Effective Bandwidth and Fast Simulation of ATM Intree Networks *

Cheng-Shang Chang †

Philip Heidelberger

Sandeep Juneja ‡

Perwez Shahabuddin

IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

November 1992

Abstract

We consider the efficient estimation, via simulation, of very low buffer overflow probabilities in certain acyclic ATM queueing networks. We apply the theory of effective bandwidths and Markov additive processes to derive an asymptotically optimal simulation scheme for estimating such probabilities for a single queue with multiple independent sources, each of which may be either a Markov modulated process or an autoregressive processes. This result extends earlier work on queues with either independent arrivals or with a single Markov modulated arrival source. The results are then extended to estimating loss probabilities for intree networks of such queues. Experimental results show that the method can provide many orders of magnitude reduction in variance in complex queueing systems that are not amenable to analysis.

*Performance Evaluation, Vol. 20, pp. 45-66, 1994.

†Current address: Dept. of EE, National Tsing Hua University, Hsinchu 30043, Taiwan, R.O.C.

‡Current Address: Department of Operations Research, Stanford University, Stanford, California 94305

1 INTRODUCTION

An open and challenging problem for high speed digital networks is to allocate bandwidths to various real-time services with time-varying rates, e.g., voice and video, such that certain grades of services, e.g., delay constraint and loss probability, are satisfied and reasonable network utilization is achieved. A tentative solution for this problem is to "characterize" services or sources with rate functions that are parameterized by the grades of services. In other words, the rate function maps to the (minimum) amount of channel capacity needed for a certain grade of service. This mapping also reduces the bandwidth allocation problem to the classical connection networks with *constant* rates, where flow control and admission control are well studied. The rate functions that characterize services are called the effective or equilibrium bandwidths in the literature ([18, 22, 16, 8, 9, 23] among many others).

In [9], a general theory of effective bandwidths is developed for the tail distributions of buffer occupancy (queue length) in a network. The effective bandwidths defined in [9] range between average rates and peak rates, i.e., peak rate assignment for the highest grade of service and average rate assignment for the lowest grade of service. It is also shown in [9] that (i) the effective bandwidths of multiplexing independent sources is less than the sum of the effective bandwidth of each source, (ii) the distributions of delays and queue lengths are bounded by exponential tails when the effective bandwidth of the input traffic is less than the channel capacity, and (iii) the effective bandwidth of output traffic is less than the effective bandwidth of the input traffic. These results, especially (ii), provide a general framework to obtain first order approximation of delays and queue lengths in ATM networks. More specifically, if q_s denotes the steady-state queue length distribution (assuming it exists), these results allow a characterization of the decay rate θ^* of the distribution of q_s, i.e., $\lim_{x\to\infty} \log Pr(q_s > x)/x = -\theta^*$.

However, these results do not provide the exact value of $Pr(q_s > x)$, and if such exact values are required, then simulation is often the only possible approach. For a switch in an ATM network, the buffer length x must be set so that $Pr(q_s > x)$ is very small. In such situations, very long simulation run lengths are required to obtain accurate estimates of the buffer loss probability $Pr(q_s > x)$. This fact has motivated much of the work on "fast simulation" techniques for queueing systems using importance sampling [19, 15]. In importance sampling, the system is simulated using a new set of input probability distributions that accelerate the occurrence of the rare event under study. Unbiased estimates of the probability of the rare event are then obtained by multiplying the output by the appropriate likelihood ratio.

In this paper, we will demonstrate that the theory of effective bandwidths can also be used to devise an efficient fast simulation technique for obtaining accurate estimates of queue length tail distributions in intree ATM networks. Specifically we consider a discrete time queueing system with constant service times and multiple independent input sources where the number of arrivals from each input source per time slot may form an autocorrelated sequence, e.g., Markov modulated or autoregressive. An intree network is an (acyclic) tree network with packets flowing from the leaves to the root of the tree; external source arrivals can also occur at any node in the tree. For such networks with many heterogeneous sources, analytic results are generally unavailable and numerical procedures are limited by the "curse of dimensionality." We will describe both an effective importance sampling change of measure, and an effective simulation sampling plan which appropriately uses the change of measure so as to produce estimates having low variance.

To derive the change of measure for a single queue, we establish the connection between the theory of effective bandwidths in [9] and the large deviations results for Markov additive processes [6, 7, 20, 29, 30]. This involves using an "exponentially twisted" distribution (see, e.g., [2, 3, 21, 25, 34, 33]) appropriately modified to account for the correlation structure. This modification, which is called the "conjugate process," involves solving a certain eigenvalue-eigenvector problem. The eigenvalue problem is shown to be the same as the problem of finding the appropriate effective bandwidths in [9]. Since the sources are independent, the overall effective bandwidth is the sum of the effective bandwidths of the individual sources. Thus each effective bandwidth and eigenvalue can be found individually, thereby avoiding the dimensionality problem. For simulation purposes, each source is twisted by an amount defined by the solution of a nonlinear equation that couples the effective bandwidths of the sources. We note that a similar concept for solving the eigenvalue problem has been applied to risk theory in [25]. To derive the change of measure for an intree network, we establish an input-output relation analogous to that in [9]. This input-output relation provides an approximation of exponentially twisted distribution for the departure process by choosing an appropriate exponentially twisted distribution for the input process.

However, the direct application of these results to estimating steady-state buffer overflow probabilities is not immediate. Since the use of importance sampling breaks down when applied over arbitrarily long time intervals [15], the regenerative method is typically used to obtain steady-state estimates. This limits the extent of the likelihood ratio to regenerative cycles, thereby resulting in a stable variance. For example, in availability models, to estimate the steady-state unavailability importance sampling can be used to estimate the expected system down time during a cycle while standard simulation is used to estimate the expected cycle length [17]; this technique was called "Measure Specific Dynamic Importance Sampling." The term dynamic arises from the fact that importance sampling is "turned off" for the rest of a cycle whenever the system reaches a failure state.

Unfortunately, the applicability of regenerative methodology for ATM networks is limited, since regenerations will occur very infrequently; for example, if there are M Markov modulated sources, then a regeneration occurs whenever the queues are at some fixed lengths (say empty) and all M sources are in a fixed state. For a large number of sources and/or queues, regenerations may be even rarer than buffer overflows. This problem is addressed in [12, 13] by changing the importance sampling so as to accelerate regenerations, or simply treating entrances to a fixed queue length configuration as approximately independent (approximate) regenerations.

In this paper, we address the problem from a firm theoretical foundation by applying the methodology described in [31] that was developed for estimating steady-state unavailability when regenerative structure is not present. We start by defining an "A-cycle" to begin whenever the process enters a fixed subset A of states. Such A-cycles are no longer iid, however $Pr(q_s > x)$ can still be expressed as the ratio of the expected amount of time that the queue length exceeds x during an A-cycle divided by the expected length of an A-cycle. These expectations are taken with respect to the stationary distribution, conditioned on being in A. Then a technique much like Measure Specific Dynamic Importance Sampling is applied. The model is run for a long time so that it is approximately "in steady-state." Whenever A is entered a "splitting" technique [19] is used in which two separate simulations of the next A-cycle are run. Importance sampling, using the appropriate conjugate processes, is performed to estimate the numerator. This importance sampling is turned off whenever the queue length exceeds x for the duration of the A-cycle. A standard simulation is then run to estimate the denominator. The end of this second simulation also serves as a probabilistically correct starting point for of the next pair of A-cycles, since where it re-enters A has (asymptotically) the stationary distribution conditioned on being in A. The method of batch means [5] is then used to obtain confidence intervals; this method correctly deals with the correlation between the A-cycles.

The rest of the paper is organized as follows. In Section 2 we first review the theory of effective bandwidths and then treat the case of a single queue. Specifically, we show that our approach is asymptotically optimal (under certain technical conditions). This extends earlier work on a single queue with either independent arrivals or with a single Markov modulated arrival source. These single queue results are then extended to intree networks in Section 3, providing a change of measure and an upper bound on the variance. While we have not proven that this is optimal, this change of measure is likely to be near optimal; to our knowledge, these are the first such results in queueing networks with correlated arrivals. Experimental results, showing many orders of magnitude reduction in variance are described in Section 4, and the paper is summarized in Section 5.

2 USING EFFECTIVE BANDWIDTHS IN FAST SIMULATION FOR A SINGLE QUEUE

In this section, we will first introduce the discrete-time queueing model for ATM networks (see [35] and references therein) and briefly review the theory of envelope processes and effective bandwidths in [9]. Then we will demonstrate how one may use the theory in fast simulation.

2.1 The model for a single queue with multiclasses of customers

Consider a discrete-time queue model with K classes of customers. The service requirements of these K class customers are one unit of time. Also, the buffer size is infinite and the server can serve c customers per unit of time. The constant c will be referred to as the capacity of the server. Let $a_k(t)$, $k = 1, \ldots, K$, be the number of class k arrivals at time t and $a(t) = \sum_{k=1}^{K} a_k(t)$ be the total number of arrivals at time t. We assume that these K arrival processes are *independent*. Denote $q(t)$ as the number of customers in the queue at time t. Then under a work-conservative policy the queue is governed by the following Lindley's equation:

$$q(t+1) = (q(t) + a(t+1) - c)^+ \tag{1}$$

where $(x)^+ \stackrel{\text{def}}{=} \max(0, x)$. Furthermore, we assume that the queue is empty at time 0.

