Stochastics and Statistics

Importance functions for restart simulation of general Jackson networks

José Villén-Altamirano *

Departamento de Matemática Aplicada, Universidad Politécnica de Madrid, Calle Arboleda s/n, 28031-Madrid, Spain

ARTICLE INFO

Keywords:
Simulation
Queueing
Rare event probabilities
RESTART

ABSTRACT

RESTART is an accelerated simulation technique that allows the evaluation of extremely low probabilities. In this method a number of simulation retrials are performed when the process enters regions of the state space where the chance of occurrence of the rare event is higher. These regions are defined by means of a function of the system state called the importance function. Guidelines for obtaining suitable importance functions and formulas for the importance function of two-stage networks were provided in previous papers. In this paper, we obtain effective importance functions for RESTART simulation of Jackson networks where the rare set is defined as the number of customers in a particular ('target') node exceeding a predefined threshold. Although some rough approximations and assumptions are used to derive the formulas of the importance functions, they are good enough to estimate accurately very low probabilities for different network topologies within short computational time.

1. Introduction

The goal of traffic engineering is to optimize a network's performance to achieve predefined requirements. The performance requirements of broadband communication networks and ultra reliable systems are often expressed in terms of events with very low probability. Developing good estimates of quality of service provided by a network, requires studying scenarios that may occur rarely during the life of the network. Analytical or numerical evaluation of low probabilities is only possible for a very restricted class of systems and although simulation is an effective means of studying such systems, acceleration methods are necessary because crude simulations require prohibitively long execution times for the accurate estimation of very low probabilities.

One such method is importance sampling. The basic idea behind this approach is to alter the probability measure governing events so that the formerly rare event occurs more often. One drawback of this technique is the difficulty of selecting an appropriate change of measure since it depends on the system being simulated. Most research has focused on finding good heuristics for particular types of models. One of the exceptions in the literature is the paper of Dupuis and Wang (2009), which deals with the construction of asymptotically optimal importance sampling schemes for queueing networks based on subsolutions. Importance sampling has difficulties to deal with large systems and/or systems with feedback.

Another method is RESTART (REpetitive Simulation Trials After Reaching Thresholds), which has a precedent of much more limited scope in the splitting method described in Kahn and Harris (1951). Villén-Altamirano and Villén-Altamirano (1991) coined the name RESTART and made a theoretical analysis that yields the variance of the estimator and the gain obtained with one threshold. A rigorous analysis of multiple thresholds was made by Villén-Altamirano and Villén-Altamirano (2002).

In the RESTART method a more frequent occurrence of a formerly rare event is achieved by performing a number of simulation retrials when the process enters regions of the state space where the importance is greater, i.e., regions where the chance of occurrence of the rare event is higher. These regions, called importance regions, are defined by comparing the value taken by a function of the system state, the importance function, with certain thresholds. Optimal values for thresholds and the number of retrials that maximize the gain obtained with RESTART were derived in Villén-Altamirano and Villén-Altamirano (2002).

The application of this method to particular models requires the choice of a suitable importance function. The problem of finding the optimal importance function has been compared by Garvels et al. (2002) with the problem of finding a good change of measure in importance sampling, because in both cases "knowledge about the behaviour of the system leading to the rare event is necessary". Glasserman et al. (1999) stated that "splitting ultimately relies on a detailed understanding of a process's rare event asymptotic, much as importance sampling does". They suggested that it may be difficult to use splitting or RESTART for systems with multi-dimensional state space. They presented bad results of a two-queue Jackson tandem network that were due to a bad choice of the importance function. Dean and Dupuis (2009) studied the construction of asymptotically optimal RESTART algorithms based on subsolutions. Unlike with importance sampling, which requires *classical* sense subsolutions, RESTART requires subsolutions only in the viscosity sense. Although this seems an interesting approach, they only presented examples of simple networks.

To obtain the importance function by means of formulas or by uncomplicated algorithms is of great interest for the application of RESTART to multi-dimensional state systems. L'Ecuyer et al. (2007) stated that "in the case of multi-dimensional state spaces, good choices of the importance function for splitting are crucial, and are definitely non-trivial to obtain in general". Tuffin and Trivedi (2000) pointed out that "the most challenging work for future research is to find and implement an efficient algorithm to determine good importance functions for defining thresholds". If such formulas or algorithms were implemented in simulators as those described in Lamers and Gorg (2002), Tuffin and Trivedi (2000), and Zimmermann et al. (2006), the method could be used by a non-specialist in rare event simulation.

In Villén-Altamirano and Villén-Altamirano (2002) the gain obtained with the application of RESTART was expressed as the optimal RESTART gain divided by several inefficiency factors, one of them reflecting the non-optimal choice of the importance function. This factor was analysed in Villén-Altamirano and Villén-Altamirano (2006) and guidelines for selecting heuristically such a function were provided. In Garvels et al. (2002), an algorithm based on reverse-time simulation was used to obtain an importance function for the two-queue Jackson tandem network, but such algorithm is difficult to implement in models with bigger state space. In Villén-Altamirano (2007) formulas for obtaining effective importance functions were provided for two-stage Markovian networks with any number of nodes in each stage.

The present paper copes with the case of Jackson networks where the rare set is defined as the number of jobs in a particular ('target') node exceeding a predefined threshold. A restriction of the Jackson network considered is that all the nodes have a distance not greater than 2 from the target node, that is, the customers leaving a node go directly to the target node (distance 1) or through only one intermediate node (distance 2). Formulas for obtaining effective importance functions will be first provided for three-stage Markovian networks with any number of nodes in each stage and then generalized for the Jackson networks with distance not greater than 2. It will be tested by simulation that the formulas are also valid for general Jackson networks without any type of restriction. Even for a network as simple as the three-queue tandem network, it is not possible to calculate the optimal importance function, see Garvels et al. (2002). Therefore, some approximations, one of them a rough approximation, are needed to derive these formulas. The goodness of the importance functions derived in the paper with such approximations is supported by the simulation results obtained: for most network topologies and loads, accurate estimations of very low probabilities (lower than 10^{-30}) can be obtained in short or moderate computational time. Let us observe that the approximations could affect the efficiency of the method, though they do not affect the correctness of the estimates. That is, longer computational time would be necessary to estimate the same probability with the same confidence if the importance function were not close to the optimal one (the importance function that minimizes the computational effort for the same confidence of the results). If the importance function were too far from the optimal one the computational time could be prohibitive.

The paper is organized as follows: Section 2 presents a review of the method and Section 3 describes the system under study. Section 4 derives the formulas of the importance function and Section 5 provides the simulation study.

2. Description of RESTART

RESTART has been described in several papers, e.g., Villén-Altamirano and Villén-Altamirano (2002, 2006), Villén-Altamirano (2007). Nevertheless it is briefly described here, and with much more detail in Supplementary material.

Let Ω denote the state space of a process $X(t)$ and A a rare subset of the state space whose probability must be estimated. A nested sequence of sets of states $C_i(C_1 \supset C_2 \supset \ldots C_M)$, is defined, which determines a partition of the state space Ω into regions $C_i - C_{i+1}$; the higher the value of i, the higher the importance of the region $C_i - C_{i+1}$. These sets are defined by means of a function $\Phi : \Omega \to \Re$, called the importance function. Thresholds $T_i(1 \leqslant i \leqslant M)$ of Φ are defined so that each set C_i is associated with $\Phi \geqslant T_i$.

