Polynomial Time Algorithms for Estimation of Rare Events in Queueing Models

Kriman, V.; Rubinstein, R.Y.

Publication date:
1995

Polynomial Time Algorithms for Estimation of Rare Events in Queueing Models

Vladimir Kriman
and **Reuven Y. Rubinstein**

Faculty of Industrial Engineering and Management
Technion—Israel Institute of Technology, Haifa 32000, Israel

Abstract

This paper presents a framework for analyzing time complexity of rare events estimators for queueing models. In particular it deals with polynomial and exponential time *switching regenerative* (SR) estimators for the steady-state probabilities of excessive backlog in the $GI/GI/1$ queue, and some of its extensions. The SR estimators are based on *large deviation theory* and *exponential change of measure*, which is parametrized by a scalar t. We show how to find the optimal value w of the parameter t, which leads to the *optimal exponential change of measure* (OECM), and we find conditions under which the OECM generates *polynomial time* estimators. We, finally, investigate the "robustness" of the proposed SR estimators, in the sense that we find how much one can perturb the optimal value w in the OECM such that the SR estimator still leads to dramatic variance reduction and still is useful in practice. Our extensive numerical results suggest that if the optimal parameter value w is perturbed up to 20%, we only lose 2–3 orders of magnitude of variance reduction compared to the orders of tenth under the optimal value w.

Keywords. Large Deviation Theory, Rare Events, Robustness, Score Function, Sensitivity Analysis, Simulation.

[0]This work was supported by the Technion V.P.R. Fundent charitable trust-non military research fund, and the Center for Economic Research at Tilburg University.

Contents

1 Introduction

Estimation of rare events is important for many modern applied systems, in particular for asynchronous transfer mode multiplexers in broadband integrated switching digital network [7]. It is well known that under the original probability measures, (crude Monte Carlo), estimation of rare events is very time consuming and therefore is extremely costly. Instead, a method based on changing the underlying distribution, called *importance sampling* (IS), to speed up the simulation is typically used [29]. In the past decade, IS has been applied to a variety of problems arising in the analysis of rare events in queueing systems (see e.g., [1], [2]- [14], [16]–[18], [19], [20], [26], [27], [30]-[36], and [38]). The main idea of IS approach is to make the occurrence of rare events more frequent by simulating the system under the new probability measure. Then, in order to obtain an unbiased estimator of the desired rare event, the simulated events are weighted by using the likelihood ratio (also called Radon-Nikodym derivative).

Note that most papers on rare events utilize large deviations (LD) theory and exponential change of measure (ECM) as the main mathematical tools (see Appendix A, Section 8 for the definition of ECM). LD techniques give the optimal change of measure (in the sense of minimal variance) *within the class of all feasible ECM.* We call it the *optimal exponential change of measure* (OECM) (see e.g., [36], [11]) The above raises the following questions: *What will be the computational cost under this 'optimal' distribution? Does expanding the class of admissible simulation distributions produces any more asymptotically efficient solutions?*, (see e.g. [30]). To answer these questions we propose a framework for simulation speedup analysis, which is based on the theory of complexity.In particular we deal with polynomial and exponential time of the so-called *switching regenerative* (SR) estimators for evaluating the steady-state probabilities of excessive backlog in the $GI/GI/1$ queue, and some of its extensions. The idea is to use initially, in each regenerative cycle, a more congested simulation distribution that will lead to quick occurrence of rare events, and then change the simulation distribution, so that the system will be driven back to the regeneration state. Empirical studies of this technique are given, for example, in [26], [20], [29] and [12]. For parallel work on complexity of rare events estinators see Asmussen and Rubinstein [6].

Identification of the OECM for complex queueing models reduces, at best, to a difficult numerical problem (see e.g., [27], [34]). Hence, typically one obtains a computational error (exceptions are some cases where we have explicit solutions, see e.g., [16], [17].) In particular we consider the following problem: *given the simulation distribution generates an exponential time IS algorithm, how high will be the computational cost T*? We show that $T = \exp(1/\epsilon z + o(1/\epsilon))$, where $\epsilon \equiv 1/x$ is a small parameter, while we obtain an analytical expression for the exponential rate z for the switching regenerative estimator. Note that our results agree with [35], in which a

rather general problem of estimating rare events is considered.

While we deal here with the steady-state behavior, we would like to mention a related transient analysis by Sadowsky [30]. He analyzes transient behavior in the $GI/GI/m$ queue and shows that within the class of all possible importance sampling distribution, the OECM has the following *strong asymptotic optimality property*: as the backlog $x \to \infty$, the computational cost (simulation time required to obtain an estimator of prescribed accuracy) T_x grows *less than exponentially fast* (we show that, in fact, it is *polynomial*) in x, and all the other $GI/GI/m$ simulation distributions incur a computational cost that grows at a *strictly positive exponential rate*! There are a few parallel works (see e.g. [23], [9]) but they do not directly deal with queueing systems.

The rest of the paper is organized as follows. In Section 2 we present a framework for complexity analysis of Monte Carlo estimators. Sections 3 and 4 deal with switching regenerative estimators and their properties. Section 5 investigates the "robustnes" of the SR estimator, in the sense that we find how much one can perturb the optimal value in the OECM such that the SR estimator still leads to dramatic variance reduction and might be useful in practice. Section 6 presents a number of extensions. In particular, it considers deterministic service times, batch arrivals, and stationary Markovian interarrival times. We also show that similar complexity results hold for sensitivities (gradients) of probabilities of rare events with respect to parameters of the interarrival and service time distributions. In Section 7 we give supporting numerical examples, a brief discussion for the future research and concluding remarks.. The definition of ECM and proofs of the main results are included in the Appendix.

2 Framework for complexity analysis via importance sampling

Consider the expected performance

$$\alpha = \mathbf{P}(\mathbf{Y} \in A) = \int_{-\infty}^{\infty} I_A(\mathbf{y}) f(\mathbf{y}) \mathrm{d}\mathbf{y} = \mathbb{E}_f[I(\mathbf{Y} \in A)], \tag{1.1}$$

where the expectation is taken with respect to the density $f(\cdot)$, and $I_A(\mathbf{y})$ is the indicator function of the set A, i.e., $I_A(\mathbf{y}) = 1$ if $\mathbf{y} \in A$ and $I_A(\mathbf{y}) = 0$ otherwise.

Let $g(\mathbf{y})$ be a probability measure density function. Assume that $g(\mathbf{y})$ dominates $f(\mathbf{y})$ in the absolutely continuous sense, that is

$$\mathrm{supp}\{f(\mathbf{y})\} \subset \mathrm{supp}\{g(\mathbf{y})\}.$$

Using the density $g(\mathbf{y})$ we can represent α in (1.1) as

$$\alpha = \int I_A(\mathbf{y}) \frac{f(\mathbf{y})}{g(\mathbf{y})} g(\mathbf{y}) \mathrm{d}\mathbf{y} = \mathbb{E}_g[I_A(\mathbf{Y}) W(\mathbf{Y})], \tag{1.2}$$

where $W(\mathbf{y}) = f(\mathbf{y})/g(\mathbf{y})$ is called the likelihood ratio (LR) or the Radon-Nikodym derivative, and the subscript g means that the expectation is taken with respect to the probability density g.

An unbiased estimator of α is given by

$$\overline{\alpha}_N(g) = \frac{1}{N}\sum_{n=1}^{N} I_n W(\mathbf{Y}_n), \tag{1.3}$$

that is, α can be estimated by taking a random sample $\mathbf{Y}_1, ..., \mathbf{Y}_N$ from a density $g(\cdot)$, and then the output ($I_n \equiv I(\mathbf{Y}_n)$) is made unbiased by multiplying by the likelihood ratio W. Sampling from a different density is called a change of measure; the density g is called the *importance sampling* density (or if $g(\mathbf{y}) = \mathrm{d}G(\mathbf{y})/\mathrm{d}\mathbf{y}$, G is called the *importance sampling* distribution); and finally, $\overline{\alpha}_N(g)$ is called the IS estimator (and the associated Monte Carlo algorithm is called the IS algorithm). In the particular case where there is no change of measure ($g = f$), we have $W = 1$, and the IS estimator reduces to the so-called crude Monte Carlo (CMC) estimator:

$$\tilde{\alpha}_N(f) = \frac{1}{N}\sum_{n=1}^{N} I_n,$$

where $\mathbf{Y}_1, ..., \mathbf{Y}_N$ is a random sample from the density $f(\cdot)$

Since essentially any dominating density g can be used for sampling, a natural problem is to find the optimal one, i.e., the density that minimizes the variance of $\overline{\alpha}_N(g)$:

$$\min_g \mathrm{Var}(\overline{\alpha}_N(g)). \tag{1.4}$$

Selecting $g(\mathbf{y}) \equiv g^*(\mathbf{y}) = f(\mathbf{y})/\alpha$ for $\mathbf{y} \in A$ and $g^*(\mathbf{y}) = 0$ otherwise results in $Z_n = I_n f(\mathbf{Y}_n)/g^*(\mathbf{Y}_n) = \alpha$ with probability one. Since the variance of a constant is zero, $g^*(\mathbf{y})$ is the optimal change of measure (and one sample from g^* gives exactly α). The optimal change of measure thus has the interpretation of being simply the original distribution, conditioned on the rare event having occurred. Unfortunately, there are several difficulties with this optimal density g^*. First, it explicitly depends on α, the unknown quantity that we are trying to estimate. If, in fact, α were known, there would be no need to run the simulation experiment at all. Second, even if α or its approximation were known, it might be impractical to sample efficiently from g^*, since typically it cannot be specified in a closed form.

Assume now that $g(\mathbf{y}) = f(\mathbf{y}, \mathbf{v}_0)$, that is, $g(\mathbf{y})$ comes from the same parametric family of distributions as $f(\mathbf{y}) = f(\mathbf{y}, \mathbf{v})$, $\mathbf{v} \in V$ comes. The parameter vector $\mathbf{v}_0$ is called the *reference* parameter. In this case, the likelihood ratio W in (1.2) reduces to

$$W(\mathbf{y}, \mathbf{v}_0) = \frac{f(\mathbf{y}, \mathbf{v})}{f(\mathbf{y}, \mathbf{v}_0)},$$

and instead of the program (1.4) we can consider the following *simpler* one

$$\min_{\mathbf{v}_0} \mathrm{Var}_{\mathbf{v}_0} \left\{ I_A(\mathbf{Y}) W(\mathbf{Y}, \mathbf{v}_0) \right\}, \tag{1.5}$$

which is the same as

$$\min_{\mathbf{v}_0} \mathcal{L}(\mathbf{v}_0) = \min_{\mathbf{v}_0} \mathbb{E}_{\mathbf{v}_0} \left\{ I_A(\mathbf{Y}) W(\mathbf{Y}_i, \mathbf{v}_0)^2 \right\}. \tag{1.6}$$

The optimal solution of this program, say $\mathbf{v}_0{}^*$, is typically not available analytically, since the variance of $I_A(\mathbf{Y})W(\mathbf{Y}, \mathbf{v}_0)$ is so. To overcome this difficulty we can instead minimize its so-called *stochastic counterpart*, namely

$$\min_{\mathbf{v}_0} \bar{\mathcal{L}}(\mathbf{v}_0) = \min_{\mathbf{v}_0} \left[N^{-1} \sum_{i=1}^{N} \left\{ I_A(\mathbf{Y}_i) W(\mathbf{Y}_i, \mathbf{v}_0)^2 \right\} \right], \tag{1.7}$$

and then take its optimal solution, say $\bar{\mathbf{v}}_{0N}^*$, as an estimator of $\mathbf{v}_0{}^*$. This yields the following algorithm.

Algorithm 2.1 :

1. *Estimate the optimal vector of reference parameters* $\mathbf{v}_0^*$ *of the program (1.5) from the solution of the stochastic counterpart (1.7).*

2. *Estimate the rare probability* α *by using estimator (1.3) with* $g(\mathbf{y}) = g(\mathbf{y}, \mathbf{v}_0^*)$.

The solution of the stochastic counterpart (1.7), however, might be time consuming, especially when the system is complex and the dimensionality of the reference parameter vector $\mathbf{v}_0$ is high. In some cases, large deviations theory (e.g., [27], [16]–[18]) provides simple algorithms for identification of the asymptotically optimal probability measure (density $g(\mathbf{y}, \mathbf{v}_0^*)$) within the parametric family of the exponential changes of measure (see Section 4 for some details). Its application, however, is limited. For example, Frater et al. [17] noted that for Jackson-type open networks the identification of the optimal exponential change of measure reduces to the solution of a complex minimax problem.

We now introduce a framework for *complexity analysis* which is related to the basic concepts of complexity theory (see e.g., [37]).

Let $\ell(x) \equiv \mathbb{E}\varphi(L, x) > 0$, where L is a sample performance, x is a deterministic parameter and ϕ is a given function, say $\varphi(L, x) = I_{\{L > x\}}$. Let $\bar{\ell}_N$ be an IS estimator of $\ell(x)$.

