Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

IMPORTANCE SAMPLING SIMULATIONS OF PHASE-TYPE QUEUES

Poul E. Heegaard

Department of Telematics
Norwegian University of Science and Technology (NTNU)
N-7491 Trondheim, Norway

Werner Sandmann

Department of Mathematics
Clausthal University of Technology
D-38678 Clausthal-Zellerfeld, Germany

ABSTRACT

Importance sampling is a variance reduction technique that is particularly well suited for simulating rare events and, more specifically, estimating rare event probabilities. Properly applied, it often results in tremendous efficiency improvements compared to direct simulation schemes, but it can also yield unbounded variance increase. Its efficiency and robustness critically rely on a suitable change of the underlying probability measure, which is highly model-dependent. In recent years, significant progress greatly broadened the classes of models successfully accessible by importance sampling, but several model classes still require further investigation. We consider importance sampling simulations of finite capacity queues where interarrival and service times are Erlang distributed. A change of measure is proposed and experimentally studied. Numerical results for loss rates due to buffer overflows indicate that the change of measure provides accurate estimates and appears promising for adaptation to other models involving phase-type distributions.

1 INTRODUCTION

Rare events though occurring with extremely small probability are often highly relevant, because they may have serious consequences such as breakdowns of manufacturing systems, technical defects, system failures in highly reliable systems, bit errors in digital communications, or packet losses in computer and communication networks. When analytical model solution techniques fail, for instance because of the model's complexity or structure, stochastic simulation still remains a viable approach. However, direct simulation of rare events is not effective, because rare events occur too infrequently in simulations to compute reliable statistical estimates in reasonable time. The simulation time to realize robust estimators with desired accuracy such as prescribed confidence interval relative half width must be reduced. Since the accuracy depends on the variance of the simulation estimators, such a simulation speed-up corresponds to variance reduction.

Importance sampling (IS) is the most widespread, potentially most powerful approach to rare event simulation and has been successfully applied in a variety of domains (Heidelberger 1995, Heegaard 1998, Srinivasan 2002, Glasserman 2004, Juneja and Shahabuddin 2006, Asmussen and Glynn 2007, Rubino and Tuffin 2009). Its basic idea is to provoke more of the rare events of interest by changing the underlying probability measure, simulating under the changed measure, and unbiasing the corresponding estimator by the likelihood ratio. The crucial point in a successful application is to find an IS estimator with much smaller variance than the direct simulation estimator. Although an optimal IS estimator always exists, in most cases it is not available and cannot be implemented as it explicitly depends on the unknown quantity to be estimated. Nevertheless, its form suggests potentially good candidates as it is reasonable that an efficient change of measure should be as close as possible to the optimal one and should have at least some properties of the optimal one.

While importance sampling for Markovian queueing models like M/M/1, Markovian tandem networks, or more general Jackson networks has been extensively studied and many insights have been gained, it was not yet systematically investigated for phase-type (PH) queues, that is queues with phase-type distributed interarrival and/or service times. However, phase-type distributions such as Cox, Erlang, Hyper- and Hypoexponential play an important role in the context of queueing models. They are particularly attractive since they allow for quite general models that still obey the Markovian property.

We study single server queues where both interarrival and service times are Erlang distributed, which might serve as a starting point for further advanced studies of other PH queues and queueing networks involving phase-type distributions. The basics of PH queues and IS are given in Sections 2 and 3, respectively. Section 4 provides a change of measure. Numerical results are presented in Section 5. Finally, Section 6 concludes the paper and outlines further research directions.

2 PHASE-TYPE QUEUES

Phase-type queues are queues where interarrival and/or service times are phase-type distributed. The concept of phase-type distributions extends the exponential distribution and is based on the connection with absorption (or similarly first-hitting) times in continuous-time Markov chains.

Definition 1. *A probability distribution on $\mathbb{R}_+ = [0,\infty)$ is a (continuous) phase-type (PH) distribution, iff it is the distribution of the time until absorption in a finite time-homogeneous continuous-time Markov chain (CTMC).*

Consider a CTMC with finite state space $\mathscr{S} = \{1,\dots,n,n+1\}$ and initial distribution $(\underline{\alpha},\alpha_{n+1})$ where the states $1,\dots,n$ are transient and the state $n+1$ is absorbing. Then the generator matrix Q of the chain can be represented as

$$Q = \begin{pmatrix} T & t \\ \underline{0} & 0 \end{pmatrix}, \quad T \in \mathbb{R}^{n\times n}, \quad t \in \mathbb{R}^n \tag{1}$$

and $(\underline{\alpha},T)$ is a representation of the PH distribution. The density, the distribution function and the moments of this PH distribution are given by $f_{(\underline{\alpha},T)}(x) = \underline{\alpha}\exp(Tx)t, F_{(\underline{\alpha},T)}(x) = 1-\underline{\alpha}\exp(Tx)\vec{1}$ and $E[X^i] = (-1)^i i!\ (\underline{\alpha}T^{-i}\vec{1}), x \geq 0, i \in \mathbb{N}$, which implies the Laplace transform $E[e^{-\vartheta X}] = \alpha_{n+1} + \underline{\alpha}(\vartheta I - T)^{-1}t, \vartheta \geq 0$. If there are multiple absorbing states and we are only interested in absorption and not in the particular state where absorption occurs, the absorbing states can be aggregated into one single absorbing class such that the concept applies similarly. Also first-hitting times are covered by modeling the state (or set of states) of interest as an absorbing state (or class).

