MANAGEMENT SCIENCE
Vol. 50, No. 7, July 2004, pp. 883–895
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5007 | 0883

informs

DOI 10.1287/mnsc.1030.0139

A Fast Cross-Entropy Method for Estimating Buffer Overflows in Queueing Networks

P. T. de Boer
Department of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, ptdeboer@cs.utwente.nl

D. P. Kroese
Department of Mathematics, University of Queensland, Brisbane 4072, Australia, kroese@maths.uq.edu.au

R. Y. Rubinstein
Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, ierrr01@ie.technion.ac.il

In this paper, we propose a fast adaptive importance sampling method for the efficient simulation of buffer overflow probabilities in queueing networks. The method comprises three stages. First, we estimate the minimum cross-entropy tilting parameter for a small buffer level; next, we use this as a starting value for the estimation of the optimal tilting parameter for the actual (large) buffer level. Finally, the tilting parameter just found is used to estimate the overflow probability of interest. We study various properties of the method in more detail for the $M/M/1$ queue and conjecture that similar properties also hold for quite general queueing networks. Numerical results support this conjecture and demonstrate the high efficiency of the proposed algorithm.

Key words: importance sampling; rare events; cross-entropy; queueing networks; simulation
History: Accepted by Paul Glasserman, stochastic models and simulation; received January 14, 2002. This paper was with the authors 6 months for 1 revision.

1. Introduction

The performance of computer and communication systems is often characterized by the probability of certain rare events. For example, the cell loss probability in asynchronous transfer mode (ATM) switches should typically be less than 10^{-9}. The performance of such systems is frequently studied through simulation. However, estimation of rare event probabilities with naive Monte Carlo techniques requires a prohibitively large number of trials in most interesting cases. One way to deal with this problem is to use importance sampling (IS). The main idea of IS, when applied to rare events, is to make its occurrence more frequent, or to "speed up" the simulation. Technically, IS aims to select a probability distribution (change of measure) that minimizes the variance of the IS estimator. Finding the right change of measure is often described by a large deviation result. This type of analysis is feasible only for relatively simple models. See also Asmussen and Rubinstein (1995) and Heidelberger (1995) for surveys and Parekh and Walrand (1989) and Frater et al. (1991) for specific results regarding queueing networks.

Because of the difficulty of analytically finding the right change of measure, several adaptive approaches have been proposed to do this. In such approaches, a simulation (under a not-yet optimal change of measure) is used to estimate what change of measure would produce a smaller (or minimal) variance, after which a new simulation is run under that change of measure. This may need to be iterated many times before the optimal change of measure has been approximated sufficiently well. The optimization step can be based on stochastic optimization techniques (al-Qaq et al. 1995; Devetsikiotis and Townsend 1993a, b) or on a more direct calculation of the optimal parameters (Lieber et al. 1997, Rubinstein 1997). Rubinstein (1999) proposes to minimize the Kullback-Leibler distance or cross-entropy (CE) instead of the estimator variance; typically this leads to explicit calculations for the new parameters rather than numerical minimization. As an aside, an attractive feature of the CE method is that it can be readily modified for solving NP-hard combinatorial optimization problems (see Alon et al. 2005; Rubinstein 1999, 2001a, b, 2002).

In this paper, we investigate an adaptive IS algorithm for the efficient simulation of buffer overflow probabilities in queueing systems, based on the CE technique discussed above. In contrast to earlier algorithms, the present one needs only three stages: In the pilot stage, we estimate the minimum CE tilting parameter for a small buffer level; next, we use this as a starting value for the estimation of the optimal

tilting parameter for the actual (large) buffer level. Finally, the tilting parameter just found is used to estimate the overflow probability of interest.

The reason why the three-stage approach works well (for arbitrary overflow levels) is that under the initial change of measure, the buffer process is unstable; moreover, this change of measure is "close" to the change of measure for the second stage. In other words, the initial tilting vector is in some sense a "good" tilting vector. We investigate these two properties, which we will call the *instability property* and the *robustness property* in more detail for the $M/M/1$ queue. A third property is the *CE optimality property*: The change of measure found using CE is close to the one that minimizes the variance. We hypothesize that these properties hold in more general networks as well. Numerical results support this conjecture and demonstrate the high efficiency of the proposed algorithm.

Compared to earlier work on IS for queueing models, this method differs in the following ways: The method from Parekh and Walrand (1989) needs a rather extensive analysis for every new model; our method is adaptive, thus obviating the need for such an analysis. This is important, for example, for integration into computer-simulation tools. In Frater et al. (1991), those calculations are much simplified, but these simplified calculations only apply to models where many of the distributions are exponential; our method does not have this limitation. Compared to the adaptive methods from al-Qaq et al. (1995) and Devetsikiotis and Townsend (1993a, b), our method needs far fewer iterations, typically just three. In de Boer (2000), de Boer et al. (2000), and de Boer and Nicola (2001), a CE-based method using a state-dependent change of measure is described. That method has the significant advantage of being able to handle models (such as those discussed in Glasserman and Kou 1995), for which state-*in*dependent tilting does not work well. However, the disadvantages of that method are greater complexity, larger number of iterations, and limitation to Markovian models. In situations where state-dependent tilting is not necessary, the method presented here is much simpler and faster.

The rest of this paper is organized as follows. In §2, we summarize the main ideas behind the adaptive approach to IS. In §3, we formulate the simulation model and give the main algorithm for simulating overflows in queueing networks. A closer investigation of the $M/M/1$ queue, with, to our knowledge, various new results, is given in §4. In §5, we demonstrate numerically the effectiveness of the algorithm by investigating various queueing models, and in §6 concluding remarks are given. Finally, some auxiliary results and proofs are given in the appendices.

2. Importance Sampling and the Cross-Entropy Method

In this section, we briefly review the ideas behind IS and the CE method. For details, see Rubinstein and Melamed (1998) and Rubinstein (1999).

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector taking values in some (measurable) space $\mathscr{X}$. Let $\{f(\cdot; \mathbf{v})\}$ be a family of probability densities on $\mathscr{X}$, with respect to some (unspecified) base measure. Here, $\mathbf{v}$ is a real-valued parameter (vector).

Let H be some (measurable) real function on $\mathscr{X}$. Suppose we wish to estimate, via simulation,

$$\gamma_{\mathbf{v}} := \mathbb{E}_{\mathbf{v}} H(\mathbf{X}),$$

where $\mathbb{E}_{\mathbf{v}}$ denotes expectation under $f(\cdot; \mathbf{v})$. In this paper, we will be mostly concerned with functions H that are indicators of certain events; for example $H(\mathbf{X}) = I_A$, with $A = \{\mathbf{X} \in \mathscr{X}_0\}$ for some subset $\mathscr{X}_0 \subset \mathscr{X}$. When the probability of A is very small, we say that A is a *rare event*.

The easiest way to estimate $\gamma_{\mathbf{v}}$ is to use crude Monte Carlo simulation: Draw a random sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ from $f(\cdot; \mathbf{v})$; then $(1/N)\sum_{i=1}^{N} H(\mathbf{X}^{(i)})$ is an unbiased estimator of $\gamma_{\mathbf{v}}$. However, this poses serious problems when H is the indicator of a rare event. In that case, a large simulation effort is required to estimate $\gamma_{\mathbf{v}}$ accurately.

An alternative is to use IS simulation: Draw a random sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ from $f(\cdot; \tilde{\mathbf{v}})$; then

$$\frac{1}{N}\sum_{i=1}^{N} H(\mathbf{X}^{(i)}) W(\mathbf{X}^{(i)}; \mathbf{v}, \tilde{\mathbf{v}}) \tag{1}$$

with *likelihood ratio*

$$W(\mathbf{X}; \mathbf{v}, \tilde{\mathbf{v}}) := \frac{f(\mathbf{X}; \mathbf{v})}{f(\mathbf{X}; \tilde{\mathbf{v}})}$$

is an unbiased estimator of $\gamma_{\mathbf{v}}$. We say that we perform the simulation under a *change of measure* parameterized by the *tilting* parameter (vector) $\tilde{\mathbf{v}}$. The aim is now to find an optimal tilting parameter ${}_*\mathbf{v}$ such that the variance, or equivalently, the second moment, of the IS estimator is minimal. In other words, we wish to find

$${}_*\mathbf{v} = \arg\min_{\tilde{\mathbf{v}}} \mathbb{E}_{\tilde{\mathbf{v}}}[H(\mathbf{X}) W(\mathbf{X}; \mathbf{v}, \tilde{\mathbf{v}})]^2. \tag{2}$$

More generally, again using the principle of IS, this is equivalent to finding

$${}_*\mathbf{v} = \arg\min_{\tilde{\mathbf{v}}} \mathbb{E}_{\mathbf{v}_j} H^2(\mathbf{X}) W(\mathbf{X}; \mathbf{v}, \tilde{\mathbf{v}}) W(\mathbf{X}; \mathbf{v}, \mathbf{v}_j), \tag{3}$$

for any tilting parameter $\mathbf{v}_j$.

