QUICK SIMULATION METHOD FOR EXCESSIVE BACKLOGS IN NETWORKS OF QUEUES

by

Shyam Parekh and Jean Walrand

Memorandum No. UCB/ERL M86/74

8 September 1986

QUICK SIMULATION METHOD FOR EXCESSIVE BACKLOGS IN NETWORKS OF QUEUES

by

Shyam Parekh and Jean Walrand

Memorandum No. UCB/ERL M86/74

8 September 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

QUICK SIMULATION METHOD FOR EXCESSIVE BACKLOGS IN NETWORKS OF QUEUES

by

Shyam Parekh and Jean Walrand

Memorandum No. UCB/ERL M86/74

8 September 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Quick Simulation Method for Excessive Backlogs in Networks of Queues[†]

Shyam Parekh and Jean Walrand

Department of Electrical Engineering and Computer Sciences
and Electronics Research Laboratory
University of California, Berkeley, CA 94720

ABSTRACT

We consider stable open Jackson networks and study the rare events of excessive backlogs. Although, these events occur rarely they can be critical, since they can impair the functioning of the network. We attempt to estimate the probability of these events by simulations. Since, the direct simulation of rare events takes a very long time, this procedure is very costly. Instead, we devise a method for changing the network to speed up the simulation of rare events. We try to pursue this idea with the help of Large Deviation theory. This approach, under certain assumptions, results in a system of differential equations which may be difficult to solve. To circumvent this, we develop a heuristic method which gives the rule for changing the network for the purpose of simulations. We illustrate, by examples, that our method of simulations can be several orders of magnitude faster than direct simulations.

September 8, 1986

[†]This research was supported in part by NSF Grant No. ECS. 8421128 and by Pacific Bell and a MICRO Grant from the state of California.

Quick Simulation Method for Excessive Backlogs in Networks of Queues[†]

Shyam Parekh and Jean Walrand

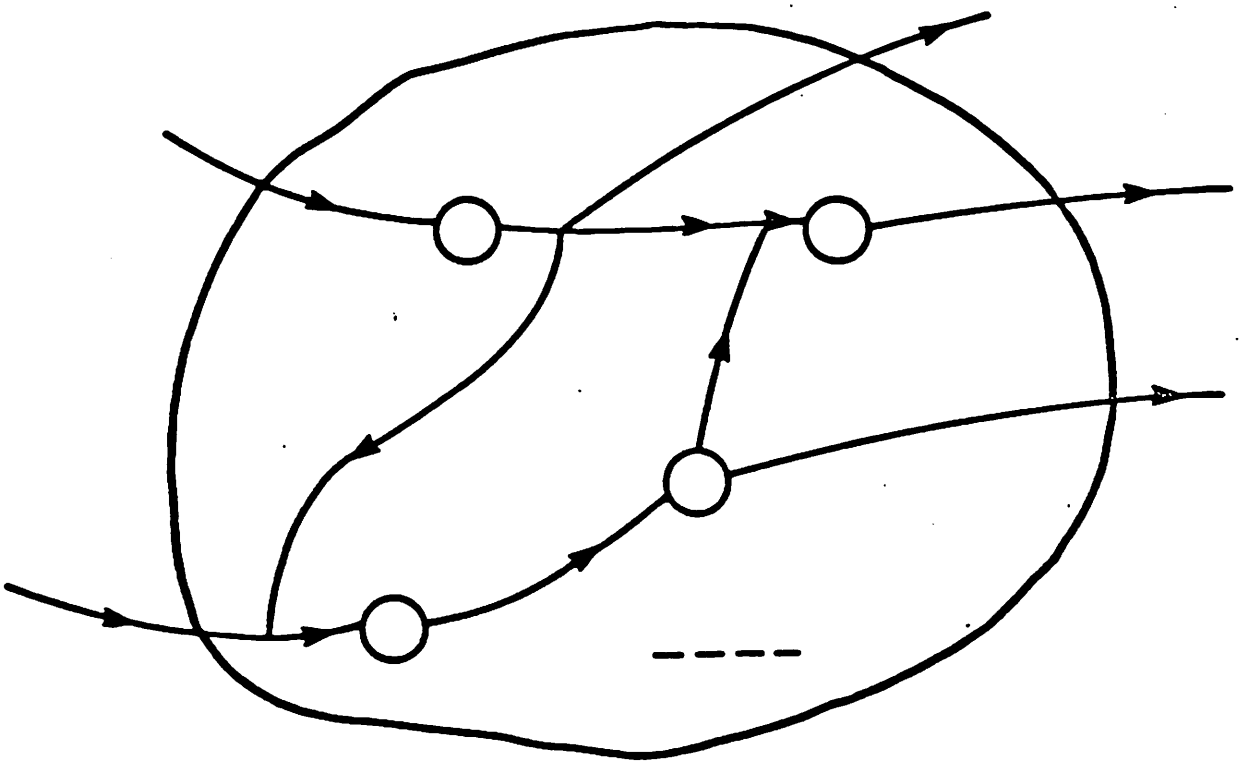
Department of Electrical Engineering and Computer Sciences
and Electronics Research Laboratory
University of California, Berkeley, CA 94720

1. Introduction.

1.1. Problem Description.

We consider arbitrary open Jackson networks (e.g. Figure (1)). A Jackson network is an interconnection of $M/M/1$ queues in which customers visit various nodes according to state and time independent (Markovian) routing probabilities. The heuristic that will be developed can be applied to networks of $GI/G/1$ queues with Markovian routing. However, most of the discussion will be limited to the case of Jackson networks, for simplicity. A network is called open, if every arriving customer leaves the system with probability 1. Let us define T as the first time that the total population in the network reaches N. In the last section we also will consider queues with finite buffers. In this case, we define T as the first time one of the queues gets full. (This case will be considered only in the last section.) In either case, T is a first passage time for the Markov state process of the network. We are interested in estimating $E_0\{T\}$, where $E_0\{T\}$ denotes the expected value of T given that the system starts empty. Notice that we are interested in the transient behavior of the system.

[†]This research was supported in part by NSF Grant No. ECS. 8421128 and by Pacific Bell and a MICRO Grant from the state of California.



Open Jackson Network.

Figure (1)

Since very little is known about the transient behavior of networks, we attempt to estimate $E_0\{T\}$ by efficient simulations. Our method of simulation, besides saving simulation time, also sheds some light on the fundamentals of the dynamics of the system.

1.2. Principle (Importance Sampling).

If the total backlog N or the buffers' capacities are large, for a stable system, the events of exceeding the system's capacity are very infrequent. Hence, direct simulations are very slow and take up a lot of computer time. Besides, there is also the difficulty of having a pseudo-random generator that can function effectively during very long simulations. The central idea is to make the rare events under investigation more frequent by changing appropriately the probability measures governing the system and performing simulations on the changed system. We then obtain our answers by translating them back to the

original system. This is done by using likelihood ratios.

1.3. Optimal Change of Measure (Largest Speed-up).

Large Deviation Theory deals with certain Markov processes and determines the asymptotic (e.g. as the backlog size N grows for an $M/M/1$ queue. See § 3) exponential rate of diminishing probabilities as a solution of a variation problem. The solution of this variation problem also gives the optimal exponential change of measure (see § 3) for simulations. An application-oriented and readable reference for this work is by Cottrell et al. [2]. Unfortunately the theory does not apply to general Jackson networks. Here a smoothness condition regarding the jump distributions (see § 3) is violated. To our knowledge, there are no known results of Large Deviation Theory for excursions of Markov processes with discontinuous kernels. To circumvent this problem, we are going to have to rely on a heuristic of Borovkov, Ruget etc. (e.g., see [3]) which gives certain tail probabilities for a $GI/G/1$ queue (see § 4). We utilize this heuristic for obtaining a change of measure that leads to substantial speed-up for simulations. We also generalize this heuristic to networks.

1.4. Outline of the Remaining Sections.

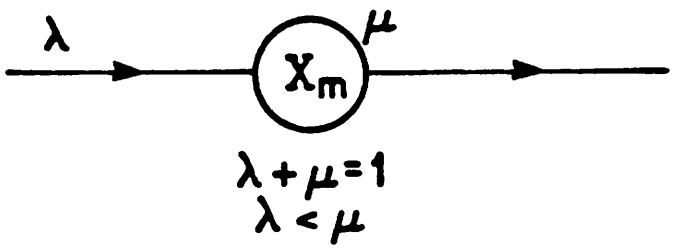
In § 2, we motivate the idea of change of measure for simulations of rare events for an $M/M/1$ queues. In § 3, we present Large Deviation theory as applied for the purpose of simulations of rare events. We also point out the difficulties in applying this theory to Jackson networks. In § 4, we present a heuristic method for obtaining an optimal change of measure for simulations of rare events for Jackson networks. Next we extend this heuristic to networks of $GI/G/1$ queues. We also apply this heuristic for the networks with finite buffers for estimating the buffer-overflow probabilities. The purpose of this section is to report the observations we have made rather than to claim new results. Our hope is that the heuristic explanations and observations presented here will motivate more research in this area. Finally, we will summarize the

results of this paper in § 5.

2. $M/M/1$ Example.

2.1. Model and Problem.

Consider an $M/M/1$ queue with arrival rate λ and service rate μ such that $\lambda < \mu$. Consider the embedded discrete time Markov chain $\{X_m, m=0,1,2,\cdots\}$, denoting the queue length, defined at the epochs of arrivals and departures of the queue. We assume, without any loss of generality, $\lambda+\mu=1$ (if not, rescale time). Figure (2) depicts such a queue. As described in § 1.1, we are interested in estimating, for large N, $E_0\{T\}$ where T denotes the first time $\{X_m\}$ reaches N.



$M/M/1$ queue.
Figure (2)

Note that the number of times $\{X_m\}$ returns to 0 before hitting N is geometrically distributed with parameter $1-\alpha$, where α is the probability that $\{X_m\}$ reaches N before returning to 0 given that it starts from 0. For large N, we can argue that

$$E_0\{T\} \approx E\{R\} \cdot E_0\{T_0\} = \frac{1-\alpha}{\alpha} \cdot E_0\{T_0\} \approx \frac{1}{\alpha} \cdot E_0\{T_0\},$$

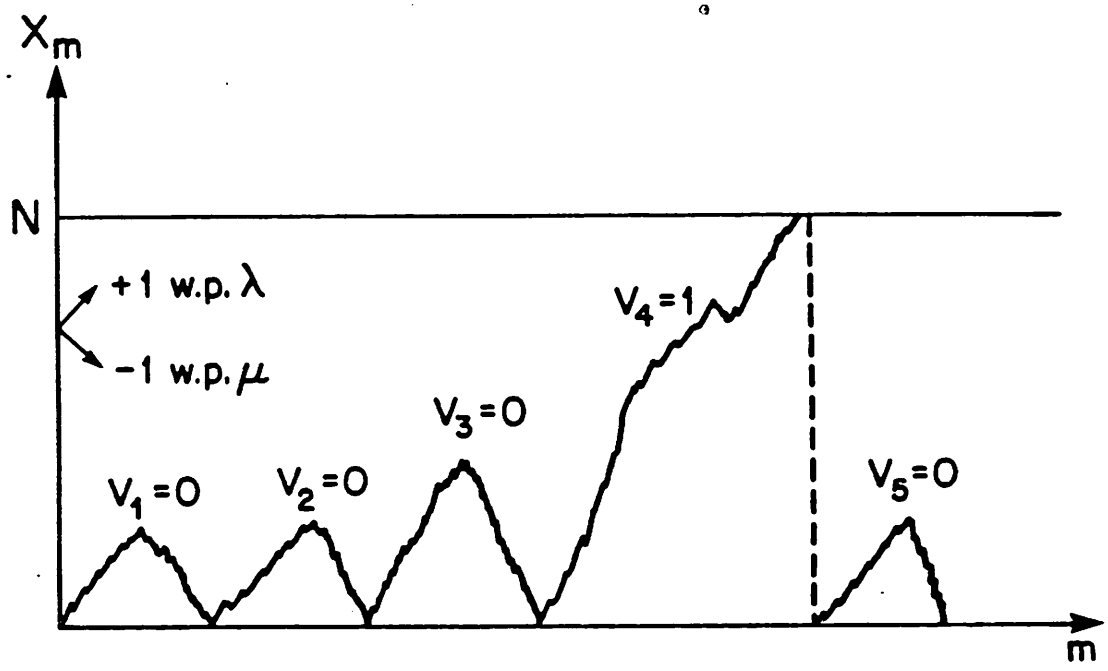
where T_0 denotes the time to hit 0 for the first time. Since, for stable systems, $E_0\{T_0\}$ can be easily estimated by direct simulations, the difficult part in estimating $E_0\{T_0\}$ is the estimation of α. So, from now on, our primary concern

will be the estimation of α.

We define a cycle as the duration starting with an empty system and ending at the instant the system, for the first time, either becomes empty again or reaches N. Let us define

$$V_k := 1\{X_m \text{ reaches } N \text{ in cycle } k\},$$

where $1\{B\}$ (or sometimes written 1_B) denotes the indicator of an event B. See Figure (3).



Realization of $\{X_m\}$.

Figure (3)

Notice that V_k's are i.i.d.. Also notice that, as shown in Figure (3), we have modified $\{X_m\}$ in that we restart $\{X_m\}$ at 0, if it exceeds N. Clearly, $\alpha = P\{V_k = 1\}$. Here we can find α by the first step method. For this define, for $0 \leq i \leq N$, $P_i =$ probability that $\{X_m\}$ hits N before 0 given that it starts from i.

Clearly, $P_0 = 0$, $P_N = 1$ and $P_1 = \alpha$. The first step equations give

$$P_i = \mu.P_{i-1}+\lambda.P_{i+1}, \quad 1 \leqslant i \leqslant N-1.$$

The solution of these linear equations can be seen to give

$$\alpha = P_1 = \frac{\frac{\mu}{\lambda}-1}{(\frac{\mu}{\lambda})^N-1}. \tag{1}$$

For future calculations, let us derive the formula for $E\{J_k\}$, where J_k denotes the number of random jumps in cycle k. Notice that J_k's are i.i.d. and that a cycle begins with a deterministic transition to 1. Let Z_i denote a jump which takes values +1 and −1 w.p. λ and μ respectively. Note that cycle k ends at N with probability α and in this case $Z_1+Z_2+\cdots+Z_{J_k} = N-1$. Similarly, cycle k ends at 0 with probability $1-\alpha$ and in this case $Z_1+Z_2+\cdots+Z_{J_k} = -1$. Then, for cycle k,

$$E\{Z_1+Z_2+\cdots+Z_{J_k}\} = \alpha.(N-1)+(1-\alpha).(-1).$$

Using Wald's identity we identify the left hand side with

$$E\{J_k\}.E\{Z_i\} = E\{J_k\}.(\lambda-\mu).$$

This gives

$$E\{J_k\} = \frac{1-N.\alpha}{\mu-\lambda}. \tag{2}$$

In the following subsections we present the idea of change of measure for estimating α by simulation.

2.2. Direct Simulation.

For direct Monte Carlo simulation, consider an unbiased and convergent estimator

$$\alpha_n := \frac{V_1+V_2+\cdots+V_n}{n}.$$

Observe that $E\{V_k\} = \alpha$ and $Var\{V_k\} = \alpha.(1-\alpha)$.

