Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

A BALANCED LIKELIHOOD RATIO APPROACH FOR ANALYZING RARE EVENTS IN A TANDEM JACKSON NETWORK

Bruce C. Shultes

Department of Mechanical, Industrial, and Nuclear Engineering
University of Cincinnati
PO Box 210116
Cincinnati, OH 45221-0116, U.S.A.

ABSTRACT

Balanced likelihood ratio importance sampling methods were originally developed for the analysis of fault-tolerant systems. This paper provides a basis for adapting this approach to analyze the rare event probability that total system size reaches a bound before returning to zero in tandem Jackson networks. An optimal importance sampling distribution for the single server case is derived through direct application of the balanced likelihood ratio approach. The generalization of this approach to larger systems is explored via a two-node tandem Jackson network. A general heuristic approach is outlined along with certain open questions whose answers could lead to a more robust solution. Asymptotic characteristics of the proposed importance sampling approach for the two-node network are discussed. Bounded relative error is only possible under certain conditions. Numerical results illustrate the benefits of the approach.

1 INTRODUCTION

The analysis of rare event probabilities in tandem Jackson networks (for an introduction to Jackson networks see Chapter 1, Serfozo 1999) has received a lot of attention in the past decade. These models are simplified versions of switched telecommunications networks and other systems that can be modeled as networks of queues, i.e., manufacturing processes and computer networks. Efficient methods for analyzing the probability that capacities or buffer sizes are not large enough are needed to accurately assess system reliability during system design. A buffer can cover the total system population or can be associated with one or more nodes in the network. Most work in this area considers a single buffer that covers the total system.

Importance sampling has been a popular approach for efficiently estimating rare event probabilities (see Heidelberger 1995). Zero-variance importance sampling, i.e., an importance sampling distribution that yields a constant value for every sample, is theoretically possible but typically impossible because it implies perfect knowledge of the performance parameter being estimated. Kuruganti and Strickland (1997) identify properties that characterize zero-variance importance sampling distributions. Unfortunately, utilizing these properties does not appear to improve the complexity of this problem. Juneja (Juneja 1993, 2001) develops these properties as a basis for identifying asymptotically optimal importance sampling distributions. Approaches for approximating a zero-variance importance sampling distribution by directly minimizing the variance of the importance sampling estimator or by minimizing the cross-entropy between the proposed importance sampling distribution and a zero variance distribution have been developed (see Rubinstein 1997, Lieber et al. 1999, and De-Boer et al. 2000).

Large deviations theory is a common tool for deriving and analyzing importance sampling estimators. Parekh and Walrand (1989) introduced a heuristic importance sampling distribution for estimating the probability that total system size reaches some bound before returning to 0 in tandem Jackson networks with single server nodes. The distribution interchanges the arrival rate with the smallest service rate in the network. The efficiency of this method for a GI/GI/m queue was established by Sadowski (1991). Frater et al. (1991) extended this approach to Jackson networks consisting of single server nodes. Justifications for the Parekh and Walrand heuristic were formalized by Tsoucas (1992). Glasserman and Kou (1995) analyzed this heuristic for tandem Jackson networks consisting of single server nodes and established necessary and sufficient conditions for this heuristic to be asymptotically efficient.

The balanced likelihood ratio approach to importance sampling (see Alexopoulos and Shultes 1998, 2001) was developed for the analysis of rare events in fault-tolerant repairable systems. In that context, importance sampling estimators have been shown to guarantee variance reduction and yield bounded relative error. Unfortunately, guaranteed variance reduction does not guarantee computa-

tional efficiency because variance reduction may come at the expense of computation time. This paper applies the general idea behind the balanced likelihood ratio approach to identify importance sampling distributions for estimating the probability that total system size reaches some bound before returning to 0 in tandem Jackson networks. The approach in this paper differs significantly from previous work. For a single node system, a closed form, zero variance importance sampling distribution for estimating this rare event probability that does not explicitly use knowledge of the true value for the unknown parameter is derived. This provides new insights into the construction of good importance sampling distributions for tandem queues. The proposed importance sampling distribution for tandem queues exploits queueing network structure that differs from the structure found in fault-tolerant repairable systems. Other properties of balanced likelihood ratio methods are maintained, including: guaranteed variance reduction over standard Monte Carlo methods, likelihood ratios that are bounded from above by one, and a distribution that adapts throughout sample paths. Some asymptotic properties of the proposed importance sampling estimators are known, while others are still under investigation. A preliminary summary of these properties is provided.

The remainder of this paper is setup as follows: Section 2 presents the model studied and provides an overview of importance sampling and issues related to identifying "good" importance sampling distributions. Section 3 provides the details of the proposed approach for a single node system. Section 4 explores the extension of this approach to tandem queues. Experimental results are discussed in Section 5 and Section 6 provides conclusions and directions for future research.

