Statistical Science
2000, Vol. 15, No. 3, 279–298

Reliabilities for Feedback Systems and Their Saddlepoint Approximation

Ronald W. Butler

Dedicated to the memory of Professor Henry Daniels

Abstract. It is shown how saddlepoint methods may be used to approximate reliabilities and failure rates of finite stochastic systems with feedback loops. Some countably infinite state systems including birth–death processes are also considered. The use of saddlepoint methods requires as input the moment generating functions (MGFs) for the system failure time distributions. Some new explicit formulas for these MGFs are given that are amenable to symbolic computation and which also make the numerical computation of saddlepoint approximations quite simple and convenient.

Key words and phrases: Failure rate, feedback, flowgraphs, reliability, saddlepoint approximation, semi-Markov systems

1. INTRODUCTION

The *reliability of a system* at time t is the probability the system is working properly, where the meaning of properly depends on the particular application. If, as often occurs in practical settings, the system has a finite or countable number of states, then they can be partitioned into two groups: those for which the system works and those for which it does not. The waiting time for system failure is the random variable W defined as the time for first passage of the system into the collection of nonworking states. Its distribution determines the reliability of the system as

$$R(t) = \Pr(W > t). \tag{1}$$

Exact computation of (1) is generally possible only for small finite systems when the system is Markov and not otherwise (see Høyland and Rausand, 1994). A much wider range of examples can be considered if the exact calculation of $R(t)$ is replaced with highly accurate approximation, as occurs when using saddlepoint methods.

There are two major purposes for this paper. First, it considers the saddlepoint approximation to $R(t)$ using the Lugannani and Rice (1980) approximation and a related approximation in Wood, Booth and Butler (1993) based upon the moment generating function (MGF) of W. The accuracy with which $R(t)$ is determined in the examples below suggests that saddlepoint approximations are highly suited for use in this particular class of applications. The reader may judge this accuracy, without sorting through all the details of the methods and examples, by examining Figures $1, 5, 7, 9$ and 16 which compare $R(t)$ with the two saddlepoint approximations, denoted as $\hat{R}(t)$ and $\breve{R}(t)$, respectively. Associated failure rates $z(t)$ and various saddlepoint approximations for them, denoted by $\hat{z}(t)$, $\tilde{z}(t)$ and $\breve{z}(t)$, are also presented in Figures $2, 6, 8, 10$ and 17. The percentage relative errors in determining $R(t)$ and $z(t)$ are given whenever possible as in Figures $3, 11$ and 18.

The second major purpose of the paper is to introduce some new simple expressions for the MGF of W. Generally the MGF of W has been very difficult to compute, particularly so when the system has feedback loops. Now, however, there is no difficulty at all if the new explicit *cofactor rules* of Section 4 are used. These cofactor formulas give the MGF of W in the context of a finite semi-Markov system. There are three separate rules which concern: (1) first passage from one state to another, (2) first return to a starting state and (3) first passage from a state to a subset of states. Each MGF expression is explicit in terms of certain cofactors and all the expressions lead to explicit saddlepoint formulas which make subsequent saddlepoint computations very simple.

Ronald W. Butler is Professor, Department of Statistics, Colorado State University, Ft. Collins, Colorado 80523 (e-mail: walrus@stat.colostate.edu).

As examples, we first introduce a $GI/M/1$ queue with Gamma interarrivals in Section 2. This is a simple system for which saddlepoint methods provide highly accurate approximations to reliability and failure rate functions. We return to this example in Section 6 to develop all its details connected with MGF determination and saddlepoint implementation and to further consider other interarrival distributions such as compound Poisson and inverse Gaussian. Section 7 considers the reliability of a redundant and repairable system (see Høyland and Rausand, 1994, Sections 4.6 and 4.7 and Chapter 6) that is semi-Markov. With the inclusion of additional system states, the system becomes Markov which allows us to compute exact reliabilities for comparison with the saddlepoint approximations. The determination of reliabilities and failure rates of finite Markov systems is reviewed in Section 8.

The concept of modular systems in which a system may be considered as a network of subsystems is introduced in Section 10. Modular usage of the cofactor rules leads to the development of recursive formulas for the MGF of W in random walks and heterogeneous birth–death processes. Practically, these recursions now allow for very simple saddlepoint approximation to the first passage and return distributions in heterogeneous birth–death processes. A load sharing queue and a null persistent random walk provide examples of such systems. The random walk example demonstrates the breakdown in accuracy of the Lugannani and Rice (1980) approximation when used with heavy-tailed distributions; the inverse Gaussian-based approximation of Wood, Booth and Butler (1993) maintains its high accuracy in this setting. The net consequence of these recursions and the cofactor rules is that reliabilities may now be approximated in a much broader class of stochastic feedback systems than has been previously considered.

A rather simple and insightful presentation of the feedback theory underlying the cofactor rules is presented in Section 9. These cofactor rules have been previously derived in Butler (1997a, 1997b) using more difficult arguments. In the latter article, they were shown to be analytically equivalent to the more complicated *Pyke–Howard rule,* developed in Pyke (1961) and Howard (1964, 1971, Sections 10.10, 11.11), as well as *Mason's loop sum formula,* developed in Mason (1953, 1956) and discussed, for example, in Whitehouse (1983) and Phillips and Harbor (1996). The work of both Mason and Howard finds its origin in the cybernetics movement at MIT that dealt explicitly with feedback systems about which we are particularly concerned.

In determining first passage distributions, saddlepoint approximations have previously been used in conjunction with Mason's loop sum formula in Butler and Huzurbazar (1993, 1997) and in conjunction with the Pyke–Howard rule in Butler and Huzurbazar (1995). The cofactor rules however are clearly simpler and easier to use than either of these previous formulas. The main difficulty with the Pyke–Howard rule is that (1) it involves $O(n)$ more computational effort where n is the number of system states, and (2) the computation of the MGF can be inaccurate and unstable, particularly near zero where it has a removable singularity. The difficulty in using Mason's loop sum formula is that it requires a listing of all the system feedback loops over which it sums. Human error in compiling such a list is a serious drawback to its use with large complicated systems. Phillips and Harbor (1996) point out that Mason's formula

> . . . must be used with extreme care, since (feedback loop) terms in either the numerator or denominator of the transfer (MGF) function can *easily* be overlooked. Furthermore, there is no method available that will give an indication in the case that terms have been overlooked.

Now, an available method for finding these overlooked terms is machine computation using the new cofactor rules which are without error. Also, in response to this same problem, Zhou, Wang and Zhao (1995) have written specialized software to automate the listing of all system feedback loops. Their complicated software, however, is no longer needed since its cofactor equivalent in (4) automatically sums over all feedback loops in its determination of the cofactors; essentially each determinant may be expressed as a permutation sum which can be shown equivalent to summing over all feedback loops. See Butler (1997b) for details of this development.

2. EXAMPLE: *GI/M/1* QUEUE

Suppose tasks arrive one at a time at a server, and that the interarrival times are distributed as independent and identically distributed (i.i.d.) Gamma(2, 2) with mean 1 and variance 1/2. Suppose a single server completes tasks at rate 2 with service distribution Exponential(2). The state of the system is the total number in queue and under service by the server. Suppose that the system is deemed to have failed when the queue length reaches 5, so that the system failure time W is the first passage time to queue length 5. The distribution of W is quite complicated but completely determined by our description of the process.

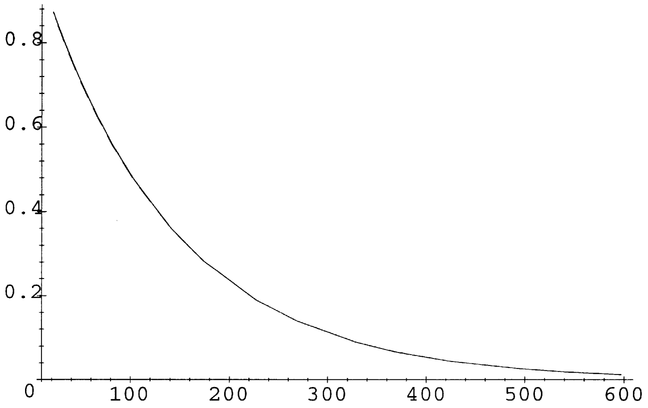

FIG. 1. *The almost indistinguishable plots of $\hat{R}(t)$ (dashed) and $R(t)$ (solid) versus t for the GI/M/1 queue with Gamma(2, 2) interarrivals.*

The complete development of this example is given in Section 6 where the exact CDF and density of W are also computed. This provides a baseline for comparison of the exact expression for $R(t)$, plotted in Figure 1 as a solid line, with the saddlepoint approximation $\hat{R}(t)$ (dashed), described in Section 5. The plot reveals hardly any graphical difference. True failure rate $z(t)$ (solid) and two saddlepoint approximations $\hat{z}(t)$ (dashed, unnormalized; see Section 5) and $\tilde{z}(t)$ (dotted; normalized) are plotted in Figure 2 to demonstrate the high accuracy of $\tilde{z}(t)$. The asymptote for $z(t)$ has $z(\infty) = 0.00741$. Figure 3 shows the relative errors (expressed as percentages) connected with the two previous graphs and plots. That is,

$$100\left(\frac{\hat{R}(t)}{R(t)} - 1\right)\% \quad \text{and} \quad 100\left(\frac{\tilde{z}(t)}{z(t)} - 1\right)\% \text{ versus } t$$

are plotted as the solid and dashed lines in Figure 3. In addition, the relative error of the inverse Gaussian-based approximation $\breve{R}(t)$ (dotted) from Wood, Booth and Butler (1993) is given.

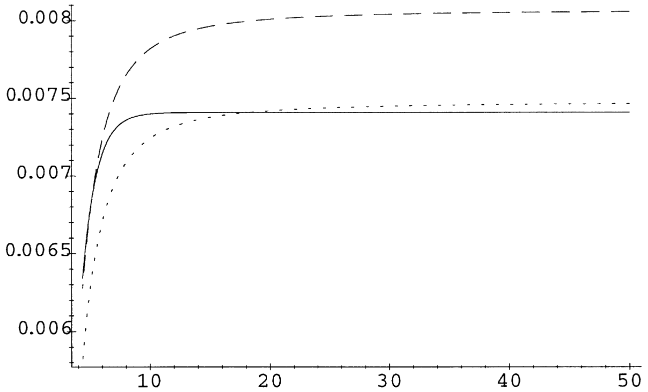

FIG. 2. *Plot of $z(t)$ (solid), $\hat{z}(t)$ (dashed) and $\tilde{z}(t)$ (dotted) versus t for the GI/M/1 queue with Gamma(2, 2) interarrivals.*

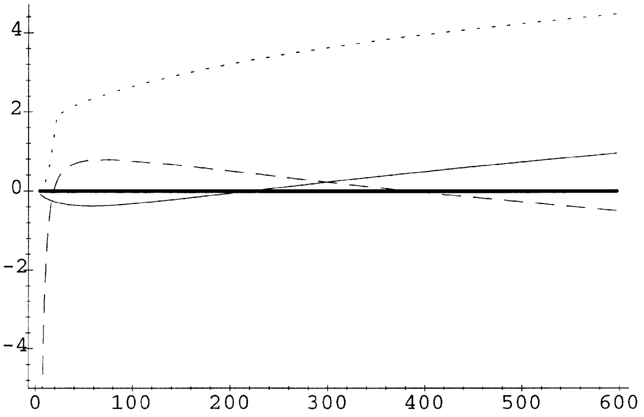

FIG. 3. *Percentage relative errors in saddlepoint approximation for $R(t)$ (solid = normal base, dotted = inverse Gaussian base) and $z(t)$ (dashed) for the GI/M/1 queue with Gamma(2, 2) interarrivals.*

3. SYSTEM PRELIMINARIES

3.1 Coin-Tossing Example

Consider how we might characterize or summarize the dynamic behavior of a finite state stochastic system with feedback. For purposes of illustration, let us also work with a specific and simple system resulting from a sequence of i.i.d. coin tosses or Bernoulli(p) trials with $p = \Pr(\text{heads})$. Suppose we are interested in the current run length of heads and, in particular, the time required to achieve the first run of three heads. Our *system* is a sequence of independent coin tosses and the system characteristic of interest is the run length of heads whose values $\{0, \ldots, 3\}$ represent the states of the system. The dynamic change of the system in state i is to either move to state $i + 1$ with probability p in time a or to move to state 0 with probability $q = 1 - p$ in time b. This dynamic behavior can be summarized in two equivalent ways: either pictorially or algebraically. Figure 4 is a *flowgraph* giving the pictorial description of dynamic change. The nodes in the flowgraph are the states and the directed branches indicate the possible state changes of the system. The system begins in *source* node B, which has only an outgoing branch, and ends in *sink* or destination node E, having only an incoming branch. *Feedback* occurs in this system because states 0, 1 and 2 may be revisited as the process evolves. The system has three *feedback loops:* $0 \to 0$, $0 \to 1 \to 0$ and $0 \to 1 \to 2 \to 0$. Each branch has associated with it a quantity we call the *branch transmittance.* This is the probability of taking the branch times the MGF for the holding time in its node of origin given that it takes that branch. The rationale for defining transmittance in such a manner becomes clear in the discussion below. The transmittances of transitions $B \to 0$ and $3 \to E$ are $1 = 1 \times e^0$ indicating that passage is certain and instantaneous. These branches can be removed if we assume that the system begins in node 0 and ends in node 3. All information about the dynamic behavior of this system appears in its flowgraph structure and branch transmittances.

