Asymptotically Optimal Importance Sampling for Product-Form Queuing Networks

KEITH W. ROSS and JIE WANG
University of Pennsylvania

Monte Carlo integration is applied to the integral representation of the normalization constant for a family of product-form multiclass queuing networks. These networks are closed with single-server, fixed-rate stations and at least one infinite-server station. Letting the population for each class go to infinity, the asymptotically optimal importance-sampling distributions are derived for estimates of the normalization constant and of the utilizations. In the normal-usage regime, in which the asymptotic utilizations at the single-server stations are strictly less than one, independent exponential sampling is asymptotically optimal for estimating the normalization constant. In the same regime, the asymptotically optimal sampling distribution for estimating utilization is complex and dependent across single-server stations. In the critical-usage regime, in which the asymptotic utilizations at the single-server stations are equal to one, truncated multivariate normal sampling is asymptotically optimal for estimating the normalization constant.

Categories and Subject Descriptors: D.4.8 [**Operating Systems**]: Performance—*queuing theory*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*probabilistic algorithms*

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Importance sampling, Monte Carlo integration, queuing networks

1. INTRODUCTION

In 1975, Baskett et al. [1975] showed that for a class of Markovian queuing networks, referred to as *product-form networks*, the equilibrium probabilities can be expressed as the product of known terms, normalized by a constant. This *normalization constant* is nontrivial to calculate for closed multichain networks. Since most performance measures of interest—including utilizations, throughputs, and derivatives of these measures—can be expressed in terms of the normalization constant, many researchers (e.g., see Conway and Georganas [1989], Lam and Lien [1983], Reiser and Kobayashi [1975], Reiser and Lavenberg [1980], Sauer and Chandi [1981]) have developed combinatorial algorithms to calculate it. Unfortunately, all of these algorithms are

Author's address: Department of Systems, University of Pennsylvania, Philadelphia, PA 19104; email: ross@eniac.seas.upenn.edu and wangj@donna.seas.upenn.edu.

nonpolynomial in the problem size and cannot be applied when the number of stations, classes, and customers are moderately large.

From the early to the mid 1980s, McKenna and Mitra [1982; 1984] and Ramakrishnan and Mitra [1982] developed an entirely different approach to calculating the normalization constant and performance measures. Their approach essentially has three steps. The first step converts the sum representation of the normalization constant into an integral representation. The second step expresses the integral as an asymptotic expansion. The third step calculates the coefficients for the leading terms of the asymptotic expansion by solving "pseudo networks," which are small networks that can be handled by the combinatorial algorithms. This approach has led to the development of a software package, PANACEA, which applies to closed multiclass queuing networks with each station being either a First-Come-First-Serve (FCFS) station or an Infinite Server (IS) station. PANACEA works remarkably well [McKenna and Mitra 1982] for moderate to large networks satisfying the *normal usage conditions*, which require that each class pass through an IS station and that the FCFS stations not be heavily loaded. However, if the normal usage conditions are not satisfied with "sufficient margin," then PANACEA will typically give upper and lower bounds for the performance measures that are unacceptably loose [Ross et al. 1991]. Morrison and Mitra [1985] extended the asymptotic expansion to critical usage for a network consisting of one IS station and one processor-sharing station with multiple classes. Mitra [1992] also studied a single-class network with multiple FCFS stations and one IS station for three traffic conditions—namely, normal usage, critical usage, and heavy usage. However, asymptotic expansions for general multiclass multistation networks do not seem to be in the existing literature.

Knessl and Tier [1990] and [1992] also studied the asymptotic expansions for closed queuing networks but in a different asymptotic regime. Instead of networks with a large number of customers, they considered networks with a large number of service stations.

Ross and Wang [1990] proposed the application of Monte Carlo summation to the calculation of the sum representation for the normalization constant and of related performance measures. Ross et al. [1991] proposed the application of Monte Carlo integration to the calculation of the integral representation for the normalization constant and of related performance measures. Monte Carlo summation can be applied to networks containing all the station types of the product-form networks [Baskett et al. 1975], including multiple-server stations with load-dependent rates. But the integral representation considered in Ross et al. and in this article only applies to networks with single-server, constant-rate stations and infinite-server stations.

The basic idea behind the Monte Carlo techniques is to sample the domain of the function being summed or integrated, evaluate the function at the samples, and then take the average to form an estimate. The success of the method greatly depends on the choice of the *importance-sampling density*. In particular, it is desirable to employ a density that is easy to sample from, but

has a small variance for the estimator. Numerical testing in Ross et al. [1991] indicates:

—*Independent exponential sampling* (to be rigorously defined in Section 2), with specific sampling parameters, is easy to sample from and gives small variance for the integral representation.
—Monte Carlo integration with independent exponential sampling typically gives smaller confidence intervals than does Monte Carlo summation (for the class of importance-sampling densities considered) for the same amount of CPU time.
—Both PANACEA and Monte Carlo integration with independent exponential sampling give excellent results when the normal usage conditions are satisfied with sufficient margin (with the amount of margin required decreasing with the population sizes).
—If the normal traffic conditions are satisfied with little margin, then the bounds of PANACEA become loose, whereas Monte Carlo Integration with independent exponential sampling continues to give excellent results.
—If the normal traffic conditions are fully violated (e.g., if there are no IS stations), then Monte Carlo integration with independent exponential sampling can still be employed, but the requisite CPU time significantly—but not exorbitantly—increases.

The purpose of this article is twofold: First, to supply some theoretical justification for the choice of independent exponential sampling when the normal usage conditions are satisfied; second, to develop a new importance-sampling density and justify it theoretically when normal usage *is not satisfied.* The theoretical justification will be given in the context of the asymptotic regime introduced by Ramakrishnan and Mitra [1982]. This regime keeps the number of stations and classes fixed, but lets the population sizes of the various classes go to infinity.

After reviewing Monte Carlo integration and the asymptotic regime of Ramakrishnan and Mitra [1982] in Section 2, we consider asymptotically optimal importance sampling for the normalization constant in Section 3. Under normal usage, we show that independent exponential sampling (with specific parameters) has an asymptotic relative variance of zero, whereas all other densities have strictly positive asymptotic variance. In Section 4 we consider asymptotically optimal importance sampling for estimating utilization under the normal usage conditions. In this case, independent exponential sampling is not asymptotically optimal. An explicit expression for the asymptotically optimal density is derived; unfortunately, this density is difficult to sample from since it calls for dependent sampling across stations. All hope is not lost, however, since the numerical testing in Ross et al. [1991] indicates that independent exponential sampling, with appropriate sampling parameters, is close to optimal when estimating utilization. In Section 4 we also outline a procedure for obtaining the asymptotically "best" sampling parameters for independent exponential sampling. In Section 5, we consider estimation of the normalization constant in the same asymptotic regime, but now

assume "critical usage" rather than normal usage. Again, the asymptotically optimal sampling density is derived: it is a multivariate normal distribution truncated to the positive quadrant. We conclude in Section 6.

We should also mention that Ross and Wang [1990] and [1992] have also applied Monte Carlo summation to product-form loss networks.

2. PRELIMINARIES

We consider closed multichain queuing networks consisting of FCFS stations and IS stations. Let J denote the number of classes and M the number of stations. Let the population size for the jth class be denoted by N_j. The service rate for each server at station m is denoted by μ_m; the service distribution at FCFS stations is assumed to be exponential. For each class j, let λ_{jm} be the relative visit ratio of a customer to service station m. Let $\rho_{jm} := (\lambda_{jm}/\mu_m)$. The state of the system is denoted by $\mathbf{n} := (n_{jm}: 1 \le j \le J, 1 \le m \le M)$, where n_{jm} denote the number of class-j customers at service station m. The set of all possible states is given by

$$\Omega := \{\mathbf{n}: n_{j1} + , \ldots, + n_{jM} = N_j, j = 1, \ldots, J\}.$$

