2013 Performance evaluation of an importance sampling technique in a Jackson network

3 authors, including:

Saeed Setayeshi

Amirkabir University of Technology (Tehran …

152 PUBLICATIONS **597** CITATIONS

SEE PROFILE

Amir Masoud Rahmani

Islamic Azad University Tehran Science and …
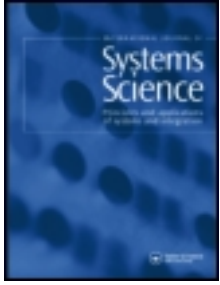
199 PUBLICATIONS **619** CITATIONS

SEE PROFILE

Performance evaluation of an importance sampling technique in a Jackson network

Ebrahim Mahdipour [a] , Amir Masoud Rahmani [a] & Saeed Setayeshi [b]

[a] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[b] Department of Medical Radiation, AmirKabir University of Technology, Tehran, Iran

PLEASE SCROLL DOWN FOR ARTICLE

Performance evaluation of an importance sampling technique in a Jackson network

Ebrahim Mahdipour[a]*, Amir Masoud Rahmani[a] and Saeed Setayeshi[b]

*[a]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran;
[b]Department of Medical Radiation, AmirKabir University of Technology, Tehran, Iran*

Importance sampling is a technique that is commonly used to speed up Monte Carlo simulation of rare events. However, little is known regarding the design of efficient importance sampling algorithms in the context of queueing networks. The standard approach, which simulates the system using an *a priori* fixed change of measure suggested by large deviation analysis, has been shown to fail in even the simplest network settings. Estimating probabilities associated with rare events has been a topic of great importance in queueing theory, and in applied probability at large. In this article, we analyse the performance of an importance sampling estimator for a rare event probability in a Jackson network. This article carries out strict deadlines to a two-node Jackson network with feedback whose arrival and service rates are modulated by an exogenous finite state Markov process. We have estimated the probability of network blocking for various sets of parameters, and also the probability of missing the deadline of customers for different loads and deadlines. We have finally shown that the probability of total population overflow may be affected by various deadline values, service rates and arrival rates.

Keywords: importance sampling; rare event; queueing network; deadline; performance evaluation

1. Introduction

Rare event simulation has piqued the interest of researchers ever since the first development of Monte Carlo techniques on computers at Los Alamos National Laboratory (Rubino and Tuffin 2009) and involves estimating extremely small but important probabilities. Despite the amount of work on the topic in the last 60 years, there are still domains needing to be explored because of new the applications (Conley 2007; Kroese, Taimre, and Botev 2011). The estimation of rare event probabilities has garnered a lot of attention in queueing theory. In the wake of their generic structure, queues have figured prominently in applied probability; for example various applications can be found in inventory, logistics, military, call centers, communication networks. The main motivation for studying these rare events lies in the fact that the events under consideration relate to situations whose happening can be very costly for queue network to face. The most common approach in rare event simulation is importance sampling. The general idea of importance sampling is to change the sampling distribution of the system under study to sample more frequently the events that are more 'important' for the simulation. Of course, using a new distribution results in a biased estimator if no correction is applied. Therefore, through multiplication with a so-called likelihood ratio (Rubino and Tuffin 2009), the simulation output is converted to the original measure.

An important benchmark problem in rare-event simulation, and importance sampling, in particular, is the problem of estimating the overflow probability of the total population of two Markovian queues in tandem (de Boer and Scheinhardt 2010). The classical papers on the application of importance sampling in queueing typically rely on a state-independent change of measure. In other words, for any system state, the probabilistic law is changed in the same manner. In de Boer (2006), the importance sampling with a state-independent change of measure has been used that is only applicable to special classes of queueing networks and/or buffer overflow probabilities and will not, in general, be asymptotically optimal for networks of queues (Dupuis and Wang 2009).

It was understood that the main problem of state-independent importance sampling schemes was that the transition rates were changed in a uniform manner, whether one of the queues was empty or not (Glasserman and Kou 1995). Thereby, it is far from certainty that the likelihood ratio is bounded on the event of interest. Therefore, the performance of importance sampling scheme proposed in Parekh and Walrand (1989) is far from ideal for some parameter values. de Boer, Nicola, and Rubinstein (2000) as well as Kroese and Nicola (2002) recommended some

*Corresponding author. Email: e.mahdipour@gmail.com

solutions to solve this problem, in which state-dependent importance sampling schemes were proposed that is, importance sampling distributions that are not uniform over the state space. The asymptotic efficiency for a state-dependent importance sampling scheme for estimating overflow probabilities in a d-node Jackson network was, firstly, proven in Dupuis, Sezer, and Wang (2007). Furthermore, a state-dependent importance sampling scheme for estimating the probability of overflow in the downstream queue of a Jacksonian two-node tandem queue was presented in (Miretskiy, Scheinhardt, and Mandjes 2010). The robustness of a discrete-time Markov chain was checked in L'ecuyer, Blanchet, Tuffin, and Glynn (2010) where state-dependent importance sampling was used to estimate first-passage time probabilities.

In all aforementioned papers, models of queueing networks have been taken account in which each customer enters the network and waits for receiving service with no time restriction. Once it receives the service from various nodes of network, it leaves the network. Based on such method, various schemes for simulating rare events in the various network models have been presented in the literature. A simple case of the above-mentioned models have been analysed in Mahdipour and Rahmani (2009) and Mahdipour, Rahmani, and Setayeshi (2009), where customers do not hurry to receive service, as just receiving the service is of importance. This type of modelling is employed in different applications such as some packet-switched telecommunication networks where no time period is guaranteed for packet transfer, and the transfer time depends on the traffic load as well as the network parameters. However, there are other models in which customers wait for service, but wait for a limited time only. In these models, the service time for a customer would be very important. In one node, it is meant; the sojourn time of a customer should not exceed a definite threshold, and if it does, the customer would not wait for receiving service from that node and leaves it. Such situations occur in the processing or merchandising of perishable goods (Barrer 1957a). Many types of military engagements are similarly characterized. An attacking airplane engaged by antiaircraft is available for 'service' – within range – for only a limited time.

The contribution of our work is twofold. First, we estimate the probability of a rare event known as total population overflow in a model of queueing network in which in some part of the system, there is no time limitation for receiving service, while the same customers should receive the needed service within a definite time once they enter another part of the system. Secondly, and more importantly, we consider the effect of various network parameters on the probability of total population overflow and show that how the probability of total population overflow affected by network parameters.

To the best of our knowledge, we have not seen such a modelling that enables us to estimate the probability of occurrence of rare events in networks where the sojourn time of customers is subject to limitation.