An analytic expression for the optimal tilting parameter ${}_*\mathbf{v}$ is typically not available. However, it

can be estimated by minimizing, possibly numerically, the estimator of the expectation in (3), leading to the approximation

$$\mathbf{v}_{j+1} = \arg\min_{\tilde{\mathbf{v}}} \sum_{i=1}^{N} H^2(\mathbf{X}^{(i)}) W(\mathbf{X}^{(i)}; \mathbf{v}, \tilde{\mathbf{v}}) W(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j), \tag{4}$$

where $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ is a random sample from $f(\cdot; \mathbf{v}_j)$. This formula forms the basis of an iterative scheme to estimate the true optimal tilting parameter.

2.1. Cross-Entropy Method

The evaluation of (4) in general involves numerical optimization, which may be quite time consuming because it requires repeated evaluation of all N samples. By replacing (2) with its CE equivalent, as introduced in Rubinstein (1999), typically (4) is replaced by an expression that can be solved analytically; that is, the updating rules for $\mathbf{v}_{j+1}$ can be given as explicit functions of the samples.

It is well known that for positive H the best possible change of measure to estimate $\gamma_{\mathbf{v}}$ is such that $\mathbf{X}$ has a density g given by

$$g(\mathbf{x}) = \frac{H(\mathbf{x}) f(\mathbf{x}; \mathbf{v})}{\gamma_{\mathbf{v}}}, \tag{5}$$

for all $\mathbf{x} \in \mathscr{X}$. However, this density may not belong to the family $\{f(\cdot; \mathbf{v})\}$. Instead of trying to find a tilting parameter ${}_*\mathbf{v}$, which minimizes the variance of the estimator (1), we could try to find a density $f(\cdot; \mathbf{v}^*)$ which, in some sense, is closest to the density given in (5). One way of doing this is by minimizing the Kullback-Leibler or CE "distance" between g and $f(\cdot; \mathbf{v}^*)$, which is given (see, e.g., Kapur and Kesavan 1992) by

$$\mathbb{E}_g \log \frac{g(\mathbf{X})}{f(\mathbf{X}; \mathbf{v}^*)}, \tag{6}$$

where $\mathbb{E}_g$ denotes expectation under g. It is not difficult to see that this is equivalent to finding

$$\mathbf{v}^* = \arg\max_{\tilde{\mathbf{v}}} \mathbb{E}_{\mathbf{v}} H(\mathbf{X}) \log f(\mathbf{X}; \tilde{\mathbf{v}}). \tag{7}$$

Analogously to (3), this is equivalent to

$$\mathbf{v}^* = \arg\max_{\tilde{\mathbf{v}}} \mathbb{E}_{\mathbf{v}_j} H(\mathbf{X}) W(\mathbf{X}; \mathbf{v}, \mathbf{v}_j) \log f(\mathbf{X}; \tilde{\mathbf{v}}), \tag{8}$$

for any tilting parameter $\mathbf{v}_j$. Similarly to (4), we may estimate $\mathbf{v}^*$ by

$$\mathbf{v}_{j+1} = \arg\max_{\tilde{\mathbf{v}}} \sum_{i=1}^{N} H(\mathbf{X}^{(i)}) W(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j) \log f(\mathbf{X}^{(i)}; \tilde{\mathbf{v}}), \tag{9}$$

where $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ is a random sample from $f(\cdot; \mathbf{v}_j)$. Because under quite mild conditions (Rubinstein and Shapiro 1993), the program

$$\max_{\tilde{\mathbf{v}}} \sum_{i=1}^{N} H(\mathbf{X}^{(i)}) W(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j) \log f(\mathbf{X}^{(i)}; \tilde{\mathbf{v}})$$

is convex and differentiable with respect to $\tilde{\mathbf{v}}$, the tilting vector $\mathbf{v}_{j+1}$ in (9) may be readily obtained by solving the following system of nonlinear equations (with respect to $\tilde{\mathbf{v}}$):

$$\sum_{i=1}^{N} H(\mathbf{X}^{(i)}) W(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j) \nabla \log f(\mathbf{X}^{(i)}; \tilde{\mathbf{v}}) = \mathbf{0}, \tag{10}$$

where the gradient is with respect to $\tilde{\mathbf{v}}$. This, of course, is provided that the expectation and differentiation operators can be interchanged (see Rubinstein and Shapiro 1993) and the function (8) is convex and differentiable with respect to $\tilde{\mathbf{v}}$.

As noted above, $\mathbf{v}_{j+1}$ can often be calculated analytically. In particular, this happens if the distributions of the random variables belong to a natural exponential family (NEF); this is demonstrated in Appendix A for a simple case and in the next section for a general queueing model.

3. Estimating Buffer Overflow Probabilities

In this section, we present the main algorithm for estimating buffer overflow probabilities in queueing networks.

Consider an open network of $GI/G/1$ queues with Markovian routing. We are interested in the probability $\gamma(l)$ of the event A that the content of a certain queue, or the combined contents of several queues, exceeds a certain level l during an interval $[0, T]$, where T is some (random) stopping time for the process $\mathbf{X}$ of interarrival times (from outside the system), service times, and routing decisions. Typically, T is the length of a busy cycle, or the first time until either the content of a queue exceeds level l or the system becomes empty.

We wish to estimate $\gamma(l)$ by using an IS procedure, in which we can change the service and interarrival time distribution at each queue. We assume that for each queue the interarrival and service time distributions belong to a NEF family that is reparametrized by the mean (vector of means) $\mathbf{v}$, as discussed in Appendix A. Note that such an IS procedure is state independent: the change of the distributions is made globally and does not vary with the state variables of the system (e.g., the content of the queues).

The idea is first to estimate the optimal tilting parameter via the iterative schemes (4) or (9) and then to use this to estimate $\gamma(l)$ via ordinary IS.

In most cases of interest, $\gamma(l)$ is a rare event probability. This means that the choice of a "good" initial tilting parameter $\mathbf{v}_0$ for the scheme (4) or (9) is crucial. For general queueing networks it is unclear what comprises a good initial guess. Obviously, the system should be instable, but it is far from trivial to determine which instable regimes are good and which are not good.

We now make three conjectures. All conjectures have been observed numerically and some can be proved in certain simple situations (see below).

1. Instability Property. The optimal tilting parameter corresponding to overflow of a *low* level l_0 (e.g., $l_0 = 3$ or $l_0 = 4$) renders the system instable.

2. Robustness Property. An optimal parameter corresponding to overflow of a *low* level l_0 is a "good" initial tilting vector for finding the optimal tilting parameter for the high level l. In other words, the estimation of the tilting parameter for the high level l is robust (insensitive) to the choice of l_0.

3. CE Optimality Property. The minimum-variance tilting parameter asymptotically coincides with the minimum CE tilting parameter.

The third property means that we can use a very simple updating formula for the tilting vectors. In particular, let $\mathbf{v} = (v_1, \ldots, v_K)$ be the (nominal) vector of means corresponding to the pdfs $(f_1, \ldots, f_K)$ of interarrival times (customers arriving to the queue from outside the system) and service times at the queues. For simplicity, we assume that the routing probabilities remain fixed; see, however, Remark 3.2. Let $H(\mathbf{X})$ be the indicator of the event A. Note that each parameter v_k corresponds to a service time or an (external) interarrival time at a certain queue. For each such service or interarrival time (indexed by k) there will be τ_k service completions/interarrivals. Denote these by $Y_{k1}, \ldots, Y_{k\tau_k}$. It follows that the density $f(\mathbf{X}; \mathbf{v})$, corresponding to the history of the process $\mathbf{X}$ during $[0, T]$, is the product

$$f(\mathbf{X}; \mathbf{v}) = \prod_{k=1}^{K} \prod_{j=1}^{\tau_k} f_k(Y_{kj}; v_k). \tag{11}$$

Thus, the likelihood ratio $W(\mathbf{X}; \mathbf{v}, \mathbf{v}_j)$, corresponding to the history of the process $\mathbf{X}$ during $[0, T]$, is the quotient of the products of the form above. Now, combining (11), (9), and Appendix A it is not difficult to see that for NEFs the components of the tilting vector should be updated as

$$v_{j+1,k} = \frac{\sum_{i=1}^{N} \left(H^{(i)}(\mathbf{X})^{(i)} W^{(i)}(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j) \sum_{j=1}^{\tau_k^{(i)}} Y_{kj}^{(i)} \right)}{\sum_{i=1}^{N} H^{(i)}(\mathbf{X}^{(i)}) W^{(i)}(\mathbf{X}^{(i)}; \mathbf{v}, \mathbf{v}_j) \tau_k^{(i)}}, \tag{12}$$

where the simulation is performed under the tilting vector $\mathbf{v}_j$.

Based on the three properties above, we now have the following algorithm:

Main Algorithm

Pilot Stage.

1. Choose an initial buffer level l_0. Choose the initial tilting vector $\mathbf{v}_0 = \mathbf{v}$.
2. Simulate N_1 paths, using the tilting vector $\mathbf{v}_0$, for overflow level l_0.
3. Find the tilting vector $\mathbf{v}_1$ from (12), for overflow level l_0.

