Analysis of an Importance Sampling Estimator for Tandem Queues

PAUL GLASSERMAN and SHING-GANG KOU
Columbia University

We analyze the performance of an importance sampling estimator for a rare-event probability in tandem Jackson networks. The rare event we consider corresponds to the network population reaching K before returning to 0, starting from 0, with K large. The estimator we study is based on interchanging the arrival rate and the smallest service rate and is therefore a generalization of the asymptotically optimal estimator for an M/M/1 queue. We examine its asymptotic performance for large K, showing that in certain parameter regions the estimator has an asymptotic efficiency property, but that in other regions it does not. The setting we consider is perhaps the simplest case of a rare-event simulation problem in which boundaries on the state space play a significant role.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics—*probabilistic algorithms*; I.6.1 [**Simulation and Modeling**]: Simulation Theory

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Importance sampling, Markov chains, queueing networks, rare events

1. INTRODUCTION

We analyze the performance of an importance sampling estimator for a rare-event probability in certain queueing networks. The probability in question is

$$p_K \triangleq P(\text{network population reaches } K \text{ before returning to } 0, \text{ starting from } 0),$$

a type of *overflow* probability if we think of K as an upper limit on the network population. The networks we consider are tandem Jackson networks —serial networks of single-server nodes with Poisson arrivals and exponen-

This work was supported by the National Science Foundation under grants MSS-9216490 and DMI-9457189.

Authors' addresses: P. Glasserman, Graduate School of Business, Columbia University, New York, NY 10027; Shing-Gang Kou, Department of Statistics, Columbia University, New York, NY 10027.

tial service times. The estimator we study, originally proposed by Parekh and Walrand [1989], is based on interchanging the arrival rate and the smallest service rate. We analyze its asymptotic performance as K becomes large, and show that in certain parameter regions the estimator has an asymptotic efficiency property, but that in other regions it does not.

To put this problem in context, we give some background. In his analysis of importance sampling for level-crossing probabilities associated with random walks, Siegmund [1976] identified the unique asymptotically optimal change of measure within a parametric class. It follows from his analysis and results in Asmussen [1982] that the asymptotically optimal change of measure for simulating large build-ups in GI/G/1 queues exponentially twists (in the sense of, e.g.,) the interarrival- and service-time distributions by a parameter $\theta_0 > 0$ satisfying $\phi_A(-\theta_0)\phi_B(\theta_0) = 1$, where ϕ_A and ϕ_B are the moment-generating functions of the interarrival- and service-times. In the case of an M/M/1 queue with arrival rate λ and service rate μ, this equation becomes

$$\left(\frac{\lambda}{\lambda + \theta_0}\right)\left(\frac{\mu}{\mu - \theta_0}\right) = 1,$$

which has $\theta_0 = \mu - \lambda$ as its only positive solution, if $\lambda < \mu$. Twisting the interarrival- and service-time distributions by this parameter has the effect of interchanging λ and μ. Thus, the asymptotically optimal change of measure for estimating p_K in an M/M/1 queue corresponds to simulating an unstable M/M/1 queue with arrival rate μ and service rate λ.

More recently, this idea has been extended to queues with complex arrival processes (see, in particular, Chang et al. [1994] and Kesidis and Walrand [1993]), to heavy-traffic simulation [Asmussen 1987, XII.7], and to multi-server queues [Sadowsky 1991]. Rather less has been accomplished in extending Siegmund's original result to networks of queues. Networks—even those that can be modeled as Markov chains—introduce difficulties not present in a single queue, primarily because of boundaries on their state spaces. The queue-length process of a single queue has a boundary at 0, but since p_K, for example, depends only on events before the first return to the origin, this boundary plays no essential role. In contrast, the boundaries in queueing networks significantly affect the form of the likelihood ratio associated with a change of measure, and make it much more difficult to identify effective importance sampling distributions. The boundaries also complicate the large-deviations analysis of rare events (see, in particular, Dupuis, Ishii, and Soner [1990]).

Based on a heuristic application of large-deviations techniques, Parekh and Walrand [1989] proposed importance sampling estimators for overflow probabilities in various Jackson networks. For tandem networks, their estimator interchanges the arrival rate and the slowest service rate, thus generalizing the M/M/1 estimator described above. They evaluated this estimator numerically and found that it generally works well. An optimization step required in Parekh and Walrand [1989] was solved in Frater et al. [1991]; the underlying large-deviations problem was further considered in Tsoucas [1992].

Our analysis of the Parekh-Walrand estimator for tandem queues includes both positive and negative conclusions. We verify that estimating p_K is difficult in the sense that, without importance sampling, the number of replications required to achieve a fixed relative error grows exponentially in K. We show that in certain parameter regions the Parekh-Walrand estimator is *asymptotically efficient*, in the sense that the number of runs required grows subexponentially; indeed, we show that it has *linearly bounded relative error*, meaning that the number of runs required grows at most linearly in K. Under an additional condition on the model parameters, the estimator has *bounded* relative error, meaning that the number of runs required is bounded in K.

However, we also show that in other parameter regions the estimator fails to be asymptotically efficient and is therefore no better, in an asymptotic sense, than straightforward simulation. In particular, in two-node systems we show that asymptotic efficiency fails when the two service rates are nearly equal and the arrival rate is small. This is consistent with a numerical observation in Parekh and Walrand [1989] and the discussion in Anantharam et al. [1990]. The intuition for this is as follows. In a system with significantly different service rates, given that a large network buildup has occurred, it has most likely occurred because of a large buildup at the bottleneck node; interchanging the arrival rate with the bottleneck service rate mimics this conditional behavior. But if the service rates are close, there are many ways for a large network population to accumulate, so importance sampling based on the interchange rule tends to be less effective. (For an example of a precise connection between a change of measure and a conditional law given a rare event, see Asmussen [1982].)

Both the positive and negative results reported here contribute to the growing area of rare-event simulation. The negative results may serve as a cautionary note on the presumption that importance-sampling distributions suggested by large-deviations calculations are automatically effective (though even this statement must be qualified because a complete large-deviations analysis of the models we consider is not available). The positive results represent, as far as we know, the only proof of asymptotic efficiency for an exponentially rare event in a setting where boundaries play a significant role. Some of the techniques we use in establishing both types of results may prove useful in other settings.

The rest of this paper is organized as follows. In section 2, we present bounds and asymptotics for p_K as a function of K; these are needed for the relative-error analysis in later sections. With no additional complication we present these results for arbitrary Jackson networks, not just tandem queues. Section 3 analyzes the performance of straightforward simulation and then introduces the importance sampling estimator in detail. Sections 4 and 5 present, respectively, necessary conditions and sufficient conditions for asymptotic efficiency.

Preliminary versions of some of the results reported here are contained in Glasserman and Kou [1993]. However, that paper considered only two-node

networks and did not establish linearly bounded relative error or bounded relative error—only asymptotic efficiency.

