Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

AN EFFICIENT PERFORMANCE EXTRAPOLATION FOR QUEUING MODELS IN TRANSIENT ANALYSIS

Mohamed A. Ahmed
Talal M. Alkhamis

Department of Statistics and Operations Research
Kuwait University
P. O. Box 5969, Safat, KUWAIT

ABSTRACT

In designing, analyzing and operating real-life complex systems, we are interested, however, not only in performance evaluation but in sensitivity analysis and optimization as well. Since most systems of practical interest are too complex to allow the analytical solution of totally realistic models, these systems must be studied by means of Monte-Carlo simulation. One problem with Monte Carlo analysis is its expensive use of computer time. To address this problem, we propose an efficient technique for estimating the expected performance of a stochastic system for various values of the parameters from a single simulation of the nominal system. This technique is based on the likelihood ratio performance extrapolation (LRPE). We provide numerical experiments that demonstrate how the proposed technique significantly outperform the likelihood ratio performance extrapolation technique in the context of the Markovian queueing models in transient analysis.

1 INTRODUCTION

Most systems of practical interest are too complex to allow the analytical solution of totally realistic models. Consequently, these systems must be studied by means of Monte-Carlo simulation. Planners typically want to know how the system will perform under various parameter settings. To determine this, a computer simulation model may be developed and then run for these parameter settings. As an example, consider a queueing network consisting of stations with buffers and multiple servers at each station. Suppose that all parameters are known and fixed except for the service rate parameters of the exponential servers. We wish to estimate the expected number of customers served (system throughput) by time T, for different service rates. A standard approach to this problem would be to use the crude Monte-Carlo approach by simulating the system at each value of the service rate. Since large-scale simulations may require great amounts of computer time and storage, appropriate statistical analysis can become quite costly.

To address this problem, we propose an efficient technique for estimating the expected performance of a stochastic system for various values of the parameters from a single simulation of the nominal system. Given the performance measure at two values of the input distribution parameters, the proposed technique provides the ability to interpolate the simulation results at different values of these parameters. This technique is based on the likelihood ratio performance extrapolation (LRPE). Arsham et al. [1989] showed that using the likelihood ratio (Radon-Nikodym derivative) approach, one can estimate simultaneously the performance measure at various parameter values from a single simulation run.

Implementation of the LRPE approach requires computation of the likelihood ratio (Radon-Nikodym derivative) of the underlying stochastic system. In this paper we use the continuous-time Markov chain frame-work to find the likelihood ratio for large classes of queuing models (see Nakayama et al. [1994]). Continuous-time Markov chains are good models for many stochastic systems, including certain queuing systems, inventory systems, and reliability and maintenance systems. While the basis of the LRPE technique (Rubinstein 1986, 1989, Glynn 1987, Reiman and Weiss 1989, L'Ecuyer 1995, 1990) has been known for some time, the technique works only for perturbations of limited size due to its high variance. In this paper, we develop an interpolating technique as an effective tool for estimating system response to parametric perturbations in simulation. We show, through extensive experimentations, that the proposed technique is an effective tool for measuring parameter sensitivity in the context of the Markovian queueing models in transient analysis. There are many instances in which the transient behavior of stochastic systems is important. Since the characteristics of most real systems change over time, the stochastic processes for those systems do not have steady-state distribution. For example, in a manufacturing system the production scheduling rules and the facility layout (e.g., number and location of machines) may change from time to time.

The rest of the paper is organized as follows. Section 2 reviews the basic idea of the Radon-Nikodym derivative approach to the so-called "what if" problem (performance extrapolation) on which the proposed interpolation technique is based. Section 3 develops the proposed technique as an efficient method for estimating the transient performance measures of stochastic systems. Section 4 gives the Radon-Nikodym derivative for some classes of Markovian queueing systems and provides our computational experiments that demonstrate the efficiency of the proposed technique. Finally, section 5 contains some concluding remarks.

2 RADON-NIKODYM DERIVATIVE APPROACH

Before proceeding further, we briefly discuss the likelihood ratio method (Radon-Nikodym derivative) approach as it is related to the proposed interpolation method to be discussed subsequently. Consider a stochastic simulation system parameterized by a real vector $\boldsymbol{\theta} \in \Theta$ of continuous parameters, where Θ is some open subset of R^n. We are interested in performance measures that are based on the behavior of the stochastic system in some time interval T, where T is a stopping time. Suppose we have independent simulation results of the system at parameter $\boldsymbol{\theta}_1 \in \Theta$ and want to estimate the transient performance measure of that system at parameter $\boldsymbol{\theta}_0 \in \Theta$, $\ell(\boldsymbol{\theta}_0 \mid \mathrm{I})$, where I represents the initial conditions used to start the simulation at time 0.