Second Stage.

1. Initialize as follows: $j := 0$ (iteration counter); choose as initial tilting vector $\mathbf{v}_0$ the resulting tilting vector ($\mathbf{v}_1$) of the pilot stage.
2. Simulate N_2 replications with the tilting vector $\mathbf{v}_j$.
3. Find the tilting vector $\mathbf{v}_{j+1}$ from (12), for overflow level l.
4. Increment j and repeat Steps 2–4, until the tilting vector has converged.

Third Stage. Estimate the probability $\gamma_{\mathbf{v}}$ via IS simulation, as in (1), with the final tilting vector obtained in the second stage.

Remark 3.1. To assess if an initial tilting vector $\mathbf{v}_0$ is "good," we have to consider how effective the second stage of the main algorithm is. Numerical evidence shows that vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$ converge accurately and quickly to the optimal tilting vector $\mathbf{v}^*$. We examine this issue further in the next section.

Remark 3.2. In the above, each random variable (and thus each element of $\mathbf{v}$) was assumed to correspond to a service or interarrival time. However, the same formalism also applies to random routing among two destinations: This involves a Bernoulli random variable, with outcomes 0 and 1 corresponding to the two destinations. The mean of this random variable is just the routing probability, which can be directly incorporated into $\mathbf{v}$, thus allowing our algorithm to also find the optimal routing probability.

4. Importance Sampling and the Cross-Entropy Method Applied to the $M/M/1$ Queue

In this section, we have a closer look at how IS and the CE method work for the $M/M/1$ queue.

Consider the probability that the queue length in an $M/M/1$ queue exceeds level l during a busy period, starting with i customers in the system at the beginning of the busy period. Denote this probability by γ_i, $i = 1, 2, \ldots, l$. Let the arrival intensity be λ and the service intensity μ. Also define $p = \lambda/(\lambda + \mu)$, $q = 1 - p$, and $\rho = p/q = \lambda/\mu$. Let $\{Y_n, n = 0, 1, 2, \ldots\}$ be the embedded Markov chain describing the number of customers in the system at arrival and departure times. Define $\mathbb{P}_i$ as the probability measure under which $\{Y_n\}$ starts at i. The corresponding expectation operator is denoted by $\mathbb{E}_i$. Let

$T = \inf\{n > 0\colon Y_n = 0 \text{ or } Y_n = l\}$. We are mainly interested in the case where we start with $i = 1$ customer in the system at the beginning of the busy period. By the classical Gambler's Ruin theorem,

$$\gamma_i = \frac{1-(q/p)^i}{1-(q/p)^l}, \qquad i = 1, \ldots, l. \tag{13}$$

We wish to estimate γ_i using a state-independent IS procedure. To define this procedure precisely it is convenient to introduce a random walk $\{S_n,\ n = 1, 2, \ldots\}$ with $S_n = X_1 + \cdots + X_n$ such that each X_k takes values 1 and -1 with probabilities, respectively,

$$\tilde{p} = \frac{p\,\mathrm{e}^{\theta}}{p\,\mathrm{e}^{\theta} + q\,\mathrm{e}^{-\theta}} \quad \text{and} \quad \tilde{q} = \frac{q\,\mathrm{e}^{-\theta}}{p\,\mathrm{e}^{\theta} + q\,\mathrm{e}^{-\theta}}. \tag{14}$$

The reader may verify from Appendix B that the distributions of X_k form a NEF with densities

$$f(x;\theta) = \mathrm{e}^{\theta x - \kappa(\theta)}\, h(x), \qquad x \in \{-1, 1\},\ \theta \geq 0,$$

with $h(1) = p$ and $h(-1) = q$, where

$$\kappa(s) = \log \mathbb{E}\,\mathrm{e}^{sX_1} = \log(p\,\mathrm{e}^{s} + q\,\mathrm{e}^{-s}). \tag{15}$$

Returning to our IS procedure, we define a change of measure $\widetilde{\mathbb{P}}_i$ such that under this measure the process $\{Y_n\}$ starts in i, and the parameters p and q are changed to $\tilde{p}$ and $\tilde{q}$ in (14). The corresponding expectation operator is denoted by $\widetilde{\mathbb{E}}_i$. Without loss of generality we assume that under $\widetilde{\mathbb{P}}_i$ the random variables $X_1, X_2, \ldots$ are i.i.d. and take values 1 and -1 with probability $\tilde{p}$ and $\tilde{q}$, respectively.

Now let A be the event that $\{Y_n\}$ reaches l before 0. Note that under $\mathbb{P}_i$ (or $\widetilde{\mathbb{P}}_i$), we can view A also as the event that $\{S_n\}$ reaches $l - i$ before $-i$. Similarly, T can be viewed as the first time that $\{S_n\}$ reaches $l - i$ or $-i$. We now have

$$\gamma_i = \mathbb{E}_i I_A = \widetilde{\mathbb{E}}_i I_A W_T,$$

with

$$W_T = \mathrm{e}^{-\theta S_T + T\kappa(\theta)}. \tag{16}$$

This last formula follows from the fact that for any fixed n the likelihood ratio of $(X_1, \ldots, X_n)$ with respect to $\mathbb{P}_i$ and $\widetilde{\mathbb{P}}_i$ is $\mathrm{e}^{-\theta S_n + n\kappa(\theta)}$.

Hence, we may estimate γ_i by simulating independent copies of the random variable $Z := I_A W_T$, and then taking the average. The question is how to choose the tilting parameter θ optimally. In the next two subsections, we examine the two approaches discussed in §2: the minimum-variance method and the CE method.

Remark 4.1. It can be shown that a zero-variance way to simulate γ_i is to use IS with a state-dependent change of measure in which

$$p_k \propto p\gamma_{k+1} \quad \text{and} \quad q_k \propto q\gamma_{k-1}, \qquad k = 1, \ldots, l-1, \tag{17}$$

where $\propto$ is the symbol for proportionality. We will use this result later on.

4.1. Minimum-Variance Method

The best possible change of measure is such that the variance of Z under the change of measure is minimal. Because $\widetilde{\mathbb{E}}_i Z = \gamma_i$, it suffices to minimize $\widetilde{\mathbb{E}}_i Z^2$. But

$$\begin{aligned}\widetilde{\mathbb{E}}_i Z^2 &= \widetilde{\mathbb{E}}_i W_T^2 I_A = \mathbb{E}_i \frac{1}{W_T} W_T^2 I_A \\ &= \mathbb{E}_i W_T I_A = \mathbb{E}_i[\mathrm{e}^{-\theta S_T + T\kappa(\theta)} \mid A]\gamma_i \\ &= \mathbb{E}_i[\mathrm{e}^{T\kappa(\theta)} \mid A]\,\mathrm{e}^{-\theta(l-i)}\,\gamma_i,\end{aligned}$$

where we have used the fact that under $\mathbb{P}_i$ we have $S_T = l - i$. It remains to find $\mathbb{E}_i[\mathrm{e}^{T\kappa(\theta)} \mid A]$ or $\mathbb{E}_i[z^T \mid A]$, for general z. Here we can use the fact that, conditioned on A, the Markov chain $\{Y_n\}$ has transition probabilities given in (17). Let $b_i(z) := \mathbb{E}_i[z^T \mid A]$. Then, by conditioning on Y_1 and using (17) we have the following recursion:

$$b_i(z) = z b_{i-1}(z) \frac{q\gamma_{i-1}}{q\gamma_{i-1} + p\gamma_{i+1}} + z b_{i+1}(z) \frac{p\gamma_{i+1}}{q\gamma_{i-1} + p\gamma_{i+1}}.$$

Noting that $\gamma_i = q\gamma_{i-1} + p\gamma_{i+1}$, and defining $a_i(z) = \gamma_i b_i(z)$, we have

$$a_{i+1}(z) - \frac{1}{zp} a_i(z) + \frac{q}{p} a_{i-1}(z) = 0, \qquad i = 1, \ldots, l-1,$$

with $a_0(z) = 0$ and $a_l(z) = 1$. This is readily solved as

$$a_i(z) = \frac{\eta_1^i - \eta_2^i}{\eta_1^l - \eta_2^l}, \qquad i = 0, 1, \ldots, l, \tag{18}$$

where

$$\eta_1 = \frac{1 + \sqrt{1 - 4z^2pq}}{2zp} \quad \text{and} \quad \eta_2 = \frac{1 - \sqrt{1 - 4z^2pq}}{2zp}. \tag{19}$$

Concluding, we have

$$\widetilde{\mathbb{E}}_i Z^2 = \mathrm{e}^{-\theta(l-i)}\, \frac{\eta_1^i - \eta_2^i}{\eta_1^l - \eta_2^l}, \tag{20}$$

with η_1 and η_2 given in (19), for $z = p\,\mathrm{e}^{\theta} + q\,\mathrm{e}^{-\theta} = \mathrm{e}^{\kappa(\theta)}$. This gives us a relatively simple explicit formula to find the optimal minimum-variance tilting parameter ${}_*\theta$.