2 BACKGROUND

Consider a tandem Jackson network with m nodes. Let $X_i(t)$ be the number of customers at node i in the network at time t, let s_i be the number of servers at node i and let b denote the total system (buffer) size. Inter-arrival times are independent exponential random variables with rate λ and customer service times at node i are exponential with rate μ_i. Assume all customers are identical, then the system can be modeled as a continuous time Markov chain (CTMC) with system state $Y(t) = (X_1(t), \ldots, X_m(t))$. The state space is irreducible and finite, so the CTMC is a regenerative process. Typically, the system empty state is used as a regeneration point.

A performance parameter of interest is the time until a buffer fills, i.e., the total system size reaches b, given that the system starts empty. Time until buffer *overflow* can be achieved by setting b one larger than the actual buffer size. Let T denote the time between visits to the empty state, and τ denote the time until system size reaches b starting from the empty state. For tandem queues, if the vector queue length process is regenerative then the time until the buffer fills can be written as

$$E[\tau] = \frac{E[\min(\tau, T)]}{P(\tau < T)},$$

as presented by Frater et al. (1991). The vector queue length process does not have to be regenerative for this ratio representation to hold, but only the regenerative case is considered here. Confidence intervals can be obtained using the same formulas as for standard regenerative ratios (see Law and Kelton 2000, pp. 531-533). To efficiently estimate $E[\tau]$ requires efficient methods for estimating $E[\min(\tau, T)]$ and $P(\tau < T)$.

2.1 Focus on Rare Event Probabilities

Attention is restricted to cases where $\lambda < \min_i \mu_i$. This is not the system stability condition since node i has s_i servers. However, if $\lambda \geq \min_i \mu_i$, then it can be difficult to estimate $E[\min(\tau, T)]$ since the empty state may not be visited frequently. The upper bound on the arrival rate implies that estimating $E[\min(\tau, T)]$ is not difficult. In contrast, estimating the rare event probability $P(\tau < T)$ is difficult. The rest of this paper focuses on estimating $P(\tau < T)$.

The quantity $P(\tau < T)$ only depends on the embedded discrete time Markov chain (DTMC). In contrast, the estimation of $E[\min(\tau, T)]$ requires the CTMC. Consider a generic system state $(x_1, \ldots, x_m)$. The total rate of event transitions out of this state is

$$r(x_1, \ldots, x_m) = \lambda + \sum_{i=1}^{m} \min(x_i, s_i)\mu_i.$$

The one-step transition probabilities for the embedded DTMC are: the probability the next event is an arrival is $\lambda / r(x_1, \ldots, x_m)$ and the probability the next event is a service completion at node i is $\min(x_i, s_i)\mu_i / r(x_1, \ldots, x_m)$.

2.2 Importance Sampling

Let Ω denote the set of all possible regenerative cycles and for each $\omega \in \Omega$, let $B(\omega)$ denote the largest system size observed within the cycle. Since the problem only requires simulation of a DTMC, the probability regenerative cycle ω is observed $P(\omega)$ is the product of one-step transition

probabilities. Importance sampling allows a new distribution P' to be defined such that $P(\omega) > 0 \Rightarrow P'(\omega) > 0$ and

$$P(\tau < T_0) = \sum_{\omega \in \Omega} 1(B(\omega) = b) \frac{P(\omega)}{P'(\omega)} P'(\omega)$$
$$= \sum_{\omega \in \Omega} 1(B(\omega) = b) L(\omega) P'(\omega),$$

where the likelihood ratio $L(\omega)$ is the Radon-Nikodym derivative of P with respect to P'. If the importance sampling distribution is defined in terms of one-step transition probabilities then the likelihood ratio $L(\omega)$ can be decomposed into a product of one-step *event likelihood ratios* associated with each individual event within the cycle.

2.2.1 Asymptotic Properties

Relative error, also known as the coefficient of variation, is the ratio of the standard deviation of an estimator over its expected value. Bounded relative error refers to the behavior of an estimator as the quantity to be estimated is reduced to zero which occurs by varying a *rarity* parameter for the system under study. An estimator yields *bounded relative error* if the relative error remains bounded as the quantity to be estimated approaches zero. This implies that the computational effort or sample size required to achieve a desired level of accuracy (relative half-width) remains bounded in the limit. There are a couple of ways to force $P(\tau < T)$ to zero. The buffer size can be increased to infinity or the arrival rate can be decreased to zero. The former approach is the most common approach used in the queueing network literature.

An estimator is *asymptotically efficient* if the relative error grows at a sub-exponential rate as the quantity to be estimated approaches zero. This implies that the number of samples required to achieve a desired level of accuracy grows at a sub-exponential rate. Several importance sampling distributions in the literature have been shown to have linearly bounded relative error when *b* is the rarity parameter (for example, see Glasserman and Kou 1995, and Kroese and Nicola 1999). Bounded relative error implies asymptotic efficiency.