The exponential distribution with parameter $\lambda > 0$ is the basic building block of any other PH distribution and is itself a PH distribution corresponding to $T = (-\lambda), \underline{\alpha} = (1), t = (\lambda)$. Hence, since state sojourn times in CTMCs are exponentially distributed, PH distributions can be viewed as extensions of the exponential distribution in that they are composed of exponentials. This extension is very general as it in fact covers a wide range of practically relevant distributions. It is well known that the class of PH distributions is closed under convolution, mixtures, minimum and maximum, that it is dense in the class of all probability distributions on $\mathbb{R}_+ = [0,\infty)$, and that all distributions on $\mathbb{R}_+$ can be arbitrarily closely approximated by an appropriate PH distribution. For proofs and many more details on PH distributions and their properties see (Neuts 1981, O'Cinneide 1990, O'Cinneide 1999, Mocanu and Commault 1999).

One major appeal of PH distributions is that they allow for Markovian modeling of systems where the random times involved are more general than exponentially distributed. In particular, it can be readily seen that phase-type queues can be modeled as multi-dimensional CTMCs with a specific structure. With a suitable ordering of states, the generator matrix is block-tridiagonal. Because of the similarity to the generator matrix of ordinary birth-death processes representing M/M/1 queues this gives rise to the notion of quasi-birth-death (QBD) processes, where the generator matrix entries are now vectors and matrices rather than scalars. Approached from a different direction, QBD processes are not restricted to modeling single server phase-type queues but also apply to, e.g., a broad class of Jackson networks.

Numerical solutions of such QBDs can be often tackled by matrix geometric methods as pioneered by Neuts and further developed by himself and others, see (Neuts 1981, Latouche and Ramaswami 1999, Bini et al. 2005). Here, we are interested in rare event simulation using importance sampling as applied to phase-type queues. Our focus in this paper is on $E_k/E_m/1$ queues where E_n denotes the Erlang distribution with n phases, which has the phase-type representation $\underline{\alpha} = (1,0,\dots,0)$ and

$$T = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ 0 & 0 & -\lambda & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \lambda \\ 0 & 0 & 0 & \cdots & -\lambda \end{pmatrix}, \quad t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \lambda \end{pmatrix}.$$

Figure 1 shows a CTMC model of a $E_k/E_m/1/N$ queue. The interarrival time is Erlang-k with intensity $\lambda > 0$, while the service time is Erlang-m with intensity $\mu > 0$. In gray the meta states states representing the number of resources (customers) in the system (server and queue) are indicated.

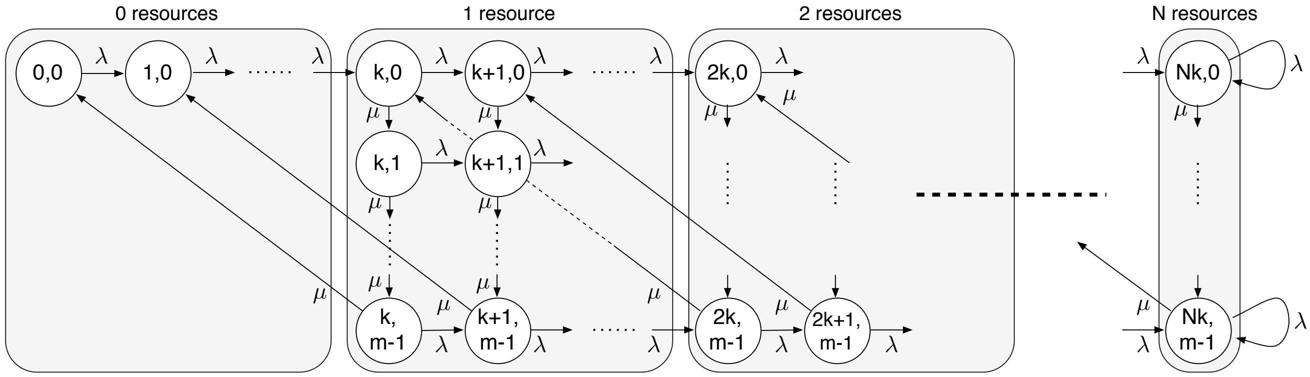


Figure 1: CTMC model for a single server queue with phase type distribution

3 IMPORTANCE SAMPLING

Consider two probability measures P and P^* on a measurable space $(\Omega, \mathscr{A})$. If P is absolutely continuous with respect to P^*, that is $\forall A \in \mathscr{A}: \; P^*(A) = 0 \;\Rightarrow\; P(A) = 0$, then the Radon-Nikodym derivative $L = dP/dP^*$ exists and importance sampling (IS) exploits that for any random variable Y on $(\Omega, \mathscr{A})$,

$$E_P[Y] = \int Y(\omega)dP = \int Y(\omega)L(\omega)dP^* = E_{P^*}[YL], \tag{2}$$

where E_P and E_{P^*} denote expectations with respect to the probability measures P and P^*, respectively. Hence, when a simulation is performed under P^* the sample mean of YL is an unbiased estimator for $E_P[Y]$. In the context of IS, the probability measure P^* is called the *IS measure*, L is called the *likelihood ratio* and choosing P^* in place of P is referred to as the *change of measure*. IS applies to very general settings and in particular to stochastic processes.