This article is organised as follows. Section 2 presents a brief review of related work. In Section 3, an analysis of the probability of the missing deadline in the two-node Jackson network and its dynamics are derived. In Section 4. the importance sampling estimator and the conditions to reach the asymptotic optimality are presented. Section 5 examines some examples to illustrate the efficacy of our model. Section 6 concludes this article.

2. Related work

Of note is (Parekh and Walrand 1989) which is one of the pioneering articles in importance sampling estimation of overflow probabilities in queueing network. Here, the estimation of the overflow probability of both a single queue, and of the total population in a network of queues is suggested to be done through a heuristically motivated state-independent change of measure. In this method, inter arrival and service time distributions of the model are simply superseded by other distributions, a replacement that stays constant during the entire simulation. For Jackson networks, the change of measure of Parekh and Walrand (1989) comes down to transpose the arrival rate with the lowest service rate. However, the general inefficiency of this change of measure was subsequently established (Glasserman and Kou 1995), while in some other cases, it conduces to an estimator with infinity variance (de Boer 2006). Nevertheless, there are few exceptions in the literature that have rigorously addressed the state-independent change of measure. In contrast, these can only be applied to special classes of queueing networks and/or buffer overflow probabilities (Dupuis and Wang 2009).

As the asymptotic optimality of state-independent change of measure for networks of queues cannot be reached (de Boer 2006), a search through the class of the dynamic (state-dependent) scheme is in order to reach optimality. Estimation of the overflow probability of the second buffer in a two-node Jackson network is achieved through a state-dependent change of measure (Kroese and Nicola 1999); the arrival and service rates are made dependent on the content of the first buffer. Some other examples of a state-dependent

change of measure are Mandjes and Ridder (2002), Heegaard (1998) and L'ecuyer et al. (2010).

A framework under which one can systematically build efficient state-dependent importance sampling schemes for simulating rare events in several models of queueing networks is presented in Dupuis et al. (2007), whose state-dependent change of measure is grounded in game theory. Through game theory, an equation for the optimal change of measure is derived, and an approximate solution to this equation is constructed. Dupuis et al. (2007) main finding endorses the asymptotic optimality of the change of measure associated to this approximate solution. The approach was illustrated by two models, of d-node Jackson networks and a two-node network with feedback. The rare event explored in the investigation was the total population overflow in the network. In the two-node network with feedback, the arrival rate of customers is considered identical and their difference lies in the service rate of the network nodes as well as the network feedback rate. Compared with Dupuis et al. (2007), we have taken four scenarios, and in each scenario, we have calculated network blocking probability for different feedback rates. We have also shown the effect of feedback rate on the network blocking probability in different scenarios implying more complete, comprehensive results than those of Dupuis et al. (2007). Moreover, we tend to model a system whose sojourn time is of importance in some parts, whereas in other parts providing service to the customers is done without any reference to sojourn time.

An analytical approximation for the overflow probability of interest that could be extended to more complicated networks than the two-node tandem queue was presented in de Boer and Scheinhardt (2010). Then, they plugged this approximation into the well-known expression for the zero-variance change of measure, arriving at a new change of measure. This new change of measure displayed a gradual transition from one measure to another, similar to the other changes of measure developed earlier; the significant difference, however, was that in their case, the transition and its width arises naturally, contrary to being inserted artificially to fix the behaviour at the boundary. Surprisingly, the transition shape is different. Likewise, the limiting change of measure at the state-space boundary was different.

In Miretskiy et al. (2010), the authors focused on estimating the probability of overflow in the downstream queue of a Jacksonian two-node tandem queue. They intended to develop a state-dependent change of measure, that proved to be asymptotically efficient. They focused on the probability of the second queue exceeding a certain predefined threshold before the system empties, and they also identified an asymptotically efficient importance sampling distribution for any initial state of the system. In Miretskiy et al. (2010), the choice of the importance sampling distribution was backed up by appealing heuristics that are rooted in large deviations theory.

3. Sojourn time and analysis of missing the deadline

As stated earlier, we tend to modelling the systems whose sojourn time is of importance in some parts, whereas in other parts providing service to the customers is done without any reference to sojourn time and some applications of this type of modelling are presented in previous section. The simplest model can be a two-node Jackson network where in the first node providing service to customers is conducted without any reference to sojourn time, while in the second node, the sojourn time of one customer should not be higher than a threshold. Therefore, once a customer receives its service from the first node receives a deadline as it enters the second queue. The deadline of a customer determines how long it can wait in the second node for service completion. If providing a service to a customer in the second node ends prior to the expire of the deadline, the customer leaves the network. However, if the deadline expires, it should return to the first queue even if it is receiving the service from the second node. Therefore, there should be a feedback from the second node to the first node.

Consider we have a two-node Jackson network with feedback (Figure 1) whose arrival process to the network is Poisson with rate λ and customers are served in the order of their arrival, or in other words, service discipline is First-Come-First-Served (FCFS). Service times are exponentially distributed with rates μ_1 and μ_2 at nodes one and two, and the arrival rate to the first queue is λ_1.

The difference between the deadline of a customer in the second queue and its arrival time from node one, referred to as a *relative deadline*, is a random variable η with mean value $\bar{\eta}$. In this article, we consider a model with deterministic customer impatience that has already been studied in Barrer (1957b). Deterministic customer impatience refers to the case where a customer whose sojourn time in the second queue is greater or equal to $\bar{\eta}$ becomes a lost customer irrespective of whether it is acquired for service or not. Thus, deterministic customer impatience has a probability distribution function described by

$$\begin{cases} D(\bar{\eta}, \tau) = 0, & \text{if } \tau < \bar{\eta} \\ D(\bar{\eta}, \tau) = 1, & \text{if } \tau \geqslant \bar{\eta}, \end{cases} \tag{1}$$
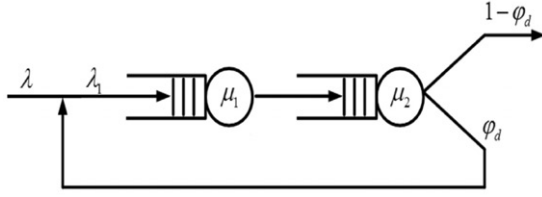
Figure 1. Two-node network with feedback.

where τ is a variable with values in the set of non-negative real numbers. In this network customer service times and relative deadlines form sequences of *iid* random variables.