The basic idea of LRPE is that $\ell(\boldsymbol{\theta}_0 \mid \mathrm{I})$ can usually be viewed as the expectation of some function of $\boldsymbol{\theta}_0$ and the sample path ω, say $h(\boldsymbol{\theta}_0, \omega)$, with respect to a probability measure $P_{\boldsymbol{\theta}_0}$. Suppose that $P_{\boldsymbol{\theta}_0}$ is absolutely continuous with respect to $P_{\boldsymbol{\theta}_1}$, i.e., for every measurable set B, if $\mathrm{P}_{\boldsymbol{\theta}_1}(B) = 0$ then $\mathrm{P}_{\theta_0}(B) = 0$. In this case, one can write

$$
\begin{aligned}
\ell(\boldsymbol{\theta}_0 \mid \mathrm{I}) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left[h(\boldsymbol{\theta}_0, \omega)\right] \\
&= \int h(\boldsymbol{\theta}_0, \omega) dP_{\boldsymbol{\theta}_0}(\omega) \\
&= \int \left[h(\boldsymbol{\theta}_0, \omega) \frac{dP_{\boldsymbol{\theta}_0}}{dP_{\boldsymbol{\theta}_1}}(\omega)\right] dP_{\boldsymbol{\theta}_1}(\omega) \\
&= \mathrm{E}_{\boldsymbol{\theta}_1}\left[h(\boldsymbol{\theta}_0, \omega) L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\right],
\end{aligned}
$$

where $L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = (dP_{\boldsymbol{\theta}_0} / dP_{\boldsymbol{\theta}_1})(\omega)$ is the Radon-Nikodym derivative of $P_{\boldsymbol{\theta}_0}$ with respect to $P_{\boldsymbol{\theta}_1}$ or the likelihood ratio of the process up to stopping time T. The subscript $\boldsymbol{\theta}_0$ in $\mathrm{E}_{\boldsymbol{\theta}_0}\left[h(\boldsymbol{\theta}_0, \omega)\right]$ means that the expectation operator is induced by $P_{\boldsymbol{\theta}_0}$.

Typically ω could be the set of values taken by a finite sequence of independent (possibly multivariate) random variables $\mathbf{Y}$ with probability density function $f(\mathbf{y}, \boldsymbol{\theta})$. For example, consider an M/M/1 queue and let $\ell(\boldsymbol{\theta}_0 \mid \mathrm{I})$ be the expected mean waiting time in the system for the first T customers in the system, provided that the initial conditions used to start the simulation at time 0 is I. In this case, ω could be the set of actual interarrival and service times and $d\,\mathrm{P}_{\boldsymbol{\theta}_0}(\omega)$ is the product of their densities. We have

$$
\begin{aligned}
\ell(\boldsymbol{\theta}_0 \mid \mathrm{I}) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left[h(\mathbf{Y}, \theta_0)\right] \qquad (1)\\
&= \int h(\mathbf{y}, \boldsymbol{\theta}_0) f(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} \\
&= \int h(\mathbf{y}, \boldsymbol{\theta}_0) \frac{f(\mathbf{y}, \boldsymbol{\theta}_0)}{f(\mathbf{y}, \boldsymbol{\theta}_1)} f(\mathbf{y}, \boldsymbol{\theta}_1) d\mathbf{y} \\
&= \mathrm{E}_{\boldsymbol{\theta}_1}\left[h(\mathbf{Y}, \boldsymbol{\theta}_0) L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\right], \qquad (2)
\end{aligned}
$$

where $L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \dfrac{f(\mathbf{y}, \boldsymbol{\theta}_0)}{f(\mathbf{y}, \boldsymbol{\theta}_1)}$ and

$$\mathrm{f}(\mathbf{y}, \boldsymbol{\theta}_0) = \prod_{j=1}^{T} f_j(\mathrm{y}_j, \boldsymbol{\theta}_0).$$

It is important to note that the original expectation of $h(\mathbf{Y})$ in (1) is taken with respect to the underlying pdf $f(\mathbf{y}, \boldsymbol{\theta}_0)$, whereas that given in (2) is taken with respect to the pdf $f(\mathbf{y}, \boldsymbol{\theta}_1)$. It follows that changing the probability density from $f(\mathbf{y}, \boldsymbol{\theta}_0)$ to $f(\mathbf{y}, \boldsymbol{\theta}_1)$, we can express the performance measure $\ell(\boldsymbol{\theta} \mid I)$ for all $\boldsymbol{\theta} \in \Theta$ as an expectation with respect to $f(\mathbf{y}, \boldsymbol{\theta}_1)$ and then estimate it accordingly. In terms of simulation, this means that in principle, one simulation at $\boldsymbol{\theta}_1$ can produce estimates of the performance measure at all "valid" values of $\boldsymbol{\theta}$.

Estimating $\ell(\boldsymbol{\theta}_0 | \mathrm{I})$ using the Radon-Nikodym approach yields computational savings, but reduces precision. By generating a sample $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ from $f(\mathbf{y}, \boldsymbol{\theta}_1)$, we can estimate $\ell(\boldsymbol{\theta}_0 \mid \mathrm{I})$ by the corresponding sample mean $\tilde{\ell}(\boldsymbol{\theta}_0 \mid \mathrm{I}) = \frac{1}{n}\sum_{i=1}^{n}\left[h(\mathbf{y}_i) L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\right]$.