Direct inspection shows that as l increases, ${}_*\theta$ decreases. Consequently, the traffic intensity under ${}_*\theta$, denoted by ${}_*\rho$, decreases with l. This is a somewhat unexpected result. Also, it is not difficult to see that as $l \to \infty$, ${}_*\theta \to \log(\mu/\lambda)$. For this asymptotic tilting parameter we have the twisted arrival and service rate $\tilde{\lambda} = \mu$ and $\tilde{\mu} = \lambda$; in other words, we interchange the original arrival and service rate. This is a well-known result (Sadowsky 1991). Note that under this change of measure the tilted traffic intensity is $\tilde{\rho} = \rho^{-1}$.

Moreover, we have in (20) that $z = 1$, $\eta_1 = \tilde{\rho}$, and $\eta_2 = 1$, so that for example

$$\widetilde{\mathbb{E}}_1 Z^2 = \rho^l \frac{\tilde{\rho} - 1}{\tilde{\rho}^l - 1}. \tag{21}$$

Note also that for any level l_0 the queue is unstable. This follows from Appendix B, or can be verified directly. We thus have

THEOREM 4.1 (INSTABILITY THEOREM). *The optimal tilted traffic intensity $\tilde{\rho}(l)$ for the buffer overflow probability in a $M/M/1$ queue is* greater than unity *regardless of the buffer size l, ($l \geq 2$). In addition, $\tilde{\rho}(l)$ decreases in l and*

$$\lim_{l \to \infty} \tilde{\rho}(l) = \rho^{-1}.$$

REMARK 4.2. Observe that $a_i(z) = \sum_{n=0}^{\infty} p_i(n) z^n$ is the generating function of the probability $p_i(n)$ of the gambler's ruin (absorption at 0) at the nth trial (Feller 1968, p. 351). In particular, we can write the generating function as

$$a_i(z) = \sum_{n=0}^{\infty} z^n \left\{ l^{-1} 2^n p^{(n-i)/2} q^{(n+i)/2} \cdot \sum_{\nu=1}^{l-1} \cos^{n-1} \frac{\pi\nu}{l} \sin \frac{\pi\nu}{l} \sin \frac{\pi i \nu}{l} \right\}. \tag{22}$$

Note that the convergence radius of this power series is $|z| \leq 1/(2\sqrt{pq})$.

4.2. Cross-Entropy Method

Let T, as before, be the first time until $\{S_n\}$ hits level $l - i$ or $-i$, and let A be the event that $l - i$ is reached before $-i$. Let $f_n(\cdot; \theta)$ be the pmf of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ under the change of measure $\widetilde{\mathbb{P}}_i$. Specifically,

$$f_n(\mathbf{x}; \theta) = \prod_{k=1}^{n} \tilde{p}^{(1+x_k)/2} \tilde{q}^{(1-x_k)/2},$$

where $\tilde{p}$ and $\tilde{q}$ are given in (14). According to (7), we have to find θ such that $\mathbb{E}_i I_A \log f_T(\mathbf{X}; \theta)$ is maximized. Now,

$$\begin{aligned} &\log f_T(\mathbf{X}; \theta) \\ &\quad = \sum_{k=1}^{T} \left[\frac{1 + X_k}{2} \log \tilde{p} + \frac{1 - X_k}{2} \log \tilde{q} \right] \\ &\quad = \tfrac{1}{2} T \log \tilde{p} + \tfrac{1}{2} S_T \log \tilde{p} + \tfrac{1}{2} T \log \tilde{q} - \tfrac{1}{2} S_T \log \tilde{q} \\ &\quad = \tfrac{1}{2} T \{\log(pq) - 2\kappa(\theta)\} + \tfrac{1}{2} S_T \{\log(p/q) + 2\theta\}. \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{E}_i I_A \log f_T(\mathbf{X}; \theta) \\ &\quad = \tfrac{1}{2} \{\log(pq) - 2\kappa(\theta)\} \mathbb{E}_i T I_A + \tfrac{1}{2} \{\log(p/q) + 2\theta\} \mathbb{E}_i S_T I_A \\ &\quad = \frac{\log(pq) - 2\kappa(\theta)}{2} a_i'(1) + \frac{\log(p/q) + 2\theta}{2} \gamma_i (l - i), \end{aligned}$$

where γ_i is given in (13) and $a_i(z)$ in (18). Consequently, we need to minimize

$$\kappa(\theta) \frac{a_i'(1)}{\gamma_i} - \theta(l - i), \tag{23}$$

where $\kappa(\theta)$ is given in (15). Note that $a_i'(1)$ depends on l and i but not on θ. For $p \neq 1/2$, we can show that for large l

$$\mathbb{E}_i[T \mid A] = \frac{a_i'(1)}{\gamma_i} = \frac{l}{|1 - 2p|} + o(l).$$

It follows that for large l, θ^* is such that $\kappa(\theta) - \theta|1 - 2p|$ is minimized. For $0 < p < 1/2$, this means that asymptotically $\theta^* = \log(q/p)$, corresponding to a change of measure where p and q are swapped. For $1/2 < p < 1$, we have $\theta^* = 0$ corresponding to the original measure.

To illustrate that the θ^* (CE method) and ${}_*\theta$ (minimum-variance method) are close, consider $\widetilde{\mathbb{E}}_i Z^2$ in (20). Now observe that

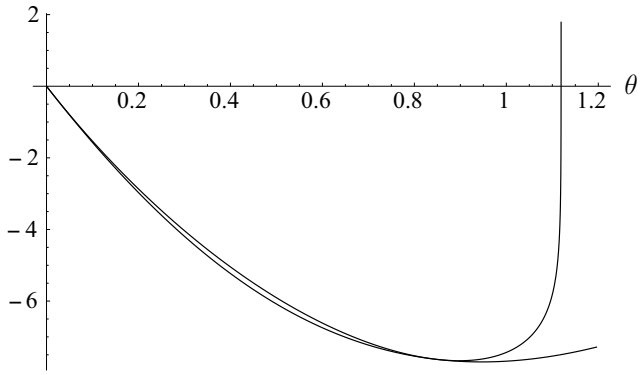
$$\widetilde{\mathbb{E}}_i Z^2 = \mathrm{e}^{(l-i)\theta} \frac{a_i(\mathrm{e}^{\kappa(\theta)})}{\gamma_i}.$$

Second, from the Taylor expansion $a_i(z) = \gamma_i + a_i'(1)(z - 1) + O((z - 1)^2)$ at $z = 1$, we obtain

$$\begin{aligned} \log \frac{a_i(\mathrm{e}^{\kappa(\theta)})}{\gamma_i} &= \log \left\{ 1 + \frac{a_i'(1)}{\gamma_i} (\mathrm{e}^{\kappa(\theta)} - 1) + O((\mathrm{e}^{\kappa(\theta)} - 1)^2) \right\} \\ &= \frac{a_i'(1)}{\gamma_i} (\mathrm{e}^{\kappa(\theta)} - 1) + O((\mathrm{e}^{\kappa(\theta)} - 1)^2) \\ &= \frac{a_i'(1)}{\gamma_i} \kappa(\theta) + O(\kappa^2(\theta)). \end{aligned}$$

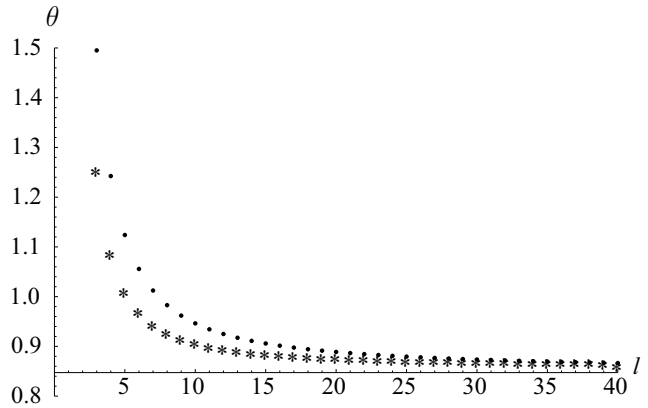
In other words, $\log \widetilde{\mathbb{E}}_i Z^2 \approx (\kappa(\theta) a_i'(1))/\gamma_i - \theta(l - i)$, and thus it is conceivable that ${}_*\theta$ is close to θ^*. The closeness of the two optimal tilting parameters is illustrated in Figures 1 and 2.

Figure 1 **The Graphs of $\log \widetilde{\mathbb{E}}_i Z^2$ and $(\kappa(\theta) a_i'(1))/\gamma_i - \theta(l - i)$; for $l = 10$, $p = 3/10$, $i = 1$**



Note. We have ${}_*\theta = 0.90$ as the argmin of the first function and $\theta^* = 0.95$ as the argmin of the second function. The asymptotically optimal tilting parameter is $\log(q/p) = \log(7/3)$.