The performance measure of interest is the tail distribution of the queue length, i.e.,

$$Pr(q(t) \geq x) \leq e^{-\theta x}, \tag{2}$$

where θ is a network design parameter to satisfy the grade of service. For instance, if the buffer size is x, then (2) implies that the probability of buffer overflow is approximately $e^{-\theta x}$. The objective is to characterize these K classes of customers such that one knows if (2) is satisfied. The characterizations used in this paper are called the *minimum envelope rates* or the *effective bandwidths* in the next subsection.

2.2 Envelope processes and effective bandwidths

In this subsection, we introduce the characterization in [9]. View the arrival process $\{a(t), t = 0, 1, 2, \ldots\}$ as a *sequence of random variables* and the characterization of $a(t)$ as a *transformation* of a sequence of random variables. Let $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. Consider the following bounding process of $a(t)$:

$$\frac{1}{\theta} \log E e^{\theta A(t_1, t_2)} \leq \hat{A}(\theta, t_2 - t_1) \ \ \forall\, t_1 \leq t_2 \tag{3}$$

The process $\hat{A}(\theta, t)$ is called an *envelope process* (EP) of $a(t)$ with respect to θ. The minimum envelope process with respect to θ (MEP) and *minimum envelope rate* of $a(t)$ with respect to θ (MER) are defined to be

$$A^*(\theta, t) = \sup_{s \geq 0} \frac{1}{\theta} \log E e^{\theta A(s, s+t)} \tag{4}$$

$$a^*(\theta) = \limsup_{t \to \infty} \frac{A^*(\theta, t)}{t}. \tag{5}$$

The MER $a^*(\theta)$ is then called the *effective bandwidth* of the process $a(t)$. An equivalent interpretation of $a^*(\theta)$ is that for every $\epsilon > 0$, there exists a constant (in t) $\hat{\sigma}(\theta)$ such that

$$A^*(\theta, t) \leq (a^*(\theta) + \epsilon)t + \hat{\sigma}(\theta). \tag{6}$$

It is shown in [9] that $a^*(\theta)$ is increasing in θ and it is between the average rate and the peak rate (if it exists). Moreover, it is continuous in θ if $a(t)$ is bounded for all t. Based on these characterizations of arrival processes, a few simple rules were derived, including superposition, a single queue performance analysis, input-output relations, etc. These rules were used to analyze or approximate the performance of ATM networks. In the next section, we will extend these characterizations and show that one may use them to run fast simulations.

2.3 Envelope processes and fast simulation for a single queue

In this section, we first extend the notion of *envelope processes* and then discuss the connections between envelope processes and fast simulations.

Definition 2.1 *we call a triplet* $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ *a* θ*-upper envelope process* ($\theta -$ *UEP*) *of* $a(t)$ *if for all* t *and* a_s, $s = 1, \ldots, t$,

$$L(\omega)\frac{\exp(\theta \sum_{s=1}^{t} a_s)}{e^{\theta \hat{a}(\theta) t}} \leq e^{\theta \hat{\sigma}(\theta)}, \tag{7}$$

where $L(\omega)$ *is the likelihood ratio with* ω *corresponding to the appropriate sample path, i.e.,*

$$L(\omega) = \frac{Pr(a(s) = a_s, s = 1, \ldots, t)}{Pr(\tilde{a}(s;\theta) = a_s, s = 1, \ldots, t)}. \tag{8}$$

We shall assume that $0 \leq \sigma(\theta) < \infty$. *Similarly, we call a triplet* $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ *a* θ*-lower envelope process* ($\theta -$ *LEP*) *if the left hand side of (7) is bounded below by* $\exp(-\theta\hat{\sigma}(\theta))$.

Note that if $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ is a $\theta -$ UEP of $a(t)$, then

$$Ee^{\theta \sum_{s=1}^{t} a(s)} = EL(\omega)e^{\theta \sum_{s=1}^{t} \tilde{a}(s;\theta)} \leq e^{\theta(\hat{a}(\theta)t + \hat{\sigma}(\theta))}. \tag{9}$$

Thus, $\hat{a}(\theta)t + \hat{\sigma}(\theta)$ is indeed an *envelope process* of $a(t)$ defined in (3) if $a(t)$ is stationary. A Similar result also holds for $\theta -$ LEP.

Now we interpret the physical meaning of the triplet $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ as follows: consider the event $\mathcal{E}^x$ the queue length exceeds x before it returns to empty given the queue is started from empty. (Estimating $Pr(q_s \geq x)$ is closely related to estimating $Pr(\mathcal{E}^x)$.) To obtain an estimate for this event, one can run N independent simulations and form the estimate $\frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{\mathcal{E}^x}(\omega_n)$, where ω_n is the sample path of the n^{th} simulation. Suppose that $Ea(t) < c$ and x is large, then the event is *rare*, i.e., most of the simulations will render a value 0. In that case, one requires sufficiently large N to achieve a reasonably good estimate of this event. If this event is rare, one can speed up the occurrences of the event. To do so, run the N simulation with the process $\tilde{a}(t;\theta)$ instead. Form the estimate $\frac{1}{N}\sum_{n=1}^{N} L(\omega_n)\mathbf{1}_{\mathcal{E}^x}(\omega_n)$. In other words, we change the probability measure Pr with a new measure $\tilde{P}r$ such that under $\tilde{P}r$ the behavior of the process $a(t)$ is the same as $\tilde{a}(t;\theta)$, i.e., for all $t \geq 1$,

$$\tilde{P}r(a(1) = a_1, \ldots, a(t) = a_t) = Pr(\tilde{a}(1;\theta) = a_1, \ldots, \tilde{a}(t;\theta) = a_t). \tag{10}$$

This technique is known as importance sampling. With the equality in (10), in this paper the likelihood ratio $L(\omega)$ will denote either $Pr(\omega)/\tilde{P}r(\omega)$ or (8) according to the appropriate context. Suppose that the event $\mathcal{E}^x$ occurs at time t. Then $A(0,t) \geq ct + x$. Thus

$$L(\omega) \leq \frac{e^{\theta\hat{\sigma}(\theta)} e^{\theta\hat{a}(\theta)t}}{e^{\theta(ct+x)}} \leq e^{\theta\hat{\sigma}(\theta)} e^{-\theta x}, \tag{11}$$

for all $\omega \in \mathcal{E}^x$ and for all $\hat{a}(\theta) \leq c$. Since the likelihood ratio is uniformly bounded for all t and for all sample paths in $\mathcal{E}^x$,

$$Pr(\mathcal{E}^x) = E\mathbf{1}_{\mathcal{E}^x}(\omega) = \tilde{E}L(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega) \leq e^{\theta\hat{\sigma}(\theta)} e^{-\theta x} \tilde{P}r(\mathcal{E}^x), \tag{12}$$

where $\tilde{E}$ denotes the expectation taken under the measure $\tilde{P}r$. Thus, under the new measure, the event $\mathcal{E}^x$ is not as *rare* as before. This also implies that when x is large, one should use the change of measure θ^* such that $\hat{a}(\theta) = c$. Overall, one can consider the set of triplets as a family of candidates for the change of measure that yields the "fastest" simulation.

Now we define what we mean by "fastest" simulation. Note that the estimator is consistent and unbiased since $\tilde{E}L(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega) = E\mathbf{1}_{\mathcal{E}^x}(\omega) = P(\mathcal{E}^x)$. For any consistent estimator that uses the change of measure, the variance is $\tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega) - Pr(\mathcal{E}^x)^2$. Thus, the variance is minimized if $\tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega)$ is minimized. Since the variance is nonnegative, $\tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega) \geq Pr(\mathcal{E}^x)^2$. Thus,

$$\lim_{x\to\infty} \frac{\log \tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega)}{\log Pr(\mathcal{E}^x)} \geq 2. \tag{13}$$

In this paper, we will say an estimator is *asymptotically optimal* if the limit in (13) is 2.

In the following, we outline the algorithm that runs a simulation to estimate the probability of the event $\mathcal{E}^x$. Later, we will show that this algorithm is indeed asymptotically optimal under certain technical conditions.

Algorithm 2.2 Suppose that each input process, $k = 1, \dots, K$, has a $\theta-$UEP $(\tilde{a}_k(t;\theta), \hat{a}_k(\theta), \hat{\sigma}_k(\theta))$.

1. Let θ^* be the solution to the equation $\sum_{k=1}^K \hat{a}_k(\theta) = c$.

2. Run N independent simulations using the processes $\tilde{a}_k(t;\theta^*)$. Form the estimator of $E\mathbf{1}_{\mathcal{E}^x}$ as $\frac{1}{N}\sum_{n=1}^N L(\omega_n)\mathbf{1}_{\mathcal{E}^x}(\omega_n)$, where L is the likelihood ratio of the original process $a(t) = \sum_{k=1}^K a_k(t)$ to the process $\tilde{a}(t;\theta^*) = \sum_{k=1}^K \tilde{a}_k(t;\theta^*)$.