The accuracy of the estimator $\tilde{\ell}(\boldsymbol{\theta}_0 \mid \mathrm{I})$ is determined by its variance $\mathrm{Var}\,\tilde{\ell}(\boldsymbol{\theta}_0 \mid \mathrm{I}) = \frac{1}{n}\mathrm{Var}_{\boldsymbol{\theta}_1}\left[h(\mathbf{Y}) L(\mathrm{T}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\right]$.

Note that the farther $\boldsymbol{\theta}_1$ is from $\boldsymbol{\theta}_0$, the higher variance of the estimator $\tilde{\ell}(\boldsymbol{\theta}_0 \mid \mathrm{I})$, i. e., the variance of the LRPE estimators grows quite fast as the length of the perturbed parameter increases.

3 THE INTERPOLATION APPROACH

In this section, we discuss the interpolation technique as an efficient method for estimating the transient performance measures. Suppose we have independent simulation results of a stochastic system at two values of the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and consider the expected performance of the system at θ_0. Then, we have

$$\begin{aligned}
\ell(\boldsymbol{\theta}_0) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left[h(\mathbf{Y},\boldsymbol{\theta}_0)\right] \\
&= \int h(\mathbf{y},\boldsymbol{\theta}_0)\, f(\mathbf{y},\boldsymbol{\theta}_0)\, d\mathbf{y} \\
&= \alpha \int h(\mathbf{y},\boldsymbol{\theta}_0)\, f(\mathbf{y},\boldsymbol{\theta}_0)\, d\mathbf{y} + (1-\alpha)\int h(\mathbf{y},\boldsymbol{\theta}_0)\, f(\mathbf{y},\boldsymbol{\theta}_0)\, d\mathbf{y} \\
&= \alpha \int h(\mathbf{y},\boldsymbol{\theta}_0) \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\boldsymbol{\theta}_1)} f(\mathbf{y},\boldsymbol{\theta}_1) d\mathbf{y} + (1-\alpha)\int h(\mathbf{y},\boldsymbol{\theta}_0) \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\boldsymbol{\theta}_2)} f(\mathbf{y},\boldsymbol{\theta}_2) d\mathbf{y} \\
&= \alpha\, \mathrm{E}_{\boldsymbol{\theta}_1}\left[h(\mathbf{Y},\boldsymbol{\theta}_0) L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)\right] + (1-\alpha)\mathrm{E}_{\boldsymbol{\theta}_2}\left[h(\mathbf{Y},\boldsymbol{\theta}_0) L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)\right]
\end{aligned}$$

An estimator of $\ell(\boldsymbol{\theta}_0)$ is

$$\tilde{\ell}(\boldsymbol{\theta}_0) = \frac{\alpha}{n}\sum_{i=1}^{n}\left[h(\mathbf{y}_i)\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)\right] + \frac{1-\alpha}{n}\sum_{i=1}^{n}\left[h(\mathbf{y}_i)\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)\right]$$

and its variance is given by

$$Var\, \tilde{\ell}(\boldsymbol{\theta}_0) = \frac{\alpha^2}{n}\mathrm{Var}_{\boldsymbol{\theta}_1}\left[h(\mathbf{Y})\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)\right] + \frac{(1-\alpha)^2}{n}\mathrm{Var}_{\boldsymbol{\theta}_2}\left[h(\mathbf{Y})\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)\right]$$

It can be shown that the value of α that minimizes $\mathrm{Var}\, \tilde{\ell}(\theta_0)$ is given by

$$\alpha^* = \frac{\mathrm{Var}_{\boldsymbol{\theta}_2}\left[h(\mathbf{Y})\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)\right]}{\mathrm{Var}_{\boldsymbol{\theta}_1}\left[h(\mathbf{Y})\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)\right] + \mathrm{Var}_{\boldsymbol{\theta}_2}\left[h(\mathbf{Y})\, L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)\right]}$$

3.1 An Exact Confidence Interval for the Interpolation Response

Let $\ell(\boldsymbol{\theta}_0)$ be the unknown quantity to be estimated for a given scalar input parameter $\boldsymbol{\theta}_0$. Let $X^{(1)} = h(\mathbf{Y},\boldsymbol{\theta}_0) L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)$ and

$\mathrm{X}^{(2)} = h(\mathbf{Y},\boldsymbol{\theta}_0) L(\mathrm{T},\boldsymbol{\theta}_0,\boldsymbol{\theta}_2)$

be unbiased estimators of $\ell(\boldsymbol{\theta}_0)$ (i.e., $E[X^{(1)}] = E[X^{(2)}] = \ell(\boldsymbol{\theta}_0)$). Define the interpolated estimator as:

$$X^{(0)}(\alpha) = \alpha X^{(2)} + (1-\alpha) X^{(1)} \;,$$

for any fixed α, $\mathrm{X}^{(0)}(\alpha)$ is an unbiased estimator of $\ell(\boldsymbol{\theta}_0)$. The value of α which minimizes Var[$X^{(0)}(\alpha)$] can easily shown to be

$$\alpha^* = \frac{\sigma^2_{X^{(1)}} - \sigma_{X^{(2)}X^{(1)}}}{\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}} - 2\sigma_{X^{(2)}X^{(1)}}}$$

where $\boldsymbol{\sigma}^2_{X^{(2)}} = Var[X^{(2)}]$, $\boldsymbol{\sigma}^2_{X^{(1)}} = Var[X^{(1)}]$, and $\boldsymbol{\sigma}_{X^{(2)}X^{(1)}} = Cov(X^{(1)}, X^{(2)})$.

