ queue.

As was the case before, the first step is to find a dominating process. In the next subsection we construct a process which dominates the time-varying queue in the unfinished workload. Using dominated CFTP [11] we achieve perfect sampling of the $M_t/G/1$ queue.

4.1 The dominating process

Proposition 2 *Construct a coupled homogeneous queue ($M/G/1$) by modifying a stable $M_t/G/1$ queue as follows. On each interval of $(k-1, k]$, $k \in \mathbb{Z}$, let the number of arrivals in the $M_t/G/1$ queue be N_k^A . In the homogeneous queue let N_k^A customers arrive uniformly on this interval. Let the service durations in the homogeneous queue be the same values in the same order as those in the time-varying queue. Denote by V_k^H the unfinished workload at time k in the homogeneous queue, and by V_k^N that in the time-varying queue. Assume both of them are initially idle at time $t_0 \in \mathbb{Z}$. Then*

$$V_k^N \leq V_k^H + 1, \quad \forall k \in \mathbb{Z}, k \geq t_0.$$

Proof We use mathematical induction to prove this proposition. Clearly $V_k^N \leq V_k^H + 1$ for $k = t_0$, since both are 0. For larger k , let ξ_k be the additional workload that arrives during the interval $(k - 1, k]$ (the same for both queues). Let η_k^N be the amount of work done on these new customers during the interval $(k - 1, k]$ in the time-varying queue, and let η_k^H be the counterpart in the homogeneous queue.

On the interval $(k - 1, k]$, unless the server finishes the remaining workload (carried from previous intervals) within that cycle, it cannot direct any capacity to serve the new arrivals on this interval. Therefore we obtain

$$\begin{aligned} V_k^N &= (V_{k-1}^N - 1)^+ + \xi_k - \eta_k^N, \\ V_k^H &= (V_{k-1}^H - 1)^+ + \xi_k - \eta_k^H. \end{aligned}$$

Note that $\eta_k^N \in [0, 1)$, and that $\eta_k^N = 0$ if $V_{k-1}^N > 1$, as there remains earlier work in the system to be done. $V_{k-1}^N + \eta_k^N \leq 1$ if $V_{k-1}^N \leq 1$, since the earlier work has been finished within a unit interval. Similar constraints hold for η_k^H .

So

$$V_k^N - V_k^H = (V_{k-1}^N - 1)^+ - (V_{k-1}^H - 1)^+ + \eta_k^H - \eta_k^N.$$

If $V_{k-1}^N \leq 1$, then $V_k^N - V_k^H \leq \eta_k^H \leq 1$, so our result holds. Otherwise $V_{k-1}^N > 1$, then $\eta_k^N = 0$, and

$$\begin{aligned} V_k^N - V_k^H &\leq V_{k-1}^N - 1 + \eta_k^H \\ &\leq V_{k-1}^H + \eta_k^H \quad (\text{by the inductive hypothesis}) \\ &\leq 1 \end{aligned}$$

which again agrees with our result. $\square$

4.2 Backward simulation of the M/G/1 queue

The previous subsection established that the upper bound of the dominating process can be estimated with the unfinished workload of the coupled homogeneous M/G/1 queue. To perform the dominated CFTP algorithm, we simulate the M/G/1 queue backwards using the time reversibility of its Processor-sharing (PS) variant. (Under the PS discipline, the customers share the server, i.e., when n customers are present, the server devotes $1/n$ of its capacity to each. The customers attain service at rate $1/n$. The first such customer leaves the system once the attained service reaches the minimum of the residual service times, unless another customer arrives first in which case the rates are reduced to $1/(n + 1)$ [1, p. 63].)

Ross [16, p. 280] showed that the M/G/1 PS model is time reversible and at stationarity the number of customers (Q) in the system has the geometric distribution

$$\Pr(Q = k) = \rho^k (1 - \rho), k = 0, 1, \dots,$$

where ρ is the occupancy; the completed (or unfinished) workload of the M/G/1 PS queue is

$$V = \sum_{i=1}^Q Y_i,$$

where the Y_i 's are i.i.d. and they follow the equilibrium distribution of the service duration.

Since the sample paths of the unfinished workload for the M/G/1 FCFS queue are identical to those of the M/G/1 PS model, we can use the latter to achieve the backward simulation of the M/G/1 FCFS queue. Sigman [18, Algorithm 1.1, Step 1] described this algorithm, which we restate as follows for the sake of completeness:

1. Simulate a steady-state draw from the M/G/1 PS queue at time 0: $Q, Y_1, \dots, Y_Q$ as specified above. If $Q = 0$, return $\mathbf{0}$ as the sampled state, otherwise continue as follows.
2. Simulate the M/G/1 queue forward in time with the PS discipline until the server becomes idle; denote this time by ζ . In this process, record the departure instants and associated service requirements for all customers.
3. Reverse time: treat each departure as an arrival, changing the signs of their times of occurrence and looking backwards in time, implement the FCFS variant of the M/G/1 queue.