Let $\mathbf{n}_m := (n_{1m}, \ldots, n_{Jm})$ be the state of service station m and $n_m := n_{1m} + \cdots + n_{Jm}$ be the number of customers present at service station m. It is well known [Baskett et al. 1975] that the equilibrium probability of being in state $\mathbf{n} \in \Omega$ has the following product form:

$$\pi(\mathbf{n}) = \frac{1}{g} \prod_{m=1}^{M} f_m(\mathbf{n}_m),$$

where

$$f_m(\mathbf{n}_m) = \begin{cases} n_m! \prod_{j=1}^{J} \dfrac{\rho_{jm}^{n_{jm}}}{n_{jm}!} & \text{if station } m \text{ is FCFS} \\ \prod_{j=1}^{J} \dfrac{\rho_{jm}^{n_{jm}}}{n_{jm}!} & \text{if station } m \text{ is IS,} \end{cases}$$

and

$$g = \sum_{\mathbf{n} \in \Omega} \prod_{m=1}^{M} f_m(\mathbf{n}_m). \tag{1}$$

The constant g is referred to as the *normalization constant*. Let g_j denote the normalization constant for the same queuing network with one less class-j customer. It is well known [Baskett et al. 1975] that the average utilization of the server by class-j customers at FCFS station m is given by

$$\text{util}_{jm} = \rho_{jm} \frac{g_j}{g}. \tag{2}$$

Let the first L, $1 \le L \le M$, service station be FCFS stations and the last $M - L$ be IS stations. McKenna and Mitra [1982] have given the following

integral representation of the normalization constant:

$$g = \frac{1}{\prod_{j=1}^{J} N_j!} \int_{\mathbf{Q}^+} e^{-\mathbf{1}'\mathbf{u}} \prod_{j=1}^{J} (\rho_{j0} + \boldsymbol{\rho}_j' \boldsymbol{u})^{N_j} \, d\boldsymbol{u}, \tag{3}$$

where

$$\begin{aligned}
\mathbf{u} &= (u_1, \ldots, u_L)' \\
\mathbf{1} &= (1, \ldots, 1)' \\
\rho_{j0} &= \sum_{m=L+1}^{M} \rho_{jm}, \qquad 1 \le j \le J \\
\boldsymbol{\rho}_j &= (\rho_{j1}, \ldots, \rho_{jL})' \\
\mathbf{Q}^+ &= \{\boldsymbol{u} \in R^L : u_l \ge 0, l = 1, \ldots, L\}.
\end{aligned}$$

It follows that

$$g_j = \frac{N_j}{\prod_{k=1}^{J} N_k!} \int_{\mathbf{Q}^+} e^{-\mathbf{1}'\boldsymbol{u}} \frac{\prod_{k=1}^{J} (\rho_{k0} + \boldsymbol{\rho}'_k \boldsymbol{u})^{N_k}}{\rho_{j0} + \boldsymbol{\rho}_j' \boldsymbol{u}} \, d\boldsymbol{u}. \tag{4}$$

2.1 Monte Carlo Integration

Monte Carlo summation was introduced in Ross and Wang [1990] as a means of estimating the sum representation of the normalization constant (1) and the performance measures expressed in terms of sums. Monte Carlo integration was introduced in Ross et al. [1991] as a means of estimating the integral representation of normalization constant (3) and the performance measures expressed in terms of integrals. Computational testing in Ross et al. [1991] indicates that Monte Carlo integration almost always gives narrower confidence intervals for the same amount of CPU time. We therefore focus on Monte Carlo integration.

We now briefly review Monte Carlo integration. Let $\mathbf{V}^i = (V_1^i, \ldots, V_L^i)$, $i = 1, 2, \ldots$, be a sequence of independent and identically distributed L-dimensional random vectors with probability density function $p(\cdot)$ defined over $\mathbf{Q}^+$. Let

$$Z^i := \frac{e^{-\mathbf{1}'\mathbf{V}^i} \prod_{j=1}^{J} \left(\rho_{j0} + \boldsymbol{\rho}_j' \mathbf{V}^i\right)^{N_j}}{p(\mathbf{V}^i)} \frac{1}{\prod_{j=1}^{J} N_j!}$$

and

$$\bar{Z}^I := \frac{1}{I} \sum_{i=1}^{I} Z^i.$$

Then $\bar{Z}^I$ is an unbiased and consistent estimator of g. Similarly, if we define

$$Z_j^i := \frac{e^{-\mathbf{1}'\mathbf{V}^i} \prod_{k=1}^{J} (\rho_{k0} + \boldsymbol{\rho}_k' \mathbf{V}^i)^{N_k}}{\left(\rho_{j0} + \boldsymbol{\rho}_j' \mathbf{V}^i\right) p(\mathbf{V}^i)} \frac{N_j}{\prod_{j=1}^{J} N_j!},$$

then

$$\Phi_{jm}^{I} := \rho_{jm} \frac{\sum_{i=1}^{I} Z_j^i}{\sum_{i=1}^{I} Z^i}$$

is a consistent estimator for util_{jm}. Although we only consider utilization, the arguments in this article easily extended to other important performance measures, such as throughput.

There is great latitude in the choice of $p(\cdot)$. The density

$$p(\boldsymbol{u}) = \frac{1}{g} e^{-1'u} \prod_{j=1}^{J} (\rho_{j0} + \boldsymbol{\rho}_j' \boldsymbol{u})^{N_j} \frac{1}{\prod_{j=1}^{J} N_j!}, \qquad \boldsymbol{u} \in \mathbf{Q}^+$$

is optimal for estimating g since it gives $Z^i = g$ at each iteration; but it is impossible to sample from without knowing g. In Ross et al. [19xx] attention is focused on independent exponential sampling densities, i.e., densities of the form

$$q_{\gamma}(\boldsymbol{u}) = \prod_{l=1}^{L} \gamma_l e^{-\gamma_l u_l}, \qquad \boldsymbol{u} \in \boldsymbol{Q}^+,$$

where the γ_l's are positive numbers and are referred to as the *sampling parameters*; let $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_L)$ be the *importance-sampling vector*. One feature of this density is that it is easy to sample from. Furthermore, numerical testing [Ross et al.] indicates that this sampling density leads to narrow confidence intervals when the sampling parameters are given by $\hat{\gamma}_l = 1 - \text{util}_l^*$, $l = 1, \ldots, L$ where util_l^* is a good approximation of the utilization at station l.

The queuing network is said to satisfy the *normal usage conditions* [Ramakrishnan and Mitra 1982] if:

$$\rho_{j0} > 0 \qquad \text{for all} \quad j = 1, \ldots, J \tag{5}$$

$$\sum_{j=1}^{J} N_j \frac{\rho_{jl}}{\rho_{j0}} < 1 \qquad \text{for all} \quad l = 1, \ldots, L. \tag{6}$$

Note that the normal usage conditions require that every class passes through an IS station and that the FCFS stations not be heavily loaded. Define

$$y_l := \sum_{j=1}^{J} N_j \frac{\rho_{jl}}{\rho_{j0}}, \qquad l = 1, \ldots, L$$

and $\mathbf{y} = (y_1, \ldots, y_L)$. The "margin" alluded to in the Introduction can be defined as $\min\{1 - y_l : l = 1, \ldots, L\}$. An approximation for the utilization at station l is given by $\text{util}_l^* = y_l$ when the normal usage conditions are satisfied [Ramakrishnan with Mitra 1982]. With this approximation for utilization, our recommended sampling parameters become

$$\hat{\gamma}_l = 1 - y_l, \qquad l = 1, \ldots, L.$$

The computational testing in Ross et al. [1991] shows that this choice of sampling parameters gives remarkably narrow confidence intervals when the normal usage conditions are satisfied.

In this article, we offer some theoretical justification of the independent exponential sampling density $q_{\gamma}(\boldsymbol{u})$ with parameters given by $\boldsymbol{\gamma} = \mathbf{1} - \mathbf{y}$. In particular, we shall show that this choice of sampling density is asymptotically optimal in a natural limiting regime whenever the normal usage conditions are satisfied. However, we shall also show that when all the y_l's are equal to unity (the critical usage condition), then the asymptotically optimal importance-sampling density is not independent exponential but instead "truncated multivariate normal." In particular, independent sampling is not asymptotically optimal in critical usage.

2.2 The Asymptotic Regime

Following Ramakrishnan and Mitra [1982], we now consider an infinite sequence of closed queuing networks indexed by N. Suppose that each of the networks has J classes, L FCFS stations, and $M - L$ IS stations. Suppose that the population size and the relative loads depend on N; in particular, for the Nth network, suppose that

$$N_j = \beta_j N, \qquad j = 1, \ldots, J,$$

and

$$\frac{\rho_{jl}}{\rho_{j0}} = \frac{\Gamma_{jl}}{N}, \qquad j = 1, \ldots, J, \; l = 1, \ldots, L,$$

where the β_j's and Γ_{jl}'s are positive constants. Thus, as $N \to \infty$ the population sizes are increasing, but the relative loads, (ρ_{jl}/ρ_{j0}), are decreasing. Note that each of the networks in the sequence satisfies the normal usage conditions if and only if

$$y_l = \sum_{j=1}^{J} \beta_j \Gamma_{jl} < 1 \qquad \text{for all } l = 1, \ldots, L. \tag{7}$$

Let $g(N)$ be the normalization constant for the Nth network, i.e.,

$$g(N) = \int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\, d\boldsymbol{u} \prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!}$$

where

$$f(\boldsymbol{u}, N) := e^{-\mathbf{1}'\mathbf{u}} \prod_{j=1}^{J} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{N}\right)^{\beta_j N}.$$

We will need to make use of the following technical result, which follows directly from Euler's formula,

$$\lim_{N \to \infty} \left(1 + \frac{x}{N}\right)^N = e^x,$$

and the inequality $(1 + (x/N))^N < e^x$.