A major difficulty with the interpolated estimator is that α^* is typically unknown, since $Var[X^{(1)}]$, $Var[X^{(2)}]$, and $Cov(X^{(1)}, X^{(2)})$ are in general unknown. Realistically, α needs to be estimated. Suppose that n independent replications of the simulation are performed, one can estimate $\ell(\boldsymbol{\theta}_0)$ by the sample average

$\tilde{\ell}(\boldsymbol{\theta}_0) = \overline{X}^{(0)}(\hat{\alpha}^*) = \hat{\alpha}^* \overline{X}^{(2)} + (1-\hat{\alpha}^*)\overline{X}^{(1)}$.

which is generally biased, since $\hat{\alpha}^*$, $\overline{X}^{(2)}$ and $\overline{X}^{(1)}$ are dependent .

Assuming that $X^{(1)}$ and $X^{(2)}$ are jointly normally distributed, we are able to construct unbiased estimators for the unknown parameters and obtain exact confidence interval estimator for these parameters. The normality assumption is reasonable, because we are dealing with cumulative statistics, and central-limit effects ensure that joint normality is, at least, asymptotically obtained. In order to construct unbiased estimator for the unknown parameters and to provide confidence interval for the interpolated estimator, we consider two cases: (i) $X^{(1)}$ and $X^{(2)}$ are independent; (ii) $X^{(1)}$ and $X^{(2)}$ are dependent.

Case 1 ($X^{(1)}$ and $X^{(2)}$ are independent). We assume $(X_1^{(1)}, X_1^{(2)}), \ldots, (\mathrm{X}_n^{(1)}, X_n^{(2)})$ denote a sequence of i.i.d bivariate normally distributed random 2 tuples. The sequence $(X_1^{(2)} - \mathrm{X}_1^{(1)},\ X_1^{(2)}) \ldots, (X_n^{(2)} - \mathrm{X}_n^{(1)},\ X_n^{(2)})$ is also i.i.d bivariate normally distributed with means (0, $\mu_{X^{(2)}}$), variances $\left(\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}},\ \sigma^2_{X^{(2)}}\right)$ and covariance $\sigma^2_{X^{(2)}}$. Under the assumption of bivariate normal distribution, one can write

$$\varepsilon_i = X_i^{(2)} - \ell(\theta_0) - \alpha^*(X_i^{(1)} - X_i^{(2)}) \quad 1 \le i \le n$$

where $\alpha^* = \dfrac{\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}}}$.

Lemma 3.1 $\{\varepsilon_i\}$ are i.i.d normal random variables independent of $\{X_i^{(2)} - X_i^{(1)}\}$ with variance

$$\sigma^2 = \sigma^2_{X^{(2)}} - 2\alpha^*\sigma^2_{X^{(2)}} + \alpha^{*2}(\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}})$$

$$= \sigma^2_{X^{(2)}}\left(1 - \sigma^2_{X^{(2)}} \Big/ (\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}})\right)$$

To estimate the unknown parameters $\ell(\theta_0)$ and α* we can use either the maximum likelihood or the least squares methods (they yield the same solution under the normality assumption).

Lemma 3.2. The estimator of $\ell(\theta_0)$, α* and σ^2 are respectively

$$\hat{\ell}(\theta_0) = \bar{X}_n^{(1)} - \hat{\alpha}^*(\bar{X}_n^{(1)},\ \bar{X}_n^{(2)}),$$

$$\hat{\alpha}^* = \frac{\hat{\sigma}^2_{X^{(2)}}}{\hat{\sigma}^2_{X^{(1)}} + \hat{\sigma}^2_{X^{(2)}}}, \text{ and}$$

$$\hat{\sigma}^2 = \hat{\sigma}^2_{X^{(1)}} - 2\hat{\alpha}^*\hat{\sigma}^2_{X^{(1)}} + \hat{\alpha}^{*2}(\hat{\sigma}^2_{X^{(1)}} + \hat{\sigma}^2_{X^{(2)}}).$$

Proof. See Fishman [1996].

Case 2 ($X^{(1)}$ and $X^{(2)}$ are dependent).