RESTART works as follows: each time the process enters a set C_i, the system state is saved and R_i trials of level i are performed. Each trial of level i is a simulation path that starts with the saved state and finishes (except the last one) when it leaves set C_i. The last trial, which continues after leaving set C_i, potentially leads to new sets of trials of level i if set C_i is reached again. A set C_{i+1} may be reached in a trial of level i and an analogous process is set in motion: R_{i+1} trials of level $i+1$ are performed, and so on. Some more notations:

- $P = \Pr\{A\}$; $C_{M+1} = A$;
- $P_{h/i}(0 \leqslant i \leqslant h \leqslant M+1)$: probability of the set C_h, knowing that the system is in a state of the set C_i. For $h \leqslant M$, as $C_h \subset C_i, P_{h/i} = \Pr\{C_h\}/\Pr\{C_i\}$;
- $r_i = \prod_{j=1}^{i} R_j, (1 \leqslant i \leqslant M)$;
- $\Omega_i(1 \leqslant i \leqslant M)$: set of possible system states x_i, when the process enters set C_i;
- $P^*_{A/x_i}(1 \leqslant i \leqslant M)$: importance of state x_i, defined as the expected number of events A in a trial of level i starting with that system state;
- $P^*_{A/i}(1 \leqslant i \leqslant M)$: expected importance when the process enters set C_i:

  $$P^*_{A/i} = E[P^*_{A/X_i}] = \int_{\Omega_i} P^*_{A/x_i}\, dF(x_i),$$

  where $F(x_i)$ is the distribution function of X_i;
- $V(P^*_{A/X_i})(1 \leqslant i \leqslant M)$: variance of the importance when the process enters set C_i:

  $$V(P^*_{A/X_i}) = \int_{\Omega_i} (P^*_{A/x_i})^2\, dF(x_i) - (P^*_{A/i})^2.$$

In Villén-Altamirano and Villén-Altamirano (2002) it is proved that the gain (also called speedup) G obtained with RESTART is given by:

$$G = \frac{1}{f_V f_O f_R f_T} \frac{1}{P(-\ln P + 1)^2}. \tag{1}$$

The term $1/(P(-\ln P+1)^2)$ can be considered the ideal gain because it is the maximum gain that can be obtained (except some rare cases). Factors f_V, f_O, f_R and f_T, all of them equal to or greater than 1 (with the exception of f_V which may be smaller than 1 in some rare cases), can be considered inefficiency factors that reduce the actual gain with respect to the ideal gain. Each factor reflects:

- f_V: inefficiency due to the non-optimal choice of the importance function.
- f_O: inefficiency due to the computer overhead of RESTART.
- f_R: inefficiency due to the non-optimal choice of the number of retrials.
- f_T: inefficiency due to the non-optimal choice of the thresholds.

In Villén-Altamirano and Villén-Altamirano (2002) criteria for minimizing factors f_O, f_R and f_T were given. A value of the factor f_R very close to 1 can be achieved if the accumulated number of trials is chosen according to:

$$r_i = \frac{1}{\sqrt{P_{i+1/1}\cdot P_{i/0}}}, \quad i = 1,\ldots,M. \tag{2}$$

Factor f_T is minimized by choosing very close thresholds, i.e., $P_{i+1/i}$ close to one. An upper bound of f_T is given by, see Villén-Altamirano and Villén-Altamirano (2002):

$$f_T \leqslant \frac{1/P_{\min} + P_{\min} - 2}{(\ln P_{\min})^2} \text{ where } P_{\min} = Min_{0\leqslant i\leqslant M}(P_{i+1/i}). \tag{3}$$

Factor f_O affects to the computational time but not to the number of events to be simulated. This factor usually takes moderate values.

In Villén-Altamirano and Villén-Altamirano (2006) factor f_V was analyzed and guidelines for choosing the importance function were provided. The following upper bound of f_V was given: $f_V \leqslant Max_{1\leqslant i\leqslant M+1}(1 + V(P^*_{A/X_i})/(P^*_{A/i})^2)$. Thus, the main concern is to minimize $V(P^*_{A/X_i})$. It can be achieved by a proper choice of the importance function.

3. System under study

A general Jackson network with any number of nodes is studied. Customer arrivals and departures are allowed in all the nodes. After being served in node l, customers can go to any node m with probability p_{lm} or they can leave the network with probability p_{l0}. The steady-state probability of the number of customers exceeding a level at a target node, $Q_{tg} \geqslant L$, is estimated. It is assumed that all the nodes have a distance from the target node not greater than 2. Let us denote K the number of nodes with distance 1 and H the number of nodes with distance 2, for any value of K and H. It will be seen later that the restriction of not allowing nodes with distance from the target node greater than 2 could be removed without affecting significantly the efficiency of the method. The correctness of the results is not affected by the removal of this restriction. Customers with independent Poisson arrivals enter each node from the outside with arrival rates $\gamma_{1i}, i = 1,\ldots,H$ to the nodes with distance 2, $\gamma_{2j}, j = 1,\ldots,K$ to the nodes with distance 1 and γ_{tg} to the target node. The total arrival rates to each node (arrivals from the outside + arrivals from the other nodes) are denoted by: $\lambda_{1i},\ i = 1,\ldots,H$, $\lambda_{2j}, j = 1,\ldots,K$ and λ_{tg}, respectively. The service times of all the nodes are assumed to be exponentially distributed with service rates $\mu_{1i}, i = 1,\ldots,H,\ \mu_{2j}, j = 1,\ldots,K$ and μ_{tg}, respectively. The buffer space in each queue is assumed to be infinite. Let us observe that $\lambda_{1i} = \gamma_{1i} + \sum_{l=1}^{H}\lambda_{1l}p_{li} + \sum_{j=1}^{K}\lambda_{2j}p_{ji} + \lambda_{tg}p_{tgi}, i = 1,\ldots,H$. Analogous equations are obtained for $\lambda_{2j}, j = 1,\ldots,K$ and λ_{tg}. The loads of the nodes are $\rho_{1i} = \lambda_{1i}/\mu_{1i}, i = 1,\ldots,H,\ \rho_{2j} = \lambda_{2j}/\mu_{2j}, j = 1,\ldots,K$ and $\rho_{tg} = \lambda_{tg}/\mu_{tg}$, respectively.

The system state $X(t)$ is given by $(Q_{11},\ldots,Q_{1H}; Q_{21},\ldots,Q_{2K}; Q_{tg})$, where Q_{1i} is the random variable indicating the number of customers in the ith queue of the nodes with distance 2 from the target node at instant t, Q_{2j} is this number in the jth queue of the nodes with distance 1 from the target node and Q_{tg} is this number in the target queue.

As a particular case of the previous model, we will study a three-stage Markovian network with H nodes in the first stage, K nodes in the second stage and N nodes in the third stage, for any value of H, K and N. Customers only enter nodes of the first stage with arrival rates $\gamma_{1i}, i = 1,\ldots,H$, after being served enter a node of the second stage with probability $p_{ij}, i = 1,\ldots,H;\ j = 1,\ldots,K$, then go to a node of the third stage with probability $p_{jl}, j = 1,\ldots,K;\ l = 1,\ldots,N$, and, finally, they leave the network. The target node is a node of the third stage. In this model $\lambda_{1i} = \lambda_{1i}, i = 1,\ldots,H$.

