Villén-Altamirano, José

Article

An improved variant of the rare event simulation method RESTART using prolonged retrials

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: Villén-Altamirano, José (2019) : An improved variant of the rare event simulation method RESTART using prolonged retrials, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 6, pp. 1-9, https://doi.org/10.1016/j.orp.2019.100108

This Version is available at:
https://hdl.handle.net/10419/246387

Operations Research Perspectives 6 (2019) 100108

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

An improved variant of the rare event simulation method RESTART using prolonged retrials

José Villén-Altamirano

Department of Applied Mathematics, Technical University of Madrid, Calle Alan Turing s/n, Madrid 28031, Spain

ARTICLE INFO

Keywords:
Simulation
Rare event probabilities
Variance reduction
RESTART
Splitting
Queuing networks

ABSTRACT

RESTART is a widely applicable accelerated simulation technique that allows the evaluation of extremely low probabilities. In this method a number of retrials (or paths) are made when the process reaches certain thresholds of a function of the system state, called the importance function. In RESTART with prolonged retrials, all but one path are cut when they drop several thresholds (rather than when they down-cross the threshold that they started from). The only path that continues collects the weight of the cut paths to keep the estimator unbiased.

In this paper a theoretical analysis of this version of the method is made. First the variances of RESTART with prolonged retrials for different degrees of prolongation are compared. Then, formulas for the computational costs of these variants are derived. It is shown that by prolonging the retrials by one or two thresholds, a significant reduction of variance with respect to RESTART is obtained in models where many thresholds can be set (for example, in communication network models). This is attained with a similar or small additional computational cost per sample, so that the gain obtained may even exceed 50%. This gain, which is achieved with no additional effort, illustrates the interest of applying these variants. Greater degrees of prolongation are not advisable because, as the formulas show, any additional reduction of variance is small and does not compensate the additional cost per sample. This would explain the bad behaviour of standard Splitting compared with RESTART.

1. Introduction

The estimation of extremely small but important probabilities is of great interest in many fields. In most of the rare event problems, the estimation based on the mathematical model cannot be made analytically due to the complexity of the model. Although simulation is an effective means of studying such systems, variance reduction techniques are necessary because standard discrete event simulations require prohibitive runtimes for the accurate estimation of very low probabilities.

Importance Sampling and RESTART/Splitting are the two main groups of methods for rare event simulation. Standard splitting (from now on simply "Splitting") was described in Kahn and Harris [10] without deriving the parameters of the method. Villén-Altamirano and Villén-Altamirano [16] proposed RESTART (Repetitive Simulation Trials After Reaching Thresholds) with one threshold and made a theoretical analysis that yields the variance of the estimator and the optimal number of retrials. The method was extended to multiple thresholds in Villén-Altamirano et al. [15]. A rigorous analysis of multiple thresholds was made by Villén-Altamirano and Villén-Altamirano [18] and optimal values for thresholds and the number of retrials that maximize the gain obtained were derived. RESTART/Splitting is increasingly used in a variety of fields and has been studied extensively in the last two decades. Examples of recent applications are: loss probabilities in queuing networks, (e.g., [9,14]), failure probabilities in ultra-reliable systems, (e.g., [12, 21], probability of a process escaping from a neighbourhood of a metastable state, (e.g., [4,11]), probabilities of potential wake encounters in air traffic management (e.g., [20]), etc. Many methods that have appeared in the literature could be considered variants of RESTART or Splitting: Subset Simulation with Splitting in mechanical engineering [6], Forward Flux Sampling in biochemical networks [1], Path Sampling [11], Adaptive multilevel splitting [2,5], Generalized multilevel splitting [3], genealogical particle analysis [7], etc.

In Splitting and RESTART a more frequent occurrence of a formerly rare event is achieved by performing a number of simulation retrials (or paths) 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 are defined by comparing the value taken by a function of the system state, the importance function,

E-mail address: jvillen@etsisi.upm.es.

https://doi.org/10.1016/j.orp.2019.100108
Received 14 December 2018; Received in revised form 25 March 2019; Accepted 26 March 2019
Available online 27 March 2019

with certain thresholds. In RESTART all but one path are cut when they down-cross the threshold where they were generated and new sets of retrials are made if the trial that continues up-crosses again that threshold. In Splitting [10], all the paths continue until the end-of-simulation condition is fulfilled but retrials are not performed if one of these trials up-crosses the threshold where the trial was generated. In RESTART with prolonged retrials, all but one the paths are cut when they drop several thresholds (instead of being cut when they down-cross the threshold that they started from), [14]. New sets of retrials are made if the only path that continues after the down-cross, then up-crosses the threshold where it was generated. This variant is studied because RESTART has greater variance than Splitting since the number of paths that reach a given level is more variable and the retrials are slightly more positively correlated since they are shorter. This correlation, due to the common history that the retrials share, decreases as their lengths increase. However, the computational cost is much greater with Splitting because much time is wasted simulating unpromising trials, and the product between variance and computational cost is also greater. If the retrials of RESTART are prolonged, the variance of the estimator decreases and the computational cost increases after a certain degree of prolongation, and so it is interesting to study the optimal degree of prolongation.

This variant of RESTART was introduced in Villén-Altamirano and Villén-Altamirano [17], where it was called RESTART with Hysteresis and an analysis of the computational cost was made for the case of only one intermediate threshold. In Villén-Altamirano et al. [15], the method was extended to multiple thresholds without further analysis. In Garvels [8] it was called RESTART with truncation, but no theoretical analysis was made. In the simulation study of Villén-Altamirano [14], a better behaviour of RESTART with some degree of prolonged retrials with respect to RESTART was observed in some systems. The need of a theoretical analysis that allows optimizing the degree of prolongation of the paths, as well as the type of systems to which it can be applied, motivated this article. In this paper we calculate both the variance of the estimators and the computational costs and compare the product between variance and computational cost of RESTART with different degrees of prolonged retrials.

The paper is organized as follows: Section 2 presents a review of the methods. Sections 3 and 4 provide a comparison of variances and costs, respectively. Section 5 describes a simulation example and Section 6 states the conclusions.

2. Description of RESTART and RESTART with prolonged retrials

RESTART has been described in detail in several papers (e.g., [18] and [13,19] and [14]). Nevertheless it is described here in order to make this paper more self-contained. This paper also describes the version involving prolonged retrials.

Let Ω denote the state space of a process $X(t)$ and A the rare set whose probability must be estimated. A nested sequence of sets of states C_i, $(C_1 \supset C_2 \supset , \ , \supset C_m)$ is defined, which determines the partitioning of the state space Ω into regions $C_i - C_{i+1}$; the higher the value of i, the greater the importance of the region $C_i - C_{i+1}$. The sets C_i are defined by means of a function, $\Phi:\Omega \to \Re$, called the importance function. Thresholds $T_i, 1 \le i \le M$ of Φ are defined so that each set C_i is associated with $\Phi \ge T_i$. Two events, B_i and D_i, are defined as follows:

B_i: event at which $\Phi \ge T_i$ having been $\Phi < T_i$ at the previous event, that is, the event at which the process enters set C_i (up-crossing threshold T_i). This definition is only valid for RESTART.

D_i: event at which $\Phi < T_i$ having been $\Phi \ge T_i$ at the previous event, that is, event at which the process leaves set C_i (down-crossing threshold T_i).

RESTART works as follows:

- A simulation path, called the main trial, is performed as if it were a crude simulation. It lasts until it reaches a predefined "end-of-simulation" condition.
- Each time an event B_1 occurs in the main trial, the system state is saved and $R_1 - 1$ retrials of level 1 are performed. Each retrial of level 1 is a simulation path that starts with the state saved at B_1 and finishes when an event D_1 occurs.
- After the $R_1 - 1$ retrials of level 1 have been performed, the main trial continues from the state saved at B_1. Note that the total number of simulated paths $[B_1, D_1)$, including the portion $[B_1, D_1)$ of the main trial, is R_1. Each of these R_1 paths is called a trial $[B_1, D_1)$. The main trial, which continues after D_1, leads to new sets of retrials of level 1 if new events B_1 occur.
- R_i trials $[B_i, D_i)$ $(1 \le i \le M)$ are performed each time an event B_i occurs in a trial $[B_{i-1}, D_{i-1})$. The number R_i is constant for each value of i.