Suppose we want to ensure that the relative error does not exceed ϵ% with probability more than β. We will call such an estimator an (ϵ,β)-confidence estimator. The normal approximation then gives

$$P\{|\alpha_{n_d}-\alpha| > \epsilon.\alpha\} \approx \beta \iff n_d \approx \frac{c^2}{\epsilon^2} \cdot \frac{Var\{V_k\}}{\alpha^2},$$

where $c = \Phi^{-1}(\beta/2)$, where Φ denotes the distribution function of a Gaussian r.v. with the mean equal to 0 and variance equal to 1. Hence $n_d \approx \gamma.(1-\alpha)/\alpha$, where $\gamma = c^2/\epsilon^2$, cycles are necessary to achieve the (ϵ,β)-confidence estimator by a direct simulation. Let T_d denote the units of simulation time required for achieving the (ϵ,β)-confidence estimator by a direct simulation. Then,

$$T_d = E\{J_k\}.n_d.$$

Since $\lambda < \mu$, for large N, $E\{J_k\} \approx 1/\mu-\lambda$ (see Eqn. (2)). Hence,

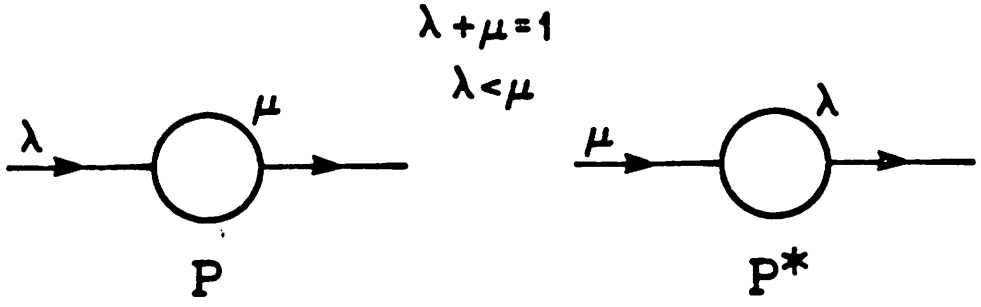
$$T_d \approx \gamma.\frac{1}{\alpha}.\frac{1}{\mu-\lambda}. \tag{3}$$

2.3. Change of Measure.

For estimating α, we propose to consider the $M/M/1$ queue with arrival rate μ and service rate λ i.e. the $M/M/1$ queue obtained by interchanging arrival rate and service rate of the original queue. Let P and P^* denote the measures induced by the corresponding Markov chains. Figure (4) shows these queues.

In simulations under the changed measure, we observe V_k's under P^*. Let L_k denote the likelihood ratio dP/dP^* during cycle k. Notice that L_k's are i.i.d. and that $E^*\{L_k.V_k\} = E\{V_k\} = \alpha$, where $E^*\{\}$ denotes the expectation under the measure P^*. Hence,

$$\alpha_n^* := \frac{L_1.V_1+L_2.V_2+\cdots+L_n.V_n}{n}$$



Change of Measure for an $M/M/1$ Queue.

Figure (4)

is also an unbiased and convergent estimator of α. As before, to achieve (ϵ,β)-confidence estimator, now the minimum number of cycles required will be

$$n_c \approx \gamma.\frac{Var^*\{L_k.V_k\}}{\alpha^2},$$

where $Var^*\{\ \}$ denotes the variance under the measure P^*. Observe that by interchanging λ and μ in Eqn. (2), we have $E^*\{J_k\} \approx N/\mu$. Let T_c denote the units of simulation time required for achieving the (ϵ,β)-confidence estimator under the changed measure. Then,

$$T_c = E^*\{J_k\}.n_c \approx \gamma.\frac{\sigma^2}{\alpha^2}.\frac{N}{\mu}, \tag{4}$$

where $\sigma^2 := Var^*\{L_k.V_k\}$.

We should point out that in reality the simulation time will be somewhat larger like $(1+\delta).T_c$, where $\delta > 0$ accounts for the time required to calculate likelihood ratios L_k's.

2.4. Comparison of T_d and T_c.

Let us define the speed-up factor $S := \frac{T_d}{T_c}$. From Eqns. (3) and (4) we get

$$S \approx \frac{1}{N}.\frac{\alpha}{\sigma^2}.\frac{1}{1-\frac{\lambda}{\mu}}. \tag{5}$$

Suppose that ω is a realization such that $V_k = 1$ and there are l departures and $N+l-1$ arrivals (not counting the first arrival) during cycle k. So, $J_k(\omega) = N+2l-1$. Let ω_k denote the section of ω that pertains to cycle k. Then, $P\{\omega_k\} = \lambda^{N+l-1}.\mu^l$ and $E^*\{\omega_k\} = \mu^{N+l-1}.\lambda^l$. Therefore,

$$L_k(\omega_k) = (\frac{\lambda}{\mu})^{N-1}.$$

This implies that, on the set $\{V_k = 1\}$,

$$L_k \equiv (\frac{\lambda}{\mu})^{N-1}. \tag{6}$$

Hence,

$$\begin{aligned}\sigma^2 &= E^*\{(L_k.V_k)^2\} - \alpha^2 \\ &= (\frac{\lambda}{\mu})^{N-1}.E^*\{L_k.V_k\} - \alpha^2 \\ &= (\frac{\lambda}{\mu})^{N-1}.\alpha - \alpha^2,\end{aligned}$$

where the second equality follows from Eqn. (6). Now using Eqn. (1), we get

$$\frac{\sigma^2}{\alpha} \approx (\frac{\lambda}{\mu})^N. \tag{7}$$

Substituting Eqn. (7) in Eqn. (5), we get

$$S \approx [N.(\frac{\lambda}{\mu})^N.(1-\frac{\lambda}{\mu})]^{-1}.$$

2.5. Example.

Consider the $M/M/1$ queue with $\lambda = 0.33$ and $\mu = 0.67$. We want to estimate α for $N = 21$. Eqn. (1) gives $\alpha = 3.583\times10^{-7}$. For $(\epsilon = 0.05, \beta = 0.05)$-confidence estimator, Eqn. (3) gives $T_d = 1.32\times10^{10}$ units (4.42×10^9 cycles), while Eqn. (4) gives $T_c = 4.96\times10^4$ units (1.58×10^3 cycles).

Our simulation experiments gave the following results.

# of cycles (n)	1000	2000	10000
α_n	0.0	0.0	0.0
α_n^*	3.440×10^{-7}	3.520×10^{-7}	3.708×10^{-7}

Example of Change of Measure.
Table (1)

Table (2) gives results of a few more simulation experiments. It also shows the time required for simulations and the corresponding number of calls to the random number generator (RNG). Table (3) gives the empirical standard deviations, means and coefficients of variation of the estimates obtained by the change of measure for the same examples as in Table (2). All the simulations are done on a VAX-750 machine. Notice that the convergence under the changed measure seems to be more rapid than predicted by Eqn. (4). This is due to the uncertainty factor introduced in the derivation of Eqn. (4) because of the use of the normal approximation.

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda = 0.20\ \mu = 0.80\ N = 15$ $\alpha = 2.794\times10^{-9}$ $\lambda^* = 0.80\ \mu^* = 0.20$						
# of Cycles (n)	5000	10000	20000	50	100	200
$\alpha_n\ (\alpha_n^*)$	0.0	0.0	0.0	2.831×10^{-9}	2.682×10^{-9}	2.663×10^{-9}
CPU Time	2.5*Sec.*	5.4*Sec.*	10.6*Sec.*	0.3*Sec.*	0.6*Sec.*	1.2*Sec.*
Calls to RNG	8550	16712	33624	800	1656	3255
Example-II $\lambda = 0.30\ \mu = 0.70\ N = 20$ $\alpha = 5.826\times10^{-8}$ $\lambda^* = 0.70\ \mu^* = 0.30$						
# of Cycles (n)	5000	10000	20000	200	300	500
$\alpha_n\ (\alpha_n^*)$	0.0	0.0	0.0	6.322×10^{-8}	5.268×10^{-8}	5.955×10^{-8}
CPU Time	3.9*Sec.*	7.6*Sec.*	16.1*Sec.*	2.0*Sec.*	2.4*Sec.*	4.6*Sec.*
Calls to RNG	12492	25426	51052	5598	7084	13684
Example-III $\lambda = 0.40\ \mu = 0.60\ N = 30$ $\alpha = 2.608\times10^{-6}$ $\lambda^* = 0.60\ \mu^* = 0.40$						
# of Cycles (n)	20000	30000	40000	1000	2000	3000
$\alpha_n\ (\alpha_n^*)$	0.0	0.0	0.0	2.910×10^{-6}	2.401×10^{-6}	2.649×10^{-6}
CPU Time	30.2*Sec.*	44.4*Sec.*	56.2*Sec.*	16.4*Sec.*	30.4*Sec.*	43.2*Sec.*
Calls to RNG	105738	151322	195760	47466	82956	127832

Simulations for an $M/M/1$ Queue.

Table (2)

Example-I $\lambda = 0.20$ $\mu = 0.80$ $N = 15$ $\alpha = 2.794\times10^{-9}$ *# of Experiments* = 20 $\lambda^* = 0.80$ $\mu^* = 0.20$		
# of Cycles (n)	100	200
Empirical Mean ($\hat{m}$)	2.744×10^{-9}	2.794×10^{-9}
Empirical Std. Dev. ($\hat{\sigma}$)	1.150×10^{-10}	1.019×10^{-10}
($\hat{\sigma}$/ $\hat{m}$)×100 %	4.1910 %	3.645 %
Example-II $\lambda = 0.70$ $\mu = 0.30$ $N = 20$ $\alpha = 5.826\times10^{-8}$ *# of Experiments* = 20 $\lambda^* = 0.70$ $\mu^* = 0.30$		
# of Cycles (n)	300	500
Empirical Mean ($\hat{m}$)	5.856×10^{-8}	5.906×10^{-8}
Empirical Std. Dev. ($\hat{\sigma}$)	2.803×10^{-9}	2.474×10^{-9}
($\hat{\sigma}$/ $\hat{m}$)×100 %	4.786 %	4.190 %
Example-III $\lambda = 0.40$ $\mu = 0.60$ $N = 30$ $\alpha = 2.608\times10^{-6}$ *# of Experiments* = 20 $\lambda^* = 0.60$ $\mu^* = 0.40$		
# of Cycles (n)	2000	3000
Empirical Mean ($\hat{m}$)	2.743×10^{-6}	2.680×10^{-6}
Empirical Std. Dev. ($\hat{\sigma}$)	2.652×10^{-7}	2.409×10^{-7}
($\hat{\sigma}$/ $\hat{m}$)×100 %	9.669 %	8.989 %

Empirical Standard Deviation for an $M/M/1$ Queue.

Table (3)

3. Large Deviation Theory and Optimal Change of Measure.

3.1. A Fundamental Theorem.

Theorem (1) (Cramér's Theorem) [5] : Let $\xi_1, \xi_2, \cdots$ be i.i.d. r.v.'s taking values in $\mathbb{R}^d$. Let F denote the distribution function (d.f.) of ξ_k and m its mean. Let P_n denote the d.f. of $(\xi_1+\xi_2+\cdots+\xi_n)/n$. We assume that the

Laplace transform of F,

$$M(s) := \int_{\mathbf{R}^d} \exp<s,z> dF(z), \quad s \in \mathbf{R}^d,$$

is finite in a neighborhood of 0. Then, P_n satisfies

(i) For each closed subset C of $\mathbf{R}^d$,

$$\limsup_{n \to \infty} \frac{1}{n}.\log P_n\{C\} \leqslant -\inf_{x \in C} h(x)$$

and

(ii) For each open subset G of $\mathbf{R}^d$,

$$\liminf_{n \to \infty} \frac{1}{n}.\log P_n\{G\} \geqslant -\inf_{x \in G} h(x),$$

where the function h, called Cramér or Legendre transform, is defined as

$$h(y) = \sup_{s \in \mathbf{R}^d} [<s,y> - \log M(s)], \quad y \in \mathbf{R}^d. \tag{8}$$

(Unless specified, logarithm is always defined to the base e.)

Interested readers can find a lucid proof of this theorem in the succinct monograph by Varadhan [5]. This theorem gives the rate of convergence for the Weak Law of Large Numbers (WLLN). This is quite easily seen from an equivalent statement of this theorem in $\mathbf{R}^1$. For this, let $a > m$, then,

$$\lim_{n \to \infty} \frac{1}{n}.\log P\{\frac{\xi_1+\xi_2+\cdots+\xi_n}{n} > a\} = -h(a). \tag{9}$$

Intuitively, Eqn. (9) states that $P\{S_n / n \approx a\} = \exp(-n.h(a)+o(n))$, where $S_n = \xi_1+\xi_2+\cdots+\xi_n$. The exponential rate of decay can be expected from the observation that if the event $\{(\xi_1+\cdots+\xi_{2n})/ 2n > a\}$ were to occur, it is most likely that it happens if and only if events $\{(\xi_1+\cdots+\xi_n) > a.n\}$ and $\{(\xi_{n+1}+\cdots+\xi_{2n}) > a.n\}$ occur. Hence, $P\{S_{2n} / 2n > a\} \approx (P\{S_n / n > a\})^2$.

Next we list some properties of the Cramér transform defined in Eqn. (8). We define

$$l(s) := \log M(s),\ s \in \mathbf{R}^d.$$

(P1) h is convex and nonnegative lower semicontinuous.

(P2) For each $b < \infty$, the set $\{u \ / \ h(u) \leqslant b\}$ is compact in $\mathbf{R}^d$.

(P3) $h(y)$ has its minimum value 0 at $y = m$, i.e., $h(m) = 0$.

(P4) l and h are convex dual of each other,

$$l(s) = \sup_u [<s,u> - h(u)].$$

(P5) Let V denote the interior of the set $\{s \in \mathbf{R}^d \ / \ M(s) < \infty\}$ and U denote the set $\{u \in \mathbf{R}^d \ / \ h(u) < \infty\}$. The derivatives h' and l' are reciprocals of each other, i.e.,

$$h'(l'(s)) = s,\ s \in V$$

and

$$l'(h'(u)) = u,\ u \in U.$$

Finally, we give a few examples of the Cramér transform that will be useful to us in the subsequent sections.

(E1) ξ_k's take values +1 and −1 w.p. p_1 and p_2 respectively. Then,

$$h(u) = \frac{1+u}{2}.\log(\frac{1+u}{2.p_1}) + \frac{1-u}{2}.\log(\frac{1-u}{2.p_2}),\ -1 \leqslant u \leqslant 1,$$

$$= \infty,\ \textit{otherwise.} \qquad (10)$$

(E2) ξ_k's are exponentially distributed with the parameter $\nu > 0$. Then,

$$h(u) = \nu.u - 1 - \log(\nu.u),\ u > 0,$$

$$= \infty,\ \textit{otherwise.} \qquad (11)$$

3.2. Slow Markov Walk.

In this subsection we present a Large Deviation Theorem due to Ventsel and Freidlin [6], regarding certain Markov chains. Cottrell et al. [2] have a good discussion on these concepts.

Consider the Markov chain $\{X_n^\epsilon\} \epsilon \mathbf{R}^d$ given by

$$X_0^\epsilon = x_0,$$

$$X_{n+1}^\epsilon = X_n^\epsilon + \epsilon . V(X_n^\epsilon, \xi_n), \quad n \geq 0, \tag{12}$$

where $\epsilon > 0$ is the parameter defining the Markov chain $\{X_n^\epsilon\}$, x_0 is the initial value, $V(.,.)$ is a function from $\mathbf{R}^d \times \mathbf{R}^1 \to \mathbf{R}^d$ and ξ_n's are i.i.d. r.v.'s. We are interested in analyzing $\{X_n^\epsilon\}$ when $\epsilon \to 0$.

Let F_x denote the d.f. of $V(x, \xi_n)$. Let

$$m(x) = \int_{\mathbf{R}^d} z dF_x(z)$$

be the mean of F_x,

$$M_x(s) := \int_{\mathbf{R}^d} \exp < s, z > dF_x(z)$$

be its Laplace transform, $l_x(s) := \log M_x(s)$ and

$$h_x(u) = \sup_{s \epsilon \mathbf{R}^d} [< s, u > - l_x(s)]$$

be its Cramér transform. We assume that

(A1) $M_x(s) < \infty$ in a neighborhood of 0 for each $x \epsilon \mathbf{R}^d$.

(A2) $d(F_{x_1}, F_{x_2}) \leq c. \| x_1 - x_2 \|$, where d is the Prohorov distance [1] and $c > 0$ is a constant, i.e., F_x is Lipschitz smooth in x.