The system state $X(t)$ is also given by $(Q_{11},\ldots,Q_{1H};\ Q_{21},\ldots,Q_{2K};\ Q_{tg})$. Let us observe that the number of customers in the queues of the third stage, except the target queue, has no influence on the rare set and is not included in $X(t)$.

4. Choice of the importance function

The importance function should lead to a small value of factor f_V. This can be achieved if the importance of all the states x_i when the process enters set C_i, P^*_{A/x_i}, is of the same order of magnitude, in which case the variance $V(P^*_{A/X_i})$ is small.

To achieve this objective, some approximations and assumptions are required since an exact evaluation of the importance of all the states is not possible even for most simple multi-dimensional systems. On the other hand, an exact evaluation is not necessary, because, as mentioned above, it is enough to know the order of magnitude of the importance of the states. For obtaining the formula of the importance function in each model, we will follow these steps:

1. First we will assume that the importance function is a linear function of the system state: $\Phi = a_{11}Q_{11} + \ldots + a_{1H}Q_{1H} + a_{21}Q_{21} + \cdots + a_{2K}Q_{2K} + Q_{tg}$. Linear importance functions led to successful results in the systems simulated in Villén-Altamirano (2007). The optimal importance function obtained numerically in Garvels et al. (2002) for the two-queue tandem network with finite capacity was approximately a linear function of the queue lengths.

2. We will evaluate the importance of the extreme (also called boundary) states x_i when the process enters set C_i, that is, the system states when only one queue is not empty. The importance of the states $(q_{11}, \ldots, 0;\ 0, \ldots, 0; 0), \ldots, (0, \ldots, 0;\ 0, \ldots, q_{2K}; 0), (0, \ldots, 0;\ 0, \ldots, 0;\ q_{tg})$ is a function of $q_{11}, \ldots, q_{2K}, q_{tg}$, respectively. If the importance of these extreme states is similar, then it is reasonable to assume that the intermediate states also have similar importance.
3. We will calculate the coefficients a_{11} as the ratio q_{tg}/q_{11} that equates the importance of the state $(0, \ldots, 0; 0, \ldots, 0;\ q_{tg})$ with the importance of $(q_{11}, \ldots, 0;\ 0, \ldots, 0; 0)$. In analogous way we will calculate all the other coefficients of the importance function.

For evaluating the importance of the extreme states in step 2, the importance of each state x_i is approximated by the probability of reaching the rare set in a trial of level i when the system state at the entry state of C_i is x_i. This probability is an underestimation of the importance of each state because the rare set can be reached more than once in that trial. The impact of this underestimation on the coefficients obtained has not to be important because, as the underestimation affects both members of the equation, the effects are at least partially compensated.

Let us first study the case $H = K = N = 1$, the well known three-queue tandem Jackson network, then generalize for a three-stage network with any value of H, K and N, and finally generalize for a general Jackson network.

4.1. Three-queue tandem Jackson network

Let Q_1, Q_2 and Q_3 be the number of customers in the first, second and third node, respectively, all of them with infinite buffer, and let γ be the arrival rate to the first queue. First we will consider the case $\rho_1 > \rho_2 > \rho_3$, given that this is the case where more difficulties may arise in the simulation because the first queue is the bottleneck, also the second queue is a bottleneck with respect to the third one. The rare set is defined as $Q_3 \geqslant L$, and we want to estimate the steady-state probability of this set.

If we let X_n denote the number of customers in the third queue, operating in equilibrium, after the nth transition (arrival or departure), then $\{X_n, n = 0, 1, 2, \ldots\}$ is a Markov Chain with state space 0,1,2,..., L,..., and transition probabilities $p_{i,i+1} = \frac{\gamma}{\gamma+\mu_3}$; $p_{i,i-1} = \frac{\mu_3}{\gamma+\mu_3} 0 < i < \infty$. If 0 and L are absorbing states, the probability that, starting in the state $q_3 < L$, the number of customers will eventually reach L is given (see Parzen (1999), Example 6A) by $\frac{\rho_3{}^{q_3}-1}{\rho_3{}^{L}-1}$, where $\rho_3 = \frac{\gamma}{\mu_3}$.

In our case the state space is $q_3 - 1, q_3, \ldots, L, \ldots$, and the absorbing states are $q_3 - 1$ and L, given that each retrial from the threshold $T = q_3$ finishes when that threshold is crossed down, that is, when there are $q_3 - 1$ customers in the third queue and the other queues are empty. The importance of $(0, 0, q_3)$ is approximated by the probability that, starting in the state q_3, the queue length will eventually reach L before it reaches to $q_3 - 1$. This probability is given by:

$$\frac{\rho_3^{-1} - 1}{\rho_3^{-L+q_3-1} - 1}. \tag{4}$$

In fact such probability is slightly lower because the initial system state is $(0, 0, q_3)$ and thus, during a short transient period the load is lower than ρ_3.

The exact calculation of the importance of the state $(q_1, 0, 0)$ is not possible, see e.g., Batchelor and Henry (2002). To study it we will make the following assumption: "In a three-queue tandem Jackson network with loads $\rho_1 > \rho_3$, if the initial system state is $(q_1, 0, 1)$ and the number of customers in the third queue reaches L before the third queue empties, q_1 customers will be in the third queue when the first one becomes empty". This assumption is partially justified by Lemma 1 of Villén-Altamirano (2007). That lemma was proved for the two-queue tandem network, but also applies to our case assuming that the transient period (while customers are in the second node) has no influence on the result. The lemma only provides lower and upper bounds for the number of customers in the third queue when the first one empties. So, it is difficult to provide even a heuristic argument to justify that such number is exactly q_1. Nevertheless, although the assumption is a very rough approximation, it leads to effective importance functions (as will be seen in Section 5) because, as mentioned above, it is not necessary that all the system states of the same threshold have the same importance. It is enough that their importance be of the same order of magnitude.

Based on this assumption and taking into account that, as long as the first queue is not empty, the load of the third queue, after a short transient period, is ρ_3/ρ_1, applying Eq. (4), the importance of the state $(q_1, 0, 1)$ is approximated by:

$$\frac{(\rho_3/\rho_1)^{-1} - 1}{(\rho_3/\rho_1)^{-q_1} - 1} \cdot \frac{\rho_3^{-q_1} - 1}{\rho_3^{-L} - 1}. \tag{5}$$

The first ratio of (5) is the probability that the third queue length first reaches q_1 prior to becoming empty, and the second ratio is the probability that, starting with q_1 customers, the third queue length first reaches L prior to becoming empty. In fact the first probability is slightly lower because during the transient period the load of the third queue is lower than ρ_3/ρ_1. The effect of this overestimation is partially compensated in Eq. (6) with the overestimation of the importance of $(0, 0, q_3)$, given by Eq. (4).