In RESTART with prolonged retrials of depth (or degree) j, RESTART-Pj, each of the $R_i - 1$ retrials of level i is a simulation path $[B_i, D_{i-j})$ that also starts with the state saved at B_i but finishes when it leaves set C_{i-j}; that is, it continues until it down-crosses the threshold $i - j$. If one of these trials again up-crosses the threshold where it was generated (or any other between $i - j + 1$ and i) a new set of retrials is not performed and that event is not considered an event B_i. If $j \ge i$, the retrials are cut when they reach the threshold 0. The main trial, which continues after leaving set C_{i-j}, potentially leads to new events B_i and so, to new sets of retrials, if it up-crosses threshold T_i after having left set C_{i-j}. If the main trial reaches the threshold 0, it collects the weight of all the retrials (which has been cut at that threshold) and thus, new sets of retrials of level 1 are performed next time the main trial up-crosses threshold T_1. Note that RESTART and Splitting could be considered as particular cases of RESTART with prolonged retrials of depth j, for $j = 0$ and $j = M$, respectively.

Fig. 1 illustrates a RESTART-P1 simulation with $M = 2$, $R_1 = 3$, $R_2 = 2$, in which the chosen importance function Φ also defines set A as $\Phi \ge L$.

Some more notation: ([18] and [19])

- $r_i = \prod_{j=1}^{i} R_j$, $1 \le i \le M$: accumulative number of trials;
- $R_0 = 1$, $r_0 = 1$, $C_0 = \Omega$, $C_{M+1} = A$;
- $P_{h/i}$ $(0 \le i \le h \le M + 1)$: probability of set C_h, knowing that the system is in a state of set C_i. As $C_h \subset C_i$, $P_{h/i} = \Pr\{C_h\}/\Pr\{C_i\}$;
- $P_{A/i} = P_{M+1/i}$;
- $P = P_{M+1/0} = P_{A/0} = \Pr\{A\}$;
- N_A: total number of events A that occur in the simulation (in the main trial or in any retrial);
- N_i^0 $(1 \le i \le M)$: number of events B_i that occur in the main trial;
- N: number of events simulated in the main trial;
- a_i $(1 \le i \le M)$: expected number of events in a trial $[B_i, D_i)$;
- X_i $(1 \le i \le M)$: random variable indicating the state of the system at

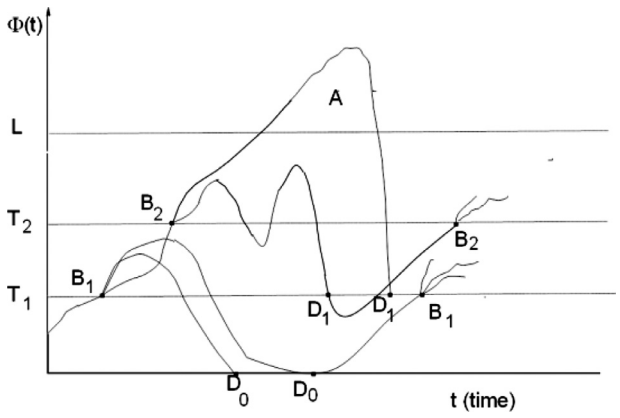

Fig. 1. Simulation with RESTART-P1.

an event B_i;

- Ω_i $(1 \le i \le M)$: set of possible system states at an event B_i;
- P^*_{A/X_i} $(1 \le i \le M)$: importance of state X_i, defined as the expected number of events A in a trial $[B_i, D_i)$ when the system state at B_i is X_i. Note that P^*_{A/X_i} is also a random variable which takes the value $P^*_{A/x_i}$when $X_i = x_i$;
- $P^*_{A/i}(1 \le i \le M)$: expected importance of an event B_i:

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

where $F(x_i)$ is the distribution function of X_i. Note that: $P^*_{A/i} = a_i \cdot P_{A/i}$;

- $V\left(P^*_{A/X_i}\right)$ $(1 \le i \le M)$:variance of the importance of an event B_i:

$$V\left(P^*_{A/X_i}\right) = E\left[\left(P^*_{A/X_i} - P^*_{A/i}\right)^2\right]$$

Set A is defined as $\Phi \ge L$ for the remainder of the work. The estimator of P is given by:

$$\hat{P} = \frac{N_A}{N \cdot r_M}$$

3. Comparison of the variance of the estimators

As $P^*_{A/i} = a_i \cdot P_{A/i}$, the variance of $\hat{P}$in a RESTART simulation with M thresholds derived in Villén-Altamirano and Villén-Altamirano [18] can be written as:

$$V(\hat{P}) = \frac{K_A P}{N}\left(\frac{1}{r_M} + \sum_{i=1}^{M} \frac{s_i P^*_{A/i}(R_i - 1)}{r_i}\right), \tag{1}$$

with

$$s_i = \frac{1}{K_A}\left(K'_i + \frac{V(P^*_{A/X_i})}{(P^*_{A/i})^2}\gamma_i\right)(1 \le i \le M), \tag{2}$$

where $K'_i = V(N_i^0)/E[N_i^0]$. The formula for s_i used in this work differs from that in Villén-Altamirano and Villén-Altamirano [18] precisely by the factor a_i.

As mentioned in the above paper, factor K'_i is a measure of the autocorrelation of the process of the occurrence of events B_i in the main trial. If the process is uncorrelated, K'_i is close to 1. In most applications, the process has a weak positive autocorrelation. As shown in Villén-Altamirano and Villén-Altamirano [17], K'_i decreases slightly as the prolongation of the retrials increases, because the distance between events B_i is greater since there are fewer B_i events. K_A is the same as K'_i but related to events A and it is not affected by the prolongation of the retrials because the expected number of events A is the same for any degree of prolongation of the retrials. Factor γ_i is a measure of the dependence of the importance of the system states X_i on events B_i occurring in the main trial. There may be some dependence between system states of close events B_i but this dependence is negligible for distant events B_i. γ_i is usually close to 1 and decreases very slightly as the retrials are prolonged, for the same reason given for K'_i. First we will compare the variances of the estimators of P with RESTART and RESTART-P1, then the variances with RESTART-P1 and RESTART-P2 and so on. In Section 4, the computational costs of the methods will be compared.

Let us use α_i to denote P^*_{A/X_i} $(1 \le i \le M)$, that is, the importance of state X_i, defined as the expected number of events A in a trial $[B_i, D_i)$ when the system state at B_i is X_i. Let β_i, γ_i and δ_i denote the expected number of events A when the system state at B_i is X_i, but in trials $[B_i, D_{i-1})$, $[B_i, D_{i-2})$ and $[B_i, D_{i-3})$ respectively.

The variance of $\hat{P}$in a RESTART-P1 simulation is also given by Eq. (1) but changing $s_i P^*_{A/i} = s_i E[\alpha_i]$ by $s_{1i} E[\beta_i]$, with

$$s_{1i} = \frac{1}{K_A}\left(K'_{1i} + \frac{V(\beta_i)}{(E[\beta_i])^2}\gamma_{1i}\right)(1 \le i \le M),$$

where $K'_{1i} = V(N_i^0)/E[N_i^0]$. Note that formulas of K'_{1i} and γ_{1i} are the same as those for K'_i and γ_i, respectively, but their values are different because N_i^0, the number of events B_i that occur in the main trial, is lower with RESTART-P1 than with RESTART.

The variance of $\hat{P}$in a RESTART-Pj simulation, for $j > 1$, is defined analogously.