Let us construct continuous time paths from the realizations of $\{X_n^\epsilon\}$. To do this, at the epochs

$$t = n.\epsilon \quad \textit{define} \quad X^\epsilon(t) := X_n^\epsilon \tag{13}$$

and interpolate piecewise linearly. Let $\mathbf{C}_T$ denote the set of the continuously piecewise differentiable functions $\phi : [0,T] \to \mathbf{R}^d$ such that $\phi(0) = x_0$ is fixed. Let

P^ϵ denote the measure induced by the Markov chain $\{X_n^\epsilon\}$ on the Borel σ-field Σ of C_T endowed with the Skorohod topology [1]. It is well known that, under some technical assumptions, the deterministic trajectory $\bar{\phi}(t)$ which solves

$$\frac{d\bar{\Phi}}{dt}(t) = m(\bar{\phi}(t))$$

$$\bar{\phi}(0) = x_0,$$

satisfies

$$\text{for } all\ \eta > 0, \text{ for } all\ T < \infty,\ P^\epsilon\{ \max_{0 \leqslant n.\epsilon \leqslant T} |X_n^\epsilon - \bar{\phi}(n.\epsilon)| > \eta\} \to 0.$$

Define the action integral

$$I(\phi) := \int_0^T h_{\phi(t)}(\phi'(t))dt.$$

Theorem (2) (Ventsel-Freidlin) [6] : Let ϕ be a path in C_T. Define a tube of diameter d around ϕ, T_d^ϵ, as the set of trajectories of $X^\epsilon(t)$ such that

$$|X^\epsilon(t) - \phi(t)| < d,\ \text{ for } all\ t \in [0,T].$$

Then, there exists δ_0 such that, for $0 < \delta < \delta_0$,

$$\lim_{\epsilon \to 0} (-\epsilon . P^\epsilon\{T_\delta^\epsilon(\phi)\}) = I(\phi) + e(\delta)$$

with

$$\lim_{\delta \to 0} e(\delta) = 0.$$

For establishing this theorem some more assumptions, besides (A1) and (A2), are necessary. However, the assumptions (A1) and (A2) are the most crucial ones. It is easy to see, technicalities aside, why Theorem (2) "makes sense". Notice that

$$\frac{X_{M+m}^\epsilon - X_M^\epsilon}{m} = \frac{1}{m}.\sum_{k=0}^{m-1} X_{M+k+1}^\epsilon - X_{M+k}^\epsilon$$

$$= \frac{\epsilon}{m}.\sum_{k=0}^{m-1} V(X_{M+k}^\epsilon, \xi_{M+k}), \tag{14}$$

where the last equality follows from the definition of $\{X_n^\epsilon\}$ in Eqn. (12). Let us now use the time-scaling defined in Eqn. (13) and denote $t = M.\epsilon$ and $dt = m.\epsilon$. This along with Eqn. (14) gives

$$\frac{X^\epsilon(t+dt)-X^\epsilon(t)}{dt} = \frac{1}{m}.\sum_{k=0}^{m-1} V(X_{M+k}^\epsilon, \xi_{M+k}).$$

Hence,

$$\frac{dX^\epsilon}{dt}(t) \approx \frac{1}{m}.\sum_{k=0}^{m-1} V(X_{M+k}^\epsilon, \xi_{M+k}).$$

Now using the assumption (A2) and Eqn. (12), we can intuitively argue that $V(\ ,\ )$'s on the right hand side are "almost" i.i.d.. Therefore, the set of trajectories having the slope approximately $\phi'(t)$ at time t will have probability (up to logarithmic equivalence)

$$\exp(-\frac{dt}{\epsilon} h_{\phi(t)}(\phi'(t))),$$

by using Theorem (1). Now arguing that the slope dX^ϵ/dt in the interval $(t, t+dt]$ is almost independent of that in the interval $(t+dt, t+2.dt]$, we get

$$\exp(-\frac{1}{\epsilon}.\int_0^T h_{\phi(t)}(\phi'(t))dt + o(\frac{1}{\epsilon}))$$

for the probability (up to logarithmic equivalence) of the set of trajectories having slope "close" to $\phi'(t)$ over the interval $[0,T]$.

Next, we give a consequence of Theorem (2) that will enable us to estimate $P^\epsilon\{S\}$ for $S \epsilon \Sigma$ whose boundary satisfies certain smoothness condition.

Corollary (1) (Ventsel-Freidlin) [6] : Let $S \epsilon \Sigma$ be such that

$$inf\{I(\phi) / \ \phi \in int(S)\} = inf\{I(\phi) / \ \phi \in cl(S)\}, \qquad (15)$$

then,

$$\lim_{\epsilon \to 0} (-\epsilon.\log P^\epsilon\{S\}) = \inf_{\phi \in S} I(\phi).$$

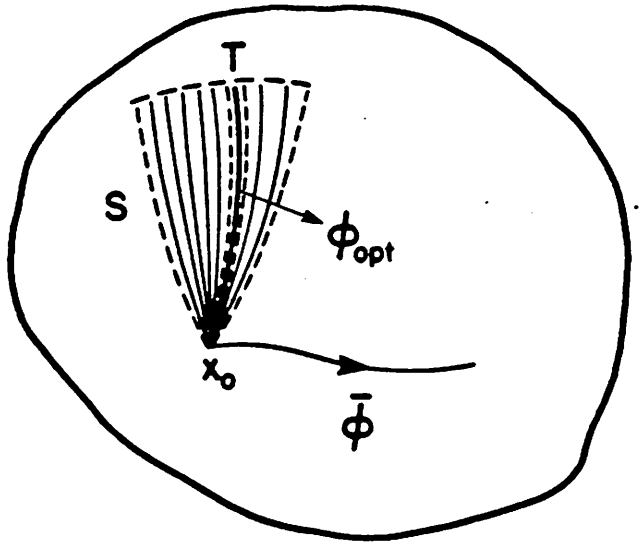
Corollary (1) suggests that

$$P^\epsilon\{S\} \approx \sum_k P^\epsilon\{T_\delta^\epsilon(\phi_k)\}$$

$$\approx \sum_k \exp(-\frac{1}{\epsilon}.I(\phi_k))$$

$$\approx \exp(-\frac{1}{\epsilon}.\inf_{\phi \epsilon S} I(\phi)),$$

where the second approximation follows from Theorem (1) and the last one follow from Corollary (1). Using lower semicontinuity of $I(\phi)$ and the condition in Eqn. (15), it is not difficult to show that $\inf_{\phi \epsilon S} I(\phi)$ is achievable. Let us denote $\operatorname{argmin}_{\phi \epsilon S} I(\phi)$ by ϕ_{opt}. We depict this setup in Figure (5).



Rare Event S and ϕ_{opt}.

Figure (5)

Suppose we are interested in the probability of the set S of trajectories which hit a "rare" set A before hitting 0 given that we start from 0. Assuming that the condition in Eqn (15) is satisfied, we need to find $\inf_{\phi \epsilon S} I(\phi)$. For this, define

$$C(x) := \inf \{ \int_0^{T(\phi)} H(\phi(t), \phi'(t))dt \ / \ \phi(0) = x, \phi \epsilon C, T(\phi) < \infty \} \qquad (16)$$

where $x = (x_{(1)}, \cdots, x_{(d)})$ and $v = (v_{(1)}, \cdots, v_{(d)})$ are vectors in $\mathbf{R}^d$, C denotes the set of the continuously piecewise differentiable functions $\phi : [0,\infty) \rightarrow \mathbf{R}^d$ and $H(\phi(t),\phi'(t)) \equiv h_{\phi(t)}(\phi'(t))$. We denote $l_x(\theta)$ by $L(x,\theta)$. Notice that ϕ_{opt} is the trajectory that achieves the infimum for $C(0)$.

The following result gives a recipe for finding ϕ_{opt}.

Theorem (3) : Assume that $C(x)$ is smooth enough to satisfy

$$\frac{\partial^2 C}{\partial x_{(i)} x_{(j)}} = \frac{\partial^2 C}{\partial x_{(j)} x_{(i)}}.$$

Define

$$\theta_{(i)}(x) = -\frac{\partial C}{\partial x_{(i)}}(x), \quad 1 \leq i \leq d. \tag{17}$$

Then, for each x that is on some $\phi \in S$,

$$L(x,\theta(x)) = 0 \tag{18}$$

and ϕ_{opt} is a solution of the following system of differential equations:

$$\frac{d\theta_{(i)}}{dt} = -\frac{\partial L}{\partial x_{(i)}}(x,\theta), \quad 1 \leq i \leq d, \tag{19}$$

$$\frac{dx_{(i)}}{dt} = \frac{\partial L}{\partial \theta_{(i)}}(x,\theta), \quad 1 \leq i \leq d. \tag{20}$$

Proof : First, we expand $C(x)$ as

$$C(x) = \inf_v \{H(x,v).\Delta t + C(x+v.\Delta t) + o(1)\}$$

$$= \inf_v \{H(x,v).\Delta t + C(x) - \sum_{i=1}^{d} v_{(i)}.\theta_{(i)}(x).\Delta t + o(\Delta t)\},$$

where we have used the definition of θ in Eqn. (17). Cancelling $C(x)$ from both the sides, dividing by Δt and letting $\Delta t \rightarrow 0$, we get

$$\inf_v \{H(x,v) - \sum_{i=1}^{d} v_{(i)}.\theta_{(i)}(x)\} = 0,$$

i.e.,

$$\sup_v \{<\theta,v> - H(x,v)\} = 0. \tag{21}$$

Using Eqn. (21) and the convex duality property (P4) of Cramér transform, § 3.1, we get

$$L(x,\theta(x)) = 0.$$

Suppose that the supremum in Eqn. (21) is achieved at $\bar{v}$, then by differentiating, we get

$$\theta_{(i)} = \frac{\partial H}{\partial v_{(i)}}(x,\bar{v}).$$

Now using the reciprocity property of l' and h', property (P5) of the Cramér transform, § 3.1, along ϕ_{opt}, we get

$$\bar{v}_{(i)} = \frac{\partial L}{\partial \theta_{(i)}}(x,\theta) \tag{22}$$

Observe from Eqn. (18) that

$$0 = \frac{dL}{dx_{(i)}}(x,\theta(x))$$

$$= \frac{\partial L}{\partial x_{(i)}}(x,\theta) + \sum_{k=1}^{d} \frac{\partial L}{\partial \theta_{(k)}}(x,\theta).\frac{\partial \theta_{(k)}}{\partial x_{(i)}}(x) \tag{23}$$

But, along ϕ_{opt},

$$\frac{d\theta_{(i)}}{dt} = \sum_{k=1}^{d} \frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x).\frac{dx_{(k)}}{dt}$$

$$= \sum_{k=1}^{d} \frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x).\frac{\partial L}{\partial \theta_{(k)}}(x,\theta), \tag{24}$$

by using Eqn. (22). Now by the assumption regarding smoothness of $C(x)$ and the definition of *theta* (x) in Eqn. (17), we get

$$\frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x) = \frac{\partial \theta_{(k)}}{\partial x_{(i)}}(x).$$

Using this in Eqn. (24), we have

$$\frac{d\theta_{(i)}}{dt} = \sum_{k=1}^{d} \frac{\partial L}{\partial \theta_{(k)}}(x,\theta).\frac{\partial \theta_{(k)}}{\partial x_{(i)}}(x).$$

Now using Eqn. (23), along ϕ_{opt}, we get

$$\frac{d\theta_{(i)}}{dt} = -\frac{\partial L}{\partial x_{(i)}}(x,\theta),\ 1 \leqslant i \leqslant d.$$

Also,

$$\frac{dx_{(i)}}{dt} = \frac{\partial L}{\partial \theta_{(i)}}(x,\theta),\ 1 \leqslant i \leqslant d,$$

along ϕ_{opt}, since $\phi'(t) = v$ from Eqn. (21). This completes the proof of the theorem.

Notice that Eqns. (19) and (20), the initial condition $x(0) = x_0$ and the terminal condition $x(T) \in \partial A$ have ϕ_{opt} as a solution. To solve for ϕ_{opt} sometimes it is convenient also to use Eqn. (18). This will be illustrated in an example in § 3.4.

Next, we explain the role played by the variable θ. For this, define a new probability measure F_x^* from F_x as

$$dF_x^*(z) := \frac{e^{<\theta_x, z>} dF_x(z)}{M_x(\theta_x)}, \tag{25}$$

where the parameter $\theta_x \in \mathbb{R}^d$. This is called the exponential change of measure with the parameter θ_x.

Suppose we want to select θ_x along ϕ_{opt} in such a way that

$$\phi'_{opt}(t) = m^*(\phi_{opt}(t)), \tag{26}$$

where $m^*(x)$ denotes the mean of F_x^*. Then,

$$m^*(x) = \int_{\mathbb{R}^d} z dF_x^*(z) = \frac{\int_{\mathbb{R}^d} z.e^{<\theta_x, z>} dF_x(z)}{M_x(\theta_x)} = \frac{M'_x(\theta_x)}{M_x(\theta_x)} = l'_x(\theta_x). \tag{27}$$

Eqns. (26) and (27) indicate that the parameter of the exponential change of measure that makes the trajectory ϕ'_{opt} most likely satisfies

$$\phi'_{opt}(t) = l'_{\phi_{opt}(t)}(\theta_{\phi_{opt}(t)}). \tag{28}$$

Recalling our notation that $L(x,\theta) = l_x(\theta)$ and comparing Eqns. (19) and (28), it is clear that the variable θ in the system of differential equations (Eqns. (19)

and (20)) represent the parameter for the exponential change of measure required to achieve the condition in Eqn. (26).

3.3. Quick Simulation Method (Optimal Exponential Change of Measure).

In this section we present the concept of importance sampling, briefly described in § 1.2, in some more detail. This idea is applied by Cottrell et al. [2] to slow Markov processes for obtaining speed-ups in simulations. Their technique is called Quick Simulation Method. We will also present a theorem due to them that will be needed in the next section.

First, we present the idea of change of measure for general Markov chains that we illustrated by an $M/M/1$ queue example in § 2.3. Let $\{X_n, n = 0,1,2,\cdots\}$ be a discrete time Markov chain and (Ω,Σ,P) be the corresponding probability space. (Ω is the collection of sample paths of $\{X_n\}$ and P is a probability measure on a σ-field Σ of Ω)

Let $S \in \Sigma$ be a rare event, i.e., $\alpha := P\{S\} << 1$. For the direct Monte Carlo simulation we have

$$\alpha_n := \frac{1}{n}.\sum_{i=1}^{n} 1_S(\omega_i) \tag{29}$$

as a convergent and unbiased estimator of α. Here ω_i's are the i.i.d. outcomes of the experiments on (Ω,Σ,P). The variance of α_n is given by

$$\frac{1}{n}.(\alpha - \alpha^2). \tag{30}$$

As described in § 2.2, we need

$$n_d \approx \gamma.\frac{Var\{\alpha_1\}}{\alpha} \tag{31}$$

experiments to get an (ϵ,β)-confidence estimator, where γ is a positive constant depending on ϵ and β.

Alternatively, we may consider a probability measure P' on the measurable space (Ω,Σ) such that P is absolutely continuous with respect to P'. The Radon-Nikodym derivative (likelihood ratio) $L := dP/dP'$ can be used to obtain

another convergent and unbiased estimator

$$\alpha'_n = \frac{1}{n}.\sum_{i=1}^{n} 1_S(\omega_i).L(\omega_i) \tag{32}$$

of α. Here ω_i are the i.i.d. outcomes of the experiments on (Ω,Σ,P'). The variance of α_n^* is given by

$$\frac{1}{n}.(\int_S L^2(\omega)dP'(\omega) - \alpha^2).$$

As in Eqn. (31), we need

$$n_c \approx \gamma.\frac{Var'\{\alpha_1\}}{\alpha} \tag{33}$$

experiments to get an (ϵ,β)-confidence estimator, where $Var'\{\ \}$ denotes the variance under the measure P'.

A comparison of Eqns. (31) and (33) shows that α'_n will be more economical if and only if $Var'\{\alpha'_1\} < Var\{\alpha_1\}$, which will be the case if and only if

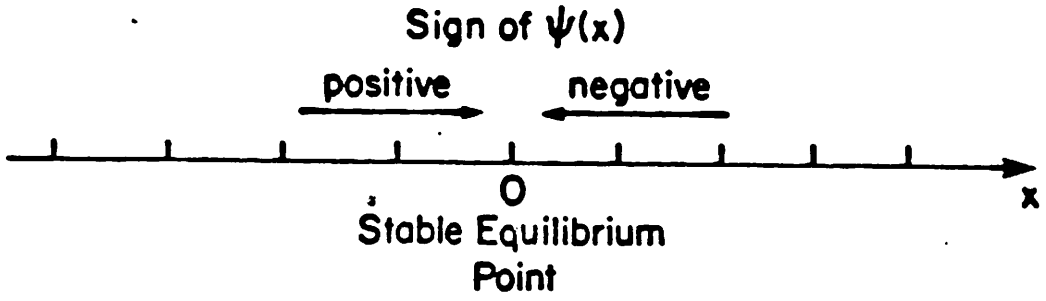
$$\int_S L^2(\omega)dP'(\omega) = \int_S L(\omega)dP(\omega) < \alpha. \tag{34}$$

Obviously, if $L(\omega) < 1$ whenever $\omega \in S$ then this condition is satisfied.