The importance of $(q_1, 0, 1)$ is a good approximation of the importance of $(q_1, 0, 0)$. Equating the importance of the extreme states $(0, 0, q_3)$ and $(q_1, 0, 0)$ given by Eqs. (4) and (5), respectively, after some algebra we obtain:

$$1 - \rho_3 = \frac{(\rho_1 - \rho_3)(\rho_3^{q_3-1} - \rho_3^{L})}{\rho_1^{q_1} - \rho_3^{q_1}} \frac{(1 - \rho_3^{q_1})}{(1 - \rho_3^{L})}. \tag{6}$$

As $1 > \rho_1 > \rho_3$, then $1 \gg \rho_1^{q_1} \gg \rho_3^{q_1}$ and $1 \gg \rho_3^{L}$. Thus, we can use the following approximation, which is good except in the cases where q_1 is close to 0 or where ρ_1 is very close to $\rho_3 : 1 - \rho_3 \simeq \frac{(\rho_1-\rho_3)\rho_3^{q_3-1}}{\rho_1^{q_1}}$, and then

$$q_3 \simeq q_1 \frac{\ln \rho_1}{\ln \rho_3} + \frac{\ln[\rho_3(1 - \rho_3)/(\rho_1 - \rho_3)]}{\ln \rho_3}.$$

The last term of the equation can be ignored except for the cases where ρ_3 is close to 1 (very improbable in rare event simulation) or where ρ_1 is very close to ρ_3. Thus, except in these cases, we can assume that if $\rho_1 > \rho_3$ then

$$q_3 \simeq q_1 \frac{\ln \rho_1}{\ln \rho_3}. \tag{7}$$

Analogously, the importance of the state $0, q_2, 0)$ is approximated by:

$$\frac{(\rho_3/\rho_2)^{-1} - 1}{(\rho_3/\rho_2)^{-q_2} - 1} \cdot \frac{\rho_3^{-q_2} - 1}{\rho_3^{-L} - 1}. \tag{8}$$

Equating the importance of the extreme states $(0, 0, q_3)$ and $(0, q_2, 0)$ given by Eqs. (4) and (8), respectively, and following the same approximations that led to Eq. (7), we obtain:

$$q_3 \simeq q_2 \frac{\ln \rho_2}{\ln \rho_3}. \tag{9}$$

Defining the importance function as

$$\Phi = \frac{\ln \rho_1}{\ln \rho_3} Q_1 + \frac{\ln \rho_2}{\ln \rho_3} Q_2 + Q_3 \tag{10}$$

fulfils the requirement that the states $(q_1, 0, 0)$ and $(0, 0, \frac{\ln \rho_1}{\ln \rho_3} q_1)$ have the same importance and also the states $(0, q_2, 0)$ and $\left(0, 0, \frac{\ln \rho_2}{\ln \rho_3} q_2\right)$, and thus Eqs. (7) and (9) hold. Let us observe that the coefficients of the importance function are the ratios of q_3/q_1 and q_3/q_2 in Eqs. (7) and (9), respectively.

With analogous reasoning, we obtain the importance function for the other five cases:

If $\rho_1 < \rho_2 < \rho_3$ or if $\rho_2 < \rho_1 < \rho_3$ then

$$\Phi = Q_1 + Q_2 + Q_3. \tag{11}$$

If $\rho_2 < \rho_3 < \rho_1$ then

$$\Phi = \frac{\ln \rho_1}{\ln \rho_3} Q_1 + Q_2 + Q_3. \tag{12}$$

If $\rho_1 < \rho_3 < \rho_2$ or if $\rho_3 < \rho_1 < \rho_2$ then

$$\Phi = \frac{\ln \rho_2}{\ln \rho_3} (Q_1 + Q_2) + Q_3. \tag{13}$$

As was seen before, if $\rho_1 > \rho_2 > \rho_3$ then $\Phi = \frac{\ln \rho_1}{\ln \rho_3} Q_1 + \frac{\ln \rho_2}{\ln \rho_3} Q_2 + Q_3$.

4.2. Three-stage Jackson networks

Analogously to Eq. (4), the importance of the state $(0, \ldots, 0; q_{tg})$ is approximated by:

$$\frac{\rho_{tg}^{-1} - 1}{\rho_{tg}^{-L+q_{tg}-1} - 1} \tag{14}$$

with ρ_{tg} given by:

$$\rho_{tg} = \frac{\sum_{i=1}^{H} \gamma_{1i} \sum_{j=1}^{K} p_{ij} p_{jtg}}{\mu_{tg}} = \frac{\lambda_{tg}}{\mu_{tg}}. \tag{15}$$

Let us see now how to obtain, approximately, the importance of the state $(0, \ldots, q_{1i}, \ldots, 0; 0), 1 \leqslant i \leqslant H$. While $q_{1i} > 0$, the load of the target queue, after a transient period, is:

$$\rho_{tgi}^{*} = \frac{\sum_{j=1}^{K} Min\left\{\mu_{1i} p_{ij} + \sum_{l \neq i} \gamma_{1l} p_{lj}, \mu_{2j}\right\} p_{jtg}}{\mu_{tg}} = \rho_{tg} \frac{\sum_{j=1}^{K} Min\left\{\mu_{1i} p_{ij} + \sum_{l \neq i} \gamma_{1l} p_{lj}, \mu_{2j}\right\} p_{jtg}}{\lambda_{tg}}. \tag{16}$$

We will first see the approximation made for the case $\rho_{tgi}^{*} < 1, \ \forall\ i$. The approximation is based on the following assumption: we will assume that the load ρ_{tg} is given by Eq. (16), i.e., the ith queue of the first stage does not become empty, until q_{1i} customers of this queue are in the target queue. This rough assumption is partially sustained by Lemma 2 of Villén-Altamirano (2007).

Let us observe that while q_{1i} customers of the ith queue enter the target queue, on average, $\left(\sum_{l \neq i} \gamma_{1l} \sum_{j=1}^{K} p_{lj} p_{jtg} / \mu_{1i} \sum_{j=1}^{K} p_{ij} p_{jtg}\right) q_{1i}$ customers from the other queues of the first stage also enter the target queue. In fact this number is slightly lower because during a short transient period the load is lower than ρ_{tgi}^{*}, given that all the queues except the ith one start empty. We define: $\alpha_{1i} = 1 + \left(\sum_{l \neq i} \gamma_{1l} \sum_{j=1}^{K} p_{lj} p_{jtg} / \mu_{1i} \sum_{j=1}^{K} p_{ij} p_{jtg}\right)$.

With the assumption made, the load ρ_{tg} is given by Eq. (16) until there are $\alpha_{1i} q_{1i}$ customers in the target queue. From $\alpha_{1i} q_{1i}$ to L, the load ρ_{tg} is given by Eq. (15). Thus the importance of the state $(0, \ldots, q_{1i}, \ldots, 0, 0), \ 1 \leqslant i \leqslant H$ is approximated by:

$$\frac{(\rho^{*}_{tgi})^{-1}-1}{(\rho^{*}_{tgi})^{-\alpha_{1i}q_{1i}}-1}\,\frac{(\rho_t)^{-\alpha_{1i}q_{1i}-1}}{(\rho_t)^{-L}-1}. \tag{17}$$

During a transient period the load of q_{tg} is lower than ρ_{tg} in (14) and also lower than ρ^{*}_{tgi} in the first ratio of (17). The effect of both overestimations of the importance of the states $(0,\ldots,0;q_{tg})$ and $(0,\ldots,q_{1i},\ldots,0;0), 1\leqslant i\leqslant H$ are at least partially compensated at Eq. (19).