3.1. Comparison of the variances with RESTART and RESTART-P1

In most cases the dependence of the system states at two consecutive events B_i is very weak. Although it is possible to find some cases with significant dependence (e.g., systems that exhibit multimodal behaviour), we will consider for the analysis that they are independent. If a trial down-crosses threshold T_i and then up-crosses the same threshold, before down-crossing threshold T_{i-1}, a new event B_i occurs (observe that it is an event B_i of RESTART, but not of RESTART-P1) and the expected number of events A in the new $[B_i, D_i)$ interval is $P^*_{A/i}$ because we will assume that the system state of B_i can be any state of Ω_i $(1 \le i \le M)$. With this assumption, as β_i denotes the expected number of events A in trials $[B_i, D_{i-1})$ when the system state at B_i is X_i, β_i is α_i plus the expected number of events A in the new $[B_i, D_i)$ intervals: $\beta_i = \alpha_i + (m_{X_i} - 1)P^*_{A/i}$, where m_{X_i}denotes the expected number of $[B_i, D_i)$ intervals in a $[B_i, D_{i-1})$ interval when the system state of the starting event B_i is X_i; Thus:

$$E[\beta_i] = E[\alpha_i + (m_{X_i} - 1)P^*_{A/i}] = m_i P^*_{A/i}$$
$$V(\beta_i) = E[(\alpha_i - P^*_{A/i} + (m_{X_i} - m_i)P^*_{A/i})^2] \simeq V(P^*_{A/X_i}). \tag{3}$$

Note that m_{X_i} is a random variable because it depends on the system state X_i at the starting event B_i and $m_i = E[m_{X_i}]$. As this dependence is very weak, all the values m_{X_i} are similar and the approximation $(m_{X_i} = m_i)$ made in Eq. (3) is accurate, as we will see below. If the importance function is well chosen, all the system states X_i have similar importance, that is, the expected number of events A in a $[B_i, D_i)$ interval is similar for the possible system states X_i at B_i. In this case, the expected number of events B_{i+1} in a $[B_i, D_i)$ interval is even more similar because the expected number of events A in different $[B_i, D_i)$ intervals is even more similar than the expected number of events A. This is because in a $[B_{i+1}, D_i)$ interval, the latter is (similar but) not the same for all posible system states X_{i+1} at B_{i+1}. Following analogous reasoning, the expected number of events B_i in a $[B_i, D_{i-1})$ interval is more similar than the expected number of events A in that interval. This means that $(m_{X_i} - m_i)$ is lower than $(\alpha_i - P^*_{A/i})$ in Eq. (3). As $(m_{X_i} - m_i)$ is multiplied by $P^*_{A/i}$, a number less than one (much smaller than one if the threshold T_i is not close to the threshold L), the product $(m_{X_i} - m_i)P^*_{A/i}$ is negligible compared to $(\alpha_i - P^*_{A/i})$. The above reasoning is also valid if the importance function is poorly chosen, but the approximation made in Eq. (3) may even be more accurate. For example, for any Jackson network the few system states for which m_{X_i} is different are those states for which one or more queues (except the target) are empty. If the network has many nodes, m_{X_i} is significative different only if several queues are empty. In the two-queue tandem network example of Section 5, the only system state with different value of m_{X_i} is the state $(0, T_k)$ for which m_{X_i} is smaller than m_i. It is not necessary that states X_i have similar importance. In that tandem queue of Section 5 if a very bad importance function, $\varphi = q_2$ is chosen, the states (100, 20) and (3, 20), for example, are in the same importance region though the first one has much more importance than the second. However, the value of $m_{X_i}$is the same for both states. Also, the only system state with different value of m_{X_i} is the state (0, 20). In these systems $(\alpha_i - P^*_{A/i})$ is very large and $(m_{X_i} - m_i)P^*_{A/i}$ very small. The previous analysis can be extended to most non-Jackson networks, but it

is difficult to ensure that the approximation is so much accurate for any system if the importance function is poorly chosen. In systems that exhibit multimodal behaviour the approximation may be less accurate.

We compare the products $s_i P^*_{A/i}$ and $s_{1i}E[\beta_i]$ with RESTART and RESTART-P1, respectively because the other terms of Eq. (1) are the same in both methods.

$$\frac{s_i P^*_{A/i}}{s_{1i}E[\beta_i]} = \frac{K_i' + \frac{V(P^*_{A/X_i})}{(P^*_{A/i})^2}\gamma_i}{m_i K_{1i}' + \frac{V(P^*_{A/X_i})}{m_i(P^*_{A/i})^2}\gamma_{1i}} = m_i \frac{K_i' + \frac{V(P^*_{A/X_i})}{(P^*_{A/i})^2}\gamma_i}{m_i^2 K_{1i}' + \frac{V(P^*_{A/X_i})}{(P^*_{A/i})^2}\gamma_{1i}}. \tag{4}$$

If there is only one system state at each B_i or if all the system states at each B_i have the same importance, then $V(P^*_{A/X_i}) = 0$ and

$$\frac{s_i P^*_{A/i}}{s_{1i}E[\beta_i]} = \frac{K_i'}{K_{1i}' m_i} \tag{5}$$

As mentioned above, factor K_i' decreases slightly as the prolongation of the retrials increases. This implies that $K_{1i}' < K_i'$. This inequality can also be proved in the following way: For each event B_i with RESTART-P1, there are $m_i^{\perp}$ events B_i with RESTART, where $m_i^{\perp}$ is the random variable "number of $[B_i, D_i)$ intervals with RESTART in a $[B_i, D_{i-1})$ interval". Let us use here N_{1i}^0 to denote N_i^0 with RESTART-P1. So, we have $N_i^0 = N_{1i}^0 \cdot m_i^{\perp}$ and $E[N_i^0] = E[N_{1i}^0]\cdot m_i$, because N_{1i}^0 and $m_i^{\perp}$ can be considered independent random variables.

$$V(N_i^0) = V(N_{1i}^0 \cdot m_i^{\perp}) = V(N_{1i}^0)m_i^2 + V(m_i^{\perp})(E[N_{1i}^0]^2) + V(N_{1i}^0)\cdot V(m_i^{\perp})$$

Thus,

$$K_i' = K_{1i}'\cdot m_i + V(m_i^{\perp})\cdot E[N_{1i}^0]/m_i + V(N_{1i}^0)\cdot V(m_i^{\perp})/E[N_{1i}^0\cdot m_i]$$

As $m_i > 1$ and the other summands of the equation are positive, it is proven that $K_{1i}' < K_i'$.

As the variance in the number of retrials is lower with RESTART-P1 than with RESTART, the ratio (5) is expected to be greater than 1. In several simulations made with the M/M/1 queue taking thresholds as close as possible, this ratio was around 1.05. However, in the example of the M/M/1 queue shown in Villén-Altamirano [14] the ratio was 1.14 due to the greater distance between thresholds. So, $m_i K_{1i}' < K_i'$ but $m_i^2 K_{1i}'$ will probably be greater than K_i' (slightly greater if m_i is small). Although $\gamma_i$is very slightly greater than γ_{1i}, the ratio of $s_i P^*_{A/i}$ to $s_{1i}E[\beta_i]$ in Eq. (4) will almost certainly be smaller than m_i, and it will be closer to m_i as m_i becomes closer to 1. Due to the term $1/r_M$ in Eq. (1), which is of the same order of magnitude as the other summands of the equation, the ratio between the variances of the estimators (RESTART/RESTART-P1) is slightly smaller than the ratio given in (4).

To gain insight into the possible values of m_i, we define the following probability: "If a trial enters a region $C_{i-1} - C_i$, after down-crossing threshold T_i we define p_i as the expected probability of down-crossing threshold T_{i-1} before up-crossing threshold T_i". Actually, the probability of down-crossing depends on the entrance state to the region $C_{i-1} - C_i$, and so p_i is defined as the expected probability.