An equivalent algebraic summarization of Figure 4 is specified in the 4×4 matrix of branch transmittances

$$\mathscr{Q}(s) = \{\mathscr{Q}_{ij}(s)\} = \begin{pmatrix} qe^{bs} & pe^{as} & 0 & 0 \\ qe^{bs} & 0 & pe^{as} & 0 \\ qe^{bs} & 0 & 0 & pe^{as} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2}$$

when accompanied by the information that 0 and 3 are the source and sink nodes. Possible transitions occur with nonzero entries in $\mathscr{Q}$ and the branch transmittances are listed in the appropriate cells. The algebraic characterization of the system in terms of $\mathscr{Q}(s)$ will be the basis for our MGF and saddlepoint computations in Sections 4 and 5.

3.2 *n*-state semi-Markov Process

We now generalize the discussion to a general n-state stochastic feedback system with state space $S = \{1, \ldots, n\}$. In this setting, flowgraphs become difficult to draw, particularly when the system has many states and/or many branch transmittances. It is therefore expedient to characterize the stochastic system in terms of the matrix of one-step branch transmittances. Whether our process is in discrete or continuous time, let the n-state system admit transitions according to the probability transition matrix (p_{ij}) where $0 \le \sum_j p_{ij} \le 1$ for all i. Suppose also that the holding time in state i is dependent not only upon i but also upon the next destination state, j, say. Then, given that passage from state i to j is assured, let the holding time in state i have MGF $\mathscr{Q}_{ij}(s)$. The $n \times n$ matrix of branch transmittances is defined as $\mathscr{Q}(s) = \{p_{ij}\mathscr{Q}_{ij}(s)\colon i, j \in S\}$. The matrix function $\mathscr{Q}(s)$ characterizes the dynamic behavior of the particular stochastic system. In addition, the collection of all systems or stochastic processes which may be so characterized is referred to as the class of n-state *semi-Markov* processes and is discussed in Pyke (1961). Our treatment below considers only nonexplosive processes; that is, processes that cannot achieve an infinite number of state changes in finite time.

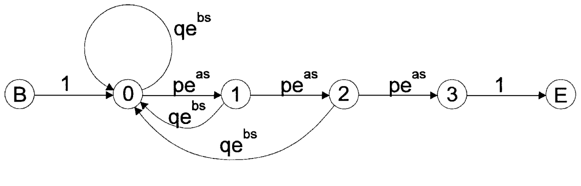

FIG. 4. *Flowgraph of the waiting time for the first occurrence of three straight heads.*

4. COFACTOR RULES

Three cofactor rules yielding first passage MGFs are introduced. The proofs are deferred to Section 9 where they are developed using the flowgraphs of their associated feedback systems.

4.1 Single Destination Cofactor Rules

First passage probabilities and MGFs are specified in the theorems below for the general n-state semi-Markov process. Without any loss in generality we take state 1 as the source and state n as the destination state.

Define the *first passage transmittance* from state 1 to state n as

$$f_{1n}\mathscr{F}_{1n}(s) = E\left\{\exp(sW_{1n})1_{(W_{1n}<\infty)}\right\}, \tag{3}$$

where W_{1n} is the first-passage time from state 1 to state n. According to this, $f_{1n} = \Pr(W_{1n} < \infty)$ is the probability of passage and $\mathscr{F}_{1n}(s)$ is the conditional MGF of W_{1n} given $W_{1n} < \infty$, or given that such passage is assured. When $f_{1n} < 1$, the distribution of W_{1n} is defective with $\Pr(W_{1n} = \infty) = 1 - f_{1n}$. The maximal convergence neighborhood for $\mathscr{F}_{1n}(s)$, about $s = 0$, is defined as the largest connected neighborhood of 0 for which the expectation in (3) is finite.

Define state i as *relevant* to first passage from state 1 to state n if it is a possible intermediate state during such passage. Designate states 1 and n as relevant if passage $1 \to n$ is possible.

THEOREM 1 (Single destination cofactor rule). *The first passage transmittance from state* 1 *to state* $n \neq 1$ *is*

$$\begin{aligned} f_{1n}\mathscr{F}_{1n}(s) &= \frac{(n,1)\text{-cofactor of } \{I_n - \mathscr{Q}(s)\}}{(n,n)\text{-cofactor of } \{I_n - \mathscr{Q}(s)\}} \\ &:= \frac{(-1)^{n+1}\left|\Psi_{n1}(s)\right|}{\left|\Psi_{nn}(s)\right|}. \end{aligned} \tag{4}$$

The ratio (4) *is well defined over an maximal convergence neighborhood of* 0 *of the form* $(-\infty, c)$ *for some* $c > 0$ *under these conditions:*

(i) *The system states* $S = \{1, \ldots, n\}$ *are exactly those relevant to passage from* $1 \to n$ *with all relevant states and no irrelevant states.*

(ii) *The intersection of the maximal convergence neighborhoods for the MGFs in the first* $n-1$ *rows of* $\mathscr{Q}(s)$ *is an open neighborhood of* 0.

For the coin-tossing example, expression (4) affords a simple computation of the first passage transmittance from state 0 to 3. Since $f_{03} = 1$, this transmittance becomes the MGF and is

$$\mathscr{F}_{03}(s) = p^3 e^{3as}\left\{1 - qe^{bs} - pqe^{(a+b)s} - p^2 qe^{(2a+b)s}\right\}^{-1}. \tag{5}$$

With $a = 1 = b$, (5) is well known and essentially given in Feller [1968, XIII.7, (7.6)] as a probability generating function. However, compare the different difficulties in derivation. Feller's development requires a good deal of understanding and thought within the context of his development of recurrent events. The derivation from (4) is an elementary and rather thoughtless exercise in symbolic computation using Maple V within which the possibility for mistake is minimal. From (5), a saddlepoint approximation for the first passage distribution function is a routine exercise.

We now consider the case of first return to state 1, a situation that has been excluded from consideration in the theorem above.

THEOREM 2 (First return cofactor rule). *The first return transmittance for state* 1 *is*

$$\begin{aligned} f_{11}\mathscr{F}_{11}(s) &= 1 - \frac{\left|I_n - \mathscr{Q}(s)\right|}{(1,1)\text{-cofactor of } \{I_n - \mathscr{Q}(s)\}} \\ &:= 1 - \frac{\left|I_n - \mathscr{Q}(s)\right|}{\left|\Psi_{11}(s)\right|}. \end{aligned} \tag{6}$$

The ratio (6) *is well defined over a maximal convergence neighborhood of* 0 *of the form* $(-\infty, c)$ *for some* $c > 0$ *under these conditions:*

(i) *The system states* $S = \{1, \ldots, n\}$ *are exactly those relevant to passage from* $1 \to 1$.

(ii) *The intersection of the maximal convergence neighborhoods for the MGFs in* $\mathscr{Q}(s)$ *is an open neighborhood of* 0.

The coin-tossing example provides a simple context for illustrating first return to state 0. State 3, achieved with three straight heads, is absorbing and therefore irrelevant to passage from $0 \to 0$ so it must be excluded when determining $\mathscr{Q}(s)$ which now becomes 3×3. The first return transmittance is computed as

$$f_{00}\mathscr{F}_{00}(s) = qe^{bs} + qe^{bs}pe^{as} + qe^{bs}(pe^{as})^2$$

using Maple V. The first return probability is computed by setting $s = 0$ and simplifies to $f_{00} = 1 - p^3$. This answer might have been expected because its complementary event is direct passage to state 1 with three straight heads with probability p^3.

4.2 Multiple Destination Cofactor Rule.

Often systems may fail upon entering any one of several states. In such instances, the first passage distribution to this collection of failure states determines the reliability of the system. In general, suppose $D = \{m+1, \ldots, n\}$ is the subset of failure states into which the system may pass. The first passage transmittance from $1 \in \{1, \ldots, m\} = C$ to D is not generally $\sum_{j \in D} f_{1j}\mathscr{F}_{1j}(s)$, with $f_{1j}\mathscr{F}_{1j}(s)$ determined as above. This is because the summing of events here is not over disjoint paths; for example, first passage to state n might pass through state $m+1$ beforehand and such possibilities are not accounted for in this sum. The next result provides a simple expression for the first passage transmittance from $1 \to D$. Denote the block form of the system transmittance matrix as

$$\mathscr{Q}(s) = \begin{pmatrix} \mathscr{Q}_{CC}(s) & \mathscr{Q}_{CD}(s) \\ \mathscr{Q}_{DC}(s) & \mathscr{Q}_{DD}(s) \end{pmatrix},$$

where $\mathscr{Q}_{CC}$ is $m \times m$ taking C into C, $\mathscr{Q}_{DD}$ is $(n-m) \times (n-m)$ taking D into D, etc. Denote the row sums of $\mathscr{Q}_{CD}(s)$ as

$$\mathscr{Q}_{CD}(s)1 = \mathscr{Q}_{C\cdot}(s) = (\mathscr{Q}_{1\cdot} \cdots \mathscr{Q}_{m\cdot})^T,$$

and let $\{I_m - \mathscr{Q}_{CC}(s)\}_{\backslash 1}$ denote the $m \times (m-1)$ matrix $I_m - \mathscr{Q}_{CC}(s)$ with its first column removed.

THEOREM 3. *The first passage transmittance from state* $1 \in C = \{1, \ldots, m\}$ *to subset* $D = \{m+1, \ldots, n\}$ *in an n-state system is*

$$f_{1D}\mathscr{F}_{1D}(s) = \frac{\left|\mathscr{Q}_{C\cdot}(s)\ \ \{I_m - \mathscr{Q}_{CC}(s)\}_{\backslash 1}\right|}{\left|I_m - \mathscr{Q}_{CC}(s)\right|}, \tag{7}$$

under the conditions specified below, where the numerator matrix is $I_m - \mathscr{Q}_{CC}(s)$ *with its first column replaced with* $\mathscr{Q}_{C\cdot}(s)$. *Expression* (7) *is well defined over an maximal convergence neighborhood of* 0 *of the form* $(-\infty, c)$ *for some* $c > 0$ *under these conditions:*

(i) *The system states* $\{1, \ldots, n\}$ *are exactly those relevant to passage from* $1 \to D$.

(ii) *The intersection of the maximal convergence neighborhoods for the MGFs in the first* m *rows of* $\mathscr{Q}(s)$ *is an open neighborhood of* 0.

The coin-tossing example provides simple verification for the validity of this expression. Take $D = \{2, 3\}$ and compute the first passage transmittance from $0 \to D$ as

$$f_{0D}\mathscr{F}_{0D}(s) =$$

$$\frac{\begin{vmatrix} 0 & -pe^{as} \\ pe^{as} & 1 \end{vmatrix}}{\begin{vmatrix} 1-qe^{bs} & -pe^{as} \\ -qe^{bs} & 1 \end{vmatrix}} = \frac{p^2e^{2as}}{1-qe^{bs}-pqe^{(a+b)s}}. \tag{8}$$

Since passage to state 3 is through 2, this transmittance must also be the first passage transmittance to state 2 and is easily shown to agree with $f_{02}\mathscr{F}_{02}(s)$ when computed from (4).

5. SADDLEPOINT APPROXIMATIONS

The previous section has provided simple formulas for the computation of first-passage MGFs. What remains is to discuss the determination of approximate distribution functions and reliabilities from these MGFs. Suppose W is a first-passage time whose transmittance $f\mathscr{F}(s)$ has been determined using one of the cofactor rules. It is only necessary to approximate the distribution of the finite portion of W since it is known to put mass $1-f$ at infinity. In general, the reliability function is

$$R(t) = \Pr(W > t) = \Pr(W > t | W < \infty) + (1-f),$$

and a saddlepoint approximation will be used to approximate the first term. Since all our numerical examples have $f = 1$, we shall suppress the second term and assume it is 0.

We may start from either an explicit formula for $\mathscr{F}(s)$, determined from symbolic computation in Maple V, or use the cofactor rules to numerically compute $\mathscr{F}(s)$ in terms of the matrix $\mathscr{D}(s)$. Saddlepoint approximation to $R(t)$ using the Lugannani and Rice (1980) approximation discussed by Daniels (1987) starts from $K(s) = \ln \mathscr{F}(s)$, the cumulant generating function (CGF) of W defined on $(-\infty, c)$ for some $c > 0$. The approximation for $R(t)$ requires that we first find the saddlepoint $\hat{s} = \hat{s}(t)$ as the unique solution to the saddlepoint equation

$$K'(\hat{s}) = t \tag{9}$$

in $(-\infty, c)$. Based upon this, then

$$\hat{R}(t) = 1 - \Phi(\hat{w}) - \phi(\hat{w})(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}), \quad t \neq E(W) = K'(0), \tag{10}$$

where Φ and ϕ are the standard normal distribution and density, and $\hat{w}$ and $\hat{u}$ depend on t implicitly as

$$\hat{w} = \operatorname{sgn}(\hat{s})\sqrt{2\{\hat{s}t - K(\hat{s})\}} \quad \text{and} \quad \hat{u} = \hat{s}\sqrt{K''(\hat{s})}. \tag{11}$$

In systems with few states, and therefore small $\mathscr{D}$, there may be some advantage in computing K symbolically from the cofactor rules. Then symbolic differentiation of K gives an expression for K' useful in solving for the saddlepoint $\hat{s}$, and further differentiation gives K'' to compute $\hat{u}$.