Definition 2.1. We say that $\bar{\ell}_N$ is an $(\epsilon,\ \delta)$-accurate estimator of $\ell(x)$ $(0 < \epsilon,\ \delta < 1)$ if

$$\mathbf{P}(\mid \bar{\ell}_N - \ell(x) \mid < \epsilon \ell(x)) > 1 - \delta. \tag{1.8}$$

□

For example, (0.05, 0.10)-accurate estimator ensures that the relative error does not exceed 5% with probability more than 90%.

Consider the *squared coefficient of variation* (SCV) of $\bar{\ell}_N$, that is

$$\kappa(x) = \frac{N\mathrm{Var}(\bar{\ell}_N)}{\ell^2(x)}.$$

By the Central Limit Theorem (CLT) we have that

$$N \approx \gamma\kappa(x),$$

where $\gamma = \Phi^{-1}(1-\delta/2)^2\epsilon^{-2}$.

Definition 2.2 An IS estimator is called $(\epsilon,\ \delta)$- *polynomial*, if (1.8) is guaranteed by a sample size (computational cost) $N = O(p(x))$ for some polynomial function $p(\cdot)$. Any IS estimator whose computational cost $N \equiv N(x)$ can not be bounded by a polynomial function is called an *exponential time* estimator. □

It follows that in order for the estimator $\bar{\ell}_N$ to be be polynomial time, it suffices that the SCV $\kappa(x)$ be bounded in x by a polynomial function $p(x)$.

For better insight into polynomial and exponential time IS estimators, consider the following simple example.

Example 2.1 Suppose we are interested in estimating $\ell = P(Y > x)$, where the random variable Y has an exponential distribution with rate v, i.e. $Y \sim f(y) = v\exp(-vy)$.

Taking into account that in this case $\ell = e^{-vx}$, it is readily seen that the SCV for the CMC estimator is

$$\kappa^2(x) \approx e^{vx},$$

provided x is large, hence the CMC estimator is exponential in x.

Let $g(y) = v_0\exp(-v_0y)$ be the IS density, and assume that we want to choose v_0 so as to minimize the variance of the LR estimator $\bar{\ell}_N(v_0)$.

The second moment of the random variable (rv) IW is

$$\mathbb{E}_g\{IW\}^* = \mathcal{L}(v_0) = \frac{v^2}{v_0}\int_{y=x}^{\infty} e^{(v_0-2v)y}\mathrm{d}y = \frac{v^2e^{-(2v-v_0)x}}{v_0(2v-v_0)}, \tag{1.9}$$

which is infinite for $v_0 \geq 2v$.

The optimal value of the reference parameter $v_0^* = v_0^*(x)$, which minimizes $\mathcal{L}(v_0)$, is

$$v_0^*(x) = v + x^{-1} - (v^2 + x^{-2})^{1/2}.$$

Suppose $x^{-1} << v$, hence $P(Y > x)$ is small, say less than 10^{-6}. In this case we have

$$v_0^*(x) \approx x^{-1} \tag{1.10}$$

and $\mathbb{E}_g[I_{\{Y>x\}}] \approx e^{-1}$.

Consider the relative efficiency $\epsilon(v_0, x)$ defined as

$$\epsilon(v_0, x) = \frac{\mathrm{Var}\{\bar{\ell}_N(v_0)\}}{\mathrm{Var}\{\tilde{\ell}_N\}}.$$

For $v_0 = v_0^*$ we obtain

$$\epsilon^* = \epsilon(v_0^*, x) \approx 0.5 x v e^{1-vx}.$$

If we take, for example, $v = 1$ and $x = 12$ then

$$\mathbb{E} I_{[12,\infty)}(L) = P(L > 12) = e^{-12} \approx 10^{-6},$$

$v_0^*(x) \approx 1/12$ and $\epsilon^* \approx 10^{-4}$. Thus, using the optimal value $v_0^* \approx 1/12$ we obtain a dramatic variance reduction, namely of the order of 10^4.

Consider now the SCV for the LR estimator $\bar{\ell}_N(v_0)$. It is not difficult to see that

$$\kappa^2(v_0, x) = \frac{v^2 e^{v_0 x}}{v_0(2v - v_0)} - 1.$$

For $v_0^* = x^{-1}$ it reduces to

$$\kappa^2(v_0^*, x) \approx 0.5 x v e. \tag{1.11}$$

That is, for large x, the SCV of the CMC and the optimal LR estimator increase in x exponentially and linearly, respectively ($\kappa^2(x) \approx e^{vx}$ and $\kappa^2(v_0^*, x) \approx 0.5 x v e$, respectively). In other words, the CMC and IS estimators can be viewed as *exponential* and *polynomial* time (the required sample size (computational cost) N is $N = O(e^x)$ and $N = O(x)$, respectively).

It is not difficult to see that if we choose $v_0 = k v_0^*$ instead of the optimal value v_0^*, then for large x we obtain

$$\kappa^2(v_0, x) = 0.5 k^{-1} x v e^k.$$

For example, if $k = 2$ (that is, $v_0 = 2v_0^* \approx 2/x$) we obtain

$$\kappa^2(v_0, x) \approx 0.25 x v e^2.$$

In this case the relative efficiency equals $\epsilon(v_0, x) \approx 0.5 e \epsilon^*$. So, perturbing v_0^* ($k = 2$) by 100% increases the variance approximately only $0.5e$ times.

Table 2.1 displays the relative efficiency $\epsilon(v_0, x)$ as function of v_0 for $v = 1$ and $x = 20$; results of table 2.1 are self explanatory.

Table 2.1 The relative efficiency $\epsilon(v_0, x)$ of the IS estimator as a function of v_0, given $v = 1$ and $x = 20$.

v_0	0.6	0.4	0.2	0.1	0.075	0.05*	0.025
$\epsilon(v_0, x)$	$4.0{\cdot}10^{-4}$	$9.6{\cdot}10^{-6}$	$3.1{\cdot}10^{-7}$	$7.8{\cdot}10^{-8}$	$6.2{\cdot}10^{-8}$	$5.5{\cdot}10^{-8}$*	$6.7{\cdot}10^{-8}$

3 Standard and switching regenerative estimators for rare events

This section deals with the so-called *switching regenerative* (SR) estimators for the evalution of p_x, the steady-state probability of excessive backlog, in a stable $GI/GI/1$ queue . Before introducing the SR estimators (§3.2), we need to cite some material on standard regenerative likelihood ratio estimators or, simply, on *standard regenerative* estimators.

3.1 Standard regenerative estimators

The steady-state probability of excessive backlog can be written as

$$p_x = \mathbf{P}_\pi(Q_n > x) = \frac{\mathbb{E}_G \sum_{n=1}^{\tau} I_{\{Q_n > x\}} W_n}{\mathbb{E}_G \sum_{n=1}^{\tau} W_n}, \tag{3.1}$$

where Q_n is the number of customers in the queue just before the n-th customer arrives at the system, I_A is the indicator function of the event A, τ is the length of a regenerative cycle, and $W_n (n = 1, \ldots, \tau)$ is the *likelihood ratio process* defined as

$$W_n = \prod_{k=1}^{n} \frac{F_A(\mathrm{d}A_k)}{G_A(\mathrm{d}A_k)} \prod_{k=1}^{D(n)} \frac{F_B(\mathrm{d}B_k)}{G_B(\mathrm{d}B_k)} \quad , \tag{3.2}$$

where $\{A_k\}$ and $\{B_k\}$ are the sequences of interarrival and service times with distributions $F_A(\cdot)$ and $F_B(\cdot)$, respectively; $G(\cdot)$ dominates the distribution $F(\cdot)$ in the absolutely continuous sense (see e.g. [29]), G is called the *importance sampling* (IS) distribution, or sometimes the *domimating* distribution (see [5]); $D(n)$ is the number of customers being served just before the n^{th} customer arrives at the queue ($D(n) < n$). Note that in order for the $GI/G/1$ queue to remain stable under the probability measure $G = (G_A, G_B)$, the inequality $\mathbb{E}_{G_A} A_k < \mathbb{E}_{G_B} B_k$ must hold. An unbiased estimator of p_x is

$$\overline{p}_{x,N} = \frac{\sum_{i=1}^{N} \sum_{n=1}^{\tau_i} I_{\{Q_{ni} > x\}} W_{ni}}{\sum_{i=1}^{N} \sum_{n=1}^{\tau_i} W_{ni}}, \tag{3.3}$$

where τ_i is the length of the i-th regenerative cycle, and N is the number of generated regenerative cycles.

Under the assumption that G_A and G_B come from a parametric family of distributions as $F_A = F_A(\mathbf{y}, \mathbf{v}_1)$ and $F_B = F_B(\mathbf{y}, \mathbf{v}_2)$ come, Asmussen, Rubinstein and Wang [5] explicitly calculated the variances of the estimator $\overline{p}_{x,N}$ for the steady-state waiting time in the $M/M/1$ queue. They showed that in order to obtain variance reduction with $\overline{p}_{x,N}$ relative to the crude Monte Carlo (CMC) method, one has to choose G_A and G_B such that the associated traffic intensity ρ_0 (under (G_A, G_B)) is moderately larger than the original traffic intensity ρ (under (F_A, F_B)). Asmussen and Rubinstein

[6] proved that the LR estimator $\overline{p}_{x,N}$ is of *exponential time*. Notice that for regenerative estimators we assume that the computational costs T is not $N = N(x)$, (see Definition 2.2) but

$$T \equiv N\mathbb{E}\tau \ ,$$

where N is the number of required cycles and τ is the cycle length. In this way, we measure the computational cost in terms of *time per customer rather than time per cycle*.

3.2 Switching regenerative (SR) estimators

A natural generalization of the LR estimator $\overline{p}_{x,N}$ in (3.3) uses *dynamic* IS distributions G_k at each step k of the cycle instead of a fixed G. In other words, let $G_{A,k}(\cdot)$ and $G_{B,k}(\cdot)$ $(k = 1, \ldots, \tau)$ be the IS distributions used for the k-th interarrival and k-th service times, respectively. That is, assume that system behavior during a cycle of length τ is driven by the set of distribution functions $\widetilde{\pi} = [(G_{A,1}, G_{B,1}), \ldots, (G_{A,\tau}, G_{B,\tau})]$, which we call the *set of IS policies*. Assume next that a policy $\widetilde{\pi}$ of choosing the probability measures during a cycle is completely defined by the system's evolution up to the last arrival prior to this moment. Then one can see that the realizations of the policies will be identical, provided the sample paths

$$[(Q_1, \widetilde{B}_1), \ldots, (Q_\tau, \widetilde{B}_\tau)], \tag{3.4}$$

are the same, where $\widetilde{B}_i$ is the remaining service time for the customer served at the moment the i-th customer arrives at the system during the cycle. Notice that we might also allow dependence of the current service and interarrival time distributions on the index of the last arriving customer; in particular, notice its relationship with a finite family of stopping times $\Xi = \{\xi_r\}$, where

$$\xi_k = \min[t,\ 0 \le t \le \tau \mid Q_t = k]. \tag{3.5}$$

(The above means that our policies $\widetilde{\pi}$ cover both the deterministic and random policies introduced in [29].) With this in hand we can extend W_n in (3.2) to

$$W_n \equiv \prod_{k=1}^{n} \frac{F_A(\mathrm{d}A_k)}{G_{A,k}(\mathrm{d}A_k)} \prod_{k=1}^{D(n)} \frac{F_B(\mathrm{d}B_k)}{G_{B,k}(\mathrm{d}B_k)}, \tag{3.6}$$

where G is redefined as

$$G \equiv \widetilde{\pi}. \tag{3.7}$$

The algorithm for estimating p_x acording to (3.3), where W_{ni} equals

$$W_{ni} = \prod_{k=1}^{n} \frac{F_A(\mathrm{d}A_{ki})}{G_{A,k,i}(\mathrm{d}A_{ki})} \prod_{k=1}^{D(n)} \frac{F_B(\mathrm{d}B_k)}{G_{B,k,i}(\mathrm{d}B_{ki})} \quad , \tag{3.8}$$

and τ_i is the length of i-th cycle and $i,\ 1 \le i \le N$, can be written as follows.

Algorithm 3.1 :

1. *Specify a policy* $\widetilde{\pi}$ *by choosing the IS distributions* $(G_{A,1}, G_{B,1}), \ldots,$ $(G_{A,\tau}, G_{B,\tau})$ *at each cycle, and generate* N *corresponding regenerative cycles.*

2. *Let*

$$\widetilde{\pi}_i = [(G_{A,1,i}, G_{B,1,i}), \ldots, (G_{A,\tau_i,i}, G_{B,\tau_i,i})].$$

be a realization of the policy $\widetilde{\pi}$ *at the* i*-th cycle. Then, compute* X_i *and* Y_i *as follows:*

$$X_i = \sum_{n=1}^{\tau_i} I_{\{Q_{ni} > x\}} W_{ni},$$

$$Y_i = \sum_{n=1}^{\tau_i} W_{ni},$$

where W_{ni} *was given in (3.8).*

3. *Calculate the point estimator* $\overline{p}_{x,N}$ *(see (3.3)) as*

$$\overline{p}_{x,N} = \frac{\overline{X}}{\overline{Y}}$$

and a $100(1-\delta)\%$ *confidence interval for* p_x

$$\widetilde{I} = [\overline{p} \pm \frac{z_\delta s}{\overline{Y} N^{1/2}}],$$

where $\overline{X} = (1/N)\sum_{i=1}^{N} X_i$ *and* $\overline{Y} = (1/N)\sum_{i=1}^{N} Y_i$,

$$\begin{aligned} s^2 &= s_{11} - 2\overline{p}s_{12} + \overline{p}^2 s_{22}, \\ s_{11} &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2, \\ s_{22} &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2, \\ s_{12} &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y}). \end{aligned}$$

We shall call the estimator (3.3), (3.8) the *switching regenerative* (SR) estimator.