Let us now take a look at the distribution of $m_i^{\perp}$ = number of $[B_i, D_i)$ intervals in a $[B_i, D_{i-1})$ interval. The random variable $m_i^{\perp}$ takes the values 1, 2,... with probabilities: $\Pr(m_i^{\perp} = 1) = p_i$,..., $\Pr(m_i^{\perp} = n) = (1 - p_i)^{n-1} p_i$. As $m_i^{\perp}$ follows a geometric distribution,

$$m_i = E[m_i^{\perp}] = \frac{1}{p_i}. \tag{6}$$

3.2. Comparison of the variances with RESTART-P1 and RESTART-P2

As in the previous section, we shall denote the expected number of $[B_i, D_{i-1})$ intervals in a $[B_i, D_{i-2})$ interval, when the system state of the starting event B_i is X_i, as m_{1X_i}.

$$E[\gamma_i] = E[\beta_i + (m_{1X_i} - 1)E[\beta_i]] = m_{1i}E[\beta_i]$$
$$V(\gamma_i) = E[\beta_i - E[\beta_i] + (m_{1X_i} - m_{1i})E[\beta_i]]^2 \simeq V(\beta_i)$$

As before, m_{1X_i} is a random variable because it depends on the system state X_i in the starting event B_i and $m_{1i} = E[m_{1X_i}]$. Giving the terms s_{2i}, K_{2i}' and γ_{2i} to denote factors s_i, K_i' and γ_i corresponding to RESTART-P2, we have:

$$\frac{s_{1i}E[\beta_i]}{s_{2i}E[\gamma_i]} = m_{1i}\frac{K_{1i}' + \frac{V(\beta_i)}{(E[\beta_i])^2}\gamma_{1i}}{m_{1i}^2 K_{2i}' + \frac{V(\beta_i)}{(E[\beta_i])^2}\gamma_{2i}}.$$

In several simulations made with the M/M/1 queue taking thresholds that are as close as possible, and also in the example of Villén-Altamirano [14], this ratio was around 1.04. As before, the ratio of variances of the estimators with RESTART-P1 and RESTART-P2 will be close to m_{1i}, and almost certainly slightly smaller.

To simplify the notation, henceforth we will assume that $p_i = p$, $m_i = m$ and $m_{1i} = m_1$ for all the regions $C_i - C_{i+1}$. This assumption is used to calculate m_1 as a function of p and does not affect the results of the comparison. The random variable $m_1^{\perp}$ = number of $[B_i, D_{i-1})$ intervals in a $[B_i, D_{i-2})$ interval has the following distribution: $\Pr(m_1^{\perp} = 1) = p + (1-p)p^2 + (1-p)^2p^3 + \cdots = \frac{p}{1-p+p^2}$. Analogously: $\Pr(m_1^{\perp} = 2) = \frac{(1-p)^2p}{(1-p+p^2)^2}$,..., $\Pr(m_1^{\perp} = n) = \frac{(1-p)^{2(n-1)}p}{(1-p+p^2)^n}$.$m_1^{\perp}$ thus also follows a geometric distribution but with parameter $p/(1-p+p^2)$, so:

$$m_1 = E[m_1^{\perp}] = \frac{1-p+p^2}{p}. \tag{7}$$

In Eq. (7) it is assumed that p is the expected probability of down-crossing threshold T_{i-1} before up-crossing threshold T_i, not only when a trial enters a region $C_{i-1} - C_i$, after down-crossing threshold T_i but also when they enter that region after up-crossing threshold T_{i-1}. If we define p' as the expected probability of the latter case, $\Pr(m_1^{\perp} = 1) = p + (1-p)p'p + (1-p)^2(p')^2p + \ldots = \frac{p}{1-p'+pp'}$,... and:

$$m_1 = E[m_1^{\perp}] = \frac{1-p'+pp'}{p}. \tag{8}$$

3.3. Comparison of the variances with RESTART-P2, RESTART-P3 and RESTART-P4

As before, the ratio between variances with RESTART-P2 and RESTART-P3 is close to m_2, which is defined as the expected number of $[B_i, D_{i-2})$ intervals in a $[B_i, D_{i-3})$ interval when the system state of the starting event B_i is X_i. Such number of intervals also follows a geometric distribution and:

$$m_2 = \frac{1 - 2p' + pp' + (p')^2 - p(p')^2 + p^2p'}{p - pp' + p^2p'}. \tag{9}$$

If $p = p$':

$$m_2 = \frac{1 - 2p + 2p^2}{p - p^2 + p^3}. \tag{10}$$

Analogously, the ratio between variances with RESTART-P3 and RESTART-P4 is close to:

$$m_3 = \frac{1 - 3p' + pp' + 3(p')^2 - 2p(p')^2 + p^2p' - p^2(p')^2 - (p')^3 + p(p')^3 + p^3p'}{p - 2pp' + p^2p' + p(p')^2 - p^2(p')^2 + p^3p'}. \tag{11}$$

If $p = p$':

$$m_3 = \frac{1 - 3p + 4p^2 - 2p^3 + p^4}{p - 2p^2 + 2p^3}. \tag{12}$$

We can see the numerical values of m, m_1, m_2 and m_3 in two practical cases. The first one corresponds to applications in which there are

no restrictions to setting as many thresholds as possible, as is usual in the study of communication networks. This setting is the optimal one and leads to relatively high values of $(1 - p)$. For example in a M/M/1 queue with arrival rate equal to 1, service rate equal to 2 and thresholds at 1,2, 3,…, $L-1$, the values of p and p' are 2/3 and the values of m, m_1, m_2 and m_3 are 1.50, 1.17, 1.07 and 1.03, respectively. The second case corresponds to applications in which it is not possible to set many thresholds. If $p = 0.99$, the value of m is 1.01 and the rest are greater than one only by fractions on the order of 10^{-4} and lower. The second case is common in reliability studies, such as the study of the Highly Reliable Markovian System (HRMS) made in Villén-Altamirano [14]. This model is often used to represent the evolution of multi-component systems in reliability settings. In the HRMS model, the system has c types of component, with n_i identical components of type i. The system works if at least r_i components of each type i work. In the simplest case, $c = 1$, the importance function is $\varphi(t)$ = Number of components failed at time t. As the thresholds are associated to failures of components, the maximum number of thresholds that can be set are $n - r$. In the first example of that paper two cases were studied, $n = 5$, $r = 1$ and $n = 10$, $r = 1$ and the unavailabilities obtained were of the order of E-13 and E-24, respectively. As we can set only 4 thresholds if $n = 5$ and 9 thresholds if $n = 10$, the distances between thresholds, $P_{i/i-1}$, are much lower than 0.01 for all i, and so p is greater than 0.99. If $c > 1$, the importance function is: $\varphi(t) = \text{Max}_I \{\text{Number of components of type } i \text{ failed at time } t/(n_i - r_i + 1)\}$. In the mentioned paper, an example with $c = 6$ was also studied and only 8 thresholds could be set to estimate a probability of the order of E-20, and so the value of p was also greater than 0.99.

In the first case, if we set as many thresholds as possible, the "distance" between two consecutive thresholds is $P_{i/i-1} = 0.5$, and the optimal number of retrials is $R_i = 1/P_{i/i-1} = 2$, for all i. As R_i has to be an integer number, $P_{i/i-1}$ must be 0.5 or lower. As $P_{i/i-1}$ decreases, p and p' increase and we observe in Eqs. (7)–(12) that the reduction of variances decreases. For example if the service rate equal to 3 in this queue, and the arrival rate and thresholds do not change, then $P_{i/i-1} = 1/3$, $R_i = 3$ and $p = p' = 3/4$. The values of m, m_1, m_2 and m_3 are 1.33, 1.08, 1.03 and 1.01, respectively, which are 11%, 8%, 4% and 2%, respectively, lower than before. In the reliability case, $P_{i/i-1}$ is much smaller, so, R and p are much greater and all the methods have similar variance.