An alternative approach in numerical computation of $\hat{R}(t)$ is to specify K' and K'' in terms of the $\mathscr{D}$ matrix. For example, in first passage from $1 \to n$, differentiation of

$$K(s) = \ln \frac{(-1)^{n+1}\left|\Psi_{n1}(s)\right|}{\left|\Psi_{nn}(s)\right|} \tag{12}$$

gives

$$K' = \operatorname{tr}\left(\Psi_{n1}^{-1}\dot{\Psi}_{n1} - \Psi_{nn}^{-1}\dot{\Psi}_{nn}\right) \tag{13}$$

with the dependencies on s suppressed and $\dot{\Psi}_{n1} := d\Psi_{n1}(s)/ds$. Furthermore,

$$K'' = \operatorname{tr}\{\Psi_{n1}^{-1}\ddot{\Psi}_{n1} - \left(\Psi_{n1}^{-1}\dot{\Psi}_{n1}\right)^2 - \Psi_{nn}^{-1}\ddot{\Psi}_{nn} + \left(\Psi_{nn}^{-1}\dot{\Psi}_{nn}\right)^2\}, \tag{14}$$

where $\ddot{\Psi}_{n1} := d^2\Psi_{n1}(s)/ds^2$. These computations are quite simple and have been successfully used by the author in systems with up to $n = 250$ states. The computation of reliabilities and first passage distributions in many difficult settings should now become routine when using these expressions.

Failure or hazard rate approximation requires an additional estimate for the density function. The saddlepoint approximation for the density of W, or $f(t) = -R'(t)$, is

$$\hat{f}(t) = \frac{1}{\sqrt{2\pi K''(\hat{s})}} \exp\left(-\frac{1}{2}\hat{w}^2\right), \tag{15}$$

as given by Daniels (1954). Combining this with $\hat{R}(t)$ provides two approximations for the failure or hazard rate $z(t) = f(t)/R(t)$:

$$\hat{z}(t) = \frac{\hat{f}(t)}{\hat{R}(t)} \quad \text{and} \quad \tilde{z}(t) = \frac{\hat{f}(t)}{\hat{R}(t)\int_0^\infty \hat{f}(u)\,du}, \tag{16}$$

the unnormalized and normalized approximations. Normalization of $\hat{f}$ will be crucial in achieving accurate approximation to $z(t)$.

The Lugannani–Rice approximation in (10) does not always succeed in accurately approximating CDFs of first passage distributions. This was first discussed in Wood, Booth and Butler (1993) where the normal-based expression in (10) was unable to accurately approximate the first return distribution of a simple random walk near to the null persistent setting ($p \approx 1/2$). Further inaccuracies of (10) for passage time distributions are discussed in Booth and Wood (1995), Booth (1994) and Butler and Huzurbazar (1995). In its place, Wood, Booth and Butler (1993) recommend an inverse Gaussian-based approximation which does not suffer such inaccuracy, at least as concerns the examples presented in the papers above. A description of this approximation is given in Appendix A. We denote the reliability approximation based on this method as $\breve{R}(t)$ and the resulting hybrid failure rate approximation as

$$\breve{z}(t) = \frac{\hat{f}(t)}{\breve{R}(t)\int_0^\infty \hat{f}(u)\,du}. \tag{17}$$

6. *GI*/*M*/1 QUEUE

Reconsider the more general structure of this introductory queue with interarrival times distributed i.i.d. with CDF G, density g and MGF $\mathscr{U}_0(s)$. A single server completes tasks at rate μ with service distribution Exponential(μ). The state of the system is the total number in queue and under service. Additional tasks that result in a queue length larger than n are turned away and the system, at that point, is said to have failed. Failure occurs when a new task arrives just after passage into state n but before the current task's service is completed, putting the system into state E, the state of failure. The system has $n+2$ states labelled $i = 0, \ldots, n$ and E. With a queue limit of $n = 4$, the transmittance matrix is

$$\mathscr{Q}(s) = \begin{pmatrix} 0 & \mathscr{U}_0(s) & 0 & 0 & 0 & 0 \\ 0 & \mathscr{Q}_{11}(s) & \mathscr{U}_0(s-\mu) & 0 & 0 & 0 \\ 0 & \mathscr{Q}_{21}(s) & \mathscr{U}_1(s-\mu) & \mathscr{U}_0(s-\mu) & 0 & 0 \\ 0 & \mathscr{Q}_{31}(s) & \mathscr{U}_2(s-\mu) & \mathscr{U}_1(s-\mu) & \mathscr{U}_0(s-\mu) & 0 \\ 0 & \mathscr{Q}_{41}(s) & \mathscr{U}_3(s-\mu) & \mathscr{U}_2(s-\mu) & \mathscr{U}_1(s-\mu) & \mathscr{U}_0(s-\mu) \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where the entries are

$$\begin{aligned} \mathscr{U}_k(s) &= \frac{\mu^k}{k!}\int_0^\infty w^k e^{sw}\,dG(w), \\ &\qquad\qquad k = 0, \ldots, n-1, \\ \mathscr{Q}_{i1}(s) &= \mathscr{U}_0(s) - \sum_{k=0}^{i-1} \mathscr{U}_k(s-\mu), \\ &\qquad\qquad i = 1, \ldots, n. \end{aligned} \tag{18}$$

Derivations of these values are in Appendix B. Consider the system of (18) with $\mu = 2$ and $n = 4$ so that the system failure time W is the first passage to a queue length of 5.

6.1 Gamma($\alpha = 2$, $\beta = 2$) Interarrivals Revisited

The transmittances in (18) are based upon the expression

$$\mathscr{U}_k(s) = \frac{\mu^k \beta^\alpha}{(\beta - s)^{k+\alpha}}\binom{k+\alpha-1}{k}, \quad k = 0, \ldots, 3.$$

For integer values of α these are rational expressions which, through the cofactor rules, determine first-passage MGFs as rational functions. Direct inversion of such rational MGFs is therefore possible by partial fraction expansion and leads to the exact calculation of $R(t)$ and $z(t)$ along with their saddlepoint approximations shown in Figures 1–3.

As determined from (4), $\mathscr{F}_{0E}(s)$ is the (multiplicative) inverse of an order 10 polynomial. Its ten roots are simple poles of $\mathscr{F}_{0E}(s)$ which, in this case, are the distinct values $\nu_1, \ldots, \nu_{10}$ consisting of four real roots and three complex conjugate pairs. Its partial fraction expansion yields $\mathscr{F}_{0E}(s) = \sum_{i=1}^{10} c_i/(\nu_i - s)$, and the density of W from direct inversion is

$$\begin{aligned} f_W(t) &= \sum_{i=1}^{10} c_i \exp(-\nu_i t) \\ &= 7.57 \times 10^{-3} e^{-7.41\times 10^{-3} t} - 4.41 \\ &\quad \times 10^{-2} e^{-0.857t} + 0.240 e^{-2.20t} \\ &\quad -0.624 e^{-3.39t} \\ &\quad +0.468 e^{-4.05t}\cos(1.04t) \\ &\quad +0.110 e^{-4.05t}\sin(1.04t) \\ &\quad -0.0544 e^{-5.05t}\cos(2.16t) \\ &\quad -0.0446 e^{-5.05t}\sin(2.16t) \\ &\quad +6.82 \times 10^{-3} e^{-5.67t}\cos(2.98t) \\ &\quad +6.49 \times 10^{-3} e^{-5.67t}\sin(2.98t). \end{aligned} \tag{19}$$

The true $R(t)$ has been computed using symbolic integration. The mean and standard deviation of this distribution, as determined from $\mathscr{F}_{0E}$, are 138 and 135. An important characteristic of the system is its long term failure rate defined as the value of $z(t)$ as $t \to \infty$. From (19), this value is $z(\infty) = 0.00741$ and may also be seen as the asymptote for the plot of $z(t)$ in Figure 2. It is also the right edge of the convergence strip for $\mathscr{F}_{0E}(s)$, a result that holds quite generally as shown in Butler and Bronson (2000). In this same figure, the accuracy of $\tilde{z}(t)$ shows the need to normalize the saddlepoint density when approximating $z(t)$.

6.2 Compound Poisson Interarrivals

Suppose the interarrival distribution of failures is $Y = \sum_{i=1}^N X_i$ where $X_1, X_2, \ldots$ are i.i.d. Exponential(λ) and N is Poisson(β) with $\beta = 1$ and restricted so $N \geq 1$; such restriction prevents Y from having a point mass at 0. Lengthy calculations show in this setting that

$$\mathscr{U}_k(s) = \frac{\mu^k e^{-\beta}\beta}{(1 - e^{-\beta})\lambda^k}(1 - s/\lambda)^{-(k+1)} \times$$
$${}_1F_1\big(k+1, 2; \beta(1 - s/\lambda)^{-1}\big), \quad s < \lambda,\ k = 0, \ldots, 3,$$

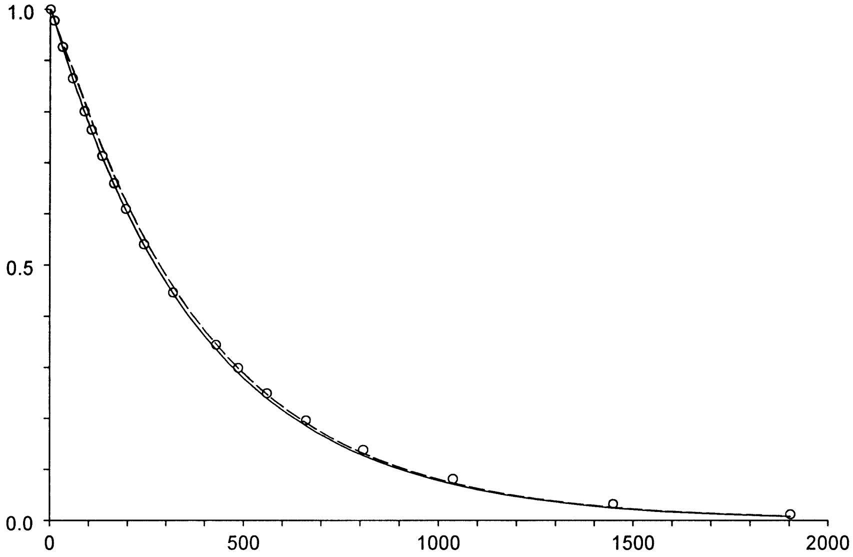

FIG. 5. *Plot of $\hat{R}(t)$ (solid), inverse Gaussian-based $\breve{R}(t)$ (dashed) and an empirical estimate of $R(t)$ (circles) versus t for the $GI/M/1$ queue with compound Poisson interarrivals.*

where ${}_1F_1$ is the confluent hypergeometric function. This expression reduces to simple forms for integer $k = 0, \dots, 3$ using the Kummer transform (13.1.27) and recurrence relation (13.4.4) of Abramowitz and Stegun (1970). The resulting failure distribution has mean 392 and standard deviation 390. The denominator of the MGF, as determined from cofactor rule (4), has its smallest positive real root as 2.56×10^{-3} which is the upper edge of its convergence strip.

Figure 5 plots the normal-based $\hat{R}(t)$ (solid) and inverse Gaussian-based $\breve{R}(t)$ (dashed) reliability approximations against simulated approximations (circle centers) using 10^6 failure runs through the system. Figure 6 compares the unnormalized $\hat{z}(t)$ (dashed) and normalized $\tilde{z}(t)$ (solid) failure rate approximations with simulated approximations (circles) using the 10^6 failure runs along with kernel density approximation for the unknown density. The normalized saddlepoint approximation agrees to high accuracy with the simulation results taking only several minutes of computing time as opposed to the several hours required for the simulation. It furthermore does not suffer from the inherent roughness problems of kernel density estimation that exist even with samples of size 10^6. The nor-

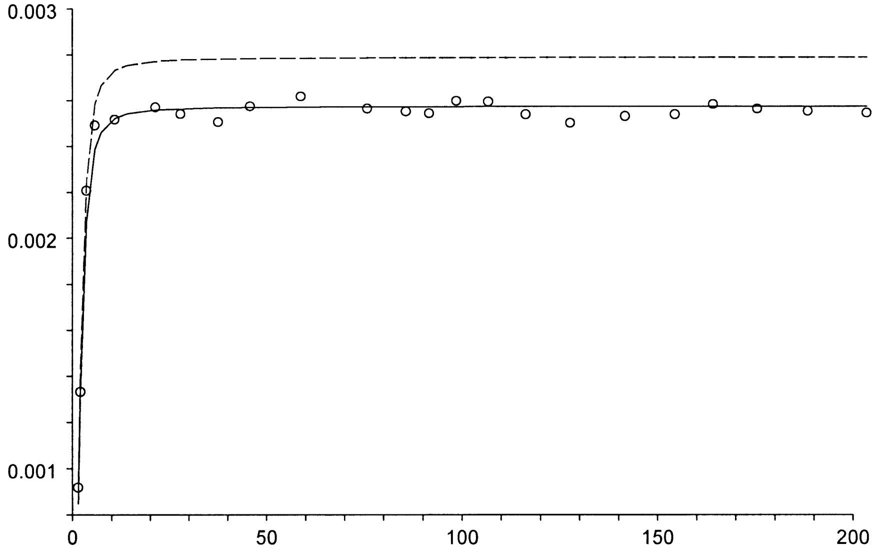

FIG. 6. *Plot of $\hat{z}(t)$ (dashed), $\tilde{z}(t)$ (solid) and an empirical approximation (circles) versus t for the $GI/M/1$ queue with compound Poisson interarrivals.*

malization constant of the saddlepoint density is 1.083.