In the previous section we discussed the Markov chain $\{X_n^\epsilon\} \in \mathbb{R}^d$, defined in Eqn. (12). We now present a theorem due to Cottrell et al. [2] that gives, for the simulation purpose, the optimality of a measure $P^{\epsilon *}$, obtained by an exponential change of measure, from P^ϵ. Their theorem is presented in [2] for the case of $\mathbb{R}^1$. However, it can be generalized to the case of $\mathbb{R}^d$.

It is assumed that the mean drift function $\psi(x) := E\{V(X_n^\epsilon,\xi_n) / \ X_n^\epsilon = x\}$ is such that the O.D.E., $x'(t) = \psi(x(t))$, with $x(0)$ specified, has 0 as a stable equilibrium point. This is illustrated in Figure (6).

Suppose that we want to estimate, for small $\epsilon > 0$, $P_0^\epsilon(S)$, probability of the event



Mean Drift Function of $\{X_n^\epsilon\}$.

Figure (6)

$S := \{\omega \,/\, \{X_n^\epsilon\}$ *exceeds* 1 *before hitting* 0 $\}$

given that $X_0^\epsilon = 0$. Let us define a probability measure $P^{\epsilon *}$ as the resultant measure when F_x^* is taken as defined by Eqn. (25), with θ_x being the solution of

$$M_x(\theta_x) = 1, \ \theta_x > 0. \tag{35}$$

Under $P^{\epsilon *}$ we have a different Markov chain $\{X_n^{\epsilon *}\} \in \mathbb{R}^d$, which can be represented as

$$X_0^{\epsilon *} = x_0,$$

$$X_{n+1}^{\epsilon *} = X_n^{\epsilon *} + \epsilon . W(X_n^{\epsilon *}, \zeta_n),$$

where ζ_n's are i.i.d. and F_x^* is the d.f. of $W(x, \zeta_n)$. The probability measure $P^{\epsilon *}$ is optimal in the sense made precise by the following theorem due to Cottrell et al. [2].

Theorem (4) (Cottrell et al.) [2] : Suppose that for the Markov chain $X_n^\epsilon \in \mathbb{R}^1$, defined in Eqn. (12), assumptions (A1) and (A2) hold. Then among all the exponential changes of measure, the transformation $P^\epsilon \rightarrow P^{\epsilon *}$ is asymptotically optimal in the sense of the variance, i.e., for $P^{\epsilon *}$

$$\lim_{\epsilon \to 0} \int_S L^2(\omega) dP^{\epsilon *}(\omega),$$

where $L = dP^\epsilon / dP^{\epsilon *}$, is minimum. (Refer to Eqn. (34) to see why this sense of optimality is meaningful.)

To end this subsection, we make some observations which are conjectural for the present. The first conjecture is regarding the transformation $P^{\epsilon} \rightarrow P^{\epsilon^*}$. We believe that the transformation $P^{\epsilon} \rightarrow P^{\epsilon^*}$ is such that it makes $L = dP^{\epsilon}/dP^{\epsilon^*}$ almost constant (under the measure P^{ϵ} and say $l = l(\epsilon)$) on the set S and $l \ll 1$. For examples, see the $M/M/1$ example of § 3.4 (particularly Eqn. (6)) and pp. 911 of [2]. This clearly ensures that

$$E^{\epsilon^*}\{(L\,1_S)^2\} \approx l.P^{\epsilon}\{S\} \ll P^{\epsilon}\{S\} = E^{\epsilon}\{(1_S)^2\},$$

where $E^{\epsilon^*}\{\ \}$ and $E^{\epsilon}\{\ \}$ denote the expectations under the measures P^{ϵ^*} and P^{ϵ} respectively. This shows why it is more efficient to simulate under P^{ϵ^*} (see Eqn. (34). Hence, if the conditional probability $P^{\epsilon}(./S)$ is concentrated on a subset S^{ϵ^*} of S then one may select P^{ϵ^*} such that it is essentially concentrated on the set S^{ϵ^*}.

Next, we point out why we believe that it suffices to make an exponential change of measure that minimizes the variance of the indicator of S times the likelihood ratio. More precisely, if we denote by J_k the simulation time of the experiment k then the expected time for a simulation for the prescribed accuracy will be $E\{J_k\}n_d$ and $E^*\{J_k\}n_c$ under the original and the exponentially changed measures respectively. Observe from Eqns. (31) and (33) that

$$\Lambda := \frac{n_c}{n_d} = \frac{Var^*\{L.1_S\}}{Var\{1_S\}}.$$

We believe that for exponential changes of measure (up to logarithmic equivalence)

$$\Lambda \approx (\kappa)^{\frac{1}{\epsilon}},$$

where $\kappa > 0$ depends on the exponential change of measure. The measure P^{ϵ^*} minimizes κ. Furthermore, we believe that

$$\Upsilon := \frac{E^*\{J_k\}}{E\{J_k\}} = O(\frac{1}{\epsilon})$$

for all the exponential changes of measure. Since we are interested in

minimizing $\Lambda.Y$, under these conjectures, our assertion, that the exponential change of measure that minimizes Λ is optimal, is self-evident for sufficiently small ϵ. We have observed the validity of these conjectures in our experiments with Jackson networks. This conjecture was seen to be true for the $M/M/1$ queue example in § 2.

3.4. Applications and Difficulties.

Consider an open Jackson network of $d > 0$ nodes with infinite buffers. Let $\{X_n, n = 0,1,2,\cdots\} \in \mathbf{R}^d$ denote the embedded discrete time Markov chain representing queue-lengths of the nodes at the epochs of the jumps in the network (arrivals, departures and transfers), where $X_n = (X_{n(1)}, X_{n(2)}, \cdots, X_{n(d)}) \in \mathbf{R}^d$ (actually, $X_n \in \mathbf{N}^d$). Let S denote the set of the realizations of $\{X_n\}$ that reach the region of the state-space where the total backlog exceeds N, $x_{(1)}+x_{(2)}+\cdots+x_{(d)} \geq N$, before hitting 0, $x_{(1)} = x_{(2)} = \cdots = x_{(d)} = 0$. We want to estimate the probability $\alpha := P_0\{S\}$, the probability of S given that $X_0 = 0$.

We can represent $\{X_n\}$ as

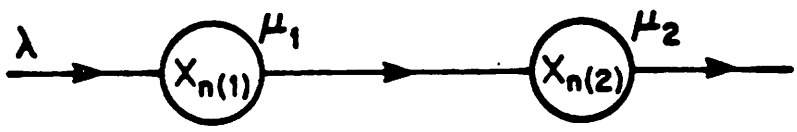
$$X_0 = x_0,$$

$$X_{n+1} = X_n + V(X_n, \xi_n), \quad n \geq 0, \tag{36}$$

where $V(x, \xi_n)$ denote the r.v. representing the jump from $X_n = x$. For example, consider $M/M/1$ queues in tandem (see Figure (7)). We assume, for stability, $\lambda < \mu_1$ and $\lambda < \mu_2$. We also assume, without any loss of generality, that $\lambda+\mu_1+\mu_2 = 1$. For simplicity, we will refer to such a system by a (λ,μ_1,μ_2)-network.

Now $\{X_n\}$ will be a Markov chain in $\mathbf{R}^2$ defined by Eqn. (36), where the distribution of $V(.,\xi_n)$ is given as follows.

$$P\{V((0,0),\xi_n) = (1,0)\} = 1,$$



$$\lambda + \mu_1 + \mu_2 = 1$$
$$\lambda < \mu_1 \quad \lambda < \mu_2$$

$M/M/1$ Queues in Tandem.

Figure (7)

$$P\{V((0,x_{(2)}),\xi_n) = (1,0)\} = \frac{\lambda}{\lambda+\mu_2},$$

$$P\{V((0,x_{(2)}),\xi_n) = (0,-1)\} = \frac{\mu_2}{\lambda+\mu_2}, \quad x_{(2)} > 0,$$

$$P\{V((x_{(1)},0),\xi_n) = (1,0)\} = \frac{\lambda}{\lambda+\mu_1},$$

$$P\{V((x_{(1)},0),\xi_n) = (-1,1)\} = \frac{\mu_1}{\lambda+\mu_1}, \quad x_{(1)} > 0,$$

$$P\{V((x_{(1)},x_{(2)}),\xi_n) = (1,0)\} = \lambda,$$

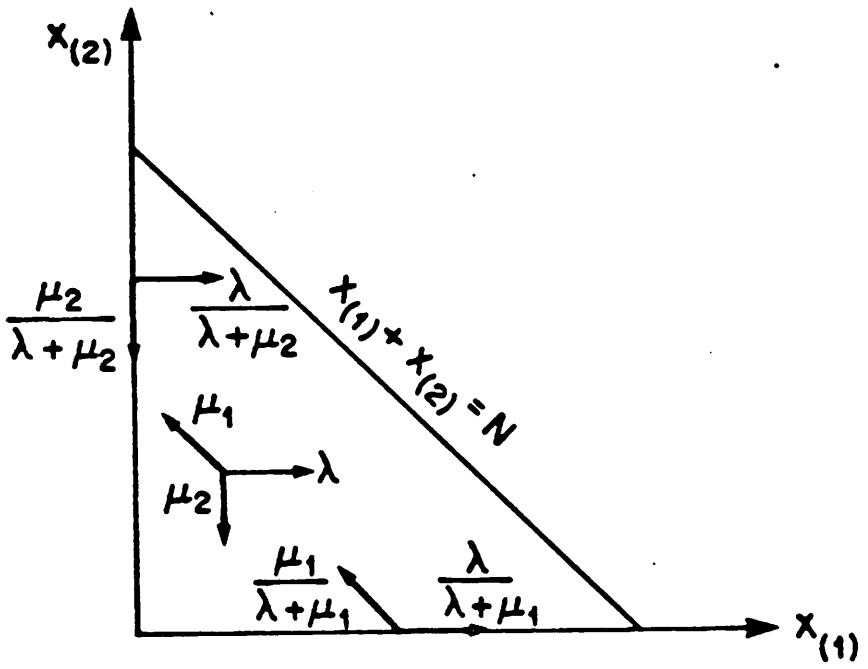
$$P\{V((x_{(1)},x_{(2)}),\xi_n) = (-1,1)\} = \mu_1,$$

$$P\{V((x_{(1)},x_{(2)}),\xi_n) = (0,-1)\} = \mu_2, \quad x_{(1)} > 0,\ x_{(2)} > 0, \tag{37}$$

These jump distributions are depicted in Figure (8).

Let us return to the discussion of general Jackson networks. It is possible to represent the embedded Markov chain $\{X_n\}$ in the form of Eqn. (12). For this define $X_n^N = X_n / N$. Then,

$$X_{n+1}^N = X_n^N + \frac{1}{N}.V(X_n,\xi_n) = X_n^N + \frac{1}{N}.V(N.X_n^N,\xi_n) = X_n^N + \frac{1}{N}.V(X_n^N,\xi_n). \tag{38}$$



Jump Distributions of $M/M/1$ Queues in Tandem.

Figure (8)

The last equality follows from the fact that in Jackson networks the distributions of $V(x,\xi_n)$ and $V(cx,\xi_n)$ are same for all x and all $c>0$. Because of Eqn. (38), we have an equivalent representation of $\{X_n\}$ which is in the same form as Eqn. (12) with $\epsilon = 1/N$. For the process $\{X_n^N\}$ we are interested in estimating $\alpha = P_0\{S^N\}$ is the set of the realizations of $\{X_n^N\}$ that reach the region of the state-space where the sum of its coordinates exceeds 1.

M/ M/ 1 Queue : Let us investigate if we can apply the ideas developed in § 3.2 and § 3.3 to an $M/M/1$. Let λ and μ be its arrival rate and service rate respectively. As usual, we assume $\lambda<\mu$ and $\lambda+\mu=1$. For the embedded Markov chain $\{X_n\}$, defined in Eqn. (36), the the distribution of $V(.,\xi_n)$ is now given by

$$P\{V(0,\xi_n)=1\}=1$$

$$P\{V(x,\xi_n)=1\}=\lambda=1-P\{V(x,\xi_n)=-1\},\ x>0. \tag{39}$$

We want to estimate $\alpha = P_0(S)$, where S now represents the set of realizations of $\{X_n\}$ that reach N before hitting 0.

Observe that the event S considered here is the same as the event that starting from 1, $\{X_n\}$ reaches N before 0. Hence, the realizations of S can be assumed to have the same jump distribution

$$P\{V(x,\xi_n)=1\}=\lambda=1-P\{V(x,\xi_n)=-1\}.$$

everywhere. For $\{X_n\}$, note that

$$M_x(s)=\lambda.e^{s}+\mu.e^{-s},\ x>0.$$

Eqn. (35) which is the same as Eqn. (18) along with the condition $\phi_{opt}(T)=1$, gives

$$\theta_x=\log(\frac{\mu}{\lambda}),\ x>0. \tag{40}$$

Eqn. (20) gives

$$\phi'_{opt}=\mu-\lambda \tag{41}$$

For the jump distribution of Eqn. (39), the example (E1) of § 3.1 (Eqn. (10)), we have

$$h_{\phi(t)}(\phi'(t))=(\mu-\lambda).\log(\frac{\mu}{\lambda}),\ t>0.$$

Now we can use Corollary (1), § 3.2, to evaluate $P_0\{S\}$. Noting that, for ϕ_{opt} defined in Eqn. (41), $T=1/(\mu-\lambda)$, we get

$$\lim_{N\to\infty}-\frac{1}{N}.\log P_0\{S\}=\log(\frac{\mu}{\lambda}).$$

This gives (up to logarithmic equivalence)

$$P_0\{S\}\approx(\frac{\lambda}{\mu})^{N}.$$

Observe that this matches well with the exact expression for $P_0\{S\}$ given by Eqn. (1).

Also observe that θ_λ given by Eqn. (40) gives the exponential change of measure (see Eqn. (25)) that corresponds to the $M/M/1$ queue with arrival rate μ and service rate λ (see Figure (4)).

From the discussion in § 3.3, it follows that, as $N \to \infty$, the variance of the estimator α'_n is minimum among all exponential changes of measure. It is not difficult to see that, for an $M/M/1$ queue $\{X_n\}$, $\{X_n^{*(\theta)}\}$ is obtained by an exponential change of measure with the parameter θ if and only if $\{X_n^{*(\theta)}\}$ is a Markov chain corresponding to an $M/M/1$ queue with some $\lambda(\theta)$ and $\mu(\theta)$ as its arrival rate and service rate respectively. Therefore, if we try to estimate $P_0\{S\}$, as $N \to \infty$, by running an $M/M/1$ queue other than that shown in Figure (4), then we will have a larger variance. Table (4) shows that even for "small" N, the minimum variance (empirically obtained) is achieved by running an $M/M/1$ queue that corresponds to the interchange of λ and μ (see Figure (4)).