We also observe that in the first case the reduction of variance of the estimators is significant between RESTART and RESTART-P1, is lower between RESTART-P1 and RESTART-P2 and decreases as the depth of prolongation increases, becoming insignificant if we prolong the retrials more than 3 or 4 thresholds. As Splitting can be defined as RESTART-PM, the reduction of variance between Splitting and RESTART-P4 is insignificant, as was observed in some examples of the simulation study made in [14].

4. Comparison of the simulation costs of each method

Let X, Y denote the expected computational costs of simulating the interval $[B_i, D_{i-1})$ by applying RESTART and RESTART-P1, respectively but only taking into account the costs of the region $C_{i-1} - C_{i+1}$. The total expected computation costs of the methods are approximately M times these values. Let c denote the expected computation cost of a trial from when it enters a region $C_i - C_{i+1}$ until it leaves it. This cost c, which is proportional to the expected number of events that occur in that region, is assumed to be the same for all regions. This assumption simplifies the comparison without affecting the results because it does not favour or penalize any of the methods. Let X_S, Y_S denote this type of cost in the region $C_{i-1} - C_{i+1}$, and X_R, Y_R the restoration costs, that is, the costs associated with saving the system state when an event B_i occurs and with restoring the system state and rescheduling the scheduled events at each retrial. So, $X = X_s + X_r$; $Y = Y_s + Y_r$. To compare the methods with deeper prolongation retrials, the expected costs of simulating the intervals $[B_i, D_{i-j})$, with $j > 1$ will be calculated.

4.1. Comparison of the simulation costs with RESTART and RESTART-P1

We calculate the expected cost of the R trials in the $[B_i, D_{i-1})$ interval. $X_s = (Rc + c)p + 2(Rc + c)(1 - p)p + 3(Rc + c)(1 - p)^2 p$.

$$X_s = \frac{(Rc + c)}{p} \cdots = (Rc + c) \sum_{n=1}^{\infty} n(1 - p)^{n-1} p$$

Note that each of the R trials counts c only once in the region $C_i - C_{i+1}$ (term Rc), since if one of them hits threshold $i + 1$, the cost c which is accounted for is the cost of ascent, because the cost of descent (after D_{i+1} occurs) is accounted for in the cost of the $[B_{i+1}, D_i)$ interval. The term c of the first summand corresponds to the only trial that continues after D_i and this trial finishes with probability p if an event D_{i-1} occurs. The product by the summation weighs these counts by their corresponding probability of occurrence, accounting for the possibility that the main trial from $C_{i-1} - C_i$ enters region $C_i - C_{i+1}$ again by up-crossing threshold T_i. Analogously:

$$Y_s = Rc + Rcp \sum_{n=0}^{\infty} (2n + 1)(1 - p)^n = \frac{2Rc}{p}. \tag{13}$$

In this case the stochastic behaviour of the R trials is the same since all of them continue after D_i occurs. For each of the R trials an event D_{i-1} occurs with probability p and there is a cost c of descent; with probability $(1 - p)$ the trial enters the region $C_i - C_{i+1}$ and there is a cost c of ascent to that region and a cost c of descent from it. This process can occur infinite times before event D_{i-1} occurs with the probabilities given in Eq. (13). Note that $2c/p$ is the expected simulation cost of each trial.

To calculate the restoration costs X_r and Y_r of the R retrials in the interval $[B_i, D_{i-1})$, let r denote the cost associated with saving/restoring the system state and rescheduling the scheduled events at each retrial. So:

$$X_r = Rr \sum_{n=1}^{\infty} n(1 - p)^{n-1} p = \frac{Rr}{p}.$$

$Y_r = Rr$, because new retrials of level i are not made with RESTART-P1 although one of the trials hits threshold T_i after event D_i occurs. Thus:

$$\frac{Y}{X} = \frac{2R\frac{c}{r} + Rp}{(R + 1)\frac{c}{r} + R} \tag{14}$$

This ratio can be greater or smaller than one, depending on the values of R, c, p and r. Moreover Y/X increases as c/r increases and if $r << c$, the cost with RESTART-P1 is greater than with RESTART since Y/X is close to $2R/(R + 1)$. $Y = X$ if $c/r = R(1 - p)/(R - 1)$. If $r >> c$, Y/X is close to p and so the cost with RESTART-P1 is lower than with RESTART. The value of r is smaller in Markovian than in non-Markovian systems. The greater the number of thresholds set, the lower the values of p, c and R. All these values will be estimated in the network example of Section 5.

4.2. Comparison of the simulation costs with RESTART-P1 and RESTART-P2

Let Y', Z define the expected computational costs of simulating the interval $[B_i, D_{i-2})$ applying RESTART-P1 and RESTART-P2, respectively. As in the previous section, $Y' = Y'_s + Y'_r$; $Z = Z_s + Z_r$, where the indexes s and r refer to simulation and restoration costs, respectively.

To calculate Y'_s and Z_s, we use Eq. (13) of the simulation costs of the interval $[B_i, D_{i-1})$, calculated in the previous section.

$$Y'_s = (Y_s + c) + (Y_s(1-p') + 2c)(1-p)p + \ldots + [nY_s(1-p') + 2nc](1-p)^n p + \ldots$$

$$Y'_s = \frac{Y_s(1-p'+pp')}{p} + \frac{c(2-p)}{p}$$

Observe that the only trial that continues after down-crossing threshold $i-1$, may up-cross the same threshold again with probability $(1-p)$ and then up-cross threshold i with probability $(1-p')$ (generating R retrials with a cost Y_S) or down-cross threshold $i-1$ with probability p'. Both cases give rise to an additional cost of $2c$. If a given trial has up-crossed threshold $i-1$ n times, then $n(1-p')$ is the expected number of times that the trial will also up-cross threshold i.

Analogously:

$$Z_s = (Y_s + Rc) + R\left[\left(\frac{Y_s}{R}(1-p') + 2c\right)(1-p)p + \ldots + \left(\frac{Y_s}{R}n(1-p') + c2n\right)(1-p)^n p + \ldots\right]$$

$$Z_s = \frac{Y_s(1-p'+pp')}{p} + \frac{Rc(2-p)}{p}. \text{ If } p = p' \text{ then}$$

$$: Y'_s = \frac{Y_s(1-p+p^2)}{p} + \frac{c(2-p)}{p}, \text{and}$$

$$Z_s = \frac{Y_s(1-p+p^2)}{p} + \frac{Rc(2-p)}{p} = \frac{2Rc(1-p+p^2)}{p^2} + \frac{Rc(2-p)}{p} \quad (15)$$

To calculate the expected restoration cost $Y'_r = E\left[Y_r^{'\perp}\right]$, we use:

$P\left(Y_r^{'\perp} = Rr\right) = \sum_{n=0}^{\infty}(1-p)^n(p')^n p = \frac{p}{1-p'+pp'}$. Then:

$$Y'_r = Rr\left(\frac{p}{1-p'+pp'} + 2\frac{(1-p)(1-p')}{1-p'+pp'}\frac{p}{1-p'+pp'} + \ldots + n\left(\frac{(1-p)(1-p')}{1-p'+pp'}\right)^{n-1}\frac{p}{1-p'+pp'} + \ldots\right).$$

As $Y_r^{'\perp}$ follows a geometric distribution, $Y'_r = Rr\frac{1-p'+pp'}{p}$.

These restoration costs could also be obtained using Markov chains. We define four states of the chain A, B, C and D which correspond to system states of the process in regions C_i, $C_{i-1} - C_i$, $C_{i-2} - C_{i-1}$ and $C_{i-3} - C_{i-2}$, respectively. The state D is absorbent because the restoration costs are considered in the interval $[B_i, D_{i-2})$. The transition matrix is given by:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1-p' & 0 & p' & 0 \\ 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The transition probability from state A to state C is one because if the process returns to A visiting only state B, there are no new restoration costs, since no new event B_i occurs. According to Markov chains theory, the average number of times that the process starting in state A visits this state is given by:

$\frac{1-p'+pp'}{p}$. So, Y'_r is obtained by multiplying this quantity by Rr.