6.3 Inverse Gaussian Interarrivals

An inverse Gaussian interarrival distribution with mean 1 and variance 1 produces a MGF of failure time converging on $(-\infty, 1.30 \times 10^{-2})$ where the right edge is a zero in the denominator of the cofactor rule. The mean is 79.1 with standard deviation 76.7. The MGF computation is based upon

$$\mathscr{U}_k(s) = \frac{e\mu^k}{k!}\sqrt{\frac{2}{\pi}}(1-2s)^{-1/2(k-1/2)}K_{k-1/2}(\sqrt{1-2s}),$$
$$s < \frac{1}{2},\ k = 0, \ldots, 3,$$

where K_ν denotes a BesselK function. For half-integer values, this function takes the simple form of the finite sum in (10.2.15) of Abramowitz and Stegun (1970). The convergence strip of the inverse Gaussian interarrival MGF is $(-\infty, 1/2]$ and not open; however, the MGF of the failure distribution has an open convergence strip due to the zero in the denominator of the cofactor rule.

Figure 7 plots $\hat{R}(t)$ (solid) and $\check{R}(t)$ (dashed) as previously described. Figure 8 plots the two normalized failure rates $\tilde{z}(t)$ (solid) and the inverse Gaussian-based hybrid $\check{z}(t)$ (dashed) in (17). Their empirical counterparts (circles) are based on 10^6 simulations making use of the generator for inverse Gaussian variates in Atkinson (1982). The normalization constant for the saddlepoint density is 1.080.

The strikingly similar behavior of the failure distributions and their MGFs for these three different interarrival distributions suggests that the accuracy and behavior of the saddlepoint approximations is due more to the system structure as expressed through the cofactor rule and less to the actual interarrival distributions used. Certainly the right tail behavior is determined by the shape of the MGF just to the left of the smallest positive root of its denominator.

7. A REDUNDANT AND REPAIRABLE SYSTEM

A partly loaded repairable system with imperfect switching is an example of a feedback system whose reliability function and failure rate may be approximated using saddlepoint methods. See Høyland and Rausand (1994, Sections 4.6, 4.7) for a discussion of this and other such examples.

Suppose that a pumping station has four equivalent pumps available for use. Under ordinary operating conditions, there are two active pumps. One is designated as the primary pump and has an Exponential(λ) failure time with failure rate λ. The second active pump is a partly loaded backup which means that it shares a reduced load with an Exponential(λ_1) failure time with $\lambda_1 < \lambda$. If any of the other pumps are not in queue for repair, then they are being held in cold standby, which means that, when activated as either the primary or backup pump, they will assume the same exponential lifetimes as their predecessors. The pumps are repairable by four independent servers and we assume that all individual repair times are Exponential(μ). The system is subject to imperfect switching in the activation of replacement pumps. Assume that each switching attempt is an independent Bernoulli($p = 1 - q$) trial and that, once the switching mechanism has failed for the first time, the system eventually fails once the currently active primary pump has failed. Finally, suppose that attempts to activate new pumps occur only following the failure of the primary active pump.

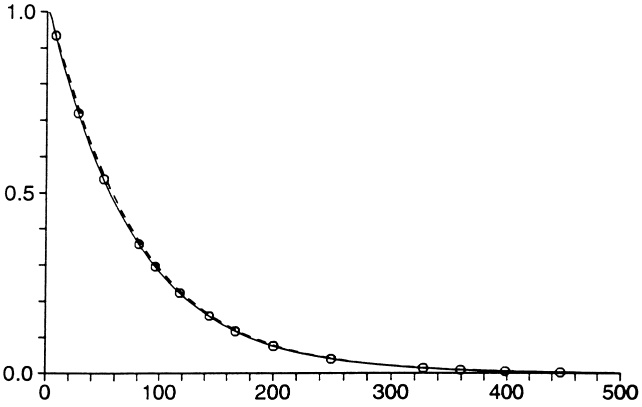

FIG. 7. *Plot of $\hat{R}(t)$ (solid), inverse Gaussian-based $\check{R}(t)$ (dashed) and an empirical estimate of $R(t)$ (circles) versus t for the $GI/M/1$ queue with inverse Gaussian interarrivals.*

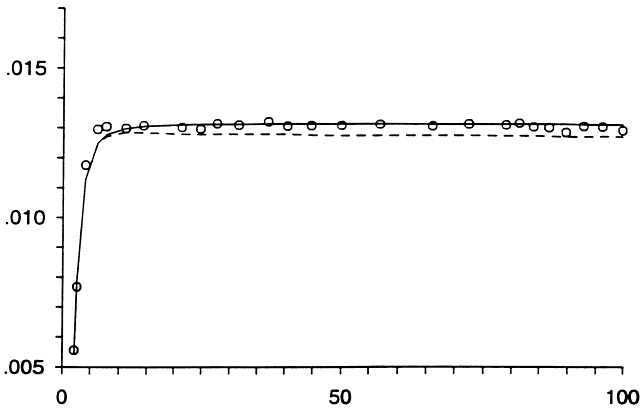

FIG. 8. *Plot of $\tilde{z}(t)$ (solid), $\check{z}(t)$ (dashed) and an empirical approximation (circles) versus t for the $GI/M/1$ queue with inverse Gaussian interarrivals.*

The behavior of this pumping system is quite complicated and the task before us is to determine the MGF of its failure time using one of the cofactor rules. There is however, some latitude in the designation of system states. Suppose that the failure of the primary pump is the trigger that initiates a change in system state. While waiting for the primary unit to fail in Exponential(λ) time, the backup pump has not failed in the interim with probability $r = \lambda/(\lambda_1 + \lambda) > 0.5$, computed as the competing risk between Exponential(λ) and Exponential(λ_1) failure times. The destination state following primary pump failure is state $j = 1, \dots, 4$ under two conditions: (1) the activations of the new primary and/or backup pumps are successful, and (2) subsequently there are j pumps undergoing repair. The destination state is failure state 5 when condition (1) fails, that is, when the switching mechanism fails to successfully activate all the required pumps. The initial state is state 0. With this designation of states, the transmittance matrix is given as (20) where the states have been ordered $0, \dots, 5$. Entry $M_\lambda(s)$ refers to the MGF of an Exponential(λ) waiting time and $N(s) = qM_\lambda + (1-r)pqM_\lambda^2$. Details for the determination of this form, specification of the entries $\{p_{ij}\}$ and treatment of the n-unit generalization are given in Appendix C. It should be clear from the large number of nonzero branch transmittances that the flowgraph of this system is very complicated and difficult to draw. The system is semi-Markov and not Markov because, while in states 0, 1 and 2, destination 5 has a different holding time than the other possible destinations:

$$(20) \qquad \mathscr{Q}(s) = \begin{pmatrix} 0 & prM_\lambda(s) & p^2(1-r)M_\lambda(s) & 0 & 0 & N(s) \\ 0 & p_{10}M_\lambda(s) & p_{11}M_\lambda(s) & p_{12}M_\lambda(s) & 0 & N(s) \\ 0 & p_{21}M_\lambda(s) & p_{22}M_\lambda(s) & p_{23}M_\lambda(s) & p_{24}M_\lambda(s) & N(s) \\ 0 & p_{31}M_\lambda(s) & p_{32}M_\lambda(s) & p_{33}M_\lambda(s) & p_{34}M_\lambda(s) & p_{3D}M_\lambda(s) \\ 0 & 0 & 0 & pM_{4\mu}(s) & 0 & qM_{4\mu}(s) \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

As a numerical example, suppose the primary unit has failure rate $\lambda = 1$ and the single backup unit has the reduced rate $\lambda_1 = 1/2$. The fixing rate is $\mu = 2$ and a successful use of the switch has probability $p = 0.95$. We consider two different types of system failure: (1) passage to state 5 so failure is due to the switching mechanism, and (2) passage to either state 4 or 5 so that failure refers to the first stoppage of all pumps.

First passage to 5. The cofactor rule in (4) gives a rational expression for $f_{05}\mathscr{F}_{05}(s)$ such that $f_{05} = 1$, so failure is ultimately assured. The mean time until failure is $\mathscr{F}'_{05}(0) = 14.8$, while the standard deviation may also be determined from the second derivative as 14.5. Approximate equality of the mean and standard deviation suggests an approximate exponential failure distribution, but the exact distribution is known to be a phase-type distribution (Aalen, 1995 and Neuts, 1981). The convergence strip of $\mathscr{F}_{05}(s)$ is $s \in (-\infty, 6.92 \times 10^{-2})$.

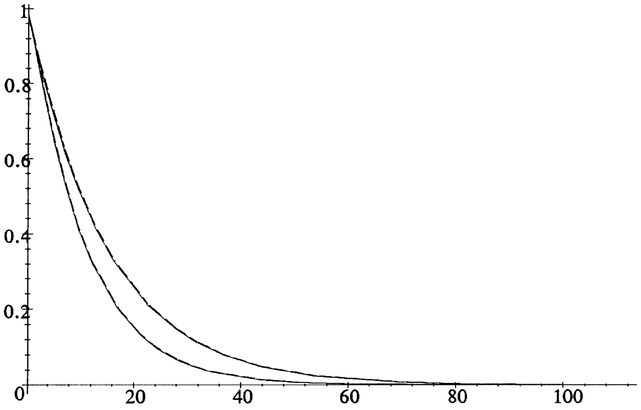

FIG. 9. *Plot of $R(t)$ (solid) and $\hat{R}(t)$ (dashed) versus t in the Markov system for the passages $0 \to 5$ (higher) and $0 \to \{4,5\}$ (lower).*

Since $\mathscr{F}_{05}(s)$ is a rational expression, the exact reliability function and failure rates can be determined. This is discussed in Section 8 in a broader context that relates the traditional solution of Markov systems to our saddlepoint approximation based upon the cofactor rules. Figure 9 plots $R(t)$ and its saddlepoint approximation $\hat{R}(t)$ versus t as the upper pair of curves and demonstrates virtually no graphical difference.

Figure 10 compares the failure rate approximations in (16) as the lower triple of curves. Numerical integration yields $\int \hat{f}(u)\,du \simeq 1.074$ which was used in computing $\tilde{z}(t)$. The true asymptote is given by $z(t) = z(\infty) = 0.0692$, as discussed in Section 8. Figure 11 shows the percentage relative errors connected with the two previous graphs as the solid and dashed lines. The limiting relative error in failure rate approximation as $t \to 0$ is determined numerically as

$$\lim_{t \to 0} 100 \left(\frac{\tilde{z}(t)}{z(t)} - 1 \right) = -23.3\%.$$

(*First passage to $D = \{4, 5\}$*). The first-passage transmittance expression in (7) yields $f_{1D}\mathscr{F}_{1D}(s)$ which is again a rational expression whose exact

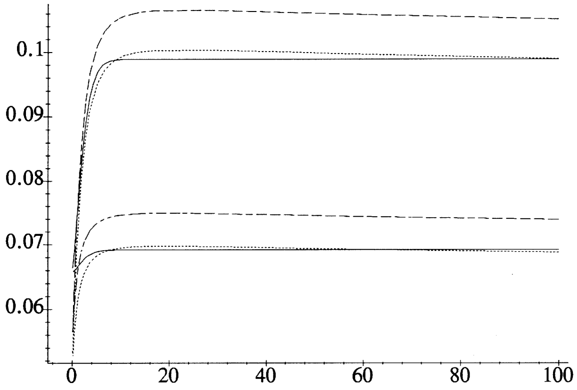

FIG. 10. *Plot of $z(t)$ (solid), $\hat{z}(t)$ (dashed) and $\tilde{z}(t)$ (dotted) versus t in the Markov system for passages $1 \to 5$ (lower) and $1 \to \{4, 5\}$ (upper).*

reliability function and failure rate can be determined and used for checking saddlepoint accuracy. The first passage transmittance has $f_{1D} = 1$, mean time to failure of $\mathscr{F}'_{1D}(0) = 11.0$, standard deviation 10.2, and convergence region $(-\infty, 9.90 \times 10^{-2})$. The remaining graphs in Figures 9–11 pertain to this example and compare the saddlepoint and exact computations. Again they demonstrate extremely close agreement. For Figure 10, the true asymptotic failure rate is $z(\infty) = 0.099$. In Figure 11, the dotted and dotted-dashed lines plot the relative error in approximating $R(t)$ and $z(t)$ and

$$\lim_{t\to 0} 100\left(\frac{\tilde{z}(t)}{z(t)} - 1\right) = -22.4\%.$$

8. FINITE MARKOV SYSTEMS

The traditional approach to computing the reliability for finite Markov systems is to use the solution to its backward Kolmogorov differential equation. This approach, while limited to Markov systems, is another useful method in reliability computation as shown in Aalen (1995). If $\mathscr{Q}(s) = \{\mathscr{Q}_{ij}(s)\}$ is the $n \times n$ matrix of system transmittances among states $S = \{1, \dots, n\}$, then the system is characterized as Markov when $\mathscr{Q}_{ij}(s) = p_{ij}(1 - s/\tau_i)^{-1}$ for all i, j; for example, holding times in state i are Exponential(τ_i) and not dependent upon the destination states. The distribution of first passage from state 1 to the subset of failure states $D = \{m+1, \dots, n\}$ is obtained as follows. Amend the system so the states in D are absorbing; then the reliability function is the probability the system is not in D at time t or