2.2.2 Variance Reduction Ratio

Variance reduction does not necessarily equate to computational savings. When comparing two sampling procedures, it is not fair to use a fixed sample size if one method utilizes more computation time than the other. Fixing the computational effort would lead to different sample sizes. For example, each regenerative cycle generated consists of many events and the number of events within each cycle can vary significantly. The *variance reduction ratio* (VRR) measures the trade-off between improved variance and the additional computation cost associated with this improvement. Specifically, to compare two estimators, a ratio of the corresponding variances is multiplied by a ratio of the corresponding computational effort, i.e., the number of events sampled to generate the variance. If a VRR is less than one, then the approach in the numerator is more efficient. If a VRR is greater than one, then the approach in the denominator is more efficient. Typically, VRRs are estimated empirically.

2.2.3 Balanced Likelihood Ratios

The proposed importance sampling approach is an extension of the balanced likelihood ratio approach developed for estimating the reliability of fault-tolerant repairable systems (see Alexopoulos and Shultes 1998, 2001). In that case, all system events are classified into two classes: events that move the system "closer" to system failure, i.e., component failure events, and events that move the system "away" from system failure, i.e., repair completions. A similar classification is possible for tandem Jackson networks: events that increase system size, i.e., arrival events, and events that reduce system size, i.e., service completions. These classifications are similar, especially for single node queueing networks.

The balanced likelihood ratio method "balances" event likelihood ratios from these two classes. Specifically, from a regenerative state, if all events fall into one of these two classes then the following properties exist:

- Every arrival event increases the system size and is followed by a sequence of service completions that cancel out the arrival event when the customer leaves the system.
- To achieve zero-variance when estimating functionals that are only non-zero when the rare-event is observed, every cycle must visit the rare-event. Hence, events that would complete a cycle before the system experiences the rare-event should have zero probability in the IS distribution.
- If an IS distribution increases the probability of arrival events, then arrival event likelihood ratios are less than one. These event likelihood ratios can be used as multipliers to reduce corresponding service probabilities. Reducing service probabilities increases arrival probabilities.
- Once buffer overflow is observed, the IS distribution can revert to standard Monte Carlo sampling probabilities. This is called Dynamic Importance Sampling (DIS) in the literature.

To summarize, each customer in the system experiences a series of events. Each event accumulates an event likelihood ratio. At any time, the product of event likelihood ratios accumulated for a customer is less than one. When a

customer leaves, then the product of the corresponding event likelihood ratios becomes one. This idea is similar to the cyclic approach proposed by Juneja (2001), but there are significant differences. Balanced likelihood ratio methods do not explicitly consider sequences of events that return a system to a fixed state. Instead, events are grouped with corresponding event likelihood ratios so that at all times within a cycle, the product of event likelihood ratios is bounded from above by one.

3 A SINGLE QUEUE

With only one node, the node index i is dropped for the rest of this section. Applying the importance sampling approach described in Section 2, there are exactly two types of events in the system: customer arrivals and service completions. The first two events within a cycle must be customer arrivals, otherwise the cycle would end before reaching the bound on system size. Each service completion will cancel out exactly one event likelihood ratio associated with a customer arrival.

The event likelihood ratio associated with an arrival when there are $x \geq 1$ customers in the system is

$$l[x] = \frac{\lambda/(\lambda + \min(x,s)\mu)}{1 - l[x-1]\left(\min(x,s)\mu/(\lambda + \min(x,s)\mu)\right)} = \frac{\lambda}{\lambda + (1 - l[x-1])\min(x,s)\mu}$$

and $l[0]$ is artificially set to zero. Similarly, the event likelihood ratio associated with a service completion when there are $x > 1$ customers in the system is

$$\bar{l}[x] = \frac{\min(x,s)\mu/(\lambda + \min(x,s)\mu)}{l[x-1]\left(\min(x,s)\mu/(\lambda + \min(x,s)\mu)\right)}$$

which equals $1/l[x-1]$. Service completions are not allowed if $x \leq 1$ and system size has not reached b. The cancellation of arrival and service event likelihood ratios guarantees that the importance sampling distribution leads to constant likelihood ratios and the constant value is $P(\tau < T)$. Notice that this is not asymptotic optimality and the distribution does not depend on the buffer size.

3.1 Asymptotic Behavior

Since the balanced likelihood ratio importance sampling distribution is optimal, it is not necessary to consider the asymptotic properties described in Section 2.2.1. However, it is interesting to examine the effects of increasing b or reducing λ on this importance sampling distribution.

The following two lemmas identify the trend in arrival event likelihood ratios for fixed b and λ. Two trends are handled simultaneously: increasing (*decreasing*).

Lemma 1 *For* $x \geq 1$, *if* $l[x-1] < (>) \lambda/\min(x,s)\mu$ *then* $l[x] < (>) \lambda/\min(x,s)\mu$.