Analogously, the importance of the state $(0,\ldots,q_{2j},\ldots,0;0),\ 1\leqslant j\leqslant K$ is approximated by:

$$\frac{(\rho^{\perp}_{tgj})^{-1}-1}{(\rho^{\perp}_{tgj})^{-\alpha_{2j}q_{2j}}-1}\,\frac{(\rho_{tg})^{-\alpha_{2j}q_{2j}}-1}{(\rho_{tg})^{-L}-1} \tag{18}$$

with:

$$\rho^{\perp}_{tgj}=\rho_{tg}\frac{\mu_{2j}p_{jtg}+\sum_{i=1}^{H}\gamma_{1i}\sum_{l\neq j}^{p_{il}p_{ltg}}}{\lambda_{tg}}\ \text{and}\ \alpha_{2j}=1+\left(\sum_{i=1}^{H}\gamma_{1i}\sum_{l\neq j}p_{il}p_{ltg}/\mu_{2j}p_{jtg}\right).$$

For determining the coefficients of the importance function, we equate the importance of the state $(0,\ldots,0;q_{tg})$, given by Eq. (14), with the importance of $(0,\ldots,q_{1i},\ldots,0;0), 1\leqslant i\leqslant H$, given by Eq. (17). After some algebra, we obtain:

$$1-\rho_{tg}=\frac{\left(\frac{\rho_{tg}}{\rho^{*}_{tgi}}-\rho_{tg}\right)\left(1-\rho_{tg}^{\alpha_{1i}q_{1i}}\right)}{\left(\frac{\rho_{tg}}{\rho^{*}_{tgi}}\right)^{\alpha_{1i}q_{1i}}-\rho_{tg}^{\alpha_{1i}q_{1i}}}\,\frac{\rho_{tg}^{q_{tg}-1}-\rho_{tg}^{L}}{1-\rho_{tg}^{L}}. \tag{19}$$

As $\rho^{*}_{tgi}<1$, we can make similar approximations to those made after Eq. (6):

$$1-\rho_{tg}\simeq\frac{\left(\frac{\rho_{tg}}{\rho^{*}_{tgi}}-\rho_{tg}\right)}{\left(\frac{\rho_{tg}}{\rho^{*}_{tgi}}\right)^{\alpha_{1i}q_{1i}}}\rho_{tg}^{q_{tg}-1}$$

and then:

$$q_{tg}\simeq\alpha_{1i}q_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}+\frac{\ln[\rho^{*}_{tgi}(1-\rho_{tg})/(1-\rho^{*}_{tgi})]}{\ln\rho_{tg}}\simeq\alpha_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}q_{1i} \tag{20}$$

given that the last term of the first equation can be ignored except for the cases where ρ^{*}_{tgi} is close to 1.

Analogously, equating the importance of the states $(0,\ldots,0;\,q_{tg})$ and $(0,\ldots,q_{2j},\ldots,0;\,0), 1\leqslant j\leqslant K$, given by Eqs. (14) and (18), respectively, assuming $\rho^{\perp}_{tgj}<1$, and following similar approximations, we obtain this relation:

$$q_{tg}\simeq\alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}q_{2j}. \tag{21}$$

From Eqs. (20) and (21) we obtain the coefficients $a_{1i}, 1\leqslant i\leqslant H$ and of the importance function. Hence this function is given by:

$$\Phi=\sum_{i=1}^{H}\alpha_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}Q_{1i}+\sum_{j=1}^{K}\alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}Q_{2j}+Q_{tg}. \tag{22}$$

As mentioned before, this function is valid for the cases where $\rho^{*}_{tgi}<1\quad\forall i$ and $\rho^{\perp}_{tgj}<1\quad\forall j$. These loads usually lead to coefficients of Q_{1i} and Q_{2j} smaller than 1. If any coefficient is greater than 1, as it does not seem reasonable that the coefficients of the queues of the first and the second stages in the importance function are greater than the coefficient of the target queue, that coefficient should be 1. Thus, the importance function for the cases where $\rho^{*}_{tgi}<1\ \forall\, i$ and $\rho^{\perp}_{tgj}<1\ \forall j$ is:

$$\Phi=\sum_{i=1}^{H}Min\left\{1,\alpha_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}\right\}Q_{1i}+\sum_{j=1}^{K}Min\left\{1,\alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}\right\}Q_{2j}+Q_{tg}. \tag{23}$$

If $\rho^{*}_{tgi}>1\ \forall\, i$ and $\rho^{\perp}_{tgj}>1\ \forall\, j$, Eq. (19) is also valid, and following analogous approximations to the previous case, we obtain the importance function:

$$\Phi=\sum_{i=1}^{H}\alpha_{1i}Q_{1i}+\sum_{j=1}^{K}\alpha_{2j}Q_{2j}+Q_t.$$

However, as the coefficients of the queues of the first and the second stages should not be greater than the coefficient of the target queue, if $\rho^{*}_{tgi}>1\ \forall\, i$ and $\rho^{\perp}_{tgj}>1\ \forall\, j$, the importance function is:

$$\Phi=\sum_{i=1}^{H}Q_{1i}+\sum_{j=1}^{K}Q_{2j}+Q_{tg}.$$

Let us observe that this equation is a particular case of Eq. (23) because if $\rho^{*}_{tgi}>1$ then $\alpha_{1i}\ln(\rho_{tg}/\rho^{*}_{tgi})/\ln(\rho_{tg})>1$, and the same occurs with the coefficients of Q_{2j}. Thus, Eq. (23) is valid for values of ρ^{*}_{tgi} and $\rho^{\perp}_{tgj}$ greater or smaller than one.

We must also take into account that, analogous to Eq. (13), if for some i,j : $Min\left\{1, \alpha_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}\right\} > Min\left\{1, \alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}\right\}$, then the coefficient of the queue i of the first stage should be given by the coefficient of the queue j of the second stage if all the customers of queue i go to queue j. Otherwise, the coefficient of the queue i should be modified proportionally to $p'_{ij} = p_{ij}/\sum_{j=1}^{K} p_{ij}$. Thus, the importance function is:

$$\Phi = \sum_{j=1}^{K} Min\left\{1, \alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}\right\}\left(1 + \sum_{i=1}^{H} x_{ij}p'_{ij}\right)Q_{2j} + Q_{tg} \tag{24}$$

with:

$$x_{ij} = \begin{cases} Q_{1i}/Q_{2j} & \text{if } Min\left\{1, \alpha_{1i}\frac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}\right\} \geqslant Min\left\{1, \alpha_{2j}\frac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}\right\} \\ \dfrac{\alpha_{1i}\dfrac{\ln(\rho_{tg}/\rho^{*}_{tgi})}{\ln\rho_{tg}}Q_{1i}}{Min\left\{1, \alpha_{2j}\dfrac{\ln(\rho_{tg}/\rho^{\perp}_{tgj})}{\ln\rho_{tg}}\right\}Q_{2j}} & \text{otherwise} \end{cases}.$$

Let us observe that Eqs. (10)–(13), (22) and (23) are particular cases of Eq. (24). Eq. (24) matches Eq. (22) in all the examples of this paper.

4.3. General Jackson networks

As mentioned in Section 3, we shall assume that K nodes are at distance 1 from the target node and that H nodes are at distance 2 from the target node, for any value of K and H, and that there are not nodes at distance greater than 2.