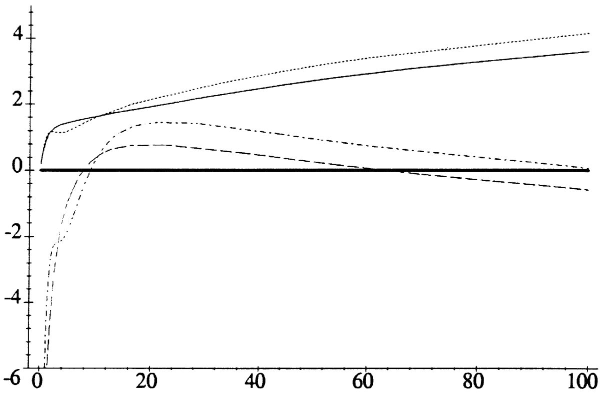

FIG. 11. *Percentage relative errors in saddlepoint approximation for $R(t)$ versus t for passage $1 \to 5$ (solid) and passage $1 \to \{4, 5\}$ (dots). Similiarly, relative error plots in approximation for $z(t)$ are the respective dashed and dot-dashed curves.*

$$R(t) = \sum_{j=1}^{m} p_{1j}(t), \tag{21}$$

with $P(t) = \{p_{ij}(t)\}$ as the solution to the appropriate backward equation of the amended system. The specifics of this would seem to require consideration of two separate cases: systems with and without states that have self-feedback loops.

When there are no self-feedback loops, denote $\mathscr{I} = (\mathscr{I}_{ij})$ as the $n \times n$ infinitesimal generator matrix in which

$$\begin{aligned} \mathscr{I}_{ij} &= \begin{cases} p_{ij}\tau_i, & \text{if } i \notin D, \\ 0, & \text{if } i \in D, \end{cases} \qquad j \neq i, \\ \mathscr{I}_{ii} &= -\sum_{j \neq i} \mathscr{I}_{ij}\,. \end{aligned} \tag{22}$$

In block form,

$$\mathscr{I} = \begin{pmatrix} \mathscr{I}_{CC} & \mathscr{I}_{CD} \\ 0 & 0 \end{pmatrix},$$

where the first m components are block C and the last $n - m$ components comprise D. According to standard theory, such as in Karlin and Taylor (1981),

$$\begin{aligned} P(t) &= \exp(t\mathscr{I}) \\ &= \begin{pmatrix} \exp(t\mathscr{I}_{CC}) & \mathscr{I}_{CC}^{-1}\{\exp(t\mathscr{I}_{CC}) - I_m\}\mathscr{I}_{CD} \\ 0 & I_{n-m} \end{pmatrix}, \end{aligned}$$

as can be shown in this instance when D consists of absorbing states. The long-term absorption probabilities are $\lim_{t\to\infty} P_{CD}(t) = -\mathscr{I}_{CC}^{-1}\mathscr{I}_{CD}$ and the system reliability is the sum of components of the first row of $\exp(t\mathscr{I}_{CC})$, computable as a function of t in Maple V.

Self-feedback loops in a Markov system need to be removed before it is possible to solve the backward equation. A self-feedback loop on state i is removed by amending all transmittances out of state i in the following manner. Passage from state $i \to j \neq i$ has transmittance $\mathscr{Q}_{ii}^k(s)\mathscr{Q}_{ij}(s)$ if the self-feedback loop is taken exactly k times; accordingly, the overall transmittance sums over k to give an (i, j)th transmittance of

$$\sum_{k=0}^{\infty} \mathscr{Q}_{ii}^{k}(s)\mathscr{Q}_{ij}(s) = \frac{\mathscr{Q}_{ij}(s)}{1-\mathscr{Q}_{ii}(s)} = \frac{p_{ij}(1-s/\tau_i)^{-1}}{1-p_{ii}(1-s/\tau_i)^{-1}} = \frac{p_{ij}}{1-p_{ii}}\left(1-\frac{s}{\tau_i(1-p_{ii})}\right)^{-1}, \quad j \neq i. \tag{23}$$

Thus, the equivalent system without self-feedback loops is again Markov with transmittances as in (23). The infinitesimal generator of this system has exactly the same form as in (22) which suggests what we might have expected: we may ignore self-loops in determining $\mathscr{I}$ for a Markov system. With or without self-feedback loops, the infinitesimal generator is computed in the same manner.

As an example, consider the partly loaded repairable system with imperfect switching from the previous section. As presented, the system is not Markov but it can be made Markov by adding in an additional state 6. This new state is entered when the switching mechanism succeeds in connecting up a new primary unit but fails in connecting a new backup unit. Transmittances of the former system into state 5, such as $N(s)$, are now split into two separate pieces, one into 6 and the other into 5. With the states ordered as $\{0, \ldots, 4, 5, 6\}$, the Markov transmittance matrix is

$$\mathscr{Q} = \begin{pmatrix} 0 & prM_\lambda & p^2(1-r)M_\lambda & 0 & 0 & qM_\lambda & (1-r)pqM_\lambda \\ 0 & p_{10}M_\lambda & p_{11}M_\lambda & p_{12}M_\lambda & 0 & qM_\lambda & (1-r)pqM_\lambda \\ 0 & p_{21}M_\lambda & p_{22}M_\lambda & p_{23}M_\lambda & p_{24}M_\lambda & qM_\lambda & (1-r)pqM_\lambda \\ 0 & p_{31}M_\lambda & p_{32}M_\lambda & p_{33}M_\lambda & p_{34}M_\lambda & p_{35}M_\lambda & p_{36}M_\lambda \\ 0 & 0 & 0 & pM_{4\mu} & 0 & qM_{4\mu} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & M_\lambda & 0 \end{pmatrix}, \tag{24}$$

where an explanation for the transition probabilities is given in Appendix C.

When considering first passage to state 5, the Markov system yields a 6×6 matrix $\mathscr{I}_{CC}$ with eigenvalues -0.0692, -1 with multiplicity 2, -8.15, and the complex conjugate pair $-0.915 \pm 0.171i$; the exact reliability as computed in (21) using Maple V is

$$\begin{aligned} R(t) = {} & 1.01e^{-0.0692t} \\ & +0.0133e^{-0.915t}\cos(0.171t) \\ & -0.0338e^{-0.915t}\sin(0.171t) \\ & -0.0220e^{-t} + 1.85 \times 10^{-6}e^{-8.15t}. \end{aligned} \tag{25}$$

The leading term determines the asymptotic order in t and has an exponent which is the right edge of the convergence strip for the rational first passage MGF $\mathscr{F}_{05}(s)$. This calculation of $R(t)$ gives the true asymptotic failure rate as

$$z(\infty) = \lim_{t\to\infty} z(t) = \lim_{t\to\infty} \frac{-R'(t)}{R(t)} = 0.0692.$$

The exact reliability and failure rate used in Figures 9–11 were based upon (25).

For first passage to $D = \{5, 6\}$, the exact reliability function is

$$\begin{aligned} R(t) = {} & 1.08e^{-0.0990t} - 0.213e^{-1.0t} \\ & +0.137e^{-0.974t}\cos(0.180t) \\ & -0.220e^{-0.974t}\sin(0.180t). \end{aligned}$$

The asymptote for the true failure rate is $z(t) = 0.0990$, the smallest real pole of $\mathscr{F}_{0D}(s)$.

9. FEEDBACK THEORY FOR FINITE STOCHASTIC SYSTEMS

The simplicity in form of the cofactor rules in Section 4 suggests that there should be elementary derivations that are much simpler than the original proofs in Butler (1997a, b). We provide these simple derivations below using methods of matrix algebra. The pictorial analogues to these derivations are flowgraphs with matrix self-feedback loops. The algebra and flowgraphs combine to offer considerable insight into the nature of stochastic systems that may have complicated collections of multiple feedback loops.

9.1 Single Destination Cofactor Rule

The first passage transmittance from $1 \to n \neq 1$ is defined as

$$f_{1n}\mathscr{F}_{1n}(s) = E\{\exp(sW_{1n})1_{(W_{1n}<\infty)}\},$$

where W_{1n} is the first passage time from state $1 \to n$, perhaps having a defective distribution. The distribution theory for W_{1n} is determined entirely by $\mathscr{Q}$, the transmittance matrix of the semi-Markov process. Since the process never leaves state n during the occurrence of W_{1n}, its distribution theory is unaltered by working with the semi-Markov process in which state n has been made absorbing. Let $\mathscr{Q}^{(n)}$

denote the transmittance matrix of such a process with

$$(26)\quad \mathscr{Q}^{(n)}(s) = \left\{ \mathscr{Q}_{ij}^{(n)}(s) \right\} := \begin{pmatrix} \mathscr{Q}_{11} & \cdots & \mathscr{Q}_{1n} \\ \vdots & & \vdots \\ \mathscr{Q}_{n-1,1} & \cdots & \mathscr{Q}_{n-1,n} \\ 0 & \cdots & 0 \end{pmatrix}.$$

Since we are considering nonexplosive processes, passage from $1 \to n$ in finite time must be in a countable number of state transitions. With N counting the number of such transitions, then

$$(27)\quad f_{1n}\mathscr{F}_{1n}(s) = \sum_{k=1}^{\infty} E\left\{ \exp(sW_{1n}) 1_{(W_{1n}<\infty \cap N=k)} \right\}.$$

Passage in 1 step contributes the term $\mathscr{Q}_{1n}^{(n)}$. Two-step transition contributes

$$\sum_{j=1}^{n-1} \mathscr{Q}_{1j}^{(n)} \mathscr{Q}_{jn}^{(n)} = \sum_{j=1}^{n} \mathscr{Q}_{1j}^{(n)} \mathscr{Q}_{jn}^{(n)},$$

which is the $(1, n)$ element of $\{\mathscr{Q}^{(n)}(s)\}^2$. More generally, the kth term of (27) is the $(1, n)$ element of $\{\mathscr{Q}^{(n)}(s)\}^k$ so that

$$\begin{aligned} f_{1n}\mathscr{F}_{1n}(s) &= \sum_{k=1}^{\infty} \left[(1, n) \text{ element of } \{\mathscr{Q}^{(n)}(s)\}^k \right] \\ (28)\qquad &= (1, n) \text{ element of } \sum_{k=1}^{\infty} \{\mathscr{Q}^{(n)}(s)\}^k \\ &= (1, n) \text{ element of} \\ &\qquad \left[\{I_n - \mathscr{Q}^{(n)}(s)\}^{-1} - I_n \right]. \end{aligned}$$

Let the cofactor matrix for $I_n - \mathscr{Q}^{(n)}(s)$ be $\{(-1)^{i+j} |\Psi_{ij}^{(n)}(s)|\}$ and use the same notation without exponent (n) for the cofactors of $I_n - \mathscr{Q}(s)$. Then, by Cramer's rule,

$$f_{1n}\mathscr{F}_{1n}(s) = \frac{(-1)^{n+1} \left|\Psi_{n1}^{(n)}(s)\right|}{\left|I_n - \mathscr{Q}^{(n)}(s)\right|} = \frac{(-1)^{n+1} \left|\Psi_{n1}(s)\right|}{\left|\Psi_{nn}(s)\right|}.$$

These arguments are valid when $I_n - \mathscr{Q}^{(n)}(s)$ is full rank. This fact has been shown in Butler (1997a, b) under the conditions of Theorem 1.