Let $(X_1^{(2)} - X_1^{(1)},\ X_1^{(1)}) \ldots, (X_n^{(2)} - X_n^{(1)},\ X_n^{(1)})$ denote a sequence of i.i.d bivariate normally distributed 2 tuple with means (0, $\mu_{X^{(2)}}$), variances $\left(\sigma^2_{X^{(2)}} + \sigma^2_{X^{(1)}} - 2\sigma_{X^{(1)}X^{(2)}},\ \sigma^2_{X^{(2)}}\right)$ and covariance $(\sigma^2_{X^{(2)}} - \sigma_{X^{(1)}X^{(2)}})$. With this assumption, one can write

$$\varepsilon_i = X_i^{(1)} - \ell(\theta_0) - \alpha^*(X_i^{(1)} - X_i^{(2)})$$

$$1 \le i \le n$$

where $\alpha^* = \dfrac{\sigma^2_{X^{(2)}} - \sigma_{X^{(1)}X^{(2)}}}{\sigma^2_{X^{(2)}} + \sigma^2_{X^{(1)}} - 2\sigma_{X^{(1)}X^{(2)}}}$ and $\{\varepsilon_i\}$ are i.i.d normal random variables independent of $\{X_i^{(2)} - X_i^{(1)}\}$ with variance

$$\sigma^2 = \sigma^2_{X^{(2)}} - \frac{(\sigma^2_{X^{(2)}} - \sigma_{X^{(1)}X^{(2)}})^3}{\sigma^2_{X^{(2)}} + \sigma^2_{X^{(1)}} - 2\sigma_{X^{(1)}X^{(2)}}}$$

$$= \sigma^2_{X^{(2)}}\left(1 - \frac{\sigma^2_{X^{(2)}} - \rho^2_{X^{(1)}X^{(2)}}\sigma^2_{X^{(1)}} - 2\sigma_{X^{(1)}X^{(2)}}}{\sigma^2_{X^{(1)}} + \sigma^2_{X^{(2)}} - 2\sigma_{X^{(1)}X^{(2)}}}\right)$$

where $\rho_{X^{(1)}X^{(2)}} = \sigma_{X^{(1)}X^{(2)}} / \sigma_{X^{(1)}}\sigma_{X^{(2)}}$.

Finally, conditional on $\{X_i^{(2)} - X_i^{(1)}\}$ i = 1, …, n, we can generate a confidence interval as follow

$$\Pr\left[\frac{\left|\breve{\ell}(\theta_0) - \ell(\theta_0)\right|}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^{n}(Z_i - \bar{Z})^2}} \le t_{n-2,1-\delta/2} \Big| \{Z_i\}\right] = 1 - \delta$$

where $Z_i = X_i^{(2)} - X_i^{(1)}$ and $t_{n-2,1-\delta/2}$ denotes the $1-\delta/2$ quantile of the student's t distribution with n-2 degree of freedom. Since the right hand side of this probability statement is constant, this confidence interval holds unconditionally, i.e., $\breve{\ell}(\theta_0) \pm t_{n-2,1-\delta/2}\sqrt{\hat{\sigma}^2 / \sum_{i=1}^{n}(Z_i - \bar{Z})^2}$ provides a 100(1-δ)% confidence interval for $\ell(\theta)$.

4 RADON-NIKODYM DERIVATIVE FOR MARKOVIAN SYSTEMS

In this section, we present the Radon Nikodym derivative for some classes of Markovian queueing systems in transient analysis. Also, we present our computational experiments that show that the proposed interpolation approach is an efficient way to estimate transient measures of performance.

4.1 M/M/1: The Classical Queueing System

The M/M/1 queue can be analyzed as a birth–death process (see, e.g. Cooper [1981], Gross and Harris [1998] and Kleinrock [1975]) by selecting the birth–death coefficients as follows:

$$\lambda_n = \lambda \qquad n = 0,\ 1,\ 2, \ldots$$

$$\mu_n = \mu \qquad n = 1,\ 2,\ 3, \ldots$$

For this queuing system, let X_k represent the state of the system at the k^{th} transition, and define the following

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \ldots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \ldots \\ . & . & . & . & . \end{bmatrix}$$

$$q(X_k) = \begin{cases} \lambda & X_k = 0 \\ (\lambda+\mu) & \text{otherwise} \end{cases}$$

The matrix Q is called the infinitesimal generator (or rate matrix or intensity matrix) of the process; and its elements, q_{ij} , give the "rates" of going from state i to state j. The ith diagonal element is usually denoted by -q_i ; q_i, gives the rate of leaving state i to any other states. The elements in each row of Q thus sum to zero. Given that the system has entered state i, the holding time in state i, is an exponential random variable with parameter $\lambda + \mu$, since it is the minimum of two exponential random variables, namely, arrival time with parameter λ and service time with parameter μ. Given that a transition occurs from state i, the probability that the transition is due to an arrival (state increases to i +1) is $\frac{\lambda}{\lambda+\mu}$ and the probability that is due to a service completion (state decreases to i-1) is $\frac{\mu}{\lambda+\mu}$. Thus we have a process that stays in state i for a time that is exponential random variable and jumps to either state i +1 or state i -1 with transition probabilities of

$$P(X_k , X_{k+1}) = \begin{cases} 1 & X_k = 0, X_{k+1} = 1 \\ \dfrac{\lambda}{\lambda+\mu} & X_{k+1} = X_k + 1 \\ \dfrac{\mu}{\lambda+\mu} & X_{k+1} = X_k - 1 \end{cases}$$

The following lemma gives the Radon-Nikodym derivative for the M/M/1 system. First, let

N_o = number of state transitions from state 0.