Before we prove the main result, i.e., the asymptotically optimality of Algorithm 2.2, we need the following two lemmas. Analogous to Lemma 2.4 in [9], one can verify easily the following results for the superposition of independent processes. (Bear in mind that $\min(x_1/y_1, x_2/y_2) \le (x_1+x_2)/(y_1+y_2) \le \max(x_1/y_1, x_2/y_2)$, $y_1, y_2 \ge 0$.)

Lemma 2.3 *If $a_k(t)$, $k = 1, \dots K$, are K independent processes and each of them has a $\theta-$UEP (resp. $\theta-$LEP) $(\tilde{a}_k(t;\theta), \hat{a}_k(\theta), \hat{\sigma}_k(\theta))$, then $a(t) = \sum_{k=1}^K a_k(t)$ has a a $\theta-$ UEP (resp. $\theta-$LEP) $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$, where $\tilde{a}(t;\theta) = \sum_{k=1}^K \tilde{a}_k(t;\theta)$, $\hat{a}(\theta) = \sum_{k=1}^K \hat{a}_k(\theta)$, and $\hat{\sigma}(\theta) = \sum_{k=1}^K \hat{\sigma}_k(\theta)$.*

In the following lemma, we derive both bounds for $Pr(\mathcal{E}^x)$ that will be used in the proof of our main theorem.

Lemma 2.4 *(i) (Upper bound) If $a(t)$ has a $\theta-$ UEP $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$, then*

$$\limsup_{x\to\infty} \frac{\log Pr(\mathcal{E}^x)}{x} \le -\theta^*, \tag{14}$$

where $\theta^* = \sup\{\theta : \hat{a}(\theta) \le c\}$.

(ii) (Lower bound) If $a(t)$ is upper bounded, i.e., $a(t) \le a_{\max} < \infty$ a.s., and $a(t)$ has a $\theta-$LEP $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ that satisfies

$$Pr(\sum_{s=1}^t \tilde{a}(s;\theta) > ct, \forall t \ge 1) > 0, \tag{15}$$

then

$$\liminf_{x\to\infty} \frac{\log Pr(\mathcal{E}^x)}{x} \ge -\theta^*, \tag{16}$$

where $\theta^* = \inf\{\theta : \hat{a}(\theta) \ge c\}$.

Proof. Since the proofs are similar, we only show (ii). Suppose that the event $\mathcal{E}^x$ occurs at time t. One has $A(0,t) \leq ct + x + a_{\max} - c$ since $a(t) \leq a_{\max}$ and $A(0,t-1) \leq c(t-1) + x$. If $\hat{a}(\theta) \geq c$,

$$L(\omega) \geq e^{-\theta\hat{\sigma}(\theta)} e^{-\theta x} e^{\theta(c - a_{\max})}. \tag{17}$$

Note that this lower bound is independent of t and it holds for all $\omega \in \mathcal{E}^x$. Since $Pr(\mathcal{E}^x) = \tilde{E}L\mathbf{1}_{\mathcal{E}^x}$, one has

$$Pr(\mathcal{E}^x) \geq e^{-\theta\hat{\sigma}(\theta)} e^{-\theta x} e^{\theta(c - a_{\max})} \tilde{P}r(\mathcal{E}^x). \tag{18}$$

Note that $\tilde{P}r(\mathcal{E}^x) \geq Pr(\sum_{s=1}^t \tilde{a}(s;\theta) > ct, \forall t \geq 1) > 0$ by (15). Taking log on both sides of (18) and letting $x \to \infty$ yields

$$\liminf_{x\to\infty} \frac{\log Pr(\mathcal{E}^x)}{x} \geq -\theta. \tag{19}$$

Choose the smallest θ such that $\hat{a}(\theta) \geq c$.

For the proof of (i), compare the argument for lower bound with the derivations of (11) and (12). $\square$

Now we show that Algorithm 2.2 is indeed asymptotically optimal under the conditions in Theorem 2.5. (In Section 2.4 we will show that the conditions of this theorem are satisfied for Markov and autoregressive arrival processes.)

Theorem 2.5 *Suppose that a triplet $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ is both a $\theta - UEP$ and $\theta - LEP$ of $a(t)$ and that $\hat{a}(\theta)$ is* strictly increasing *and* differentiable *for $\theta > 0$, i.e., $(\hat{a}(\theta))' > 0$ for $\theta > 0$.*

(i) The sequence $\{\sum_{s=1}^t a(s), t \geq 1\}$ obeys the large deviation principle (see [6]) with the rate function $I(v) = \sup_\theta\{\theta v - \theta\hat{a}(\theta)\}$.

(ii) $(\theta\hat{a}(\theta))'$ is increasing in θ, and

$$\lim_{t\to\infty} \frac{\sum_{s=1}^t \tilde{a}(s;\theta)}{t} = (\theta\hat{a}(\theta))' \ \ a.s. \tag{20}$$

(iii) Let θ^* be the unique solution of $\hat{a}(\theta) = c$ (if any). If, furthermore, $a(t)$ is upper bounded by a constant $a_{\max}$ and $a(t;\theta^*)$ is stationary *and* ergodic*, then*

$$\lim_{x\to\infty} \frac{\log Pr(\mathcal{E}^x)}{x} = -\theta^*. \tag{21}$$

(iv) Under (iii), the estimator in algorithm 2.2 is asymptotically optimal.

We note that the condition $\hat{a}(\theta)$ is *strictly increasing* excludes the constant arrival case $a(t) = a$ for all t. In such a case, $\hat{a}(\theta) = a$ and $\tilde{a}(t;\theta) = a(t) = a$. Since the constant arrival case is trivial, it will not be discussed further.

Proof. (i) From (9) and a similar result for $\theta -$ LEP, it follows that

$$\lim_{t\to\infty} (1/t)\log(E\exp(\theta A(0,t)) = \theta\hat{a}(\theta). \tag{22}$$

The large deviation principle then follows immediately from the Gärtner-Ellis theorem (see [6]).

(ii) From (22), it follows that $(\theta\hat{a}(\theta))$ is convex since the log of a moment generating function is convex. Thus, $(\theta\hat{a}(\theta))'$ is increasing.

Let $v = (\theta\hat{a}(\theta))'$ and $\tilde{A}(0,t;\theta) = \sum_{s=1}^{t}\tilde{a}(s;\theta)$. We will prove that for all $\epsilon > 0$, there exists a $\delta > 0$ such that $Pr(|\tilde{A}(0,t;\theta)/t - v| > \epsilon) < Ke^{-\delta t}$ for some finite constant K and for all t. Thus, $\sum_{t=1}^{\infty} Pr(|\tilde{A}(0,t;\theta)/t - v| > \epsilon) < \infty$. The strong law of large number then follows from Borel-Cantelli's lemma.

From Chernoff's bound, it follows that for $y \geq 0$

$$Pr(\tilde{A}(0,t;\theta) > (v+\epsilon)t) \leq E\exp(y\tilde{A}(0,t;\theta))\exp(-y(v+\epsilon)t). \tag{23}$$

Since $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ is a $\theta -$ LEP of $a(t)$,

$$E\exp(y\tilde{A}(0,t;\theta)) = E\frac{1}{L(\omega)}\exp(yA(0,t)) \leq \exp\left[-\theta(\hat{a}(\theta)t + \hat{\sigma}(\theta))\right]E\exp[(\theta+y)A(0,t)]. \tag{24}$$

Using the assumption that $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ is also a $\theta -$ UEP of $a(t)$, we have from (9) that

$$E\exp(y\tilde{A}(0,t;\theta)) \leq \exp\left[-\theta(\hat{a}(\theta)t + \hat{\sigma}(\theta))\right]\exp[(\theta+y)(\hat{a}(\theta+y)t + \hat{\sigma}(\theta+y))]. \tag{25}$$

Thus,

$$\begin{aligned} & Pr(\tilde{A}(0,t;\theta) > (v+\epsilon)t) \\ & \leq \exp\left[-\theta\hat{\sigma}(\theta) + (\theta+y)\hat{\sigma}(\theta+y)\right]\exp\left[(-\theta\hat{a}(\theta) + (\theta+y)\hat{a}(\theta+y) - y(v+\epsilon))t\right]. \end{aligned} \tag{26}$$

Since $(\theta\hat{a}(\theta))' = v$, one can choose y small enough such that the exponent of the second term is strictly negative. A similar argument can be used for $Pr(\tilde{A}(0,t;\theta) < (v-\epsilon)t)$.

(iii) It suffices to prove the condition in (15) of Lemma 2.4. Since $\hat{a}(\theta^*) = c$ and $\hat{a}(\theta)$ is strictly increasing, $(\theta^*\hat{a}(\theta^*))' > c$. It then follows from (ii) and the stationarity and ergodicity of $\tilde{a}(t;\theta^*)$

that $E\tilde{a}(t;\theta^*) > c$. Thus, there exists a $\epsilon > 0$ such that $E\tilde{a}(t;\theta^*) > c + \epsilon$. Note that

$$
\begin{aligned}
& Pr(\sum_{s=1}^{t} \tilde{a}(s;\theta^*) > ct, \forall t \geq 1) \\
& \geq Pr(\sum_{s=1}^{t} \tilde{a}(s;\theta^*) \geq (c+\epsilon)t, \forall t \geq 1) = Pr(\max_{t\geq 1}[0, \sum_{s=1}^{t} c + \epsilon - \tilde{a}(s;\theta^*)] = 0). \qquad (27)
\end{aligned}
$$

Consider a $G/G/1$ queue with interarrival times $\tilde{a}(-t;\theta^*)$ (the extended time reversed sequence) and constant service times $c+\epsilon$. The last term in (27) is simply the probability that a customer sees an empty queue. Since $E\tilde{a}(t;\theta^*) > c+\epsilon$, it then follows from the classical queueing theory that this probability is strictly larger than 0 (see [4] (2.4.1)).