LEMMA 1. *For any* $\boldsymbol{u} \in \mathbf{Q}^+$, *we have*

$$f(\boldsymbol{u}, N) \leq e^{-(\mathbf{1}-\mathbf{y})'\boldsymbol{u}}$$

and

$$\lim_{N\to\infty} f(\boldsymbol{u}, N) = e^{-(\mathbf{1}-\mathbf{y})'\boldsymbol{u}}.$$

Let $\text{util}_{jl}(N)$ denote the utilization of class j of the server at FCFS station l for the Nth network. It follows from (2)–(4) that

$$\text{util}_{jl}(N) = y_{jl}\frac{\int_{\mathbf{Q}^+}\left[f(\boldsymbol{u}, N)/f_j(\boldsymbol{u}, N)\right] d\boldsymbol{u}}{\int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\, d\boldsymbol{u}}, \tag{8}$$

where $y_{jl} := \beta_j \Gamma_{jl}$ and

$$f_j(\boldsymbol{u}, N) := 1 + \frac{\sum_l \Gamma_{jl} u_l}{N}.$$

It is not difficult to show from (8) and Lemma 1 that

$$\lim_{N\to\infty} \text{util}_{jl}(N) = y_{jl},$$

which was first established in Ramakrishnan and Mitra [1982].

3. ESTIMATING THE NORMALIZATION CONSTANT: NORMAL USAGE

In this section we study, in the context of the asymptotic regime defined in Section 2.2, the relative variance of our Monte Carlo estimate for the normalization constant as $N \to \infty$. Throughout this section we suppose that the normal usage condition (7) is satisfied, i.e., we assume that $\mathbf{y} < \mathbf{1}$.

Suppose that for the Nth network we use an importance-sampling density $p(\boldsymbol{u}, N)$, $\boldsymbol{u} \in \mathbf{Q}^+$. For a fixed N, our estimate for the normalization constant g is

$$\bar{Z}^I = \frac{1}{I}\sum_{i=1}^{I} \frac{f(\mathbf{V}^i, N)}{p(\mathbf{V}^i, N)}\left(\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!}\right).$$

The relative variance of this estimate is

$$\overline{\text{var}(N)} := \frac{\text{var}(\bar{Z}^I)}{E[\bar{Z}^I]^2},$$

where the expectations are taken with respect to the density $p(\boldsymbol{u}, N)$, $\boldsymbol{u} \in \mathbf{Q}^+$. Note that the relative variance is the square of the coefficient of variation. As mentioned earlier, $p(\boldsymbol{u}, N)$ can be chosen so that the associated relative variance is zero; however, this distribution is difficult to sample from. In this section we will determine the asymptotically optimal density; fortunately, it turns out to be easy to sample from. Recall that $q_\gamma(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbf{Q}^+$, is the independent exponential density with sampling vector $\boldsymbol{\gamma}$.

PROPOSITION 1. *Suppose that for some* $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$ *with* $\gamma_l > 0$, $l = 1, \ldots, L$,

$$\lim_{N\to\infty} p(\boldsymbol{u}, N) = q_\gamma(\boldsymbol{u}) \qquad \textit{for all } \boldsymbol{u} \in \mathbf{Q}^+.$$

(i) *If* $\gamma_l \geq 2(1 - y_l)$ *for some* $l = 1, \ldots, L$, *then*

$$\lim_{N \to \infty} \overline{\text{var}(N)} = \infty;$$

(ii) *If* $\gamma_l < 2(1 - y_l)$ *for all* $l = 1, \ldots, L$, *and if for some* N' *and integrable function* $g(\cdot)$

$$\frac{f^2(\boldsymbol{u}, N)}{p(\boldsymbol{u}, N)} \leq g(\boldsymbol{u}), \qquad \boldsymbol{u} \in \mathbf{Q}^+, \quad N \geq N', \tag{9}$$

then

$$\lim_{N \to \infty} \overline{\text{var}(N)} = \frac{1}{I}\left(\prod_{l=1}^{L} \frac{(1 - y_l)^2}{\gamma_l[2(1 - y_l) - \gamma_l]} - 1 \right).$$

PROOF. Let

$$X^i(N) := \frac{f(\mathbf{V}^i, N)}{p(\mathbf{V}^i, N)}$$

and note that

$$\begin{aligned} \overline{\text{var}(N)} &= \frac{1}{I} \frac{\text{var}(X^i(N))}{E[X^i(N)]^2} \\ &= \frac{1}{I}\left(\frac{E[X^i(N)^2]}{E[X^i(N)]^2} - 1 \right). \end{aligned} \tag{10}$$

By the Dominated Convergence Theorem [Royden 1968] and Lemma 1 we have

$$\begin{aligned} \lim_{N \to \infty} E[X^i(N)] &= \lim_{N \to \infty} \int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\, d\boldsymbol{u} \\ &= \int_{\mathbf{Q}^+} e^{-(1-y)'U}\, d\boldsymbol{u} \\ &= \prod_{l=1}^{L} \frac{1}{1 - y_l}. \end{aligned} \tag{11}$$

Moreover, Fatou's Lemma [Royden 1968] gives

$$\begin{aligned} \lim_{\overline{N \to \infty}} E[X^i(N)^2] &= \lim_{\overline{N \to \infty}} \int_{\mathbf{Q}^+} \frac{f^2(\boldsymbol{u}, N)}{p(\boldsymbol{u}, N)}\, d\boldsymbol{u} \\ &\geq \int_{\mathbf{Q}^+} \lim_{\overline{N \to \infty}} \frac{f^2(\boldsymbol{u}, N)}{p(\boldsymbol{u}, N)}\, d\boldsymbol{u} \\ &= \frac{1}{\prod_{l=1}^{L} \gamma_l} \int_{\mathbf{Q}^+} e^{-[2\mathbf{1} - \gamma - 2\mathbf{y}]'u}\, d\boldsymbol{u}. \end{aligned} \tag{12}$$

But the RHS of (12) is infinite whenever $\gamma_l \geq 2(1 - y_l)$ for some $l = 1, \ldots, L$. Thus, (i) follows from (10), (11), and (12). The condition (9) permits us to invoke also the Dominated Convergence Theorem for $E[X^i(N)^2]$:

$$\lim_{N\to\infty} E[X^i(N)^2] = \frac{1}{\prod_{l=1}^{L}\gamma_l}\int_{\mathbf{Q}^+} e^{-[2\mathbf{1}-\boldsymbol{\gamma}-2\mathbf{y}]'\boldsymbol{u}}\,d\boldsymbol{u}$$
$$= \prod_{l=1}^{L}\frac{1}{\gamma_l(2-\gamma_l-2y_l)}. \tag{13}$$

Thus (ii) follows from (10), (11), and (13). □

COROLLARY 1. *Suppose that for each N the sampling density $p(\boldsymbol{u}, N)$ is an independent exponential density with parameter $\gamma(N)$, i.e.,*

$$p(\boldsymbol{u}, N) = q_{\gamma(N)'(u)}, \qquad \boldsymbol{u} \in \mathbf{Q}^+.$$

Further suppose $\lim_{N\to\infty}\gamma(N) = \boldsymbol{\gamma}$. Then

$$\lim_{N\to\infty}\overline{\mathrm{var}(N)} = \begin{cases} \dfrac{1}{I}\left[\displaystyle\prod_{l=1}^{L}\frac{(1-y_l)^2}{\gamma_l[2(1-y_l)-\gamma_l]} - 1\right] & \text{if } 0 < \gamma_l < (1-y_l) \text{ for all } l \\ \infty & \text{otherwise.} \end{cases}$$

In particular, if $\boldsymbol{\gamma} = \mathbf{1} - \mathbf{y}$, then $\lim_{N\to\infty}\overline{\mathrm{var}(N)} = 0$.