The flowgraph analogue to this argument is given in Figure 12. The numbers below the nodes give the dimensions of the nodes with B and E as one-dimensional source and sink nodes. There are two different versions of node S, each representing an n-dimensional node comprised of all the system states $\{1, \ldots, n\}$. Vector ξ_i is the $n \times 1$ indicator vector of its ith component. Thus, branch transmittance $\xi_1^T = (1, 0^T)$ effectively inserts the system from source into state 1, the first component of the first S, with no time delay. Transition from the first S to the second version of S has the $n \times n$ matrix branch transmittance $\mathscr{Q}(s)$. Upon reaching the second S, the system may feedback about this matrix node with transmittance $\mathscr{Q}^{(n)}$ given in (26) for $k \geq 0$ loops. The final branch transmittance $\xi_n = (0^T, 1)^T$ removes the system from the nth component of S without time delay. The summation computation in (27) and (28) is equivalent to summing over all mutually exclusive paths from node B to E and is

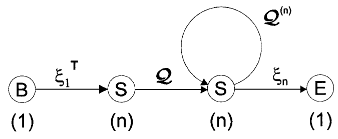

FIG. 12. *A matrix self-loop of the state space $S = \{1, \ldots, n\}$ into itself with the various nodal dimensions given in parentheses below.*

$$\begin{aligned} &\xi_1^T \mathscr{Q}(s) \sum_{k=0}^{\infty} \left\{ \mathscr{Q}^{(n)}(s) \right\}^k \xi_n \\ (29)\quad &= \xi_1^T \sum_{k=1}^{\infty} \left\{ \mathscr{Q}^{(n)}(s) \right\}^k \xi_n \\ &= \xi_1^T \left[\left\{ I_n - \mathscr{Q}^{(n)}(s) \right\}^{-1} - I_n \right] \xi_n \\ &= (1, n) \text{ element of } \left[\{I_n - \mathscr{Q}^{(n)}(s)\}^{-1} - I_n \right], \end{aligned}$$

which gives the cofactor rule. Two versions of matrix node S are required in the flowgraph of Figure 12 because the transition $1 \to n \neq 1$ requires at least one state transition in state space S. The matrix self-feedback loop alone, with transmittance $\left\{ I_n - \mathscr{Q}^{(n)}(s) \right\}^{-1}$, would allow for the possibility of passage with 0 transitions; thus the elimination of the first S node would yield the same answer but the flowgraph algebra would then be incorrect.

9.2 First Return Cofactor Rule

The flowgraph in Figure 12 is easily modified to yield a very simple derivation of the first return cofactor rule from state $1 \to 1$. For the matrix self-feedback loop of $S \to S$, use instead the matrix transmittance $\mathscr{Q}^{(1)}(s)$ defined as $\mathscr{Q}(s)$ with the first row set to 0 so that state 1 is absorbing. According to the flowgraph, passage $B \to S$ starts the system in state 1. Passage $S \to S$ with transmittance $\mathscr{Q}(s)$ is the mandatory first step into some state in S. Thereafter, the system accumulates time by passing through the feedback loop with transmittance $\mathscr{Q}^{(1)}(s)$. Once the system enters state 1, time stops and transmittance ξ_1 takes the system from $1 \in S \to E$. This yields the first return transmittance,

$$f_{11}\mathscr{F}_{11}(s) = \xi_1^T \mathscr{Q}(s) \sum_{k=0}^{\infty} \left\{ \mathscr{Q}^{(1)}(s) \right\}^k \xi_1$$

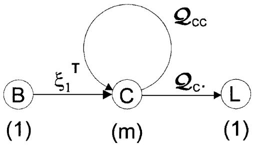

FIG. 13. *Multiple destination flowgraph with destination states lumped into a single state.*

$$= \xi_1^T \mathscr{Q}(s) \left\{ I_n - \mathscr{Q}^{(1)}(s) \right\}^{-1} \xi_1$$
$$= (1,1) \text{ element of } \mathscr{Q}(s) \left\{ I_n - \mathscr{Q}^{(1)}(s) \right\}^{-1}.$$

Using a cofactor expansion of the first column of $\left\{ I_n - \mathscr{Q}^{(1)}(s) \right\}^{-1}$ gives

$$\begin{aligned} f_{11}\mathscr{F}_{11}(s) &= \sum_{j=1}^{n} \frac{\mathscr{Q}_{1j}(s)(-1)^{1+j} \left| \Psi_{1j}^{(1)}(s) \right|}{\left| I_n - \mathscr{Q}^{(1)}(s) \right|} \\ &= \sum_{j=1}^{n} \frac{\mathscr{Q}_{1j}(s)(-1)^{1+j} \left| \Psi_{1j}(s) \right|}{\left| \Psi_{11}(s) \right|} \\ &= -\sum_{j=1}^{n} \frac{\{I_n - \mathscr{Q}(s)\}_{1j} (-1)^{1+j} \left| \Psi_{1j}(s) \right|}{\left| \Psi_{11}(s) \right|} \\ &\quad + \frac{(-1)^2 \left| \Psi_{11}(s) \right|}{\left| \Psi_{11}(s) \right|} \\ &= -\frac{\left| I_n - \mathscr{Q}(s) \right|}{\left| \Psi_{11}(s) \right|} + 1. \end{aligned}$$

The last line results upon recognizing that the numerator is the cofactor expansion of $\left| I_n - \mathscr{Q}(s) \right|$.

9.3 Multiple Destination Cofactor Rule

The first passage transmittance from $1 \in C = \{1, \ldots, m\} \to D = \{m+1, \ldots, n\}$ is determined from the flowgraph in Figure 13. State L is a one-dimensional lumped state representing the collection of $n-m$ states in D. One-step branch transmittances into this lumped state from vector node C are the components of $m \times 1$ vector $\mathscr{Q}_{C\cdot} := \mathscr{Q}_{CD}1$. It should be intuitively clear that the first passage transmittance is most easily determined by basing its derivation upon the lumped state and its associated branch transmittance $\mathscr{Q}_{C\cdot}$; formal justification for working with the lumped state is given in Appendix D.

From Figure 13, with ξ_1 as the $m \times 1$ indicator of the first component,

$$f_{1D}\mathscr{F}_{1D}(s) = \xi_1^T \left\{ I_m - \mathscr{Q}_{CC}(s) \right\}^{-1} \mathscr{Q}_{C\cdot}$$

or the first component of the vector $\left\{ I_m - \mathscr{Q}_{CC}(s) \right\}^{-1} \mathscr{Q}_{C\cdot}$. This is the expression in (7) by Cramér's rule.

9.4 Other Properties

The derivations of these cofactor rules and their motivation using flowgraphs provide an understanding of these formulas as sums over distinct paths. Each of the three expressions is concerned with computing the expectation of $\exp(sW)$ where W is a finite waiting time for first passage. The first-passage transmittances are developed as sums of contributions over mutually exclusive events and the partitioning into these disjoint events is given pictorially through the matrix feedback loops of Figures 12 and 13.

Suppose that there are no feedback loops in the n-state system and no states of the system can be repeated. Such a system is a *cascading system* and the cofactor rules simplify further in this context.

COROLLARY 4. *Consider the n-state system of Theorem* 1 *and suppose it is also cascading. Then the first passage transmittance from* $1 \to n$ *in* (4) *further simplifies to*

$$(30) \qquad f_{1n}\mathscr{F}_{1n}(s) = (-1)^{n+1} \left| \Psi_{n1}(s) \right|, \quad n \neq 1.$$

PROOF. For a cascading system the entries of $\mathscr{Q}(s)$ on or below the diagonal are zero; thus $\left| \Psi_{nn}(s) \right| = 1$. □

The cofactor in (30) has the physical interpretation as the sum over all cascading paths from $1 \to n$ each of which cannot feed back upon itself.

A similar result holds for first passage to a subset of states D.

COROLLARY 5. *Consider the n-state system of Theorem 3 and suppose it is also cascading in subset C. Then the first passage transmittance from* $1 \in C \to D$ *in* (7) *further simplifies to*

$$f_{1D}\mathscr{F}_{1D}(s) = \left| \mathscr{Q}_{C\cdot}(s) \quad \{I_m - \mathscr{Q}_{CC}(s)\}_{\backslash 1} \right|.$$

The matrix is $I_m - \mathscr{Q}_{CC}(s)$ with its first column replaced with $\mathscr{Q}_{C\cdot}(s)$, the row sums of the $m \times (n-m)$ block $\mathscr{Q}_{CD}(s)$.

Suppose the system starts in state i with prior probability π_i for $i \in S$ so that the prior state vector is $\pi^T = (\pi_1, \ldots, \pi_n)$ with $\pi^T 1 \leq 1$. This is a situation in which each individual system state has its own input. The first passage transmittance to state n can be computed using the superposition or additivity of the inputs.

THEOREM 6. *Suppose an n-state system consists of all states relevant to passage from* $\{1, \ldots, n-1\}$

$\to n$. Let $\pi_{-n}^T = (\pi_1, \ldots, \pi_{n-1}, 0)$. The first passage transmittance to state n is

$$(31)\quad f_{\pi n}\mathscr{F}_{\pi n}(s) = \frac{1}{|\Psi_{nn}(s)|}\begin{vmatrix} \{I_n - \mathscr{Q}(s)\}_{-n} \\ \pi_{-n}^T \end{vmatrix},$$

where $\{I_n - \mathscr{Q}(s)\}_{-n}$ is $(n-1) \times n$ and defined as $I_n - \mathscr{Q}(s)$ with its nth row removed.

PROOF. The flowgraph in Figure 12 describes the system; however, now the input branch transmittance from $B \to S$ is π_{-n}^T rather than ξ_1^T as shown. The computation is therefore

$$f_{\pi n}\mathscr{F}_{\pi n}(s) = \pi_{-n}^T\left[\left\{I_n - \mathscr{Q}^{(n)}(s)\right\}^{-1} - I_n\right]\xi_n,$$

which reduces to (31) using the arguments given in (29).

10. MODULAR SYSTEMS

In all the systems discussed above, the branch transmittances could have been first-passage transmittances of subsystems whose detailed structures have been suppressed in order to make the overall system and its flowgraph look simpler. This is the modular concept by which systems are perceived as comprised of subsystems and further subsubsystems in an effort to provide hierarchical organization to a system with many states. The modular concept is extremely powerful and general and has deeper implications for the analysis of large complex systems when used in conjunction with the aforementioned cofactor rules.

A simple example in which the modular concept provides a simple first passage analysis concerns the first return to state 0 in a birth-death process as first considered in Butler and Huzurbazar (1995). Consider the infinite state system in which S is the set of integers. Split state 0 into two versions: source state 0 to which the system cannot return, and sink state 0′, the version to which the system can return. Create the module, or first passage transmittance, representing first return to state 1 from higher numbered states and denote it as $\mathscr{U}_{11}^+$. Figure 14 shows this modular system. The cofactor rule applied to this flowgraph gives

$$f_{00'}\mathscr{F}_{00'}(s) = \frac{\mathscr{Q}_{01}(s)\mathscr{Q}_{10}(s)}{1 - \mathscr{U}_{11}^+(s)}.$$

More generally if the module for first return to state i from the right has transmittance $\mathscr{U}_{ii}^+(s)$, then a recursion relation for the modular transmittances is

$$(32)\quad \mathscr{U}_{ii}^+(s) = \frac{\mathscr{Q}_{i,\,i+1}(s)\mathscr{Q}_{i+1,\,i}(s)}{1 - \mathscr{U}_{i+1,i+1}^+(s)}, \quad i = 0, 1, \ldots .$$

This yields a recursive computation of $f_{00'}\mathscr{F}_{00'}(s)$ as $\mathscr{U}_{00}^+(s)$.

A two-sided birth–death process that considers first return to state 0 by allowing for negatively numbered states would consist of two modules connected in parallel: Figure 14 as one module and its mirror image as the second module over the negative states. If $\mathscr{U}_{ii}^-(s)$ denotes the first return transmittance to state i from lower-numbered states, then

$$(33)\quad \mathscr{U}_{ii}^-(s) = \frac{\mathscr{Q}_{i,\,i-1}(s)\mathscr{Q}_{i-1,\,i}(s)}{1 - \mathscr{U}_{i-1,\,i-1}^-(s)}, \quad i = 0, -1, \ldots$$

is the mirror image recursion. The overall first return transmittance is now $\mathscr{U}_{00}^+(s) + \mathscr{U}_{00}^-(s)$. In the setting of a homogeneous random walk with $\mathscr{Q}_{i,\,i+1}(s) = pe^s$ and $\mathscr{Q}_{i,\,i-1}(s) = qe^s$, then

$$\mathscr{U}_{ii}^+(s) = \mathscr{U}_{i+1,\,i+1}^+(s) = \mathscr{U}_{ii}^-(s) = \mathscr{U}_{i+1,\,i+1}^-(s) \quad \forall_i$$

and the common value can be resolved from recursions (32) and (33) by extracting the negative solution in the quadratic equations. This yields the known result

$$(34)\qquad \mathscr{U}_{00}^-(s) = \mathscr{U}_{00}^+(s) = \tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 4pqe^{2s}}$$

found by Feller [1968, Section XI.3, (3.13)] from a different point of view involving the renewal theory of generating functions. The numerical example of the next section is a two-sided random walk with $p = 1/2$ and Exponential(1) holding times so that $\mathscr{Q}_{i,\,i-1}(s) = \mathscr{Q}_{i,\,i+1}(s) = 1/2(1 - s)^{-1}$. If the walk starts at 0, the first return transmittance is

$$(35)\qquad \begin{aligned} &f_{00'}\mathscr{F}_{00'}(s) \\ &\quad = \frac{1/2(1-s)^{-2}}{1-\left(1/2-1/2\sqrt{1-(1-s)^{-2}}\right)}, \quad s \le 0. \end{aligned}$$

Modular system concepts also give the first-passage transmittance from state 0 to n in a general birth–death process. The modular transmittances are shown in the first system of Figure 15 as $\mathscr{V}_{0,\,n-1}$, the first-passage transmittance from state 0 to $n-1$, and $\mathscr{U}_{n-1,\,n-1}^-$, the first return transmittance to state $n-1$ from the lower numbered states.