4 Properties of the switching regenerative estimator

In this section we characterize a subclass $\Pi^{**} \subset \Pi$ of IS policies, which gives rise to the polynomial time SR estimators of the type (3.3), (3.8) (see Theorem 4.1). We start with the following definition

Definition 4.1. We say that a policy $\widetilde{\pi}$ belongs to the class Π if the following conditions are satisfied:

1. For all $\widetilde{\pi} = [(G_{A,k}, G_{B,k}),\ k = 1, \ldots, \tau] \in \Pi$, we have that $G_{A,k} \in \mathcal{G}_A$ and $G_{B,k} \in \mathcal{G}_B$, where $\mathcal{G}_A$ and $\mathcal{G}_B$ are sets of the dominating distribution functions for F_A and F_B, respectively; $0 < \mathbb{E}_{G_{A,k}} A^2 < \infty$, $0 < \mathbb{E}_{G_{B,k}} B^2 < \infty$, $k = 1, ..., \tau$.

2. $G_{B,k}(\cdot) \equiv G_{B,j}(\cdot)$ if $R(k) = R(j)$, that is, we allow a change of probability measure only at the arrival epochs k, $1 \le k \le \tau$. (Here $R(k)$ is the number of the last arriving customer just before the k-th customer starts service). Moreover, we assume that the probability measure between the k-th and $k+1$ arrivals, $G_{A,k}$ and $G_{B,j}$, $j = D(k)+1, \ldots, D(k+1)$, are completely defined by k, $\mathcal{F}_k$ and $\Xi = \{\xi_r,\ r = 1, \ldots, r_0\}$, where $\mathcal{F}_k = \mathcal{F}_k(Q_k, \widetilde{B}_k)$ is the σ-algebra; and ξ_k is a finite set Ξ of stopping times. (See (3.4) and (3.5) for the definition of $\widetilde{B}_k$ and ξ_k, respectively.)

3. $\mathbb{E}_{G_{A,k}} A_k - \mathbb{E}_{G_{B,j}} B_j < 0,\ D(k) < j < D(k+1),\ 1 \le k \le \xi_1 \equiv \xi$; that is, the simulating system is *unstable* until the occurence of the first overflow, and $\mathbf{P}_{\widetilde{\pi}}(\xi < \infty) = 1$ for all x.

4. $\mathbb{E}_{G_{A,k}} A_k - \mathbb{E}_{G_{B,j}} B_j > 0,\ D(k) < j < D(k+1),\ k \in [\xi, \ldots, \tau] \backslash I_3$, where $I_3 \subset [\xi, \ldots, \tau]$ is a finite set; that is, after the first overflow till the end of the cycle, the system becomes *stable* again. excluding, perhaps, a finite number of steps, $\mathbf{P}(\tau - \xi < \infty; \xi < \tau) = 1$ and $\lim_{x\to\infty} P(Q_t = \psi x) = 0$, for $1 < \psi < \infty$.

5. All parameters of $G_{A,k}$ and $G_{B,k}$ are independent of x.

Definition 4.2. Let $\pi^* = [(G^*_{A,1}, G^*_{B,1}), \ldots, (G^*_{A,\tau}, G^*_{B,\tau})] \in \Pi$ be a policy with

$$G^*_{A,k} = \begin{cases} F_A^{(w)}, & 1 \le k < \xi, \\ F_A, & \xi \le k \le \tau, \end{cases} \tag{4.1}$$

and

$$G^*_{B,k} = \begin{cases} F_B^{(-w)}, & 1 \le R(k) < \xi, \\ F_B, & \xi \le R(k) \le \tau, \end{cases} \tag{4.2}$$

respectively, where $F_A^{(w)}$ and $F_B^{(-w)}$ are the *optimal exponential change of the measures* (OECM) for the interarrival and service time distributions, defined as

$$F_A^{(w)}(\mathrm{d}x) = \exp(-wx - \Lambda_A(w))F_A(\mathrm{d}x) \tag{4.3}$$

and

$$F_B^{(-w)}(\mathrm{d}x) = \exp(wx - \Lambda_B(-w))F_B(\mathrm{d}x), \tag{4.4}$$

respectively; in (4.3) $\Lambda_X(w) = \log(\mathbb{E}e^{-wX})$ denotes the cumulant function, w is a unique positive solution of the equation

$$\Lambda(w) \equiv \Lambda_A(w) + \Lambda_B(-w) = 0$$

(for more details see Appendix A, Section 8); and as before, $R(k)$ is the number of the customers arriving at the queue just before the k-th customer starts its service (clearly $R(k) > \xi$ implies $k > \xi - x$).

As examples of $F_X^{(t)}$ for $t \in \{t > 0,\ \Lambda_X(t) < \infty\}$ let :

(a) $Y \sim \exp(v)$, then $F^{(t)} = \exp(v + t)$.

(b) $Y \sim \mathrm{Gamma}(\lambda, \beta)$, then $F^{(t)} = \mathrm{Gamma}(\lambda + t, \beta)$.

(c) $Y \sim N(\mu, \sigma^2)$, then $F^{(t)} = N(\mu - 2\sigma^2 t, \sigma^2)$.

(d) $Y \sim Geometric(p)$, then $F^{(t)} = Geometric(1 - (1-p)e^{-t})$.

Note that under the policy π^* the LR process W_{ni} in (3.8) reduces (see also formula (4.5.6) of Rubinstein and Shapiro [29]) to

$$W_{ni} = \begin{cases} \prod_{k=1}^{n} \frac{F_A(\mathrm{d}A_{ki})}{G^*_{A,k}(\mathrm{d}A_{ki})} \prod_{k=1}^{D(n)} \frac{F_B(\mathrm{d}B_k)}{G^*_{B,k}(\mathrm{d}B_{ki})}, & 1 \le n \le \xi \\ \prod_{k=1}^{\xi} \frac{F_A(\mathrm{d}A_{ki})}{G^*_{A,k}(\mathrm{d}A_{ki})} \prod_{k=1}^{D(\xi)} \frac{F_B(\mathrm{d}B_k)}{G^*_{B,k}(\mathrm{d}B_{ki})}, & \xi \le n \le \tau\ . \end{cases} \tag{4.5}$$

Formulas (4.1), (4.2), (4.5) mean that from the beginning of the cycle until the first overflow we use the OECM pair $(F_A^{(w)}, F_B^{(-w)})$, and then accomplish the cycle by switching to the original distributions.

Definition 4.3. Let Π^* be a class of policies $\widetilde{\pi} \in \Pi$ such that

$$G_{A,k} = F_A^{(w)},\ k \in [1, \ldots, \xi_x] \backslash I_1,\ \text{(a.s.)}$$

and

$$G_{B,k} = F_B^{(-w)},\ R(k) \in [1, \ldots, \xi_x] \backslash I_2,\ \text{(a.s.)}$$

where $I_1 \subset [1, \ldots, \xi_x]$ and $I_2 \subset [1, \ldots, D(\xi_x)]$ are sets with $o(x)$ elements.

Theorem 4.1 *Assume that the following conditions hold:*

Condition A1: $0 < \mathbb{E}B < \mathbb{E}A < \infty$ *and* $\mathbf{P}(B - A > 0) > 0$;

Condition A2: *There exists a scalar* w *satisfying*

$$\Lambda(w) \equiv \Lambda_A(w) + \Lambda_B(-w) = 0, \tag{4.6}$$

such that $\Lambda_A(w) < \infty$, $\Lambda_B(-w) < \infty$, *and* $0 < \Lambda'(w) < \infty$ *(see also Appendix A, Section 8). Then*

$$\pi^* \in \Pi^{**}, \tag{4.7}$$

$$\Pi^{**} \subset \Pi^*, \tag{4.8}$$

where $\Pi^{**} \subset \Pi$ *is the subclass of* Π *which generates the polynomial time IS estimators.* □

Proof. The proof of Theorem 4.1 is given in Appendix B (Section 8). □

Formulas (4.7) and (4.8) of Theorem 4.1 can be interpreted as follows. Formula (4.7) states that the estimator (3.3), (4.5) based on policy π^* is polynomial, while formula (4.8) states that for Algorithm 3.1 to be polynomial it is necessary that $\widetilde{\pi} \in \Pi^*$, that is, it is necessary to use the OECM until the first overflow ξ excluding, perhaps, $o(x)$ steps.

It is our belief that Theorem 4.1 is the first result characterizing the complexity of switching regenerative estimators and specifying a subclass $\Pi^{**} \in \Pi$ that generates polynomial time estimators.

We shall call the parameter w in (4.1), (4.2), (4.3), (4.4) satisfying (4.6) the *optimal parameter* (OP) in the OECM's $F_A^{(w)}$ and $F_B^{(-w)}$, respectively.

Assume that the original interarrival and service time pdfs in the $GI/G/1$ queue come from the following exponential family

$$f(x, \mathbf{v}) = a(\mathbf{v}) \exp(b(\mathbf{v})c(x))d(x), \tag{4.9}$$

where $c(x)$ is bounded by a polynomial function and $a(\mathbf{v})$ and $b(\mathbf{v})$ are twice continuously differentiable functions with the parameter vector $\mathbf{v}$. Taking into account (4.3), (4.4) and the fact that the OP w is a scalar, it follows that the OECM $(F_A^{(w)} = F_A^{(w)}(\mathbf{v}_{01}^*), F_B^{(-w)} = F_B^{(w)}(\mathbf{v}_{02}^*))$ coincides with the *original distributions* $(F_A(\mathbf{v}_1),\ F_B(\mathbf{v}_2))$, up to a *single* parameter; that is, *only a single* parameter, in the vector $\mathbf{v}_{01}^*$ of the OECM $F_A^{(w)}(\mathbf{v}_{01}^*)$, differs from the vector $\mathbf{v}_1$ of the original interarrival time distribution $F_A(\mathbf{v}_1)$, and similar for the service time distribution.

Table 4.1 presents the optimal parameter w and the optimal reference parameter vector $\mathbf{v}_0^* = (\mathbf{v}_{01}^*, \mathbf{v}_{02}^*)$ for several commonly used exponential families.

Table 4.1 The optimal parameter w and the optimal reference parameter vector $\mathbf{v}_0^* = (\mathbf{v}_{01}^*, \mathbf{v}_{02}^*)$ for several commonly used exponential families.

$F_A,\ F_B$	$\mathbf{v}_1,\ \mathbf{v}_2$	$\mathbf{v}_{01}^*,\ \mathbf{v}_{02}^*$	w
$\exp(\cdot)$	$\lambda,\ \mu$	$\mu,\ \lambda$	$\mu - \lambda$
$\text{Gamma}(\cdot,\cdot)$	$(\lambda,\beta),\ (\mu,\gamma)$	$(\lambda+w,\beta),\ (\mu-w,\gamma)$	$w>0:\ (\lambda+w)^\beta(\mu-w)^\gamma = \lambda^\beta\mu^\gamma$
$\text{Poisson}(\cdot)$	$\lambda,\ \mu$	$\mu,\ \lambda$	$\log(\lambda\mu)$

5 Robustness of switching estimators

In this section we address the following robustness question: how much can one perturb the optimal parameter w in the OECM such that the SR estimators (3.3), (4.5) still obtain low variance. To answer this question we argue as follows. We first replace $F_A^{(w)}$ and $F_B^{(-w)}$ in (4.3), (4.4) by G'_A and G'_B, respectively, assuming that $(G'_A, G'_B) \neq (F_A^{(w)}, F_B^{(-w)})$.

Definition 4.4. Let Π' be a subclass of the IS policies,

$$\widetilde{\pi} = [(G_{A,1}, G_{B,1}), \ldots, (G_{A,\tau}, G_{B,\tau})] \in \Pi',$$

where

$$G_{A,k} = \begin{cases} G'_A & 1 \le k < \xi_x, \\ F_A, & \xi_x \le k \le \tau, \end{cases}$$

and

$$G_{B,k} = \begin{cases} G'_B, & 1 \le R(k) < \xi_x, \\ F_B, & \xi_x \le R(k) \le \tau. \end{cases}$$

It follows from Theorem 5.1 that in this case the estimator (3.3), (4.5) has exponential time complexity. Clearly, when $G'_A = F_A^{(w)}$ and $G'_B = F_B^{(-w)}$, we have $\widetilde{\pi} \equiv \pi^*$ and, thus a polynomial time estimator (3.3), (4.5) with the exponential rate z of the computational cost $T_{\widetilde{\pi}}$ ($T_{\widetilde{\pi}} = \exp(z_{\widetilde{\pi}} x + o(x))$) equal 0.