For Markov chains, either in discrete or continuous time, the probability measures P and P^* are path distributions and absolute continuity corresponds to the condition that all paths containing the event of interest that are possible under the original measure P must remain possible under the changed measure P^*. In continuous time this can be obviously achieved by the condition that for all positive transition rates in the original model the corresponding transition rates under IS are positive. Similarly, this condition applies for transition probabilities in case of discrete time. For details on IS for stochastic processes including Markov processes and generalized semi-Markov processes (GSMP) see (Glynn and Iglehart 1989).

Estimating the probability $\gamma := P(\mathscr{R})$ of a rare event $\mathscr{R}$ by IS is performed via $P(\mathscr{R}) = E_P[I(\mathscr{R})] = E_{P^*}[I(\mathscr{R})L]$ such that the sample mean of $I(\mathscr{R})L$ under probability measure P^* provides an unbiased estimator for $P(\mathscr{R})$. From

$$\sigma^2_{P^*}(I(\mathscr{R})L) = E_{P^*}[(I(\mathscr{R})L - \gamma)^2] = E_{P^*}[(I(\mathscr{R})L^2] - \gamma^2, \tag{3}$$

where $\sigma^2_{P^*}$ denotes the variance under P^*, it can be readily seen that the optimal zero-variance IS measure is determined by the original probability conditioned on the occurrence of the rare event, $P^*(A) = P(A|\mathscr{R}) = P(A \cap \mathscr{R})/P(\mathscr{R})$ for all events $A \in \mathscr{A}$. In other words, the choice of P^* such that $P^*(A) = P(A)/P(\mathscr{R})$ for all $A \subseteq \mathscr{R}$ and $P^*(A) = 0$ otherwise, provides a zero-variance estimator for $P(\mathscr{R})$. Unfortunately, it explicitly depends on the unknown rare event probability and is therefore not available for implementation. Even if it was it may take a form that makes sampling from it difficult or impossible.

Therefore, the key is to find a change of measure such that the corresponding IS estimator possesses certain efficiency or robustness properties, which are typically focused on the asymptotic behavior of estimators as the probability of the rare event approaches zero. The most prominent such properties are bounded relative error and asymptotic optimality but many others are reasonable, see, e.g., (Tuffin 1999, Blanchet et al. 2007, Sandmann 2007, L'Ecuyer et al. 2009).

Suppose we want to estimate the probability $\gamma_b := P(\mathscr{R}_b)$ of a rare event $\mathscr{R}_b$ that depends on some 'rarity parameter' $b \in \mathbb{N}$ such as, e.g., the overflow of a buffer of size b. Hence, we consider a sequence of events $\mathscr{R}_1, \mathscr{R}_2, \ldots$ where $P(\mathscr{R}_1) \geq P(\mathscr{R}_2) \geq \cdots$ and $\lim_{b\to\infty} P(\mathscr{R}_b) = 0$. Denote the unbiased IS estimator for γ_b by $\hat{\gamma}_b$ and the associated likelihood ratio by L_b.

Definition 2. *The sequence* $(\hat{\gamma}_b)_{b\in\mathbb{N}}$ *of IS estimators possesses bounded relative error (BRE), or equivalently bounded relative variance (BRV), iff*

$$\exists c > 0: \lim_{b\to\infty} \frac{\sigma_{P^*}(\hat{\gamma}_b)}{\gamma_b} \leq c < \infty. \tag{4}$$

Definition 3. *The sequence* $(\hat{\gamma}_b)_{b\in\mathbb{N}}$ *of IS estimators possesses asymptotic optimality (AO), also known as logarithmic efficiency (LE), iff*

$$\lim_{b\to\infty} \frac{\ln E_{P^*}[(I(\mathscr{R})L^2]}{\ln \gamma_b} = 2. \tag{5}$$

Note that BRE/BRV implies AO/LE but in general the converse is not true such that AO/LE is weaker than BRE/BRV.

4 CHANGE OF MEASURE IN $E_k/E_m/1/N$ MODELS

We consider $E_k/E_m/1/N$ models, where the change of measure is done by choosing alternative rates λ^* and μ^* in place of λ and μ. Note that with IS all rates may be state dependent even if they are state independent in the original system. Although it is known that in general an efficient change of measure should be state dependent there are cases where a state independent change of measure provides accurate estimates. In some cases, a state independent change of measure can even provide BRE or AO. In any case, the great advantage of a state independent change of measure is the simplicity of implementation in simulators.

We show that state independent changes of measure are efficient in IS simulations of $E_k/E_m/1/N$ queues. Our approach to obtain good state independent changes of measure is focusing on the likelihood ratio along cycles within sample paths, which is known as a critical determinant of IS estimators' efficiency. For the optimal zero-variance estimator the likelihood ratio is constant on all paths containing the rare event of interest, which implies that the likelihood ratio on all cycles on such paths is equal to one. According to (Juneja 2001), if the likelihood ratio is constant on all paths containing $\mathscr{R}_b$ and $P^*(\mathscr{R}_b)$ is lower bounded by some constant $\beta > 0$ ($\mathscr{R}$ is not rare under P^* in the sense that it does not converge to zero as b converges to infinity), then the corresponding IS estimators possess BRE and AO. This has motivated the cyclic approach in (Juneja 2001) and a similar idea based on the so-called likelihood ratio conditions in (Sandmann 2004a, Sandmann 2004b).