Theorem 3.1: *If $s_1, s_2, \ldots, s_k$ are mutually independent, identically distributed exponential random variables with parameter μ, then the random variable $s_1 + s_2 + \cdots + s_k$ has an Erlang-k distribution with parameter μ.*

Suppose s_i is the service time of ith customer in the second node, and there are n customers in the second queue. As service time at node two is exponentially distributed with rate μ_2 and according to Theorem 3.1, $S(n) = s_1 + s_2 + \cdots + s_n$ has the following probability distribution function:

$$F_{S(n)}(\mu_2, \tau) = P(S(n) \leq \tau) = 1 - e^{-\mu_2 \tau} \sum_{i=0}^{n-1} \frac{(\mu_2 \tau)^i}{i!},$$
$$\text{if} \quad n > 0. \quad (2)$$

Let V be the time an arriving customer from node one with an infinite (no) deadline must wait before it completes its service at node two. V is called the *offered sojourn time* in the network and its distribution function is $F_V(\tau)$. Therefore, the *probability of missing the deadline* is

$$\varphi_d = P(V > \bar{\eta}) = \int_0^\infty D(\tau)\, \mathrm{d}F_V(\tau) = \int_{\bar{\eta}}^\infty f_V(\tau)\mathrm{d}\tau. \quad (3)$$

φ_d represents the steady-state probability that a customer misses its deadline and f_V is the probability density function of V.

To calculate the offered sojourn time, first the conditional offered sojourn time should be calculated as outlined below. Let V_n be the time an arriving customer from node one with an infinite (no) deadline must wait before it completes its service at node two, given it finds n customers at node two. V_n is called the *conditional offered sojourn time*, given there are n customers at node two. Clearly the probability distribution function $V_n, F_{V_n}(.)$, is given by

$$F_{V_0}(\tau) = P(V_0 \leq \tau) = 1 - e^{-\mu_2 \tau}$$
$$F_{V_n}(\tau) = P(V_n \leq \tau) = P(V_{n-1} + S(1) \leq \tau \mid V_{n-1} \leq \bar{\eta}),$$
$$\text{if} \quad n > 0 \quad (4)$$

Applying the definition of conditional probability to (4) yields

$$P(V_n \leq \tau) = \frac{P(V_{n-1} + S(1) \leq \tau,\ V_{n-1} \leq \bar{\eta})}{P(V_{n-1} \leq \bar{\eta})}. \quad (5)$$

Moreover, from the above definitions, it is easy to verify that $V_n = S(n+1)$, which gives

$$P(V_{n-1} + S(1) \leq \tau,\ V_{n-1} \leq \bar{\eta}) = P(S(1) \leq \tau - \bar{\eta}). \quad (6)$$

According to (2) and (6), we have

$$P(V_{n-1} + S(1) \leq \tau,\ V_{n-1} \leq \bar{\eta}) = 1 - e^{-\mu_2(\tau - \bar{\eta})} \quad (7)$$

and

$$P(V_{n-1} \leq \bar{\eta}) = 1 - e^{-\mu_2 \bar{\eta}} \sum_{i=0}^{n-1} \frac{(\mu_2 \bar{\eta})^i}{i!}. \quad (8)$$

Therefore, the probability distribution function of V_n is given by

$$F_{V_n}(\tau) = \frac{1 - e^{-\mu_2(\tau - \bar{\eta})}}{1 - e^{-\mu_2 \bar{\eta}} \sum_{i=0}^{n-1} \frac{(\mu_2 \bar{\eta})^i}{i!}}, \quad \text{if} \quad n > 0 \quad (9)$$

or, equivalently, the probability density function of V_n is

$$f_{V_n}(\tau) = \frac{\mu_2 e^{-\mu_2(\tau - \bar{\eta})}}{1 - e^{-\mu_2 \bar{\eta}} \sum_{i=0}^{n-1} \frac{(\mu_2 \bar{\eta})^i}{i!}}, \quad \text{if} \quad n > 0. \quad (10)$$

Let ψ_n be the probability that one of the customers in the second node misses its deadline, given there are n customers in the second node. ψ_n also is called *conditional loss rate* for n customers and is given by

$$\psi_0 = 0$$
$$\psi_n = \mu_2 \frac{F_{S(n-1)}(\mu_2, \bar{\eta}) - F_{S(n)}(\mu_2, \bar{\eta})}{F_{S(n)}(\mu_2, \bar{\eta})}, \quad \text{if} \quad n > 0$$
$$(11)$$

whose proof can be found in Movaghar (2006).

We assume that the second queue has a finite capacity K and q_n is the steady-state probability that there are n customers in the second queue, then we have the following steady-state recursive equations:

$$-\lambda_1 q_0 + (\mu_2 + \psi_1)q_1 = 0$$
$$\lambda_1 q_{n-1} - (\lambda_1 + \mu_2 + \psi_n)q_n + (\mu_2 + \psi_{n+1})q_{n+1} = 0$$
$$(12)$$

the normalising condition is

$$\sum_{n=0}^{K} q_n = 1 \quad (13)$$

which gives us

$$q_0 = \frac{1}{\sum_{i=1}^{K} \frac{\lambda_1^i}{\prod_{j=1}^{i}(\mu_2 + \psi_j)}} \quad (14)$$

and

$$q_n = \frac{\lambda_1^n}{\prod_{i=1}^n (\mu_2 + \psi_i)} q_0. \tag{15}$$

The probability density function of customer offered sojourn time can be determined as

$$f_V(\tau) = \sum_{n=0}^{K-1} \frac{\mathrm{d}F_{V_n}(\tau)}{\mathrm{d}\tau} q_n \tag{16}$$

and a closed-form solution for φ_d is derived as

$$\varphi_d = \sum_{n=0}^{K-1} q_n P(V_n > \bar{\eta}) = 1 - \sum_{n=0}^{K-1} q_n P(V_n \le \bar{\eta}) \tag{17}$$

Finally, by applying (8) and (15) to (17) the probability of missing the deadline can be computed as follows:

$$\varphi_d = 1 - \sum_{n=0}^{K-1} \left(\lambda_1^n \frac{1 - e^{-\mu_2 \bar{\eta}} \sum_{i=0}^n \frac{(\mu_2 \bar{\eta})^i}{i!}}{\prod_{i=1}^n (\mu_2 + \psi_i) \sum_{i=1}^K \frac{\lambda_1^i}{\prod_{j=1}^i (\mu_2 + \psi_j)}} \right). \tag{18}$$