Next we derive explicit expressions for the exponential rate z of the computational cost $T_{\widetilde{\pi}}$ under $(G'_A, G'_B) \neq (F_A^{(w)}, F_B^{(-w)})$. This z will characterize the robustness of our IS estimator. In addition, this exponential rate z might give useful insight into time complexity of IS estimators for more general queueing systems. We define the following new (but not probabilistic!) finite measures:

$$K_A(\mathrm{d}A) \equiv K_A(G_A)(\mathrm{d}A) \equiv \frac{F_A}{G_A}(\mathrm{d}A) F_A(\mathrm{d}A) \tag{5.1}$$

and

$$K_B(\mathrm{d}B) \equiv K_B(G_B)(\mathrm{d}B) \equiv \frac{F_B}{G_B}(\mathrm{d}B)F_B(\mathrm{d}B). \tag{5.2}$$

We also define

$$\Lambda_{K_A}(u) = \log \mathbb{I}_{K_A} e^{-uA_k} \equiv \log \int_0^\infty e^{-uA_k} K_A(\mathrm{d}A_k)$$

and

$$\Lambda_{K_B}(u) = \log \mathbb{I}_{K_B} e^{-uB_k} \equiv \log \int_0^\infty e^{-uB_k} K_A(\mathrm{d}B_k).$$

Note that the symbol $\mathbb{I}$ is used here instead of the conventional symbol $\mathbb{E}$ (for expectation); the functions $\Lambda_{K_A}(u)$ and $\Lambda_{K_B}(u)$ have a meaning similar to the cumulants Λ_A and Λ_B, respectively.

We finally define $v \equiv \sup[u \mid \Lambda_K(u) \le 0]$, where

$$\Lambda_K(u) = \Lambda_{K_A}(u) + \Lambda_{K_B}(-u),$$

and we assume that the following regularity condition holds.

Condition A3. There exists a value u such that $\Lambda_K(u) \le 0$.

Note that since $\Lambda(u) \to \infty$ as $u \to \infty$, condition **A3** directly implies that there exists a unique value v such that

$$\Lambda_K(v) = 0 \quad \text{and} \quad 0 < \Lambda'(v) < \infty. \tag{5.3}$$

Theorem 5.1 *Assume that conditions* **A1**–**A3** *hold. Then for any policy* $\widetilde{\pi} \in \Pi'$ *the exponential rate* z *equals*

$$z_{\widetilde{\pi}} \equiv \lim_{x \to \infty} x^{-1} \log T_{\widetilde{\pi}} = \Lambda_{K_A}(v) - 2\Lambda_A(w). \tag{5.4}$$

□

Proof. The proof of Theorem 5.1 is given in Appendix B (Section 8). □

Corollary 5.1 Assume that conditions **A1**–**A3** hold. Let $\widetilde{\pi}_t = \widetilde{\pi}(G'_A, G'_B) \in \Pi'$ with $G'_A = F_A^{(t)}$ and $G'_B = F_B^{(-t)}$. Then

$$z_{\widetilde{\pi}_t} = \Lambda_A(t) + \Lambda_A(v - t) - 2\Lambda_A(w). \tag{5.5}$$

□

Proof. Corollary 5.1 directly follows from (5.4) by substituting $G'_A = F_A^{(t)}$ and $G'_B = F_B^{(-t)}$. □

Using (5.5), we obtain various efficiency characteristics of the SR estimator (3.3), (4.5), namely the computational cost $T_{\widetilde{\pi}_t}$, the exponential rate

$z_{\widetilde{\pi}}$, the optimal speedup factor $S_{opt} = T_F/T_{\pi^*}$ $(z_{\pi^*} = 0)$, and the speedup factor $S_t = T_F/T_{\widetilde{\pi}_t}$.

Table 5.1 presents (analytic) efficiency values of the SR estimator (3.3), (4.5) under both the optimal parameter w and the perturbed parameter t, as functions of the traffic intensity ρ, for the probabilities of excessive backlog $p_x = 10^{-15}$ in the $M/M/1$ queue with the service rate $\mu = 1$. In particular (under the OECM $(F_A^{(w)}, F_B^{(-w)}$, $w = \mu - \lambda)$ it represents S_{opt}, called the *optimal speedup factor*, and (under the ECM $(F_A^{(t)}, F_B^{(-t)}$) it represents S_t and $z_{\widetilde{\pi}_t}$, called the *speedup factor* $S_t = T_F/T_{\widetilde{\pi}_t}$ and the *exponential rate*. Notice that both S_t and $z_{\widetilde{\pi}_t}$ were calculated for t being *less* than the optimal value w by 20%. Notice also that the values x in the last row of the table correspond to $p_x = 10^{-15}$.

It follows from the table that the OECM leads to dramatic variance reduction (11–13 orders of speedup out of 15 orders), and that perturbing w by 20% $(t < w)$, we *lose only 2–3 orders of speedup.*

Table 5.1 The efficiencies S_{opt}, S_t and $z_{\widetilde{\pi}_t}$ of the SR estimator (3.3), (4.5) as functions of the traffic intensity ρ for the probability of excessive backlog $p_x = 10^{-15}$ in the $M/M/1$ queue with $\mu = 1$.

ρ	0.2	0.4	0.6	0.8
w	0.8	0.6	0.4	0.2
S_{opt}	10^{13}	10^{12}	10^{12}	10^{11}
t	0.64	0.48	0.32	0.16
S_t	10^{11}	10^{10}	10^{9}	10^{8}
z	$1.0{\cdot}10^{-1}$	$5.6{\cdot}10^{-2}$	$3.1{\cdot}10^{-2}$	$1.3{\cdot}10^{-2}$
x	21	36	62	107

6 Extensions

Here we present some extensions of the results of Sections 4 and 5. In particular we consider estimation of:

(i) rare events in the $GI/D/1$ queue (§6.1),

(ii) rare events in the $GI/G/1$ queue with dependent arrivals (§6.2),

(iii) rare events in the $GI/G/1$ queue with batch arrivals (§6.3),

(iv) sensitivities (derivatives) of the probability of the excessive backlog $p_x(\mathbf{v})$ with respect to the parameter vector $\mathbf{v}$ of the interarrival time distributions (§6.4).

6.1 The $GI/D/1$ queue

By definition, this queue has deterministic constant service times, independent and identically distributed interarrivals with the pdf $f_A(A) \equiv \mathrm{d}F_A(A)/\mathrm{d}A$. Presented in [18] is a heuristic result on quick simulation of rare events in the $M/D/1$ queue and its extensions to batch arrivals with exponential distribution. The problem here is that the deterministic service time excludes the possibility of introducing a corresponding change of measure with the concomitant LR process W_n in the standard way. To circumvent this problem, we use the so-called *"push out"* (PO) technique introduced in [28]. It is shown in [28] how to incorporate the PO technique for the $GI/D/1$ queue in the standard regenerative setting, in order to reduce the original problem to an equivalent associated one. In short, one merely needs to "push out" the deterministic service time parameter, say d, to the interarrival time random variable $Y \sim f_A(y)$, then introduce a new auxiliary random variable $\tilde{Y} = A - d$ distributed $\tilde{f}_A(y+d)$ and a new auxiliary sample performance $\tilde{Q}_n(\tilde{\mathbf{Y}}_n)$, $\tilde{\mathbf{Y}}_n = (\tilde{Y}_1, ..., \tilde{Y}_n)$, (instead of f_A and $Q_n(\mathbf{A}_n)$, $\mathbf{A}_n = (A_1, ..., A_n)$, respectively) and incorporate them into the following asymptotical consistent regenerative estimator of p_x :

$$\overline{p}_{x,N} = \frac{\sum_{i=1}^{N} \sum_{n=1}^{\tau_i} I_{\{\tilde{Q}_{ni} > x\}} W_{ni}}{\sum_{i=1}^{N} \sum_{n=1}^{\tau_i} W_{ni}} \quad , \tag{6.1}$$

where (similar to 3.3) τ_i is the length of the i-th regenerative cycle, N is the number of regenerative cycles, $\tilde{\mathbf{Y}}_{ni} = (\tilde{Y}_{1i}, ..., \tilde{Y}_{ni})$, $n = 1, ..., \tau_i$; $i = 1, ..., N$; $\tilde{Q}_{ni}(\tilde{\mathbf{Y}}_{ni}) = Q_{ni}(\mathbf{Y}_{ni})$;

$$W_{ni} = W_{ni}(\tilde{\mathbf{Y}}_{ni}) = \prod_{k=1}^{n} \frac{\tilde{f}(\tilde{Y}_{ki})}{\tilde{g}(\tilde{Y}_{ki})} \quad ,$$

$\tilde{\mathbf{Y}}_{ni}$ is a random sample from $\tilde{g}$ which is a dominating pdf for $\tilde{f}$.

In order to adopt the "push out" method, we need to replace the OECM in the SR estimator of the type (3.3), (4.5) by a new OECM defined below, with all other data remaining the same. (Similar results can be obtained for the $D/GI/1$ queue.) It is not difficult to see that under conditions **A1**–**A2** we have

$$w = w_1 \equiv \sup[\theta \geq 0 \mid \Lambda_{\tilde{Y}}(\theta) < 0],$$

where $\Lambda_{\tilde{Y}}(\theta) = \mathbb{E}e^{-\theta \tilde{Y}}$ with $\tilde{Y} \sim \tilde{f}(\tilde{Y}) = f_A(y+d)$. Let

$$k(y) \equiv \frac{\tilde{f}(y)^2}{\tilde{g}(y)} \quad ,$$

where $\tilde{g}(\cdot)$ is a dominating pdf for $\tilde{f}(\cdot)$, and

$$\Lambda_k(u) \equiv \mathbb{I}_k e^{-uy} \equiv \log \int_0^\infty e^{-uy} k(y) \mathrm{d}y.$$

Then the following proposition follows.

Proposition 6.1 *Consider an IS estimator based on the policy* $\widetilde{\pi} \in \Pi'$. *Assume that conditions* **A1**, **A2** *hold, and there exists a value* u *such that* $\Lambda(u) < 0$. *Let* v_1 *be the solution of the equation* $\Lambda_k(u) = 0,\ 0 < \Lambda_k'(u) < \infty$. *Then the exponential part* $z_{\widetilde{\pi}}$ *of the computational cost* $T_{\widetilde{\pi}}$ *is given by*

$$z_{\widetilde{\pi}} = d(2w_1 - v_1).$$

□

Proof The proof of this and the following two propositions (namely propositions 6.2 and 6.3) are omitted because they are similar to the proof of Theorem 5.1. (Their proofs are available from the authors.)

Notice that for the IS pdf $\tilde{g}(y) = g^{(w_1)}(y + d)$, we obtain $z_{\widetilde{\pi}} = 0$ and hence, a polynomial time estimator (6.1).

6.2 Rare events with dependent arrivals

First, we consider the case with interarrival times driven by stationary uniform recurrent Markov chains (see [9]), which admit a strong Perron-Frobeneous theory (see e.g., [8]). Note that the class of such distributions covers the TES (Transform-Expand-Sample) processes (see e.g., [25]) which by itself captures a wide variety of sample path behaviors with autocorrelation input sequences; TES are used for modelling video and data sources.

We now introduce the following notations for the cumulant function and ECM for the interarrival distribution given by stationary uniform recurrent Markov chains:

- $\Lambda_A(\theta) = \log(\lambda_A(\theta))$, where $\lambda_A(\theta)$ is a unique simple real eigenvalue of the kernel $\exp(-\theta A_k)F_A(\mathrm{d}b, a)$ with the following properties: $\lambda_A(\theta)$ is analytic and strictly convex, and the related eigenfunction $\phi(a;\theta)$ is bounded.

- The interarrival time distribution equals

$$F_A^{(\theta)}(a, \mathrm{d}b) = e^{-\theta b - \Lambda_A(\theta)} \frac{\phi(b;\theta)}{\phi(a;\theta)} F(a, \mathrm{d}b). \tag{6.2}$$

Consider now a single-server queue with iid service times, but with interarrival time distributions given by a uniform recurrent stationary (time-homogeneous) Markov Chain. We call such a queue the $GM/GI/1$ queue. In this case we define a *cycle* differently from the standard regenerative cycle: we call a *cycle* a piece of trajectory starting the system at a steady-state regime, provided the system is empty, and terminating the cycle on the first occasion when the system is empty again.

It is important to note that in this case we need to simulate (in parallel) two sample path (trajectories), namely one with a change of measure and

one without a change of measure (CMC). When the second process runs long enough to be treated as in the steady state, we use the instants when the system is empty to start a cycle of the first process by an IS simulation. Note that unlike the standard regenerative cycles, these cycles are not independent. However, the ratio estimator remains asymptotically unbiased, provided the cycles are initialized under a steady-state distribution (see [20]). The associated IS algorithm is the same as Algorithm 3.1.