$\lambda = 0.3$ $\mu = 0.7$ N = 15 $\alpha = 4.030\times10^{-6}$ # of Experiments = 100 # of Cycles (n) = 500		$\lambda = 0.2$ $\mu = 0.8$ N = 8 $\alpha = 4.578\times10^{-5}$ # of Experiments = 100 # of Cycles (n) = 500	
$\lambda(\theta) = 1.0 - \mu(\theta)$	Empirical Std. Dev. ($\tilde{\sigma}$)	$\lambda(\theta) = 1.0 - \mu(\theta)$	Empirical Std. Dev. ($\tilde{\sigma}$)
0.60	1.518×10^{-7}	0.70	7.036×10^{-7}
0.62	1.050×10^{-7}	0.72	6.458×10^{-7}
0.64	9.039×10^{-8}	0.74	5.402×10^{-7}
0.66	6.301×10^{-8}	0.76	4.633×10^{-7}
0.68	6.399×10^{-8}	0.78	3.253×10^{-7}
0.70	5.355×10^{-8}	0.80	3.212×10^{-7}
0.72	6.281×10^{-8}	0.82	3.646×10^{-7}
0.74	5.387×10^{-8}	0.84	3.548×10^{-7}
0.76	6.527×10^{-8}	0.86	4.272×10^{-7}
0.78	1.307×10^{-7}	0.88	6.952×10^{-7}

Empirical Standard Deviation for an $M/M/1$ Queue
as a Function of $\lambda(\theta)$.
Table (4)

Let us point out that to get ϕ_{opt} and hence the optimal change of measure it is not necessary to go through the derivation of the general method that led us to Eqns. (19) and (20). Recall from Corollary (1) that we need to minimize

$$\int_0^T h_{\phi(t)}(\phi'(t))dt.$$

First observe that in this example $h_{\phi(t)}(\phi'(t)) = h(\phi'(t))$, since the jump distributions are identical. Since $h(.)$ is convex (see the property (P1) of the Cramér transform, § 3.1) and $\phi(0) = 0$ and $\phi(T) = 1$,

$$\frac{1}{T}.\int_h(\phi'(t)) \geq h(\frac{1}{T}).$$

So, we find that

$$\inf_{\phi \in S} \int_S h_{\phi(t)}(\phi'(t))dt = \inf_{T>0} T.h(\frac{1}{T}).$$

Now, here h is given by the example (E1) of § 3.1 (Eqn. (10)). The minimization on the right hand side gives us the same result as before, namely,

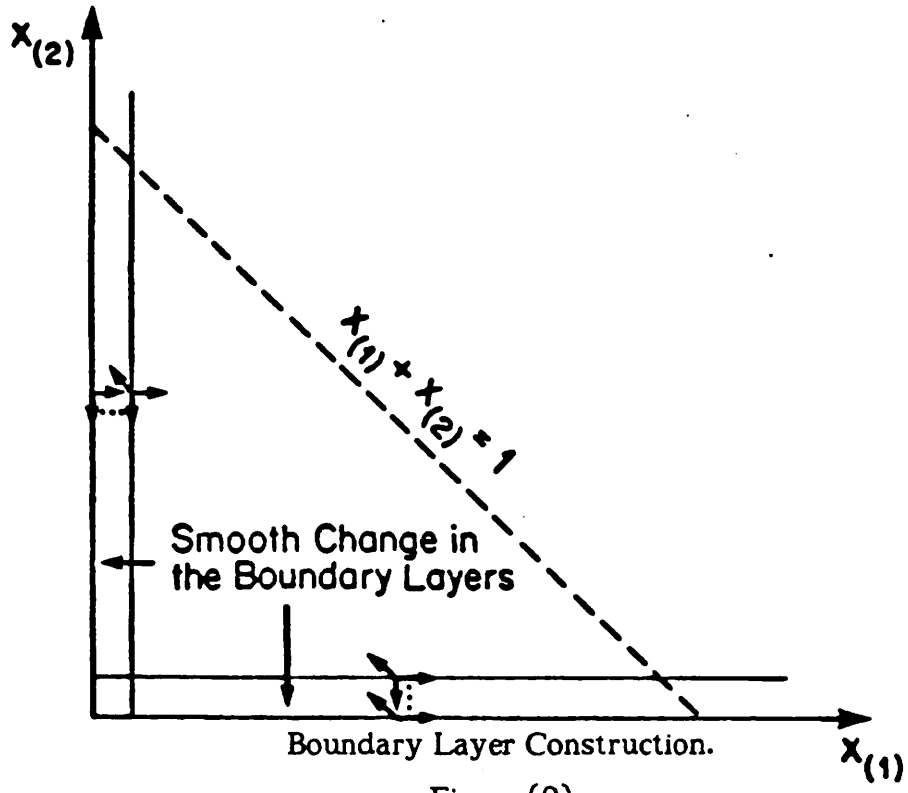
$$T = \frac{1}{\mu - \lambda}.$$

This also gives that under the optimal exponential change of measure the drift should be $\mu - \lambda$, which again corresponds to the interchange of λ and μ (see Figure (4)).

M/ M/ 1 Queues in Tandem : We consider a (λ, μ_1, μ_2)-network defined in the beginning of this section. (See Figures (7) and (8) and Eqn. (37).) This simple Jackson network illustrates the difficulty in applying the results of the previous two sections to Jackson networks.

Observe from Figure (10) that the jump distributions change abruptly near $x_{(1)}$-axis (second queue empty) and $x_{(2)}$-axis (first queue empty) if we move from these axes to R, the interior region (both the queues non-empty). This violates the smoothness assumption (A2) of § 3.2. Hence, the results of the previous two sections are not applicable here.

As a remedy to this difficulty, we ma y consider a process which has jump distributions modified near the boundaries ($x_{(1)}$-axis and $x_{(2)}$-axis) such that over a thin layer they make smooth transitions. We call such a construction a boundary layer construction. Intuitively, by modifying jump distributions a little, it is plausible that $P_0\{S\}$ (defined in the beginning of this subsection) does not change much. For the scaled process $\{X_n^N\}$ this construction is illustrated in Figure (9).

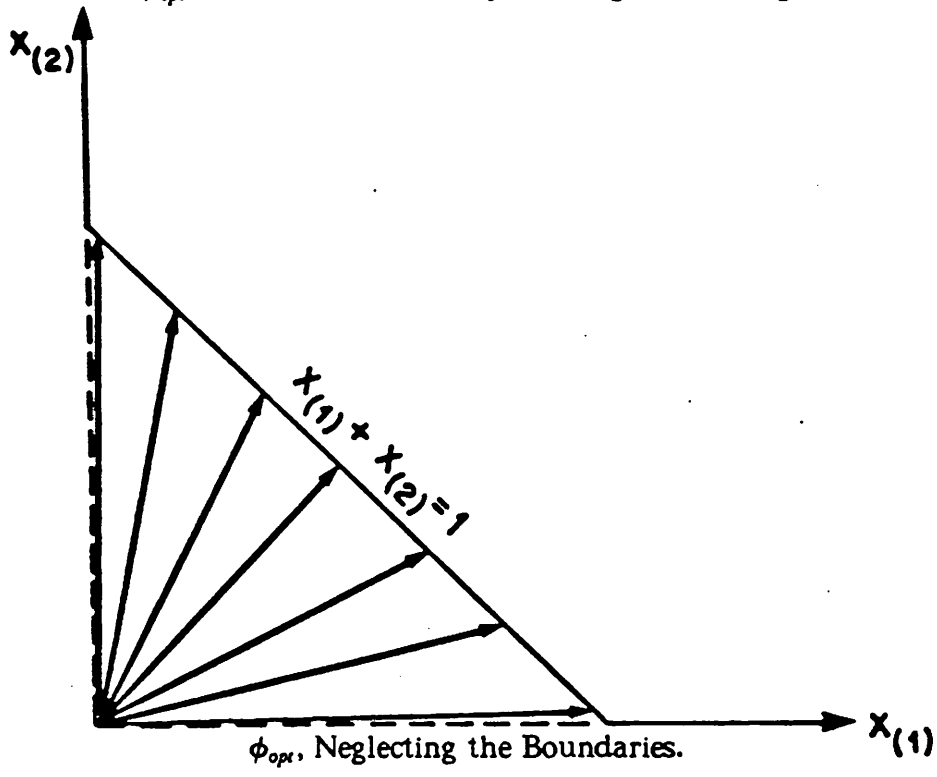


Boundary Layer Construction.

Figure (9)

If such a construction were indeed valid, we could use Eqns. (18), (19) and (20) to find ϕ_{opt} and $P_0\{S\}$ by the Quick Simulation Method. However, we find this numerical approach rather formidable because of the need to solve a system of differential equations with mixed initial and terminal conditions. Hence, we will not pursue further this approach here.

Next, we show that the boundaries in this example are indeed important. Suppose we assume that the jump distributions are identical everywhere to that of the interior region R. Then, from Eqns. (19) and (20), we see that $\phi'(t)$ is constant. Hence, ϕ_{opt} will be one of the rays through R (see Figure (10)).



ϕ_{opt}, Neglecting the Boundaries.

Figure (10)

From Eqn. (16) we get $C(x) = 0$ if $x = (x_{(1)},x_{(2)}) \in \partial A$, where $\partial A := \{x = (x_{(1)},x_{(2)}) \in \mathbf{R}^2 \,/\, x_{(1)} \geq 0, x_{(2)} \geq 0 \text{ and } x_{(1)}+x_{(2)} = 1\}$. Hence, $C((x_{(1)},1-x_{(1)})) = 0$, $0 \leq x_{(1)} \leq 1$. Therefore, for $x \in \partial A$,

$$0 = \frac{dC}{dx_{(1)}}((x_{(1)},1-x_{(1)})) = \frac{\partial C}{\partial x_{(1)}}((x_{(1)},1-x_{(1)})) - \frac{\partial C}{\partial x_{(2)}}((x_{(1)},1-x_{(1)})).$$

So, it follows that, for $x \in \partial A$,

$$\frac{\partial C}{\partial x_{(1)}} = \frac{\partial C}{\partial x_{(2)}} \tag{42}$$

From Eqn. (19), we see that θ, the parameter for the exponential change of

measure, is constant along ϕ_{opt}. Then, from Eqns. (17) and (42), we have

$$\theta_{(1)} = \theta_{(2)} \quad along\ \phi_{opt}.$$

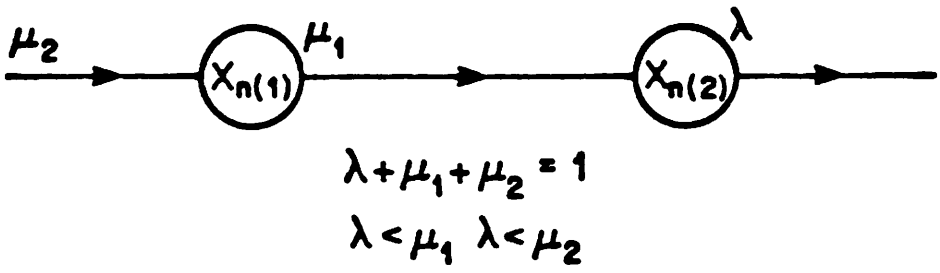
In the present case

$$l_x(s) = \log(\lambda.e^{s(1)} + \mu_1.e^{-s(1)+s(2)} + \mu_2.e^{-s(2)})$$

Solving $l_x(\theta) = 0$ (see Eqn. (18)), with the constraint that $\theta_{(1)} = \theta_{(2)}$ ($\neq 0$), we get

$$\theta_{(1)} = \theta_{(2)} = \log(\frac{\mu_2}{\lambda}).$$

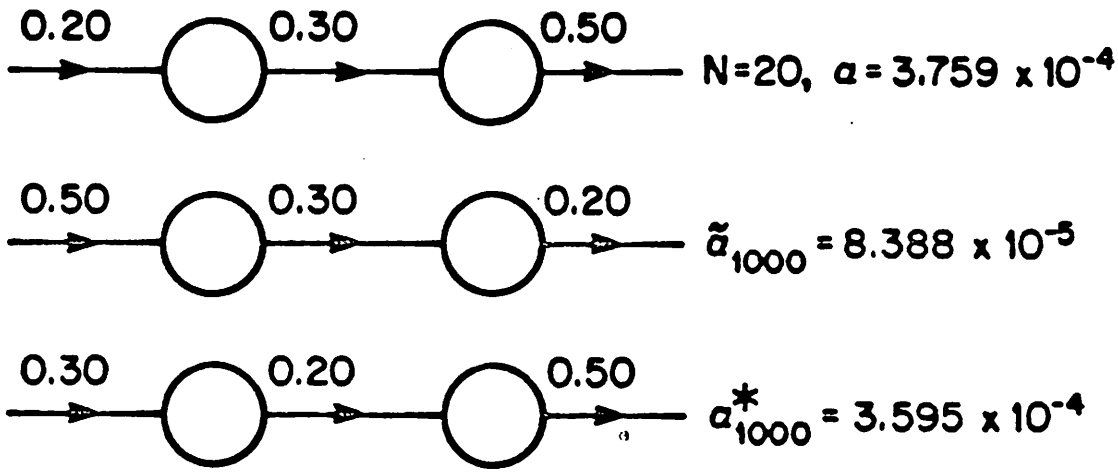
The exponential change of measure with the parameter θ can be seen to give the (μ_2,μ_1,λ)-network (see Figure (11)).



Change of Measure, Neglecting the Boundaries.

Figure (11)

It is easy to verify by simulations that above is not an optimal exponential change of measure for a (λ,μ_1,μ_2)-network. For example, for the $(\lambda = 0.20,\mu_1 = 0.30,\mu_2 = 0.50)$ -network and $N = 20$, $\alpha = P_0\{S\}$ is found by solving the first step equations numerically to be 3.759×10^{-4}. If we simulate the (0.50,0.30,0.20)-network, as suggested by the above discussion (see Figure (11)), we get $\tilde{\alpha}_{1000} = 8.388\times10^{-5}$, while simulating the (0.30,0.20,0.50)-network



Comparison of Changes of Measure.
Figure (12)

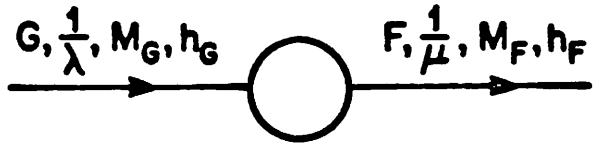
we get $\alpha^{\bullet}_{1000} = 3.595\times10^{-4}$. This example is illustrated in Figure (12). Note that the (0.30,0.20,0.50)-network is also obtained from the original network by an exponential change of measure. In the next section we will present a heuristic that will justify the optimality of this change of measure.

4. Simulation of Events of Excessive Backlog - A Heuristic Approach.

The purpose of this section is to report the observations we have made rather than to claim new results. Our hope is that the heuristic explanations and observations presented here will motivate more research in this area. Some limiting cases for our heuristic are reported in § 4.4.

4.1. Heuristic of Borovkov, Ruget [3] etc. for a *GI/ G/* 1 Queue and its Application to Simulations.

Consider a *GI/ G/* 1 Queue shown in Figure (13). Let G and F denote the



GI / G / 1 Queue.

Figure (13)

interarrival and service time d.f.'s respectively. Let M_D and h_D denote the Laplace and Cramér transforms of a d.f. D. Let $1/\lambda$ and $1/\mu$ denote the means of G and F respectively. For stability, we assume $1/\lambda > 1/\mu$. Let P denote the measure induced by the stochastic process describing the queue. We want to calculate α, the probability of the backlog exceeding N in a cycle, i.e., the probability of hitting N before returning to 0 given that the system starts empty. Let S denote the event that the system reaches N before hitting 0. Then, $\alpha = P_0\{S\}$.