(iv) Observe from (11) that

$$\tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega) \leq e^{2\theta^*\hat{\sigma}(\theta^*)}e^{-2\theta^* x}. \qquad (28)$$

It then follows from (iii) that

$$\lim_{x\to\infty} \frac{\log \tilde{E}L^2(\omega)\mathbf{1}_{\mathcal{E}^x}(\omega)}{\log Pr(\mathcal{E}^x)} \leq 2. \qquad (29)$$

In conjunction with (13), one concludes that the estimator using $\theta = \theta^*$ is asymptotically optimal. $\square$

2.4 Examples

In this subsection, we present the $\theta-$UEP and $\theta-$LEP processes for finite state Markov arrival Processes and autoregressive processes. These processes are related to the θ-*conjugate* processes studied in [6, 29, 30, 25] and they will be shown to satisfy the conditions in Theorem 2.5 and thus Algorithm 2.2 is asymptotically optimal.

Example 2.6 Consider an M-state irreducible and aperiodic Markov chain $x(t)$. A process $a(t)$ is called a *Markov Arrival Process* (MAP) $a(t)$ if $(a(t), x(t))$ a Markov chain with the transition kernel, $P_{i,j}(a) = Pr(a(t+1) \leq a, x(t+1) = j | x(t) = i)$. Let $\Gamma(\theta)$ be an $M \times M$ matrix with $\Gamma_{i,j}(\theta) = \int e^{\theta a} dP_{i,j}(a)$, $\theta \geq 0$. Then it can be easily verified from the backward equation that

$$\lim_{t\to\infty} \frac{1}{t} \log Ee^{\theta A(0,t)} = \log(\mathrm{sp}(\Gamma(\theta))),$$

where $\text{sp}(\Gamma(\theta))$ is the spectral radius of the matrix $\Gamma(\theta)$, i.e., the maximal eigenvalue of the *nonnegative* matrix $\Gamma(\theta)$. Thus, the MER or the effective bandwidths, $a^*(\theta) = \frac{1}{\theta}\log(\text{sp}(\Gamma(\theta)))$. Moreover, $a^*(\theta)$ is strictly increasing and differentiable since $\log(\text{sp}(\Gamma(\theta)))$ is analytic and strictly convex except the case $a^*(\theta)$ is a constant for all θ (see [20, 28]).

The θ-conjugate process of an MAP is also an MAP $(\tilde{a}(t;\theta), \tilde{x}(t;\theta))$ with the transition kernel

$$\begin{aligned} &Pr(\tilde{a}(t+1;\theta) = a, \tilde{x}(t+1;\theta) = j|\tilde{x}(t;\theta) = i) \\ &= \frac{1}{\text{sp}(\Gamma(\theta))}\frac{f_j(\theta)}{f_i(\theta)}e^{\theta a} Pr(a(t+1) = a, x(t+1) = j|x(t) = i), \end{aligned} \tag{30}$$

where the transpose of the row vector $f(\theta) = (f_1(\theta), \ldots, f_M(\theta))$ is the right eigenvector of the matrix $\Gamma(\theta)$ corresponding to the eigenvalue $\text{sp}(\Gamma(\theta))$, i.e., $\text{sp}(\Gamma(\theta))f(\theta)^T = \Gamma(\theta)f(\theta)^T$. Since $x(t)$ is irreducible and aperiodic, the vector $(f_1(\theta), \ldots, f_M(\theta))$ is positive and it can be normalized to $\sum_{i=1}^M f_i(\theta) = 1$.

Now we will show that the likelihood ratio $L(\omega)$ is uniformly bounded. Consider a particular sample path ω of t units of time, starting at $x(0) = \tilde{x}(0;\theta) = i_0$ for some initial state i_0. For $s = 1, \ldots, t$, let i_s be the state of ω at time s and a_s be the number of arrivals of ω at time s. Then

$$\begin{aligned} L(\omega) &= \frac{Pr(a(1) = a_1, \ldots, a(t) = a_t)}{Pr(\tilde{a}(1;\theta) = a_1, \ldots, \tilde{a}(t;\theta) = a_t)} \\ &= \frac{\prod_{s=0}^{t-1} Pr(a(s+1) = a_{s+1}, x(s+1) = i_{s+1}|x(s) = i_s)}{\prod_{s=0}^{t-1} Pr(\tilde{a}(s+1;\theta) = a_{s+1}, \tilde{x}(s+1;\theta) = i_{s+1}|\tilde{x}(s;\theta) = i_s)} = \frac{f_{i_0}(\theta)}{f_{i_t}(\theta)}\frac{\text{sp}(\Gamma(\theta))^t}{\exp(\theta\sum_{s=1}^t a_s)}. \end{aligned}$$

Thus,

$$\min_{i,j}\frac{f_i(\theta)}{f_j(\theta)} \le L(\omega)\frac{\exp(\theta\sum_{s=1}^t a_s)}{e^{\theta a^*(\theta)t}} \le \max_{i,j}\frac{f_i(\theta)}{f_j(\theta)}. \tag{31}$$

Note that both the lower and upper bounds are independent of the sample path. It then follows from (31) that $(\tilde{a}(t;\theta), a^*(\theta), \frac{1}{\theta}\log(\max_{i,j}\frac{f_i(\theta)}{f_j(\theta)}))$ is both a $\theta-$UEP and a $\theta-$LEP of $a(t)$. $\Box$

As a special case of MAP, consider a Markov modulated process (MMP) $a(t) = y_{x(t)}(t)$, where $x(t)$ is an irreducible and aperiodic Markov chain on the states $\{1, \ldots, M\}$ with the transition matrix $\boldsymbol{r}$, i.e., $r_{i,j}$ is the transition probability from state i to state j, and $\{y_i(t), t = 0, 1, \ldots\}$, $i = 1, \ldots, M$, are M sequences of i.i.d. random variables with the moment generating functions $\phi_i(\theta) = E\exp(\theta y_i(0))$. Clearly, $Pr(a(t+1) \le a, x(t+1) = j|x(t) = i) = Pr(y_j(t+1) \le a)r_{i,j}$ and thus

$$a^*(\theta) = \frac{1}{\theta}\log\text{sp}(\boldsymbol{r}\boldsymbol{\phi}(\theta)), \tag{32}$$

where $\boldsymbol{\phi}(\theta)$ is the diagonal matrix diag$\{\phi_1(\theta), \ldots, \phi_M(\theta)\}$ and sp$(\boldsymbol{r\phi}(\theta))$ is the spectral radius of the matrix $\boldsymbol{r\phi}(\theta)$.

The θ-conjugate process of an MMP is also an MMP, $\tilde{a}(t;\theta) = \tilde{y}_{\tilde{x}(t;\theta)}(t;\theta)$, where $\tilde{x}(t;\theta)$ is a M-state Markov process with the transition matrix $\tilde{r}$, and $\{\tilde{y}_i(t), t = 0, 1, \ldots\}$, $i = 1, \ldots, M$, are M sequences of i.i.d. random variables. The distribution of $\tilde{y}_i(t;\theta)$ is related to that of $y_i(t)$ via

$$Pr(\tilde{y}_i(t;\theta) = x) = \frac{e^{\theta x}}{\phi_i(\theta)} Pr(y_i(t) = x). \tag{33}$$

The transition matrix $\tilde{r}$ is obtained as follows:

$$\tilde{r}_{i,j} = \frac{1}{\text{sp}(\boldsymbol{r\phi}(\theta))} \frac{1}{f_i(\theta)} r_{i,j} \phi_j(\theta) f_j(\theta), \tag{34}$$

where the transpose of the row vector $f(\theta) = (f_1(\theta), \ldots, f_M(\theta))$ is the right eigenvector of the matrix $\boldsymbol{r\phi}(\theta)$ corresponding to the eigenvalue sp$(\boldsymbol{r\phi}(\theta))$.

For the usual voice model, one has for a two-state Markov chain that

$$a^*(\theta) = \frac{1}{\theta} \log \Big(\frac{r_{11}\phi_1(\theta) + r_{22}\phi_2(\theta) + \sqrt{(r_{11}\phi_1(\theta) - r_{22}\phi_2(\theta))^2 + 4r_{12}r_{21}\phi_1(\theta)\phi_2(\theta)}}{2} \Big). \tag{35}$$

In particular, if $a(t)$ is a sequence of i.i.d. random variables, then $a^*(\theta) = (1/\theta)\log(E\exp(\theta a(0)))$ The effective bandwidths, $a^*(\theta)$, for i.i.d. random variables and two-state Markov chains were reported in [18, 22, 16].

We note that $x(t)$ in the definition of the MAP needs not to be a finite state Markov chain. The state space can be extended to a general space, e.g., $\mathcal{R}^n$. However, this extension in general complicates the problem of solving the maximal eigenvalue of an Markov chain on a general state space. An interesting example where one can solve the maximal eigenvalue problem is the autoregressive process. Note that the autoregressive process has been used as model for video in [27].