PROOF. It suffices to show that (9) is satisfied when $0 < \gamma_l < 2(1 - y_l)$ for all l. This is a straightforward exercise that is left to the reader. □

PROPOSITION 2. *Suppose that $p(\boldsymbol{u}, N)$, $\boldsymbol{u} \in \mathbf{Q}^+$, converges pointwise to some density $p(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbf{Q}^+$, such that $p(\boldsymbol{u}) \neq q_{1-\mathbf{y}}(\boldsymbol{u})$ for some $\boldsymbol{u} \in \mathbf{Q}^+$. Then $\lim_{N\to\infty}\overline{\mathrm{var}(N)} > 0$.*

PROOF. Recall the definition of $X^i(N)$ in the proof of Proposition 1. From Fatou's Lemma and (11) we have

$$\lim_{N\to\infty}\overline{\mathrm{var}(X^i(N))} \geq \int_{\mathbf{Q}^+} p(\boldsymbol{u})\left\{\frac{e^{-(1-\mathbf{y})'\boldsymbol{u}}}{p(\boldsymbol{u})} - \prod_{l=1}^{L}\frac{1}{1-y_l}\right\}^2 d\boldsymbol{u}$$
$$= \int_{\mathbf{Q}^+}\frac{1}{p(\boldsymbol{u})\prod_{l=1}^{L}(1-y_l)^2}$$
$$\times\left\{\prod_{l=1}^{L}(1-y_l)e^{-(1-y_l)u_l} - p(\boldsymbol{u})\right\}^2 d\boldsymbol{u}.$$

Since $p(\boldsymbol{u})$ is continuous, the RHS of the above expression is equal to zero if and only if $p(\boldsymbol{u}) = q_{1-\mathbf{y}}(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbf{Q}^+$. The proof is completed upon invoking (10) and (11). □

The above results give us great insight into the proper choice of sampling density $p(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbf{Q}^+$. In particular, Corollary 1 and Proposition 2 indicate

that independent exponential sampling with $\gamma = \mathbf{1} - \mathbf{y}$ is likely a good candidate for estimating the normalization constant when the population sizes are large and the normal usage conditions (5) and (6) are satisfied. Under the same conditions, independent exponential sampling with $\gamma_l > 2(1 - y_l)$ for some l is likely a bad choice. Numerical testing in Ross et al. [1991] validates these claims.

Suppose that an independent exponential sampling density $q_\gamma(\boldsymbol{u})$ is used for the Nth network. The problem of minimizing $\text{var}(\bar{Z}^I)$ over $\gamma \in \mathbf{Q}^+$ with N fixed is considered in Ross et al. [1991]. It is shown that there exists a unique $\gamma^*(N) \in [0, 2]^L$ that minimizes $\text{var}(\bar{Z}^I)$. We refer to $\gamma^*(N)$ as the *optimal sampling vector for network N*. It is also shown in Ross et al. that the optimal sampling vector $\gamma^*(N)$ can be determined from the solution of a fixed-point equation. However, the empirical testing in Ross et al. indicates that $\gamma(N) = \mathbf{1} - \mathbf{y}$ gives a variance that is almost as small as that given by $\gamma^*(N)$ (even for small population sizes). The heuristic $\gamma(N) = \mathbf{1} - \mathbf{y}$ is further justified by the following result.

COROLLARY 2.

$$\lim_{N \to \infty} \gamma^*(N) = \mathbf{1} - \mathbf{y}.$$

PROOF. If the result is not true, then there exists a subsequence $\{\gamma^*(N_k)\}$ converging to a point $\gamma^* \neq \mathbf{1} - \mathbf{y}$. By Corollary 1, it follows that there exists an $\epsilon > 0$ and a $K > 0$ such that for all $k \geq K$ the variance associated with $q_{\gamma^*(N_k)^{(u)}}$ is greater than ϵ. Further, it also follows from Corollary 1 that the variance associated with $q_{1-y}(\boldsymbol{u})$ is less than ϵ for all N greater than some N'. Choosing $N \geq N'$ and $N = N_k$ for some $k \geq K$, we arrive at a contradiction since $\gamma^*(N)$ minimizes the variance for the Nth network. □

Note that all of the results of this section hold true for an arbitrary number of replications I.

4. ESTIMATING UTILIZATION: NORMAL USAGE

In this section we study, in the context of the asymptotic regime defined in Section 2.2, the variance of our Monte Carlo estimate of utilization as $N \to \infty$. We again suppose that the normal usage condition (7) is satisfied. In contrast with the estimate of the normalization constant, the asymptotic variance for utilization involves the number of Monte Carlo iterations, I, in a complex manner. Therefore, in order to gain insight into the performance of the Monte Carlo scheme, we shall study the variance of the estimator with both N and I going to infinity.

Throughout this section let k and m be a fixed class and fixed FCFS station, respectively. In order to limit notation, throughout this section we assume that the sampling density $p(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbf{Q}^+$, does not depend on N. All expectations in this section are with respect to this density. We further assume that V_l^i has finite first and second moments for $l = 1, \ldots, L$. Our

estimate for util_{km} becomes

$$\Phi^I(N) = y_{km} \frac{\Sigma_{i=1}^I Y^i(N)}{\Sigma_{i=1}^I X^i(N)},$$

where

$$X^i(N) = \frac{f(\mathbf{V}^i, N)}{p(\mathbf{V}^i)}$$

and

$$Y^i(N) = \frac{f(\mathbf{V}^i, N)/f_k(\mathbf{V}^i, N)}{p(\mathbf{V}^i)}.$$

Let $\Gamma_k' := (\Gamma_{k1}, \ldots, \Gamma_{kL})$. The proof of the following result can be found in the Appendix.

LEMMA 2.

$$E[\Phi^I(N)] = y_{km} - y_{km} \frac{E[W^I]}{N} + o\left(\frac{1}{N}\right) \tag{14}$$

and

$$\text{var}(\Phi^I(N)) = y_{km}^2 \frac{\text{var}(W^I)}{N^2} + o\left(\frac{1}{N^2}\right) \tag{15}$$

where

$$W^I := \frac{\Sigma_{i=1}^I e^{-(1-\mathbf{y})'\mathbf{V}^i} \Gamma_k' \mathbf{V}^i / p(\mathbf{V}^i)}{\Sigma_{i=1}^I e^{-(1-\mathbf{y})'\mathbf{V}^i} / p(\mathbf{V}^i)}.$$

From Lemma 2 we observe that, for large N, the variance of our Monte Carlo estimate can be made relatively small by choosing a sampling density $p(\boldsymbol{u})$ that minimizes $\text{var}(W^I)$. We now consider minimizing $\text{var}(W^I)$ for large I. Define the random variables

$$B := \frac{e^{-(1-\mathbf{y})'\mathbf{V}^i}}{p(\mathbf{V}^i)}$$

and

$$A := \Gamma_k' \mathbf{V}^i B.$$

The proof of the following result can be found in the Appendix.

PROPOSITION 3. *If* $E[A^2] < \infty$ *and* $E[B^2] < \infty$, *then*

$$\lim_{I\to\infty}\lim_{N\to\infty}\sqrt{I}N\left[\Phi^I(N) - \text{util}_{km}(N)\right] \overset{\text{dist.}}{=} \lim_{N\to\infty}\lim_{I\to\infty}\sqrt{I}N\left[\Phi^I(N) - \text{util}_{km}(N)\right]$$
$$\overset{\text{dist.}}{=} N\left[0, h(p)y_{km}^2\prod_{l=1}^{L}(1-y_l)^2\right],$$

where

$$h(p) := E\left[\left\{B\frac{E[A]}{E[B]} - A\right\}^2\right]$$
$$= \int_{\mathbf{Q}^+}\frac{e^{-2(1-\mathbf{y})'\mathbf{u}}\{\sum_{l=1}^{L}\Gamma_{kl}[u_l - (1/1-y_l)]\}^2}{p(\mathbf{u})}\,d\mathbf{u}.$$

Now consider the problem of minimizing $h(p)$ over densities $p(\mathbf{u})$, $\mathbf{u}\in\mathbf{Q}^+$. Let

$$\phi(\mathbf{u}) := e^{-(1-\mathbf{y})'\mathbf{u}}\left|\sum_{l=1}^{L}\Gamma_{kl}\left(u_l - \frac{1}{1-y_l}\right)\right|, \qquad \mathbf{u}\in\mathbf{Q}^+,$$

and consider the problem of estimating $b = \int_{\mathbf{Q}^+}\phi(\mathbf{u})\,d\mathbf{u}$. If we use Monte Carlo integration with importance-sampling density $p(\mathbf{u})$, the $p(\mathbf{u})$ that minimizes the variance of the estimator is $p(\mathbf{u}) = \phi(\mathbf{u})/b$ (e.g., see Kalos and Whitlock [1986]). But minimizing this variance is equivalent to minimizing

$$\int_{\mathbf{Q}^+}\frac{\Phi(\mathbf{u})^2}{p(\mathbf{u})}\,d\mathbf{u}.$$

Thus, the sampling density $p(\cdot)$ that minimizes $h(p)$ is given by

$$p(\mathbf{u}) = \frac{e^{-(1-\mathbf{y})'\mathbf{u}}\left|\sum_{l=1}^{L}\Gamma_{kl}(u_l - 1/(1-y_l))\right|}{\int_{\mathbf{Q}^+}e^{-(1-\mathbf{y})'\mathbf{u}}\left|\sum_{l=1}^{L}\Gamma_{kl}(u_l - 1/(1-y_l))\right|\,d\mathbf{u}}, \qquad \mathbf{u}\in\mathbf{Q}^+.$$

Note that, in contrast with estimating the normalization constant, the asymptotically optimal density does not call for independent sampling. Unfortunately, it is difficult, if not impossible, to efficiently generate samples from the above distribution. We are therefore motivated to consider minimizing $h(p)$ over a class of densities from which samples are easily drawn.