N_u = number of state transitions from state X_k to state X_k+1.

N_d = number of state transitions from state X_k to state X_k-1.

Note that N_o+N_u represent the number of arrivals by time T and N_d represents the number of departures by time T.

Lemma 4.1.1 For the M/M/1 system, let T be the stopping time. The likelihood ratio with respect to parameter λ_0 is given by

$$L(T,\lambda,\lambda_0) = \left(\frac{\lambda}{\lambda_0}\right)^{N_o+N_u} \exp\{-T(\lambda-\lambda_0)\}$$

and with respect to μ_0 is given by

$$L(T,\mu,\mu_0) = \left(\frac{\mu}{\mu_0}\right)^{N_d} * \exp\left\{-(\mu-\mu_0)(\sum_{k\in N-N_0} t_k + 1_{\{X_{N(T)}\neq 0\}}[T-T_{N(T)}])\right\}$$

Proof Let N(T) denote the number of transitions up to time T. The likelihood of the sample path up to time T under parameter λ and μ is

$$\Gamma(T,\mu,\lambda) = \prod_{k=0}^{N(T)-1} q(X_k,\lambda,\mu)\exp\{-q(X_k,\lambda,\mu)t_k\} * P(X_k,X_{k+1})\exp\{-q(X_{N(T)},\lambda,\mu)(T-T_{N(T)})\}$$

$$= \lambda^{N_0}\exp\left\{-\lambda\sum_{k\in N_0} t_k\right\}\cdot(\lambda+\mu)^{N_u}\exp\left\{-(\lambda+\mu)\sum_{k\in N_u} t_k\right\}\left(\frac{\lambda}{\lambda+\mu}\right)^{N_u}$$
$$\cdot(\lambda+\mu)^{N_d}\exp\left\{-(\lambda+\mu)\sum_{k\in N_d} t_k\right\}\left(\frac{\mu}{\lambda+\mu}\right)^{N_d}\left[\exp\{-(\lambda+\mu-1_{\{X_{N(T)}=0\}}\cdot\mu)\}(T-T_{N(T)})\right]$$
$$= \lambda^{N_0}\exp\{-\lambda T\}\cdot\exp\left\{-\mu\sum_{k\in N(T)-N_0} t_k\right\}\cdot\lambda^{N_u}\cdot\mu^{N_d}\cdot\exp\{-\mu\cdot 1_{\{X_{N(T)}\neq 0\}}(T-T_{N(T)})\}$$

For a given parameter value λ_0, the Likelihood ratio is given by

$$L(T,\lambda_0,\lambda) = \frac{\Gamma(T,\lambda)}{\Gamma(T,\lambda_0)} = \left(\frac{\lambda}{\lambda_0}\right)^{N_o+N_u}\exp\{-T(\lambda-\lambda_0)\}$$

For a given parameter value μ_0, the Likelihood ratio is given by

$$L(T,\mu_0,\mu) = \frac{\Gamma(T,\mu)}{\Gamma(T,\mu_0)} = \left(\frac{\mu}{\mu_0}\right)^{N_d}\exp\left\{-(\mu-\mu_0)(\sum_{k\in N(T)-N_0} t_k + 1_{\{X_{N(T)}\neq 0\}}[T-T_{N(T)}]\right\}$$

Example 4.1.2 Consider an M/M/1 system which starts empty and runs for time T. We are interested in knowing the average number in the system at time T, for various arrival rates. Assume we are not aware that this can be solved analytically. We simulate at arrival rate λ and use LRPE to estimate the average number in the system at time T for several other arrival rates $\lambda_0 = \lambda + \Delta$. It has been shown that as Δ increases, the variance of the LRPE estimate increases very rapidly (Rubinstein [1989]; Arsham et al. [1989]). For the interpolation approach we simulate at arrival rates λ_1 and λ_2 to estimate the average number in the system at time T for several other arrival rates λ_0.

The M/M/1 model was simulated using the following parameters. Arrival rates $\lambda_1 = 1$ and $\lambda_2 = 2$, service rate $\mu = 2$, and T = 3. The LRPE was applied to estimate the average number in the system at time T for 9 perturbed arrival rates between 1 and 2. Table 1 presents the crude Monte Carlo simulation for the average number in the system at time T, LRPE estimates for the nine rates from λ_1 and λ_2, and the interpolation estimates with their corresponding variances.