In particular, a state independent change of measure for Markovian models has BRE/BRV as long as the rare event is lower bounded by some constant and all simulated cycles of the underlying CTMC or its embedded DTMC, respectively, contribute to the likelihood ratio with a factor of one. Obviously, the likelihood ratio condition alone is not sufficient for BRE since it trivially holds for the original measure as IS measure. But it seems reasonable to enforce it as far as possible.

We consider cycles in the CTMC representation of $E_k/E_m/1/N$ models depicted in Figure 1. A cycle is defined by a sequence of states where the first and last state are identical and all other states are different, $\tilde{X} = (X_0, X_1, \ldots, X_n), n > 0$ where $X_0 = X_n$ and $X_i \neq X_j$ for $i, j = 1, \ldots, n, i \neq j$. In an $E_k/E_m/1$ model any shortest cycle consists of k arrival phases and m service phases and all other cycles are built of multiple shortest cycles such that considering shortest cycles suffices. We call a cycle an *interior cycle* if it does not contain any of the states $(i, 0)$, $i = 0, \ldots, k-1$, that is states of the boundary meta state representing an empty system. Observe that all states in the interior cycles have either a phase change in the arrival or service process as their next state, and hence the outgoing rate is $\lambda + \mu$ for all states. The probability (considering the embedded DTMC) of shortest interior cycles $\tilde{X}$ in the original model is

$$p_{\tilde{X}} = \left(\frac{\lambda}{\lambda+\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^m \tag{6}$$

and the corresponding probability of these cycles in the IS model is

$$p^*_{\tilde{X}} = \left(\frac{\lambda^*}{\lambda^*+\mu^*}\right)^k \left(\frac{\mu^*}{\lambda^*+\mu^*}\right)^m. \tag{7}$$

Now, without loss of generality we let $\lambda+\mu=1$. Futhermore, we let $\lambda^*+\mu^*=1$ to obtain a unique solution. Then, the likelihood ratio L of $\tilde{X}$ is

$$L_{\tilde{X}} = \frac{p_{\tilde{X}}}{p^*_{\tilde{X}}} = \frac{\lambda^k \mu^m}{(\lambda^*)^k(\mu^*)^m}. \tag{8}$$

From the equations above we see that in an $E_k/E_m/1/N$ model the simulated cycles have a likelihood ratio $L_{\tilde{X}}=1$ for all interior cycles when

$$\begin{aligned} \lambda^k\mu^m &= (\lambda^*)^k(\mu^*)^m \\ &\Updownarrow \\ \lambda^k(1-\lambda)^m &= (\lambda^*)^k(1-\lambda^*)^m \end{aligned} \tag{9}$$

For $k=m$, this yields

$$\begin{aligned} \lambda^k\mu^k &= (\lambda^*)^k(\mu^*)^k \\ &\Updownarrow \\ \lambda\mu &= \lambda^*\mu^* \end{aligned} \tag{10}$$

and, assuming that $\lambda<\mu$ and $\lambda^*>\mu^*$, we obtain $\lambda^*=\mu, \mu^*=\lambda$, which is the well known asymptotically optimal state independent change of measure for the M/M/1 queue that also possesses BRE. For $k\neq m$, equation (9) can be rearranged to

$$\lambda^* = \lambda\left(\frac{1-\lambda}{1-\lambda^*}\right)^{m/k} \tag{11}$$

which has no simple closed form solution. It can be solved numerically but in order to simplify implementation in a simulator and to facilitate a potentially generalizable way for models where a numerical solution is hard to obtain, we also propose an approximation that turns out to work particularly well when λ is close to zero. The approximation is based on the assumption that a possible or approximate solution λ^*_a to (9) might be of the form

$$\begin{matrix} \lambda^k \propto (\mu^*_a)^m \\ \mu^m \propto (\lambda^*_a)^k \end{matrix} \;\Leftrightarrow\; \begin{matrix} \lambda^*_a = C\cdot\mu^{m/k} \\ \mu^*_a = C\cdot\lambda^{k/m} \end{matrix}$$

With the normalization constant $C=(\lambda^{k/m}+\mu^{m/k})^{-1}$ we obtain $\lambda^*_a+\mu^*_a=1$ and apply the approximation

$$\lambda^* \approx \lambda^*_a := \frac{\mu^{m/k}}{\mu^{m/k}+\lambda^{k/m}}. \tag{12}$$

Figure 2 shows a comparison between the numerical solution λ^*_n of (9) and the approximation λ^*_a in (12). They are plotted as a function of λ. The upper (blue) is the numerical solution while the lower (red) is the approximation. Observe that the closer λ is to zero the better the approximation, that is the approximation approaches the numerical solution in light traffic as the traffic intensity becomes small.