For the remainder of this section we suppose that independent exponential sampling is used, i.e., we suppose $p(\mathbf{u}) = q_{\boldsymbol{\gamma}}(\mathbf{u})$, $\mathbf{u}\in\mathbf{Q}^+$. With slight abuse of notation, we write $h(\boldsymbol{\gamma}) = h(q_{\boldsymbol{\gamma}})$.

COROLLARY 3. *Suppose that* $0 < \gamma_l < 2(1 - y_l)$ *for* $l = 1, \ldots, L$. *Then*

$$h(\boldsymbol{\gamma}) = \frac{1}{\prod_{l=1\gamma_l}^{L}[2 - 2y_l - \gamma_l]} \left\{ \left[\sum_l \frac{\Gamma_{kl}(1 - \gamma_l - y_l)}{(1 - y_l)(2 - \gamma_l - 2y_l)} \right]^2 + \sum_l \frac{\Gamma_{kl}^2}{(2 - \gamma_l - 2y_l)^2} \right\}. \tag{16}$$

Furthermore, $h(\boldsymbol{\gamma})$ *is strictly convex in the region*

$$\Lambda := \{\boldsymbol{\gamma} : 0 < \gamma_l < 2(1 - y_l), l = 1, \ldots, L\}$$

and

$$\lim_{\gamma_l \to 2(1-y_l)^-} h(\boldsymbol{\gamma}) = \lim_{\gamma_l \to 0^+} h(\boldsymbol{\gamma}) = \infty, \qquad l = 1, \ldots, L.$$

Consequently, $h(\boldsymbol{\gamma})$ *has a unique minimum in* Λ.

PROOF. The condition $\gamma_l < 2(1 - y_l)$ for all $l = 1, \ldots, L$ implies that $\text{var}(A) < \infty$ and $\text{var}(B) < \infty$. From Proposition 3 we therefore have

$$h(\boldsymbol{\gamma}) = \int_{\mathbf{Q}^+} \left[\prod_{l=1}^{L} \frac{e^{-(2-2y_l-\gamma_l)u_l}}{\gamma_l} \right] \left[\sum_{l=1}^{L} \Gamma_{kl} \left(u_l - \frac{1}{1 - y_l} \right) \right]^2 d\mathbf{u}, \tag{17}$$

from which (16) follows after evaluating the integral and some algebraic manipulation. The term $\prod_{l=1}^{L} e^{-(2-2y_l-\gamma_l)u_l}/\gamma_l$ is a convex function of $\boldsymbol{\gamma}$. Thus, from (17), $h(\boldsymbol{\gamma})$ is convex. The last statement follows directly from (16). □

Corollary 3 gives us great insight into the proper choice of sampling parameters when both N and I are large. At the very least, the sampling parameters $\boldsymbol{\gamma}$ should satisfy $0 < \gamma_l < 2(1 - y_l)$ for all $l = 1, \ldots, L$. The convex function $h(\boldsymbol{\gamma})$ of L variables can be minimized using, for example, gradient descent methods [Hillier and Lieberman 1986].

Note that the vector that minimizes $h(\boldsymbol{\gamma})$ is not equal to $\mathbf{1} - \mathbf{y}$, the asymptotically optimal importance-sampling parameter for estimating the normalization constant (see Corollary 1). However, numerical testing indicates that this vector, although asymptotically suboptimal, produces a remarkably small variance for estimating utilization.

5. ESTIMATING THE NORMALIZATION CONSTANT: CRITICAL USAGE

In this section we study Monte Carlo integration, in the asymptotic regime of Section 2.2, but no longer suppose that the normal usage conditions are satisfied. In particular, we now consider the critical usage case, $\mathbf{y} = \mathbf{1}$, under an asymptotic regime that is slightly more detailed than that considered by

Ramakrishnan and Mitra [1982]. More specifically, we assume that

$$N_j = \beta_j N + \alpha_j \sqrt{N}, \qquad j = 1, \ldots, J, \tag{18}$$

$$\frac{\rho_{jl}}{\rho_{j0}} = \frac{\Gamma_{jl}}{N}, \qquad j = 1, \ldots, J,\ l = 1, \ldots, L, \tag{19}$$

$$\sum_{j=1}^{J} \beta_j \Gamma_{jl} = 1, \qquad l = 1, \ldots, L, \tag{20}$$

where the β_j's and Γ_{jl}'s are positive constants and the α_j's are arbitrary constants.

Let

$$g(t, x) := \left[e^{-x} \left(1 + \frac{x}{t} \right)^t \right]^t.$$

We shall need the following technical result, which is proved in the Appendix.

LEMMA 3. (i) *For any* $x > 0$,

$$\lim_{t \to \infty} g(t, x) = e^{-x^2/2}.$$

(ii) $g(t, x)$ *is decreasing in* t *for* $t > 0$.

(iii) *For all* $0 < c < t$,

$$e^{-x^2/2} \le g(t, x) \le e^{-cx} \left(1 + \frac{x}{c} \right)^{c^2}.$$

Under our asymptotic assumptions, the normalization constant is given by

$$g(N) = \left[\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!} \right] \int_{\mathbf{Q}^+} e^{-\mathbf{1}'\boldsymbol{u}} \prod_{j=1}^{J} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{N} \right)^{\beta_j N + \alpha_j \sqrt{N}} d\boldsymbol{u}.$$

Now, with the change in variables, $u_l / \sqrt{N} \to u_l$, for $l = 1, \ldots, L$, we have

$$g(N) = \left[\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!} \right] N^{L/2} \int_{\mathbf{Q}^+} e^{-\sqrt{N}\mathbf{1}'\boldsymbol{u}} \prod_{j} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}} \right)^{\beta_j N + \alpha_j \sqrt{N}} d\boldsymbol{u}. \tag{21}$$

Our focus is now on estimating the integral in (21). To this end, let

$$f(\boldsymbol{u}, N) := e^{-\sqrt{N}\mathbf{1}'\boldsymbol{u}} \prod_{j} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}} \right)^{\beta_j N + \alpha_j \sqrt{N}}.$$

(The $f(\boldsymbol{u}, N)$ defined above is analogous to—but not the same as—the function $f(\boldsymbol{u}, N)$ defined in Section 2 for normal usage.) Also let

$$\boldsymbol{\Gamma}_j := (\Gamma_{j1}, \ldots, \Gamma_{jL})'$$
$$\boldsymbol{\Gamma} := \sum_j \beta_j \boldsymbol{\Gamma}_j \boldsymbol{\Gamma}_j'.$$

It is easily verified that $\boldsymbol{\Gamma}$ is positive semidefinite; we hereafter assume that $\boldsymbol{\Gamma}$ is positive definite. Hence it has an inverse $\boldsymbol{\Gamma}^{-1}$ that is also positive definite. Therefore, there exists a lower-triangular invertible matrix $\mathbf{C}$ such that $\mathbf{C}\mathbf{C}' = \boldsymbol{\Gamma}^{-1}$; hence $\mathbf{C}'\boldsymbol{\Gamma}\mathbf{C} = I$. Also let

$$z_l := \sum_j \alpha_j \Gamma_{jl}, \qquad l = 1, \ldots, L$$
$$\boldsymbol{z} = (z_1, \ldots, z_L)'$$
$$\boldsymbol{u}_0 := \boldsymbol{\Gamma}^{-1}\boldsymbol{z}.$$

Throughout this section we suppose that $\{\mathbf{V}^i\}$ is a sequence of i.i.d. random vectors on $\mathbf{Q}^+$ with density

$$p^*(\boldsymbol{u}) = \frac{e^{-(1/2)(\boldsymbol{u}-\boldsymbol{u}_0)'\boldsymbol{\Gamma}(\boldsymbol{u}-\boldsymbol{u}_0)}}{\int_{\mathbf{Q}^+} e^{-(1/2)(\boldsymbol{u}-\boldsymbol{u}_0)'\boldsymbol{\Gamma}(\boldsymbol{u}-\boldsymbol{u}_0)}\, d\boldsymbol{u}}, \qquad \boldsymbol{u} \in \mathbf{Q}^+.$$

Note that $p^*(\boldsymbol{u})$ is the density of a multivariate normal distribution with mean vector $\boldsymbol{u}_0$ and covariance matrix $\boldsymbol{\Gamma}^{-1}$ *truncated* to $\mathbf{Q}^+$. In particular, the components of $\mathbf{V}^i$ are *not* independent. For the Nth network, our estimate for the normalization constant at the ith iteration is now given by

$$Z^i = \frac{f(\mathbf{V}^i, N)}{p^*(\mathbf{V}^i)} N^{L/2} \left[\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!} \right].$$

As before, let $\bar{Z}^I$ be the same mean, and let $\overline{\mathrm{var}(N)}$ be the associated relative variance.