Proposition 6.2 *Consider a $GM/GI/1$ queue. Let*

$$K_A(a, db) \equiv \frac{F_A(a, db)}{G_A(a, db)} F_A(a, db),$$

where $G_A(\cdot,\cdot)$ dominates the distribution $F_A(\cdot,\cdot)$ and $\Lambda_{K_A}(u) \equiv \log(\lambda_{K_A}(u))$. Furthermore, let $\lambda_{K_A}(u)$ be defined as a unique simple real eigenvalue of the kernel $\exp(\theta A_k)\ K_A(dy, x)$ having the following properties: $\lambda_{K_A}(u)$ is analytic and strictly convex and the related eigenfunction $\phi_{K_A}(x;\theta)$ is bounded. Assume, finally, that there exists a unique value u such that $\Lambda_K(u) < 0$. Then, under conditions **A1**–**A2**, *the estimator (3.3), (4.5) based on the IS policy $\widetilde{\pi} \in \Pi'$ with the simulation distributions $G_A(a, db)$ (stationary Markov chain) and $G_B(b)$ possesses the following exponential rate of the computational cost:*

$$z_{\widetilde{\pi}} = \Lambda_{K_A}(v_2) - 2\Lambda_A(w_2)\ , \tag{6.3}$$

where v_2 and w_2 are the solutions of the equations $\Lambda_K(v_2) = 0,\ 0 < \Lambda'_K(v_2) < \infty$, and $\Lambda(w_2) = 0,\ 0 < \Lambda'(w_2) < \infty$, respectively. □

For the particular case $G_A = F_A^{(w_2)}$ and $G_B = F_B^{(-w_2)}$ we have $z_{\widetilde{\pi}} = 0$ and hence a polynomial time IS estimator.

Example 6.1 Consider the case where the interarrival process $\{A_k\}$ is represented by an AR(1) process:

$$A_{k+1} = rA_k + (1 - r)X,$$

$0 \le r \le 1,\ X \sim \exp(\lambda)$. For this case the real eigenevalue and the associated eigenvector are given by

$$\lambda_A(\theta) = \frac{\lambda}{\lambda + \theta}$$

and

$$\phi(b;\theta) = e^{\frac{b\lambda r}{1-r}}\ .$$

Substituting these expressions into (6.2), we see that the ECM (in particular the OECM) for this case is the same as for the independnent interarrivals with distribution $\exp(\lambda)$. In other words, the family of the exponential changes of measure for this case is the family of independent exponential distributions. □

6.3 Rare events in $GI/G/1$ queue with batch arrivals

Let T be a random variable denoting the length of the batch (i.e., T is the number of customers in the batch). Assume that T is distributed according to a discrete distribution, denoted by $F_T(\cdot)$. We call such a queue the $GB/GI/1$ queue. Let us make the following assumptions.

Condition B1: (analogous to **A1**) $0 < \mathbb{E}B/\mathbb{E}T < \mathbb{E}A < \infty$ and $\mathbf{P}(BT - A > 0) > 0$; $\Lambda_T(\nu) \equiv \log \mathbb{E}[\exp(-\nu T)] < \infty$ for all ν.

Condition B2. There exists w_3 such that $\Lambda(w_3) = 0$ and $0 < \Lambda'(w_3) < \infty$, where we define

$$\Lambda(\theta) = \Lambda_A(\theta) + \Lambda_T(-\Lambda_B(-\theta)).$$

Condition B3 Let $K_T(n) = F_T(n)^2/G_T(n)$, $\Lambda_{K_T}(u) = \log \mathbb{I}_{K_T} \exp(-sT) = \log \sum_{n=0}^{\infty} \exp(-sn) K_T(n)$ and $\Lambda_K(\theta) = \Lambda_{K_A}(u) + \Lambda_{K_T}(-\Lambda_{K_B}(-u))$. Then there exists a value v_3 such that $\Lambda_K(v_3) = 0$ and $0 < \Lambda_K(v_3) < \infty$.

Proposition 6.3 *Under the conditions* **B1**–**B3** *the exponential part of the computational cost in the $GB/GI/1$ queue is given by*

$$z_{\widetilde{\pi}} = \Lambda_{K_A}(v_3) - 2\Lambda_A(w_3).$$

□

For the particular case with $G_A = F_A^{(w_3)}$, $G_B = F_B^{(-w_3)}$, and $G_N = F_T^{(\nu)}$ where $\nu = -\Lambda_B^{(-w_3)}$, we have $z_{\widetilde{\pi}} = 0$ and hence a polynomial time IS estimator.

Example 6.2 Consider the $M/M/1$ queue with batch arrivals and assume that the batch length is distributed Poisson; that is, $A \sim \exp(\lambda)$, $B \sim \exp(\mu)$, and $T \sim \text{Poisson}(\gamma)$. In this case the ECM (with the parameter θ) is given by $G_A = F_A^{(\theta)} = \exp(\lambda + \theta)$, $G_B = F_B^{(-\theta)} = \exp(\mu - \theta)$, and $G_T = F_T^{(\nu)} = \text{Poisson}(\gamma(e^{-\nu} - 1))$. With $\nu = -\Lambda_B(-\theta)$ the distribution $F_T^{(\nu)}$ reduces to $\text{Poisson}(\gamma(\mu/(\mu - \theta) - 1))$. By Proposition 6.3 we have a polynomial time IS estimator for $\theta = w_3$, where w_3 satisfies

$$\frac{\lambda}{\lambda + w_3} e^{\gamma(\frac{\mu}{\mu - w_3} - 1)} = 1.$$

□

6.4 Sensitivity analysis

Consider now the sensitivity (gradient) of p_x with respect to the parameter vector $\mathbf{v}$ of the interarrival time pdf $f_A(x, \mathbf{v})$. Assume that $f_A(x, \mathbf{v})$ belongs to the exponential family (4.9). Differentiating $\bar{p}_x$ in (3.3) with respect to $\mathbf{v}$, we obtain the following asymptotically consistent estimator of ∇p_x [29]:

$$\overline{\nabla}_{\mathbf{v}} p_x \equiv \overline{p}_x^{(1)} - \overline{p}_x \overline{p}_x^{(2)} \ , \tag{6.4}$$

where

$$\overline{p}_x = \frac{\sum_{i=1}^{N} \sum_{t=1}^{\tau_i} I_{ti} W_{ti}}{\sum_{t=1}^{\tau} W_{ti}}, \quad \overline{p}^{(1)} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{\tau_i} I_{ti} \boldsymbol{S}_{ti} W_{ti}}{\sum_{t=1}^{\tau} W_{ti}},$$

$$\overline{p}^{(2)} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{\tau_i} \boldsymbol{S}_{ti} W_{ti}}{\sum_{t=1}^{\tau} W_{ti}}, \quad I_{ti} = I_{\{Q_{ti} > x\}},$$

and (see (4.9))

$$\boldsymbol{S}_{ti} = t \frac{\nabla_{\mathbf{v}} a(\mathbf{v})}{a(\mathbf{v})} + \nabla_{\mathbf{v}} b(\mathbf{v}) \sum_{k=1}^{t} c(A_{ki}). \tag{6.5}$$

Note that $\boldsymbol{S}_t$ is called the *score function process* (see e.g., [29]).

Proposition 6.4 *Under the conditions of Theorem 5.1 the computational cost* $T_{\widetilde{\pi}}^{(1)} = \exp(z_{\widetilde{\pi}}^{(1)}(x) + o(x))$ *of the estimator (6.4) coincides with* $T_{\widetilde{\pi}} = \exp(z_{\widetilde{\pi}}(x) + o(x))$ *of the SR estimator (3.3), (4.5).*

□

Proof. In Appendix B (Section 8) we present a sketch of the proof of Proposition 6.4, which is similar to that of Theorem 5.1. □

It follows from Proposition 6.4 that for efficient estimation of both p_x and ∇p_x simultaneously, we can use the same change of measures. In particular, π^* results in polynomial time IS estimator.

7 Numerical examples and concluding remarks

7.1 Numerical examples

Here we present numerical results on the efficiency of the SR estimator (3.3), (4.5) for the $M/M/1$ queue (with the service rate $\mu = 1$) and some of its extensions. In all cases we simulated 10^6 customers and estimated the probability p_x of the excessive backlog x. We run the SR estimator with $\widetilde{\pi} \in \Pi'$ (see Definition 4.4) for the following two cases: with the optimal parameter w, corresponding to the distributions $(G'_A, G'_B) = (F_A^{(w)}, F_B^{(-w)})$ and with the perturbed values of w, denoted by t and corresponding to $(G'_A, G'_B) = (F_A^{(t)}, F_B^{(-t)})$. In all our tables the values in the first column marked with $*$ correspond to w.

Table 7.1 presents the computational cost (simulation time) $\overline{T}_{\widetilde{\pi}}$ required to obtain an $(0.5, 0.5)$-accurate estimators (see Definition 2.1) for; the ratio $R_t = \overline{T}_{\widetilde{\pi}_t}/\overline{T}_{\pi^*} = S_{opt}/S_t$, called the *loss factor*; and the exponential rate z, (approximation of $\exp(zx)$), as functions of t (and the relative perturbation $\delta = (t-w)/w$), while estimating the probability $p_x = 5.34 \cdot 10^{-10}$ $(x = 40)$ in the $M/M/1$ queue with $\rho = 0.6$.

Table 7.1 The efficiency of the RS estimator (3.3), (4.5) as a function of the perturbed parameter t for the $M/M/1$ queue with $\rho = 0.6,\ \mu = 1$, and $p_x = 5.34 \cdot 10^{-10}$ $(x = 40)$.

t	δ	$\overline{T}_{\widetilde{\pi}}$	R_t	$\exp(zx)$	z
0.30	-0.25	$1.20 \cdot 10^6$	4.5	6.1	$4.53 \cdot 10^{-2}$
0.35	-0.125	$5.87 \cdot 10^5$	2.2	1.7	$1.33 \cdot 10^{-2}$
0.40*	0.00	$2.76 \cdot 10^5$	1.0	1.0	0.0
0.43	0.075	$4.27 \cdot 10^5$	1.6	1.3	$1.36 \cdot 10^{-2}$
0.46	0.15	$1.52 \cdot 10^6$	5.5	5.2	$4.13 \cdot 10^{-2}$

Tables 7.2, 7.3 and 7.4 present data similar to those of Table 7.1 for the $M/D/1$ queue with $\rho = 0.3,\ d = 1$, and $p_x = 1.04 \cdot 10^{-9}$ $(x = 20)$; the $MM/M/1$ queue with $\rho = 0.5,\ \mu = 1,\ p_x = 5.34 \cdot 10^{-10}$ $(x = 40)$ and the interarrival sequence being the following $AR(1)$ sequence:

$$A_{k+1} = \theta A_k + (1 - \theta)X,\ X \sim \exp(\lambda),\ \theta = 0.3;$$

and the derivative $\nabla_\lambda p_x$ with respect to the arrival rate λ in the $M/M/1$ queue with $\rho = 0.3,\ \mu = 1,\ p_x = 1.44 \cdot 10^{-16}$ $(x = 30)$, $\nabla_\lambda p_x = 1.42 \cdot 10^{-14}$, respectively.

Table 7.2 The efficiency of the RS estimator (3.3), (4.5) for the $M/D/1$ queue with $\rho = 0.3,\ d = 1$, and $p_x = 1.04 \cdot 10^{-9}$ $(x = 20)$.

t	δ	$\overline{T}_{\widetilde{\pi}}$	R_t	$\exp(zx)$	z
1.3	-0.368	$1.11 \cdot 10^7$	39.42	27.1	$1.65 \cdot 10^{-1}$
1.6	-0.22	$1.55 \cdot 10^5$	5.48	3.46	$6.20 \cdot 10^{-2}$
2.0645*	0.0	$2.84 \cdot 10^4$	1.0	1.0	0.0
2.4	0.19	$7.51 \cdot 10^4$	2.65	2.27	$4.10 \cdot 10^{-2}$
2.7	0.31	$1.06 \cdot 10^7$	37.43	48.4	$1.94 \cdot 10^{-2}$

Table 7.3 The efficiency of the RS estimator (3.3), (4.5) for the $MM/M/1$ queue with $\rho = 0.5,\ \mu = 1,\ \theta = 0.3$ and $p_x = 5.34 \cdot 10^{-10}$ $(x = 40)$.

t	δ	$\overline{T}_{\widetilde{\pi}}$	R_t	$\exp(zx)$	z
0.40	-0.2	$1.08 \cdot 10^6$	4.2	3.51	$1.22 \cdot 10^{-2}$
0.45	-0.1	$5.13 \cdot 10^5$	2.0	1.44	$4.19 \cdot 10^{-2}$
0.50*	0.0	$2.55 \cdot 10^5$	1.0	1.0	0.0
0.54	0.08	$5.43 \cdot 10^5$	2.1	1.45	$1.23 \cdot 10^{-2}$
0.58	0.16	$1.91 \cdot 10^6$	7.5	10.3	$7.78 \cdot 10^{-2}$

Table 7.4 The efficiency of the RS estimator (6.4), (4.5)
for the $M/M/1$ queue
with $\rho = 0.3,\ \mu = 1,\ p_x = 1.44 \cdot 10^{-16}$ $(x = 30)$, and $\nabla_\lambda p_x = 1.42 \cdot 10^{-14}$.

t	δ	$\overline{T}_{\widetilde{\pi}}$	R_t	$\exp(zx)$	z
0.50	-0.285	$1.03{\cdot}10^7$	40.7	53.9	$1.33{\cdot}10^{-2}$
0.60	-0.143	$8.75{\cdot}10^5$	3.75	3.53	$4.21{\cdot}10^{-2}$
0.70^*	0.0	$2.53{\cdot}10^5$	1.0	1.0	0.0
0.75	0.07	$3.68{\cdot}10^5$	1.45	1.9	$2.14 \cdot 10^{-2}$
0.80	0.143	$3.17{\cdot}10^7$	124.9	171.2	$1.71 \cdot 10^{-1}$

It follows from these tables that our simulation results are in agreement with the theoretical ones. In particular, the RS estimator (3.3), (4.5) is robust with respect to small and moderate perturbations in w, in the sense that for the relative perturbation $\mid \delta \mid < 0.2$ we still have dramatic variance reduction. Similar results were obtained for the $GI/G/1$ queue with different interarrival and service time distributions.