3.1. *The system dynamics*

Let $N = \{N(k): k = 0, 1, 2, \ldots\}$ be the embedded discrete time Markov chain that represents the queue lengths at the transition epochs of the network, and suppose that $N(k) = (N_1(k), N_2(k))$ where $N_i(k)$ is the length of the queue at node i after the kth transition. Obviously, N can only take values in $\mathbb{Z}_+^2$ and its dynamics can be modelled by $N(k+1) = N(k) + \pi[N(k), Y(k+1)]$, where $\{Y(k)\}$ are *iid* random variables taking values in $\Omega = \{\omega_0 = (1, 0), \omega_1 = (-1, 1), \omega_2 = (0, -1), \omega_3 = (1, -1)\}$ and the mapping π is defined for every $N = (N_1, N_2) \in \mathbb{Z}_+^2$ as

$$\pi[N, y] = \begin{cases} 0 \text{ if } N_1 = 0 \text{ and } y = \omega_1 \\ 0 \text{ if } N_2 = 0 \text{ and } y = \omega_2 \text{ or } \omega_3. \\ y \text{ otherwise} \end{cases} \tag{19}$$

The distribution of N is completely determined by the sequence $Y = \{Y(k)\}$. Define $\mathrm{P}^+(\Omega) = \{\theta = (\theta_0, \theta_1, \theta_2, \theta_3)\}$, θ is a probability measure on Ω and $\theta_i = \theta[\omega_i]$ for every $i = 0, 1, 2, 3$. Because of the stability condition we suppose $\lambda < (\mu_1 \wedge \mu_2)(1 - \varphi_d)$, and without loss of generality, $\lambda + \mu_1 + \mu_2 = 1$. Therefore, under the original probability measure $\mathbb{P}$, the distribution of $Y(k)$ is just

$$\Theta = (\lambda, \mu_1, (1 - \varphi_d)\mu_2, \varphi_d \mu_2) \in \mathrm{P}^+(\Omega) \tag{20}$$

To be more precise, for a given threshold n, define the scaled state process $X^{(n)} = N/n$, where N is defined as above. Since the definition of π implies $\pi[nx, y] =$

$\pi[x, y]$ for every $x \in \mathbb{R}_+^2$, it is not difficult to see that $X^{(n)}$ satisfies the equation

$$X^{(n)}(k+1) = X^{(n)}(k) + \frac{1}{n}\pi[X^{(n)}(k), Y(k+1)] \tag{21}$$

with initial condition $X^{(n)}(0) = N(0)/n = 0$.

4. The importance sampling estimator and its asymptotic optimality

It was our interest to explore the probability p_n that $N_1 + N_2$ reaches n before return to 0, given that the system is initially empty. Asymptotically, as n rises, it is known that p_n plummets exponentially fast (see Proposition 3.1 in Dupuis and Wang 2009), at some rate

$$\xi = -\lim_{n \to \infty} \frac{1}{n} \log p_n. \tag{22}$$

Define the hitting times

$$T_n = \inf\{k \ge 0 : X_1^n(k) + X_2^n(k) = 1\}$$
$$T_0 = \inf\{k \ge 0 : X_1^n(k) + X_2^n(k) = 0\}. \tag{23}$$

Thus, $p_n = \mathbb{P}[T_n < \infty | X^{(n)}(0) = 0]$. Let A_n be the event of interest, that is

$$A_n = \{X_1^n + X_2^n \text{ reaches 1 before returning to 0}\} \tag{24}$$

and A_n obeys some probability measure $\mathbb{P}$.

If simulations under the original probability measure $\mathbb{P}$, starting at $X^{(n)}(0) = 0$, are run, $\mathbb{E}(1_{A_n})$ must be simulated. However, as the simulation needed to be speed up, the importance sampling technique was utilised in which the model under an alternative probability measure was simulated, and the simulation output was translated back into the original measure by multiplying with a likelihood ratio. In order to estimate $\mathbb{P}(A_n)$, samples under a (state-dependent) alternative probability measure $\mathbb{Q}$ were produced by importance sampling such that $\mathbb{P} \ll \mathbb{Q}$ and an estimator is formed by averaging independent replications of

$$\hat{p}_n = \mathbb{E}^{\mathbb{Q}}\left[1_{A_n} \frac{\mathbb{P}(A_n)}{\mathbb{Q}(A_n)} \right], \tag{25}$$

where $\mathbb{E}^{\mathbb{Q}}$ defines an expectation under the new measure $\mathbb{Q}$, 1_{A_n} is the indicator function of the event A_n in any simulation run and $\frac{\mathbb{P}(A_n)}{\mathbb{Q}(A_n)}$ is the likelihood ratio or Radon-Nikodým derivative (Asmussen and Glynn 2007) of the path under investigation, which is defined as follows:

$$\frac{\mathbb{P}(A_n)}{\mathbb{Q}(A_n)} = \prod_{k=0}^{T_n-1} \frac{\Theta[Y(k+1)]}{\bar{\Theta}^n[Y(k+1)|X^{(n)}(k)]}. \tag{26}$$

The importance sampling generates $\{Y(k)\}$ as follows. The conditional probability of $Y(k+1)=\omega_i$, $P(Y(k+1)=\omega_i|Y(j), \ j=1,2,\ldots,k)$, is just $\bar{\Theta}^n[\omega_i| X^{(n)}(k)]$ for each $i=0,1,2,3$. The importance sampling scheme we consider use state-dependent change of measure that can be characterised by stochastic kernels $\bar{\Theta}^n[\cdot|\cdot]$ on Ω given $\mathbb{R}^2_+$, that is $\bar{\Theta}^n[\cdot|x]\in \mathrm{P}^+(\Omega)$ for every $x\in\mathbb{R}^2_+$. It is obvious that $\bar{\Theta}^n[\omega_i|x]$ is the probability of a transition in direction ω_i if the current state of the process $X^{(n)}$ is x. The importance sampling estimator is just

$$\hat{p}_n = 1_{A_n}\cdot\prod_{k=0}^{T_n-1}\frac{\Theta[Y(k+1)]}{\bar{\Theta}^n[Y(k+1)|X^{(n)}(k)]}. \tag{27}$$

The second moment of $\hat{p}_n$ equals $\mathbb{E}_{\mathbb{P}}[\hat{p}_n]=\mathbb{E}_{\mathbb{Q}}[\hat{p}_n^2]$. The aim is to select a stochastic kernel $\bar{\Theta}^n$ so that this second moment (whence the variance of $\hat{p}_n$) is as small as possible (much less than ξ). Another important point for $\bar{\Theta}^n$ is simplicity and easy implementation. Before the construction of the importance sampling technique, definition of some notations and terminology are in order. Define

$$D = \{(x_1, x_2): x_1\geq 0, x_1+x_2 < 1\}$$
$$\delta_1 = \{0, x_2\}: 0 < x_2 < 1\}$$
$$\delta_2 = \{x_1, 0\}: 0 < x_1 < 1\}$$
$$\delta_e = \{x_1, x_2\}: x_1\geq 0, x_1+x_2 = 1\}$$
$$\bar{D} = D\cup\delta_1\cup\delta_2\cup\delta_e \tag{28}$$

sometimes δ_e is called the 'exit boundary'.