Figure 2 **Optimal Tilting Parameters ${}_*\theta$ (Stars) and θ^* (Dots) for Various Values of l, with $p = 3/10$ and $i = 1$**



Note. $\theta^*(l) > {}_*\theta(l)$. Also, ${}_*\theta(2) = \theta^*(2) = \infty$.

An alternative way to find the optimal tilting parameter is to use the fact that we are dealing here with a NEF. Using a similar argument as in (A2), we obtain the following simple formula for the optimal CE parameter if we reparametrize the NEF via the mean v:

$$v^* = \frac{\mathbb{E}_i I_A \sum_{k=1}^{T} X_k}{\mathbb{E}_i I_A T}.$$

On A, $\sum_{k=1}^{T} X_k$ is simply $l - i$, so $\mathbb{E}_i I_A \sum_{k=1}^{T} X_k = (l-i)\gamma_i$. Also, we saw before that $\mathbb{E}_i I_A T$ is equal to $a_i'(1)$. Consequently,

$$v^* = \frac{(l-i)\gamma_i}{a_i'(1)}.$$

This is in accordance with finding θ^* by minimizing (23), or solving

$$\kappa'(\theta^*)\frac{a_i'(1)}{\gamma_i} - (l-i) = 0,$$

because $v^* = \kappa'(\theta^*)$ by definition (see Appendix A).

4.3. Robustness Property

Consider the main algorithm. In the pilot stage we obtain an initial tilting parameter $v_1(l_0)$ via the estimator

$$\frac{\sum_{k=1}^{N_1} I_{A_0}^{(k)}(l_0 - i)}{\sum_{k=1}^{N_1} I_{A_0}^{(k)} T^{(k)}}, \tag{24}$$

where the simulation is carried out under the original measure (i.e., with $\theta = 0$). Here, the $I_{A_0}^{(k)}$ are the indicators of the event that l_0 is reached before 0. The estimator above is a ratio estimator, that is, an estimator of the form

$$R_N := \frac{\sum_{k=1}^{N} U_i}{\sum_{k=1}^{N} V_i},$$

where $(U_1, V_1), (U_2, V_2), \ldots$ are i.i.d. It is well known (see, for example, Asmussen et al. 1994 and Rubinstein and Melamed 1998) that if $\mathbb{E}U$ and $\mathbb{E}V$ are finite, then the estimator R_N converges with probability 1 to $r := \mathbb{E}U/\mathbb{E}V$, as $N \to \infty$. Moreover, if $\mathbb{E}U^2$, $\mathbb{E}V^2$, and $\mathbb{E}UV$ are finite, then $\sqrt{N}(R_N - r)$ converges in distribution to a $\mathsf{N}(0, \sigma^2)$ distribution, where

$$\sigma^2 = \frac{1}{(\mathbb{E}V)^2}\operatorname{Var} U + \frac{(\mathbb{E}U)^2}{(\mathbb{E}V)^4}\operatorname{Var} V - 2\frac{\mathbb{E}U}{(\mathbb{E}V)^3}\operatorname{Cov}(U, V).$$

For the estimator in (24), this means that if we can show that $\mathbb{E}_i I_{A_0} T < \infty$, then the estimator converges with probability 1 to the optimal $v^*(l_0)$. But this follows from the fact that $\mathbb{E}_i I_{A_0} T = a_i'(1)$; and because a_i has convergence radius larger than 1 if $p \neq 1/2$, then $a_i'(1)$ must be finite (for $p \neq 1/2$). Asymptotic normality for the first stage follows in the same way.

Next, we consider the second stage. Here, we start with some tilting parameter θ obtained via the pilot stage. Suppose we estimate $v^*(l)$ via one iteration. The estimator is given by

$$\frac{(l-i)\sum_{k=1}^{N_2} I_A^{(k)} W_{T^{(k)}}}{\sum_{k=1}^{N_2} I_A^{(k)} W_{T^{(k)}} T^{(k)}}, \tag{25}$$

where the simulation is carried out under the tilted measure with tilting parameter θ. To show that this ratio estimator has the consistency and asymptotic normality property, we have to show that $\widetilde{\mathbb{E}}_i U^2$, $\widetilde{\mathbb{E}}_i V^2$, and $\widetilde{\mathbb{E}}_i UV$ are all finite, with $U := I_A W$ and $V := I_A W T$. Using the definition of $a_i(z)$ and the fact that $I_A W = I_A \, \mathrm{e}^{-\theta(l-i)} \, \mathrm{e}^{T\kappa(\theta)}$, we have

$$\begin{aligned}
\widetilde{\mathbb{E}}_i U &= \mathbb{E}_i I_A = \gamma_i, \\
\widetilde{\mathbb{E}}_i V &= \mathbb{E}_i I_A T = a_i'(1), \\
\widetilde{\mathbb{E}}_i U^2 &= \mathbb{E}_i I_A W = \mathrm{e}^{-\theta(l-i)}\, a_i(\mathrm{e}^{\kappa(\theta)}), \\
\widetilde{\mathbb{E}}_i V^2 &= \mathbb{E}_i I_A W T^2 = \mathrm{e}^{-\theta(l-i)}\, a_i''(\mathrm{e}^{\kappa(\theta)})\, \mathrm{e}^{2\kappa(\theta)} \\
&\quad + \mathrm{e}^{-\theta(l-i)}\, a_i'(\mathrm{e}^{\kappa(\theta)})\, \mathrm{e}^{\kappa(\theta)}, \\
\widetilde{\mathbb{E}}_i UV &= \mathbb{E}_i I_A W T = \mathrm{e}^{-\theta(l-i)}\, \mathrm{e}^{\kappa(\theta)}\, a_i'(\mathrm{e}^{\kappa(\theta)}).
\end{aligned}$$

It follows that a sufficient condition for asymptotic normality is that

$$\mathrm{e}^{\kappa(\theta)} < \frac{1}{2\sqrt{pq}}, \tag{26}$$

because if $\mathrm{e}^{\kappa(\theta)}$ is less than the convergence radius of the power series $a_i(z)$, then all derivatives of a_i exist at $\mathrm{e}^{\kappa(\theta)}$. Moreover, if $\mathrm{e}^{\kappa(\theta)}$ is larger than the convergence radius, then all derivatives must be ∞. Note that condition (26) holds if at the first stage l_0 is large enough, that is, when $\kappa(\theta)$ is close enough to 0.

Theorem 4.2 (Robustness Theorem). *Suppose θ_0 is the optimal tilting parameter for estimating the buffer overflow probability in an $M/M/1$ queue with a (low) overflow level l_0. Consider simulating this same $M/M/1$ queue*

but with a higher overflow level l, and tilted with θ_0. If this simulation is used to estimate the tilting parameter θ^* (using (25)*), that would in turn be CE optimal for estimating the overflow probability at level l, and this estimate of θ^* has finite variance for all $l > l_0$ for all l_0 sufficiently large (namely, such that* (26) *is satisfied*).

Remark 4.3. In general (for cases other than $M/M/1$), it will not be easy to find the "cut-off" value for l_0. However, practical experiments (see §5) suggest that actually a stronger robustness holds: The condition on l_0 does not seem to be necessary, and the estimator's variance is not just finite, but small enough to be practical.

As an example, suppose we wish to estimate for $p = 0.3$ the optimal tilting parameter θ^* for $l = 20$, by simulating the $M/M/1$ queue under the tilting parameter θ (obtained from the pilot run). It follows from (26) that the ratio estimator is asymptotically normal if $\theta < 1.03$. From Figure 2 we see that any sufficiently accurate pilot stage with initial level $l_0 \geq 7$ makes $\theta < 1.03$ and thus brings the second stage in a region where the sufficient condition (26) holds. For $l_0 \leq 6$, both the numerator and the denominator of the ratio estimator (24) have infinite variance, and it is unclear what consequences this has for their ratio, which will be used as the tilting parameter for the next iteration. Note also that for $l_0 = 2$ the optimal θ is ∞.

5. Simulation Results

In §§5.1–5.3, we give some numerical examples of the application of our main algorithm. These examples are used to illustrate the three properties we have discussed above.

5.1. Single $M/M/1$ Queue

As a first example, we consider the $M/M/1$ queue, with arrival rate $\lambda = 0.3$, service rate $\mu = 0.7$, and overflow level (buffer size) $l = 20$.

The results are presented in Table 1. The table has one row for every simulation run (iteration), listing the number of busy cycles (replications) simulated, the values (in principle) of the tilting parameters $\mathbf{v}_k$, and the estimate for the overflow probability found in that simulation run along with its relative error (RE). In the present model, all distributions are exponential, and tilting them exponentially gives again an exponential distribution. Therefore, instead of listing the tilting parameters $\mathbf{v}_k$ explicitly, we prefer to show the resulting rates, because these are more intuitive. The same applies to routing probabilities in later examples.