Our extensive numerical results also suggest that the optimal parameter w, once found, can be used for estimating *simultaneously all probabilities* p_x (for different values x) of the order 10^{-2} or less.

7.2 Further research

We now give some guidance on how to approximate the optimal parameter w (for the above $GI/G/1$ queue and its extensions) without resorting to the solution of the equation of type

$$\Lambda(t) \equiv \Lambda_A(t) + \Lambda_B(-t) = 0. \tag{7.1}$$

To do so, consider minimization of the variance of the switching regenerative estimator (3.3), (4.5), but with the OP w in (4.5) replaced by t, that is, consider the following program:

$$\min_t \mathrm{Var}\{\bar{p}_{x,N}(t)\}. \tag{7.2}$$

It readily follows from the definition of $T_{\widetilde{\pi}_t}$ that asymptotically in N the optimal solution of (7.2) coincides with the optimal solution w of the program

$$\min_t T_{\widetilde{\pi}_t}. \tag{7.3}$$

where $\widetilde{\pi}_t = \widetilde{\pi}(G'_A, G'_B) \in \Pi'$ and $G'_A = F_A^{(t)},\ G'_B = F_B^{(-t)}$. One of our main goals will be to show that the SR estimator (3.3), (4.5) still remains polynomial, if we replace the OECM pair $(F_A^{(w)}(y_1, \mathbf{v}_1), F_B^{(-w)}(y_2, \mathbf{v}_2))$ in W_{ni} (see (4.5)) by the following pair $(G^*_A, G^*_B) = (F_A(y_1, \mathbf{v}^*_{01}), F_B(y_2, \mathbf{v}^*_{02}))$,

where the pair $(\mathbf{v}_{01}^*, \mathbf{v}_{02}^*)$ represents the optimal solution of the program (7.2), but with t replaced by the vector $(\mathbf{v}_{01}, \mathbf{v}_{02})$.

In practice, however, the exact optimal solution $(\mathbf{v}_{01}^*, \mathbf{v}_{02}^*)$ of the program

$$\min_{(\mathbf{v}_{01}, \mathbf{v}_{02})} \mathrm{Var}\{\bar{p}_{x,N}(\mathbf{v}_{01}, \mathbf{v}_{02})\} \tag{7.4}$$

is not available, so it must be estimated by simulation,that is, it is the solution of the following stochastic counterpart, (see e.g., [29]):

$$\min_{(\mathbf{v}_{01}, \mathbf{v}_{02})} \{\tilde{\mathrm{Var}}\bar{p}_{x,N}(\mathbf{v}_{01}, \mathbf{v}_{02})\}, \tag{7.5}$$

Here $\tilde{\mathrm{Var}}(\cdot)$ is the sample variance of $\mathrm{Var}(\cdot)$.

Problem (7.5) can be readily extended to more general queueing models, such as

$$\min_{\mathbf{v}_0} \{\tilde{\mathrm{Var}}\bar{p}_{x,N}(\mathbf{v}_0)\}.$$

Let $\hat{\mathbf{v}}_{0N}^*$ be the optimal solution of this program. One of our further main goals is to show that the SR estimators of the type (3.3), (4.5) with the estimated optimal value $\hat{\mathbf{v}}_{0N}^*$ in the LR $W_{ni}(\mathbf{y}, \hat{\mathbf{v}}_{0N}^*)$ will still possess reasonably low variance. Such optimism relies on:

(a) the robustness and low variance of the SR estimator (3.3), (4.5) with the OP w;

(b) the convergence of $\hat{\mathbf{v}}_{0N}^*$ to $\mathbf{v}_0^*$ (see [**?**]),from which it follows that one can always take a sample N (perhaps, large enough) such that

$$||\hat{\mathbf{v}}_{0N}^* - \mathbf{v}_0^*||/||\mathbf{v}_0^*|| \leq 0.2$$

holds with high probability. Note that 0.2 is associated with the robustness of the SR estimator (3.3), (4.5), in the sense that for relative perturbations $|\delta| = |(t-w)|/w \leq 0.2$ it still has manageable variance.

Concluding remarks

In this paper we presented a framework for complexity analysis of rare event estimators. In particular we defined polynomial and exponential time SR estimators for the evaluation of the steady-state probabilities of excessive backlog p_x in the $GI/GI/1$ queue, and some of its extensions. The proposed SR estimators are based on large deviation theory and use of exponential change of measure, parametrized by a scalar t. We showed how to find the optimal value w of the parameter t, which results in the optimal exponential change of measure (OECM) and found conditions (see Theorem 4.1) under which the SR estimator (3.3), (4.5) is polynomial. We investigated the robustness of the proposed SR estimators, in the sense that we found

how much one can perturb the optimal value w while the SR estimator still results in dramatic variance reduction and remains useful in practice. In particular, our numerical results suggest that if the optimal parameter w is perturbed up to 20%, we only loose 2-3 orders of magnitude of variance reduction, compared with the orders of10 under the optimal value w.

8 Appendies

Appendix A: Optimal exponential change of measure

Large deviation theory allows us to identify the exponential part of statistics of rare events, namely:

$$\lim_{x\to\infty} \frac{1}{x} \log \alpha_x \sim I$$

or

$$\alpha_x \approx \exp(-xI) \quad \text{(UTLE)} \quad ,$$

where UTLE is the acronym for 'up to logarithmic equivalence'.

We shall use LD results that deal with identifying the so-called optimal exponential change of measure (OECM). We now explain this concept referring to the $GI/GI/1$ queue case.

Let X be a random variable (rv) with distribution $F_X(\cdot)$. Let

$$\Lambda_X(\theta) = \log(\mathbb{E}e^{-\theta X}), \theta \in R^1$$

denote the cumulant function, and define

$$\mathcal{D}_X \equiv \{\theta \in R^1 \mid \Lambda_X(\theta) < \infty\}.$$

For any $\theta \in \mathcal{D}_X$ define the exponential change of measure as follows:

$$F_X^{(\theta)}(\mathrm{d}x) = \exp(-\theta x - \Lambda_X(\theta)) F_X(\mathrm{d}x).$$

The $GI/GI/1$ queue is determined by iid interarrival and service times sequences $\{A_k\}$ and $\{B_k\}$. Define

$$\Lambda(\theta) = \Lambda_A(\theta) + \Lambda_B(-\theta)$$

and

$$w = \sup[\theta \in \mathcal{D}_A \cap \mathcal{D}_B \mid \Lambda(\theta) \leq 0].$$

The classical LD result 'states' that if $0 < w < \infty$, $\Lambda(w) = 0$, and $\Lambda'(w) < \infty$, then

$$P_\pi(Q_n \geq x) \approx \exp(x\Lambda(w)) \quad \text{(UTLE)} \quad .$$

Moreover,

$$\theta = w, \tag{8.1}$$

specifies by

$$F_A^{(\theta)}(\mathrm{d}x) = \exp(-\theta x - \Lambda_A(\theta))F_A(\mathrm{d}x), \tag{8.2}$$

and

$$F_B^{(\theta)}(\mathrm{d}x) = \exp(\theta x - \Lambda_B(-\theta))F_B(\mathrm{d}x) \tag{8.3}$$

specify the optimal (in terms of minimal variances of the IS estimators) exponential change of measure within the class of all feasible ($\theta \in \mathcal{D}_A \cap \mathcal{D}_B$) exponential changes of measure (called the optimal exponential change of measure (OECM)).

Recall that OECM results in an unstable queueing process (see e.g., [17] for some discussions). For example, for the $M/M/1$ queue with traffic intensity $\rho < 1$ the OECM corresponds to the $M/M/1$ queue with traffic intensity $\rho_0 \equiv \lambda_0/\mu_0 = \rho^{-1} = \mu/\lambda > 1$.

Appendix B: Proofs of main results

Proof of Theorem 4.1. (sketch)

Part A. Using the well known formula for the variance of ratio estimators (see e.g., [29]) we have (after simple transformations) that the SCV $\kappa(x)$ of the SR estimator (3.3), (4.5) is given by

$$\kappa(x) = \frac{\mathbb{E}_{\widetilde{\pi}} X^2}{(\mathbb{E}_{\widetilde{\pi}} X)^2} + \frac{\mathbb{E}_{\widetilde{\pi}} Y^2}{(\mathbb{E}_{\widetilde{\pi}} Y)^2} - 2\frac{\mathbb{E}_{\widetilde{\pi}} XY}{\mathbb{E}_{\widetilde{\pi}} X \mathbb{E}_{\widetilde{\pi}} Y} = \kappa_1 + \kappa_2 + \kappa_3, \tag{8.4}$$

where $X \equiv \sum_{t=1}^{\tau} I_t W_t$, $Y \equiv \sum_{t=1}^{\tau} W_t$, $I_t \equiv I_{[x,\infty)}(Q_t)$,

$$W_t = \frac{\prod_{k=1}^{D(t)} F_B(\mathrm{d}B_k) \prod_{k=1}^{t} F_A(\mathrm{d}A_k)}{\prod_{k=1}^{D(t)} G_{B,k}(\mathrm{d}B_k) \prod_{k=1}^{t} G_{A,k}(\mathrm{d}A_k)},$$

where $\prod_{k=i}^{j} C_k \equiv 1$ and $\sum_{k=i}^{j} C_k \equiv 0$ for any sequence $\{C_k\}$ if $j < i$. Using the IS policy $\widetilde{\pi} = \pi^*$ (see Definition 4.2), we obtain

$$\begin{aligned} W_t &= \exp(x\Lambda_A(w) + wZ_t), \quad t = \xi, \\ &= W_\xi, \quad \xi < t < \tau, \end{aligned} \tag{8.5}$$

where

$$Z_t \equiv \sum_{k=1}^{t} A_k - \sum_{k=1}^{t-x} B_k. \tag{8.6}$$

Now consider each term of the right hand side of (8.4) separately:

$$\begin{aligned} \mathbb{E}_{\pi^*} X^2 &= \mathbb{E}_{\pi^*}[(\sum_{t=\xi}^{\tau} I_t W_t)^2; \xi \le \tau] \\ &= \mathbb{E}_{\pi^*}[W_\xi^2 \sum_{t=\xi}^{\tau} I_t; \xi \le \tau] + 2\mathbb{E}_{\pi^*}[\sum_{t=\xi}^{\tau} I_t W_t \sum_{s=t+1}^{\tau} I_s W_s; \xi \le \tau] \\ &= \mathbb{E}_{\pi^*}[e^{2Z_\xi w} h_x; \xi \le \tau] e^{2\Lambda_A(w)x}, \end{aligned} \tag{8.7}$$

where $h_x = \mathbb{E}[\sum_{t=\xi}^{\tau} I_t(1 + 2\sum_{s=t+1}^{\tau} I_s) \mid \mathcal{F}_\xi]$, withs $\mathcal{F}_\xi = \sigma(\widetilde{B}_\xi)$. Hence we can write $h_x \equiv h_x(\widetilde{B}_\xi)$.

Lemma 8.1. Under assumptions **A1**, **A2** we have

1. The pointwise monotone limit $h_x(\cdot) \to h(\cdot)$, where $h(\cdot)$ is a measurable bounded function.
2. $$\lim_{x\to\infty} \mathbb{E}_w[e^{2Z_\xi w} h(\widetilde{B}_\xi); \xi \le \tau] = O(1) \lim_{x\to\infty} \mathbb{E}[e^{2Z_\xi w}; \xi \le \tau].$$

Proof. The proof of this lemma is, in fact, a specification of the proofs of Lemma 5.2 and Lemma 5.4 in [33]. □

Write now

$$\xi = \min[1 \le t \le \tau \mid Z_t < 0].$$

By the renewal theory (see e.g., [3] or [15]) Z_ξ has a limiting distribution

$$\lim_{x\to\infty} \mathbb{E}_w[e^{2Z_\xi w}; \xi \le \tau] = \psi,$$

where ψ is a constant $0 < \psi < \infty$. Thus, $\mathbb{E}_{\pi^*} X^2 = O(1)\exp(2\Lambda_A(w)x)$. Noting that $(\mathbb{E}X)^2 = (\mathbb{E}\tau \mathbf{P}_\pi(Q \ge x))^2 = \widetilde{O}(1)\exp(2\Lambda_A(w)x)$ (see e.g., [3]), we obtain that κ_1 is bounded by a constant function in x.