PROPOSITION 4. *$\bar{Z}^I$ is an unbiased estimate for $g(N)$ and*

$$\lim_{N\to\infty} \overline{\mathrm{var}(N)} = 0.$$

PROOF. The unbiasedness is obvious. To show that the relative variance goes to zero, it suffices to show that

$$\lim_{N\to\infty} \frac{E\left[\left\{\dfrac{f(\mathbf{V}^i, N)}{p^*(\mathbf{V}^i)}\right\}^2\right]}{\left\{E\left[\dfrac{f(\mathbf{V}^i, N)}{p^*(\mathbf{V}^i)}\right]\right\}^2} = 1. \tag{22}$$

Since $\Sigma_j \beta_j \Gamma_{jl} = 1$, we have

$$f(\boldsymbol{u}, N) = e^{-\sqrt{N}\Sigma_l u_l} \prod_j \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}}\right)^{\beta_j N + \alpha_j \sqrt{N}}$$

$$= \left\{ e^{-\Sigma_l \Sigma_j \beta_j \Gamma_{jl} u_l} \prod_j \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}}\right)^{\beta_j \sqrt{N}} + \alpha_j \right\}^{\sqrt{N}}$$

$$= \prod_j \left\{ \left[e^{-\Sigma_l \Gamma_{jl} u_l} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}}\right) \right]^{\sqrt{N}} \right\}^{\sqrt{N}\beta_j} \left(1 + \frac{\Sigma_l \Gamma_{jl} u_l}{\sqrt{N}}\right)^{\alpha_j \sqrt{N}}.$$

By Lemma 3 (i), we have

$$\lim_{N \to \infty} f(\boldsymbol{u}, N) = e^{-(1/2)\Sigma_j \beta_j (\Sigma_j \Gamma_{jl} u_l)^2 + \Sigma_j \alpha_j \Sigma_l \Gamma_{jl} u_l}$$

$$= e^{-(1/2)\boldsymbol{u}'\Gamma\boldsymbol{u} + \boldsymbol{z}'\boldsymbol{u}}$$

$$= a e^{-(1/2)(\boldsymbol{u} - \boldsymbol{u}_0)'\Gamma(\boldsymbol{u} - \boldsymbol{u}_0)},$$

where

$$a := e^{(1/2)\boldsymbol{z}'\Gamma^{-1}\boldsymbol{z}}.$$

By Lemma 3 (iii),

$$f(\boldsymbol{u}, N) \le \prod_j e^{-(c\beta_j - \max_k\{\alpha_k, 0\})\Sigma_l \Gamma_{jl} u_l} \left(1 + \frac{\Sigma_j \Gamma_{jl} u_l}{c}\right)^{c^2 \beta_j}$$

when $\sqrt{N} > c$. If we choose $c > \max_k\{\alpha_k, 0\}/\beta_j$, for $j = 1, \ldots, J$, the Dominated Convergence Theorem implies

$$\lim_{N \to \infty} E\left[\frac{f(\mathbf{V}^i, N)}{p^*(\mathbf{V}^i)}\right] = \lim_{N \to \infty} \int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\, d\boldsymbol{u}$$

$$= \int_{\mathbf{Q}^+} a e^{-(1/2)(\boldsymbol{u} - \boldsymbol{u}_0)'\Gamma(\boldsymbol{u} - \boldsymbol{u}_0)}\, d\boldsymbol{u},$$

and

$$\lim_{N \to \infty} E\left[\left\{\frac{f(\mathbf{V}^i, N)}{p^*(\mathbf{V}^i)}\right\}^2\right] = \lim_{N \to \infty} \int_{\mathbf{Q}^+} \frac{f(\boldsymbol{u}, N)^2}{p^*(\boldsymbol{u})}\, d\boldsymbol{u}$$

$$= \lim_{N \to \infty} \left[\int_{\mathbf{Q}^+} e^{-(1/2)(\boldsymbol{u} - \boldsymbol{u}_0)'\Gamma(\boldsymbol{u} - \boldsymbol{u}_0)}\, d\boldsymbol{u}\right]$$

$$\times \int_{\mathbf{Q}^+} \frac{g(N, \boldsymbol{u})^2}{e^{-(1/2)(\boldsymbol{u} - \boldsymbol{u}_0)'\Gamma(\boldsymbol{u} - \boldsymbol{u}_0)}}\, d\boldsymbol{u}$$

$$= \left[\int_{\mathbf{Q}^+} e^{-(1/2)(\boldsymbol{u} - \boldsymbol{u}_0)'\Gamma(\boldsymbol{u} - \boldsymbol{u}_0)}\, d\boldsymbol{u}\right] a^2$$

$$\times \int_{\mathbf{Q}^+} e^{-(1/2)(u-u_0)'\Gamma(u-u_0)}\, du$$

$$= \left[a \int_{\mathbf{Q}^+} e^{-(1/2)(u-u_0)'\Gamma(u-u_0)}\, du\right]^2,$$

from which (22) directly follows. □

From Proposition 4 we know that the importance-sampling density consisting of a specific multivariate normal density truncated to $\mathbf{Q}^+$ is asymptotically optimal in the sense that the relative variance for the normalization constant goes to zero as $N \to \infty$. One can similarly establish the critical-usage analog of Proposition 2: Any density other than $p^*(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbf{Q}^+$, gives an asymptotic relative variance strictly greater than zero.

We conclude this section by discussing another approach to estimating the normalization constant in critical usage. Making the change of variables, $\mathbf{C}^{-1}(\boldsymbol{u} - \boldsymbol{u}_0) \to \boldsymbol{u}$, we have

$$g(N) = \left[\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!}\right] N^{L/2} \int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\, d\boldsymbol{u}$$

$$= \left[\prod_{j=1}^{J} \frac{\rho_{j0}^{N_j}}{N_j!}\right] N^{L/2} |\det \mathbf{C}| \int_{\Theta} f(\mathbf{C}\boldsymbol{u} + \boldsymbol{u}_0, N)\, d\boldsymbol{u}, \tag{23}$$

where

$$\Theta := \{\boldsymbol{u} \in \mathbf{R}^L : \mathbf{C}\boldsymbol{u} \geq -\boldsymbol{u}_0\}.$$

Now consider estimating the integral in (23). It can be shown (see proof of Proposition 4) that

$$\lim_{N \to \infty} f(\mathbf{C}\boldsymbol{u} + \boldsymbol{u}_0, N) = ae^{-(1/2)u'u}, \qquad \boldsymbol{u} \in \Theta.$$

Therefore, in order to estimate the integral by Monte Carlo methods, a natural importance-sampling density is given by

$$\hat{p}(\boldsymbol{u}) = \frac{e^{-(1/2)u'u}}{\int_{\Theta} e^{-(1/2)u'u}\, du}, \qquad \boldsymbol{u} \in \Theta.$$

Note that this sampling density corresponds to *independent standard normal components truncated to* Θ. Mimicking the proof of Proposition 4 it can be shown that the associated relative variance approaches zero as $N \to \infty$.

A theory for estimating utilization under critical usage can also be developed; in analogy with the results in Section 4, truncated multivariate normal densities are not asymptotically optimal.

6. CONCLUSIONS

We have shown that independent exponential sampling with sampling vector $\mathbf{1} - \mathbf{y}$ is asymptotically optimal for estimating the normalization constant under normal usage. As mentioned earlier, numerical testing in Ross et al. [1991] indicates that this density gives remarkably small variances for both

the normalization constant and the utilization when N is finite and when the normal usage condition is satisfied. However, as the "margin" for the normal usage condition becomes smaller, the associated variances increase significantly (but not exorbitantly). Therefore, when the "margin" becomes close or equal to zero, there is motivation to consider densities other than independent exponentials; Proposition 4 strongly suggests the truncated multivariate normal density. But in order to apply this density, two issues must first be resolved.

The first issue concerns the proper choice of the β_j's, α_j's, Γ_{jl}'s, and N for a given network with finite traffic parameters. Clearly, these parameters must be chosen so that the constraints (18)–(19) are satisfied. But the parameters should also be chosen so that N is large, in order to invoke the asymptotic result, and so that (20) is satisfied, in order to claim critical usage. McKenna and Mitra [1982] provided a method for choosing the β_j's, Γ_{jl}'s, and N for networks in normal usage. However, due to the presence of α_j's in the asymptotic regime in (18)–(19), and the restrictions induced by critical usage, a new method is needed to select the β_j's. This issue will be addressed in future research.