Let X_i^D denote the i^{th} i.i.d. copy of a random variable distributed with the d.f. D. Then, we let X_i^G denote the i^{th} interarrival time and X_i^F denote the i^{th} virtual service time. Consider the subset of S where the system reaches N at time T and the average interarrival and the virtual service times are $1/\lambda'$ and $1/\mu'$ respectively with $1/\lambda' < 1/\mu'$. Now, by Cramér's Theorem, Theorem (1), (up to logarithmic equivalence)

$$P\{X_1^G + \cdots + X_{\lambda' T}^G \approx T\} = P\{\frac{X_1^G + \cdots + X_{\lambda' T}^G}{\lambda' T} \approx \frac{1}{\lambda'}\} \approx \exp(-\lambda' T h_G(\frac{1}{\lambda'})).$$

Similarly, (up to logarithmic equivalence)

$$P\{X_1^F + \cdots + X_{\mu' T}^F \approx T\} \approx \exp(-\mu' T h_F(\frac{1}{\mu'})).$$

Since $1/\lambda' < 1/\mu'$, for large T, we assume that most of the virtual services were the actual services. Then, $T \approx N/(\lambda'-\mu')$. Since, the interarrival times and the virtual service times are independent, (up to logarithmic equivalence)

$$\alpha \equiv P_0\{S\} \approx \sum_T \sum_{\substack{\lambda',\mu' \\ \lambda' > \mu' \geq 0 \\ N = T.(\lambda'-\mu')}} \exp\{-T.(\lambda'.h_G(\frac{1}{\lambda'})+\mu'.h_F(\frac{1}{\mu'}))\}$$

$$= \sum_{\lambda' > \mu' \geq 0} \exp\{-\frac{N}{\lambda'-\mu'}.(\lambda'.h_G(\frac{1}{\lambda'})+\mu'.h_F(\frac{1}{\mu'}))\}.$$

Hence, for large N, (up to logarithmic equivalence)

$$\alpha \approx \exp\{-N. \inf_{\lambda' > \mu' \geq 0} [\frac{1}{\lambda'-\mu'}.(\lambda'.h_G(\frac{1}{\lambda'})+\mu'.h_F(\frac{1}{\mu'}))]\}. \tag{43}$$

To obtain the exponent, we differentiate

$$\frac{1}{\lambda'-\mu'}.(\lambda'.h_G(\frac{1}{\lambda'})+\mu'.h_F(\frac{1}{\mu'}))$$

with respect to λ' and μ' and equate the results to 0. This gives

$$h_G(\frac{1}{\lambda'})+h_F(\frac{1}{\mu'}) = (\frac{1}{\lambda'}-\frac{1}{\mu'}).h'_G(\frac{1}{\lambda'}) = (\frac{1}{\mu'}-\frac{1}{\lambda'}).h'_F(\frac{1}{\mu'}). \tag{44}$$

Suppose that λ^* and μ^* achieve the infimum. Then, from Eqn. (2.44),

$$-h'_G(\frac{1}{\lambda^*}) = h'_F(\frac{1}{\mu^*}) = \theta^* \ (say). \tag{45}$$

We can argue from the convexity of h_G and h_F that $\theta^* > 0$. Also, from Eqn. (44), we have

$$\theta^*.\frac{1}{\lambda^*}+h_G(\frac{1}{\lambda^*}) = \theta^*.\frac{1}{\mu^*}-h_F(\frac{1}{\mu^*}). \tag{46}$$

From the convex duality property (P4) of the Cramér transform (see § 3.1) and Eqns. (45) and (46), we have

$$\log M_G(-\theta^*) = -\theta^*.\frac{1}{\lambda^*}-h_G(\frac{1}{\lambda^*})$$

and

$$\log M_F(\theta^*) = \theta^* \cdot \frac{1}{\mu^*} - h_F\left(\frac{1}{\mu^*}\right). \tag{47}$$

Therefore,

$$\log M_G(-\theta^*) = -\log M_F(\theta^*), \tag{48}$$

i.e., the conditions for determining θ^* are

$$\theta^* > 0 \text{ and } M_F(\theta^*).M_G(-\theta^*) = 1. \tag{49}$$

From Eqns. (43) and (47), for large N, we also have (up to logarithmic equivalence)

$$\alpha \approx \exp(-N.\log M_F(\theta^*)). \tag{50}$$

Let G^* denote the measure obtained by an exponential change of measure from G such that its mean is $1/\lambda^*$, i.e., the parameter for the exponential change of measure, θ_G^*, satisfies

$$dG^*(z) = \frac{e^{\theta_G^* z} dG(z)}{M_G(\theta_G^*)}$$

and

$$\frac{1}{\lambda^*} = \frac{\int z.e^{\theta_G^* z} dG(z)}{M_G(\theta_G^*)} = \frac{d}{d\theta} \log M_G(\theta_G^*).$$

Using Eqn. (45) and the property of reciprocity of the derivatives of the Cramér and the log-Laplace transforms (property P(5) of the Cramér transform, § 3.1), we get

$$\theta_G^* = -\theta^*.$$

Similarly, let F^* denote the measure obtained by an exponential change of measure from F such that its mean is $1/\mu^*$. Then, the required parameter for the exponential change of measure, θ_F^*, can be seen to satisfy

$$\theta_F^* = \theta^*.$$

Now define a transformed $GI/G/1$ queue with G^* and F^* as its interarrival time and service time d.f.'s respectively. Let P^* denote the measure induced by the transformed stochastic process.

The definitions of λ^*, μ^* and P^* suggest that, for large N,

$$\frac{dP}{dP^*} \ll 1$$

almost everywhere (under measure P) on the event S. Then, Eqn. (34) indicates that it will be faster to estimate α under the measure P^* than under P.

M/ M/ 1 Example : Let λ and μ $(0 < \lambda < \mu)$ denote arrival and service rates. If D is the exponential d.f. with the mean $1/\nu$ then we denote by M_ν and h_ν its Laplace and Cramér transforms respectively. Recall that

$$M_\nu(s) = \frac{\nu}{\nu - s}, \quad s. < \nu,$$

$$= \infty, \quad \textit{otherwise.}$$

Eqn. (49) gives

$$\theta^* > 0 \text{ and } \frac{\lambda}{\lambda+\theta^*} \cdot \frac{\mu}{\mu-\theta^*} = 1. \tag{51}$$

It is easily checked that the solution of Eqn. (51) is

$$\theta^* = \mu - \lambda.$$

Then, Eqn. (50) gives (up to logarithmic equivalence)

$$\alpha = \left(\frac{\lambda}{\mu}\right)^N.$$

Observe that this matches well with the exact expression for α given by Eqn. (1).

Also, calculations of G^* and F^*, as defined above, show that the transformed $M/M/1$ queue for the purpose of estimating α by simulations is the one that corresponds to the interchange of λ and μ. Recall that we had obtained the same exponential change of measure by applying Large Deviation theory in § 3.4.

To end this subsection, we point out that we can have similar heuristic as above for the embedded Markov chain $\{X_n, n = 1,2,\cdots\}$ of an $M/M/1$ queue defined at the epochs of the arrivals and departures. As usual, we assume $\lambda+\mu = 1$. We can couple paths of any $M/M/1$ queue to those which have either an arrival or a virtual departure every unit of time and have probabilities λ and μ respectively. For the embedded Markov chain, consider the paths for which the event S occurs in T transitions (arrivals and virtual departures) and λ' and μ' proportions of arrivals and virtual departures respectively. Now we can have a heuristic similar to the one that led us to Eqn. (43) where now we can restrict our minimization to the set of proportions λ' and μ' such that $\lambda' > \mu'$, i.e., $\lambda'+\mu' = 1$ and $\lambda' > \mu' \geq 0$. Therefore, for simulating the embedded Markov chain to estimate α, we should use the embedded Markov chain of the $M/M/1$ queue for which λ and μ have been interchanged.

4.2. Extension to Simple Jackson Networks ($M/M/1$ Queues in Tandem and in Parallel).

As in § 3.4, for an open Jackson network of $d > 0$ nodes with infinite buffers, let $\{X_n, n = 0,1,2,\cdots\} \in \mathbf{R}^d$ denote the embedded discrete time Markov chain representing queue-lengths of the nodes at the epochs of the jumps in the network (arrivals, departures and transfers). We want to estimate $\alpha \equiv P_0\{S\}$, where S is the set of the realizations of $\{X_n\}$ that reach the region of the state-space where the total backlog exceeds N, before hitting 0.

M/ M/ 1 Queues in Tandem : For the embedded Markov chain $\{X_n\} \in \mathbf{R}^2$, Eqn. (37) gives the jump distributions. Recall that we have uniformized the Markov chain, i.e., $\lambda+\mu_1+\mu_2 = 1$.

Consider the paths of S which require T transitions and have λ', μ'_1 and μ'_2 proportions for the arrivals, virtual departures from the first queue and that from the second queue respectively. Continuing the same line of heuristic as in § 4.1, we can write (up to logarithmic equivalence)

$$\alpha \approx \sum_{\substack{\lambda' > 0, \mu'_1 \geq 0, \mu'_2 \geq 0 \\ \lambda'+\mu'_1+\mu'_2 = 1 \\ \lambda' > \mu'_1 \text{ or } \lambda' > \mu'_2}} \exp\{-T(\lambda', \mu'_1, \mu'_2).(\lambda' h_\lambda(\frac{1}{\lambda'}) + \mu'_1 h_{\mu_1}(\frac{1}{\mu'_1}) + \mu'_2 h_{\mu_2}(\frac{1}{\mu'_2}))\},$$

where $T(\lambda',\mu'_1,\mu'_2)$ is the total number of transitions (which equals the number of time units due to the uniformization) required for the realizations belonging to S with λ', μ'_1 and μ'_2 proportions of arrivals and virtual services from the queues respectively.

It can be heuristically argued that, for large N and when $\lambda' > \mu'_1$ or $\lambda' > \mu'_2$, $T(\lambda',\mu'_1,\mu'_2) \approx N.R(\lambda',\mu'_1,\mu'_2)$, where

$$R = \begin{cases} 1/(\lambda'-\mu'_1), \text{ if } \lambda' > \mu'_1 \text{ and } \mu'_1 \leq \mu'_2, \\ 1/(\lambda'-\mu'_2), \textit{otherwise.} \end{cases}$$

Therefore, for large N, (up to logarithmic equivalence)

$$\alpha = \exp\{-N. \inf_{\substack{\lambda'>0,\mu'_1\geq 0,\mu'_2\geq 0 \\ \lambda'+\mu'_1+\mu'_2=1 \\ \lambda'>\mu'_1 \text{ or } \lambda'>\mu'_2}} [R(\lambda',\mu'_1,\mu'_2).(\lambda'.h_\lambda(\frac{1}{\lambda'})+ \mu'_1.h_{\mu_1}(\frac{1}{\mu'_1})+\mu'_2.h_{\mu_2}(\frac{1}{\mu'_2}))]\}. \quad (52)$$

Numerical minimization gives λ^*, μ_1^* and μ_2^* that correspond to the interchange of λ with the smallest of μ_1 and μ_2. (For the limiting case where $\mu_1 = \mu_2$, § 4.4.) As explained for the case of an $M/M/1$ queue in § 4.1, to estimate α, it will be faster to simulate the embedded Markov chain of the $(\lambda^*,\mu_1^*,\mu_2^*)$-network.

Table (5) lists some illustrations of simulation speed-ups when simulated under the transformed system. It also shows the time required for the direct computation of α by solving the first step equations, the time required for a simulation and the corresponding number of calls to the random number generator (RNG).

Table (6) gives the empirical standard deviations, means and coefficients of variation of the estimates obtained by the change of measure for the same examples as in Table (5). All the simulations were done on a VAX-750 machine and the first step equations were solved using the IMSL routine LEQT2F.

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda = 0.05$ $\mu_1 = 0.10$ $\mu_2 = 0.85$ $N = 15$ $\alpha = 3.459\times10^{-5}$ *CPU Time* = 61.1*Sec.* $\lambda^* = 0.10$ $\mu_1^* = 0.05$ $\mu_2^* = 0.85$						
# of Cycles (n)	10000	20000	40000	200	500	1000
α_n (α_n^*)	0.0	0.0	0.0	3.338×10^{-5}	3.577×10^{-5}	3.448×10^{-5}
CPU Time	17.0*Sec.*	33.3*Sec.*	69.9*Sec.*	2.5*Sec.*	5.6*Sec.*	10.4*Sec.*
Calls to RNG	52769	109573	216395	5512	13595	26303
Example-II $\lambda = 0.10$ $\mu_1 = 0.50$ $\mu_2 = 0.40$ $N = 13$ $\alpha = 2.104\times10^{-7}$ *CPU Time* = 29.6*Sec.* $\lambda^* = 0.40$ $\mu_1^* = 0.50$ $\mu_2^* = 0.10$						
# of Cycles (n)	20000	30000	50000	700	1000	1500
α_n (α_n^*)	0.0	0.0	0.0	1.979×10^{-7}	2.159×10^{-7}	1.594×10^{-7}
CPU Time	25.4*Sec.*	38.1*Sec.*	67.3*Sec.*	7.5*Sec.*	11.7*Sec.*	16.9*Sec.*
Calls to RNG	79816	120270	200917	18920	27529	40763
Example-III $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $N = 20$ $\alpha = 3.759\times10^{-4}$ *CPU Time* = 310.3*Sec.* $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$						
# of Cycles (n)	5000	10000	20000	300	500	1000
α_n (α_n^*)	2.000×10^{-4}	1.000×10^{-4}	7.500×10^{-4}	3.848×10^{-4}	3.734×10^{-4}	3.595×10^{-4}
CPU Time	24.6*Sec.*	48.1*Sec.*	92.3*Sec.*	7.5*Sec.*	12.2*Sec.*	23.0*Sec.*
Calls to RNG	72234	144913	286539	18006	29854	56489

Simulations for $M/M/1$ Queues in Tandem.

Table (5)

Example-I $\lambda = 0.05$ $\mu_1 = 0.10$ $\mu_2 = 0.85$ $N = 15$ $\alpha = 3.459\times10^{-5}$ *# of Experiments* = 20 $\lambda^* = 0.10$ $\mu_1^* = 0.05$ $\mu_2^* = 0.85$		
# of Cycles (n)	500	1000
Empirical Mean ($\hat{m}$)	3.493×10^{-5}	3.385×10^{-5}
Empirical Std. Dev. ($\hat{\sigma}$)	8.971×10^{-7}	7.985×10^{-7}
($\hat{\sigma}$/ $\hat{m}$)×100 %	2.568 %	2.359 %
Example-II $\lambda = 0.10$ $\mu_1 = 0.50$ $\mu_2 = 0.40$ $N = 13$ $\alpha = 2.104\times10^{-7}$ *# of Experiments* = 20 $\lambda^* = 0.40$ $\mu_1^* = 0.50$ $\mu_2^* = 0.10$		
# of Cycles (n)	700	1500
Empirical Mean ($\hat{m}$)	2.223×10^{-7}	2.116×10^{-7}
Empirical Std. Dev. ($\hat{\sigma}$)	2.320×10^{-8}	1.610×10^{-8}
($\hat{\sigma}$/ $\hat{m}$)×100 %	10.437 %	7.608 %
Example-III $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $N = 20$ $\alpha = 3.759\times10^{-4}$ *# of Experiments* = 20 $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$		
# of Cycles (n)	500	1000
Empirical Mean ($\hat{m}$)	3.765×10^{-4}	3.805×10^{-4}
Empirical Std. Dev. ($\hat{\sigma}$)	2.481×10^{-5}	2.095×10^{-5}
($\hat{\sigma}$/ $\hat{m}$)×100 %	6.588 %	5.500 %

Empirical Standard Deviation for $M/M/1$ Queues in Tandem.

Table (6)

M/ M/ 1 Queues in Parallel : Consider two $M/M/1$ queues in parallel with λ_i and μ_i, $i = 1,2$, as their arrival and service rates respectively. We assume that $\lambda_i < \mu_i, i = 1,2$, and $\lambda_1+\mu_1+\lambda_2+\mu_2 = 1$. We denote such a system by $(\lambda_1,\mu_1 \mid \lambda_2,\mu_2)$-network.

As for $M/M/1$ queues in tandem, we can argue heuristically that (up to logarithmic equivalence), for large N,

$$\alpha \approx \exp\{-N. \inf_{\substack{\lambda'_1 \geq 0, \mu'_1 \geq 0, \lambda'_2 \geq 0, \mu'_2 \geq 0 \\ \lambda'_1+\mu'_1+\lambda'_2+\mu'_2 = 1 \\ \lambda'_1 > \mu'_1 \text{ or } \lambda'_2 > \mu'_2}} [R(\lambda'_1,\mu'_1,\lambda'_2,\mu'_2).(\lambda'_1 h_{\lambda_1}(\frac{1}{\lambda'_1})+$$

$$\mu'_1 h_{\mu_1}(\frac{1}{\mu'_1})+\lambda'_2 h_{\lambda_2}(\frac{1}{\lambda'_2})+\mu'_2 h_{\mu_2}(\frac{1}{\mu'_2}))]\}, \quad (53)$$

where (when $\lambda'_1 > \mu'_1$ or $\lambda'_2 > \mu'_2$)

$$R = \begin{cases} 1/(\lambda'_1-\mu'_1), \text{ if } \lambda'_1 > \mu'_1 \text{ and } \lambda'_2 \leq \mu'_2, \\ 1/((\lambda'_1-\mu'_1)+(\lambda'_2-\mu'_2)), \text{ if } \lambda'_1 > \mu'_1 \text{ and } \lambda'_2 > \mu'_2, \\ 1/(\lambda'_2-\mu'_2), \textit{otherwise.} \end{cases}$$

Numerical minimization gives $\lambda_1^*, \mu_1^*, \lambda_2^*$ and μ_2^* that correspond to the interchange of λ_i and μ_i with the larger traffic intensity λ_i/μ_i. (For the limiting case where $\lambda_1/\mu_1 = \lambda_2/\mu_2$, see § 4.4.) Hence, for estimating α, we simulate the embedded Markov chain corresponding to the $(\lambda_1^*,\mu_1^* \mid \lambda_2^*,\mu_2^*)$-network.