Proof Assume $l[x-1] < (>) \lambda/\min(x,s)\mu$. Then

$$\lambda - l[x-1]\min(x,s)\mu > (<) 0, \text{ and}$$

$$\lambda + (1 - l[x-1])\min(x,s)\mu > (<) \min(x,s)\mu, \text{ then}$$

$$l[x] = \frac{\lambda}{\lambda + (1 - l[x-1])\min(x,s)\mu} < (>) \frac{\lambda}{\min(x,s)\mu}. \quad \blacksquare$$

Lemma 2 *For* $x \geq 1$, *if* $l[x-1] < (>) \lambda/\min(x,s)\mu$ *then* $l[x] > (<) l[x-1]$.

Proof Assume $l[x-1] < (>) \lambda/\min(x,s)\mu$. Then

$$(l[x-1]\min(x,s)\mu - \lambda)(l[x-1] - 1) > (<) 0.$$

Multiplying this equation out and rearranging terms yields

$$l[x] = \frac{\lambda}{\lambda + (1 - l[x-1])\min(x,s)\mu} > (<) l[x-1]. \quad \blacksquare$$

The proof of Lemma 3 relies on two observations. Lemma 2 implies that arrival event likelihood ratios strictly increase (*decrease*) until the bound in Lemma 1 is reached. Lemma 1 implies that arrival event likelihood ratios never cross over the bound.

Lemma 3 $\lim_{x \to \infty} l[x] = \lambda/s\mu$.

Proof Lemmas 1 and 2 imply that the sequence of $l[x]$ must converge to some limit. For $x \geq s$, the event likelihood ratios are either monotonically increasing or monotonically decreasing. There are three possibilities. Either the $l[x]$ converge to $\lambda/s\mu$, they converge to a value $p < \lambda/s\mu$, or they converge to a value $p > \lambda/s\mu$. The latter two cases are symmetric, so it is sufficient to only consider one of them. Consider the case $p < \lambda/s\mu$. The proof is by contradiction. Suppose that the $l[x]$ are converging to a limit $p < \lambda/s\mu$. This implies that for any $\varepsilon > 0$ there should be an infinite sub-sequence of event likelihood ratios that are greater than p-ε and less than p (see Bartle 1976, pp. 90-92). To complete the proof, a contradiction is reached by constructing an $\bar{\varepsilon} > 0$ such that any event likelihood ratio in $(p - \bar{\varepsilon}, p)$ leads to a subsequent event likelihood ratio that is larger than p.

If there exists an $\bar{\varepsilon} > 0$ such that any event likelihood ratio in $(p - \bar{\varepsilon}, p)$ leads to a subsequent event likelihood ratio that is larger than p, then

$$\frac{\lambda}{\lambda + (1-(p-\bar{\varepsilon}))s\mu} > p.$$

Solving this equation for $\bar{\varepsilon} > 0$ yields

$$\bar{\varepsilon} < \left(\frac{1-p}{p}\right)\left(\frac{\lambda - ps\mu}{s\mu}\right).$$

The right hand side is positive. ■

3.1.1 Result: Optimality as b Tends to Infinity

The limit specified in Lemma 3 is independent of b and may not be achieved for small buffer sizes. However, if the buffer size is increased without bound, then the event likelihood ratios converge to the limit in Lemma 3 and the limiting importance sampling arrival and service completion probabilities are the same as those obtained if one swaps the arrival rate with the service rate.

3.1.2 Result: Optimality as λ Tends to Zero

Decreasing the arrival rate or increasing the service rate are alternate ways to force buffer overflows to be more rare. If either of these approaches is considered, then the balanced likelihood ratio method forces the importance sampling probability associated with customer arrivals to approach one. This occurs because the event likelihood ratios associated with arrivals approach zero. This has the same effect as swapping the arrival and service rates in these alternate limits.

4 TANDEM QUEUES

In tandem queues, the approach developed for the single node case requires several modifications. The most significant change is that events fall into m+1 classes instead of just two. Arrival events in tandem queues are similar to arrival events in the single node case. Service completions at the last (m^{th}) node are similar to service completions in the single node case. Unfortunately, service completion at any of the other m-1 nodes are new types of events. These events do not directly move the system closer to nor further away from the bound on system size b. There is an indirect influence in that customers move closer to departure with each of these events.

It is not possible to ignore these m-1 intermediate nodes in the importance sampling distribution. The reason is simple. The goal is to make arrival events more likely. Arrival events are always possible and if an importance sampling distribution increases the likelihood of arrival events then other events must have reduced likelihoods. At various points in time, any mixture of the m different service completions may be possible. Consequently, the importance sampling distribution has to be flexible enough to make any or all m different service completion events less likely. To accomplish this, an arrival event is "paired" with the sequence of m service completions required to push a customer through the tandem network.

In the single node case, arrivals are forced under IS if there is only one customer in the system and the rare event has not been observed. In tandem queues, arrivals are forced under IS if there is only one customer in the system, the customer is at the last node, and the rare event has not been observed. This difference creates a problem because there is no obvious way to start using importance sampling within a cycle. An initial "pseudo" event likelihood ratio of c will be constructed to handle this issue.