4.2 M/M/1/s: Queuing System with Truncation

For this queuing system define the following

$$\lambda_n = \begin{cases} \lambda & n < s \\ 0 & n \geq s \end{cases}$$

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \ldots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \ldots \\ . & . & . & . & . \end{bmatrix}$$

$$q(X_k) = \begin{cases} \lambda & X_k = 0 \\ (\lambda+\mu) & \text{otherwise} \end{cases}$$

$$P(X_k, X_{k+1}) = \begin{cases} 1 & \begin{cases} X_k = 0, \ X_{k+1} = 1 \\ X_k = s, \ X_{k+1} = s-1 \end{cases} \\ \dfrac{\lambda}{\lambda+\mu} & X_{k+1} = X_k + 1 \\ \dfrac{\mu}{\lambda+\mu} & X_{k+1} = X_k - 1 \end{cases}$$

Let N_o = number of state transitions from state 0.
N_k = number of state transitions from state k.
N_u = number of state transitions from state X_k to X_k+1
N_d = number of state transitions from state X_k to state X_k-1.

Lemma 4.2.2 For the M\M\1\s system, let T be the stopping time. The likelihood ratio with respect to parameter λ_0 is given by

$$L(T,\lambda,\lambda_0) = \left(\frac{\lambda}{\lambda_0}\right)^{N_o+N_u} \exp\left\{-(\lambda-\lambda_0)(T-\sum_{j\in N_k} t_j)\right\} \cdot \exp\left\{-(\lambda-\lambda_0)(T-T_{N(T)})\cdot 1_{\{X_{N(T)}=k\}}\right\}$$

and with respect to μ_0 is given by

$$L(T,\mu,\mu_0) = \left(\frac{\mu}{\mu_0}\right)^{N_k+N_d} \exp\left\{-(\mu-\mu_0)\sum_{j\notin N_0} t_j\right\} \cdot \exp\left\{-(1-1_{\{X_{N(T)}=0\}})(\mu-\mu_0)(T-T_{N(T)})\right\}.$$

Example 4.2.1 Consider an MM/1/s System which starts empty and runs for time T. We are interested in knowing the probability that the server is busy at time T, for various arrival rates. The M/M/1/s model was simulated using the following parameters. Arrival rates λ_1 =0.1 and λ_2 = 0.23, service rate μ = 1, T = 50, s = 5. The LRPE was applied to estimate the probability that the server is busy at time T for 9 perturbed arrival rates λ_0 between 0.1 and 0.23. The analytical solution for the probability that the server is busy was calculated using the randomization technique for computing transient solution of Markov process (Gross and Miller [1984]). Table 2 presents the analytical solution for the probability that the server is busy at time t, LRPE estimates for the nine rates from λ_1 and λ_2, and the interpolation estimates with their variances.

Table 1: M/M/1 Example

λ_0	CMS		LRPE from λ_1		LRBE from λ_2		Interpolation	
	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$
1.1	0.881	1.313	0.882	1.673	0.804	1.688	0.843	0.839
1.2	1.074	1.875	1.001	2.595	1.028	1.569	1.018	0.978
1.3	1.217	2.084	1.121	4.180	1.133	1.307	1.130	0.996
1.4	1.342	2.385	1.196	6.496	1.213	1.302	1.210	1.084
1.5	1.440	2.346	1.329	10.437	1.420	1.276	1.410	1.137
1.6	1.642	2.830	1.327	16.162	1.570	1.377	1.551	1.269
1.7	1.766	3.175	1.456	25.000	1.747	1.683	1.729	1.577
1.8	1.956	3.656	1.565	37.188	1.888	2.173	1.870	2.053
1.9	2.155	4.011	1.580	55.030	2.078	2.954	2.053	2.803

4.3 M/M/∞ (Responsive Servers Queuing System)

For this queuing system define the following

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \dots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & \dots \\ . & . & . & . & . \end{bmatrix}$$

$$q(X_k) = \lambda + \mu X_k$$

$$P(X_k, X_{k+1}) = \begin{cases} \dfrac{\lambda}{\lambda+\mu X_k} & X_{k+1} = X_k + 1 \\ \dfrac{\mu}{\lambda+\mu X_k} & X_{k+1} = X_k - 1 \end{cases}$$

Let

N_u = number of state transitions from state X_k to state X_k+1.

N_d = number of state transitions from state X_k to state X_k-1.

Lemma 4.3.1 Let T be the stopping time for the system, then the likelihood ratio for the M\M\∞ system with respect to parameter λ_0 is given by

$$L(T,\lambda,\lambda_0) = \left(\frac{\lambda}{\lambda_0}\right)^{N_u} \exp\left\{-\sum_{k=0}^{N(T)-1}(\lambda-\lambda_0)t_k\right\}$$

and with respect to μ_0 is given by

$$L(T,\mu,\mu_0) = \left(\frac{\mu}{\mu_0}\right)^{N_d} * \exp\left\{-(\mu-\mu_0)\left(\sum_{k=0}^{N(T)-1}(X_k)t_k + X_{N(T)}(T-T_{N(T)})\right)\right\}.$$

Example 4.3.1 Consider an M/M/∞ system which starts empty and runs for time T. We are interested in knowing the average number in the system at time T, for various arrival rates. The M/M/∞ model was simulated using the following parameters. Arrival rates $\lambda_1 = 0.1$ and $\lambda_2 = 0.2$, service rate $\mu = 0.3$, and T = 10. The LRPE was applied to estimate the average number in the system at time T for 5 perturbed arrival rates as shown in table 4. The analytical solution for the average number in the system at time T was calculated using the transient derivation for M/M/∞ model. Table 3 presents the analytical solution for the average number in the system at time T, the crude Monte-Carlo simulation, LRPE estimates for the 5 rates from λ_1 and λ_2, and the interpolation estimates with their variances.