Table (7) lists some illustrations of simulation speed-ups when simulated under the transformed system. It also shows the time required for the direct computation of α by solving the first step equations, the time required for a simulation and the corresponding number of calls to the random number generator (RNG).

Table (8) gives the empirical standard deviations, means and coefficients of variation of the estimates obtained by the change of measure for the same examples as in Table (7). All the simulations were done on a VAX-750 machine and the first step equations were solved using the IMSL routine LEQT2F.

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda_1 = 0.10\ \mu_1 = 0.20\ \lambda_2 = 0.30\ \mu_2 = 0.40\ N = 23$ $\alpha = 1.213\times10^{-3}$ *CPU Time* = 624.4*Sec.* $\lambda_1^* = 0.10\ \mu_1^* = 0.20\ \lambda_2^* = 0.40\ \mu_2^* = 0.30$						
# of Cycles (n)	10000	20000	30000	2000	2500	3000
α_n (α_n^*)	1.000×10^{-3}	1.350×10^{-3}	1.267×10^{-3}	1.228×10^{-3}	1.188×10^{-3}	1.120×10^{-3}
CPU Time	57.0*Sec.*	120.4*Sec.*	173.9*Sec.*	41.6*Sec.*	52.5*Sec.*	60.4*Sec.*
Calls to RNG	160278	321031	482256	99576	127937	145472
Example-II $\lambda_1 = 0.10\ \mu_1 = 0.40\ \lambda_2 = 0.15\ \mu_2 = 0.35\ N = 18$ $\alpha = 6.935\times10^{-7}$ *CPU Time* = 158.1*Sec.* $\lambda_1^* = 0.10\ \mu_1^* = 0.40\ \lambda_2^* = 0.35\ \mu_2^* = 0.15$						
# of Cycles (n)	10000	20000	50000	3000	5000	6000
α_n (α_n^*)	0.0	0.0	0.0	7.016×10^{-7}	6.264×10^{-7}	6.483×10^{-7}
CPU Time	16.7*Sec.*	36.0*Sec.*	84.1*Sec.*	42.3*Sec.*	68.0*Sec.*	80.7*Sec.*
Calls to RNG	46758	94602	236618	93984	150190	179600
Example-III $\lambda_1 = 0.08\ \mu_1 = 0.12\ \lambda_2 = 0.20\ \mu_2 = 0.60\ N = 23$ $\alpha = 4.635\times10^{-5}$ *CPU Time* = 635.0*Sec.* $\lambda_1^* = 0.12\ \mu_1^* = 0.08\ \lambda_2^* = 0.20\ \mu_2^* = 0.60$						
# of Cycles (n)	10000	20000	50000	3000	4000	5000
α_n (α_n^*)	0.0	5.000×10^{-5}	0.0	4.725×10^{-5}	4.435×10^{-5}	4.725×10^{-5}
CPU Time	31.2*Sec.*	66.9*Sec.*	154.3*Sec.*	68.1*Sec.*	85.0*Sec.*	110.9*Sec.*
Calls to RNG	87180	186083	417832	153426	202164	262938

Simulations for $M/M/1$ Queues in Parallel.

Table (7)

Example-I $\lambda_1 = 0.10$ $\mu_1 = 0.20$ $\lambda_2 = 0.30$ $\mu_2 = 0.40$ $N = 23$ $\alpha = 1.213\times10^{-3}$ *# of Experiments* = 20 $\lambda_1^* = 0.10$ $\mu_1^* = 0.20$ $\lambda_2^* = 0.40$ $\mu_2^* = 0.30$		
# of Cycles (n)	2500	3000
Empirical Mean ($\hat{m}$)	1.193×10^{-3}	1.250×10^{-3}
Empirical Std. Dev. ($\hat{\sigma}$)	9.634×10^{-5}	7.571×10^{-5}
($\hat{\sigma}$/ $\hat{m}$)×100 %	8.072 %	7.571 %
Example-II $\lambda_1 = 0.10$ $\mu_1 = 0.40$ $\lambda_2 = 0.15$ $\mu_2 = 0.35$ $N = 18$ $\alpha = 6.935\times10^{-7}$ *# of Experiments* = 20 $\lambda_1^* = 0.10$ $\mu_1^* = 0.40$ $\lambda_2^* = 0.35$ $\mu_2^* = 0.15$		
# of Cycles (n)	5000	6000
Empirical Mean ($\hat{m}$)	6.905×10^{-7}	6.992×10^{-7}
Empirical Std. Dev. ($\hat{\sigma}$)	6.793×10^{-8}	5.028×10^{-8}
($\hat{\sigma}$/ $\hat{m}$)×100 %	9.839 %	7.191 %
Example-III $\lambda_1 = 0.08$ $\mu_1 = 0.12$ $\lambda_2 = 0.20$ $\mu_2 = 0.60$ $N = 23$ $\alpha = 4.635\times10^{-5}$ *# of Experiments* = 20 $\lambda_1^* = 0.12$ $\mu_1^* = 0.08$ $\lambda_2^* = 0.20$ $\mu_2^* = 0.60$		
# of Cycles (n)	3000	5000
Empirical Mean ($\hat{m}$)	4.471×10^{-5}	4.623×10^{-5}
Empirical Std. Dev. ($\hat{\sigma}$)	2.998×10^{-6}	2.004×10^{-6}
($\hat{\sigma}$/ $\hat{m}$)×100 %	6.705 %	4.334 %

Empirical Standard Deviation for $M/M/1$ Queues in Parallel.

Table (8)

4.3. Extension to Networks with Routing.

In § 4.2, we extended our heuristic to $M/M/1$ queues in tandem and in Parallel. In this subsection we will extend it further to networks where probabilistic routing may be present. By doing so, we will have extended the heuristic to arbitrary open Jackson networks.

For this purpose, we need the following theorem due to Sanov [4].

Theorem (5) (Sanov) [4] : Let Z_i, $i \geq 1$, be random variables whose possible values are $a_1, \cdots, a_n$ with $p_1, \cdots, p_n$ as respective probabilities. For

$N > 1$, define $m_i(N) :=$ # of Z_k's, $1 \leq k \leq N$, that are equal to a_i. Define the relative frequency

$$\nu_i := \frac{m_i(N)}{N}, \quad 1 \leq i \leq n.$$

Let $q_1, \cdots, q_n$ be real numbers satisfying $q_i \geq 0$, $1 \leq i \leq n$ and $q_1 + \cdots + q_n = 1$. Then,

$$\lim_{N \to \infty} \frac{1}{N} . \log P\{ |\nu_1(N) - q_1| \leq \epsilon, \cdots, |\nu_n(N) - q_n| \leq \epsilon \} = -K(\mathbf{q},\mathbf{p}) + e(\epsilon),$$

where

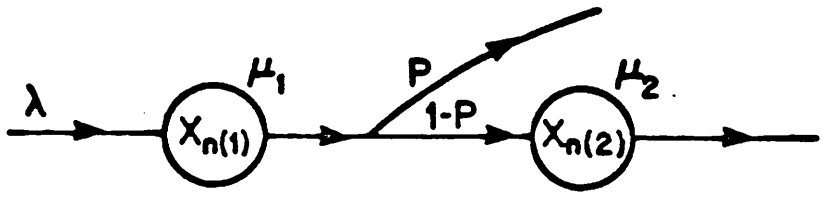
$$K(\mathbf{q},\mathbf{p}) = \sum_{i=1}^{n} q_i . \log(\frac{q_i}{p_i})$$

and the term $e(\epsilon)$ is $O(\epsilon.\log(1/\epsilon))$. (If $q_i > 0$, $1 \leq i \leq n$, then $O(\epsilon.\log(1/\epsilon))$ can be replaced by $O(\epsilon)$.)

Above theorem suggests that (up to logarithmic equivalence)

$$P\{m_1(N) \approx q_1.N, \cdots, m_n(N) \approx q_n.N\} \approx \exp(-N.K(\mathbf{q},\mathbf{p})).$$

Now consider the network shown in Figure (14).



$$\lambda + \mu_1 + \mu_2 = 1$$
$$\lambda < \mu_1 \quad \lambda \cdot (1-P) < \mu_2$$

Example of a Network with Routing.

Figure (14)

For stability, we assume that $\lambda < \mu_1$ and $\lambda.(1-p) < \mu_2$. We also assume, without any loss of generality, that $\lambda+\mu_1+\mu_2 = 1$. We consider the embedded Markov chain $\{X_n\}$. See the beginning of § 4.2 for the description of the problem.

As in the cases of $M/M/1$ queues in tandem and in parallel, consider the paths of S which require T transitions, have λ', μ'_1, and μ'_2 proportions for the arrivals, virtual departures from the first queue and that from the second queue respectively and have p' and $1-p'$ proportions of customers routed out of the network and to the second queue respectively from the outout of the first queue. Then, as in the last subsection, we can argue heuristically that, for large N, (up to logarithmic equivalence)

$$\alpha \approx \exp\{-N. \inf_{\substack{\lambda' \geq 0, \mu'_1 \geq 0, \mu'_2 \geq 0, 0 \leq p' \leq 1 \\ \lambda'+\mu'_1+\mu'_2 = 1 \\ \lambda' > \mu'_1 \text{ or } \lambda'.(1-p') > \mu'_2}} [R(\lambda', \mu'_1, \mu'_2, p').(\lambda'.h_\lambda(\frac{1}{\lambda'}) + \mu'_1.h_{\mu_1}(\frac{1}{\mu'_1}) + \mu'_2.h_{\mu_2}(\frac{1}{\mu'_2}) + K(\mathbf{p}', \mathbf{p}))]\}, \qquad (54)$$

where

$$K(\mathbf{p}', \mathbf{p}) = p'.\log(\frac{p'}{p}) + (1-p').\log(\frac{1-p'}{1-p})$$

and (when $\lambda' > \mu'_1$ or $\lambda'.(1-p') > \mu'_2$)

$$R = \begin{cases} 1/(\lambda'-\mu'_1), \text{ if } \lambda' > \mu'_1 \text{ and } \mu'_1.(1-p') \leq \mu'_2, \\ 1/((\lambda'-\mu'_1)+(\mu'_1.(1-p')-\mu'_2)), \text{ if } \lambda' > \mu'_1 \text{ and } \mu'_1.(1-p') > \mu'_2, \\ 1/((\lambda'.(1-p')-\mu'_2), \textit{otherwise}. \end{cases}$$

Numerical minimization gives us λ^*, μ_1^*, μ_2^* and p^* as the parameters of the network obtained by an optimal exponential change of measure. Examples show that the node with higher traffic intensity blows up while the other one remains stable. The limiting case occurs when the traffic intensities are equal (see § 4.4).

Table (9) lists some illustrations of simulation speed-ups when simulated under the transformed system. It also shows the time required for the direct computation of α by solving the first step equations, the time required for a simulation and the corresponding number of calls to the random number generator (RNG).

Table (10) gives the empirical standard deviations, means and coefficients of variation of the estimates obtained by the change of measure for the same examples as in Table (9). All the simulations were done on a VAX-750 machine and the first step equations were solved using the IMSL routine LEQT2F.

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda = 0.20\ \mu_1 = 0.30\ \mu_2 = 0.50\ p = 0.10\ N = 20$ $\alpha = 3.269\times10^{-4}$ *CPU Time* = 288.8*Sec.* $\lambda^* = 0.30\ \mu_1^* = 0.20\ \mu_2^* = 0.50\ p^* = 0.10$						
# of Cycles (n)	5000	10000	15000	1000	1500	2000
$\alpha_n\ (\alpha_n^*)$	0.0	2.000×10^{-4}	2.667×10^{-4}	3.097×10^{-4}	3.479×10^{-4}	3.272×10^{-4}
CPU Time	28.6*Sec.*	60.1*Sec.*	79.4*Sec.*	25.8*Sec.*	39.3*Sec.*	53.0*Sec.*
Calls to RNG	84434	170263	245106	67941	106382	138763
Example-II $\lambda = 0.20\ \mu_1 = 0.60\ \mu_2 = 0.20\ p = 0.50\ N = 20$ $\alpha = 2.349\times10^{-6}$ *CPU Time* = 281.7*Sec.* $\lambda^* = 0.30\ \mu_1^* = 0.60\ \mu_2^* = 0.10\ p^* = 0.33$						
# of Cycles (n)	5000	10000	20000	500	1000	1500
$\alpha_n\ (\alpha_n^*)$	0.0	0.0	0.0	2.441×10^{-6}	2.447×10^{-6}	2.363×10^{-6}
CPU Time	15.7*Sec.*	31.3*Sec.*	63.4*Sec.*	14.4*Sec.*	31.8*Sec.*	45.0*Sec.*
Calls to RNG	45873	91234	193429	37768	80186	119749
Example-III $\lambda = 0.10\ \mu_1 = 0.70\ \mu_2 = 0.20\ p = 0.20\ N = 20$ $\alpha = 2.390\times10^{-8}$ *CPU Time* = 295.0*Sec.* $\lambda^* = 0.22\ \mu_1^* = 0.70\ \mu_2^* = 0.08\ p^* = 0.09$						
# of Cycles (n)	10000	30000	50000	1000	2000	5000
$\alpha_n\ (\alpha_n^*)$	0.0	0.0	0.0	2.364×10^{-8}	2.595×10^{-8}	2.425×10^{-8}
CPU Time	20.7*Sec.*	60.3*Sec.*	101.8*Sec.*	25.2*Sec.*	57.2*Sec.*	79.3*Sec.*
Calls to RNG	63703	193930	319010	68535	141488	210525

Example of a Network with Routing (see Figure (14)).

Table (9)

Example-I $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $p = 0.10$ $N = 20$ $\alpha = 3.269\times10^{-4}$ *# of Experiments* = 20 $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$ $p^* = 0.10$		
# of Cycles (n)	1500	2000
Empirical Mean ($\hat{m}$)	3.194×10^{-4}	3.255×10^{-4}
Empirical Std. Dev. ($\hat{\sigma}$)	1.370×10^{-5}	1.011×10^{-5}
($\hat{\sigma}$/ $\hat{m}$)×100 %	4.288 %	3.107 %
Example-II $\lambda = 0.20$ $\mu_1 = 0.60$ $\mu_2 = 0.20$ $p = 0.50$ $N = 13$ $\alpha = 2.349\times10^{-6}$ *# of Experiments* = 20 $\lambda^* = 0.30$ $\mu_1^* = 0.60$ $\mu_2^* = 0.10$ $p^* = 0.33$		
# of Cycles (n)	1000	1500
Empirical Mean ($\hat{m}$)	2.366×10^{-6}	2.333×10^{-6}
Empirical Std. Dev. ($\hat{\sigma}$)	1.485×10^{-7}	1.294×10^{-7}
($\hat{\sigma}$/ $\hat{m}$)×100 %	6.278 %	5.546 %
Example-III $\lambda = 0.10$ $\mu_1 = 0.70$ $\mu_2 = 0.20$ $p = 0.20$ $N = 20$ $\alpha = 2.390\times10^{-8}$ *# of Experiments* = 20 $\lambda^* = 0.22$ $\mu_1^* = 0.70$ $\mu_2^* = 0.08$ $p^* = 0.09$		
# of Cycles (n)	2000	3000
Empirical Mean ($\hat{m}$)	2.405×10^{-8}	2.390×10^{-8}
Empirical Std. Dev. ($\hat{\sigma}$)	5.110×10^{-10}	4.357×10^{-10}
($\hat{\sigma}$/ $\hat{m}$)×100 %	2.125 %	1.823 %

Empirical Standard Deviation for the Examples of Table (9).