2. OVERFLOW PROBABILITIES IN JACKSON NETWORKS

In this section, we derive upper and lower bounds on the overflow probability p_K for arbitrary (stable) Jackson networks. The lower bounds, in particular, are necessary for the asymptotic analysis of the importance-sampling estimator given in subsequent sections. The development in this section proceeds in two steps: we first bound the *stationary* probability of the overflow *set*—the set of states with job population K; we then use a regenerative argument to convert the stationary bounds to bounds on the transient probability p_K.

2.1 Stationary Overflow

Consider a d-node Jackson network with arrival rate λ and service rates $\mu = (\mu_1, \ldots, \mu_d)$. Arrivals join node i with probability q_i; departures from node i join node j with probability P_{ij} and leave the network with probability $1 - \sum_j P_{ij}$. The matrix P is irreducible and has spectral radius less than 1. Let $q = (q_1, \ldots, q_d)$ and suppose throughout that

$$\left[\lambda q(I - P)^{-1}\right]_i < \mu_i, \qquad i = 1, \ldots, d, \tag{1}$$

so that the network is stable. The utilization parameters $\rho = (\rho_1, \ldots, \rho_d)$ are given by

$$\rho_i = \left[\lambda q(I - P)^{-1}\right]_i / \mu_i, \qquad i = 1, \ldots, d.$$

When (1) holds, the vector queue-length process has stationary distribution π given by

$$\pi(x) = \prod_{i=1}^{d} (1 - \rho_i)\rho_i^{x_i}, \qquad x \in Z_+^d; \tag{2}$$

see, e.g., Kelly [1979].

Define the *overflow* set

$$C_K = \{x \in Z_+^d : x_1 + \cdots + x_d = K\},$$

the set of states in which the network population is exactly K. We bound p_K by first bounding $\pi(C_K)$. Let

$$\rho_* = \max_i \rho_i,$$

the utilization of the most highly utilized node in the network.

LEMMA 2.1. $\pi(C_K) \geq \rho_*^K \prod_{i=1}^{d}(1 - \rho_i)$.

This follows from the fact that the state in which there are K jobs at a maximal-utilization node and no jobs anywhere else is an element of C_K.

LEMMA 2.2. *There is a constant* $c \le 1$ *such that* $\pi(C_k) \le c\rho_*^K(K+1)^{d-1}$. *Moreover, if* $d = 2$, *then*

$$\pi(C_K) \le c_1 \rho_*^K, \qquad \text{if } \rho_1 \ne \rho_2;$$
$$\pi(C_K) \le c_2 \rho_*^K (K+1), \qquad \text{if } \rho_1 = \rho_2,$$

where c_1, c_2 *are constants.*

PROOF. For the general case, we have

$$\begin{aligned}
\pi(C_k) &= \sum_{x_1 + \cdots + x_d = K} \prod_{i=1}^{d} (1-\rho_i)\rho_i^{x_i} \\
&= \prod_{i=1}^{d} (1-\rho_i)\rho_*^K \sum_{x_1 + \cdots + x_d = K} \prod_{i=1}^{d} (\rho_i/\rho_*)^{x_i} \\
&\le \prod_{i=1}^{d} (1-\rho_i)\rho_*^K \sum_{x_1 + \cdots + x_d = K} 1 \\
&\le \prod_{i=1}^{d} (1-\rho_i)\rho_*^K (K+1)^{d-1} \\
&= c\rho_*^K (K+1)^{d-1}.
\end{aligned} \tag{3}$$

For the special case $d = 2$, the result for $\rho_1 = \rho_2$ follows directly from (3). If, instead, we have $\rho_1 < \rho_2$, then (3) becomes

$$\pi(C_K) = (1-\rho_1)(1-\rho_2)\rho_2^K \sum_{x_1=0}^{K} (\rho_1/\rho_2)^{x_1} \le c_1 \rho_2^K.$$

The case $\rho_2 < \rho_1$ works the same way. □

2.2 Transient Overflow

Under the stability condition (1), the origin is positive recurrent for the queue-length process $X_t = (X_t^1, \ldots, X_t^d)$, $t \ge 0$. Let T_0 denote the time of the first return to zero, let E_0 denote expectation for X started at the origin, and let $c_0 = E_0[T_0]$ be the expected length of a 0-cycle. A standard result on regenerative processes (e.g., Asmussen 1987, p. 126) then asserts that

$$\pi(C_K) = c_0^{-1} E_0\left[\int_0^{T_0} 1_{\{X_t \in C_K\}}\, dt\right]. \tag{4}$$

We use this to prove the following result, applying an argument of Anantharam [1989].

THEOREM 2.3. *For all* $K \ge 1$,

$$b_1 \rho_*^K K^{-1} \le p_K \le b_2 \rho_*^K (K+1)^d,$$

where b_1, b_2 *are constants.*

PROOF. Let

$$I_K = \int_0^{T_0} 1_{\{X_t \in C_K\}}\, dt$$

and observe that

$$E_0[I_K] = p_K E_0[I_K \mid I_K > 0],$$

so (4) implies

$$p_K = \frac{c_0 \pi(C_K)}{E_0[I_K \mid I_K > 0]}. \tag{5}$$

We can therefore bound p_K using our bounds on $\pi(C_K)$, provided we can bound $E_0[I_K \mid I_K > 0]$.

Once the network population reaches K, at least one service time must elapse before the last exit from C_K in the cycle. Thus,

$$0 < \min_\iota \mu_\iota^{-1} \leq E_0[I_K \mid I_K > 0]. \tag{6}$$

To get an upper bound, let

$$t^* = \max\left\{E_x[T_0] : \sum_\iota x_\iota = 1\right\},$$

where E_x denotes expectation starting from state x. Thus, t^* bounds the expected time to empty the network from any state in which there is exactly one job in the network. The stability condition (1) implies $t^* < \infty$. Moreover, given that C_K is reached, I_K is bounded above by the time to reach zero. We claim that the expected time to reach zero from C_K is bounded by Kt^*. This will follow if we can show that the expected time to reach C_{K-1} from C_K is bounded by t^*. For this, consider the following sample-path argument. The time to reach C_{K-1} from C_K is bounded above by the time to reach C_{K-1} with all but one of the original K jobs frozen, and all newly arriving jobs given preemptive priority over the frozen jobs. But this is the same as the time to empty the system from an initial population of 1, and thus has expectation bounded by t^*. We conclude that

$$E_0[I_K \mid I_K > 0] \leq Kt^*. \tag{7}$$

Combining (5)–(7) with our bounds on $\pi(C_K)$ concludes the proof. □

Remark. An immediate consequence of the bounds in Theorem 2.3 is the logarithmic limit

$$\lim_{K \to \infty} \frac{1}{K} \log p_K = -\log \rho_*, \tag{8}$$

which was proved in Glasserman and Kou [1993].