5 CONCLUSION

We have presented an interpolation technique that uses the likelihood performance extrapolation approach to estimate the expected performance of a stochastic system for various values of the input parameters from a single simulation of the nominal system. We have introduced the Radon Nikodym derivative for some classes of Markovian queueing systems in transient analysis. We have shown, through extensive experimentations, that the proposed technique is an effective tool for measuring parameter sensitivity in the context of the Markovian queueing models in transient analysis.

Table 2: M/M/1/5 Example

λ_0	Analytical.	LRPE from λ_1		LRPE from λ_2		Interpolation	
		$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$
0.113	0.114	0.112	0.009	0.111	0.025	0.117	0.007
0.126	0.127	0.120	0.011	0.125	0.021	0.131	0.007
0.139	0.140	0.126	0.013	0.126	0.021	0.145	0.007
0.152	0.153	0.129	0.015	0.155	0.015	0.159	0.008
0.165	0.166	0.128	0.017	0.169	0.014	0.172	0.008
0.178	0.179	0.129	0.019	0.182	0.012	0.184	0.008
0.191	0.192	0.118	0.020	0.195	0.012	0.195	0.009

Table 3: M\M\∞ Example

λ_0	Anal.	CMS		LRPE from λ_1		LRBE from λ_2		Interpolation	
		$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$	$\tilde{\ell}(\lambda_0)$	Var $\tilde{\ell}(\lambda_0)$
0.11	0.348	0.337	0.329	0.350	0.421	0.347	0.218	0.348	0.143
0.15	0.498	0.501	0.499	0.499	1.862	0.497	0.409	0.498	0.335
0.19	0.602	0.602	0.603	0.593	4.407	0.601	0.537	0.600	0.478
0.05	0.158	0.158	0.155	0.161	0.087	0.157	0.100	0.159	0.046
0.40	1.267	1.258	1.256	1.033	151.91	1.359	68.613	1.258	47.265

REFERENCES

H. Arsham, A. Feuerverger, D. L. Mcleish, J. Kreimer, R.Y. Rubinstein, Sensitivity Analysis and the "What If" Problem in Simulation Analysis, Mathl. Comput. Modeling 12 (2) (1989) 193-219.

M. K. Nakayama, A. Goyal, P.W. Glynn, Likelihood Ratio Sensitivity Analysis for Markovian Models of Highly Dependable Systems. Operations Research, 42 (1994) 137-157.

R.Y. Rubinstein, The Score Function Approach for Sensitivity Analysis of Computer Simulation Models. Math Comput Sim 28, (1986) 351-379.

R.Y. Rubinstein, Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models. Operations Research 37, (1989) 72-81.

P. W. Glynn, Likelihood Ratio Gradient Estimation: An Overview. In Proceedings of the 1987 Winter Simulation Conference, A. Thesen, H. Grant and W D Kelton (eds.). IEEE Press, (1987) 366-375.

M. I. Reiman , A. Weiss , Sensitivity Analysis for Simulations Via Likelihood Ratios. Operations Research 37 (1989) 830-844.

P. L'Ecuyer, Note: On the Interchange of Derivative and Expectation for Likelihood Ratio Derivative Estimators, Management Science, 41 (1995) 738-748.

P. L'Ecuyer, A Unified View of the IPA, SF and LR Gradient Estimation Techniques. Management Science, 36 (1990) 1364-1383.

G. S. Fishman, Monte Carlo: Concepts, Algorithms, and Applications, Springer-Verlag, New York 1996.

R. B. Cooper, Introduction to Queueing Theory, 2nd Ed., North Holland 1981.

D. Gross, C. M. Harris, Fundamentals of Queueing Theory, 3rd Ed., Wiley, New York 1998.

L. Kleinrock, Queueing Systems, Wiley, New York 1975.

D. Gross, D. R. Miller, The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. Operations Research, 32, 2 (1984) 343-363.

AUTHOR BIOGRAPHIES

MOHAMED A. AHMED is an associate professor in the Department of Statistics and Operations Research at Kuwait University. He received a B.S. in Production Engineering from Helwan University-Egypt, a M.S. and a Ph.D. in Operations Research from George Washington University. His research interests lie in the area of inventory control, simulation, stochastic optimization, and applied probability. His e-mail address is `wakeel@kuc01.kuniv.edu.kw`.

TALAL M. ALKHAMIS is an associate professor in the Department of Statistics and Operations Research at Kuwait University. He received a B.S. in Computer Science & Statistics from Kuwait University, a M.S. in Computer Science, and a Ph.D. in Operations Research from the Florida Institute of Technology. His research interests include discrete Optimization, meta-heuristics, and stochastic optimization. His e-mail address is `alkhamis@kuc01.kuniv.edu.kw`.