Table (10)

4.4. Some Observations.

(I) On M/ M/ 1 Queues in Tandem :

(a) If the set of arguments for the minimization in Eqn. (52) is not unique, i.e., if there are more than one set of parameters $(\lambda^*,\mu_1^*,\mu_2^*)$ then, even for large N, it is not possible to have a single most dominant "tube" of paths in S. This case occurs when $\mu_1 = \mu_2$. For example, for $(\lambda = 0.20,\mu_1 = 0.40,\mu_2 = 0.40)$-network, we get (0.40,0.40,0.20) and (0.40,0.20,0.40) as two sets of optimal parameters. In

this limiting case the speed-up due to the change of measure is less than that for the examples shown in Table (5) (at least for the "small" N's that were feasible for us to consider). e.g. for $N = 20$, $\alpha = 1.812\times10^{-5}$. After simulating (0.40,0.20,0.40)-network for 20000 cycles we obtained $\alpha_n^* = 1.764\times10^{-5}$ as an estimate. Our estimates had intolerable errors for less number of cycles. In summary, if $\mu_1 = \mu_2$ then we have observed speed-ups as compared to the direct simulations but they are less than that for the examples of Table (5). Furthermore, if μ_1 and μ_2 are not much apart then we need larger N's to isolate the dominant "tube" of S. In this case, we may not get very reliable estimates for small N's without running the simulations for relatively more (as compared to the numbers in Table (5)) number of cycles under the changed measure.

(b) It follows from the result of Weber [7] that (λ,μ_1,μ_2)-network and (λ,μ_2,μ_1)-network have identical α's for all N. For sufficiently large N, we have observed that it may be better to start with (λ,μ_1,μ_2)-network with $\mu_1 \geq \mu_2$. Then the corresponding $(\lambda^*,\mu_1^*,\mu_2^*)$-network as given by our heuristic will be the interchange of λ with μ_2. For example, for $(\lambda = 0.10,\mu_1 = 0.50,\mu_2 = 0.40)$-network $\alpha = 1.327\times10^{-14}$ for $N = 25$. A simulation of (0.40,0.50,0.10)-network gives 1.265×10^{-14} after 20000 cycles while a simulation of (0.40,0.10,0.50)-network gives 1.114×10^{-14} after 40000 cycles.

(II) On M/ M/ 1 Queues in Parallel :

As in the previous observation, we have the limiting case when the set of arguments for the minimization in Eqn. (53) is not unique. In this case, it is not clear which one is the optimal set of arguments. For example, the $(\lambda_1 = 0.20,\mu_1 = 0.30 \mid \lambda_2 = 0.20,\mu_2 = 0.30)$-network has three sets of minimizing arguments, namely, (0.30,0.20 | 0.20,0.30), (0.30,0.20 | 0.30,0.20) and (0.20,0.30 | 0.30,0.20). For $N = 25$, $\alpha = 4.156\times10^{-4}$ After 20000 cycles, these networks gave 3.337×10^{-4}, 3.855×10^{-4} and 3.100×10^{-4} respectively. It seems to us that in this limiting case, it might be faster to simulate the network where for both the queues arrival and service rates have been interchanged. This

observation also suggests that if the traffic intensities of the two queues are not much apart then we will require larger N's to single out the dominant "tube" of paths of S.

(III) On the Network shown in Figure (14) :

If the traffic intensities of the two queues are not much apart the we need larger N's for our method of simulation to be effective.

4.5. Extension to Networks of GI/G/1 Queues.

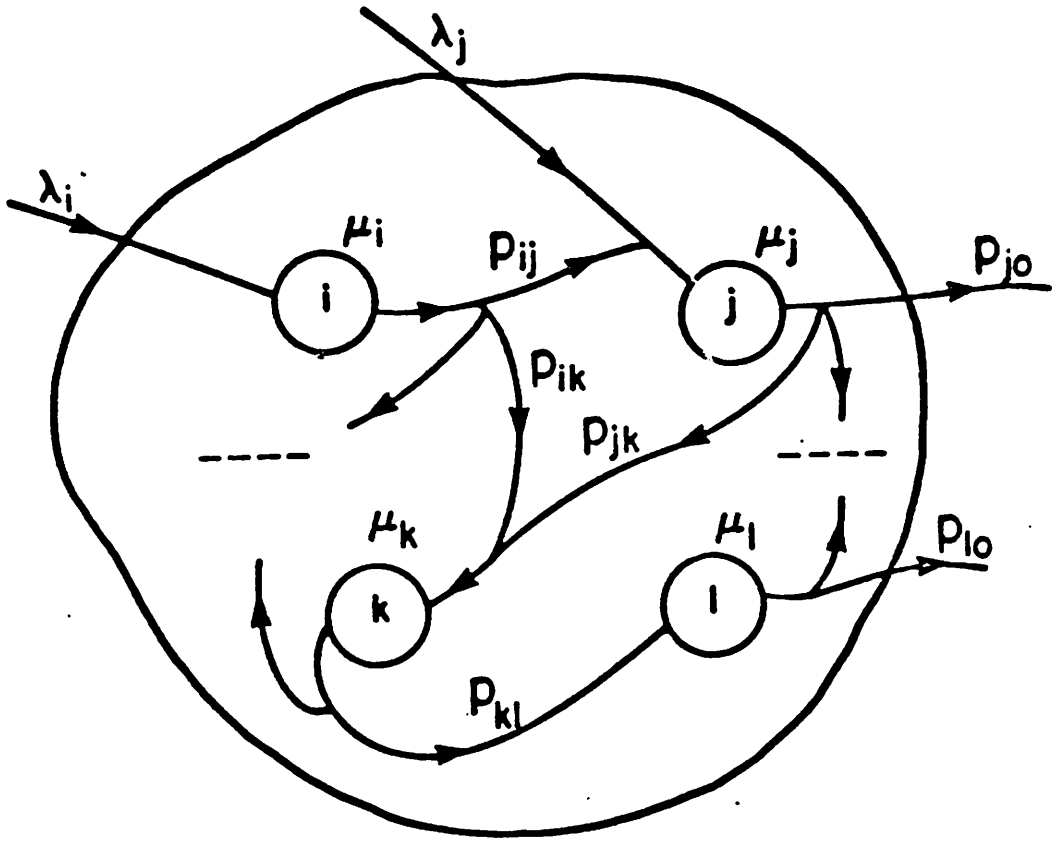
In this subsection, we extend the heuristic of the previous four subsections to networks of $GI/G/1$ queues. Observe that for estimating α, we no longer have an embedded Markov chain to work with. Now we have to simulate the network in real time, i.e., by generating various random times (service times and interarrival times).

Consider a general open network of $GI/G/1$ queues shown in Figure (15).

Suppose there are $d > 0$ nodes. Let $1/\lambda_i, 1 \leq i \leq d$ and $1/\mu_i, 1 \leq i \leq d$ denote the means of G_i, the interarrival time d.f. of the external input process to the node i and F_i, the service time d.f. at the node i. Let p_{ij} denote the probability of routing from the node i to the node j. By p_{io} we denote the probability of leaving the network after the service completion at the node i.

Consider the paths of S which require T time units to have the backlog build up to N, have $1/\lambda'_i$ and $1/\mu'_i$ average interarrival times and virtual service times respectively and have $\mathbf{P}' = \{p'_{ij}\}$ as the apparent routing probabilities. Let λ' and μ' denote the d-dimensional vectors $\{\lambda'_i\}$ and $\{\mu'_i\}$ respectively. Let $\gamma' = \{\gamma'_i\}$ denote the effective rate for this paths which we can find approximately (because μ'_i's are the virtual service rates) by solving the "flow balance equations"

$$\gamma'_i = \lambda'_i + \sum_{j=1}^{d} min(\gamma'_j, \mu'_j).p'_{ji}, \quad 1 \leq i \leq d. \tag{55}$$



Network of *GI/ G/* 1 Queues.

Figure (15)

As in the previous subsections, we can argue heuristically to get the following relationship between T, γ' and μ'.

$$T \approx N.R,$$

where

$$R = \frac{1}{\sum_{i=1}^{d}(\gamma'_i-\mu'_i).1\{\gamma'_i > \mu'_i\}}.$$

Finally, the same line of heuristic gives (up to logarithmic equivalence)

$$\alpha = \sum_{\lambda',\mu',P'} \exp\{-N.H(\lambda',\mu',P')\},$$

where

$$H(\lambda',\mu',P') = R.\sum_{i=1}^{d}\lambda'_i.h_{A_i}(\frac{1}{\lambda'_i})+\sum_{i=1}^{d}\mu'_i.h_{S_i}(\frac{1}{\mu'_i})+\sum_{i=1}^{d}\min(\gamma'_i,\mu'_i).K(\mathbf{p}'_i,\mathbf{p}_i)$$

and $\mathbf{p}'_i$ and $\mathbf{p}_i$ are the i^{th} rows of the matrices P' and P respectively.

Hence, for large N, (up to logarithmic equivalence)

$$\alpha = \exp\{-N.H^*\},$$

where

$$H^* = \inf_{\lambda',\mu',P'} H(\lambda',\mu',P')$$

with γ' given by Eqn. (55).

Let λ^*, μ^* and P^* denote the arguments achieving this infimum. Define new service time distributions F_i^*'s by

$$dF_i^*(z) = \frac{e^{\theta \cdot z} dF_i(z)}{E\{e^{\theta \cdot F_i}\}},$$

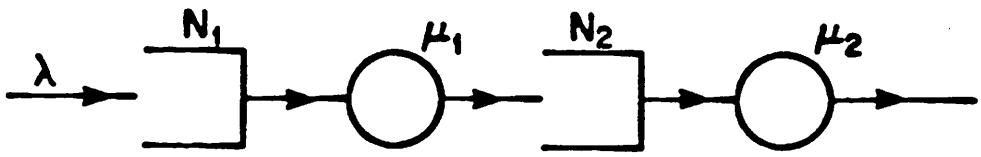
where θ is such that it satisfies $\int z dF_i(z) = 1/\mu_i^*$. Similarly define new interarrival time distributions G_i^*'s by

$$dG_i^*(z) = \frac{e^{\theta \cdot z} dG_i(z)}{E\{e^{\theta \cdot G_i}\}},$$

where θ is such that it satisfies $\int z dG_i(z) = 1/\lambda_i^*$. Then, for large N, we propose to use the network of $GI/G/1$ queues with the parameters λ^*, μ^* and P^* for estimating α.

4.6. Extension to Networks of Queues with Finite Buffers and Optimal Design Problem.

We can extend the heuristic of the previous subsections to networks of queues with finite buffers. We illustrate this by considering two $M/M/1$ queues with finite buffers in tandem. Let λ, μ_1 and μ_2 denote arrival and service rates of the queues respectively. Let N_1 and N_2 be the sizes of the buffers at the first and second queue respectively. Such a network is shown in Figure (16). Let $N_1+N_2 = N$ and $\beta = N_1/N$.



$$\lambda + \mu_1 + \mu_2 = 1$$
$$N_1 + N_2 = N \qquad \lambda < \mu_1 \qquad \lambda < \mu_2$$

M / M / 1 Queues with Finite Buffers in Tandem.

Figure (16)

Now α denotes the probability that one of the two queues exceeds its buffer capacity before returning to 0 given that the system starts empty. The heuristic of § 4.2 can be modified to give (up to logarithmic equivalence)

$$\alpha \approx \sum_{\substack{\lambda'>0,\mu'_1\geq 0,\mu'_2\geq 0 \\ \lambda'+\mu'_1+\mu'_2=1 \\ \lambda'>\mu'_1 \text{ or } \lambda'>\mu'_2}} \exp\{-T(\lambda',\mu'_1,\mu'_2).(\lambda'.h_\lambda(\frac{1}{\lambda'})+\mu'_1.h_{\mu_1}(\frac{1}{\mu'_1})+\mu'_2.h_{\mu_2}(\frac{1}{\mu'_2}))\},$$

where, $T(\lambda',\mu'_1,\mu'_2) \approx N.R(\lambda',\mu'_1,\mu'_2)$ with (when $\lambda' > \mu'_1$ or $\lambda' > \mu'_2$)

$$R = \begin{cases} \beta/(\lambda'-\mu'_1), \text{ if } \lambda' > \mu'_1 \text{ and } \mu'_1 \leq \mu'_2, \\ \min(\beta/(\lambda'-\mu'_1), (1-\beta)/(\mu'_1-\mu'_2)), \text{ if } \lambda' > \mu'_1 \text{ and } \mu'_1 > \mu'_2, \\ (1-\beta)/(\lambda'-\mu'_2), \textit{otherwise.} \end{cases}$$

Then, for large N, (up to logarithmic equivalence)

$$\alpha = \exp\{-N. \inf_{\substack{\lambda'>0,\mu'_1\geq 0,\mu'_2\geq 0 \\ \lambda'+\mu'_1+\mu'_2=1 \\ \lambda'>\mu'_1 \text{ or } \lambda'>\mu'_2}} [R(\lambda',\mu'_1,\mu'_2).(\lambda'.h_\lambda(\frac{1}{\lambda'})+$$

$$\mu'_1.h_{\mu_1}(\frac{1}{\mu'_1})+\mu'_2.h_{\mu_2}(\frac{1}{\mu'_2}))]\}. \qquad (56)$$

We propose to simulate the system with the parameters λ^*, μ_1^* and μ_2^*, which are the arguments giving the minimization in Eqn. (56).

We finally describe how one can use this simulation technique for an optimal design problem where we want to allocate N_1 and N_2 with the constraint $N_1+N_2 = N$ such that α is minimized. There is a well-known conjecture (amply justified by numerical examples) that the function $\alpha(N_1) := \alpha(N_1, N-N_1)$, the probability α when the buffer allocation is N_1 and $N-N_1$, is of the cup-shape. More precisely, on the range $1 \leq N_1 \leq N-1$, $\alpha(N_1)$ has either a unique minimum or two minima at the consecutive points N_1^* and N_1^*+1. The latter case occurs when $\mu_1 = \mu_2$ and $N = N_1+N_2$ is an odd integer. In that case, the minima of $\alpha(N_1)$ occur at two integers adjacent to $N/2$. We can use this observation to select N_1 and N_2 optimally by finding $\alpha(N_1)$'s by our simulation technique. By estimating $\alpha(N_1)$'s, say, for $N_1 = k$ and $N_1 = k+1$, we can decide if we need to check below k or above $k+1$ for finding the minimum value of α.

5. Conclusions.

In this paper, we have used some techniques inspired by Large Deviation theory for obtaining a simulation method for events of excessive backlogs in networks of queues that is much faster than the direct Monte Carlo simulation method. We have seen that the classical Large Deviation results of Ventsel and Freidlin are not directly applicable to networks of queues because of discontinuous kernels. To circumvent this difficulty, a heuristic method based on the work by Borovkov, Ruget etc. for a $GI/G/1$ queue has been developed for the simulation purpose and has also been extended to open networks of $GI/G/1$ queues as well as to networks of queues with finite buffers. Further work is needed to justify analytically our heuristic method and also to connect the transient and steady state behaviors for rare events in networks of queues.

References:

[1] Billingsley, P. (1968) Convergence of Probability Measures. J. Wiley & Sons.

[2] Cottrell, M., Fort, J.-C. and Malgouyres, G. (1983) Large Deviations and Rare Events in the Study of Stochastic Algorithms. IEEE Trans. A.C., vol. AC-28, No. 9, 907-920.

[3] Ruget, G. (1979) Quelques Occurences Des Grands Ecarts Dans La Littérature Electronique". Asterisque 68, Soc. Math. France, 187-199.

[4] Sanov, I. (1957) On the Probability of Large Deviation of Random Variables. Sel. Trans. Math. Statist. Prob., I, 213-244.

[5] Varadhan, S.R.S. (1984) Large Deviations and Applications. SIAM.

[6] Ventsel, A.D. (1976) Rough Limit Theorems on Large Deviation for Markov Stochastic Processes - II. Theory Prob. Appl. (USSR), vol. 21, 499-512.

[7] Weber, R. (1979) The Interchangeability of ./ M / 1 Queues in Series. J. App. Prob. 16, 690-695.