Note that not even in the $E_k/E_m/1/N$ models as discussed in this paper all cycles will obey likelihood ratios equal to one by applying the state independent changes of measure outlined above. Cycles that include a state in $\{(i,0)\}$, $i=0,\cdots,k-1$ will deviate from the conditions but fortunately they are few. Furthermore, observe that the transition intensities λ of $(i,0)$, $i=0,\cdots,k-1$ are not changed in the IS model, $\lambda^*=\lambda$ because these states only have one outgoing transition (see Figure 1). In this sense one might even refer to a state dependent change of measure as it depends on whether the model state is within or without the boundary meta state representing the empty system. However, of course, it is essentially state independent and we prefer the interpretation that IS is switched off when the system is empty. In the following section the simulation efficiency is demonstrated using different state independent changes of measures, including both the numerical solution λ^*_n and the approximation λ^*_a. Numerical examples indicate that the proposed change of measure is efficient.

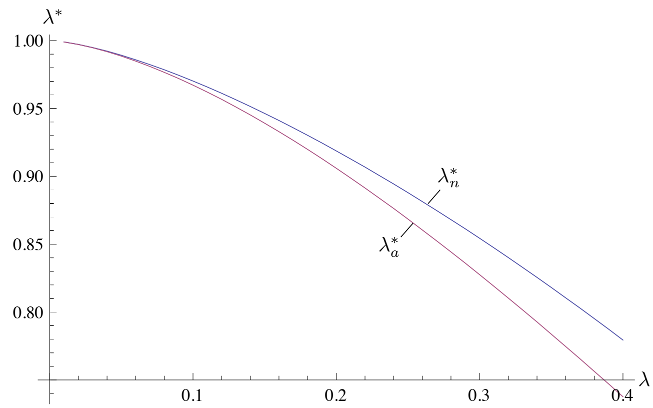


Figure 2: Comparison of numerical λ_n^* (blue, upper) and approximate λ_a^* (red, lower)

5 NUMERICAL EXAMPLES

We have conducted numerical examples for $E_k/E_m/1/N$ models with different combinations of k and m and with variable queue length N. The rare event of interest is the probability of buffer overflow or packet loss in a communication network. All simulation experiments consist of 20 replications of 20,000 regenerative cycles. All experiments have been conducted on a MacBook Pro (2.16 Ghz Core 2 Duo, 2 GB RAM, OS X 10.5) using the programming language Simula (Kirkerud 1989) with the DEMOS class (Birtwistle 1997).

5.1 Importance Sampling Parameters

Table 1 summarizes the different combinations of parameters that have been studied by simulation. The table includes the state independent changes of measure, both λ_n^* and λ_a^*. The results in the following include a selection that represents typical behavior that has been observed and the selection is indicated by light gray in the table of parameters. In all simulations the sensitivity of the change of measure around the proposed "optimal" value is studied empirically.

Table 1: Importance sampling parameters with $\lambda = 0.2$ and $\mu = 0.8$

k	m	λ^*	μ^*	λ_a^*	μ_a^*
1	1	0.800	0.200	0.800	0.200
2	2	0.800	0.200	0.800	0.200
2	3	0.626	0.374	0.677	0.323
3	2	0.919	0.081	0.906	0.094
3	3	0.800	0.200	0.800	0.200
5	10	0.488	0.512	0.589	0.411
5	5	0.800	0.200	0.800	0.200
10	5	0.966	0.034	0.967	0.043
10	10	0.800	0.200	0.800	0.200

5.2 Results

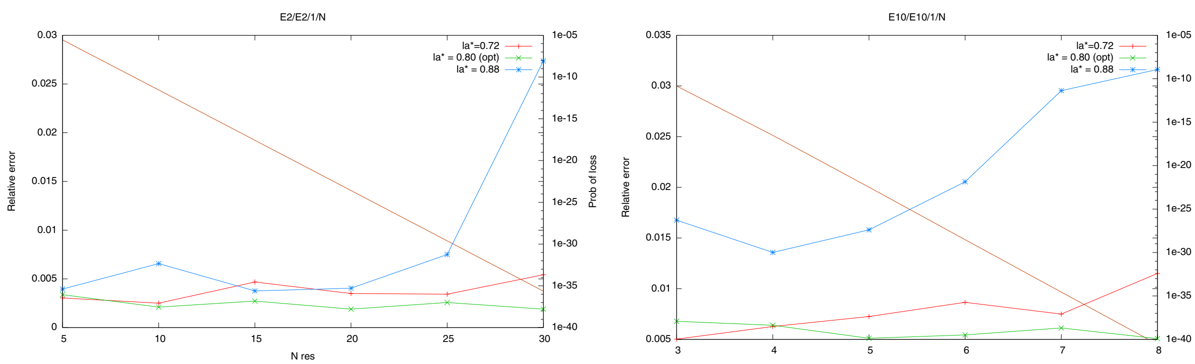
The first set of simulations assumes the same number of phases in the interarrival and service times, i.e. $k = m = 2$ and $k = m = 10$. The simulation results given in Tables 2 and 3 show the estimated loss ratio and the standard error for different values of λ^*. The results in gray are from using the "optimal" change of measure. In Figure 3 the relative error is given as a function of the finite queue size N, while Figure 4 shows the relative error times the CPU time consumption as a function of the finite queue size N. The straight red line is the loss ratio plotted against the right axis.