Example 2.7 Consider a standard autoregressive process $a(t)$ that is driven by a sequence of $M \times 1$ i.i.d. random vectors $u(t) = (u_1(t), \ldots, u_M(t))$:

$$a(t) = vx(t), \quad x(t) = \boldsymbol{r}x(t-1) + u(t),$$

where $\boldsymbol{r}$ is an $M \times M$ constant matrix, v is a $1 \times M$ constant vector, and $x(t)$'s are $M \times 1$ state vectors. We will assume that sp$(\boldsymbol{r}) < 1$ so that $\boldsymbol{I} - \boldsymbol{r}$ is invertible. Clearly, an autoregressive process is an MMP with a *continuous* state space on $\mathcal{R}^M$. Note that

$$\sum_{s=1}^{t} a(s) = \sum_{s=1}^{t} \upsilon \sum_{j=0}^{t-s} \boldsymbol{r}^j u(s) + \sum_{j=1}^{t} \upsilon \boldsymbol{r}^j x_0 \tag{36}$$

if $x(0) = x_0$. Thus, one has the MER or the effective bandwidths $a^*(\theta) = \frac{1}{\theta} \log E e^{\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} u(t)}$ in this example. It is easy to see that $a^*(\theta)$ is strictly increasing and differentiable since the log of a moment generating function is analytic and strictly convex except the case with constant arrivals.

The θ-conjugate process of an autoregressive process is again an autoregressive process with the following form:

$$\tilde{a}(t;\theta) = \upsilon \tilde{x}(t;\theta), \quad \tilde{x}(t;\theta) = \boldsymbol{r} \tilde{x}(t-1;\theta) + \tilde{u}(t;\theta),$$

where the distribution of $\tilde{u}(t;\theta)$ is related to that of $u(t)$ through the equation

$$Pr(\tilde{u}(t;\theta) = u_t) = \frac{e^{\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} u_t}}{E e^{\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} u(t)}} Pr(u(t) = u_t). \tag{37}$$

for any $M \times 1$ vector $u_t = (u_{t,1}, \ldots, u_{t,M})$.

We can also show that the likelihood ratio $L(\omega)$ is uniformly bounded. Consider a particular sample path ω of t units of time , starting $x(0) = \tilde{x}(0;\theta) = x_0$. Since both $\{u(t), t \geq 1\}$ and $\{\tilde{u}(t;\theta), t \geq 1\}$ are sequences of i.i.d. random vectors,

$$L(\omega) = \frac{Pr(u(s) = u_s, s = 1, \ldots, t)}{Pr(\tilde{u}(s;\theta) = u_s, s = 1, \ldots, t)} = \frac{(e^{\theta a^*(\theta)})^t}{e^{\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} \sum_{s=1}^{t} u_s}}.$$

Applying 36 yields

$$L(\omega) \frac{e^{\theta \sum_{s=1}^{t} a_s}}{e^{\theta a^*(\theta) t}} = e^{-\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} \sum_{s=1}^{t} \boldsymbol{r}^{t-s+1} u_s} e^{\theta \upsilon \sum_{j=1}^{t} \boldsymbol{r}^j x_0}. \tag{38}$$

To obtain a uniformly bounded likelihood ratio, we assume that the matrices υ and $\boldsymbol{r}$ are nonnegative and that $u(t)$ is nonnegative and bounded, i.e., there exists a vector $u_{\max}$ such that $\mathbf{0} \leq u(t) \leq u_{\max}$ a.s., where $\mathbf{0}$ is a zero vector with the appropriate dimension. Note that

$$\begin{aligned} 1 &\geq \exp(-\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} \sum_{s=1}^{t} \boldsymbol{r}^{t-s+1} u_s) \geq \exp(-\theta \upsilon (\boldsymbol{I} - \boldsymbol{r})^{-1} \sum_{s=1}^{t} \boldsymbol{r}^{t-s+1} u_{\max}) \\ &= \exp(-\theta \upsilon (\boldsymbol{r} - \boldsymbol{r}^{t+1}) u_{\max}) \geq \exp(-\theta \upsilon \boldsymbol{r} u_{\max}) \end{aligned} \tag{39}$$

and that (assuming $\mathbf{0} \leq x_0 \leq u_{\max}$)

$$e^{\theta v(\boldsymbol{I}-\boldsymbol{r})^{-1}u_{\max}} \geq e^{\theta v \sum_{j=1}^{t} \boldsymbol{r}^j x_0} \geq 1.$$

Thus, $(\tilde{a}(t;\theta), a^*(\theta), v(\boldsymbol{I}-\boldsymbol{r})^{-1}u_{\max})$ is both a a $\theta-$UEP and a $\theta-$LEP of $a(t)$. $\square$

In particular, if $a(t+1) = \alpha a(t) + u(t)$ with $u(t)$ being a sequence of i.i.d. random variables with supports on $[0, u_{\max}]$, then the θ-conjugate process is driven by $\tilde{u}(t;\theta)$ that satisfies

$$Pr(\tilde{u}(t;\theta) = u) = \frac{e^{\theta u/(1-\alpha)}}{Ee^{\theta u(t)/(1-\alpha)}} Pr(u(t) = u). \tag{40}$$

The likelihood ratio for sample paths of t units of time, starting from $0 \leq x(0) = \tilde{x}(0;\theta) = x_0 \leq \max$, is uniformly bounded as follows:

$$e^{-\theta(1-\alpha)^{-1}u_{\max}} \leq L(\omega)\frac{\exp(\theta \sum_{s=1}^{t} a_s)}{e^{\theta a^*(\theta)t}} \leq e^{\theta(1-\alpha)^{-1}u_{\max}}. \tag{41}$$

From these two examples and Lemma 2.3, it follows that if $a(t)$ is a superposition of bounded processes that are either an MAP in Example 2.6 or an autoregressive process in Example 2.7, then it satisfies the conditions in Theorem 2.5 with $\hat{a}(\theta) = a^*(\theta)$.

3 FAST SIMULATION OF INTREE NETWORKS

In this previous section, we have discussed the fast simulation method for a single queue, using the notion of $\theta-$UEP. Note that the departure processes of the queues at the higher level are simply the input processes of the queues at the lower level. Thus, if one can obtain $\theta-$UEP of the departure processes from the $\theta-$UEP of the input processes, then one can use Algorithm 2.2 to simulate the queues in the downstream. In the following theorem, we establish this input-output relation.

Let $\{b(t), t = 1, 2, \ldots\}$, denote the number of departures from a single queue at time t.

Theorem 3.1 *Consider a discrete-time queue as in §2.1 with $q(0) = q_0$. Suppose that the input process $a(t)$ has a θ_1-UEP $(\tilde{a}(t;\theta_1), \hat{a}(\theta_1), \hat{\sigma}(\theta_1))$ and that $\hat{a}(\theta_1)$ is strictly increasing and differentiable for $\theta_1 > 0$. Let $\tilde{\theta}$ be a solution of $(\theta_1\hat{a}(\theta_1))' = c$. Define a function $\psi(\theta)$ to be $\psi(\theta) = \hat{a}(\theta)$ for $\theta \leq \tilde{\theta}$ and $\psi(\theta) = c - \frac{\tilde{\theta}}{\theta}(c - \hat{a}(\tilde{\theta}))$ for $\theta > \tilde{\theta}$. Then the departure process $b(t)$ has a θ_2-UEP $(\tilde{b}(t;\theta_2), \psi(\theta_2), (1/\theta_2)(\theta_1 q_0 + \theta_1\hat{\sigma}(\theta_1)))$ with $\theta_1 = \min[\theta_2, \tilde{\theta}]$, and $\tilde{b}(t;\theta_2)$ is the departure process when the input process is $\tilde{a}(t;\theta_1)$. Moreover, $\psi(\theta_2)$ is strictly increasing and differentiable for $\theta_2 > 0$.*

The reason why $\psi(\theta)$ is defined separately can be explained as follows: recall from Theorem 2.5 that $\tilde{a}(t;\theta_1)$ has the mean $(\theta_1\hat{a}(\theta_1))'$ (if it is also a $\theta_1 -$ LEP). Thus, when $(\theta_1\hat{a}(\theta_1))' < c$, the queue is "stable" and the departure process should have the same characteristics as the input process. When $(\theta_1\hat{a}(\theta_1))' \geq c$, the queue will build up and the departure process is truncated at the rate c. The exact form of $\psi(\theta)$ can then be obtained by convex transforms as in [23] which will be reported in the full paper.

Intuitively, this input-output relation provides an approximation of exponentially twisted distribution for the departure process by choosing appropriate exponentially twisted distribution for the input process. That $\theta_1 = \min[\theta_2, \tilde{\theta}]$ means that one should not overdrive the queue in simulation.