Details can be found in Algorithm 2, where $H(\cdot)$ stands for the spread distribution (the length-biased distribution of a randomly selected service duration) of the generic service duration B . It has c.d.f. [18, Remark 1.1].

$$H(x) = \mu \int_0^x \overline{G}(y) dy - \mu x \overline{G}(x), \quad (4.1)$$

where $\overline{G}(x) = 1 - G(x)$, $x \geq 0$.

In Algorithm 2, the vector INSTANTS is used to store the departure instants (when running the PS model forward) and arrival instants (after being reversed), and SERVICES the associated service requirements. They are initially empty. When looking backwards, ζ is the stationary age of the busy period of the M/G/1 FCFS queue.

It is well known (for example, [12, pp. 213–214]) that

$$\mathbb{E}(\zeta) = \frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)(1 - \rho)^2} \quad (4.2)$$

so we need the existence of $\mathbb{E}(B^2)$ in order for this algorithm to work in finite time.

4.3 Algorithm for perfect sampling of the $M_t/G/1$ queue

The perfect sampling of the $M_t/G/1$ queue is performed using dominated CFTP. Imagine that an $M_t/G/1$ queue and the coupled dominating process $\{V_k^H + 1\}_{k \in \mathbb{Z}}$,

Algorithm 2 M/G/1 FCFS queue backward simulation [18]

```

1: Initialize vectors INSTANTS and SERVICES to be empty.
2: Simulate  $Q \sim \text{Geom}(1 - \rho)$ 
3: if  $Q = 0$  then
4:   return 0
5: else
6:   for  $i = 1$  to  $Q$  do
7:     Simulate  $U_i \sim \text{Unif}(0, 1)$  # standard uniform distribution
8:     Simulate  $X_i \sim H(\cdot)$  #  $X$ : a vector of associated service requirements
9:      $Y_i \leftarrow U_i X_i$  #  $Y$ : a vector of residual works
10:   end for
11:    $t \leftarrow 0$  #  $t$  traces the event instants
12:   Simulate  $a \sim \text{Exp}(\lambda)$  #  $a$ : time to the earliest arrival instant from  $t$ 
13:   while  $Q > 0$  do
14:      $j \leftarrow \arg \min_i \{Y_i, i = 1, \dots, Q\}$ 
15:      $d \leftarrow QY_j$  #  $d$ : time to the earliest departure instant from  $t$ 
16:     if  $d < a$  then
17:        $t \leftarrow t + d; a \leftarrow a - d$  # departure event
18:       for  $i = 1$  to  $Q$  do
19:          $Y_i \leftarrow Y_i - d/Q$ 
20:       end for
21:       Append  $t$  to INSTANTS and  $X_j$  to SERVICES
22:       Remove the  $j^{\text{th}}$  entries from  $Y$  and  $X$ 
23:        $Q \leftarrow Q - 1$ 
24:     else
25:        $t \leftarrow t + a$  # arrival event
26:       for  $i = 1$  to  $Q$  do
27:          $Y_i \leftarrow Y_i - a/Q$ 
28:       end for
29:        $Q \leftarrow Q + 1$  # it implies an element is appended to  $X$ , same for  $Y$ 
30:       Simulate  $X_Q \sim G(\cdot)$ 
31:        $Y_Q \leftarrow X_Q$ 
32:       Simulate  $a \sim \text{Exp}(\lambda)$ 
33:     end if
34:   end while
35:    $\zeta \leftarrow$  the last element in INSTANTS
36:   Change signs of INSTANTS, reverse orders of INSTANTS and SERVICES
37:   return  $-\zeta$ , INSTANTS and SERVICES
38: end if

```

specified in Proposition 2, were started from infinitely long ago, i.e. $t_0 = -\infty$, with empty states, so that at time 0, they must be in steady-state. Since the coupled M/G/1 queue can be simulated backwards (Sect. 4.2), this algorithm can be described as follows:

1. Starting from time 0, simulate backwards the M/G/1 queue until it becomes idle for the first time. Denote this time by $-\zeta \in \mathbb{R}$.
2. Continue simulating the M/G/1 queue backwards until time $-\zeta_1$, determined as follows. It should be the start of a busy cycle, and the summation of the lengths of the idle periods on the interval $(\lceil -\zeta_1 \rceil, \lfloor -\zeta \rfloor)$ should exceed 2. Record the homogeneous arrival instants and service durations on the interval $(\lceil -\zeta_1 \rceil, 0)$.
3. Use equation (3.8) to determine the corresponding time-varying arrival instants on the interval $(\lceil -\zeta_1 \rceil, 0)$ from the arrival instants recorded above.