Table 2: Loss ratio mean and standard error in $E_2/E_2/1/N$-queue for different λ^*

λ^*	$N=5$		$N=10$		$N=15$		$N=20$		$N=25$		$N=30$	
	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.
0.72	2.83	0.008.6	2.85	0.0072	2.69	0.0126	2.59	0.0091	2.45	0.0084	2.25	0.0122
0.80	2.82	0.0095	2.81	0.0059	2.68	0.0073	2.57	0.0049	2.43	0.0063	2.33	0.0044
0.88	2.83	0.0112	2.80	0.0184	2.71	0.0102	2.47	0.0100	2.28	0.0171	2.32	0.0635
	[×e-06]		[×e-12]		[×e-18]		[×e-24]		[×e-30]		[×e-36]	

Table 3: Loss ratio mean and standard error in $E_{10}/E_{10}/1/N$-queue for different λ^*

λ^*	$N=3$		$N=4$		$N=5$		$N=6$		$N=7$		$N=8$	
	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.
0.72	1.42	0.007	3.08	0.019	3.35	0.024	3.04	0.026	2.75	0.021	2.65	0.031
0.80	1.44	0.010	3.08	0.020	3.31	0.017	2.98	0.016	2.79	0.017	2.66	0.014
0.88	1.45	0.0024	3.05	0.041	3.27	0.052	2.94	0.060	2.60	0.077	2.40	0.076
	[×e-11]		[×e-17]		[×e-23]		[×e-29]		[×e-35]		[×e-41]	



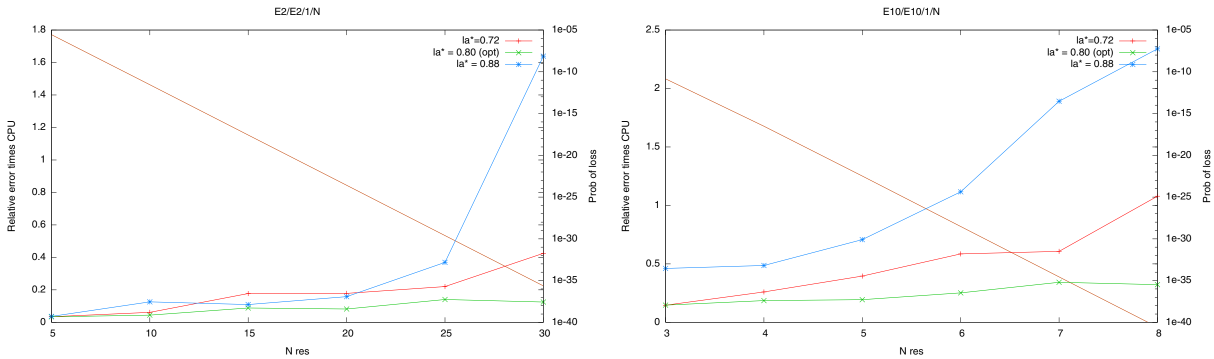
(a) $k=2,\ m=2$ (b) $k=10,\ m=10$

Figure 3: The relative error for different N when $k=m$

The second set of simulations includes $k=3$, $m=2$ and $k=5$, $m=10$. The simulation results are given in Tables 4 and 5. The results in gray are from using the numerically obtained change of measure while the light gray are results using the approximation. The estimated loss ratio and the standard error are included. In Figure 5 the relative error is given as a function of the finite queue size N, while Figure 6 shows the relative error times the CPU time consumption as a function of the finite queue size N.

5.3 Key Observations

The results clearly show that the "optimal" state independent change of measure is efficient in the case with $k=m$ in an $E_k/E_m/1/N$ queue. The relative error is insensitive to the rarity of the problem, while the CPU consumptions times relative error show a slow linearly increase. The sensitivity experiments indicate that it seems to be indeed the optimal state independent change of measure in the sense that it is superior for all studied cases, including the cases where $k \neq m$. The proposed approximation for obtaining the IS parameters works particularly well when the λ is close to one.



(a) $k = 2,\ m = 2$ (b) $k = 10,\ m = 10$

Figure 4: The CPU consumption times relative error for different N when $k = m$

Table 4: Loss ratio mean and standard error in $E_3/E_2/1/N$-queue for different λ^*

λ^*	$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 25$		$N = 30$	
	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.
0.800	1.82	0.0061	2.40	0.0111	3.12	0.0356	4.84	0.0370	4.13	0.0273	4.06	0.2560
0.880	1.81	0.0041	2.47	0.0057	2.89	0.0059	3.37	0.0076	3.92	0.0181	4.68	0.0097
0.906	1.81	0.0034	2.47	0.0049	2.89	0.0065	3.43	0.0073	3.94	0.0077	4.60	0.0101
0.919	1.81	0.0054	2.46	0.0045	2.91	0.0054	3.37	0.0076	4.00	0.0071	4.59	0.0092
0.940	1.81	0.0042	2.46	0.0056	2.88	0.0070	3.37	0.0106	4.00	0.0109	4.69	0.0116
0.960	1.82	0.0085	2.43	0.0075	2.85	0.0104	3.41	0.0101	4.10	0.0181	4.25	0.0137
	[×e-09]		[×e-19]		[×e-29]		[×e-39]		[×e-49]		[×e-59]	

6 CONCLUSION AND FURTHER RESEARCH

We have studied importance sampling for estimating the probability of buffer overflows or loss rates in $E_k/E_m/1$ systems. A state independent change of measure strategy assuring that the likelihood ratio equals one on most cycles (all except for cycles involving states representing the empty system) has been obtained with the proviso that arrival phase rates should be increased while service phase rates should be decreased.