If $p = p'$ then: $Y'_r = Rr\frac{1-p+p^2}{p}$.

This methodology will also be followed to obtain the restoration costs in the following section. However it cannot be applied to calculate the simulation costs Y'_s because if the process returns to state A after visiting state B the simulation costs will depend on the previous states of the process. If the process has previously visited state C, retrials will be made and so the simulation costs will be R times greater than if the process had not visited C. Hence the process is not Markovian.

The restoration costs of RESTART-P2 are: $Z_r = Rr$. Thus:

$$\frac{Z}{Y'} = \frac{2R\frac{c}{r}\frac{(1-p'+pp')}{p^2} + R\frac{c}{r}\frac{(2-p)}{p} + R}{2R\frac{c}{r}\frac{(1-p'+pp')}{p^2} + \frac{c}{r}\frac{(2-p)}{p} + R\frac{(1-p'+pp')}{p}}. \quad (16)$$

If $p = p'$ the same formula is obtained but replacing $(1-p'+pp')$ by $(1-p+p^2)$.

4.3. Comparison of the simulation costs with RESTART-P2 and RESTART-P3

Let Z' and T define the expected computational costs of simulating the interval $[B_i, D_{i-3})$ applying RESTART-P2 and RESTART-P3, respectively. As in previous sections, $Z' = Z'_s + Z'_r$; $T = T_s + T_r$. The formulas of T and Z' will be obtained for the case $p = p'$.

Let W denote the expected simulation costs from when a trial enters region $C_{i-2} - C_{i-1}$ (up-crossing threshold $i-2$ or down-crossing threshold $i-1$) until it down-crosses threshold $i-2$. Then:

$W = c + (1-p)(Z_s(1-p) + Wp + c)$, where Z_S is given by Eq. (15). Thus: $W = \frac{Z_s(1-p)^2 + c(2-p)}{1-p+p^2}$.

$$Z'_s = (Z_s + c) + \sum_{n=1}^{\infty} n(W+c)(1-p)^n p = \frac{Z_s(1-2p+2p^2) + c(3-4p+2p^2)}{p-p^2+p^3}.$$

Observe that the only trial that continues after down-crossing threshold $i-2$, may up-cross the same threshold with probability $(1-p)$ and that each time carries out such an up-crossing, the additional simulation cost is W.

Analogously:

$$T_s = \frac{Z_s(1-2p+2p^2) + Rc(3-4p+2p^2)}{p-p^2+p^3}$$

The restoration costs Z'_r are calculated using the methodology of Markov chains, as explained in a previous section. In this case a new state E which corresponds to system states of the process in region $C_{i-4} - C_{i-3}$ is defined. This state is absorbent and state D is transient. The transition matrix is given by:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The restoration costs Z'_r are Rr multiplied by the average number of times the process starting in state A visits that state:

$$Z'_r = Rr\frac{1-2p+2p^2}{p-p^2+p^3}.$$

The restoration costs of RESTART-P3 are: $T_r = Rr$. Thus:

$$\frac{T}{Z'} = \frac{\left(\frac{2R\frac{c}{r}(1-p+p^2)}{p^2} + \frac{R\frac{c}{r}(2-p)}{p}\right)(1-2p+2p^2) + R\frac{c}{r}(3-4p+2p^2) + R(p-p^2+p^3)}{\left(\frac{2R\frac{c}{r}(1-p+p^2)}{p^2} + \frac{R\frac{c}{r}(2-p)}{p}\right)(1-2p+2p^2) + \frac{c}{r}(3-4p+2p^2) + R(1-2p+2p^2)}. \quad (17)$$

The formula of T/Z' has not been obtained for the case $p \neq p'$ but as will be seen in the numerical example of Section 5, the results of Z/Y' are similar whether $p = p'$ or whether p is 10 or 20% smaller than p'.

We can see the numerical values of the above ratios in the M/M/1 queue described at the end of Section 3.3, where $p = p' = 2/3$ and $R = 2$. The value of Y/X is $(4c/r + 4/3)/(3c/r + 2)$, the value of Z/Y' is $(11c/r + 2)/(9c/r + 7/3)$ and the value of T/Z' is $(8.556c/r + 1.037)/(7.333c/r + 1.111)$. If $c = r$, $Y/X = 1.07$, $Z/Y' = 1.15$ and $T/Z' = 1.14$. If $c = 2r$, $Y/X = 1.17$, $Z/Y' = 1.18$ and $T/Z' = 1.15$. If $c = 0.5r$, $Y/X = 0.95$, $Z/Y' = 1.10$ and $T/Z' = 1.11$. Since c decreases as the thresholds become closer to each other, the last case is the most realistic one when many thresholds can be set. In this queue only one event

Table 1
Gain obtained by the variants of RESTART with respect to RESTART for different values of the parameters.

	RESTART-P1	RESTART-P2	RESTART-P3
$p = 0.65, R = 2, c/r = 0.3$	1.78	2.02	2.01
$p = 0.65, R = 2, c/r = 0.5$	1.63	1.77	1.73
$p = 0.65, R = 3, c/r = 0.3$	1.72	1.87	1.80
$p = 0.65, R = 3, c/r = 0.5$	1.55	1.60	1.50
$p = 0.75, R = 3, c/r = 0.3$	1.38	1.34	1.21
$p = 0.75, R = 3, c/r = 0.5$	1.27	1.18	1.04
$p = 0.75, R = 4, c/r = 0.3$	1.36	1.29	1.15
$p = 0.75, R = 4, c/r = 0.5$	1.24	1.12	0.97

(arrival or end of service) occurs in each region $C_i - C_{i+1}$ and so, $c < r$. In the second case of this queue where $p = p' = 3/4$ and $R = 3$, if $c = 0.5r$, $Y/X = 1.05$, $Z/Y' = 1.17$ and $T/Z' = 1.16$.

As can be seen, the ratio of the computational costs between RESTART-P$_i$ + 1 and RESTART-P$_i$ increases as i increases, for realistic values of the ratio c/r. This is an expected result because the higher the value of i, the greater the possibility that the process will up and down cross thresholds many times before all the paths (except one) are cut. Hence, the ratio of the simulation costs (which is greater than 1) increases. The ratio of the restoration costs (which is lower than 1) also increases because the restoration costs significantly decrease when the retrials are prolonged by one threshold; however, as more thresholds are prolonged, the restoration costs decrease more slowly.

The gain obtained by RESTART-P$_j$ with respect to RESTART is defined as the ratio of the products between variance and computational cost of RESTART and RESTART-P$_j$. It is interesting to study how the gain changes with respect to the significant parameters of the method. The figures of Table 1 are obtained with Eqs. (6), (7), (10), (14), (16), and (17).

It is seen that the smaller the ratio c/r, the greater the gain when the retrials are prolonged. This is because, as indicated above, if the retrials are prolonged, the simulation costs increase and restoration costs are reduced because there are fewer B_i events. Therefore, the greater the costs of restoration in relation to those of simulation, the greater the gain. With Markovian systems the restoration costs are lower than with non-Markovian systems because the rescheduling is straightforward. Hence, the gain obtained when the retrials are prolonged is greater with non-Markovian systems.

By increasing R the gain decreases when the retrials are prolonged since, as can be observed in the Eqs. (6), (7), (10), (14), (16), and (17), the cost ratios between RESTART-P$_i$ + 1 and RESTART-P$_i$ increase while the variances do not change. This decrease will be more significant the longer the retrials last.