Following the same reasoning as for three-stage networks, the same importance functions given by Eq. (24) is obtained, but with the loads and values of $\alpha_{1i}, i = 1,\ldots,H$ and $\alpha_{2j}, j = 1,\ldots,K$ given by the following formulas:

$$\rho_{tg} = \frac{\gamma_{tg} + \sum_{j=1}^{K}\lambda_{2j}p_{jtg} + \lambda_{tg}p_{tgtg}}{\mu_{tg}} = \frac{\lambda_{tg}}{\mu_{tg}}$$

$$\rho^{*}_{tgi} = \rho_{tg}\frac{\gamma_{tg} + \sum_{j=1}^{K} Min\{\lambda_{2j} + (\mu_{1i} - \lambda_{1i})p_{ij}, \mu_{2j}\}p_{jtg} + \lambda_{tg}p_{tgtg}}{\lambda_{tg}}$$

$$\rho^{\perp}_{tgj} = \frac{\gamma_{tg} + \mu_{2j}p_{jtg} + \sum_{l\neq j}\lambda_{2l}p_{ltg} + \lambda_{tg}p_{tgtg}}{\mu_{tg}} = \rho_{tg}\left(1 + \frac{(\mu_{2j} - \lambda_{2j})p_{jtg}}{\lambda_{tg}}\right)$$

$$\alpha_{1i} = 1 + \frac{\sum_{l\neq i}\gamma_{1l}\sum_{j=1}^{K}p_{lj}p_{jtg} + \sum_{j=1}^{K}\gamma_{2j}p_{jtg} + \gamma_{tg}}{\mu_{1i}\sum_{j=1}^{K}p_{ij}p_{jtg}}; \quad \alpha_{2j} = 1 + \frac{\sum_{i=1}^{H}\gamma_{1i}\sum_{l\neq j}p_{il}p_{ltg} + \sum_{l\neq j}\gamma_{2l}p_{ltg} + \gamma_{tg}}{\mu_{2j}p_{jtg}}.$$

Let us observe that in the previous formulas of α_{1i} and α_{2j} we have just taken into account the customers that go to the target node through the shortest way. For example, a customer that enters a node at distance 1 can go directly to the target node or can go there through one or more intermediate nodes and formulas only consider the first case. This approximation affects in a similar manner to the numerator (except the arrival from the outside to the target node γ_{tg}) and denominator of α_{1i} and α_{2j} and thus it seems a good approximation that makes simpler the formulas of those coefficients. Analogous approximations have been made in the formulas of ρ^{*}_{tgi} and $\rho^{\perp}_{tgj}$. It will be checked by simulation that the formula of the importance function leads to good results in a case where the shortest way has small probability to occur.

Following analogous reasoning, formulas for Jackson networks without any restriction could be obtained. Nevertheless, it is not worth using more complicated formulas for networks with nodes at distance greater than 2 from the target node because the dependence between these nodes and the target node is usually very weak and can be ignored. Hence, the formulas of the importance function should be applied only to nodes at distances 1 and 2 from the target. It will be checked by simulation that the above formulas of the importance function also lead to good results when they are applied to networks with some nodes at distance greater than 2 and also when they are applied to networks for which some of the assumptions made for deriving the importance function do not hold.

As a summary of the formulas given in this paper, the importance function of any Jackson network can be calculated by Eq. (24) with the loads and values of $\alpha_{1i}, i = 1,\ldots,H$ and $\alpha_{2j}, j = 1,\ldots,K$ given in this section.

5. Test cases

We conduct several simulation experiments on Jackson networks with different topologies and loads. The rare set A is defined in most cases as $Q_{tg} \geqslant 70$, where Q_{tg} is the number of customer at the target node. The steady-state probability of A is of the order of 10^{-34} in those examples. The reason for simulating such small probabilities is to show the goodness of the importance functions obtained in the paper, given that, if it is possible to estimate accurately such small probabilities with short or moderate computational time, it will take much less time to estimate more realistic probabilities.

Thresholds T_i were set for every integer value of Φ between 2 (in some cases 3) and l, where l varies between 71 and 75 depending on the case being simulated. Let us observe that the rare set can be reached in retrials from the last l-69 regions $C_i - C_{i+1}$ (C_M if $i = M$). Pilot runs (one or two for each case) were made to set the number of retrials. We proceed as following: we set (for example) the thresholds 2, 3, 4, ..., 74 and we make a pilot simulation. This simulation gives us the optimal number of retrials according to Eq. (2), following the guidelines given in Villén-Altamirano and Villén-Altamirano (2002) for rounding to integer values. If the number of retrials from one threshold (in the pilot simulation) is 1, we eliminate such threshold. If the number of retrials from the last threshold is greater than 5, it is possible to set an additional threshold. The number of retrials obtained R_i is 2 or 3 in all cases. The confidence interval width is evaluated using the batch

Table 1
Results for the three-queue tandem network. Relative error = 0.1. Rare set probability: $P(Q_{tg} \geqslant 70) = 3.99 \times 10^{-34}$, $\rho_{tg} = 1/3$, $\Phi = aQ_1 + bQ_2 + Q_3$.

$\hat{P}$	ρ_1	ρ_2	a	b	Events (millions)	Time (min)	Gain (events)	f_V	Gain (time)	f_O
4.17×10^{-34}	2/3	1/2	0.37	0.63	210	69.0	1.0×10^{28}	37.4	2.5×10^{27}	3.9
3.97×10^{-34}	1/3	1/2	0.63	0.63	30	10.7	7.0×10^{28}	5.4	1.8×10^{28}	3.9
4.15×10^{-34}	1/2	1/4	0.63	1	68	26.6	3.1×10^{28}	12.2	8.0×10^{27}	3.8
4.03×10^{-34}	1/5	1/4	1	1	3.6	1.0	6.0×10^{29}	0.6	1.6×10^{29}	3.7
3.88×10^{-34}	1/3	1/3	1	1	67	18.4	3.2×10^{28}	11.8	5.5×10^{27}	3.7

Table 3
Results for Jackson networks with 7 nodes. Relative error = 0.1. Rare set probability: $P(Q_{tg} \geqslant 70) = 8.8 \times 10^{-35}$, $\rho_{tg} = 0.3262$, $\Phi = a\sum_{i=1}^{4} Q_i + b\sum_{j=5}^{6} Q_j + Q_{tg}$.

$\hat{P}$	ρ_{1i}	ρ_{2j}	a	b	Events (millions)	Time (min)	Gain (events)	f_V	Gain (time)	f_O
9.1×10^{-35}	0.50	0.41	0.23	0.45	41	13.7	3.7×10^{29}	4.3	8.3×10^{28}	4.4
8.5×10^{-35}	0.32	0.41	0.36	0.45	13	3.8	1.2×10^{30}	1.4	2.8×10^{29}	4.2
8.9×10^{-35}	0.28	0.30	0.40	0.61	7	2.2	2.1×10^{30}	0.8	5.1×10^{29}	4.2

means method. (It could also be used the independent replication method). Each batch finishes after a fix number of arrivals (usually between 500,000 and 2,000,000). After each batch the half width of the 95% confidence interval divided by the estimate (relative error) is calculated and the simulation finishes when the relative error is smaller than 0.1. For each case we made 3 simulations, and we wrote in the tables the results corresponding to the median of the computational times. All the experiments were run on a Pentium(R) D CPU 3.01 GHz.

5.1. Three-queue tandem network

Customers arrive to the first queue of this network with arrival rate equal to 1, then go to the second queue, then to the third and then they leave the network. The importance function Φ was chosen according to Eqs. (10)–(13), that is, $\Phi = aQ_1 + bQ_2 + Q_3$, with values of a and b depending on the loads ρ_1, ρ_2 and ρ_3. The results, which are summarized in Table 1, show that very accurate estimates were obtained in short or moderate computational time.