The function $W(x)$ which is defined for all $x\in\bar{D}$, plays a pivotal role in the change of measure. This function is an approximate solution to a set of equations derived using game theory. The construction of function $W(x)$ happens in three steps. First, three affine functions $\bar{W}_k^\delta(x)$ are formed based on some δ parameter as follows,

$$\bar{W}_1^\delta(x) = \langle r_1, x\rangle + 2\xi - \delta$$
$$\bar{W}_2^\delta(x) = \langle r_2, x\rangle + 2\xi - 2\delta$$
$$\bar{W}_3^\delta(x) = \langle r_3, x\rangle + 2\xi - (1+2\xi/\alpha)\delta. \tag{29}$$

Keep in mind that $\langle r_i, x\rangle$, $i=1,2,3$, is the inner product between r_i and x, where the value of r_i is rendered by $r_1=(-2\xi, -2\xi)$, $r_2=(-2\xi, -2(\alpha-\xi))$, $r_3=(0, 0)$. The value of ξ and α are derived from (Glasserman and Kou 1995)

$$\xi = \log\frac{(1-\varphi_d)(\mu_1\wedge\mu_2)}{\lambda} \tag{30}$$

$$\alpha = \begin{cases}\log[\mu_1/(\mu_1+\lambda-(1-\varphi_d)\mu_2)], & \text{if } \mu_1\geq\mu_2\\ \log[\mu_1/(\lambda+\varphi_d\mu_1)], & \text{if } \mu_1\leq\mu_2\end{cases} \tag{31}$$

The main feature of these affine functions is that they meet a condition derived using game theory, which is $\mathbb{H}(D\bar{W}_k^\delta)\geq 0$ (with equality for $k=1$) where $D\bar{W}_k^\delta = r_k$ is the gradient of $\bar{W}_k^\delta(x)$ and $\mathbb{H}$ defines a function which is known as the Hamiltonian. The precise definition and meaning of it which can be found in Dupuis et al. (2007) are inconsequential here but should be noted that its form may be found below.

Secondly, the minimum of these three affine functions is calculated, which produces a piecewise affine function $\bar{W}^\delta = \bar{W}_1^\delta\wedge\bar{W}_2^\delta\wedge\bar{W}_3^\delta$. Note that depending on which of the three functions $\bar{W}_k^\delta$ achieves the minimum, the set $\bar{D}$ is broken down into three regions. With each of these regions, the corresponding vector r_k determines a constant (not state-dependent) change of measure associated to the related region.

Finally, in order to make the resultant function W smooth along the boundaries of three subsets of $\bar{D}$ and therefore, to avoid the sudden transition from one type of measure to another, a mollification procedure is utilised, as the path of the process $X^{(n)}(k)$ traverses $\bar{D}$. Our mollification in Dupuis et al. (2007) is given, which is parameterised by ε:

$$W^{\varepsilon,\delta}(x) = -\varepsilon\log\sum_{k=1}^{3}e^{-\bar{W}_k^\delta(x)/\varepsilon} \tag{32}$$

where $W^{\varepsilon,\delta}(x)$ is the same function $W(x)$.

Keep in mind that $W^{\varepsilon,\delta}(x)$ is converged on the function $\bar{W}^\delta(x)$ as $\varepsilon\to 0$. Furthermore, the smoothness of $W^{\varepsilon,\delta}(x)$ along the aforementioned boundaries is designated by the value of ε. The state-dependent change of measure in each state x is largely associated to $DW^{\varepsilon,\delta}(x)$ which is the gradient of $W^{\varepsilon,\delta}(x)$ in x. As a matter of fact, this gradient can be presented as a state-dependent weighted average of the vectors r_k:

$$DW^{\varepsilon,\delta}(x) = \sum_{k=1}^{3}\rho_k(x)r_k \tag{33}$$

with

$$\rho_k(x) = \frac{e^{-\bar{W}_k^\delta(x)/\varepsilon}}{\sum_{j=1}^{3}e^{-\bar{W}_j^\delta(x)/\varepsilon}}. \tag{34}$$

As it was given in Proposition 3.2 in Dupuis et al. (2007), a change of measure is associated to each vector $p=(p_1,p_2)$ as follows

$$\bar{\Theta}^*((p_1,p_2)) = \Upsilon((p_1,p_2))\cdot(\lambda e^{\frac{-p_1}{2}}, \mu_1 e^{\frac{p_1-p_2}{2}},$$
$$(1-\varphi_d)\mu_2 e^{\frac{-p_2}{2}}, \varphi_d\mu_2 e^{\frac{p_2-p_1}{2}})$$
$$\bar{\Theta}_{\delta_1}^*((p_1,p_2)) = \Upsilon_1((p_1,p_2))\cdot(\lambda e^{\frac{-p_1}{2}}, \mu_1,$$
$$(1-\varphi_d)\mu_2 e^{\frac{-p_2}{2}}, \varphi_d\mu_2 e^{\frac{p_2-p_1}{2}})$$
$$\bar{\Theta}_{\delta_2}^*((p_1,p_2)) = \Upsilon_2((p_1,p_2))\cdot(\lambda e^{\frac{-p_1}{2}}, \mu_1 e^{\frac{p_1-p_2}{2}},$$
$$(1-\varphi_d)\mu_2, \varphi_d\mu_2), \tag{35}$$

where $\Upsilon((p_1, p_2))$, $\Upsilon_i((p_1, p_2))$ are normalizing constants so that all these vectors are probability vectors. Moreover, $\mathbb{H}((p_1, p_2)) = 2\log \Upsilon((p_1, p_2))$ and $\mathbb{H}_{\delta_i}((p_1, p_2)) = 2\log \Upsilon_i((p_1, p_2))$.