Proof. We need to show that

$$L(\omega)\frac{e^{\theta_2 B(0,t)}}{e^{\theta_2 \psi(\theta_2)}} \leq e^{\theta_1 q_0} e^{\theta_1 \hat{\sigma}(\theta_1)}. \tag{42}$$

Consider the following two cases:

Case 1. $\theta_2 \leq \tilde{\theta}$. In this case, $\psi(\theta_2) = \hat{a}(\theta_2)$ and $\theta_1 = \theta_2$. Thus,

$$\frac{e^{\theta_2 B(0,t)}}{e^{\theta_2 \psi(\theta_2)t}}\frac{e^{\theta_1 \hat{a}(\theta_1)}}{e^{\theta_1 A(0,t)}} = \frac{e^{\theta_2 B(0,t)}}{e^{\theta_2 A(0,t)}} \leq e^{\theta_2 q_0} \tag{43}$$

since $B(0,t) \leq A(0,t) + q_0$. Multiplying both sides of (7) with the inequality in (43) yields the inequality in (42).

Case 2. $\theta_2 > \tilde{\theta}$. In this case, $\psi(\theta_2) = c - \frac{\tilde{\theta}}{\theta_2}(c - \hat{a}(\tilde{\theta}))$ and $\theta_1 = \tilde{\theta}$. Thus,

$$\frac{e^{\theta_2 B(0,t)}}{e^{\theta_2 \psi(\theta_2)t}}\frac{e^{\theta_1 \hat{a}(\theta_1)}}{e^{\theta_1 A(0,t)}} = \frac{e^{\tilde{\theta} B(0,t)}}{e^{\tilde{\theta} A(0,t)}}\frac{e^{(\theta_2 - \tilde{\theta})B(0,t)}}{e^{(\theta_2 - \tilde{\theta})ct}} \leq e^{\tilde{\theta} q_0} \tag{44}$$

since $B(0,t) \leq A(0,t) + q_0$ and $B(0,t) \leq ct$. Multiplying both sides of (7) with the inequality in (44) yields the inequality in (42).

That $\psi(\theta_2)$ is strictly increasing and differentiable follows from the same properties of $\hat{a}(\theta_1)$ except at $\theta_2 = \tilde{\theta}$. At $\theta_2 = \tilde{\theta}$, the left derivative of $\psi(\theta_2)$ is $(\hat{a}(\tilde{\theta}))'$ and the right derivative is $(c - \hat{a}(\tilde{\theta}))/\tilde{\theta}$. Since $\tilde{\theta}$ is a solution of $(\theta\hat{a}(\theta))' = c$, we conclude that $\psi(\theta_2)$ is also differentiable at $\tilde{\theta}$. □

We note that if $(\theta\hat{a}(\theta))' < c$ for all θ, we can simply let $\tilde{\theta} = \infty$. Also, if $(\theta_1\hat{a}(\theta_1))'$ is strictly increasing, e.g., the MAP and the autoregressive processes in Examples 2.6 and 2.7, then one can verify that $(\theta_2\psi(\theta_2))'$ is also strictly increasing up to c. Thus, in our simulations later on, the solution of the equation $(\theta\hat{a}(\theta))' = c$ is always *unique*.

We also note that any function in the family of functions, $\{\psi(\theta,\theta_0);\theta_0\}$ with $\psi(\theta,\theta_0) = \hat{a}(\theta)$ for $\theta \le \theta_0$ and $\psi(\theta,\theta_0) = c - \frac{\theta_0}{\theta}(c - \hat{a}(\theta_0))$ for $\theta > \theta_0$, would be a $\theta -$ UEP of the departure process as in Theorem 3.1. However, one can show that when the solution of the equation $(\theta\hat{a}(\theta))' = c$ is unique, $\psi(\theta)$ (with $\theta_0 = \tilde{\theta}$) is the best one since $\psi(\theta) = \inf_{\theta_0}\psi(\theta,\theta_0)$.

Consider an intree network with J queues. Label the J queues in the *postorder* (left subtree, right subtree, root). Denote $q_j(t)$ and c_j the queue length of queue j at time t and its capacity, respectively. Assume that each queue follows the Lindley's equation as the discrete-time queue in §2.1. If the output process of queue i is an input process of queue j, we say queue j is a *successor* of queue i and queue i is a *predecessor* of queue j. Let $S(j)$ (resp. $P(j)$) be the set of queues that are *successors* (resp. predecessors) of queue j. Since the network is an intree, $S(i)$ only contains one element. Also, we have from the postorder labelling that $i < j$ for all $i \in P(j)$ and $S(j) > j$. Beside the arrivals from the queues in the predecessors, there are k_j external arrival processes to queue j with $a_{j,k}(t)$, $k = 1,\dots,k_j$, denoting the number of arrivals at time t. Assume that all the external arrival processes are *independent* and that the external input process $a_{j,k}(t)$ has a $\theta -$ UEP $(\tilde{a}(t;\theta),\hat{a}(\theta),\hat{\sigma}(\theta))$. Our objective here is to to find appropriate changes of measures of the external processes to efficiently estimate the performance at queue J (the root). In so doing, we need to decide the parameters $\theta^*_{j,k}$ so that $a_{j,k}(t)$ is simulated by $\tilde{a}_{j,k}(t;\theta^*_{j,k})$. The following is an algorithm for finding appropriate $\theta^*_{j,k}$.

Algorithm 3.2

1. For $j = 1,\dots,J$, define recursively the following functions:

$$\hat{a}_j(\theta) = \sum_{k=1}^{k_j}\hat{a}_{j,k}(\theta) + \sum_{k\in P(j)}\psi_k(\theta) \tag{45}$$

$$\psi_j(\theta) = \begin{cases} \hat{a}_j(\theta), & if\ \ \theta \le \tilde{\theta}_j; \\ c_j - \frac{\tilde{\theta}_j}{\theta}(c_j - \hat{a}_j(\tilde{\theta}_j)), & otherwise, \end{cases} \tag{46}$$

where $\tilde{\theta}_j$ is a solution of $(\theta\hat{a}_j(\theta))' = c_j$.

2. Solve the unique solution θ^*_J from the equation $\hat{a}_J(\theta) = c_J$ ($\theta^*_J = \infty$ if $\hat{a}_J(\theta) < c_J$ for all θ).

3. For $j = J,\dots,1$, define recursively $\theta^*_j = \min[\theta^*_{S(j)},\tilde{\theta}_j]$.

4. Let $\theta^*_{j,k} = \theta^*_j$ for all $k = 1, \ldots, k_j, j = 1, \ldots, J$.

Once $\theta^*_{j,k}$ are determined, we can run the simulation and form the estimator as in Algorithm 2.2. Unfortunately, we are not able to show the above algorithm is asymptotically optimal since we only have the upper bound, i.e., θ − UEP (cf. Lemma 2.4). What we can say is that the second moment and the variance of the estimate decays at rate $\exp(-2\theta^*_J x)$ (see (28)) and that $Pr(\mathcal{E}^x)$ (for queue J) decays at most at the rate $\exp(-\theta^*_J x)$ (see Lemma 2.4(i)). One would expect improvement via Algorithm 3.2 if $Pr(\mathcal{E}^x)$ decays at least at the rate $\exp(-2\theta^*_J x)$. As demonstrated in the next section, Algorithm 3.1 indeed shows significant variance reduction.

4 EXPERIMENTAL RESULTS

In the previous sections, we have shown how one chooses the appropriate changes of measures for MAP and autoregressive processes in estimating $Pr(\mathcal{E}^x)$. In this section, we will further discuss the simulation methodology for estimating the steady state probabilities, e.g. tail distributions of queue lengths and loss probabilities. We will also report various experiment results for loss probabilities for both a single queue case and an intree network case.

4.1 Simulation methodology for steady state probabilities

Now we consider the problem of estimating the steady state probability $\alpha = P(q_s \geq x)$ through simulation. The standard way of estimating steady state measures is the method of batch means (see [5]). If the system is regenerative, then the regenerative method can also be used (see [11]). For the MMP case, if we aggregate the input processes and the queue length into one process, then the resulting process is a finite state Markov chain and thus a regenerative process (any state of the aggregated Markov chain may be used as the regenerative state). However if we have many input processes then regenerative cycles will be typically very long, and this method is not very efficient. In such cases it is advisable to use the method of batch means. For rare event simulation in which we want to use importance sampling, the application of the batch means method is not straightforward. We use a method similar to that which was used in [31] for estimating the steady state unavailability through importance sampling in non-Markovian highly dependable systems. Let π denote the steady state distribution of the aggregated input process at the times when the queue is empty. Let $\{T_{0i}, i \geq 1\}$ be the time between successive instants when the queue is empty. We will define an "A-cycle" to be the process between such times. Define $Y = \int_0^{T_0 - 1} \mathbf{1}_{\{q(t) \geq x\}} dt$. Then (see, for example, [10])

$$\alpha = Pr(q_s \geq x) = \frac{E_\pi Y}{E_\pi T_0} \tag{47}$$

where the subscript in the expectation sign indicates that at time $t = 0$ the aggregated input process is distributed with π. Note that the denominator is easy to estimate using standard (naive) simulation, but the numerator includes the rare event $\{q(t) \geq x\}$ and an appropriate change of measure is needed. Note that the main problem is to make the process go along the most likely path to level x. For this, the same change of measure that was used for estimating $Pr(\mathcal{E}^x)$ can be used. (Note that $Y > 0$ iff $\mathbf{1}_{\mathcal{E}^x} = 1$.) Once the process hits x in an A-cycle, then we turn the importance sampling off for the remainder of the cycle. Hence we write

$$E_\pi Y = \tilde{E}_\pi Y L \tag{48}$$

where now the $\tilde{E}$ denotes the change of measure using the $\theta^* -$ UEP defined in Algorithm 2.2 until the process hits x and then using the original process for the remainder of the cycle.