The value of p depends on both the system being simulated and the distance between thresholds. Since p and $P_{i/i-1}$ are negatively correlated, the increase in p produces the same effect as a decrease in $P_{i/i-1}$: the gain decreases when the retrials are prolonged because the simulation costs increase more than the restoration costs.

5. Application example

Consider the two-queue Jackson tandem network shown in Fig. 2. Messages with Poisson arrival enter the first queue and, after being served, enter the second one. The mean arrival rate is $\lambda = 1$ and the service time is exponentially distributed in each queue with service rates of $\mu_1 = 2$ and $\mu_2 = 3$, respectively. The buffer space at each queue is assumed to be infinite. This model has received considerable attention in the rare event literature, see e.g., Villén-Altamirano and Villén-Altamirano [19] and 11 references therein. The difficulty of applying accelerated simulation techniques arises when the first queue is the bottleneck and the rare set definition is related to the value of Q_2. So, the rare set A is defined as $Q_2 \geq L$, with $L = 30$.

For this model we will estimate the variances and the simulation costs of the different variants applying the formulas obtained in previous sections, comparing them with the simulation results.

To estimate the values of p and p', we use the importance funtion derived in Villén-Altamirano [13]: $\Phi = \frac{\ln \rho_1}{\ln \rho_2} Q_1 + Q_2 = 0.63 Q_1 + Q_2$ (ρ_1 and ρ_2 are the loads of the nodes), setting the thresholds at every integer number. This is the optimal setting, because the thresholds are as close as possible and it is not possible to cross more than one threshold in one step. Note that when a message leaves node 2, the importance function decreases by one unit and the process down-crosses a threshold. If the process down-crosses thresholds T_i entering region $C_{i-1} - C_i$, we assume that the value X of the importance function follows a uniform distribution in the interval $[T_{i-1}, T_i)$ of length 1, because the only information available is that X is in that interval. With this assumption we calculate the expected probability of down-crossing thresholds T_{i-1}: $p_x = 29/36$ if $x \in [T_{i-1}, T_{i-1} + 0.26)$, $p_x = 27/36$ if $x \in [T_{i-1} + 0.26, T_{i-1} + 0.37)$, $p_x = 24/36$ if $x \in [T_{i-1} + 0.37, T_{i-1} + 0.63)$, and $p_x = 18/36$ if $x \in [T_{i-1} + 0.63, T_{i-1} + 1)$. So the expected value is $p = 0.65$ and the standard deviation 0.14. However, if the process has entered region $C_{i-1} - C_i$, by up-crossing thresholds T_{i-1}, it is due to a new arrival with a probability of 0.46 or to an end of service at node 1 with a probability of 0.54. In the first case we assume that the value X of the importance function follows a uniform distribution in the interval $[T_{i-1}, T_{i-1} + 0.63)$, and in the second case in the interval $[T_{i-1}, T_{i-1} + 0.37)$. The expected probability of down-crossing thresholds T_{i-1} is $p' = 0.77$ and the standard deviation 0.05.

According to Eqs. (6), (7), (10), and (12) derived in Section 3, that is assuming $p = p' = 0.65$: $m = 1.54$, $m_1 = 1.19$, $m_2 = 1.09$, $m_3 = 1.04$. According to Eqs. (8), (9), and (11), that is assuming $p = 0.65$, $p' = 0.77$: $m_1 = 1.12$, $m_2 = 1.04$, $m_3 = 1.01$. The value of m does not change if $p \neq p'$ Remember that m should be close to the ratio between the variances with RESTART and RESTART-P1, and m_i should be close to this ratio between RESTART-Pi and RESTART-Pi + 1, for $i = 1, 2, 3$.

According to Eqs. (14), (16), and (17), the ratio of the simulation costs for $p = p' = 0.65$ and $R = 3$ are: $Y/X = (6c/r + 1.95)/(4c/r + 3)$, $Z/Y' = (17.20c/r + 3)/(13.05c/r + 3.57)$ and $T/Z' = (13.11c/r + 1.51)/(10.62c/r + 1.63)$. If $c = r$, $Y/X = 1.14$, $Z/Y' = 1.22$ and $T/Z' = 1.19$. If $c/r = 0.5$, $Y/X = 0.99$, $Z/Y' = 1.15$ and $T/Z' = 1.16$. If $c/r = 0.3$, $Y/X = 0.89$, $Z/Y' = 1.09$ and $T/Z' = 1.13$. Assuming $p = 0.65$ and $p' = 0.77$, Y/X does not change, $Z/Y' = (16.60c/r + 3)/(12.45c/r + 3.37)$ while T/Z' has not been calculated. If $c = r$, $Z/Y' = 1.24$, if $c/r = 0.5$, $Z/Y' = 1.18$ and if $c/r = 0.3$, $Z/Y' = 1.12$.

We now compare the above theoretical results with the results of the simulation of this model for estimating $P(Q_2 \geq 30) = 4.86$ E-15, with the different variants of RESTART. We shall denote the observed simulation cost per sample of RESTART,..., RESTART-P4 as x, y, z, t, v, respectively. Each sample finishes when all the retrials reach a given end time, $t = 30{,}000$. After each sample the relative error (half width of the confidence interval divided by the estimate) is calculated and the simulation finishes when the error is smaller than 0.005. To minimize the influence of the randomness in the comparisons, five simulation runs were performed for each case and only the results corresponding to the median of the computational times are given in the table. All the

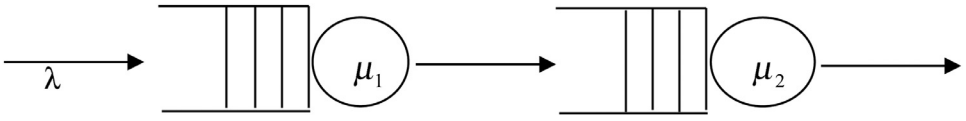

Fig. 2. Two-queue tandem network.

Table 2
Simulation results for the two-queue Jackson tandem network $P = P(Q_2 \geq 30)$, $\rho_1 = 1/2$, $\rho_2 = 1/3$, $\lambda = 1$, $t = 30{,}000$, relative error = 0.005.

Method	$\hat{p}$	Samples	Time (sec)	Variance ratios	Cost ratios
RESTART	4.85 E-15	3909	2906	$m = 1.29$	$y/x = 0.93$
RESTART-P1	4.86 E-15	3032	2107	$m_1 = 1.14$	$z/y = 1.13$
RESTART-P2	4.86 E-15	2660	2091	$m_2 = 1.07$	$t/z = 1.17$
RESTART-P3	4.87 E-15	2476	2287	$m_3 = 1.01$	$v/t = 1.40$
RESTART-P4	4.85 E-15	2458	3188		

Table 3
Gain obtained by the variants of RESTART with respect to RESTART for different networks.

Method	Jackson 3	Non-Jackson 3	Jackson 7	Non-Jackson 7
RESTART-P1	1.54	1.52	1.33	1.33
RESTART-P2	1.41	1.52	1.33	1.65

experiments were run on a personal computer with a 3.20 GHz Intel Core i7 processor and 16 GB RAM. The results are given in Table 2.

All the methods give estimates of P that are very close to the analytic result. As can be seen, the variance decreases as the retrials are prolonged since a lower number of samples is needed to obtain the same relative error as the depth of prolongation increases. However, the decrease is small for RESTART-Pi, with $i > 2$. These observations agree with the theoretical results. The observed values of m_1, m_2 and m_3 are close to those derived in Section 3, even if we assume that $p = p$'. As it is more realistic to assume that $p \neq p$', the differences are smaller in this case (lower than 3%). However, the observed value of m is 16% less than the value derived in Eq. (6). As mentioned in that section, this occurs when the value of m is much greater than one (1.54 in this case).