The current state x may determine the vector p which can be interpreted as $DW^{\varepsilon,\delta}(x)$. We can obtain a change of measure from a given function $W^{\varepsilon,\delta}(x)$. For each state x, Equation (34) can be utilised for each of the components $\bar{W}_k^\delta(x)$ to calculate the weighting function $\rho_k(x)$, and subsequently define $\bar{\Theta}[\omega_i]$ as the corresponding weighted average of the $\bar{\Theta}_k^* = \bar{\Theta}^*(D\bar{W}_k^\delta) = \bar{\Theta}^*(r_k)$, $k = 1, 2, 3$, which are computed using (35) $p = D\bar{W}_k^\delta = r_k$. Define a stochastic kernel $\bar{\Theta}^{\varepsilon,\delta}[\cdot|\cdot]$ by

$$\bar{\Theta}^{\varepsilon,\delta}[\cdot|x] = \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta} \bar{\Theta}^*(r_k), \quad \text{if} \quad x \in D \quad (36)$$

and

$$\bar{\Theta}^{\varepsilon,\delta}[\cdot|x] = \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta} \bar{\Theta}_{\delta_i}^*(r_k), \quad \text{if} \quad x \in \delta_i, i = 1, 2. \quad (37)$$

Theorem 4.1: *The importance sampling estimator $\hat{\rho}_n$ is asymptotically optimal if $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$.*

The proof of Theorem 4.1 is the same as Theorem 2 in de Boer and Scheinhardt (2007) and is omitted. We now allow ε and δ to vary depending on n and denote them by ε_n, δ_n in lieu of using a fixed pair of parameters ε, δ and let $\bar{\Theta}^n[\cdot|\cdot] = \bar{\Theta}^{\varepsilon_n,\varepsilon_n}[\cdot|\cdot]$. On the basis of Theorem 4.1, the asymptotic optimality of the importance sampling estimator is reached if $\delta_n \to 0$,

$\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$ and a good option is to set $\delta_n = -\varepsilon_n\log \varepsilon_n$.

5. Numerical results and discussion

In this study, the aforementioned two-node Jackson network with feedback for various parameters is analysed including customer arrival rate, service rate to customers in the first and second nodes, and also the feedback rate of customers from the second node to the first one. In this analysis, some parameters related to the network efficiency such as the probability of network blocking and the probability of missing the deadline of the second queue customers are estimated. As stated earlier, the network has a common buffer structure that can keep 20 customers in it at most.

5.1. Network blocking probability analysis

Figure 2 illustrates the network blocking probability for the set of parameters $\lambda = 0.1$, $\mu_1 = 0.55$, $\mu_2 = 0.35$ and $\lambda = 0.1$, $\mu_1 = 0.35$, $\mu_2 = 0.55$.

The selection of these two sets of parameters lies in the fact that in the first state, the second node is the network's bottleneck, whereas in the second state, the first node is network's bottleneck. In this study, we have changed the parameter feedback from 0.05–0.7 with 0.05 increments, and have estimated the blocking probability of each step for both states. Figure 2 also shows how the network blocking probability increases as the network feedback probability increases. It is shown that for φ_d in the range of 0.05–0.33 the change
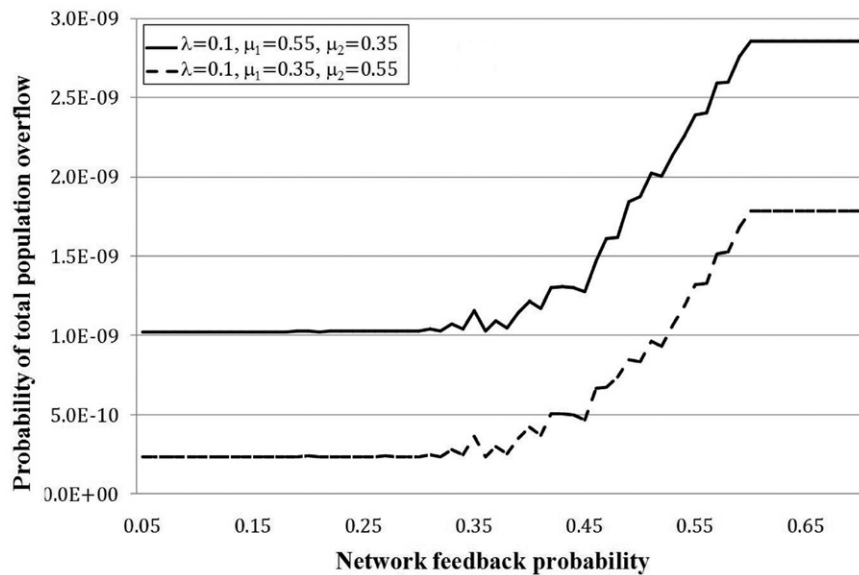
Figure 2. Estimation of the blocking probability.

in blocking probability is insignificant, while for φ_d in the range of 0.33–0.6 the network blocking probability rises, but it would be insignificant in the range of 0.6–0.7. As shown in the figure, such observations are repeated for the two parameter sets λ, μ_1 and μ_2. Another conclusion obtained from Figure 2 is that the network blocking probability is different for various sets of parameters, as the network with parameters $\lambda = 0.1$, $\mu_1 = 0.35$, $\mu_2 = 0.55$ has lower blocking probability over the network with parameters $\lambda = 0.1$, $\mu_1 = 0.55$, $\mu_2 = 0.35$. In other words, in the case that the second node is the network bottleneck, more customers wait in the system and thereby the network blocking rises. And if the first node is the network bottleneck, the network blocking probability will decrease. Such a difference lies in the fact that when the first node is the network bottleneck, providing service to customers in the first node is slower than that of the second node. Likewise, when the second node is the network bottleneck, the first node provides service faster than the second node.

Analysing the status of the second node in the network explains this condition. After receiving service from the second node with probability $1 - \varphi_d$, the customers leave the network or return to the first node with probability φ_d. If one customer leaves the network, the number of existing customers in the network will decrease, and the network remains far from the blocking state, leading to lower a blocking probability. If the second node is faster than the first node, the blocking probability will be more than the former state. It is due to the state of the second node in the network, as all customers that receive service from

the first node are placed at the second waiting line. As these customers rapidly leave the first node, and the second node works slower than the first node, there will be more customers at the second line, leading to more customers in the system, where it increases the network blocking probability. Because of the slower speed of providing service in the second node in this state compared with the former state, fewer customers receive service in this node, and the number of customers in the system will increase gradually.

In the next step, we repeated the experiment for states $\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.7$ and $\lambda = 0.1$, $\mu_1 = 0.7$, $\mu_2 = 0.2$, we obtained the network blocking probability. As shown in Figure 3, the results of this experiment are similar to the first one.

5.2. Analysis of probability of missing deadline for customers in the second queue

In this experiment, we have analysed the effect of different values of the deadline of customers in the second queue, λ and μ_2 on network feedback rate, and have estimated the network feedback probability for three values of different $\bar{\eta}$ for customer's deadline with varying λ and μ_2. As indicated in Section 3, $\bar{\eta}$ is the mean value of the random variable η known as the relative deadline.