A consequence of these differences is that it is not clear whether an optimal importance sampling distribution can be achieved by applying the balanced likelihood ratio approach. However, it should be possible to identify heuristics that work well.

4.1 A Two-Node Tandem Queue

The balanced likelihood ratio approach for tandem queues is illustrated through the two-node case. To mimic the properties of the optimal importance sampling distribution constructed in Section 3, we construct a balanced likelihood ratio approach that yields event likelihood ratios that are strictly less than one. This property guarantees that variance reduction is achieved, but does not guarantee a variance reduction ratio larger than one. The key idea is that each arrival event generates an event likelihood ratio. This event likelihood ratio can be split, via a square root, for use as a multiplier for the probability of one service completion at node 1 and one service completion at node 2.

Before formalizing the approach, consider the following four possible scenarios an arriving customer can encounter in a two-node network: (1) the system is empty, (2) all customers are waiting for service at node 1, (3) all customers are waiting for service at node 2, (4) some customers are waiting for service at node 1 and node 2. The trend in event likelihood ratios can be identified in each case. Without loss of generality, the following 4 points are discussed in the context of single server queues (s=1). Let l^2 be a generic arrival event likelihood ratio.

Case 1: The system is empty. The event likelihood ratio for the next event is one since the next event has to be an arrival. This event likelihood ratio is artificially replaced by c in the implementation, because an event likelihood ratio of one does not allow service completion probabilities associated with this arrival to reduced.

Case 2: All customers in the system are at node 1. The constraint

$$\frac{\lambda/(\lambda+\mu_1)}{1-l\left(\mu_1/(\lambda+\mu_1)\right)}=\frac{\lambda}{\lambda+(1-l)\mu_1}\le l^2$$

on all event likelihood ratios for customer arrival events establishes an upper bound on event likelihood ratios. This constraint allows l to be used as a multiplier for the probability of completing service at each of the two service nodes making these events less likely. Forcing $l < 1$ and solving this constraint for l leads to the inequality

$$l\ge\frac{\lambda+\sqrt{\lambda(\lambda+4\mu_1)}}{2\mu_1}.$$

If l is smaller than this lower bound then the subsequent event likelihood ratio will be larger than l^2.

Case 3: All customers are waiting for service at node 2. This is the same as Case 2 with μ_1 replaced by μ_2.

Case 4: Some customers are waiting for service at node 1 and node 2. This case is the same as Case 2 with μ_1 replaced by $\mu_1+\mu_2$.

Combining the bounds from Cases 2-4 yields

$$l\ge\frac{\lambda+\sqrt{\lambda\left(\lambda+4\min(\mu_1,\mu_2)\right)}}{2\min(\mu_1,\mu_2)}=\underline{l}.$$

If the lower bound on l holds at equality, then the inequality $\lambda<\min(\mu_1,\mu_2)/2$ must hold in order for all event likelihood ratios associated with arrivals to be less than one.

If an event likelihood ratio smaller than $\underline{l}^2$ is used, say l', then the next arrival event likelihood ratio will be larger than l'. This trend continues until an arrival event likelihood ratio larger than $\underline{l}^2$ is observed at which point the trend reverses.

The bounds on l can be extended to multi-server nodes by replacing $\min(\mu_1,\mu_2)$ with an appropriate multi-server rate. To accomplish this, the bound becomes state dependent. Let S_k denote the set of system states with k customers in the system. With single servers, the smallest possible service rate is always $\min(\mu_1,\mu_2)$. For multiple servers this turns into

$$\min_{(x_1,\ldots,x_m)\in S_k}\left(\sum_i\min(x_i,s_i)\mu_i\right).$$

Since the bound must hold for all k, the set of cases covered by the proposed importance sampling distribution still requires $\lambda<\min(\mu_1,\mu_2)/2$. However, the limit of the event likelihood ratios can vary with buffer size b.

Define two stacks: l_1 for node 1 service completion probability multipliers, and l_2 for node 2 multipliers. Initially each stack contains one multiplier, $\sqrt{c}$ on l_1 and 0 on l_2 where the 0 guarantees that an arrival occurs if the system is in state $(0,1)$ and system size has not reached b. Let $\dagger_i$, for i=1 and 2, denote the multiplier at the head of stack i. The event likelihood ratio associated with an arrival event when the system is in state $Y(t)=(x_1,x_2)$ is

$$\begin{aligned}l^2&=\frac{\lambda}{\left(\lambda+\sum_i\min(x_i,s_i)\mu_i\right)}\Big/\left(1-\frac{\sum_j\dagger_j\min(x_j,s_j)\mu_j}{\left(\lambda+\sum_i\min(x_i,s_i)\mu_i\right)}\right)\\&=\frac{\lambda}{\lambda+\sum_i(1-\dagger_i)\min(x_i,s_i)\mu_i}\end{aligned}$$

and then if $(x_1,x_2)\neq(0,0)$, then l is pushed onto both stacks. Similarly, the event likelihood ratio for a service completion at node i is

$$\frac{\min(x_i,s_i)\mu_i}{\left(\lambda+\sum_j\min(x_j,s_j)\mu_j\right)}\Big/\frac{\dagger_i\min(x_i,s_i)\mu_i}{\left(\lambda+\sum_j\min(x_j,s_j)\mu_j\right)}=1/\dagger_i.$$

This procedure for specifying importance sampling probabilities is complete except for the selection of the initial event likelihood ratio c.