It seems plausible that $E_0[I_K \mid I_K > 0]$ is in fact bounded by a constant independent of K; Anantharam and Ganesh [to appear] have proved such a result for a different type of overflow set. An improvement on the bound in (7) from $O(K)$ to $O(1)$ would allow us to remove the factor K^{-1} from the lower bound in Theorem 2.3, and this has implications for the analysis in Section 5. We point out one case in which such an improvement is readily available:

PROPOSITION 2.4. *In a d-node tandem network satisfying*

$$\frac{1}{\lambda} > \sum_{i=1}^{d} \frac{1}{\mu_i}, \tag{9}$$

we have $p_K \geq b\rho_*^K$, *with b a constant.*

PROOF. Construct an associated queue with arrival rate λ and service times equal to the sum of d exponential random variables with parameters $\mu_1, \ldots, \mu_d$; under condition (9); this queue is stable. It can be coupled to the tandem network so that its queue length Y_t is never less than the total population $X_t^1 + \cdots + X_t^d$ of the tandem system. Hence, $E_0[I_K \mid I_K > 0]$ is bounded above by the time Y_t spends at or above level K, given that it reaches level K before returning to zero. This, in turn, is bounded by the expectation of the last time $Y_t \geq K$ before reaching 0, starting from K, which is finite and bounded in K [Janson 1986, Theorem 1(ii)]. It follows that $E_0[I_K | I_K > 0]$ is bounded independent of K. □

3. DIRECT SIMULATION AND IMPORTANCE SAMPLING

3.1 Relative Error

Define the *relative error* of any unbiased estimator of p_K to be the ratio of the standard deviation of the estimator and p_K. The direct-simulation estimator of p_K generates sample paths of the vector queue-length process X (starting from the origin) and returns the indicator $1_{\{T_K < T_0\}}$, where T_K is the time of the first visit to C_K. This indicator is clearly unbiased, and the sample mean of n replications of the indicator has variance $(p_K - p_K^2)/n$. For large K, its relative error is therefore

$$RE = \frac{\sqrt{p_K - p_K^2}}{p_K\sqrt{n}} \approx \frac{1}{\sqrt{np_K}},$$

since $p_K^2 \ll p_K$. But from Theorem 2.3 we find that

$$\frac{1}{\sqrt{np_K}} \geq \frac{1}{\sqrt{nb_2\,\rho_*^K(K+1)^d}}.$$

It follows that for large K, the number of replications n_δ required to achieve a relative error δ will obey

$$n_\delta \geq \frac{1}{b_2\,\rho_*^K\delta^2(K+1)^d};$$

i.e., the number of replications required to achieve a fixed relative error grows exponentially in K.

This analysis shows that a precise estimation of p_K by direct simulation becomes infeasible for large K. The source of the difficulty is the fact that the second moment of the indicator $\mathbf{1}_{\{T_K < T_0\}}$ is just p_K itself; in particular, then, the first and second moments of the direct estimator vanish at the same rate, resulting in a relative error that increases exponentially in K.

Let us contrast this with the best possible performance for any estimator. Denote by m_K the second moment of an unbiased estimator of p_K; then $m_K \geq p_K^2$, so

$$\liminf_{K\to\infty} \frac{\log m_K}{\log p_K} \geq 2.$$

The estimator is *asymptotically efficient* if

$$\limsup_{K\to\infty} \frac{\log m_K}{\log p_K} \leq 2. \tag{10}$$

The second moment of an asymptotically efficient estimator vanishes at twice the exponential rate of p_K itself. Consequently, its relative error increases at a subexponential rate. More precisely, (10) and the convergence of $K^{-1} \log p_K$ together imply that $K^{-1}(\frac{1}{2}\log m_K - \log p_K) \to 0$. Since $RE \leq \sqrt{m_K}/p_K$, this implies that $\limsup K^{-1} \log RE \leq 0$; i.e.,

$$\log(RE) - \epsilon K \to -\infty, \qquad \forall \epsilon > 0,$$

which is to say that $RE = o(e^{\epsilon K})$ for all $\epsilon > 0$. Because RE grows at a subexponential rate, the number of replications required to achieve a specified relative error grows at a subexponential rate as well. This property is sometimes called *asymptotic optimality*.

The best possible performance for an asymptotically efficient estimator is that it have *bounded* relative error, for then the number of replications required to achieve a specified RE is bounded in K. A somewhat weaker requirement, which nevertheless represents a significant improvement over unqualified asymptotic efficiency, is that the relative error be *linearly* bounded in K, or *polynomially* bounded by a polynomial of known degree. The analysis in the rest of the paper is devoted to identifying conditions under which a particular estimator of p_K is asymptotically efficient and has bounded or linearly bounded relative error.

3.2 The Importance Sampling Estimator

For the rest of the paper we restrict attention to tandem queues with arrival rate λ and consecutive service rates $\mu_1, \ldots, \mu_d$. We always assume

$$\mu_d = \min_i \mu_i; \tag{11}$$

changing the order of the service rates does not change p_K [Weber 1979] so this assumption entails no essential loss of generality. The importance sampling estimator we study interchanges λ and μ_d to make the network unstable and thus the overflow event less rare.

For simplicity, we work with the discrete-time chain embedded at the transition epochs of the continuous-time queue-length process. Thus, let X_n^i denote the number of jobs at node i just after the nth transition, $i = 1, \ldots, d$. Take as initial state $X_0 = (1, 0, \ldots, 0)$, the only state reachable from the origin in one transition. Let T_K be the smallest n for which $\sum_i X_n^i = K$ and

T_0 the smallest n for which $\sum_\iota X_n^\iota = 0$. Thus, $p_K = P_{1,0,\ \ ,0}(T_K < T_0)$, where the subscript on P indicates the starting state. Throughout, we adopt the convention that

$$\lambda + \mu_1 + \cdots + \mu_d = 1. \tag{12}$$

Let $A_1, A_2, \ldots, A_{2^{d-1}-1}$ be the distinct, nonempty subsets of $\{1, \ldots, d-1\}$. For each $i = 1, \ldots, 2^{d-1} - 1$, let

$$N_n^\iota = |\{0 \le k < n : X_k^j > 0 \Leftrightarrow j \in A_\iota\}|,$$

the number of visits before time n to the boundary on which the nodes in A_ι are nonempty but node d and all other nodes not in A_ι are empty. Define constants

$$a_\iota = \left(\mu_d + \sum_{j \in A_\iota} \mu_j\right) \Bigg/ \left(\lambda + \sum_{j \in A_\iota} \mu_j\right), \qquad i = 1, \ldots, 2^{d-1} - 1.$$

Let $\overline{P}, \overline{E}$ denote probability and expectation for the new system in which λ and μ_d are interchanged. The likelihood ratio for $\{(X_k^1, \ldots, X_k^d),\ 0 \le k \le n\}$ relating to the original $(\lambda, \mu_1, \ldots, \mu_d)$ system to the new system with parameters $(\mu_d, \mu_1, \ldots, \lambda)$ is given by

$$L_n = \left(\frac{\lambda}{\mu_d}\right)^{X_n^1 + \quad + X_n^d - 1} \cdot \left(\prod_{\iota=1}^{2^{d-1}-1} a_\iota^{N_n^\iota}\right). \tag{13}$$

More specifically, we have the following result. (In the statement of the proposition and throughout the paper, for any event G the notation $\overline{E}[\cdot; G]$ denotes $\overline{E}[\cdot \mathbf{1}_G]$, where $\mathbf{1}_G$ is the indicator of G.)