The results of this study show that for a specific deadline for customers in the second queue, as the proportion of μ_2 to λ rises, that is as the existing load on the second node decreases, then the network feedback probability decreases. The reason is that as
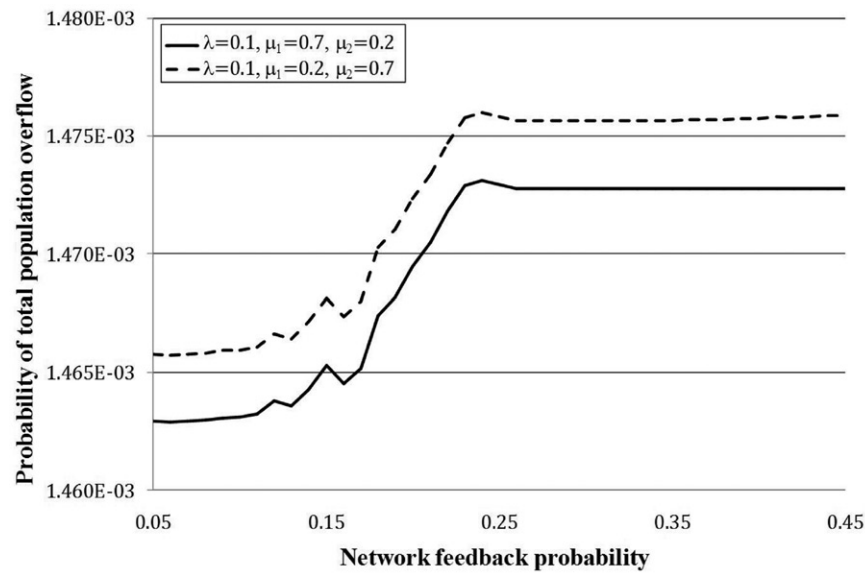

Figure 3. Blocking probability in two-node Jackson network with feedback.

the existing load on the second node decreases, the fewer number of customers in the second queue wait for service, and as a result the population of this node will be lower arriving at lower deadline rate of customers. In addition, the increase of load on the second node makes customer deadlines longer. In Figure 4, it is illustrated that as the customer's deadline decreases, the network feedback rate increases. Figure 5 illustrates the customer's deadline probability for various rates of customer entry to the network and various values of the customer's deadline. In this study,

for $\mu_2 = 0.4$ and three different values of λ, the customer's deadline probability is estimated, and varying values for customer's deadline are considered in each state. This experiment shows how the increase in the customer's deadline at the network lowers the probability of the deadline. It also illustrates that increase of customer's arrival rate to the network, the network feedback probability increases too. Figure 4 displays customer's deadline probability for $\lambda = 0.1$ and different service rate to customers in the second node, and for different customer's deadline. In this
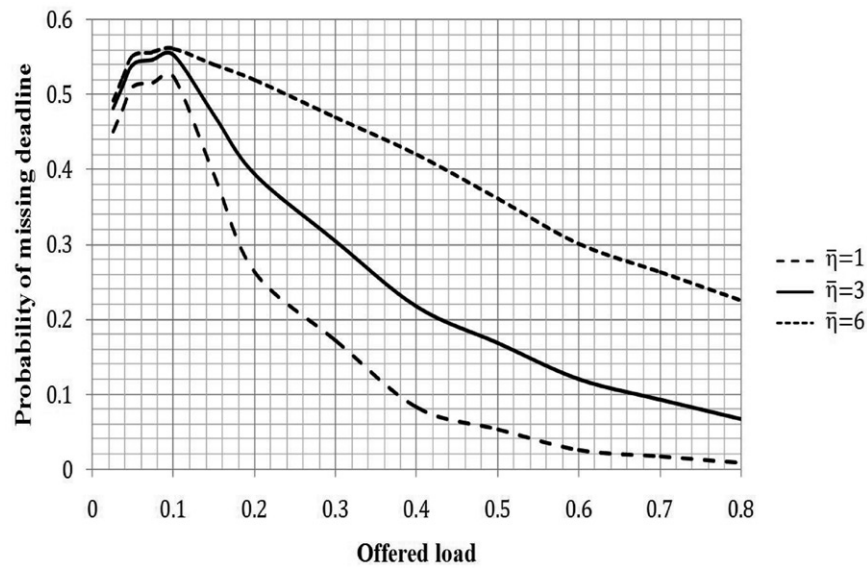

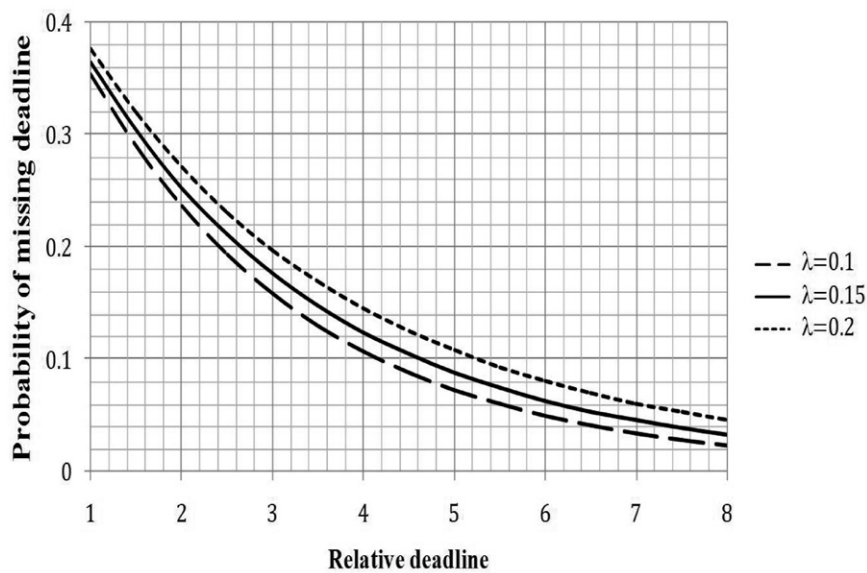Figure 4. Customer's deadline probability for various loads.

Figure 5. Customer's deadline probability for arrival rate and varying deadlines.

experiment, the customer's deadline probability for varying values of μ_2 is estimated, while varying values are considered for customer's deadline. This experiment shows how the increase of the network's customer's deadline lowers the network feedback rate, and it also displays the effect of increasing service rate on the network feedback probability.