4.1.1 Initial Event Likelihood Ratios

There are numerous approaches for specifying an initial event likelihood ratio. The key is that the likelihood ratio associated with a regenerative cycle should not exceed one. One approach that satisfies this condition is to let $c=\max\left(\underline{l}^2,\lambda/(\lambda+\mu_2)\right)$. This construction guarantees that all arrival event likelihood ratios throughout the regenerative cycle are less than or equal to c.

4.1.2 Asymptotic Behavior

With the specified choice for c, an upper bound on the likelihood ratio for an entire cycle is $c^{(b-1)/2}$. This corresponds to a path that hits b when all but one customer is at the second node. Similarly, let $d=\min\left(\underline{l}^2,\lambda/(\lambda+\mu_2)\right)$. A lower

bound on cycle likelihood ratios is $d^{(b-1)}$. The difference between these bounds is

$$c^{(b-1)/2} - d^{b-1} = c^{(b-1)/2}\left(1 - \left(\frac{d}{\sqrt{c}}\right)^{b-1}\right)$$

which is an upper bound for the absolute deviation between any observation and the sample mean. Taking the limit as $b \to \infty$ forces this bound to approach zero. In the limit, the proposed importance sampling distribution is a zero-variance distribution. The proposed importance sampling distribution only achieves bounded relative error if the rate of convergence to the zero-variance distribution is faster than the rate the expected value is converging to zero. It is not clear whether this is true.

If the rarity parameter is the arrival rate for customers λ, then the balanced likelihood ratio importance sampling distribution converges to a zero-variance distribution. It can be shown (omitted for brevity) that, in this limit, the balanced likelihood ratio estimator yields bounded relative error. If the rarity parameter is the maximum system size b, then it is not clear whether bounded relative error is achieved or not. Bounded relative error and asymptotic efficiency are discussed further in Section 5.

4.2 A Modification

The asymptotics of the proposed importance sampling distribution provide a bound on the size of the range of values that can be observed from a single sample path. Relative to the quantity being estimated, this range can be large. Of greater concern is that the largest values can correspond to sequences of events that occur with minimally biased event probabilities. Hence relatively large values may be rare under importance sampling implying that large sample sizes may be needed to reach steady-state in the simulation. This issue also applies to existing IS distributions. A modification is proposed to reduce this problem.

Observe that it is not necessary to bias all service probabilities. To increase the probability of an arrival event, it is sufficient to reduce the probability of only one service event. This leads to the following modification. Consider states where both nodes are busy. For these states, instead of multiplying the probability of a service completion at each node by a multiplier, leave the service completion probability at node 1 alone and multiply the service completion probability at node 2 by a multiplier. If a service completion at node 1 occurs, then the corresponding event likelihood ratio is one. But a multiplier was placed on stack l_1 for this event. This multiplier can be popped off stack l_1 and placed on a new stack l_3. Multipliers on stack l_3 can be used as a second multiplier for service completions at node 2. Let $\overline{l_3}$ be the head of stack l_3 when it is not empty and one otherwise.

The proposed modification changes the form of arrival event likelihood ratios. If the second node is busy then arrival event likelihood ratios have the following form:

$$l = \frac{\dfrac{\lambda}{\left(\lambda + \sum_j \min\left(x_j, s_j\right)\mu_j\right)}}{1 - \dfrac{\min\left(x_1, s_1\right)\mu_1 + l_2 l_3 \min\left(x_2, s_2\right)\mu_2}{\left(\lambda + \sum_j \min\left(x_j, s_j\right)\mu_j\right)}} = \frac{\lambda}{\lambda + \left(1 - l_2 l_3\right)\min\left(x_2, s_2\right)\mu_2}.$$

This implies that importance sampling probabilities for service completions at node 1 revert back to the original probabilities (i.e., no importance sampling) and importance sampling probabilities for service completions at node 2 become

$$\frac{l_2 l_3 \min\left(x_2, s_2\right)\mu_2}{\lambda + \sum_j \min\left(x_j, s_j\right)\mu_j}.$$

If the second node is idle, then event likelihood ratios are the same as before (i.e., without the modification).