It should be noted that there is a difference between the simulation of the $M/M/1$ queue as performed here and the analysis in §4. In the analysis, it was assumed that the simulation is done in terms of a discrete-time Markov chain: Basically, samples are drawn from a Bernoulli distribution to decide whether the next event is an arrival or a service completion. In contrast, the present simulation uses a continuous-time Markov chain: Two independent exponential distributions are sampled, one for determining the time of the next arrival, the other for determining service durations. Obviously, both representations are valid and thus should lead to the same estimate for the overflow probability. The reason the actual simulations are done with a continuous-time model is that this formulation is more in line with the one in §3, and it is easily generalized to non-Markovian models.

Table 1 **Simulation Results for the $M/M/1$ Queue for $l = 20$**

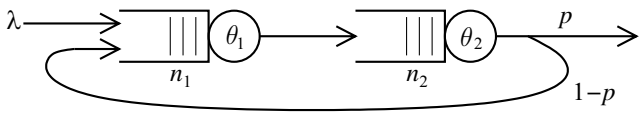
Iteration	Repl.	λ	μ	Estimate	RE
$l_0 = 2$					
1	100	0.300	0.700	—	—
2	1,000	1.406	0.449	$8.309 \cdot 10^{-9}$	0.3031
3	1,000	1.004	0.319	$4.643 \cdot 10^{-8}$	0.1332
4	1,000	0.787	0.275	$5.286 \cdot 10^{-8}$	0.0514
5	1,000	0.743	0.298	$5.597 \cdot 10^{-8}$	0.0419
6	1,000	0.729	0.296	$5.952 \cdot 10^{-8}$	0.0406
$l_0 = 8$					
1	10,000	0.300	0.700	—	—
2	1,000	0.805	0.285	$5.609 \cdot 10^{-8}$	0.0573
3	1,000	0.728	0.294	$6.148 \cdot 10^{-8}$	0.0398
4	1,000	0.723	0.299	$5.940 \cdot 10^{-8}$	0.0406
5	1,000	0.716	0.296	$6.211 \cdot 10^{-8}$	0.0385

Table 1 shows results for two different values of the overflow level l_0 in the pilot run, namely 2 and 8. The former is the minimum that can work; for $l_0 = 1$, the system would already have reached the "rare" target event in its initial state. In the case with $l_0 = 8$, the overflow in the pilot run is rather rare, so a large number of replications are needed to observe it a reasonable number of times (16 in this experiment).

The results for the case $l_0 = 8$ show that three iterations can indeed be enough. The first (pilot run) makes the system unstable; that is, the λ and μ that the pilot run calculates as optimal for the second iteration are such that $\lambda > \mu$. The second run does not yet yield an optimal (i.e., low RE) estimate of the overflow probability, because it uses a tilting found in the first iteration, which is thus optimal for an overflow level of 8 rather than 20. However, the second run does find optimal values for λ and μ to be used in the third iteration; the third iteration achieves a relative error of 0.0398, and further iterations do not significantly improve this.

In the case of $l_0 = 2$, things look a bit different. Clearly, five iterations are needed here before λ and μ are sufficiently close to their final values to achieve a low relative error. This is not surprising.

Figure 3 Two Queues in Tandem with Feedback



At the end of §4 it was noted that if l_0 is chosen too low, the estimator for the tilting parameter becomes the ratio of two infinite-variance estimates, and thus has unknown behavior. That was calculated for the discrete-time simulation, but it seems reasonable to expect similar problems in the continuous-time counterpart. The present simulation results suggest that the estimator for the tilting vector is biased in this situation, causing more iterations to be needed. With every iteration we move closer to the correct tilting and thus away from the "problematic" region.

Looking at these two examples, it may at first glance seem beneficial to choose a high l_0, because it saves one or two iterations; however, this comes at the cost of needing more replications in the pilot run.

Finally, for this simple model, the overflow probability can also be calculated directly, giving $\gamma_1 = 5.826 \cdot 10^{-8}$, which confirms the correctness of the simulation results.

5.2. Two Non-Markovian Queues with Random Feedback

As a second example, we consider the network depicted in Figure 3. It consists of two queues in tandem, where customers departing from the second queue either leave the network (with probability p) or go back to the first queue (with probability $1-p$). We are interested in the probability that the total number of customers in the network exceeds some high level, 50 in this example, during one busy cycle.

Interestingly, for this model (and in general, any model with random feedback) we cannot work with $l_0 = 2$, as we could in the single $M/M/1$ queue. The reason for this is the following. Consider using $l_0 = 2$. This means that after starting the busy cycle with 1 customer in the network, we are interested in the probability of reaching a state where 2 customers are in the network, before the network becomes empty. So, until the overflow, there will be always exactly 1 customer in the network: if less than 1, the busy cycle would end, and if more than 1, the overflow would happen. Therefore, no departures from the system can occur on a sample path to the overflow. Consequently, if a service completion ever happens at the second queue on the sample path, the customer leaving that queue *must* be routed back to the first queue, otherwise the busy cycle would end. Therefore, we will observe customers being routed back to the first queue with probability 1, which then becomes the value of the routing probability for the next iteration due to the CE algorithm. Once a routing probability has become 1, later iterations will never observe the alternative routing decision, so the probability will remain 1. So using a pilot run with $l_0 = 2$ forces the routing probability to be 1 in all later iterations, which is incorrect if $l > 2$ in those iterations.

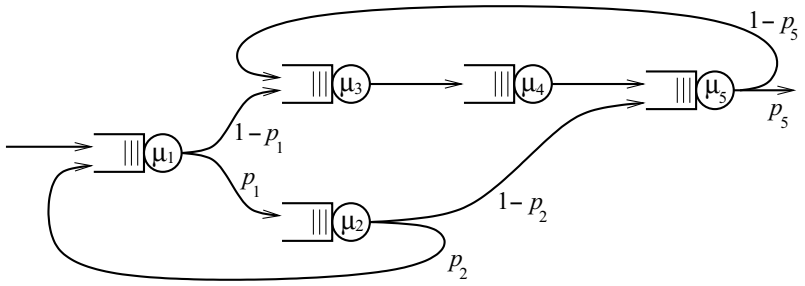
In this example, the interarrival time distribution is a two-stage Erlang distribution, with exponential parameter $\lambda = 0.2$. The service time distribution of the first server is uniform on $[0, 3.333]$, and the second server's service time has a Weibull distribution with shape parameter $= 2$, scaled such that the average service duration is 2.5. The results are shown in Table 2. In this table, θ_1 and θ_2 are the exponential tilting factors applied to the non-Markovian service time distributions; basically, these are the θ from (A1).

The algorithm converges quickly, already reaching the final accuracy in the third iteration. No numeri-

Table 2 Simulation Results for the Non-Markovian Network for $l = 50$

Iteration	Repl.	λ	θ_1	θ_2	p	Estimate	RE
$l_0 = 3$							
1	100	0.2000	0.0000	0.0000	0.5000	—	—
2	10^4	0.3423	−0.0237	0.2373	0.1778	$1.855 \cdot 10^{-25}$	0.2097
3	10^4	0.3622	−0.0256	0.1440	0.2312	$1.697 \cdot 10^{-25}$	0.0130
4	10^4	0.3596	−0.0000	0.1579	0.2340	$1.641 \cdot 10^{-25}$	0.0105
5	10^4	0.3600	−0.0028	0.1588	0.2341	$1.653 \cdot 10^{-25}$	0.0115
6	10^6	0.3594	0.0000	0.1591	0.2343	$1.657 \cdot 10^{-25}$	0.0011
$l_0 = 7$							
1	10^4	0.2000	0.0000	0.0000	0.5000	—	—
2	10^4	0.3675	0.0000	0.1531	0.2241	$1.640 \cdot 10^{-25}$	0.0123
3	10^4	0.3602	0.0000	0.1587	0.2345	$1.633 \cdot 10^{-25}$	0.0110
4	10^4	0.3599	0.0000	0.1593	0.2343	$1.670 \cdot 10^{-25}$	0.0106
5	10^4	0.3603	0.0000	0.1578	0.2340	$1.682 \cdot 10^{-25}$	0.0123
6	10^4	0.3606	−0.0000	0.1586	0.2352	$1.651 \cdot 10^{-25}$	0.0105
7	10^6	0.3603	−0.0023	0.1586	0.2348	$1.658 \cdot 10^{-25}$	0.0012

Figure 4 A Five-Node Jackson Network



cal results are available for validation; therefore, we did the last iteration with 100 times more replications to see whether relative error decreases appropriately (i.e., by a factor of $\sqrt{100} = 10$). The fact that this is indeed the case gives confidence.

Tilting parameters for the model considered here could also be calculated using the heuristic method from (Parekh and Walrand 1989). However, this would be a very tedious numerical calculation, involving minimization of a function available only as the numerical maximum of a function involving the error function (in this case; for other distributions, this could be different). This calculation could be done in principle, but would clearly be much more complicated than the rather straightforward adaptive simulation procedure used here.

5.3. Five-Node Jackson Network

As a final example, consider the estimation of the overflow probability of the total population of the five-node Jackson network with random routing depicted in Figure 4.