6. Conclusions

In this article, we deal with the concept of a deadline on a two-node Jackson network with feedback in which arrival and service rates are modulated by an exogenous finite state Markov process. Based on the definition of the deadline for customers in the second queue, we have calculated the probability of missing the deadline, and have shown how the feedback rate of the network is affected by the deadline value. The results have used in an importance sampling estimator. It was found that an increase in the probability of missing deadline raises the probability of total population overflow. Likewise, it was known that how the probability of total population overflow affected by service rates in the network.

Acknowledgements

Notes on contributors

Ebrahim Mahdipour received his BS in Computer Engineering from IAU University, Dezful, in 2004, the MS in Computer Engineering from IAU University, Tehran, in 2006 and he is now pursuing the PhD degree in Computer Engineering at the IAU University, Tehran. His research interests are in the areas of network security, computer networks, modelling and performance evaluation.

Amir Masoud Rahmani received his BS in Computer Engineering from Amir Kabir University, Tehran, in 1996, the MS in Computer Engineering from Sharif University of Technology, Tehran, in 1998 and the PhD degree in Computer Engineering from IAU University, Tehran, in 2005. Currently, he is an Associate Professor in the Department of Computer Engineering at the IAU University. He is the author/co-author of more than 120 publications in technical journals and conferences. His research interests are in the areas of distributed systems, ad hoc and wireless sensor networks and evolutionary computing.

Dr Saeed Setayeshi, BSc, MSc, MA, MASc, PhD (Electrical & Computer Engineering, TUNS, Canada, 1993) is an Associate Professor and teaching in the Department of Medical Radiation Engineering, Amirkabir University of Technology (Tehran Polytechnics). He is also presenting graduate courses in Department of Computer Engineering, Research & Science Branch of IAU as an Invited Professor. He has founded AL and Complex System Researches for the first time there and supervised many graduate students in the areas of using CA and LA. He has published more than 150 papers in ISI Journals and Conferences and five books. Dr Setayeshi's research interests are in the areas of AI & AL, intelligent control (neural-fuzzy-expert-GA-CA-LA), adaptive signal processing, digital speech processing, agent based modeling, artificial society, social evolution, wealth distribution, knowledge based systems and dynamics of complex systems.

References

Asmussen, S., and Glynn, P.W. (2007), *Stochastic Simulation: Algorithms and Analysis*, New York, USA: Springer.

Barrer, D.Y. (1957a), 'Queuing with Impatient Customers and Indifferent Clerks', *Operations Research*, 5, 644–649.

Barrer, D.Y. (1957b), 'Queuing with Impatient Customers and Ordered Service', *Operations Research*, 5, 650–656.

Conley, W. (2007), 'Simulation Optimisation and Correlation with Multi Stage Monte Carlo Optimisation', *International Journal of Systems Science*, 38, 1013–1019.

de Boer, P.T. (2006), 'Analysis of State-independent Importance-sampling Measures for the Twonode Tandem Queue', *ACM Transactions on Modeling and Computer Simulation*, 16, 225–250.

de Boer, P.T., Nicola, V.F., and Rubinstein, R.Y. (2000), 'Adaptive Importance Sampling Simulation of Queueing Networks', in *2000 Winter Simulation Conference*, Orlando, FL, pp. 646–655.

de Boer, P.T., and Scheinhardt, W.R.W. (2007), 'Alternative Proof and Interpretations for a Recent State-dependent Importance Sampling Scheme', *Queueing Systems*, 57, 61–69.

de Boer, P.T., and Scheinhardt, W.R.W. (2010), 'A New, Analysis-based, Change of Measure for Tandem Queues', in *Eighth International Workshop on Rare-event Simulation*, June, Cambridge University, UK, pp. 12–13.

Dupuis, P., Sezer, A.D., and Wang, H. (2007), 'Dynamic Importance Sampling for Queueing Networks', *The Annals of Applied Probability*, 17, 1306–1346.

Dupuis, P., and Wang, H. (2009), 'Importance Sampling for Jackson Networks', *Queueing Systems: Theory and Applications*, 62, 113–157.

Glasserman, P., and Kou, S.G. (1995), 'Analysis of an Importance Sampling Estimator for Tandem Queues', *ACM Transactions on Modeling and Computer Simulation*, 5, 22–42.

Heegaard, P.E. (1998), 'A Scheme for Adaptive Biasing in Importance Sampling', *AE International Journal of Electronics and Communication*, 52, 113–212.

Kroese, D.P., and Nicola, V.F. (1999), 'Efficient Simulation of a Tandem Jackson Network', in *Second International Workshop on Rare Event Simulation*, March, University of Twente, The Netherlands, pp. 197–211.

Kroese, D.P., and Nicola, V.F. (2002), 'Efficient Simulation of a Tandem Jackson Network', *ACM Transactions on Modeling and Computer Simulation*, 12, 119–141.

Kroese, D.P., Taimre, T., and Botev, Z. (2011), *Handbook of Monte Carlo Methods*, New York, USA: John Wiley & Sons.

L'ecuyer, P., Blanchet, J.H., Tuffin, B., and Glynn, P.W. (2010), 'Asymptotic Robustness of Estimators in Rare-event Simulation', *ACM Transactions on Modeling and Computer Simulation*, 20, Article no. 6.

Mahdipour, E., and Rahmani, A. (2009), 'Importance Sampling for a Two-node Jackson Network with Customer Impatience Until the End of Service', in *2009 International Conference on Future Networks*, Bangkok, Thailand, pp. 137–141.

Mahdipour, E., Rahmani, A., and Setayeshi, S. (2009), 'Importance Sampling for a Markov Modulated Queuing Network with Customer Impatience Until the End of Service', *Informatica Economica*, 13, 106–118.

Mandjes, M., and Ridder, A. (2002), 'A Large Deviations Analysis of the Transient of a Queue with Many Markov Fluid Inputs: Approximations and Fast Simulation', *ACM Transactions on Modeling and Computer Simulation*, 12, 1–26.

Miretskiy, D., Scheinhardt, W., and Mandjes, M. (2010), 'State-dependent Importance Sampling for a Jackson Tandem Network', *ACM Transactions on Modeling and Computer Simulation*, 20, Article no. 15.

Movaghar, A. (2006), 'On Queueing with Customer Impatience Until the End of Service', *Stochastic Models*, 22, 149–173.

Parekh, S., and Walrand, J. (1989), 'A Quick Simulation Method for Excessive Backlogs in Networks of Queues', *IEEE Transactions on Automatic Control*, 34, 54–66.

Rubino, G., and Tuffin, B. (2009), *Rare Event Simulation using Monte Carlo Methods*, West Sussex, UK: John Wiley & Sons.