4.2.1 Asymptotic Behavior

The proposed modifications reduce the upper bound on likelihood ratios for an entire cycle to $c^{b-1/2}$ because service completions at node 1 are only biased when node 2 is idle. The bound on the absolute deviation between any observation and the sample mean is

$$c^{b-1/2} - d^{b-1} = c^{b-1}\left(c^{1/2} - \left(d/c\right)^{b-1}\right).$$

Taking limits yields the same results obtained for the importance sampling distribution without the modification. However, the rate of convergence to the zero-variance distribution is increased, if $b \to \infty$, $\lambda \to 0$ since the exponent on c increases by $b/2$.

4.3 More Than Two Nodes

Issues raised in the development of the balanced likelihood ratio approach for the two-node case carry over to cases with more nodes. It appears that the bound on λ, required to guarantee arrival event likelihood ratios are smaller than one, generalizes to $\lambda < \min\left(\mu_1, \ldots, \mu_m\right)/m$. This has not been derived analytically, but has been validated numerically for several cases. To generalize the upper bound on arrival

event likelihood ratios, a root to the following equation, between zero and one, can be found numerically

$$\frac{\lambda}{\lambda + (1-l)\min(\mu_1,...,\mu_m)} = l^m.$$

The modified balanced likelihood ratio approach for the two node system can be extended and the asymptotic properties in these cases are the same as the two-node case.

5 NUMERICAL RESULTS

Three test cases are considered. All cases have single servers at each node in the network. The first case satisfies and the other cases violate the necessary condition derived by Glasserman and Kou (1995) for the heuristic solution (Swap) of interchanging the arrival rate with the smallest service rate to be asymptotically efficient. All simulation results are based on 10,000,000 regenerative cycles. Each simulation run provides an estimate for $P(\tau_0 < T_0)$ (the Mean), a 95% confidence interval half-width (Halfwidth), and the corresponding relative error (RE). Computation times (CPU) are displayed in terms of the average number of events per regenerative cycle. Standard Monte Carlo estimates are unavailable at this sample size. Consequently, all VRRs compare balanced likelihood ratio methods to the Swap heuristic. The proposed balanced likelihood ratio method without modification (BLR1) is omitted for brevity and the method with modification (BLR2) is shown. All simulations were implemented in C++ and run on a PC.

The first test case has an arrival rate of 0.18, the service rate at node 1 is 0.42, and the service rate at node 2 is 0.4. Results for b = 25 are displayed in Table 1, and results for b = 100 are displayed in Table 2.

Table 1: $P(\tau_0 < T_0)$ Estimates for Test Case 1 with b = 25

Method	Mean	Halfwidth	RE	CPU	VRR
Swap	3.83e-8	1.88e-9	0.05	72	-
BLR2	3.67e-8	1.84e-9	0.03	103	0.74

Table 2: $P(\tau_0 < T_0)$ Estimates for Test Case 1 with b = 100

Method	Mean	Halfwidth	RE	CPU	VRR
Swap	4.13E-34	2.20E-35	0.03	319	-
BLR2	4.58E-34	1.09E-34	0.12	432	0.03

For test case 1 with b = 25, BLR2 is approximately equivalent to the Swap heuristic in performance. When b = 100, Swap appears to perform much better than BLR2, but this is misleading. The variance estimate from the Swap heuristic is unstable in that it grew by a factor of 10 when the sample size was changed from 1,000,000 to 10,000,000. All BLR cycles are forced to reach system size b which leads to larger CPU times than Swap. The initial likelihood ratio for BLR, c = 0.870, limits the biasing of service events.

The second test case considers a significantly smaller arrival rate than the first. Test case 2 has arrival rate 0.03, node 1 service rate 0.5, and node 2 service rate 0.47. Results for b = 25 are displayed in Table 3, and results for b = 100 are displayed in Table 4.

Table 3: $P(\tau_0 < T_0)$ Estimates for Test Case 2 with b = 25

Method	Mean	Halfwidth	RE	CPU	VRR
Swap	2.46E-28	1.48E-29	0.03	48	-
BLR2	2.57E-28	2.85E-31	0.00	49	2642

Table 4: $P(\tau_0 < T_0)$ Estimates for Test Case 2 with b = 100

Method	Mean	Halfwidth	RE	CPU	VRR
Swap	6.11E-118	4.61e-119	0.04	204	-
BLR2	7.77E-118	2.40e-119	0.03	211	3.57

The relative errors for BLR2 are significantly smaller in case 2 than for case 1. This supports the proposition that BLR methods yield bounded relative error as the arrival rate approaches zero. In addition, VRRs of 2642 and 3.57 indicate that BLR2 is more efficient than the Swap heuristic for this test case. The initial event likelihood ratio for BLR is c = .082.

The relative errors in test cases 1 and 2 grow linearly in the buffer size for BLR2. There does not appear to be any pattern in the relative errors for the Swap heuristic. Throughout all of these experiments, sample size is an issue. The variance estimates from BLR2 are stable, but the variance estimates from the Swap heuristic typically vary greatly within the first 10,000,000 cycles.