PROPOSITION 3.1. *With the notation above,*

$$p_K \equiv P_{1,0,\ \ ,0}(T_K < T_0) = \overline{E}_{1,0,\ \ ,0}[L_{T_K}; T_K < T_0],$$

the subscripts indicating the starting state.

This result is a version of Wald's likelihood ratio identity in the particular form put forth in Glynn and Iglehart [1989] for Markov chains, so we omit the proof.

Remark. In Section 5, it will be important to keep in mind that none of the N_n^ι counts visits to the origin, since the A_ι are nonempty. Every visit to the origin is directly followed by a visit to $(1, 0, \ldots, 0)$ under both the old and new measures. Consequently, the likelihood ratio associated with this transition is just 1, and does not contribute to L_n.

From Proposition 3.1 and (13) we find that

$$p_K = (\lambda/\mu_d)^{K-1} \overline{E}_{1,0,\ \ ,0}\left[\prod_\iota a_\iota^{N_{T_K}^\iota}; T_K < T_0\right].$$

Thus, the estimator of p_K obtained from simulation under the new measure consists of independent replications of

$$\hat{p}_K \triangleq (\lambda/\mu_d)^{K-1} \prod_\iota a_\iota^{N_{T_K}^\iota} \mathbf{1}_{\{T_K < T_0\}}, \tag{14}$$

from initial state $(1, 0, \ldots, 0)$. Comparing this expression with (8) and noting that $\rho_* = \lambda/\mu_d$, we see that $\hat{p}_K$ in effect estimates the correction to the asymptotic result (8) for finite K. The analysis of this estimator is complicated by the fact that the a_i are strictly greater than 1. The presence of these factors embodies the difficulty introduced in the problem by the boundaries corresponding to states in which the last node is empty. A consequence of Proposition 3.1 and (8) is

$$K^{-1} \log \bar{E}_{1,0,\ldots,0}\left[\prod_i a_i^{N_{T_K}^i}; T_K < T_0\right] \to 0,$$

though *a priori* it is not even obvious that the $\bar{E}_{1,0,\ldots,0}[a_i^{N_{T_K}^i}; T_K < T_0]$ are finite.

4. NECESSARY CONDITIONS FOR ASYMPTOTIC EFFICIENCY

The second moment of the estimator $\hat{p}_K$ in (14) is

$$\bar{E}_{1,0,\ldots,0}\left[L_{T_K}^2; T_K < T_0\right] = (\lambda/\mu_d)^{2K-2} \bar{E}_{1,0,\ldots,0}\left[\prod_i a_i^{2N_{T_K}^i}; T_K < T_0\right]. \quad (15)$$

Thus, in light of (8) and (10), our estimator is asymptotically efficient only if

$$\limsup_{K\to\infty} K^{-1} \log \bar{E}_{1,0,\ldots,0}\left[\prod_i a^{2N_{T_K}^i}; T_K < T_0\right] = 0. \quad (16)$$

We will see that this is not always the case.

4.1 Two-Node Networks

We begin by examining the case $d = 2$ because it is the simplest. In two-node networks there is only one boundary to consider—the set of states in which the second node is empty. Thus, we may simply write

$$a \equiv a_1 = \frac{\mu_2 + \mu_1}{\lambda + \mu_1},$$

and let N_n be the number of visits to the *horizontal* boundary $A_1 = \{(x_1, x_2): x_1 > 0, x_2 = 0\}$ before time n. We now have

PROPOSITION 4.1. *A necessary condition for asymptotic efficiency is* $\mu_2 \le 1/(\mu_1 + 2)$; *in particular,* μ_2 *cannot exceed* $\sqrt{2} - 1$.

PROOF. We obtain a lower bound on $\bar{E}_{1,0}[a^{2N_{T_K}}; T_K < T_0]$ by considering a single sample path in the event $\{T_K < T_0\}$. Consider the path that moves $K - 1$ steps to the right from $(1, 0)$ to hit $(K, 0)$. This path has probability $(\mu_2/(\mu_1 + \mu_2))^{K-1}$; on this path, $N_{T_K} = K - 1$. Thus,

$$\bar{E}_{1,0}\left[a^{2N_{T_K}}; T_K < T_0\right] \ge \left(\frac{\mu_2}{\mu_2 + \mu_1}\right)^{K-1} \left(\frac{\mu_2 + \mu_1}{\lambda + \mu_1}\right)^{2K-2}$$

and our necessary condition (16) for asymptotic efficiency entails

$$\left(\frac{\mu_2}{\mu_2 + \mu_1}\right)\left(\frac{\mu_2 + \mu_1}{\lambda + \mu_1}\right)^2 \leq 1. \tag{17}$$

This simplifies to $\mu_2 \leq 1/(\mu_1 + 2)$. The largest value of μ_2 satisfying this inequality and also the condition $\mu_2 \leq \mu_1$ required by (11) is $\sqrt{2} - 1$. □

Our necessary condition is consistent with observations in Parekh and Walrand [1989], based on numerical experiments, that the estimation problem is most difficult when μ_2 is close to μ_1. Anantharam et al. [1990] propose an entirely different estimator for the case $\mu_1 = \mu_2$.

4.2 Multinode Systems

We now derive necessary conditions for asymptotic efficiency in multinode systems. The argument we use is the same as that used to prove Proposition 4.1: we identify sample paths with sufficiently high probability and sufficiently large likelihood ratios that they are consistent with asymptotic efficiency only in certain parameter ranges. The presence of additional nodes provides considerably more flexibility in the choice of paths.

To state the next result, we introduce the notation

$$A_{ij} = A_i \cup \{j\}, \qquad A_{ijk} = A_i \cup \{j,k\} \qquad i = 1,\ldots,2^{d-1}-1; \quad j,k = 1,\ldots,d.$$

With this we have the following theorem.

THEOREM 4.2. *The following conditions are necessary for asymptotic efficiency*:

$$\begin{aligned}\mu_d\left(\mu_d + \sum_{r\in A_{ij}} \mu_r\right)\prod_{k=1}^{j-1} \mu_k\left(\mu_d + \sum_{r\in A_{ijk}} \mu_r\right) \\ \leq \left(\lambda + \sum_{r\in A_{ij}} \mu_r\right)^{2^{j-1}} \prod_{k=1}^{2^{j-1}} \left(\lambda + \sum_{r\in A_{ijk}} \mu_r\right)^2,\end{aligned} \tag{18}$$

for all $i = 1,\ldots,2^{d-1}-1$ *and* $j = 1,\ldots,d-1$.