The second issue concerns the development of efficient procedures for sampling from the truncated multivariate normal density $p^*(\boldsymbol{u})$ or from its "cousin" $\hat{p}(\boldsymbol{u})$. There are two possible approaches. The first is to generate $\mathbf{V}^i$ from the multivariate normal distribution corresponding to the numerator of $p^*(\boldsymbol{u})$ on $\mathbf{R}^L$ without truncation and then apply the rejection method. This approach will be efficient when a significant fraction of the mass of the untruncated distribution is on $\mathbf{Q}^+$. The second approach is to generate samples over Θ from a distribution that is easy to sample from and is similar to but not the same as $\hat{p}(\boldsymbol{u})$. In this case none of the samples will be wasted. However, since the samples are not from $\hat{p}(\mathbf{u})$, we would expect the variances of the estimates to suffer to some extent. Both of these methods will be addressed in a future paper.

APPENDIX

PROOF OF LEMMA 2. Since $X^i(N) = Y^i(N)(1 + \Gamma_k^t \mathbf{V}^i/N)$, we have

$$y_{km} - \Phi^I(N) = y_{km} \frac{\sum_{i=1}^I [X^i(N) - Y^i(N)]}{\sum_{i=1}^I X^i(N)}$$

$$= \frac{y_{km} \sum_{i=1}^I Y^i(N) \Gamma_k^t \mathbf{V}^i}{\sum_{i=1}^I X^i(N)}. \tag{24}$$

Thus,

$$\lim_{N \to \infty} N\left[y_{km} - \Phi^I(N) \right] = y_{km} W^I. \tag{25}$$

For fixed I, $\max_{1 \le i \le I} \Gamma_k' \mathbf{V}^i$ has a finite expected value (due to our assumption on $p(\boldsymbol{u})$); furthermore, from (20) we have

$$N|y_{km} - \Phi^I(N)| \le y_{km} \max_{1 \le i \le I} \Gamma_k' \mathbf{V}^i.$$

Hence, from the Dominated Convergence Theorem and (25) we have

$$\begin{aligned} \lim_{N \to \infty} N\{y_{km} - E[\Phi^I(N)]\} &= E\Big[\lim_{N \to \infty} N\{y_{km} - \Phi^I(N)\}\Big] \\ &= y_{km} E[W^I], \end{aligned} \tag{26}$$

and the first result is established.

For the second result, note that

$$\begin{aligned} &\Phi^I(N) - E[\Phi^I(N)] \\ &\quad = \frac{\sum_{i=1}^I Y^i(N)\{y_{km} - E[\Phi^I(N)](1 + \Gamma_k' \mathbf{V}^i/N)\}}{\sum_{i=1}^I X^i(N)} \\ &\quad = \frac{1}{N} \frac{\sum_{i=1}^I Y^i(N)\{N(y_{km} - E[\Phi^I(N)]) - E[\Phi^I(N)]\Gamma_k' \mathbf{V}^i\}}{\sum_{i=1}^I X^i(N)}. \end{aligned}$$

Combining this with (14) gives

$$\lim_{N \to \infty} N^2 \{\Phi^I(N) - E[\Phi^I(N)]\}^2 = y_{km}^2 \{W^I - E[W^I]\}^2.$$

Noting $(a + b)^2 \le 2(a^2 + b^2)$, then from (25) and (26) we have, for N sufficiently large,

$$\begin{aligned} N^2 \{\Phi^I(N) - E[\Phi^I(N)]\}^2 &\le 2N^2 (\Phi^I(N) - y_{km})^2 + 2N^2 (y_{km} - E[\Phi^I(N)])^2 \\ &\le 2y_{km}^2 \Big(\max_{1 \le i \le I} \Gamma_k' \mathbf{V}^i\Big)^2 + 2\big(y_{km} E[W^I] + 1\big)^2. \end{aligned}$$

Since $W^I \le \max_{1 \le i \le I} \Gamma_k' \mathbf{V}^i$, and from our assumption, we also have $E[(\max_{1 \le i \le I} \Gamma_k' \mathbf{V}^i)^2] < \infty$, the RHS of the last expression has finite expected value for any fixed I. Hence by the Dominated Convergence Theorem, (25) and (26),

$$\begin{aligned} \lim_{N \to \infty} N^2 \operatorname{var}(\Phi^I(N)) &= E\Big[\lim_{N \to \infty} N^2 \{\Phi^I(N) - E[\Phi^I(N)]\}^2\Big] \\ &= y_{km}^2 \operatorname{var}(W^I). \end{aligned}$$

□

PROOF OF PROPOSITION 3. We will need the following result [Serfling 1980] for ratio estimates: Suppose each of $\{X^i\}$ and $\{Y^i\}$ is an independently and identically distributed sequence such that

$$E\big[(X^i)^2\big] < \infty, \quad E\big[(Y^i)^2\big] < \infty.$$

Let

$$\hat{R}_n = \frac{\sum_{i=1}^{n} X^i}{\sum_{i=1}^{n} Y^i},$$

$$R = \frac{E[X^i]}{E[Y^i]}.$$

Then

$$\lim_{n\to\infty} \sqrt{n}\left[\hat{R}_n - R\right] \overset{dist.}{=} N[0, \sigma^2] \tag{27}$$

where

$$\sigma^2 := R^2\left\{\frac{E\left[(X^i)^2\right]}{E[X^i]^2} + \frac{E\left[(Y^i)^2\right]}{E[Y^i]^2} - 2\frac{E[X^iY^i]}{E[X^i]E[Y^i]}\right\}.$$

It follows from (8) that

$$\begin{aligned}\lim_{N\to\infty} N[y_{km} - \text{util}_{km}(N)] &= \lim_{N\to\infty} y_{km} N\left[1 - \frac{\int_{\mathbf{Q}^+}(f(\boldsymbol{u}, N)/f_k(\boldsymbol{u}, N))\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\,d\boldsymbol{u}}\right] \\ &= y_{km} \lim_{N\to\infty} \frac{\int_{\mathbf{Q}^+}(f(\boldsymbol{u}, N)\boldsymbol{\Gamma}_k'\boldsymbol{u}/f_k(\boldsymbol{u}, N))\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} f(\boldsymbol{u}, N)\,d\boldsymbol{u}} \\ &= y_{km} \frac{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}} \boldsymbol{\Gamma}'_k \boldsymbol{u}\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\,d\boldsymbol{u}}.\end{aligned}$$

Therefore, from (25) we have

$$\begin{aligned}\lim_{N\to\infty} N\left[\Phi^I(N) - \text{util}_{km}(N)\right] &= \lim_{N\to\infty} N\left[\Phi^I(N) - y_{km}\right] \\ &\quad + \lim_{N\to\infty} N[y_{km} - \text{util}_{km}(N)] \\ &= -y_{km}W^I + y_{km}\frac{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\boldsymbol{\Gamma}_k'\boldsymbol{u}\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\,d\boldsymbol{u}} \\ &= -y_{km}\left[W^I - \frac{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\boldsymbol{\Gamma}_k'\boldsymbol{u}\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\,d\boldsymbol{u}}\right].\end{aligned} \tag{28}$$

Recalling the definition of A and B, we have from (27)

$$\begin{aligned}\lim_{I\to\infty} \sqrt{I}\left[W^I - \frac{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\boldsymbol{\Gamma}_k'\boldsymbol{u}\,d\boldsymbol{u}}{\int_{\mathbf{Q}^+} e^{-(1-\mathbf{y})'\boldsymbol{u}}\,d\boldsymbol{u}}\right] &\overset{\text{dist.}}{=} N\left[0, \frac{E\left[\left\{B\frac{E[A]}{E[B]} - A\right\}^2\right]}{E[B]^2}\right] \\ &= N\left[0, h(p)\prod_{l=1}^{L}(1-y_l)^2\right].\end{aligned} \tag{29}$$

Hence, from (28) and (29), we have

$$\lim_{I\to\infty}\lim_{N\to\infty}\sqrt{I}N\left[\Phi^I(N)-\mathrm{util}_{km}(N)\right]\overset{\text{dist.}}{=}N\left[0,h(p)y_{km}^2\prod_{l=1}^{L}(1-y_l)^2\right].$$