4. Starting from $\lceil -\zeta_1 \rceil$ with unfinished workload $V_{\lceil -\zeta_1 \rceil}^N = V_{\lceil -\zeta_1 \rceil}^H + 1$, run the time-varying queue forward with the collection of time-varying arrival instants and the service durations from the previous steps, and output V_0^N as a stationary draw of the unfinished workload at time 0 of the $M_t/G/1$ queue.

Proposition 3 *By following the algorithm for simulating the $M_t/G/1$ queue specified above, the output of V_0^N is a stationary draw of the unfinished workload at time 0.*

Proof Assume an $M_t/G/1$ queue and an $M/G/1$ queue were started infinitely long ago and coupled in the way described in Proposition 2. So $V_k^N \leq V_k^H + 1, \forall k \in \mathbb{Z}$.

The algorithm as constructed above ensures that there exists idle time on interval $(\lceil -\zeta_1 \rceil, \lfloor -\zeta \rfloor)$ in the coupled time-varying queue. When mapping the homogeneous arrivals to time-varying ones, the maximum work originally finished after $\lfloor -\zeta \rfloor$ that could be rearranged prior to $\lfloor -\zeta \rfloor$ is $V_{\lfloor -\zeta \rfloor}^H$. However, since V_t^H becomes zero on the interval $[\lfloor -\zeta \rfloor, \lfloor -\zeta \rfloor + 1)$, it follows that $V_{\lfloor -\zeta \rfloor}^H < 1$. According to the dominance, it follows that $V_{\lceil -\zeta_1 \rceil}^N \leq V_{\lceil -\zeta_1 \rceil}^H + 1$. Thus based on the homogeneous system the extra work introduced from the customers arriving prior to $\lceil -\zeta_1 \rceil$ is no greater than 1. Consequently, the 2 units of idle time on interval $(\lceil -\zeta_1 \rceil, \lfloor -\zeta \rfloor)$ are guaranteed by the algorithm to be sufficient to ensure that there is idle time for the time-varying queue on this interval.

Finally, this idle time existence ensures the coalescence of all possible chains of the time-varying queue. At time $\lceil -\zeta_1 \rceil$, all possible chains of the $M_t/G/1$ queue have unfinished workload ranging from 0 to $V_{\lceil -\zeta_1 \rceil}^H + 1$. They are driven by the same arrival instants and associated homogeneous service durations, and it is easy to see that this coupling is monotone. So when the upper bound chain (corresponding to $V_{\lceil -\zeta_1 \rceil}^N = V_{\lceil -\zeta_1 \rceil}^H + 1$) becomes zero, so do all the coupled chains, ensuring coalescence. The usual CFTP argument then establishes our result. $\square$

4.4 Examples

Here we illustrate sampling the stationary unfinished workload in the $M_t/G/1$ queue with Pareto and Erlang distributions of the service durations. In both cases, the arrival rates are

$$\lambda(t) = 3 + 3 \sin(2\pi t).$$

The service rates are $\mu = 4 = 1/\mathbb{E}(B)$.

- In the Pareto case the c.d.f. of service duration distribution has the form

$$G(x) = 1 - \left(\frac{\theta}{x + \theta} \right)^\alpha, x > 0$$

with $\alpha = 5$ and hence $\theta = \frac{\alpha-1}{\mu} = 1$.

- For the Erlang case

$$B \sim \Gamma(2, \theta),$$

Fig. 3 Average unfinished workloads and 95 % confidence intervals (*areas in gray shadow*) in two $M_t/G/1$ queues with Pareto and Erlang distributions of the service durations. They involve 2 cycles and 100 points are drawn in each cycle. For each point we generate 10,000 samples

where $\Gamma(\alpha, \theta)$ stands for the standard Gamma distribution with shape parameter α and rate θ . Here we have

$$\theta = \alpha\mu = 8.$$

The average unfinished workload and 95 % confidence intervals are illustrated in Fig. 3. It is evident from the figure that they behave in periodic patterns with the same periodic lengths as the arrival processes.

5 Conclusions

In this paper, we have presented algorithms for perfect sampling of the $M_t/M_t/1$ and $M_t/G/1$ queues, where the time-varying ingredients have periodic patterns. The stationary queue length and virtual waiting time (unfinished workload) have time-dependent distributions.