For $k = m$ the obtained change of measure generalized in a straightforward manner the asymptotically optimal exponential change of measure for $M/M/1$ (i.e. $E_1/E_1/1$) in which the arrival and the service rate are switched. For $k \neq m$ the IS rates are obtained numerically. In both cases, our numerical results indicate the efficiency of the change of measure. Empirical sensitivity studies suggest that the change of measure might be the optimal state independent one. We have also proposed an approximation for the case with $k \neq m$ in place of numerically obtaining the IS rates, which works well, in particular for λ^* close to one. It seems that the relative error remains is bounded, and the relative error times CPU consumption increases linearly with the rarity (and system capacity N)

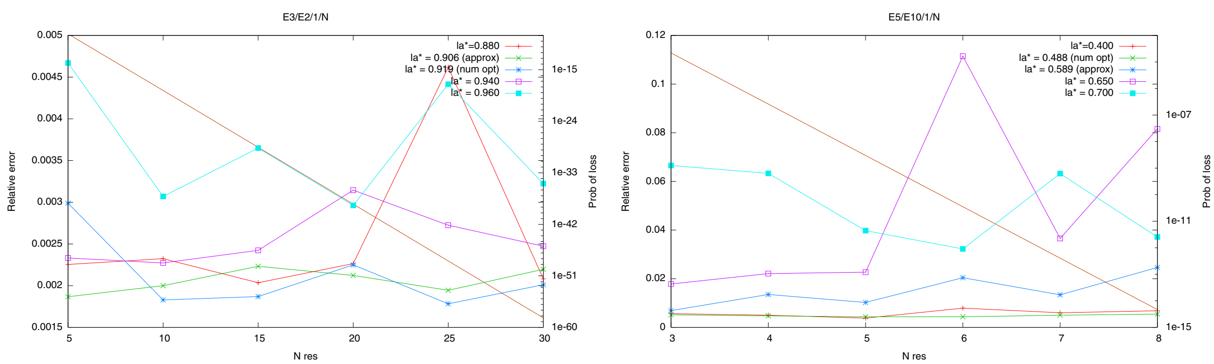
Theoretical investigation of the proposed change of measure is desirable such as proving or disproving its robustness properties and its optimality within the class of state independent changes of measure. Further research includes extensions to other phase-type distributions such as hypoexponential, hyperexponential and Cox distributions as well as extensions to multi-server and multi-class queues and queueing networks starting with tandem networks.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation: Algorithms and analysis*. Springer.
Bini, D. A., G. Latouche, and B. Meini. 2005. *Numerical methods for structured Markov chains*. Oxford University Press.
Birtwistle, G. 1997. *Demos - a system for discrete event modelling on simula*. University of Leeds.

Table 5: Loss ratio mean and standard error in $E_5/E_{10}/1/N$-queue for different λ^*

λ^*	$N=3$		$N=4$		$N=5$		$N=6$		$N=7$		$N=8$	
	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.	mean	std.err.
0.400	2.28	0.0128	2.74	0.0137	3.12	0.0118	3.58	0.0283	4.41	0.0265	4.80	0.0328
0.488	2.18	0.0112	2.60	0.0123	2.99	0.0130	3.50	0.0152	4.05	0.0203	4.67	0.0255
0.589	2.12	0.0146	2.61	0.0353	2.93	0.0300	3.35	0.0684	3.77	0.0504	4.79	0.1180
0.650	2.16	0.0385	2.50	0.0553	2.54	0.0576	4.01	0.4470	3.31	0.1210	3.42	0.2790
0.700	2.54	0.1690	2.78	0.1760	1.94	0.0771	1.84	0.0593	1.74	0.1100	1.36	0.0505
0.800	1.63	0.0609	1.43	0.0606	0.95	0.2170	2.32	0.4640	0.43	0.0604	0.13	0.0010
	[×e-05]		[×e-07]		[×e-09]		[×e-11]		[×e-13]		[×e-15]	



(a) $k=3,\ m=2$ (b) $k=5,\ m=10$

Figure 5: The relative error for different N when $k \neq m$

Blanchet, J. H., P. W. Glynn, P. L'Ecuyer, W. Sandmann, and B. Tuffin. 2007. Asymptotic robustness of estimators in rare-event simulation. In *Proc. 2007 INFORMS Simulation Society Research Workshop*.

Glasserman, P. 2004. *Monte Carlo methods in financial engineering*. Springer.

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

Heegaard, P. E. 1998. *Efficient simulation of network performance by importance sampling*. Ph. D. thesis, Norwegian University of Science and Technology.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5 (1): 43–85.

Juneja, S. 2001. Importance sampling and the cyclic approach. *Operations Research* 49 (6): 900–912.

Juneja, S., and P. Shahabuddin. 2006. Rare event simulation techniques: An introduction and recent advances. In *Simulation*, ed. S. G. Henderson and B. L. Nelson, Handbooks in Operations Research and Management Science, 291–350. Amsterdam, The Netherlands: Elsevier. Chapter 11.