Now we can adopt the following procedure. We first run a few A-cycles of the original process so that it reaches steady state and the aggregated input process has (approximately) the distribution π at the times when the queue is empty. After that, at the beginning of each A-cycle, we start two parallel processes, one with the change of measure and the other without the change of measure. The ones with the change of measure are used to get samples of Y and L. The ones without the change of measure are used to get samples of T_0 and ensures that the aggregated process again has distribution π the next time the queue is empty. Let us say we run N such sets of parallel A-cycles to get the samples Y_i, L_i and T_{0i}, $1 \leq i \leq N$, of Y, L and T_0 respectively. Then $\hat{\alpha} = \sum_{i=1}^N Y_i L_i / \sum_{i=1}^N T_{0i}$ is a natural estimator for α. But since the successive Y_i's (also L_i's and T_{0i}'s) are not independent, care has to be taken in estimating the variance constant in the central limit theorem (CLT) for $\hat{\alpha}$. We use the method of batch means.

In this method we group the successive Y_i's (L_i's and T_{0i}'s) into b batches, each batch containing k successive Y_i's (L_i's and T_{0i}'s) i.e., $kb = N$. Let $\delta_j = \sum_{i=(j-1)k+1}^{jk} Y_i L_i / k$ and $\gamma_j = \sum_{i=(j-1)k+1}^{jk} T_{0i}/k$. Then we can estimate α by $\hat{\alpha} = \hat{\delta}/\hat{\gamma}$ where $\hat{\delta} = \sum_{j=1}^b \delta_j / b$ and $\hat{\gamma} = \sum_{i=1}^b \gamma_j / b$. Note that this is the same estimator as before. However, now we are able to construct confidence intervals. The method of batch means works on the assumption that for k sufficiently large the successive δ_j's (γ_j's) are (approximately) normally distributed and independent. In that case, for b also large, $\sqrt{b}(\hat{\alpha} - \alpha) \approx N(0, \sigma^2)$ where

$$\sigma^2 = \frac{Var(\delta_j) + 2\alpha Cov(\delta_j, \gamma_j) + \alpha^2 Var(\gamma_j)}{E(\gamma_j)}. \tag{49}$$

The normal approximation can then be used to construct confidence intervals. In particular, for sufficiently large values of b and k an approximate 99% confidence interval for α is given by $(\hat{\alpha} - 2.56\sqrt{\sigma^2/b}, \hat{\alpha} + 2.56\sqrt{\sigma^2/b})$. The relative error (RE) is defined to be $2.56\sqrt{\sigma^2/b}/\alpha$. In practice, since σ^2 is unknown, we estimate it by estimating $E(\gamma_j)$ by $\hat{\gamma}$, $Var(\delta_j)$ by $(\sum_{j=1}^{b} \delta_j^2/b - \hat{\delta}^2)$, $Var(\gamma_j)$ by $(\sum_{j=1}^{b} \gamma_j^2/b - \hat{\gamma}^2)$, $Cov(\delta_j, \gamma_j)$ by $(\sum_{j=1}^{b} \gamma_j \delta_j/b - \hat{\gamma}\hat{\delta})$ and α by $\hat{\alpha}$ and then using (49).

4.2 Experimental results for loss probabilities in single queues

In this subsection, we estimate the loss probability of a single queue with buffer size x and capacity c. We slightly change the recursion in (1) to take into account the fact that all customers that arrive when the buffer size is x are lost, i.e., the queue size never exceeds x. The new recursion is given by

$$q(t+1) = (\min(q(t) + a(t+1), x) - c)^+ \tag{50}$$

Also, instead of estimating $P(q_s > x)$ we estimate the steady state loss probability, i.e., Y is now $\int_{t=0}^{T_0-1} (q(t) + a(t+1) - x)^+ dt$. Note that for Y to be positive, the process, starting from a state where the queue length is zero (and the Markov arrival process has state distribution π) has to cross level $x - c$ (which is approximately equal to x for x large) before the queue length becomes zero again. Also, note that the new recursion is the same as the old recursion until the process crosses x. So intuitively speaking, the change of measure for estimating $P(q_s > x)$ under the previous dynamics, should also work for the new quantity of interest under the new dynamics (given by (50)).

We consider two single queue examples, one with MMP arrival processes and the other with autoregressive arrival processes. For the first example we took 16 heterogeneous MMP sources. Each source is a two-state MMP process taking values in $\{1, 2\}$. The r_{11}'s and r_{22}'s of the 16 sources were set to

$$(0.9, 0.9, 0.9, 0.9, 0.8, 0.8, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.7, 0.6, 0.6, 0.5)$$

and

$$(0.7, 0.8, 0.6, 0.5, 0.3, 0.6, 0.6, 0.4, 0.8, 0.5, 0.9, 0.6, 0.5, 0.8, 0.6, 0.9)$$

respectively. Note that for all the sources $r_{11} + r_{22} - 1 > 0$ so that the individual input streams are positively correlated (as is natural to assume in practice). For each source, the number of arrivals in state 1 and state 2 are 0 and 1 respectively, i.e., $y_1(t) = 0$ and $y_2(t) = 1$ for all t. The steady-state total input rate can then be calculated to be 6.51. The capacity c was taken to be 8 so that the traffic intensity for the queue was 0.814. Using Algorithm 2.2, step 1, the optimal θ^* came out to be 0.34375. In this case the aggregated Markov chain has $2^{16} \times x$ states and is therefore difficult to solve by numerical (non-simulation) techniques.

x	Importance Sampling Estimate	Standard Simulation Estimate
10	$1.16 \times 10^{-2} \pm 2.8\%$	$1.17 \times 10^{-2} \pm 3.6\%$
15	$1.78 \times 10^{-3} \pm 3.3\%$	$1.63 \times 10^{-3} \pm 8.4\%$
20	$3.17 \times 10^{-4} \pm 3.2\%$	$3.14 \times 10^{-4} \pm 18.5\%$
40	$3.76 \times 10^{-7} \pm 3.4\%$	$1.27 \times 10^{-7} \pm 259\%^{*}$
80	$5.24 \times 10^{-13} \pm 3.4\%$	$0.00\pm ?$

Table 1: Estimates of steady-state loss probability in the single queue with MMP arrival.

For the autoregressive input processes we chose two types of sources, with 8 sources of each type. The sources generate arrival streams of the form $a(t+1) = \alpha a(t) + u(t)$ where $\alpha < 1$ and $u(t)$ is a sequence of i.i.d random variables with supports on $[0, u_{\max}]$. For the first type we took $\alpha = 0.6$ and $u(t)$ uniformly distributed on $[0, 0.4]$. For the second type we took $\alpha = 0.5$ and $u(t)$ uniformly distributed on $[0, 0.6]$. The steady state total input rate can then be calculated to be 8.8. The c was taken to be 10 so that the traffic intensity for the queue was 0.88. The optimal θ^* for the change of measure came out to be 1.5. This is a continuous state Markovian process and so it is again very difficult to solve numerically for the loss probability.

We ran experiments for several values of x for both the MMP case and the autoregressive case. These are reported in Table 1 and Table 2 respectively. In each case we first ran the process under the original measure for 300 A-cycles for it to reach steady state. For the importance sampling case we simulated for a total of 120000 A-cycles (60000 each, for the numerator and the denominator); 2000 batches with 60 A-cycles in each batch (the variance was fairly insensitive to the batch size for the same total number of A-cycles). Note how in this case the RE remains almost constant as x increases. As the value of x increases, the average number of events per A-cycle increase, leading to an increase in the computational effort, but it is still orders of magnitude better than standard simulation. (For example, in Table 1, the $x = 10$ entry took 16 CPU seconds whereas the $x = 80$ entry took 149 CPU seconds). For comparison purposes we also ran standard simulations (i.e., using (47) without the change of measure) for each x for the same CPU time (we again use a batch size of 60) as in simulation with importance sampling. Note how the RE error in this case blows up as x increases (the $*$ indicates that the estimate had not yet stabilized). In fact for the autoregressive case for $x = 15$, we found out (after running long enough so that the variance had stabilized) that the standard simulation will have to run 2500 times longer to get an estimate of the same RE as through importance sampling. Also, as shown in the table, for some of the higher values of x we did not get any samples of the event of interest.

x	Importance Sampling Estimate	Standard Simulation Estimate
10	$1.31 \times 10^{-3} \pm 3.6\%$	$1.35 \times 10^{-3} \pm 4.2\%$
15	$2.22 \times 10^{-7} \pm 4.2\%$	$3.96 \times 10^{-7} \pm 184.0\%^{*}$
20	$1.11 \times 10^{-10} \pm 4.2\%$	$0.00 \pm ?$
30	$3.07 \times 10^{-17} \pm 4.4\%$	$0.00 \pm ?$

Table 2: Estimates of steady-state loss probability in the single queue with autoregressive arrival.