For the $M_t/M_t/1$ queue with relatively light load, i.e. $\sup \lambda(s) < \inf \mu(t)$, we implemented perfect sampling with the dominated CFTP by constructing a homogeneous $M/M/1$ queue with arrival and service rates of $\sup \lambda(s)$ and $\inf \mu(t)$, respectively. In the general case where $\bar{\lambda} < \bar{\mu}$, we constructed a discrete dominating process, and applied the regenerative method to get the time-dependent steady-state draws of the time-varying queue.

Perfect sampling of the $M_t/G/1$ queue was achieved using dominated CFTP. The dominating process was computed based on the unfinished workload of the coupled

homogeneous queue ($M/G/1$), which can be simulated backwards using its Processor-sharing variant. Unlike most CFTP implementations, our algorithm allows direct calculation of the (random) starting time in the past, rather than the usual trial and error approach. This is in common with the dominated CFTP algorithm for the $M_t/M_t/1$ queue, and also with the dominated CFTP algorithm by [18]. As with other CFTP algorithms, it yields a draw from the stationary distribution at time 0. Other time points in the cycle are simulated simply by adjusting the origin of the clock.

Acknowledgments This work was supported by NSERC Discovery Grants to Drs. Murdoch and Stanford.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

1. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
2. Asmussen, S., Glynn, P.: Stochastic Simulation: Algorithms and Analysis. Springer, New York (2007)
3. Asmussen, S., Glynn, P.W., Thorisson, H.: Stationarity detection in the initial transient problem. *ACM Trans. Model. Comput. Simul.* **2**, 130–157 (1992)
4. Asmussen, S., Thorisson, H.: A Markov chain approach to periodic queues. *J. Appl. Probab.* **24**(1), 215–225 (1987)
5. Blanchet, J., Dong, J.: Perfect sampling for infinite server and loss systems. *Adv. Appl. Probab.* (2014, forthcoming)
6. Chaudhry, M.L., Gupta, U.C.: Queue-length and waiting-time distributions of discrete-time $GI^X/Geom/1$ queueing systems with early and late arrivals. *Queueing Syst.* **25**, 307–324 (1997)
7. Connor, S.B., Kendall, W.S.: Perfect simulation of $M/G/c$ queues. [arXiv:1402.7248v1](https://arxiv.org/abs/1402.7248v1) (2014)
8. Ensor, K.B., Glynn, P.W.: Simulating the maximum of a random walk. *J. Stat. Plan. Inference* **85**, 127–135 (2000)
9. Harrison, J.M., Lemoine, A.J.: Limit theorems for periodic queues. *J. Appl. Probab.* **14**(3), 566–576 (1977)
10. Hasofer, A.M.: On the single-server queue with non-homogeneous poisson input and general service time. *J. Appl. Probab.* **1**(2), 369–384 (1964)
11. Kendall, W.S., Møller, J.: Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. Appl. Probab.* **32**(3), 844–865 (2000)
12. Kleinrock, L.: Queueing Systems volume 1: Theory. Wiley, New York (1975)
13. Margolius, B.H.: A sample path analysis of the $M_t/M_t/c$ queue. *Queueing Syst.* **31**, 59–93 (1999)
14. Margolius, B.H.: Transient and periodic solution to the time-inhomogeneous quasi-birth death process. *Queueing Syst.* **56**, 183–194 (2007)
15. Propp, J.G., Wilson, D.B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algorithms* **9**, 223–252 (1996)
16. Ross, S.M.: Stochastic Processes, 2nd edn. Wiley, New York (1996)
17. Ross, S.M.: Introduction to Probability Models, 10th edn. Academic Press, Burlington (2010)
18. Sigman, K.: Exact simulation of the stationary distribution of the FIFO $M/G/c$ queue. *J. Appl. Probab.* **48A**, 209–213 (2011)
19. Sigman, K.: Exact simulation of the stationary distribution of the FIFO $M/G/c$ queue: the general case for $\rho < c$. *Queueing Syst.* **70**, 37–43 (2012)
20. Xiong, Y., Murdoch, D.J., Stanford, D.A.: Perfect and nearly perfect sampling of work-conserving queues (2014, Submitted)
21. Zeifman, A., Leorato, S., Orsingher, E., Satin, Y., Shilova, G.: Some universal limits for nonhomogeneous birth and death processes. *Queueing Syst.* **52**, 139–151 (2006)
22. Zhang, J.: The transient solution of time-dependent $M/M/1$ queues. *IEEE Trans. Inf. Theory* **37**, 1690–1696 (1991)