PROOF. Each of the $(d-1)\times(2^{d-1}-1)$ cases in (18) is established using the argument in Proposition 4.1 for a particular sequence of sample paths. For each i and j, the inequality in (18) corresponds to the following path: the first job advances to the highest-index node in A_i; a second job arrives and advances to the node in A_i with the next highest index; this continues until each node in A_i has one job. Subsequently, jobs arrive and advance through the network, accumulating at node j, without any of the other nodes in A_i emptying. This continues until there are $K - |A_{ij}|$ jobs at node j, and 1 job at each of the other nodes in A_{ij}. The path is completed with the arrival of a Kth job at node 1. A straightforward but tedious calculation shows that the $\bar{P}$-probability of this path is

$$M_1 \cdot \left[\left(\frac{\mu_d}{\mu_d + \sum_{r\in A_{ij}} \mu_r}\right)\prod_{k=1}^{j-1}\left(\frac{\mu_k}{\mu_d + \sum_{r\in A_{ijk}} \mu_r}\right)\right]^{K-|A_{ij}|},$$

and that the likelihood ratio for this path is

$$M_2 \cdot \left[\left(\frac{\mu_d + \sum_{r \in A_{i,j}} \mu_r}{\lambda + \sum_{r \in A_{i,j}} \mu_r} \right) \prod_{k=1}^{j-1} \left(\frac{\mu_d + \sum_{r \in A_{i,jr}} \mu_r}{\lambda + \sum_{r \in A_{i,j}} \mu_r} \right) \right]^{K - |A_{i,j}|},$$

where M_1 and M_2 represent positive terms independent of K. Consequently,

$$\begin{aligned} \bar{E}_{1,0,\ldots,0}\left[L_{T_K}^2 ; T_K < T_0 \right] \\ \geq M_1 M_2 \cdot \left[\left(\frac{\mu_d}{\mu_d + \sum_{r \in A_{i,j}} \mu_r} \right) \prod_{k=1}^{j-1} \left(\frac{\mu_k}{\mu_d + \sum_{r \in A_{i,jk}} \mu_r} \right) \right]^{K - |A_{i,j}|} \\ \times \left[\left(\frac{\mu_d + \sum_{r \in A_{i,j}} \mu_r}{\lambda + \sum_{r \in A_{i,j}} \mu_r} \right)^2 \prod_{k=1}^{j-1} \left(\frac{\mu_d + \sum_{r \in A_{i,jr}} \mu_r}{\lambda + \sum_{r \in A_{i,jr}} \mu_r} \right)^2 \right]^{K - |A_{i,j}|}. \end{aligned}$$

A necessary condition is, then, that the product of the terms in square brackets be no greater than 1, which yields (18) after some algebraic simplification. □

With $A_i = \{1\}$ and $j = 1$, (18) becomes $\mu_1 \mu_d + \mu_d^2 \leq (\mu_1 + \lambda)^2$, which matches (17) when $d = 2$. Examples of parameters violating this necessary condition are $(\lambda, \mu_1, \mu_2, \mu_3) = (0.10, 0.30, 0.32, 0.28)$, for $d = 3$, and $(\lambda, \mu_1, \mu_2, \mu_3, \mu_4) = (0.09, 0.23, 0.227, 0.227, 0.226)$, for $d = 4$.

5. SUFFICIENT CONDITIONS FOR ASYMPTOTIC EFFICIENCY

Lower bounds on the second moment of the estimator give necessary conditions for asymptotic efficiency; to get sufficient conditions, we need upper bounds. We begin by making two modifications that result in upper bounds: we omit the indicator $\mathbf{1}_{\{T_K < T_0\}}$ in (15), and we replace the $N_{T_K}^i$ with

$$N = \lim_{n \to \infty} \sum_i N_n^i,$$

the *total* number of transitions on boundaries in which the last node is empty; N is almost surely finite under the new measure because the last node is unstable in the $(\mu_d, \mu_1, \ldots, \mu_{d-1}, \lambda)$ system, and thus empty for only finitely many transitions. If we can show that

$$\bar{E}_{1,0,\ldots,0}[a^{2N}] < \infty, \tag{19}$$

with

$$a = \max_i a_i \tag{20}$$

it will follow that the expectation on the right side of (15) is bounded uniformly in K, and this will lead to asymptotic efficiency.

5.1 A Gauge-Function Lemma

To establish (19), we prove a general result on moment-generating functions of functionals of Markov chains, sometimes called *gauge functions*. Lemma 5.1, below, is a counterpart of a result in Simon [1979, p. 117] for Brownian motion. Simon cites earlier work in the mathematical-physics literature, and points out that the result extends to general strong Markov processes; for completeness, we include a proof of the result in the form required for our setting. This result could prove useful in the analysis of other likelihood ratios for Markov chains with boundaries.

LEMMA 5.1. *Let q be a nonnegative, real-valued function on the state space of X. If*

$$\gamma \equiv \sup_x \bar{E}_x\left[\sum_{n=0}^{\infty} q(X_n)\right] < 1,$$

then

$$\sup_x \bar{E}_x\left[\exp\left\{\sum_{n=0}^{\infty} q(X_n)\right\}\right] \le (1-\gamma)^{-1} < \infty.$$

PROOF. We have

$$\begin{aligned}
\bar{E}_x\left[\sum_{0\le n_1\le n_2}^{\infty} q(X_{n_1})q(X_{n_2})\right] &= \bar{E}_x\left[\bar{E}_x\left[\sum_{0\le n_1\le n_2}^{\infty} q(X_{n_1})q(X_{n_2}) \mid X_1,\ldots,X_{n_1}\right]\right] \\
&= \sum_{n_1=0}^{\infty} \bar{E}_x\left[q(X_{n_1})\bar{E}_{X_{n_1}}\left[\sum_{J=0}^{\infty} q(X_J)\right]\right] \\
&\le \bar{E}_x\left[\sum_{n_1=0}^{\infty} q(X_{n_1})\right]\gamma \\
&\le \gamma^2.
\end{aligned}$$

Thus, by symmetry,

$$\frac{1}{2}\bar{E}_x\left[\left(\sum_{n=1}^{\infty} q(X_n)\right)^2\right] \le \gamma^2,$$

and by an analogous argument

$$\frac{1}{k!}\bar{E}_x\left[\left(\sum_{n=1}^{\infty} q(X_n)\right)^k\right] \le \gamma^k,$$

for all $k = 0, 1, 2, \ldots$. By the monotone convergence theorem, we thus have

$$\bar{E}_x\left[\exp\left\{\sum_{n=1}^{\infty} q(X_n)\right\}\right] \le \sum_{k=0}^{\infty} \gamma^k = (1-\gamma)^{-1}. \qquad \square$$

The particular form of this result we employ is the following:

COROLLARY 5.2. *For* $\theta > 1$, *if*

$$\log(\theta)\left(\sup_x \overline{E}_x[N]\right) < 1,$$

then

$$\sup_x \overline{E}_x[\theta^N] < \infty.$$

PROOF. Take $q(x) = \log \theta \cdot 1_{\{x_d = 0\}}$ in Lemma 5.1. □

To establish asymptotic efficiency, we need to verify the hypothesis of Corollary 5.2 for $\theta = a^2$, with a as in (20). We give particular attention to the case $d = 2$, then generalize to arbitrarily many nodes.