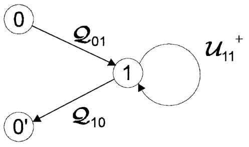

FIG. 14. *Modular system for first return to state 0 in a birth–death process.*

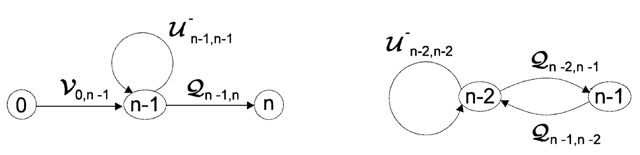

FIG. 15. *Modular system for first passage to state n in a birth–death process.*

Applying the cofactor rule, then

$$\mathscr{V}_{0,n}(s) = \frac{\mathscr{V}_{0,n-1}(s)\mathscr{Q}_{n-1,n}(s)}{1-\mathscr{U}^-_{n-1,n-1}(s)} \tag{36}$$

is a recursion of $\mathscr{V}_{0,n}$ in n. Its computation also depends on a recursion of $\mathscr{U}^-_{n-1,n-1}$ in n given by the second modular system of Figure 15 as

$$\mathscr{U}^-_{n-1,n-1}(s) = \frac{\mathscr{Q}_{n-1,n-2}(s)\mathscr{Q}_{n-2,n-1}(s)}{1-\mathscr{U}^-_{n-2,n-2}(s)}. \tag{37}$$

Recursions (36) and (37) together provide simple computations of $\mathscr{V}_{0,n}$ even for large values of n such as $n = 1000$. Saddlepoint methods also require computation of the first two derivatives, $\mathscr{V}'_{0,n}$ and $\mathscr{V}''_{0,n}$. These can also be computed recursively by analytically differentiating the two recursions and carrying through the six-dimensional recursions of the terms

$$\mathscr{U}^-_{n-1,n-1}\ \mathscr{U}^{-\prime}_{n-1,n-1}\ \mathscr{U}^{-\prime\prime}_{n-1,n-1}\ \mathscr{V}_{0,n}\ \mathscr{V}'_{0,n}\ \mathscr{V}''_{0,n}\ .$$

The only limitation in their use is the stability and accuracy of computations resulting from the accruement of roundoff error, a limitation found in all recursive computations.

In the homogeneous random walk setting, the expression for $\mathscr{U}^-_{n-1,n-1}$ in (34) can be used to show that the recursion in (36) is

$$\mathscr{V}_{0,n}(s) = \frac{pe^s}{1-(1/2-1/2\sqrt{1-4pqe^{2s}})}\mathscr{V}_{0,n-1}(s)$$
$$= (\frac{1-\sqrt{1-4pqe^{2s}}}{2qs})^n.$$

This is a known result in Feller [1968, Section XI.3, (3.6) and part (d)] derived from a quite different point of view.

10.1 Load Sharing Repairable System with Perfect Switching

Repairable systems are feedback systems of practical importance that are most often characterized as queueing systems. Some flowgraph examples of queues are given in Butler and Huzurbazar (1993, 1995) and include finite and infinite state $M/M/$p queues and an $M/G/1$ queue. The load sharing queue considered here assumes that working units share a common workload. If there are n units among which i are working, let λ_i be the common failure rate of each working unit and Exponential(λ_i) the common failure distribution. When, for example, tasks arrive at rate λ_0, the choice of $\lambda_i = \lambda_0/i$ matches the service and arrival rates of tasks. Suppose perfect switching and m independent servers with common fixing rate μ and distribution Exponential(μ). The system is a nonhomogeneous birth–death process. Passage out of state i, representing the state in which i units have failed, is a competition between the failure of $n-i$ working units with failure rates λ_{n-i} and the repair of $i_m = \min(i, m)$ units with repair rate μ; the branch transmittances are

$$\mathscr{Q}_{i,i-1}(s) = \frac{i_m\mu}{t_i}M_{t_i}(s)$$

and

$$\mathscr{Q}_{i,i+1}(s) = \frac{(n-i)\lambda_{n-i}}{t_i}M_{t_i}(s), \quad i = 0,\ldots,n,$$

where $t_i = i_m\mu + (n-i)\lambda_{n-i}$. System failure occurs in state n and the first passage transmittance to state n is $\mathscr{V}_{0,n}$ computed from the recursions along with saddlepoint ingredients $\mathscr{V}'_{0,n}\ \mathscr{V}''_{0,n}$.

10.2 Null Persistent Random Walk

This is a system with a countably infinite number of states on the integers. A simple random walk starts at zero and has an Exponential(1) holding time in each state. Consider the first return time to state zero with transmittance (35) which is also the MGF of first return due to null persistence of the recurrence. Its convergence strip is $(-\infty, 0]$ which is not open; however, the saddlepoint equation can always be solved for $t \in (0, \infty)$ and saddlepoint approximations are available.

Figure 16 plots $\breve{R}(t)$ (dashed) versus an empirical estimate (circles) based on simulation of 10^6 returns; the smaller inset plot shows the accuracy for

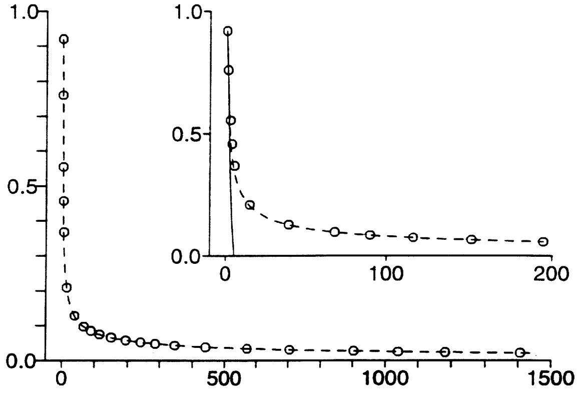

FIG. 16. *Plot of inverse Gaussian-based $\breve{R}(t)$ (dashed), $\hat{R}(t)$ (solid) and an empirical estimate of $R(t)$ (circles) versus t for the first return time of the random walk.*

$t \in (0, 200)$ and also plots $\hat{R}(t)$ as the solid line. The Lugannani and Rice approximation fails miserably with reliability estimates turning negative at $t = 5$ and continuing to decrease to -15 at $t = 1516$. By contrast, the inverse Gaussian-based approximation $\check{R}(t)$ agrees quite closely with the simulation. At $t = 8$ the value of $\hat{\alpha}$ for the inverse Gaussian base has grown to 10^3 and continues to increase out to 10^{10} at $t = 1516$. This suggests that the inverse Gaussian base is approaching the stable law with index 1/2, a heavy-tailed distribution with mean ∞. The accuracy of this approximation for this continuous waiting distribution agrees with its accuracy when used (with continuity correction) to approximate the analogous discrete random walk in which holding times are fixed at 1 (Wood, Booth and Butler, 1993 and Booth and Wood, 1995). Figure 17 plots $\check{z}(t)$ (solid) versus t and empirical estimates (circles). The accuracy of the normal-based density approximation, as the numerator of $\check{z}(t)$, is reflected in this plot and contrasts sharply with the extreme inaccuracy of the Lugannani and Rice approximation, its CDF counterpart. The normalization constant for the saddlepoint density is 1.000.

This example has been used to illustrate and contrast the failure of the Lugannani and Rice approximation with the success of the inverse Gaussian based approximation for heavy-tailed distributions. The particular heavy-tailed distribution was also chosen because there are explicit expressions for its Bessel density and CDF. Feller [1971, Section XIV.7, (6.16)] gives the density as

$$f(t) = t^{-1} I_1(t) e^{-t}, \quad t > 0, \tag{38}$$

where I_1 is a BesselI function of order 1. Integration, using Abramowitz and Stegun (1970, 11.3.14), yields reliability

$$R(t) = (I_0(t) + I_1(t)) e^{-t}, \quad t > 0. \tag{39}$$

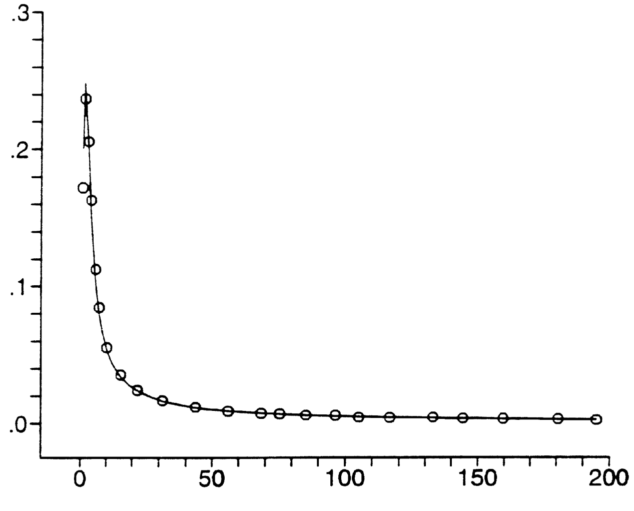

FIG. 17. *Plot of $\check{z}(t)$ (solid) and an empirical approximation (circles) versus t for the first return time of the random walk.*

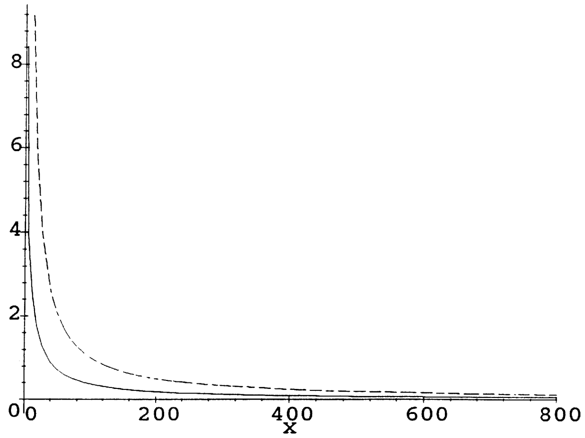

FIG. 18. *Plot of the exact percentage relative error of $\check{R}(t)$ (solid) and $\check{z}(t)$ (dashed) versus t.*

Figure 18 plots the percentage relative errors of $\check{R}(t)$ and $\check{z}(t)$ as compared with their true counterparts determined from (38) and (39). Beyond $t = 800$, these errors continue to decrease to the respective errors of 0.05% and 0.03% at $t = 1516$.

The example provides a strong argument for routine computation of the inverse Gaussian-based approximation as a companion to the Lugannani and Rice approximation. Appendix A demonstrates that this involves hardly any further computation. One of its most important roles, however, is in determining whether or not the Lugannani and Rice approximation can be trusted. From past numerical experience, what can be said about the Lugannani and Rice approximation is that it is either extremely accurate when "working properly" or extremely bad as seen with this example. Our recommendation is to trust it when the inverse Gaussian-based approximation is close but otherwise not. If these two approximations should differ considerably, then further comparison to the integrated saddlepoint density is in order, which could require considerably more computational effort.

APPENDIX

A. Inverse Gaussian-Based Saddlepoint Approximation

Let Γ_α and γ_α be the CDF and density of an inverse Gaussian distribution, with mean parameter $\alpha > 0$ and variance α^3, to be used as the base distribution of the approximation. Then

$$\Pr\{W > t\} \simeq \check{R}(t)$$

$$= 1 - \Gamma_{\hat{\alpha}}(\hat{\xi}) - \gamma_{\hat{\alpha}}(\hat{\xi}) \left(\frac{1}{w_{\hat{\xi}}} - \frac{1}{\hat{u}_{\hat{\xi}}} \right), \quad t \neq K'(0),$$

where $\hat{\alpha}$, $\hat{\xi}$, $w_{\hat{\xi}}$ and $\hat{u}_{\hat{\xi}}$ are chosen in the following way. First $\hat{\alpha}$ is determined as the value of α matching the third standardized cumulants of the tilted inverse Gaussian base with the tilted true distribution under approximation; this yields

$$\hat{\alpha} = \frac{\{K'''(\hat{s})\}^2}{3\{K''(\hat{s})\}^3}\left(3 + \hat{w}\sqrt{\frac{\{K'''(\hat{s})\}^2}{\{K''(\hat{s})\}^3}}\right)^{-1}, \tag{40}$$

where $\hat{w}$ is given in (11). This simplifies expression (20) in Wood, Booth and Butler (1993). The value $\hat{\xi}$ is

$$\hat{\xi} = \hat{\alpha} + \tfrac{1}{2}\hat{\alpha}^2\left(\hat{w}^2 + \hat{w}\sqrt{\hat{w}^2 + 4\hat{\alpha}^{-1}}\right), \tag{41}$$

which simplifies expression (27) in Wood, Booth and Butler. Then $w_{\hat{\xi}}$ is

$$w_{\hat{\xi}} = \tfrac{1}{2}\left(\hat{\alpha}^{-2} - \hat{\xi}^{-2}\right)$$

and

$$\hat{u}_{\hat{\xi}} = \hat{u}\left\{K''_{I\hat{\alpha}}\left(w_{\hat{\xi}}\right)\right\}^{-1/2},$$

where $\hat{u}$ is given in (11) and $K_{I\alpha}$ is the cumulant generating function of γ_α. Our use of the approximation in this form assumes that $K'''(\hat{s}) \geq 0$, which has always been the case when dealing with first passage CGFs. The expression for $\hat{\alpha}$ in (40) must also be positive, which was always the case except in the extreme left tail of W. In such instances, it suffices to simply set $\hat{\alpha}$ equal to its first positive value in the left tail occurring over the time grid used in plotting.