5.3.1. One Bottleneck. We first simulate this network at a parameter setting where server 3 is the bottleneck queue; it has a load of 0.2, while the other servers have a load of 0.1. These parameters are as follows: $\lambda = 3$, $\mu_1 = 40$, $\mu_2 = 20$, $\mu_3 = 25$, $\mu_4 = 50$, $\mu_5 = 60$, with all routing probabilities equal to 0.5. The overflow level during the pilot run l_0 was set to 5; this level is reached by about 1% of all sample paths under the original measure.

The results are shown in Table 3. For an overflow level of 80 the method still converges fine; and although the relative error tends to vary notably among further iterations, the estimates do appear to be consistent. We have repeated the simulation for various overflow levels and have observed that the relative error does not increase much between $l = 20$ and $l = 80$, suggesting that the method is asymptotically efficient.

Finally, it is noteworthy that the parameters found by the CE procedure are close to those calculated by the method of Frater et al. (1991) (based in turn on Parekh and Walrand 1989), which are

$$\lambda' = 13, \quad \mu_1' = 40, \quad p_1' = \tfrac{11}{26} \approx 0.423, \quad \mu_2' = 20,$$
$$p_2' = \tfrac{13}{22} \approx 0.591, \quad \mu_3' = 15, \quad \mu_4' = 50, \quad \mu_5' = 60,$$
$$p_5' = \tfrac{1}{6} \approx 0.167.$$

5.3.2. Equal Loads. Next, we simulate the same network, but with all servers having an equal load (=0.1), with $\lambda = 3$, $\mu_1 = 40$, $\mu_2 = 20$, $\mu_3 = 50$, $\mu_4 = 50$, $\mu_5 = 60$, and all routing probabilities again equal to 0.5.

The simulation results are presented in Table 4. We note that a substantially larger number of replications is needed per simulation than in the previous case. Still, the basic observations from this paper hold: The first iteration makes the system unstable, and then after one or two more iterations the final accuracy is obtained.

We note that a more efficient simulation of this case is possible at the expense of complexity, by using a state-dependent change of measure (de Boer 2000 or de Boer and Nicola 2001).

Table 3 Simulation Results for the Five-Node Network with One Bottleneck; $l_0 = 5$, $l = 80$

Iteration	Repl.	λ	μ_1	p_1	μ_2	p_2	μ_3	μ_4	μ_5	p_5	Estimate	RE
1	10^5	3.0	40.0	0.500	20.0	0.500	25.0	50.0	60.0	0.500	—	—
2	10^5	10.5	36.9	0.535	17.4	0.641	20.0	45.0	55.2	0.215	$2.510 \cdot 10^{-55}$	0.3631
3	10^5	13.3	36.7	0.464	19.4	0.564	16.7	47.6	57.8	0.185	$8.026 \cdot 10^{-55}$	0.0604
4	10^5	13.0	39.8	0.433	19.8	0.589	15.3	49.8	59.8	0.168	$7.822 \cdot 10^{-55}$	0.0235
5	10^5	12.9	39.7	0.431	19.6	0.595	15.3	49.4	59.6	0.170	$7.495 \cdot 10^{-55}$	0.0144
6	10^5	13.0	39.7	0.431	19.7	0.594	15.4	49.6	59.5	0.168	$7.686 \cdot 10^{-55}$	0.0477
7	10^5	13.0	39.7	0.430	19.7	0.594	15.4	49.7	59.4	0.170	$7.602 \cdot 10^{-55}$	0.0170

Table 4 **Simulation Results for the Five-Node Network with Equally Loaded Queues;** $l_0 = 4$, $l = 20$

Iteration	Repl.	λ	μ_1	p_1	μ_2	p_2	μ_3	μ_4	μ_5	p_5	Estimate	RE
1	10^7	3.0	40.0	0.500	20.0	0.500	50.0	50.0	60.0	0.500	—	—
2	10^7	12.3	35.6	0.575	16.0	0.647	44.0	43.1	51.8	0.193	$7.942 \cdot 10^{-16}$	0.0272
3	10^7	20.5	35.0	0.588	14.8	0.667	43.2	42.1	50.0	0.160	$7.759 \cdot 10^{-16}$	0.0164
4	10^7	20.6	35.3	0.581	15.1	0.658	43.1	42.0	50.0	0.161	$7.875 \cdot 10^{-16}$	0.0231
5	10^7	20.6	35.2	0.586	15.0	0.651	43.1	42.1	50.4	0.165	$7.659 \cdot 10^{-16}$	0.0102
6	10^7	21.0	35.0	0.580	15.3	0.656	43.1	41.9	50.5	0.160	$7.522 \cdot 10^{-16}$	0.0090
7	10^7	21.0	34.9	0.578	15.4	0.657	43.1	42.0	50.5	0.157	$7.679 \cdot 10^{-16}$	0.0132

6. Concluding Remarks

In this paper, we have presented an efficient CE method for estimating buffer overflow probabilities in queueing networks via simulation. For an $M/M/1$ queue we have proved analytically two properties (instability and robustness) and provided strong evidence for the third property (CE optimality), and we have conjectured that these three properties also hold for more general queueing networks. We have also explained why the method works well in terms of the three properties. Numerical results support this conjecture and demonstrate the high efficiency of the proposed algorithm for queueing networks up to five queues.

The simulation method used is in principle well known from earlier work (Parekh and Walrand 1989): IS with a state-independent exponential change of measure. As a consequence, our method in principle handles the same classes of models as earlier work: networks of a (possibly large) number of queues, with random routing, with the constraint that a state-independent change of measure should be sufficient (the latter constraint excludes models like those in Glasserman and Kou 1995). However, there are models for which earlier approaches may not be suitable for practical reasons, while ours is. This can either be due to the number of iterations required (our method usually needs only three) or to the complexity of the calculations involved (as with the example in §5.2).

Some issues for further research are the following:

- Extension of the proofs of the three properties to more general queueing models.
- Further investigation of the behavior of the ratio estimators of type (24) for the $M/M/1$ queue and more general queueing models.
- Finding conditions under which a state-independent change of measure, as used in this method, can or cannot lead to an (asymptotically) efficient simulation.

Acknowledgments

The authors thank Søren Asmussen for his helpful comments and for the proof of Appendix B, and Zinovy Landsman for introducing them to the natural exponential family. The third author acknowledges the support of the Binational Science Foundation (grant 191-574) for this project.

Appendix A. Natural Exponential Families

Consider a univariate family of distributions with densities (pmfs, pdfs) $\{f_\theta,\ \theta \in \Theta\}$, for some subset $\Theta \subset \mathbb{R}$. The family is said to be a NEF if

$$f_\theta(x) = \mathrm{e}^{x\theta - \kappa(\theta)}\, h(x), \tag{A1}$$

where h is a positive (normalization) function (Morris 1982, Jorgensen 1997).

For example, if we take $\theta = \lambda/\sigma^2$ and $\kappa(\theta) = \sigma^2\theta^2/2$, then f_θ is the density of the $\mathsf{N}(\lambda, \sigma^2)$ distribution, where σ^2 is fixed. Similarly, for $\theta = -\lambda$ and $\kappa(\theta) = -r\log(-\theta) = -r\log\lambda$, we obtain the class of gamma distributions with shape parameter r (fixed) and scale parameter λ. Note that in the latter case $\Theta = (-\infty, 0)$. There are many NEFs. In fact, every distribution with pdf f_0 for which the moment generating function exists in a neighborhood of 0 generates its own NEF by letting κ be the cumulant function

$$\kappa(\theta) = \log \int \mathrm{e}^{\theta x}\, f_0(x)\, dx$$

and by defining

$$f_\theta(x) = \mathrm{e}^{\theta x - \kappa(\theta)}\, f_0(x),$$

with Θ the largest interval for which the cumulant function exists. We say that f_θ is obtained from f_0 by an exponential twist/tilt with twisting/tilting parameter θ.