5.2 Two-Node Networks

In a two-node tandem network, we have

$$a = \frac{\mu_1 + \mu_2}{\mu_1 + \lambda},$$

and N counts the total number of visits to the horizontal boundary

$$A = \{(x_1, x_2) : x_1 > 0, x_2 = 0\}.$$

(Notice that A does not include the origin; see the remark that follows Proposition 3.1.) To apply Corollary 5.2, we need to bound $\overline{E}_x[N]$. We do this through a sequence of lemmas, beginning with a known result (see, e.g., Asmussen 1987, pp. 90–91):

LEMMA 5.3. *Consider an* $M/M/1$ *queue with service rate* λ *and arrival rate* μ, $\lambda < \mu$. *If the system starts with one job present, the probability that it ever empties is* λ/μ.

Now consider again the (μ_2, μ_1, λ) tandem system. Let

$$T_+ = \inf\{n \geq 0 : X_n \notin A\},$$

so that T_+ is the first time the process is outside the horizontal boundary. Let

$$T_A = \inf\{n \geq T_+ : X_n \in A\},$$

so that T_A is the time of the first visit to A after at least one visit to the complement of A. When $\mu_2 < \mu_1$, let $\overline{P}_{\tilde{\pi},1}$ denote the law of the process started in state $(X_0^1, 1)$, with X_0^1 having distribution $\tilde{\pi}$ and

$$\tilde{\pi}(k) = \left(\frac{\mu_2}{\mu_1}\right)^k \left(1 - \frac{\mu_2}{\mu_1}\right), \qquad k = 0, 1, \ldots .$$

This is the stationary distribution for the first queue in the (μ_2, μ_1, λ) system.

LEMMA 5.4. *If* $\mu_2 < \mu_1$, *then* $\bar{P}_{\tilde{\pi},1}(T_A < \infty) = \lambda/\mu_2$.

PROOF. When the first queue is given its stationary distribution, its departure process becomes Poisson with rate μ_2 and the second queue becomes M/M/1 with arrival rate μ_2 and service rate λ; thus, the result follows from Lemma 5.3. □

LEMMA 5.5. *The following decrease in* i *for* $i = 1, 2 \ldots : \bar{P}_{i,0}(T_A < \infty)$ *and* $\bar{P}_{i,0}(N > n)$ *for all* $n \geq 1$.

PROOF. Both claims follow from straightforward sample-path arguments. For the second claim, start two copies of the vector queue-length process in states $(i, 0)$ and $(i + 1, 0)$, respectively. The number of jobs at node 2 in the second copy is never less than that in the first. Hence, by the time the second process makes its nth visit to A, the first process has already done so. The first claim works the same way. □

A consequence of Lemma 5.5 is that $\bar{P}_{i,0}(T_A < \infty)$ is maximized over $i \geq 1$ at $i = 1$. Let p^* be a constant for which

$$\bar{P}_{1,0}(T_A < \infty) \leq p^*.$$

We now have

LEMMA 5.6. *If* $p^* < 1$, *then* $\sup_x \bar{E}_x[N] \leq (1 - p^*)^{-1}(\mu_1 + \mu_2)/\mu_1$.

PROOF. For the supremum over x it suffices to consider $x \in A$, and then in light of Lemma 5.5 it suffices to consider the single starting state $(1, 0)$. From this state, the total number of returns to A preceded by visits to the complement of A is stochastically bounded by a geometric random variable with parameter p^*. At each return to A, the process makes geometrically many transitions along the boundary, with parameter $\mu_2/(\mu_1 + \mu_2)$. Thus, the total number of transitions on A is stochastically bounded by the sum of geometrically many geometric random variables, all independent of each other. The mean number of visits is bounded by the product of the means of the two geometric distributions. □

Now we bound p^*. Starting from state $(1, 0)$, the process makes geometrically many transitions along the boundary before leaving it. In particular,

$$\bar{P}_{1,0}(T_A < \infty) = \bar{P}_{v,1}(T_A < \infty),$$

where

$$v(k) = \left(\frac{\mu_2}{\mu_1 + \mu_2}\right)^k \left(\frac{\mu_1}{\mu_1 + \mu_2}\right), \qquad k = 0, 1, \ldots .$$

If $\mu_2 < \mu_1$, then the distribution $\tilde{\pi}$ exists and we have

$$\begin{aligned}\overline{P}_{v,1}(T_A < \infty) &= \sum_k v(k)\overline{P}_{k,1}(T_A < \infty) \qquad (21)\\ &= \sum_k \tilde{\pi}(k)\overline{P}_{k,1}(T_A < \infty)(v(k)/\tilde{\pi}(k))\\ &\le \left(\sum_k \tilde{\pi}(k)\overline{P}_{k,1}(T_A < \infty)\right) \sup_k (v(k)/\tilde{\pi}(k))\\ &= \overline{P}_{\tilde{\pi},1}(T_A < \infty) \sup_k (v(k)/\tilde{\pi}(k))\\ &= (\lambda/\mu_2) \sup_k (v(k)/\tilde{\pi}(k)), \qquad (22)\end{aligned}$$

the last equality following from Lemma 5.4. Next, observe that

$$\frac{v(k)}{\tilde{\pi}(k)} = \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^k \frac{\mu_1^2}{(\mu_1 + \mu_2)(\mu_1 - \mu_2)} \le \frac{\mu_1^2}{\mu_1^2 - \mu_2^2}.$$

Thus, we may set

$$p^* = \frac{\lambda\mu_1^2}{\mu_2(\mu_1^2 - \mu_2^2)}.$$

Whenever this is less than 1, we get the bound in Lemma 5.6. Thus, we arrive at our main result for two-node networks:

THEOREM 5.7. *If* $\mu_2 < \mu_1$, $p^* < 1$ *and*

$$(1 - p^*)^{-1}\left(\frac{\mu_1 + \mu_2}{\mu_1}\right)\left(2\log\left(\frac{\mu_1 + \mu_2}{\mu_1 + \lambda}\right)\right) < 1,$$

then the estimator in (14) is asymptotically efficient and has linearly bounded relative error. If, in addition, (9) holds, then the estimator has bounded relative error.