Kirkerud, B. 1989. *Object-oriented programming with simula*. Addison Wesley.

Latouche, G., and V. Ramaswami. 1999. *Introduction to matrix analytic methods in stochastic modeling*. SIAM.

L'Ecuyer, P., J. H. Blanchet, B. Tuffin, and P. W. Glynn. 2009. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*. to appear.

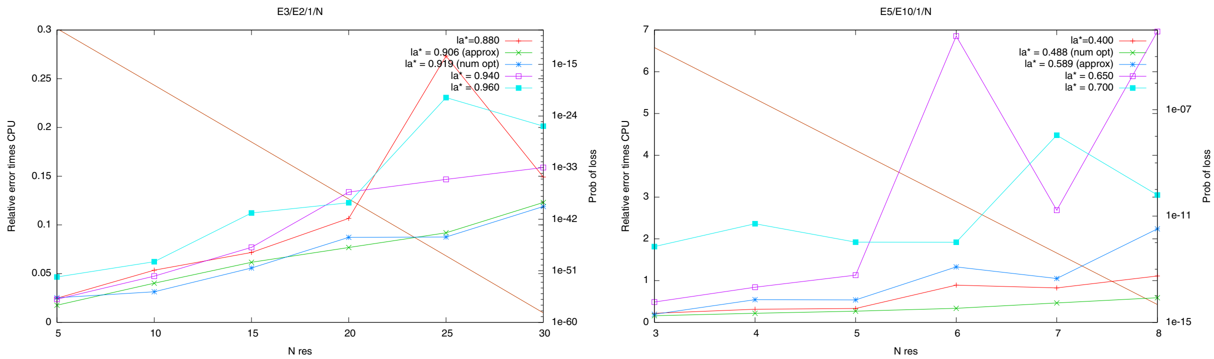
Mocanu, S., and C. Commault. 1999. Sparse representations of phase-type distributions. *Stochastic Models* 15 (4): 759–778.

Neuts, M. F. 1981. *Matrix-geometric solutions in stochastic models: An algorithmic approach*. John Hopkins University Press.

O'Cinneide, C. A. 1990. Characterization of phase-type distributions. *Stochastic Models* 6 (1): 1–57.

O'Cinneide, C. A. 1999. Phase-type distributions: Open problems and a few properties. *Stochastic Models* 15 (4): 731–757.

Rubino, G., and B. Tuffin. (Eds.) 2009. *Rare event simulation using Monte Carlo methods*. John Wiley & Sons.



(a) $k=3,\ m=2$ (b) $k=5,\ m=10$

Figure 6: The CPU time consumption times the relative error for different N when $k \neq m$

Sandmann, W. 2004a. Fast simulation of excessive population size in tandem Jackson networks. In *Proc. 12th IEEE Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS*, 347–354.
Sandmann, W. 2004b. State-independent importance sampling for Markovian tandem queues. In *Proc. 5th Int. Workshop on Rare Event Simulation and related Combinatorial Optimization Problems, RESIM/COP*.
Sandmann, W. 2007. Efficiency of importance sampling estimators. *Journal of Simulation* 1 (2): 137–145.
Srinivasan, R. 2002. *Importance sampling: Applications in communications and detection*. Springer.
Tuffin, B. 1999. Bounded normal approximation in simulations of highly reliable Markovian systems. *Journal of Applied Probability* 36 (4): 974–986.

AUTHOR BIOGRAPHIES

POUL E. HEEGAARD is Associate Professor and Head of Department at Department of Telematics, Norwegian University of Science and Technology (NTNU). Heegaard received his Siv.ing. (M.S.E.E. in '89) and his Dr. Ing. (PhD in '98) degrees from the University of Trondheim (now NTNU). Heegaard was a Research Scientist and Senior Scientist at SINTEF Telecom and Informatics (1989-1999) and Senior Research Scientist at Telenor R&I (1999-2009). In the academic year 2007/08 he was a visiting professor at Duke University, Durham, NC. His research interests cover performance, dependability and survivability evaluation of communication systems. Special interests are rare event simulation techniques, IP network monitoring and modeling, and distributed, autonomous and adaptive management and routing in communication networks and services. His e-mail address is poul.heegaard@item.ntnu.no, and his web page can be found at <www.item.ntnu.no/ poulh/>.

WERNER SANDMANN is an Associate Professor of Stochastics in the Department of Mathematics at Clausthal University of Technology (Germany). He studied Computer Science and Mathematics at the University of Bonn (Germany) where he received his diploma in Computer Science and his PhD (Dr. rer. nat.) in 1998 and 2004, respectively. From 1998 to 2003, he was a Scientific Member of the Computer Science Department at the University of Bonn. Since 2004, he has been an Assistant Professor of Computer Science at the University of Bamberg (Germany) where he is now on a temporary leave for his current position. His research interests are in applied and computational stochastics, in particular stochastic modeling and system analysis with applications to performance evaluation of computer and communications networks, operations research, and systems biology. His e-mail address is werner.sandmann@tu-clausthal.de, and his web page can be found at <http://www.math.tu-clausthal.de/ ws09/>.