The third test case considers an arrival rate between the first two. The arrival rate is 0.12, the service rate at node 1 is 0.42, and the service rate at node 2 is 0.46. Results for b = 25 are displayed in Table 5. BLR2 outperforms the Swap heuristic with a VRR of 2.58. The initial event likelihood ratio for BLR is c = .485.

Table 5: $P(\tau_0 < T_0)$ Estimates for Test Case 3 with b = 25

Method	Mean	Halfwidth	RE	CPU	VRR
Swap	4.84e-13	6.62e-14	0.03	44	-
BLR2	6.44e-13	3.27e-14	0.00	70	2.58

6 CONCLUSIONS

The balanced likelihood ratio approach yields a zero variance importance sampling distribution for a single node system when computing the probability the total system size reaches a bound within a regenerative cycle. Applying the balanced likelihood ratio method to tandem queues is more challenging. The two importance sampling distributions proposed have some nice theoretical properties, but only BLR2 appears to work reasonably well in practice.

Performance of BLR2 appears to be consistent for different buffer sizes and improves as the arrival rate decreases. This improvement should also occur if the number of servers at nodes increased. Numerous issues remain open for the application of the balanced likelihood ratio approach in general queueing systems.

Within the Markovian framework, the proposed importance sampling distributions represent a first step in the development of general balanced likelihood ratio methods. Asymptotic properties including rates of convergence to steady-state need to be more fully understood. Restrictions on model parameters need to be relaxed and variations on the proposed approaches that consistently yield variance reduction ratios larger than one are sought.

REFERENCES

Alexopoulos, C. and B. C. Shultes. 1998. The balanced likelihood ratio method for estimating performance measures of highly reliable systems. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 1479-1486. Piscataway, New Jersey: IEEE.

Alexopoulos, C. and B. C. Shultes. 2001. Estimating reliability measures for highly dependable systems, using balanced likelihood ratios. *IEEE Transactions on Reliability* 50 (3): 265-280.

Bartle, R. G. 1976. *The Elements of Real Analysis*. 2nd edition. New York: John Wiley & Sons.

De Boer, P. T., V. F. Nicola, and R. Y. Rubinstein. 2000. Adaptive importance sampling of queueing networks. In *Proceedings of the 1998 Winter Simulation Conference*, ed. J. A. Joines, R. R. Burton, K. Kang, and P. A. Fishwick, 646-655. Piscataway, New Jersey: IEEE.

Frater, M. R., T. M. Lennon ,and B. D. O. Anderson. 1991. Optimally efficient estimation of statistics of rare events in queueing networks. *IEEE Transactions on Automatic Control* 36 (12): 1395-1405.

Glasserman, P. and S.-G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation* 5 (1): 22-42.

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

Juneja, S. 1993. Efficient rare event simulation of stochastic systems. Ph. D. Thesis, Department of Operations Research, Stanford University, Palo Alto, California.

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

Kroese, D. P. and V. F. Nicola. 1999. Efficient simulation of a tandem network. In *Proceedings of the 1998 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 411-419. Piscataway, New Jersey: IEEE.

Kuruganti, I. and S. Strickland. 1997. Optimal importance sampling for Markovian systems with applications to tandem queues. *Mathematics and Computers in Simulation* 44: 61-79.

Law, A. and W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd edition. New York: McGraw-Hill.

Lieber, D., A. Nemirovskii, and R. Y. Rubinstein. 1999. A fast Monte Carlo method for evaluating reliability indices. *IEEE Transactions on Reliability* 48 (3): 256-261.

Parekh, S. and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34 (1): 54-66.

Rubinstein, R. Y. 1997. Optimization of computer simulation models with rare events. *European Journal of Operational Research* 99: 89-112.

Sadowski, J. S. 1991. Large deviation theory and efficient simulation of excessive backlogs in a GI/GI/m queue. *IEEE Transactions on Automatic Control* 36 (12): 1383-1394.

Serfozo, R. 1999. *Introduction to Stochastic Networks*. New York: Springer.

Tsoucas, P. 1992. Rare events in series of queues. *Journal of Applied Probability* 29 (1): 168-175.

AUTHOR BIOGRAPHY

BRUCE C. SHULTES is an Assistant Professor of Industrial Engineering at University of Cincinnati. Previously, he spent two years at the Naval Postgraduate School as an Assistant Research Professor and a National Research Council Research Associate. He received his B.S. in Applied Mathematics from Carnegie Mellon University, his M.S. in Management Science from Case Western Reserve University, and his Ph.D. in Industrial Engineering (Stochastic Systems) from Georgia Institute of Technology. His research interests are in the areas of applied probability, simulation, analysis of stochastic systems, algorithms, and software engineering. His recent research efforts include efficient algorithms for: analyzing rare events, estimating ANSI form tolerances, and quickly assessing the scheduling feasibility of tasks. He is a member of IIE, INFORMS, and the INFORMS College on Simulation. His email address is <Bruce.Shultes@uc.edu>.