PROOF. If the hypotheses of the theorem hold, then from Corollary 5.2, Lemma 5.6, and (20), we conclude that $\overline{E}[a^{2N}; T_K < T_0]$ is bounded uniformly in K, and therefore that the second moment of (14) is bounded by a constant times $(\lambda/\mu_2)^{2(K-1)}$. Combining this with the lower bound on p_K in Theorem 2.3 proves asymptotic efficiency and shows that

$$RE \le \text{constant} \cdot \frac{\sqrt{(\lambda/\mu_2)^{2(K-1)}}}{(\lambda/\mu_2)^K K^{-1}} \le \text{constant} \cdot K.$$

If (9) holds, then by Proposition 2.4 the factor K may be omitted, resulting in a uniform bound on the relative error. □

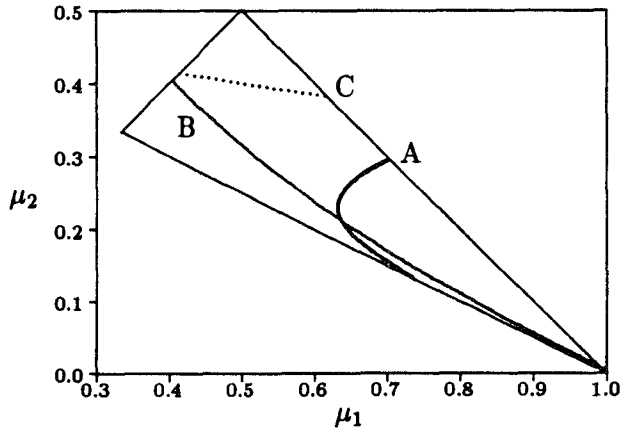

Fig. 1. Conditions for asymptotic efficiency and bounded relative error in a two-node system

The conclusions of Proposition 4.1 and Theorem 5.7 are illustrated in Figure 1. The triangular region in the figure is the set of possible parameter values (μ_1, μ_2) for a two-node system with $\lambda + \mu_1 + \mu_2 = 1$. Points to the right of curve A satisfy our sufficient condition for asymptotic efficiency. Points above curve B satisfy condition (9), so in the intersection of these regions the estimator has bounded relative error. Points above curve C violate our necessary condition for asymptotic efficiency.

Remark. A sharper bound on (21) is provided by the value of the optimization problem

$$\begin{aligned} \text{maximize} \quad & \sum_k v(k)a_k \\ \text{subject to} \quad & \sum_k \pi(k)a_k = \lambda/\mu_2 \\ & 0 \le a_k \le 1. \end{aligned}$$

Because $v(k)$ decreases with k, this problem is easy to solve. However, we have found numerically that it very rarely improves the simpler bound in Lemma 5.6.

5.3 Multinode Networks

Our overall approach to proving asymptotic efficiency for multinode networks is the same as that used for two-node networks, but there are some important differences in the details. The two-node network offers simplifying features not present in the general case, so the results we obtain here, even when specialized to $d = 2$, are not as effective as those in the previous subsection.

In the multinode setting, we take

$$A = \bigcup_\iota A_\iota = \left\{(x_1, \ldots, x_d) : x_d = 0, \sum_\iota x_\iota > 0\right\},$$

and (as before) let N be the total number of visits to A, and T_A the time of the first visit to A preceded by at least one visit to the complement of A. We bound $\sup_x \overline{E}_x[N]$ by bounding $\sup_{x \in A} \overline{P}_x(T_A < \infty)$ (leading to a bound on the number of *distinct* visits to A) and by bounding the number of transitions on A at each distinct visit. Each of these steps is somewhat more complicated in the multinode setting than it was before. Throughout this section, we restrict attention to the case

$$\mu_d < \mu_1, \ldots, \mu_{d-1},$$

so that under the new measure only the last node is unstable.

LEMMA 5.8. $\sup_{x \in A} \overline{P}_x(T_A < \infty) \leq p^*$, *where*

$$p^* = \frac{\lambda(\mu_1 + \cdots + \mu_d)}{\mu_d^2 \tilde{\pi}(0)}, \tag{23}$$

and

$$\tilde{\pi}(0) = \prod_{i=1}^{d-1} \left(1 - \frac{\mu_d}{\mu_i}\right).$$

PROOF. Much as in the two-node case, the supremum over $x \in A$ of $\overline{P}_x(T_A < \infty)$ is easily seen to be attained at $x = (0, 0, \ldots, 1, 0)$, so we consider this starting state. Notice that $\overline{P}_{0,0,\ldots,1,0}(T_A < \infty) = \overline{P}_{v,1}(T_A < \infty)$ where v is the distribution of the first $d-1$ components of the state at the arrival of the first job to node d, starting from state $(0, 0, \ldots, 1, 0)$. The first arrival to node d is the first departure from the subnetwork consisting of nodes $1, \ldots, d-1$. Let us write $\tilde{X}_t = (\tilde{X}_t^1, \ldots, \tilde{X}_t^{d-1})$ for the (right-continuous) continuous-time queue-length process of this subnetwork, and let T_d be the time of the first departure from this subnetwork. For any $\tilde{x} = (x_1, \ldots, x_{d-1})$,

$$v(\tilde{x}) = \overline{P}_{0,\ldots,1,0}\left(\tilde{X}_{T_d} = \tilde{x}\right),$$

the subscript on $\overline{P}$ still referring to the full d-dimensional state. Let $\tilde{x}' = \tilde{x} + (0, \ldots, 0, 1)$; then

$$\begin{aligned}\overline{P}_{0,\ldots,1,0}\left(\tilde{X}_{T_d} = \tilde{x}\right) &\leq \overline{P}_{0,\ldots,1,0}\left(\tilde{X} \text{ visits } \tilde{x}' \text{ before } T_d\right) \\ &\leq \overline{P}_0\left(\tilde{X} \text{ visits } \tilde{x} \text{ before } T_d\right).\end{aligned}$$

Moreover, if we let $\tilde{T}_0$ be the time of the first return of $\tilde{X}$ to the origin, then

$$\overline{P}_0\left(\tilde{X} \text{ visits } \tilde{x} \text{ before } T_d\right) \leq \overline{P}_0\left(\tilde{X} \text{ visits } \tilde{x} \text{ before } T_0\right).$$

Thus, we have shown that

$$v(x_1, \ldots, x_{d-1}) \leq \overline{P}\left(\tilde{X} \text{ visits } (x_1, \ldots, x_{d-1}) \text{ in a 0-cycle}\right), \tag{24}$$

the 0-cycle referring to the process $\tilde{X}$. Now let

$$\tilde{\pi}(x_1, \ldots, x_{d-1}) = \prod_{i=1}^{d-1} \left(\frac{\mu_d}{\mu_i}\right)^{x_i} \left(1 - \frac{\mu_d}{\mu_i}\right)$$

denote the stationary distribution of $\tilde{X}$. The mean length of a 0-cycle for $\tilde{X}$ is $1/(\mu_d \tilde{\pi}(0))$; given that $\tilde{X}$ reaches $(x_1, \ldots, x_{d-1})$ in a cycle, the expected time it spends there is no less than $1/(\mu_1 + \cdots + \mu_d)$, the minimal mean holding time in any state. It follows from the regenerative representation of $\tilde{\pi}$ that

$$\begin{aligned}\bar{P}\big(\tilde{X} \text{ visits } \tilde{x} = (x_1, \ldots, x_{d-1}) \text{ in a 0-cycle}\big) \\ = \frac{\tilde{\pi}(x_1, \ldots, x_{d-1}) \bar{E}_0\big[\tilde{T}_0\big]}{\bar{E}_0\big[\text{time spent in } \tilde{x} \mid \tilde{X} \text{ visits } \tilde{x} \text{ in cycle}\big]}.\end{aligned}$$