To evaluate approximately the gain in events or the gain in time with respect to a crude simulation, the data of the crude simulations were estimated by extrapolating the measured values for $P(Q_{tg} \geqslant 14) = 2.09 \times 10^{-07}$. To estimate this probability with a relative error equal to 0.1, it was necessary to simulate the arrival of 4009 million customers and it took 5.6 h of computational time in the same PC. The extrapolation was made taking into account that with crude simulation the relative error for estimating a probability P with n batches (samples) is proportional to $\sqrt{P(1-P)/n}/P \simeq 1/\sqrt{nP}$. Thus, the number of samples for achieving a given relative error is inversely proportional to the probability that is to be estimated.

It is interesting to compare the measured gain with the theoretical one derived from Eq. (1). If we assume $f_V = 1$ and $f_O = 1$ in Eq. (1), the theoretical gain is equal to 3.74×10^{29} (taking the values of r_i given by Eq. (2) and thus $f_R = 1$, and f_T is equal to its bound in Eq. (3) evaluated for $P_{min} = P^{1/M}$). We see that the theoretical gain (for $f_V = 1, f_O = 1$) is, in the first row, 37.4 times the actual gain in events. Given that the gain in events is not affected by the factor f_O, the value 37.4 can be taken as an estimate of f_V. Finally, the factor f_O can be estimated as the ratio between the gain in events and the gain in time. Values of f_O up to 3.9 were obtained, as shown in Table 1.

The low values of f_V show that the choice of the importance function is appropriate and that the application is close to the optimal, at least for most of the tested cases. We observe that the worst result is obtained when $\rho_1 > \rho_2 > \rho_3$, as we mentioned in Section 4.1, but even in this case the computational time is moderate.

To study the goodness of the approximations that led to Eqs. (10)–(13) and also to analyse the sensitivity of the results to changes in the importance function, the same network was studied with the importance function $\Phi = aQ_1 + bQ_2 + Q_3$ for different values of a and b. The best results were obtained with the importance functions given by Eqs. (10)–(13) except in the last case: $\rho_1 = \rho_2 = \rho_3$, where slightly better results were obtained for $a = 0.9$, $b = 0.9$. These results support the goodness of the importance functions given by Eqs. (10)–(13) and seem to indicate that the approximations made for deriving the equations are good enough for the desired purpose.

As regards the sensitivity, the values of f_V for values of $a$and b up to 10% lower than those given by the formulas, became less than double. However, worse results were obtained for values of $a$and b greater than those given by the formulas.

Garvels et al. (2002) estimated the importance function using the time-reversal method with the restriction of assuming finite capacity of the first two queues, but they could not estimate accurately overflow probabilities lower than 10^{-15}. The method cannot be generalized to bigger networks because, as they mentioned, the bigger state space results in a greater variance of the estimates of the importance function.

Remark 1. The results (Table 2) and comments corresponding to the three-stage Jackson network can be seen in Supplementary material.

5.2. General Jackson networks

Although it is not possible to simulate all the Jackson networks to prove that the importance function given by Eq. (24) is always effective, we have selected test cases that a priori could have some difficulties. If the importance function is effective for these cases, it is supposed that it will be also effective for most Jackson networks.

5.2.1. Jackson networks with all nodes at distance lower than 3

Let us now consider a Jackson network with 7 nodes. Customers from the outside arrive at any node of the network at a rate $\gamma_i = 1,\ i = 1,\ldots,7$. After being served in each node, a customer leaves the network with probability 0.2. Otherwise the customer goes to another node in accordance to the following transition matrix:

	1	2	3	4	5	6	*tg*
1	0.1	0.1	0.1	0.1	0.2	0.2	0
2	0.1	0.1	0.1	0.1	0.2	0.2	0
3	0.1	0.1	0.1	0.1	0.2	0.2	0
4	0.1	0.1	0.1	0.1	0.2	0.2	0
5	0.1	0.1	0.1	0.1	0	0.1	0.3
6	0.1	0.1	0.1	0.1	0.1	0	0.3
tg	0.1	0.1	0.1	0.1	0.1	0.1	0.2

The overall arrival rate to each node is $\lambda_i = 4.5,\ i = 1,\ldots,4;\ \lambda_5 = \lambda_6 = 5.73; \lambda_{tg} = 5.55$. The rare set is also $Q_{tg} \geqslant 70$, where Q_{tg} is the target queue. The load of the target node is 0.3262. We can observe in the matrix that nodes 5 and 6 are at distance 1 from the target node, and that nodes 1 to 4 are at distance 2 from the target node. Nodes 5 and 6 have the same load, and the same occurs with nodes 1 to 4.

The importance function Φ was chosen according to Eq. (24) with the coefficients given in Section 4.3. The gain and the factors f_V and f_O are estimated as in Section 5.1. The results, which are summarized in Table 3, show that very accurate results were obtained in short computational time.

We can also observe that the worst results are obtained when $\rho_{1i} > \rho_{2j} > \rho_{tg}$, while the best results are obtained when the bottleneck is the target queue. The results are better than in the previous networks because many customers of the other queues never go to the target queue. Thus, the dependence of the target queue on the queue length of the other queues is weaker and the efficiency of RESTART is greater. Unlike with importance sampling, the efficiency of RESTART may improve with the complexity of the systems and it does not seem affected by the feedback. Let us observe that the splitting technique described in Kahn and Harris (1951) would not be effective for estimating steady-state probabilities in complex networks because the trials that include regions of less importance are not killed and so much time would be wasted, see Villén-Altamirano and Villén-Altamirano (2006). The two variants of splitting that use resplits, described in L'Ecuyer et al. (2007), might also be appropriate for this problem, given that those variants are closer to RESTART method than to splitting.

In the three cases the best results are obtained with coefficients *a* and *b* around 20–25% greater than those given by the formulas derived in the paper. Nevertheless, very good results are also obtained with the importance function given by Eq. (24). The low values of f_V achieved show that the application is close to the optimal one, at least for the tested cases. The robustness of the choice of these coefficients is greater than in previous networks, because acceptable results (values of f_V less than double those of the best results) are obtained for values of *a* and *b* up to 20% lower or greater than the optimal ones.

5.2.2. Large Jackson networks with some nodes at distances greater than 2

We studied a Jackson network with 15 nodes: 4 of them at distance 3 from the target node, 5 at distance 2 and 5 at distance 1. After being served at each node, customers could leave the network with a probability of 0.2 or could go to another eight nodes with a probability of 0.1 for each. Twenty minutes of computational time was necessary to estimate a probability of the order of 10^{-33} and 4 min to estimate a probability of the order of 10^{-17}. It seems that large networks are not a problem and that it is not necessary to take into account in the importance function the nodes that are at distance greater than 2 from the target node. As mentioned above, RESTART may improve with the complexity of the systems because of the weaker dependence of the target queue on the length of the other queues.

Remark 2. A more detailed analysis of the influence of nodes at distances greater than 2 on the importance function can be seen in Supplementary material. It is concluded in this material that to take into account in the importance function the nodes at distance 3 improves very slightly the efficiency obtained with Eq. (24). However, if the IF does not contain the nodes at distance 2, the efficiency is much worse. In this material it is also fully specified the network with 15 nodes.