Consider κ_3. We have (see. [5])

$$\begin{aligned}
\mathbb{E}Y^2 &= \mathbb{E}_{\pi^*}[(\sum_{s=t}^{\tau} W_t)^2] \\
&= \mathbb{E}_{\pi^*}[(\sum_{s=t}^{\tau} W_t^2)] + 2\mathbb{E}_{\pi^*}[\sum_{t=1}^{\tau} W_t \sum_{s=t+1}^{\tau} W_s] \\
&= \mathbb{E}_{\pi^*}[(\sum_{s=t}^{\tau} W_t^2)(1 + 2\mathbb{E}[\tau - t \mid \mathcal{F}_t]]. \qquad (8.8)
\end{aligned}$$

Since $t \ge D(t)$, $S_t \equiv \sum_{k=1}^{t} A_k - \sum_{k=1}^{D(t)} B_k \le 0$, $1 \le t \le \tau$, and $\Lambda_A(w) < 0$, we obtain $W_t \equiv \exp((t - D(t))\Lambda_A(w) + wS_t) \le 1$, $1 \le t \le \xi$, and thus

$$\begin{aligned}
\mathbb{E}_{\pi^*}Y^2 &\le \mathbb{E}_{\pi^*}[\sum_{t=1}^{\tau} W_t] + 2\mathbb{E}_{\pi^*}[\sum_{t=1}^{\tau} W_t \mathbb{E}[\tau - s \mid \mathcal{F}_t]] \\
&= \mathbb{E}\tau + 2\mathbb{E}[\sum_{t=1}^{\tau} \mathbb{E}[\tau - t \mid \mathcal{F}_t]] \\
&\le \mathbb{E}\tau + 2(\mathbb{E}\tau)^2 = O(1).
\end{aligned}$$

Then $\mathbb{E}_{\pi^*}Y = \mathbb{E}\tau$; hence, $\kappa_2(x) = O(1)$.

Noting next that $N < \gamma(\kappa_1(x)+\kappa_2(x))$, we have $N = O(1)$. (Recall that N is the number of cycles required to obtain an estimator of (ϵ,δ)-accuracy and $\gamma = \Phi^{-1}(1-\delta/2)^2\epsilon^{-2}$.) It is not difficult to show that

$$\lim_{x\to\infty} P_{\pi^*}(\xi \le \psi x; \xi < \tau) = 1, \tag{8.9}$$

where $\psi < \infty$. Indeed, note that the waiting time of the $(\xi-1)$-th customer

$$L_\xi = \sum_{k=1}^{\xi-1}(B_k - A_k) < \sum_{k=\xi-x}^{\xi-1} B_k$$

provided $\xi < \tau$. This implies

$$\sum_{k=1}^{\xi-1}(B'_k - A_k) < \sum_{k=\xi-x}^{\xi-1} B'_k,$$

where

$$\begin{aligned} B'_k &= B_k, \quad 1 \le k \le \xi - x \\ &= S_k, \quad k > \xi - x \end{aligned}$$

with $S_k \sim G_B$. It is clear that $\xi = \xi(x) > x$. Hence, by the strong law of large numbers (SLLN) we obtain

$$\xi^{-1}\sum_{k=1}^{\xi}(B'_k - A_k) \to \mathbb{E}_{G_B}B - \mathbb{E}_{G_A}A \equiv \delta_1 > 0$$

(a.s.) as $x \to \infty$. Besides, by the SLLN we get

$$x^{-1}\sum_{k=\xi-x}^{\xi-1} B'_k \to \mathbb{E}_{G_B}B \equiv \delta_2$$

(a.s.) as $x \to \infty$. Hence, $\lim_{x\to\infty} P_{\pi^*}(\xi\delta_1 < x\delta_2; \xi < \tau) = 0$ for $0 < \delta_2/\delta_1 < \infty$. Moreover, using the CLT we can obtain by routine evaluation $P_{\pi^*}(\xi/x = \widetilde{\psi}_1; \xi < \tau) = O(\exp(-r_1\widetilde{\psi}_1 x))$ for $\delta_2/\delta_1 < \widetilde{\psi}_1 < \infty$, $r > 0$. Therefore

$$\mathbb{E}_{\pi^*}[\xi; \xi < \tau] \le \psi_1 x.$$

Similarly,

$$\mathbb{E}_{\pi^*}[\tau - \xi; \tau > \xi] \le \psi_2 x, \tag{8.10}$$

where ψ_1 and ψ_2 are some finite constants. We also note that

$$P_{\pi^*}(\tau = \psi x; \tau < \xi) = P_{\pi^*}(\sum_{k=1}^{\psi x}(B_k - A_k) < 0; \tau < \xi).$$

But $(\psi x)^{-1}\sum_{k=1}^{\psi x}(B_k - A_k) \to \mathbb{E}_{G_B}B_k - \mathbb{E}_{G_A}A_k > 0$ (a.s.) as $x \to \infty$. Hence by the CLT,

$$P_{\pi^*}(\tau = \psi x; \tau < x) = O(\exp(-r_2\psi x))$$

with $0 < \psi < \infty$, $0 < r_2 < \infty$ for large x, and therefore $\mathbb{E}[\tau; \tau < \xi] < \psi_3 x$, $\psi_3 < \infty$. Thus we have

$$\mathbb{E}_{\pi^*}\tau = \mathbb{E}_{\pi^*}[\xi; \tau > \xi] + \mathbb{E}_{\pi^*}[\tau - \xi; \tau > \xi] + \mathbb{E}_{\pi^*}[\tau; \tau \le \xi].$$

Hence, $\mathbb{E}_{\pi^*}\tau \le (\psi_1 + \psi_2 + \psi_3)x$.

Finally, $T_{\pi^*} = N\mathbb{E}_{\pi^*}\tau = O(x)$. $\square$

Part B. Note that according to (8.4), (8.7) and (8.8), we have

$$\frac{\mathbb{E}_{\widetilde{\pi}}[X^2]}{\mathbb{E}_{\widetilde{\pi}}[Y^2]} > \frac{\mathbb{E}_{\widetilde{\pi}}[\sum_{t=1}^{\tau} I_t W_t^2]}{\mathbb{E}_{\widetilde{\pi}}[\sum_{t=1}^{\tau} W_t^2(1 + 2\mathbb{E}[\tau - t \mid \mathcal{F}_t]]}.$$

Then for $\widetilde{\pi} \in \Pi$ and $H_t \equiv \mathbb{E}[\tau - t \mid \mathcal{F}_t]$ we have $H_t = O(Q_t)$, given $\mathcal{F}_t = (Q_t, \widetilde{B}_t)$ with $Q_t \to \infty$ and $\widetilde{B}_t < \infty$ (similar to (8.10)). Taking into account that, according to the definition of $\widetilde{\pi} \in \Pi$, we have

$$P_{\widetilde{\pi}}(Q_t = \widetilde{\psi}_1 x) = O(\exp(-r\widetilde{\psi}_1 x))$$

for $\widetilde{\psi}_1 > 1$, $0 < r < \infty$., and hence

$$\mathbb{E}_{\widetilde{\pi}}[W_t^2 H_t] < \psi x \mathbb{E}_{\widetilde{\pi}}[W_t^2]$$

where $\psi < \infty$.

Thus

$$\frac{\mathbb{E}_{\widetilde{\pi}}[X^2]}{\mathbb{E}_{\widetilde{\pi}}[Y^2]} > (1 + 2\psi x)^{-1}\frac{\mathbb{E}_{\widetilde{\pi}}[\sum_{t=1}^{\tau} I_t W_t^2]}{\mathbb{E}_{\widetilde{\pi}}[\sum_{t=1}^{\tau} W_t^2]}.$$

Now we introduce the following notation (similar to (5.1)–(5.2)):

$$K_{A,k}(\mathrm{d}A) = \frac{F_{A,k}(\mathrm{d}A)}{G_{A,k}(\mathrm{d}A)}F_{A,k}(\mathrm{d}A),$$

$$K_{B,k}(\mathrm{d}B) = \frac{F_{B,k}(\mathrm{d}B)}{G_{B,k}(\mathrm{d}B)}F_{B,k}(\mathrm{d}B),$$

and we let $\widetilde{\pi}_K$ and $\mathbb{I}_{\widetilde{\pi}_K}$ with K_A and K_B be defined in the same way as $\widetilde{\pi}$ and $\mathbb{E}_{\widetilde{\pi}}$ with G_A and G_B, respectively. Now we can write

$$\frac{\mathbb{E}_{\widetilde{\pi}}[X^2]}{\mathbb{E}_{\widetilde{\pi}}[Y^2]} > (1 + 2\psi x)^{-1}\frac{\mathbb{I}_{\widetilde{\pi}_K}[\sum_{t=1}^{\tau} I_t]}{\mathbb{I}_{\widetilde{\pi}_K}[\sum_{t=1}^{\tau} 1]}.$$

Assume without loss of generality that $\mathbb{I}_{\widetilde{\pi}_K}\tau < \infty$. Using the basic result for regenerative processes, (see e.g., [3], (5.1.1)) we obtain

$$\frac{\mathbb{E}_{\widetilde{\pi}}[X^2]}{\mathbb{E}_{\widetilde{\pi}}[Y^2]} > (1+2\psi x)^{-1}\mathbb{I}_{\pi_K} I_t.$$

, where the subscript π_K indicates that the 'expectation' $\mathbb{I}_{\pi_K}$ is taken under the steady-state regime, provided the behavior of the queue is 'driven' by the IS policy $\widetilde{\pi}_K$. Now using the fact that $I_t = I_{[-\infty,0)}(\max_{s,\ 1\le s\le t}\widetilde{Q}_s)$, where $\widetilde{Q}_{t+1} = \widetilde{Q}_t + 1 - D_t$, D_t is the number of customers departing from the queue between the t-th and $t+1$ arrivals,s given it is not empty during this period. (It is clear that $Q_{t+1} = \max[Q_t + 1 - D_t, 0]$, i.e., $\widetilde{Q}_t$ is an unreflected modification of Q_t.) Denote $\widetilde{\xi} \equiv \min[t \mid \widetilde{Q}_t > x]$. We obtain (see, e.g., [15]) that $\mathbb{I}_{\pi_K} I_t = \mathbb{I}_{\widetilde{\pi}_K}[\widetilde{\xi} < \infty]$. Thus, we have

$$\frac{\mathbb{E}_{\widetilde{\pi}}[X^2]}{\mathbb{E}_{\widetilde{\pi}}[Y^2]} > (1+2\psi x)^{-1}\mathbb{I}_{\widetilde{\pi}_K}[\widetilde{\xi} < \infty].$$

Parallel to the proof of Lemma 3 in [23] we obtain

$$\mathbb{I}_{\widetilde{\pi}_K}[I_{[0,\infty)}(\widetilde{\xi})] > \mathbb{I}_{\widetilde{\pi}_{K_u}}[I_{[0,\infty)}(\widetilde{\xi})\exp(P_u)],$$

where

$$P_u = \sum_{t=1}^{\widetilde{\xi}}(\Lambda_{K_A,t}(u) + \sum_{t=1}^{\widetilde{\xi}_{-x}}\Lambda_{K_B,t}(-u)); \tag{8.11}$$

the subscript $\widetilde{\pi}_{K_u}$ indicates that we use $K_A^{(u)}$ and $K_B^{(-u)}$ instead of K_A and K_B. One can choose u such that

$$\lim_{x\to\infty}\mathbb{I}_{\widetilde{\pi}_{K_u}}[I_{[0,\infty)}(\widetilde{\xi})] > \psi_1 > 0$$

(see e.g. [24], page 227 or [23], Lemma 3).

By Jensen's inequality, $\Lambda_{K_{A,t}}(2u) \ge 2\Lambda_A(w)$ and $\Lambda_{K_{B,t}}(-2u) \ge 2\Lambda_B(-w)$ with equalities if and only if $G_{A,t} = F^{(w)}$ and $G_{B,t} = F^{(-w)}$. Let $\widetilde{\pi} \notin \Pi^*$, then

$$\lim_{x\to\infty}\frac{P_u - 2\Lambda_A(w)}{x} = \widetilde{\psi} > 0.$$

Thus,

$$\frac{\kappa_1}{\kappa_2} > \frac{\widetilde{O}(1)}{1+2\psi x}\frac{\mathbb{E}_{\widetilde{\pi}}X^2}{\mathbb{E}_{\widetilde{\pi}}Y^2\exp(2\Lambda_A(w)x)} = \frac{\widetilde{O}(1)}{1+2\psi x}\exp(zx), \tag{8.12}$$

where $z > 0$. Finally, note that by Cauchy's inequality

$$T_{\widetilde{\pi}} \ge N \ge \gamma(\sqrt{\kappa_1} - \sqrt{\kappa_2})^2.$$

We have proved that $\widetilde{\pi} \notin \Pi^*$ implies $\widetilde{\pi} \notin \Pi^{**}$, which directly implies statement B of the theorem. $\Box$.