Thus,

$$\begin{aligned}\bar{P}\big(\tilde{X} \text{ visits } \tilde{x} = (x_1, \ldots, x_{d-1}) \text{ in a 0-cycle}\big) \\ \le \tilde{\pi}(x_1, \ldots, x_{d-1}) \frac{\mu_1 + \cdots + \mu_{d-1}}{\mu_d \tilde{\pi}(0)},\end{aligned}$$

which together with (24) yields

$$\frac{v(x_1, \ldots, x_{d-1})}{\tilde{\pi}(x_1, \ldots, x_{d-1})} \le (\mu_1 + \cdots + \mu_d)/(\mu_d \tilde{\pi}(0)).$$

Arguing just as in (22), we conclude that p^* in (23) is indeed an upper bound on $\bar{P}_{0, \ldots, 1, 0}(T_A < \infty)$. □

We now have

THEOREM 5.9. *If* $p^* < 1$ *and*

$$\frac{1}{1 - p^*}\left[(d-1) + \left(\frac{\mu_d}{\mu_1} + \cdots + \frac{\mu_d}{\mu_{d-1}}\right)\right] \cdot 2\log(a) < 1,$$

then the estimator in (14) is asymptotically efficient and has linearly bounded relative error. If, in addition, (9) holds, then the estimator has bounded relative error.

PROOF. It suffices to show that the term in square brackets is an upper bound on the expected number of transitions on A at each distinct visit to A; once we establish that, the rest of the argument is just as in Theorem 5.7.

The expected number of transitions until the first exit from A is clearly maximized (over A) at the initial state $(1, 0, \ldots, 0)$; the required number of transitions can be made pathwise smaller from any other initial state. Starting from $(1, 0, \ldots, 0)$, the first exit from A occurs when the job at node 1 reaches node d. The number of transitions required for this job to complete service at each node i, $i = 1, \ldots, d-1$, is stochastically bounded by a geo-

metric random variable with mean $(\mu_i + \mu_d)/\mu_i$, so the expected number of transitions for the job to reach node d is bounded by

$$\frac{\mu_1 + \mu_d}{\mu_1} + \cdots + \frac{\mu_{d-1} + \mu_d}{\mu_{d-1}},$$

which is equivalent to the expression in square brackets in the statement of the theorem. □

Examples of parameter values satisfying the hypotheses of Theorem 5.9 are $(\lambda, \mu_1, \mu_2, \mu_3) = (0.0012, 0.4431, 0.4873, 0.0684)$, and (0.0001, 0.8, 0.17, 0.0299).

6. CONCLUSION

We have analyzed an importance sampling estimator for overflow probabilities in tandem Jackson networks that generalizes the asymptotically optimal estimator for M/M/1 queues. We have shown that in certain parameter regions this estimator has linearly bounded and even purely bounded relative error, but that in other regions it is not even asymptotically efficient. These results follow from upper and lower bounds on the first and second moments of the estimator.

The techniques used to bound the second moment may be of broader applicability in the analysis of importance sampling estimators, so we briefly summarize them. We obtained lower bounds by evaluating the square of the estimator on individual paths; well-chosen paths have sufficiently high probability and a sufficiently large squared-estimator value to rule out asymptotic efficiency. We obtained upper bounds by separating the contributions to the likelihood ratio due to the interior of the state space and the boundaries. The contribution of the interior is easily handled, because in its interior the queue-length process is spatially homogeneous. The boundaries, however, contribute factors to the likelihood ratio that are exponential in the number of transitions on boundaries; these are more problematic. A "gauge function lemma" bounds the expectations of these exponential factors in terms of the expected number of visits to the boundaries. This part of the analysis seems fairly general. The final step—bounding the expected number of visits to boundaries—exploits particular features of the model.

REFERENCES

ANANTHARAM, V. 1989. The optimal buffer allocation problem. *IEEE Trans. Inf. Theor. 35*, 721–725.

ANANTHARAM, V., AND GANESH, A. Correctness within a constant of an optimal buffer allocation rule of thumb. *IEEE Trans. Inf. Theor.* (to appear).

ANANTHARAM, V., HEIDELBERGER, P., AND TSOUCAS, P. 1990. Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. IBM Res. Rep. RC16820. Yorktown Heights, N.Y.

ASMUSSEN, S. 1982. Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue. *Adv. Appl. Prob. 14*, 143–170.

ASMUSSEN, S. 1987. *Applied Probability and Queues*. Wiley, New York.

BUCKLEW, J. A. 1990. *Large Deviations Techniques in Decision, Simulation, and Estimation*. Wiley, New York.

CHANG, C. S., HEIDELBERGER, P., JUNEJA, S., AND SHAHABUDDIN, P. 1994 Effective bandwidth and fast simulation of ATM Intree Networks. *Perform. Eval. 20*, 45–65.

DUPUIS, P., ISHII, H., AND SONER, H. M. 1990. A viscosity solution approach to the asymptotic analysis of queueing systems. *Ann. Probab. 18*, 226–255.

FRATER, M. R., LENON, T. M., AND ANDERSON, B. D. O. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Trans. Autom. Control TAC-36*, 1395–1405.

GLASSERMAN, P., AND KOU, S.-G. 1993. Overflow probabilities in Jackson networks. In *Proceedings of the 32nd IEEE Conference on Decision and Control*. IEEE Press, New York, 3178–3182.

GLYNN, P. W., AND IGLEHART, D. L. 1989. Importance sampling for stochastic simulations. *Manage. Sci. 35*, 1367–1392.

JANSON, S. 1986. Moments for first-passage and last-exit times, the minimum, and related quantities for random walks with positive drift. *Adv. Appl. Probab. 18*, 865–879.

KELLY, F. J. 1979. *Reversibility and Stochastic Networks*. Wiley, New York.

KESIDIS, G., AND WALRAND, J. 1993. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. *ACM Trans. Model. Comput. Simul. 3*, 269–276.

PAREKH, S., AND WALRAND, J. 1989. Quick simulation of rare events in networks. *IEEE Trans. Autom. Control TAC-34*, 54–66.

SADOWSKY, J. S. 1991. Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queues. *IEEE Trans. Autom. Control TAC-36*, 1383–1394.

SIEGMUND, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist. 4*, 673–684.

SIMON, B. 1979. *Functional Integration and Quantum Physics*. Academic Press, New York.

TSOUCAS, P. 1992. Rare events in series of queues. *J. Appl. Probab. 29*, 168–175.

WEBER, R. R. 1979. The interchangeability of M/M/1 queues in series. *J. Appl. Probab. 16*, 690–695.

Received June 1994; revised January 1995; accepted February 1995