B. *GI*/*M*/1 Queue

Let $W \sim G$ be the interarrival time and Exponential(μ) the fixing distribution. The branch transmittance from state $i = 1, \ldots, n$ to state $j = 2, \ldots, i+1$, is

$$\begin{aligned}\mathcal{Q}_{ij}(s) &= E^W\{\Pr(\text{complete } i - j + 1 \\ &\qquad \text{tasks during time } W\,|W)e^{sW}\} \\ &= E^W\left\{\frac{(\mu W)^{i-j+1}}{(i-j+1)!}e^{-\mu W}e^{sW}\right\},\end{aligned}$$

which leads to the values in (18). Passage to state $j = 1$ is the complementary event to those above and is computed as above by replacing the probability with $1 - \sum_{k=0}^{i-1}\Pr$ (complete k tasks in time $W\,|W$) to give $\mathcal{Q}_{i1}(s) = \mathscr{U}_0(s) - \sum_{j=2}^{i+1}\mathcal{Q}_{ij}(s)$.

C. Partly Loaded Repairable System with Imperfect Switching

The transmittance matrix of this system is derived in the more general setting with n operating units (pumps) instead of 4 and with an equal number of servers. The absorbing state due to switching failure is now state $n+1$. Passage out of state i, in which i failed units are under repair, depends upon (1) how many units can be fixed before the primary unit fails in time $P \sim$ Exponential(λ), and (2) how the backup failure time $B \sim$ Exponential(λ_1) compares to P. Let Z_i denote the number of units fixed prior to primary unit failure. The transition probabilities depend on $g_{i1}(z) = \Pr(Z_i = z, P < B)$, which involves an event with a single switching, and $g_{i2}(z) = \Pr(Z_i = z, P > B)$, an event requiring two switchings. Before deriving each of these, we summarize the transition probabilities in terms of these mass functions. The arguments for getting these probabilities involve two essential issues: (1) accounting for whether $P < B$ or $B < P$ and (2) requiring a single switching (w.p. p) in the former setting and two switchings (w.p. p^2) in the latter.

From state $i = 1, \ldots, n-2$,

$$\mathcal{Q}_{ij}(s) = \begin{cases} pg_{i1}(i)M_\lambda(s), & j = 1, \\ \{pg_{i1}(i-j+1) & \\ \quad + p^2 g_{i2}(i-j+2)\}M_\lambda(s), & j = 2, \ldots, i+1, \\ p^2 g_{i2}(0)M_\lambda(s), & j = i+2, \\ 0, & \text{otherwise}, \\ N(s) = (1-r)pqM_\lambda^2(s), + qM_\lambda(s), & j = n+1. \end{cases}$$

From state 0, $\mathcal{Q}_{00} = 0 = \mathcal{Q}_{03} = \cdots = \mathcal{Q}_{0n}$, $\mathcal{Q}_{01}(s) = prM_\lambda(s)$, and $\mathcal{Q}_{02}(s) = p^2(1-r)M_\lambda(s)$, $\mathcal{Q}_{0,n+1}(s) = N(s)$.

From state $n-1$, $\mathcal{Q}_{n-1,\,0} = 0$ and

$$\mathcal{Q}_{n-1,\,j}(s) = \begin{cases} p^2 g_{n-1}(n-j)M_\lambda(s), & j = 1, \ldots, n-2, \\ pg_{n-1}(1)M_\lambda(s), & j = n-1, \\ g_{n-1}(0)M_\lambda(s), & j = n, \\ pq\left\{\sum_{z\geq 2} g_{n-1}(z)\right\}M_\lambda^2(s) & \\ \quad + q\left\{\sum_{z\geq 1} g_{n-1}(z)\right\}M_\lambda(s), & j = n+1, \end{cases}$$

where $g_i(z) = \Pr\{Z_i = z\} = g_{i1}(z) + g_{i2}(z)$.

From state n, $\mathcal{Q}_{nj} = 0$ for $j \neq n-1, n+1$; $\mathcal{Q}_{n,n-1}(s) = pM_{n\mu}(s)$; and $\mathcal{Q}_{n,n+1}(s) = qM_{n\mu}(s)$.

Derivation of $g_{i1}(z)$ and $g_{i2}(z)$ for $z = 0, \ldots, i$: Use $Z_i|P \sim$ Binomial$(i, 1 - e^{-\mu P})$ so that

$$\begin{aligned} g_{i1}(z) &= E^P\{\Pr(Z_i = z|P)\Pr(B > P|P)\} \\ &= \int_0^\infty \binom{i}{z}(1 - e^{-p\mu})^z(e^{-p\mu})^{i-z}e^{-p\lambda_1}\lambda e^{-p\lambda}\,dp \qquad (42) \\ &= \binom{i}{z}\lambda\sum_{k=0}^{z}(-1)^k\binom{z}{k}\{(i+k-z)\mu + \lambda_1 + \lambda\}^{-1} \end{aligned}$$

from the Binomial formula. The same argument but without the term $\Pr(B > P|P) = e^{-p\lambda_1}$ gives

$$g_i(z) = \binom{i}{z}\lambda\sum_{k=0}^{z}(-1)^k\binom{z}{k}\{(i+k-z)\mu+\lambda\}^{-1},$$

which is (43) with λ_1 set to 0. Now $g_{i2}(z) = g_i(z) - g_{i1}(z)$.

The Markov representation of this system, when state $n+2$ is included (state 6 in the numerical example) is the same as the above with the following qualifications:

$$\left.\begin{aligned}\mathscr{Q}_{i,\,n+2} &= (1-r)pqM_\lambda,\\ \mathscr{Q}_{i,\,n+1} &= qM_\lambda,\end{aligned}\right.\qquad i = 0,\dots,n-2,$$

$$\mathscr{Q}_{n-1,n+2} = pq\left\{\sum_{z\geq 2} g_{n-1}(z)\right\}M_\lambda(s)$$

and

$$\mathscr{Q}_{n-1,\,n+1} = q\left\{\sum_{z\geq 1} g_{n-1}(z)\right\}M_\lambda(s).$$

Passage out of state $n+2$ is only to state $n+1$ with transmittance M_λ.

D. Legitimacy of Lumping Destination States

Create state $n+1$ to which the system passes instantaneously from any state in D. Now consider the process with $n+1$ states characterized by the transmittance

$$\mathscr{Q}^*(s) = \begin{pmatrix}\mathscr{Q}_{CC}(s) & \mathscr{Q}_{CD}(s) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{1}\\ \mathbf{0}^T & \mathbf{0}^T & 0\end{pmatrix},$$

where the blocks correspond to C, D, and $n+1$ with dimensions $m, n-m$, and 1 and bold indicates a vector or matrix of zeros or ones. Since passage is instantaneous and certain to state $n+1$ from D, the waiting time for passage $1 \to n+1$ in this system is the same as passage from $1 \to D$ in the original n-state system. Formally,

$$f_{1D}\mathscr{F}_{1D}(s) = f^*_{1,\,n+1}\mathscr{F}^*_{1,\,n+1}(s),$$

where stars refer to characteristics of the $(n+1)$-state system. From (4),

$$f^*_{1,\,n+1}\mathscr{F}^*_{1,\,n+1}(s) = \frac{(-1)^{n+2}\begin{vmatrix}\{I_m-\mathscr{Q}_{CC}(s)\}_{\backslash 1} & -\mathscr{Q}_{CD}(s) & \mathbf{0}\\ \mathbf{0} & I_{n-m} & \mathbf{-1}\end{vmatrix}}{\begin{vmatrix}I_m-\mathscr{Q}_{CC}(s) & -\mathscr{Q}_{CD}(s)\\ \mathbf{0} & I_{n-m}\end{vmatrix}}.$$

To evaluate the numerator, we move the last column and place it as the first column; this entails $n-1$ column exchanges with a sign change for each interchange. Thus

$$\begin{aligned}f^*_{1,\,n+1}\mathscr{F}^*_{1,\,n+1}(s) &= (-1)^{n+2}\times(-1)^{n-1}\\ &\quad\times\frac{\begin{vmatrix}\mathbf{0} & \{I_m-\mathscr{Q}_{CC}(s)\}_{\backslash 1} & -\mathscr{Q}_{CD}(s)\\ \mathbf{-1} & \mathbf{0} & I_{n-m}\end{vmatrix}}{|I_m-\mathscr{Q}_{CC}(s)|}\\ &= \frac{(-1)\left|\,[\mathbf{0}\quad\{I_m-\mathscr{Q}_{CC}(s)\}_{\backslash 1}]+\mathscr{Q}_{CD}(s)(\mathbf{-1},\mathbf{0})\,\right|}{|I_m-\mathscr{Q}_{CC}(s)|},\end{aligned}\tag{43}$$

where the numerator determinant has been evaluated using its block structure. Adding the block components in the numerator of the last expression leads to the multiple state transmittance in (7).

ACKNOWLEDGMENTS

This research was supported by NSF Grants DMS-93-04274 and DMS-96-25396. The author is grateful for the help and comments of Robert Paige, James Booth and Douglas Bronson.

REFERENCES

AALEN, O. O. (1995). Phase type distributions in survival analysis. *Scand. J. Statist.* **22** 447–463.

ABRAMOWITZ, M. and STEGUN, I. A. (1970). *Handbook of Mathematical Functions*, 9th ed. Dover, New York.

ATKINSON, A. C. (1982). The simulation of generalized inverse Gaussian and generalized hyperbolic random variables. *SIAM J. Sci. Comput.* **3** 502–515.

BOOTH, J. G. (1994). A note on the accuracy of two saddlepoint tail approximations. In *Proceedings of the Section on Physical and Engineering Sciences* 56–58. Amer. Statist. Assoc., Alexandria, VA.

BOOTH, J. G. and WOOD, A. T. A. (1995). An example in which the Lugannani and Rice saddlepoint formula fails. *Statist. Probab. Lett.* **23** 53–61.

BUTLER, R. W. (1997a). System reliability, flowgraphs, and saddlepoint approximation. Technical report, Dept. Statistics, Colorado State Univ.

BUTLER, R. W. (1997b). First passage distributions in semi-Markov processes and their saddlepoint approximation. In *Data Analysis from Statistical Foundations* (E. Saleh, ed.). Nova Science Publishers, Huntington, NY. To appear.

BUTLER, R. W. and BRONSON, D. A. (2000). Bootstrapping survival times in stochastic systems using saddlepoint approximations. Technical report, Dept. Statistics, Colorado State Univ.

BUTLER, R. W. and HUZURBAZAR, A. V. (1993). Prediction in stochastic networks. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session, Firenze*. Book 3, Topic 23.

BUTLER, R. W. and HUZURBAZAR, A. V. (1995). Bayesian prediction of waiting times in stochastic models. *Canad. J. Statist.* To appear.

BUTLER, R. W. and HUZURBAZAR, A. V. (1997). Stochastic network models for survival analysis. *J. Amer. Statist. Assoc.* **92** 246–257.

DANIELS, H. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631–650.

DANIELS, H. (1987). Tail probability approximations. *Internat. Statist. Rev.* **55** 37–48.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**. Wiley, New York.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

HOWARD, R. A. (1964). System analysis of semi-Markov processes. *IEEE Trans. Military Electronics* **8** 114–124.

HOWARD, R. A. (1971). *Dynamic Probabilistic Systems* **2**: *Semi-Markov and Decision Processes*. Wiley, New York.

HØYLAND, A. and RAUSAND, M. (1994). *System Reliability Theory, Models and Statistical Methods*. Wiley, New York.

KARLIN, S. and TAYLOR, H. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

LUGANNANI, R. and RICE, S. (1980). Saddlepoint approximations for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490.

MASON, S. J. (1953). Feedback theory: Some properties of signal flow graphs. *Proc. Institute of Radio Engineers* **41** 1144–1156.

MASON, S. J. (1956). Feedback theory—Further properties of signal flow graphs. *Proc. Institute of Radio Engineers* **44** 920–926.

NEUTS, M. F. (1981). *Matrix-geometric Solutions in Stochastic Models.* Johns Hopkins Univ. Press.

PHILLIPS, C. L. and HARBOR, R. D. (1996). *Feedback Control Systems.* Prentice Hall, Englewood-Cliffs, New Jersey.

PYKE, R. (1961). Markov renewal processes with finitely many states. *Ann. Math. Statist.* **32** 1243–1259.

WHITEHOUSE, G. (1983). Flowgraph analysis. *Encyclopedia of Statistical Science* **3**. Wiley, New York.

WOOD, A. T. A., BOOTH, J. G. and BUTLER, R. W. (1993). Saddlepoint approximations to the CDF for some statistics with non-normal limit distributions. *J. Amer. Statist. Assoc.* **88** 680–686.

ZHOU, M. C., WANG, C.-H. and ZHAO, X. (1995). Automating Mason's rule and its application to analysis of stochastic Petri nets. *IEEE Trans. Control Systems Tech.* **3** 238–244.