Proof of Theorem 5.1. (sketch) Parallel to the proof of Theorem 4.1, Part B for $\widetilde{\pi} \in \Pi'$, and using that $P_v = \Lambda_{K_A}(v)x$ in (8.11), we obtain that

$$\frac{\kappa_1}{\kappa_2} > \frac{\widetilde{O}(1)}{1+2\psi x} \exp(x(\Lambda_{K,A}(v) - 2\Lambda_A(w)))$$

; that is, $z_{\widetilde{\pi}} > \Lambda_{K,A}(v) - 2\Lambda_A(w) \geq 0$. (For $z_{\widetilde{\pi}} > 0$ we have by (8.12) that κ_1 is the dominating term in (8.4).) Then taking into account that $\widetilde{B}_\xi$ has a limiting distribution (by the renewal theory) and $h(\widetilde{B}) = \mathbb{E}[\sum_{t=\xi}^{\tau} I_t(1 + 2\sum_{s=t+1}^{\tau} I_s) \mid \widetilde{B}_\xi]$ is a bounded function (by Lemma 8.1) we have

$$\mathbb{E}_{\widetilde{\pi}}[h(\widetilde{B}_\xi) \mid \mathcal{F}_\xi] < \psi < \infty \quad \text{(a.s.)}.$$

Thus we obtain an upper bound

$$\kappa_1 < \exp(-2\Lambda_A(w)x)\mathbb{E}_{\widetilde{\pi}}[W_\xi^2 h(\widetilde{B}_\xi)] < \psi \exp(-2\Lambda_A(w)x)\mathbb{E}_{\widetilde{\pi}}[W_\xi^2]$$

. Then under **A1**–**A3** we have by Theorem 3.1 [30]

$$\mathbb{E}_{\widetilde{\pi}}[W_\xi^2] = \widetilde{\psi} \exp(x\Lambda_{K,A}(v),$$

where $0 < \widetilde{\psi} < \infty$. Finally, we obtain $z_{\widetilde{\pi}} = \Lambda_{K,A}(v) - 2\Lambda_A(w)$. $\Box$

We omit the proofs of Propositions 6.1–6.3, since they differ from the proof of Theorems 4.1 and 5.1 only in their routine calculations.

Proof of Proposition 6.4 (sketch).
We can write $\nabla p_x \equiv \nabla_{\mathbf{v}} p_x$ as follows (see [29]):

$$\begin{aligned} \nabla p_x &= \frac{\mathbb{E}\sum_{t=1}^{\tau} I_t S_t}{\mathbb{E}\tau} - \frac{\mathbb{E}\sum_{t=1}^{\tau} I_t}{\mathbb{E}\tau}\frac{\mathbb{E}\sum_{t=1}^{\tau} S_t}{\mathbb{E}\tau} \\ &\equiv p_x^{(1)} - p_x p^{(2)}. \end{aligned}$$

Let $T_{\widetilde{\pi}}^{(1)}$, $T_{\widetilde{\pi}}$, $T_{\widetilde{\pi}}^1$, $T_{\widetilde{\pi}}^2$, and $T_{\widetilde{\pi}}^{0,2}$ be simulation costs required to obtain IS estimators of prescribed accuracy for ∇p_x, p_x, $p_x^{(1)}$, $p_x^{(2)}$, and $p_x p_x^{(2)}$, respectively. It is clear that

$$\max[T_{\widetilde{\pi}_1}, T_{\widetilde{\pi}_2}^{0,2}] < T_{\widetilde{\pi}_i}^{0,2} < \psi_1 T_{\widetilde{\pi}_1} + \psi_2 T_F^{0,2}, \ i = 1,2.$$

Here $\widetilde{\pi}_i$, $i = 1,2$ are certain IS policies from Π, ψ_i, $i = 1,2$ are some finite constants, and $T_F^{0,2} = T_{\widetilde{\pi}_2}^{0,2}$ for the CMC estimator. Taking into account that $T_F^{0,2}$ does not depend on x, we obtain

$$T_{\widetilde{\pi}} \leq T_{\widetilde{\pi}}^{0,2} = O(T_{\widetilde{\pi}})$$

for large x. This implies that

$$T_{\widetilde{\pi}} < T_{\widetilde{\pi}}^{(1)} < \psi_3 T_{\widetilde{\pi}}^1 + \psi_4 T_{\widetilde{\pi}},$$

when $\psi_i < \infty,\ i = 3, 4$. It remains to prove that the simualtion cost $T_{\widetilde{\pi}}^1 = O(T_{\widetilde{\pi}})$. Similar to (8.4) we have

$$\begin{aligned} T_{\widetilde{\pi}}^1 = \mathbb{E}_{\widetilde{\pi}} \tau\gamma\kappa(x) &= \frac{\mathbb{E}_{\widetilde{\pi}} X_1^2}{(\mathbb{E}_{\widetilde{\pi}} X_1)^2} + \frac{\mathbb{E}_{\widetilde{\pi}} Y^2}{(\mathbb{E}_{\widetilde{\pi}} Y)^2} - \frac{2\mathbb{E}_{\widetilde{\pi}} X_1 Y}{\mathbb{E}_{\widetilde{\pi}} X_1 \mathbb{E}_{\widetilde{\pi}} Y} \\ &= \kappa_{1,1} + \kappa_{2,1} + \kappa_3, \end{aligned} \tag{8.13}$$

where $X_1 \equiv \sum_{t=1}^{\tau} I_t S_t W_t$, and $Y \equiv \sum_{t=1}^{\tau} W_t$. We will show that $\kappa_{1,1} = O(\kappa_1)$ (see (8.4)) which implies by the proof of Theorem 4.1 (see in particular (8.12)) that $T_{\widetilde{\pi}}^1 = O(T_{\widetilde{\pi}})$.

From the proof of Theorem 3.1 we have that $\lim_{x\to\infty} P_{\widetilde{\pi}}(\xi \le \tau < \psi_2 x) = 1$. Then by the definitions of $f_A(\cdot)$ and S_t we have

$$S_t = \psi_3 t + \psi_4 \sum_{k=1}^{t} c(A_k)$$

and $\mathbb{E}c(A_k)^2 < \infty$. Hence,

$$\lim_{x\to\infty} P_{\widetilde{\pi}}(\mid S_t \mid < \psi_6 x) = 1$$

for $t \le \tau$. Also by the CLT we obtain, using routine evalution, that

$$\lim_{x\to\infty} P_{\widetilde{\pi}}(\mid S_t \mid > \psi_7 x) = O(\exp(-r\psi_7 x)),$$

where $r > 0$. Therefore

$$\mathbb{E}_{\widetilde{\pi}}(X_1^2) = \mathbb{E}_{\widetilde{\pi}}((\sum_{t=1}^{\tau} I_t W_t S_t)^2) = O(x^2)\mathbb{E}(X^2).$$

Also we have

$$\mathbb{E}_{\widetilde{\pi}} X_1 \mathbb{E}_{\widetilde{\pi}}((\sum_{t=1}^{\tau} I_t W_t S_t)) = (\nabla p_x - p_x p^{(2)})\mathbb{E}\tau = \widetilde{O}(x p_x) = \widetilde{O}(x)\mathbb{E}X.$$

Hence we get

$$\kappa_{1,1} = O(\kappa_1).$$

The statement of the proposition directly follows from the last equation. □

Acknowledgement

We would like thank Adam Schwartz at Technion and Jack Kleijnen at Tilburg Unviersity for several valuable suggestions on the earlier draft of this work.

References

[1] V. Anantharam, P. Heidelberger and P. Tsoucas (1990) Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. Research report, IBM Research Division.

[2] S. Asmussen (1985) Conjugate processes and the simulation of ruin problems, *Stoch. Proc. Appl.,* 20, 213–229.

[3] S. Asmussen (1987) *Applied Probability and Queues,* John Wiley & Sons, New York.

[4] S. Asmussen (1989) Risk theory in a Markovian environment, *Scand. Actuarial J.,* 69–100.

[5] S. Asmussen, R.Y. Rubinstein and C.L. Wang (1992) Regenerative rare events simulation via likelihood ratios, to be published in the *Journal of Appl. Prob.*

[6] S. Asmussen and R.Y. Rubinstein (1994) Complexity properties of steady-state rare events simulation in queueing models. Submitted to *Frontiers in Queueing.*

[7] V.A. Bolotin, J.G. Kappel and P.J. Kuehn (1991) Teletraffic analysis of ATM systems, *IEEE Journal Select. Areas Commun.,* 9, 281–283.

[8] J.A. Bucklew (1990) *Large Deviation Techniques in Decision, Simulation, and Estimation*, John Wiley & Sons, New York.

[9] J.A. Bucklew, P. Ney and and J.S. Sadowsky (1990) Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains, *J. Appl. Prob.*, 27, 44-59.

[10] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin (1992) Effective bandwidth and fast simulation of ATM intree networks. Research report, IBM Research Division, T.J. Watson Research Center, NY.

[11] M. Cottrel, J.C. Fort and G. Malgoures (1983) Large deviations and rare events in the study of stochastic algorithms, *IEEE Trans. Automatic Control,* Ac-28, 907-918.

[12] M. Devetsikiotis and K.R. Townsend (1992) On the efficient simulation of large communication networks using importance sampling, in *Proceedings of IEEE Globecom '92,* IEEE Computer Society.

[13] M. Devetsikiotis and K.R. Townsend (1992) A dynamic importance sampling methodology for the efficient estimation of rare events probabilities in regenerative simulations of queueing systems, in *Proceedings of IEEE Globecom '92,* 1290–1297, IEEE Computer Society.

[14] M. Devetsikiotis and K.R. Townsend (1993) Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks, preprint.

[15] W. Feller (1966) *An Introduction to Probability Theory and its Application,* Volume 2, John Wiley & Sons, New York.

[16] M.R. Frater, T.M. Lennon and B.D.O. Anderson (1991) Optimally efficient estimation of the statistics of rare events in queueing networks, *IEEE Trans. on Auto. Control,* AC-36, 1395–1405.

[17] M.R. Frater and B.D.O. Anderson (1989) Fast estimation of the statistics of excessive backlogs in tandem networks of queues, *Australian Telecommunication Research,* 23, 49-55.

[18] M.R. Frater, J. Walrand and B.D.O. Anderson (1990) Optimality efficient estimation of the buffer overflow in queues with deterministic service times, *Australian Telecommunication Research,* 24, 1–8.

[19] P.W. Glynn and D.L. Iglehart (1989). Importance sampling for stochastic simulations, *Management Sci.*, 35 (11), 1367–1392.

[20] P. Heidelberger (1993) Fast simulation of rare events in queueing and reliability models. IBM Research Report RC 19028, Yorktown heights, New York. Preliminary version published in *Performance Evaluation of Computer and Communications Systems*. Springer Lecture Notes in Computer Science **729**, 165–202.

[21] J.F.C. Kingman (1966) On the algebra of queues, *Ann. Math. Statis.,* 3, pp. 285–326.

[22] V. Kriman (1993) Sensitivity analysis of $GI/GI/m/B$ queues with respect to buffer size by the score function method, to be published in *Stochastic Models*

[23] T. Lehtonen and H. Nyrhinen (1992) Simulating level crossing probabilities by importance sampling, *Adv. Appl. Prob.,* December.

[24] A. Martin-Lof (1986) Entropy, a useful concept in risk theory, *Scand. Actuarial J.,* 223-235.

[25] B. Melamed (1991) TES: A class of methods for generating autocorrelated uniform variates, ORSA J. on Computing, 3, 317-329.

[26] V.F. Nicola, P. Shahabuddin, P. Heidelberger and P.W. Glynn (1993) Fast simulation of steady-state availability in non-Markovian highly dependable systems, in *Proceedings of the Twentieth International Symposium on Fault-Tolerant Computing,* 491-498, IEEE Computer Society Press.

[27] S. Parekh and J. Walrand (1989) A quick simulation method for excessive backlogs in networks of queues, *IEEE Trans. Automat. Contr.*, 54–66

[28] R.Y. Rubinstein (1992) The 'push out' method for sensitivity analysis of discrete event systems, *Annals of Operations Research*, 39.

[29] R.Y. Rubinstein and A. Shapiro (1993) *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method,* John Wiley & Sons.

[30] Sadowsky, J.S. (1991) Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue, *IEEE Trans. Automat. Contr.*, 36, 1383–1394.

[31] Sadowsky, J.S. and Bucklew, J.S. (1990) On large deviations theory and asymptotically efficient Monte Carlo estimation, *IEEE Trans. Inform. Theory*, 36, 579–588.

[32] Sadowsky, J.S. and Szpankowski W. (1992) The probability of large queue length and waiting times in a heterogeneous multiserver queue. Part I: Tight limits. Manuscript, School of Electrical Engineering and Department of Computer Science, Purdue University, West Lafayette, Indiana 47907 USA.

[33] J. S. Sadowsky and W. Szpankowski W. (1992) The probability of large queue length and waiting times in a heterogeneous multiserver queue. Part II: Positive recurrence and logarithmic limits. Manuscript, School of Electrical Engineering and Department of Computer Science, Purdue University, West Lafayette, Indiana 47907 USA.

[34] J. S. Sadowsky (1993) On the optimality and stability of exponential twisting in Monte Carlo simulation. To appear in *IEEE Transaction on Information Theory,* **IT-39**, 119-128.

[35] J.S. Sadowsky (1994) Monte Carlo Estimation of Large Deviation Probabilities. Manuscript, School of Electrical Engineering and Department of Computer Science, Purdue University, West Lafayette, Indiana 47907 USA.

[36] A. Shwartz and A. Weiss (1992) *Large Deviation for Performance Analysis: Queues, Communication and Computers.* Manuscript, Department of Electrical Engineering, Technion-IIT, Haifa, Israel.

[37] L.J. Stockmeyer (1992) Computational complexity, *Handbooks in OR & MS,* Vol. 3

[38] P. Tsoucas (1992) Rare events in series of queues, *Journal of Appl. Prob.,* 168–175.