As regards the cost per sample, the lowest computation cost was obtained with RESTART-P1, due to the much lower restoration costs (Y_r) than those arising from RESTART since the number of events B_i is lower. This compensates the greater simulation costs (Y_s), which increase with the prolongation of the retrials. For greater depths of prolongation the restoration costs decrease slowly while the simulation costs increase rapidly. So, the computational cost per sample with RESTART-P2 is 13% greater than with RESTART-P1 and the same cost with RESTART-P4 is 40% greater than with RESTART-P3. The observed computational cost ratios are close to the theoretical values for $c/r = 0.40$.

The gain obtained with these variants of RESTART, that is, the reduction of cost multiplied by variance with respect to RESTART, measured as the ratio between computational times, is around 38%–39% for RESTART-P1 and P2 and around 27% for RESTART-P3 (Table 2).

In order to validate the theoretical results obtained in this paper, we will compare the theoretical gain of RESTART and its variants P1 and P2 with the gain observed in four queuing networks simulated in Villén-Altamirano [14]. In that paper it was observed that the gain obtained with RESTART-P3 and P4 was always smaller than the gain obtained with P1 and P2, which agrees with the theoretical results obtained in this article, as can be seen in Table 1. Table 3 gives the results of the four networks studied in that paper but in a different format. Jackson 3 is a three-queue Jackson tandem network with $\lambda = 2$, $\mu_1 = 3$, $\mu_2 = 4$ and $\mu_3 = 6$. Non Jackson 3 is the same network with the same loads of the nodes but with hyper-exponential inter-arrival times and Erlang service times. The topology and the parameters of seven node networks are given in the paper mentioned above.

It is difficult to calculate the exact value of p and c/r in these networks. In Jackson 3, the load of the target queue is 1/3, the average distance between thresholds $P_{i/i-1}$ is close to this value and so, R is close to 3. The observed gain with RESTART-P1 almost matches the theoretical gain for $p = 0.65$ and $c/r = 0.5$, which are plausible values. The observed gain with P2 should be similar to that observed with P1, as in the network of 2 nodes, but is a 8.5% lower. The difference can only be explained by the random nature of the results of the simulation. In Jackson 7, the load of the target queue is lower than 1/3, $P_{i/i-1} < 1/3$ and so, $R > 3$ and surely $p > 0.65$. These values of the parameters could explain the lower gain observed in this network. The observed gain is significantly increased by simulating the respective non-Jackson networks with RESTART-P2, which is in accordance with the theoretical results since c/r is lower. A slight increase in the gain observed with P1 was also expected, but the results are similar. Finally, the HRMS systems discussed at the end of Section 3 were also simulated in Villén-Altamirano [14], and slightly better results were obtained with RESTART than with its variants, which is in accordance with the theoretical results since the values of R and p are very high.

6. Conclusions

Formulas for the computational time ratios and for the ratios of variances between RESTART and its variants, with prolonged retrials, were obtained. The observed computational cost and variance ratios were close to the theoretical values in the classical two-queue Jackson tandem network example and in other networks studied in Villén-Altamirano [14].

The reduction of variance is substantial when the retrials are prolonged by one or two thresholds in models where many thresholds can be set, but any further reduction is much less pronounced if the retrials are prolonged more than two thresholds. By contrast, the computational cost is similar (or even slightly lower) when the retrials are prolonged by one threshold but increases significantly as the degree of prolongation increases. As a consequence, RESTART-P1 and RESTART-P2 need smaller computational times than RESTART in these models to obtain estimates with the same relative error. The gain obtained is around 38%–39% in the example. This gain, which is achieved with no additional effort, illustrates the interest of applying these variants. However, those computational times are similar with all methods in models where the distance between thresholds is much greater.

The very slight reduction of variance and the significant increase in computational cost when the retrials are prolonged by more than 3 thresholds explains the poor performance of Splitting (compared with RESTART) observed in the simulation study of Villén-Altamirano [14] because Splitting can be considered a particular case of this method (RESTART-PM) in which the retrials are prolonged until the first threshold is down-crossed.

Acknowledgement

This research was partially supported by Comisión Interministerial de Ciencia y Tecnología (CYCIT), grant MTM2017-86875-C3-3-R.

References

[1] Allen RJ, Warren PB, ten Wolde PR. Sampling rare switching events in biochemical networks. Phys Rev Lett 2005;94:018104.
[2] Bréhier C-E, Gazeau M, Goudenège L, Lelièvre T, Rousset M. Unbiasedness of some generalized adaptive multilevel splitting algorithms. Ann Appl Probab 2016;26(6):3559–601.
[3] Botev ZI, L'Ecuyer P, Rubino G, Simard R, Tuffin B. Static network reliability estimation via generalized splitting. Informs J Comput 2013;25(1):56–71.
[4] Buijsrogge A, Dupuis P, Snarski M. Importance sampling versus RESTART for simulation over long time intervals. Presentation at the 12th international workshop on rare-event simulation (RESIM 2018). 2018.
[5] Cérou F, Guyader A. Adaptive multilevel splitting for rare event analysis. Stoch Anal Appl 2007;25(2):417–43.
[6] Ching J, Au SK, Beck JL. Reliability estimation for dynamical systems subject to stochastic excitation using subset simulation with Splitting. Comput Methods Appl Mech Eng 2005;194:1557–79.
[7] Del Moral P, Garnier J. Genealogical particle analysis of rare events. Ann Appl Probab 2005;15(4):2496–534.
[8] Garvels MJJ. The splitting method in rare event simulation PhD Thesis University of Twente; 2000

[9] Hyytiä E, Righter R. Evaluating rare events in mission critical dispatching systems. Proc. 30th international teletraffic congress 2018. https://doi.org/10.1109/ITC30.2018.00010.
[10] Kahn H, Harris TE. Estimation of particle transmission by random sampling. Natl Bur Stand Appl Math Ser 1951;12:27–30.
[11] Swenson DWH, Prinz JH, Noe F, Chodera JD, Bolhuis PG. OpenPathSampling: a Python framework for path sampling simulations. I. Basics. J Chem Theory Comput 2019;15(2):837–56. https://doi.org/10.1021/acs.jctc.8b00626.
[12] Turati P, Cammi A, Lorenzi S, Pedroni N, Zio E. Adaptive simulation for failure identification in the advanced lead fast reactor European demonstrator. Prog Nucl Energy 2018;103:176–90.
[13] Villén-Altamirano J. Importance function for RESTART simulation of general Jackson networks. Eur J Oper Res 2010;203(1):156–65.
[14] Villén-Altamirano J. RESTART vs. Splitting: a comparative study. Perform Eval 2018;121–122:38–47.
[15] Villén-Altamirano M, Martínez-Marrón A, Gamo JL, Fernández-Cuesta F. Enhancement of the accelerated simulation method RESTART by considering multiple thresholds. Proc. 14th international teletraffic congress. 1994. p. 797–810.
[16] Villén-Altamirano M, Villén-Altamirano J. RESTART: a method for accelerating rare event simulations. Proc. 13th international teletraffic congress. Amsterdam: Elsevier Science Publisher; 1991. p. 71–6.
[17] Villén-Altamirano M, Villén-Altamirano J. Accelerated simulation of rare event using RESTART method with hysteresis. Proc. ITC specialists' seminar on telecommunication services for developing economies. 1991. p. 240–51.
[18] Villén-Altamirano M, Villén-Altamirano J. Analysis of RESTART simulation: theoretical basis and sensitivity study. Eur Trans Telecommun 2002;13(4):373–86.
[19] Villén-Altamirano M, Villén-Altamirano J. On the efficiency of RESTART for multidimensional state systems. ACM Trans Model Comput Simul 2006;16(3):251–79.
[20] Zare-Noghabi A, Shortle JF. Rare event simulations for potential wake encounters. Proc. 2017 winter simulation conference. 2017. p. 2554–65.
[21] Zimmermann A, Maciel P. Dependability evaluation of AFDX real-time avionic communication networks. Proc. 22nd IEEE pacific rim international symposium on dependable computing. 2017. p. 22–5.