5.2.3. Networks with strong feedback

We consider a 2-node Jackson network with a external arrival rate at each node equal to 1. Customers departing any of the two nodes join the other node with a probability of 0.8, or leave the network with a probability of 0.2. The load of the target node is 1/3, the load of the other node is 2/3 and $\Phi = 0.36Q_1 + Q_2$. Thirty minutes of computational time was necessary to estimate a probability of the order of 10^{-34} and less than 4 min for a probability of the order of 10^{-15}.

Let us now study a Jackson network with 7 nodes but with stronger feedback than the network of Section 5.2.1. The arrival rate from the outside and the probability of leaving the network are the same as in the previous network. The transition matrix is the same as before except that the last two rows in this network are: (0.1, 0.1, 0.1, 0.1, 0.3, 0, 0.1) and (0.1, 0.1, 0.1, 0.1, 0.2, 0.2, 0) respectively. The loads of the nodes are: (0.5, 0.5, 0.5, 0.5, 0.42, 0.40, 0.31), and the importance function given by Eq. (24) is: $\Phi = 0.26(Q_1 + Q_2 + Q_3 + Q_4) + 0.51Q_5 + 0.37Q_6 + Q_7$. Note that customers served at node 6 go with a probability of 0.3 to node 5 (and from there to the target node with a probability of 0.3) and with a probability of 0.1 directly to the target node. As was indicated in Section 4.3, the formulas of the importance function only take into account the customers that go to the target node by the shortest way, and we wanted to check whether the fact that the probability of the shortest way is smaller for customers of node 6 might affect the results. Less than five minutes of computational time was needed to estimate a probability of the order of 10^{-36}. This result is better than that obtained for the first case of Section 5.2.1.

These examples corroborate that, unlike in importance sampling, the efficiency of RESTART is not affected by the feedback, while the second example suggests that the assumption of considering the shortest way to the target node (for deriving the importance function) does not affect the efficiency, either.

5.2.4. Networks with high dependencies and some nodes at distances greater than 2

In Section 5.1 we studied networks for which the dependency of the target queue on the queue length of the other queues was very high because all the customers of the other queues had to go to the target queue. We now consider another example with the same high dependency but with several nodes at distances greater than 2 and with a quite lower load of the target queue. The test case is a six-queue Jackson tandem network, in which the first five nodes have the same load (2/3) and the last (target) node has a lower load (1/3). Thirty minutes of computational time was needed to estimate a probability of the order of 10^{-15} with the importance function given by Eq. (24). A heuristic importance function, $\Phi = 0.15Q_1 + 0.2Q_2 + 0.25Q_3 + 0.31Q_4 + 0.37Q_5 + Q_6$, that takes into account the dependence of the target node on all the nodes and gives lower weights to the nodes that are farther from the target node, provided an accurate estimation of the same probability with 10 min of computational time.

These results show that the worst cases are networks with very high dependencies, in which the target queue has a much lower load than the other queues. But even in a case as "problematic" as the six-queue tandem network, the importance function given by Eq. (24) allows probabilities of the order of 10^{-15} to be estimated with moderate computer times. Although the importance function given by Eq. (24) can be improved for some specific networks, it seems to be good enough for estimating very low probabilities with short or moderate computational times for most (perhaps all) Jackson networks.

6. Conclusions

The choice of the importance function is the most critical feature when the method RESTART is applied to multi-dimensional systems. This paper has focused on finding formulas of effective importance functions for three-stage Markovian networks with any number of nodes and for more general Markovian networks. These formulas could be implemented in tools used for fast simulation of general Jackson networks.

Different types of networks with different loads of the nodes have been simulated and buffer overflow probabilities much lower than those needed in practical problems (around 10^{-34}) have been accurately estimated within reasonable computational work. The efficiency of RESTART may improve with the complexity of the systems because the dependence of the target queue on the queue length of the other queues is weaker. The efficiency does not seem affected by the feedback.

To obtain the formulas of the importance functions it has been necessary to make certain approximations and assumptions. Although some of them are rough approximations, they lead to effective importance functions: in many cases the best simulation results were obtained with the coefficients of the importance function given by the formulas and in the other cases the best coefficients were close to those obtained with the formulas. Let us observe that the approximations could affect to the efficiency of the method, but they do not affect to the correctness of the estimates.

The importance function has been derived equating the importance of one extreme state with the importance of each of the other extreme states. It would be interesting to see whether the importance of the extreme states would be affected in a similar manner when the interarrivals and/or services times are not exponentially distributed. As a consequence, the importance function derived for Jackson networks would be fit for other networks. Future research must determine the types of non-Markovian multi-dimensional networks for which the importance function given by Eq. (24) is valid.

Acknowledgement

I should like to thank Mr. Iván García-Jiménez for his assistance in writing the simulation program and Mr. Manuel Villén-Altamirano for his helpful comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2009.07.013.

References

Batchelor, M.T., Henry, B.I., 2002. Exact solution for random walks on the triangular lattice with absorbing boundaries. Journal of Physics A: Mathematical and General 35, 5951–5961.
Dean, T., Dupuis, P., 2009. The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. Annals of Operation Research.
Dupuis, P., Wang, H., 2009. Importance sampling for Jackson networks. Queueing Systems: Theory and Applications 62 (1,2), 113–157.
Garvels, M.J.J., Ommeren, J.K.C.W., Kroese, D.P., 2002. On the importance function in splitting simulation. European Transaction on Telecommunications 13 (4), 363–371.
Glasserman, P., Heidelberger, P., Shahabuddin, P., Zajic, T., 1999. Multilevel splitting for estimating rare event probabilities. Operation Research 47, 585–600.
Kahn, H., Harris, T.E., 1951. Estimation of particle transmission by random sampling. National Bureau of Standards Applied Mathematics Series 12, 27–30.
Lamers, E., Gorg, C., 2002. Distributed RESTART with MuSICS: rare event simulation on a network of workstations. In: Proceedings of the Fourth Workshop on Rare Event Simulation, Madrid, pp. 117–127.
L'Ecuyer, P., Demers, V., Tuffin, B., 2007. Rare events, splitting and quasi-Monte Carlo. ACM Transaction on Modelling and Computer Simulation 17 (2). Article 9.
Parzen, E., 1999. Stochastic Processes. SIAM, Philadelphia.
Tuffin, B., Trivedi, K.S., 2000. Implementation of importance splitting techniques in stochastic Petri net package. Lecture Notes in Computer Science 1786, 216–229.
Villén-Altamirano, J., 2007. Rare event RESTART simulation of two-stage networks. European Journal of Operational Research 179 (1), 148–159.
Villén-Altamirano, M., Villén-Altamirano, J., 1991. RESTART: a method for accelerating rare event simulations. In: Proceedings of the 13th International Teletraffic Congress. Elsevier Science Publisher, Amsterdam, pp. 71–76.
Villén-Altamirano, M., Villén-Altamirano, J., 2002. Analysis of RESTART simulation: theoretical basis and sensitivity study. European Transaction on Telecommunications 13 (4), 373–386.
Villén-Altamirano, M., Villén-Altamirano, J., 2006. On the efficiency of RESTART for multi-dimensional state systems. ACM Transaction on Modelling and Computer Simulation 16 (3), 251–279.
Zimmermann, A., Knoke, M., Huck, A., Hommel, G., 2006. Towards version 4.0 of TimeNET. In: Proceedings of the 13th GI/ITG Conference on Measurement, Modelling and Evaluation of Computer and Communication Systems, Nürnberg, Germany, pp. 477–480.