Now let X have a distribution in some NEF $\{f_\theta\}$. It is not difficult to see that

$$v := \mathbb{E}_\theta X = \kappa'(\theta) \quad \text{and} \quad \mathrm{Var}_\theta\, X = \kappa''(\theta).$$

Because κ' is increasing, we may reparametrize the family using the mean v. In particular, to the NEF above corresponds a family $\{g_v\}$ such that for each pair (θ, v) satisfying $\kappa'(\theta) = v$ we have $g_v = f_\theta$. For example, for the NEF corresponding to the gamma distribution discussed above we have $\kappa'(\theta) = -r/\theta = v$, and hence

$$g_v(x) = \mathrm{e}^{\theta x + r\log(-\theta)}\, h(x) = \mathrm{e}^{-(r/v)x} \left(\frac{r}{v}\right)^r h(x).$$

Now consider (8) for the case where X is a random variable from a NEF $\{f(\cdot; v)\}$, reparametrized by the mean v. Hence,

$$f(x; v) = f_{\theta(v)}(x) = g_v(x) = \exp\big(\theta(v)x - \kappa(\theta(v))\big)h(x),$$

where $\theta(v)$ is some differentiable function of v. We wish to maximize, with respect to $\tilde{v}$, the function D

defined as

$$D(\tilde{v}) = \mathbb{E}_{v_j} H(X) W(X; v, v_j) \log f(X; \tilde{v}).$$

Solving $D'(\tilde{v}) = 0$ for $\tilde{v}$ gives

$$\begin{aligned}\mathbb{E}_{v_j} H(X) W(X; v, v_j) \{\theta'(\tilde{v}) X - \kappa'(\theta(\tilde{v}))\theta'(\tilde{v})\} \\ = \mathbb{E}_{v_j} H(X) W(X; v, v_j) \theta'(\tilde{v})(X - \tilde{v}) = 0,\end{aligned}$$

which is solved for $\tilde{v} = v^*$, with

$$v^* = \frac{\mathbb{E}_{v_j} H(X) W(X; v, v_j) X}{\mathbb{E}_{v_j} H(X) W(X; v, v_j)}. \tag{A2}$$

That v^* is a global maximum follows from the convexity of D and the fact that

$$D''(v^*) = -\theta'(v^*) \mathbb{E}_v H(X) < 0,$$

because $\theta'(v^*) = 1/\operatorname{Var}_{v^*}(X) > 0$.

Similarly, the sample version of (A2) is given by

$$\widehat{v^*} = \frac{\sum_{i=1}^N H(X^{(i)}) W(X^{(i)}; v, v_j) X^{(i)}}{\sum_{i=1}^N H(X^{(i)}) W(X^{(i)}; v, v_j)},$$

where $X^{(1)}, \ldots, X^{(N)}$ is a random sample from $f(\cdot; v_j)$.

Appendix B. Instability for Random Walks

Consider a random walk $\{S_n,\ n = 1, 2, \ldots\}$ with $S_n = X_1 + \cdots + X_n$ such that the common distribution of the X_k belongs to a NEF indexed by θ, and generated by some density f_0, as in Appendix A. Let $\kappa(\theta) = \log \mathbb{E}_0 \mathrm{e}^{\theta S_1}$ be the corresponding cumulant function.

Let $T = \inf\{n > 0 \colon S_n > x \text{ or } S_n \le 0\}$. Suppose we are interested in simulating $\gamma = \mathbb{P}_\theta(A)$, with $A = \{S_T > x\}$, using IS with tilting parameter $\tilde{\theta}$. We will show that if an optimal $\tilde{\theta} = {}_*\theta$ (in the sense of minimizing the variance) exists, then necessarily $\kappa'({}_*\theta) > 0$ so that the IS distribution has positive drift (in queueing terminology, $\rho > 1$).

The estimator to be simulated from $\mathbb{P}_{\tilde{\theta}}$ is $Z(\tilde{\theta}) = W_T(\theta, \tilde{\theta}) I_A$, where

$$W_n(\theta, \tilde{\theta}) = \exp\bigl\{(\theta - \tilde{\theta}) S_n - n(\kappa(\theta) - \kappa(\tilde{\theta}))\bigr\}.$$

Let θ_1 be arbitrary with $\kappa'(\theta_1) < 0$ and let $\theta_2 > \theta_1$ be defined by $\kappa(\theta_2) = \kappa(\theta_1)$. Because $\kappa'(\theta_2) > 0$ by convexity, our result will then follow if we can show that $\operatorname{Var}_{\theta_2} Z(\theta_2) < \operatorname{Var}_{\theta_1} Z(\theta_1)$, which, because the means are both γ, is the same as $\mathbb{E}_{\theta_1} Z^2(\theta_1) < \mathbb{E}_{\theta_2} Z^2(\theta_2)$. But

$$\begin{aligned}\mathbb{E}_{\theta_2} Z^2(\theta_2) &= \mathbb{E}_{\theta_2}[W_T^2(\theta, \theta_2); A] \\ &= \mathbb{E}_\theta[W_T(\theta_2, \theta) W_T^2(\theta, \theta_2); A] \\ &= \mathbb{E}_\theta[W_T(\theta, \theta_2); A] \\ &= \mathbb{E}_\theta\bigl[\exp\{(\theta - \theta_2) S_T - T(\kappa(\theta) - \kappa(\theta_2))\}; A\bigr]\end{aligned}$$

Similarly,

$$\mathbb{E}_{\theta_1} Z^2(\theta_1) = \mathbb{E}_\theta\bigl[\exp\{(\theta - \theta_1) S_T - T(\kappa(\theta) - \kappa(\theta_1))\}; A\bigr].$$

The result now follows from $\theta_2 > \theta_1$, $S_T > 0$, and $\kappa(\theta_1) = \kappa(\theta_2)$.

Remark B.1. Note that the "instability property" above is applicable in a queueing theory context provided that we can write the process of interest as a random walk. Examples are the actual waiting time process in a $GI/G/1$ queue or the process describing the number of customers in the system just before arrival times in a $G/M/1$ queue.

References

Alon, G., D. P. Kroese, T. Raviv, R. Y. Rubinstein. 2005. Application of the cross entropy method for buffer allocation problem in simulation-based environment. *Ann. Oper. Res.* Forthcoming.

Asmussen, S., R. Y. Rubinstein. 1995. Complexity properties of steady-state rare events simulation in queueing models. J. Dshalalow, ed. *Advances in Queueing: Theory, Methods and Open Problems*, Vol. 1. CRC Press, 429–462.

Asmussen, S., R. Y. Rubinstein, C.-L. Wang. 1994. Regenerative rare event simulation via likelihood ratios. *J. Appl. Probab.* **31**(3) 797–815.

de Boer, P. T. 2000. Analysis and efficient simulation of queueing models of telecommunications systems. Ph.D. thesis, University of Twente, Enschede, The Netherlands.

de Boer, P. T., V. F. Nicola. 2001. Adaptive state-dependent importance sampling simulation of Markovian queueing networks. *Eur. Trans. Telecomm.* **13**(4) 303–315.

de Boer, P. T., V. F. Nicola, R. Y. Rubinstein. 2000. Adaptive importance sampling simulation of queueing networks. J. A. Joines, R. R. Barton, K. Kang, P. A. Fishwick, eds. *Proc. 2000 Winter Simulation Conf.*, Orlando, FL, 646–655.

Devetsikiotis, M., J. K. Townsend. 1993a. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Trans. Networking* **1** 293–305.

Devetsikiotis, M., J. K. Townsend. 1993b. An algorithmic approach to the optimization of importance sampling parameters in digital communication system simulation. *IEEE Trans. Comm.* **41** 1464–1473.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications I*, 3rd ed. John Wiley and Sons, New York.

Frater, M. R., T. M. Lennon, B. D. O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Trans. Automat. Control* **36** 1395–1405.

Glasserman, P., S.-G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Model. Comput. Simulation* **5**(1) 22–42.

Jorgensen, B. 1997. *The Theory of Dispersion Models*. Chapman and Hall, London, U.K.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simulation* **5**(1) 43–85.

Kapur, J. N., H. K. Kesavan. 1992. *Entropy Optimization Principles with Applications*. Academic Press, New York.

Kriman, V., R. Y. Rubinstein. 1997. Polynomial time algorithms for estimation of rare events in queueing models. J. Dshalalow, ed. *Frontiers in Queueing: Models and Applications in Science and Engineering*. CRC Press, New York, 421–448.

Lieber, D., R. Y. Rubinstein, D. Elmakis. 1997. Quick estimation of rare events in stochastic networks. *IEEE Trans. Reliability* **46**(2) 254–265.

Morris, C. N. 1982. Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**(1) 65–80.

Parekh, S., J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Control* **34** 54–66.

Qaq, W. A. al-, M. Devetsikiotis, J. K. Townsend. 1995. Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems. *IEEE Trans. Comm.* **43** 2975–2985.

Rubinstein, R. Y. 1997. Optimization of computer simulation models with rare events. *Eur. J. Oper. Res.* **99** 89–112.

Rubinstein, R. Y. 1999. The simulated entropy method for combinatorial and continuous optimization. *Methodology Comput. Appl. Probab.* **2** 127–190.

Rubinstein, R. Y. 2001a. Combinatorial optimization via cross-entropy. S. Gass, C. Harris, eds. *Encyclopedia of Operations Research and Management Sciences*. Kluwer, Dordrecht, The Netherlands, 102–106.

Rubinstein, R. Y. 2001b. Combinatorial optimization, cross-entropy, ants, and rare events. S. Uryasev, P. M. Pardalos, eds. *Stochastic Optimization: Algorithms and Applications*. Kluwer, Dordrecht, The Netherlands, 304–358.

Rubinstein, R. Y. 2002. Cross-entropy and rare events for maximal cut and partition problems. *ACM Trans. Model. Comput. Simulation* (TOMACS) **12**(1) 27–53.

Rubinstein, R. Y., B. Melamed. 1998. *Modern Simulation and Modeling*. John Wiley and Sons, New York.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*. John Wiley and Sons, New York.

Sadowsky, J. S. 1991. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. Automat. Control* **36**(12) 1383–1394.