It remains to show that the above results holds if we interchange the limits. From (27), for any N,

$$\lim_{I\to\infty}N\sqrt{I}\left[\Phi^I(N)-\mathrm{util}_{km}(N)\right]\overset{\text{dist.}}{=}N\left[0,N^2\sigma(p,N)^2\right],$$

where

$$\sigma(p,N)^2=\mathrm{util}_{km}(N)^2\left\{\frac{E[X^i(N)^2]}{E[X^i(N)]^2}+\frac{E[Y^i(N)^2]}{E[Y^i(N)]^2}-2\frac{E[X^i(N)Y^i(N)]}{E[X^i(N)]E[Y^i(N)]}\right\}.$$

Hence it is sufficient to show that for any x,

$$\lim_{N\to\infty}\int_{-\infty}^{x}e^{-t^2/2N^2\sigma(p,N)^2}\,dt=\int_{-\infty}^{x}e^{-t^2/2h(p)y_{km}^2\Pi_l(1-y_l)^2}\,dt. \tag{30}$$

Note that

$$\begin{aligned}\lim_{N\to\infty}N^2\sigma(p,N)^2&=\lim_{N\to\infty}N^2\mathrm{util}_{km}(N)^2\left\{\frac{E[X^i(N)^2]}{E[X^i(N)]^2}+\frac{E[Y^i(N)^2]}{E[Y^i(N)]^2}-2\frac{E[X^i(N)Y^i(N)]}{E[X^i(N)]E[Y^i(N)]}\right\}\\&=y_{km}^2\frac{E\left[\{B(E[A]/E[B])-A\}^2\right]}{E[B]^2}\\&=h(p)y_{km}^2\prod_l(1-y_l)^2.\end{aligned}$$

Since

$$\begin{aligned}&N^2\mathrm{util}_{km}(N)^2\left\{\frac{E[X^i(N)^2]}{E[X^i(N)]^2}+\frac{E[Y^i(N)^2]}{E[Y^i(N)]^2}-2\frac{E[X^i(N)Y^i(N)]}{E[X^i(N)]E[Y^i(N)]}\right\}\\&\quad=N^2\mathrm{util}_{km}(N)^2\frac{E\{E[Y^i(N)]X^i(N)-Y^i(N)E[X^i(N)]\}^2}{(E[X^i(N)])^4}\\&\quad=\frac{\mathrm{util}_{km}(N)^2}{(E[X^i(N)])^4}E\{N(E[X^i(N)]-E[Y^i(N)])Y^i(N)\end{aligned}$$

$$-E[Y^i(N)]N[X^i(N) - Y^i(N)]\}^2$$
$$\leq \frac{y_{km}^2}{(E[X^i(1)])^4}(2E[A^2]E[B]^2 + 2E[A]^2E[B^2]),$$

by Dominated Convergence Theorem, we have proved (30). □

PROOF OF LEMMA 3. For (i) it is sufficient to show that

$$\lim_{t\to\infty} \ln g(t, x) = -\frac{x^2}{2}.$$

Note that for any $0 < a < 1$,

$$a - \frac{a^2}{2} < \ln(1 + a) < a - \frac{a^2}{2} + \frac{a^3}{3};$$

hence, when $t > x$,

$$\ln g(t, x) = -tx + t^2\ln\left(1 + \frac{x}{t}\right)$$
$$\geq -tx + t^2\left(\frac{x}{t} - \frac{x^2}{2t^2}\right)$$
$$= -\frac{x^2}{2}.$$

On the other hand,

$$-tx + t^2\ln\left(1 + \frac{x}{t}\right) \leq -tx + t^2\left(\frac{x}{t} - \frac{x^2}{2t^2} + \frac{x^3}{3t^3}\right)$$
$$= -\frac{x^2}{2} + \frac{x^3}{3t}.$$

For (ii) notice that $g(t, x) = e^{-tx+t^2\ln(1+(x/t))}$, and for any $x > 0$,

$$\frac{\partial g(t, x)}{\partial t} = xg(t, x)\left[-1 - \frac{1}{1 + x/t} + 2\frac{t}{x}\ln\left(1 + \frac{x}{t}\right)\right].$$

Let

$$h(t, x) := -1 - \frac{1}{1 + x/t} + 2\frac{t}{x}\ln\left(1 + \frac{x}{t}\right).$$

We show that $h(t, x) < 0$ for any $t > 0$. First, note that $\lim_{t\to\infty} h(t, x) = 0$ and $\lim_{t\to 0} h(t, x) = -1$. Thus, it is sufficient to show that $(\partial h(t, x))/(\partial t) >$

0. Now,

$$\frac{\partial h(t,x)}{\partial t} = -\frac{x/t^2}{(1+x/t)^2} + \frac{2}{x}\frac{\ln(1+x/t)}{x/t} - \frac{2}{t}\frac{1}{1+x/t}$$
$$= \frac{1}{x(1+x/t)^2}\phi\left(\frac{x}{t}\right),$$

where $\phi(y) := -y^2 + 2(1+y)^2\ln(1+y) - 2y(1+y)$. We only need to show $\phi(y) > 0$ for $y > 0$. Note

$$\phi(y) = -3y^2 - 2y + 2(1+y)^2\ln(1+y),$$
$$\phi(0) = 0, \tag{31}$$
$$\phi'(y) = -4y + 4(1+y)\ln(1+y),$$
$$\phi'(0) = 0, \tag{32}$$
$$\phi''(y) = 4\ln(1+y) > 0. \tag{33}$$

For $y > 0$, we have, from (32) and (33), $\phi'(y) > 0$, which when combined with (31) gives $\phi(y) > 0$.

Finally (iii) immediately follows (ii). □

REFERENCES

BASKETT, F., CHANDY, M., MUNTZ, R., AND PALACIOS, J. 1975. Open, closed and mixed networks of queues with different classes of customers. *J. ACM 22*, 2, 248–260.

CONWAY, A. E., AND GEORGANAS, N. D. 1989. *Queueing Networks—Exact Computational Algorithms: A Unified Theory Based on Decomposition and Aggregation.* MIT Press, Cambridge, Mass.

HILLIER, F. S., AND LIEBERMAN, G. J. 1986. *Introduction to Operations Research.* Holden-Day, Oakland, Calif.

KALOS, M., AND WHITLOCK, P. A. 1986. *Monte Carlo Methods*, Vol. 1. *Basics.* John Wiley, New York.

KNESSL, C., AND TIER, C. 1992. Asymptotic expansions for large closed queueing networks with multiple job classes. *IEEE Trans. Comput. 41*, 4, 480–488.

KNESSL, C., AND TIER, C. 1990. Asymptotic expansions for large closed queueing networks. *J. ACM 37*, 4, 144–174.

LAM, S. S., AND LIEN, Y. L. 1983. A tree convoluted algorithm for the solution of the queueing networks. *Commun. ACM 26*, 3, 203–215.

MCKENNA, J., AND MITRA, D. 1984. Asymptotic expansions and integral representations of moments of queue lengths in closed Markovian networks. *J. ACM 31*, 2, 346–360.

MCKENNA, J., AND MITRA, D. 1982. Integral representations and asymptotic expansions for closed Markovian queueing networks: Normal usage. *Bell Syst. Tech. J. 61*, 5, 661–683.

MITRA, D. 1992. Asymptotic optimal design of congestion control for high speed data networks. *IEEE Trans. Commun. 40*, 2, 301–311.

MORRISON, J. A., AND MITRA, D. 1985. Heavy-usage asymptotic expansions for the waiting time in closed processor-sharing systems with multiple classes. *Adv. Appl. Prob. 17*, 1, 163–185.

RAMAKRISHNAN, K. G., AND MITRA, D. 1982. An overview of PANACEA, a software package for analyzing Markovian queueing networks. *Bell Syst. Tech. J. 61*, 10, 2849–2872.

REISER, M., AND KOBAYASHI, H. 1975. Queueing networks with multiple closed chains: Theory and computational algorithms. *IBM J. Res. Dev. 19*, 3, 283–294.

REISER, M., AND LAVENBERG, S. S. 1980. Mean value analysis of closed multichain queueing networks. *J. ACM 27*, 2, 313–322.

ROSS, K. W., AND WANG, J. 1992. Monte Carlo summation applied to product-form loss networks. *Prob. Eng. Inf. Sci. 6*, 3, 323–348.

ROSS, K. W., AND WANG, J. 1990. Solving product form stochastic networks with Monte Carlo summation. In *Proceedings of the 1990 Winter Simulation Conference*. ACM/IEEE, New York, 270–275.

ROSS, K. W., TSANG, D. H. K., AND WANG, J. 1991. Monte Carlo summation and integration applied to multichain queueing networks. Submitted, *J. ACM*.

ROYDEN, H. L. 1968. *Real Analysis*. Macmillan, New York.

SAUER, C. H., AND CHANDI, K. M. 1981. *Computer Systems Performance Modeling*. Prentice-Hall, Englewood Cliffs, N.J.

SERFLING, R. 1980. *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

Received September 1992; revised March 1993; accepted June 1993