4.3 Simulation Methodology and Experimental Results for Intree Networks

Next we consider intree networks. For the first example we consider a model with two queues, where the input of the first queue feeds into the second queue. There are a total five (two state MMP) arrival processes, the first three of which input into the first queue and last two input into the second queue. The r_{11}'s and the r_{22}'s for the five arrival processes were taken to be (0.8,0.5,0.7,0.6,0.7) and (0.5,0.6,0.8,0.6,0.6) respectively thus making the external arrival rates 1.44 and 0.93 respectively. The c's for the two queues were chosen to be 2 and 3 respectively.

Since we have already shown how to estimate the loss probability in the first queue, we now estimate only the loss probability in the second queue. Using Algorithm 3.2, steps 1 to 4, the $\tilde{\theta}_2$ and θ_2^* came out to be 0.4375 and 0.625 respectively. In this case an A-cycle is defined to start whenever the second queue becomes empty. For the importance sampling estimate, we use the same simulation run length, batch size and initial transient runs as for the single queue case. As before, we also run standard simulations for the same CPU times as the importance sampling cases.

Results from experiments have been reported in Table 3. The buffer size of the first queue, x_1, is fixed at 25 and we vary the buffer size, x_2, in the second queue. Again note the orders of magnitude increase in efficiency of importance sampling over standard simulation. Also note how the RE is nearly constant over a large range of values for the estimate.

For another illustration of the importance sampling heuristic we ran a second intree network case, so that for this case $\tilde{\theta}_2 > \theta_2^*$. Again, the output from the first queue feeds into the second queue. Now we have 8 identical MMP sources feeding into the first queue ($r_{11} = 0.8$ and $r_{22} = 0.5$) and 2 identical MMP sources ($r_{11} = 0.6$ and $r_{22} = 0.7$) feeding into the second queue. The c for both the queues was taken to be 4. The $\tilde{\theta}_2$ and $\theta^*{}_2$ came out to be 0.4688 and 0.2734 respectively. Results for experiments (where, again, $x_1 = 25$) are presented in Table 4. The results have the same bounded RE characteristic as before.

x_2	Importance Sampling Estimate	Standard Simulation Estimate
10	$1.98 \times 10^{-4} \pm 6.1\%$	$1.87 \times 10^{-4} \pm 29.4\%$
25	$5.80 \times 10^{-9} \pm 7.3\%$	0.00±?
50	$3.71 \times 10^{-16} \pm 10.1\%$	0.00±?

Table 3: Estimates of steady-state loss probability in the intree network ($\tilde{\theta}_2 < \theta_2^*$)

.

x_2	Importance Sampling Estimate	Standard Simulation Estimate
10	$1.00 \times 10^{-2} \pm 3.6\%$	$1.01 \times 10^{-2} \pm 4.3\%$
25	$1.50 \times 10^{-4} \pm 3.9\%$	$1.81 \times 10^{-4} \pm 25.12\%$
50	$1.60 \times 10^{-7} \pm 3.9\%$	0.00±?

Table 4: Estimates of steady-state loss probability in the intree network ($\tilde{\theta}_2 > \theta_2^*$)

.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we have demonstrated how one may use the theory of effective bandwidths in [9] to run fast simulations for estimating the loss probabilities of ATM intree networks. Two types of arrival processes are considered: Markov arrival processes in Example 2.6 and autoregressive processes in Example 2.7. The algorithms for finding appropriate changes of measures are outlined in Algorithm 2.2 for a single queue and Algorithm 3.2 for an intree network, respectively. For the single queue case, we have shown that our algorithm is indeed asymptotically optimal. We have also shown how the change of measure can be used to estimate the steady state probabilities without relying regenerative structure. For an intree network, we have also justified the use of Algorithm 3.1 by various experiments that show significant variance reduction. An interesting topic for future research is if one could show the optimality of Algorithm 3.1 from a theoretical foundation. To be specific, one needs to establish the lower bound in Theorem 3.1.

Other topics for future research includes generalizations of our methodology to general ATM networks, e.g. feedforward networks with Markovian routing. Also, one might extend our methodology for discrete-time models to continuous-time models. For instance, one might be able to provide a theoretical foundation for the heuristic change of measure recently proposed by [24] for two-state Markov fluids in a single queue.

References

[1] D. Anick, D. Mitra and M.M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.,* Vol. 61, pp. 1871-1894, 1982.

[2] S. Asmussen, "Conjugate processes and the simulation of ruin problems," *Stochastic Processes and their Applications* Vol. 20, pp. 213-229, 1985.

[3] S. Asmussen, "Risk theory in a Markovian environment," *Scand. Actuarial J.*, pp. 69-100, 1989.

[4] F. Baccelli and P. Bremaud. *Elements of Queueing Theory.* New York: Springer Verlag, 1990.

[5] P. Bratley, B.L. Fox and L.E. Schrage. *A Guide to Simulation* Second Edition. New York: Springer Verlag, 1987.

[6] J. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc., 1990.

[7] J.A. Bucklew, P. Ney, and J.S. Sadowsky, "Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains," *J. Appl. Prob.* Vol. 27, pp. 44-99, 1990.

[8] C.S. Chang, "Stability, queue length and delay, Part I: Deterministic queueing networks," IBM RC 17708, 1992. Part of the report is published in *Proceedings of the IEEE CDC'92 Conference*, pp. 999-1004, Tucson, Arizona, 1992.

[9] C.S. Chang, "Stability, queue length and delay, Part II: Stochastic queueing networks," IBM RC 17709, 1992. Part of the report is published in *Proceedings of the IEEE CDC'92 Conference*, pp. 1005-1010, Tucson, Arizona, 1992.

[10] R. Cogburn, "A uniform theory for sums of Markov chain transition probabilities, " *The Annals of Probability*, Vol. 3, pp.191-214, 1975.

[11] M.A. Crane and D.L. Iglehart, " Simulating stable stochastic systems III: regenerative processes and discrete event simulation," *Operations Research*, Vol. 23, pp. 33-45, 1975.

[12] M. Devetsikiotis and K.R. Townsend, "A dynamic importance sampling methodology for the efficient estimation of rare event probabilities in regenerative simulations of queueing systems," *Proceedings of the IEEE ICC '92 Conference*, pp. 1290-1297, 1992.

[13] M. Devetsikiotis and K.R. Townsend, "On the efficient simulation of large communication networks using importance sampling," Preprint, 1992.

[14] M.R. Frater, M., T.M. Lenon, and B.D.O. Anderson, "Optimally efficient estimation of the statistics of rare events in queueing networks," *IEEE Transactions on Automatic Control* Vol. 36, pp. 1395-1405, 1991.

[15] P.W. Glynn and D.L. Iglehart. 1989. "Importance sampling for stochastic simulations," *Management Science* Vol. 35, pp. 1367-1392, 1989.

[16] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[17] A. Goyal, P. Shahabuddin, P. Heidelberger, V.F. Nicola and P.W. Glynn. "A unified framework for simulating Markovian models of highly reliable systems," *IEEE Transactions on Computers* Vol. C-41, pp. 36-51, 1992.

[18] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.,* Vol. 9, pp. 968-981, 1991.

[19] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods.* London: Methuen and Co., Ltd., 1964.

[20] I. Iscoe, P. Ney and E. Nummelin, "Large deviations of uniformly recurrent Markov Additive Processes," *Adv. Appl. Math.*, Vol. 6, pp. 373-412, 1985.

[21] S. Juneja, "Optimality in fast simulation and uniformly bounded likelihood ratios." preprint, 1992.

[22] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[23] G. Kesidis, J. Walrand, and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *Preprint*, 1992.

[24] G. Kesidis and J. Walrand, "Quick simulation of ATM buffers," *Proceedings of the IEEE CDC'92 Conference*, pp. 1018-1019, Tucson, Arizona, 1992.

[25] T. Lehtonen and H. Nyrhinen, "On asymptotically efficient simulation of ruin probabilities in a Markovian environment," preprint, 1991.

[26] R.M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Camb. Phil. Soc.*, Vol. 58, pp. 497-520, 1962.

[27] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications", *IEEE Trans. Commun.*, Vol. 36, pp. 834-844, 1988.

[28] H.D. Miller, "A convexity property in the theory of random variables defined on a finite Markov chain," *Ann. Math. Stat.*, Vol. 32, pp. 1260-1270, 1961.

[29] P. Ney and E. Nummelin, "Markov additive processes I. Eigenvalue properties and limit theorems," *Ann. Prob.*, Vol. 15, pp. 561-592, 1987.

[30] P. Ney and E. Nummelin, "Markov additive processes II. Large deviations," *Ann. Prob.*, Vol. 15, pp. 593-609, 1987.

[31] V.F. Nicola, P. Shahabuddin, P. Heidelberger and P.W. Glynn, "Fast Simulation of Steady-State Availability in Non-Markovian Highly Dependable Systems, " IBM Research Report, 1992.

[32] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Transactions on Automatic Control* Vol. 34, pp. 54-56, 1989.

[33] J.S. Sadowsky, "Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue," *IEEE Transactions on Automatic Control* Vol. 36, pp. 1383-1394, 1991.

[34] D. Siegmund, "Importance sampling in the Monte Carlo study of sequential tests," *The Annals of Statistics* Vol. 4, pp. 673-684, 1976.

[35] K. Sohraby, "On the asymptotic behavior of heterogeneous statistical multiplexer with applications," *INFOCOM'92*, Florence, Italy.