Proceedings of the 1987 Winter Simulation Conference
A. Thesen, H. Grant, W. David Kelton (eds.)

SIMTERPOLATIONS: ESTIMATING AN ENTIRE QUEUEING FUNCTION FROM A SINGLE SAMPLE PATH

Martin I. Reiman
AT&T Bell Laboratories

Burton Simon
University of Colorado at Denver

J. Stanford Willie
US West Advanced Technologies

ABSTRACT

We present an overview of a method for estimating an entire queueing function, $f(\lambda), 0 \le \lambda < c$, from a single simulation sample path, where λ is the arrival rate to the system and c is the system capacity. For example, $f(\lambda)$ could be the average sojourn time in a queueing network as a function of λ. Recently methods have been developed that allow one to simultaneously estimate $f(0)$, $f'(0)$, $f(\lambda^*)$ $f'(\lambda^*)$, and h (the heavy traffic limit of f) based on the sample path from a single simulation experiment in which the arrival rate to the system is λ^*. 'Standard' simulation methodology has generally focused on obtaining only the point estimate of $f(\lambda^*)$ from this one sample path. The computational costs associated with obtaining all five estimates (as well as an estimate of the asymptotic covariance of the estimates) is only slightly higher than the costs associated with obtaining an estimate of $f(\lambda^*)$ alone. We propose a regenerative simulation methodology to construct estimates of $f(0)$, $f'(0)$, $f(\lambda^*)$ $f'(\lambda^*)$, and h and an approximation to the joint distribution of the estimates. We then outline a method for fitting a polynomial to a 'normalized' version of the estimates. A reverse normalization of the fitted polynomial yields an estimate of $f(\lambda), 0 \le \lambda < c$.

1. INTRODUCTION

Often one is interested in a function of the arrival rate to an open queueing system, $f(\lambda), 0 \le \lambda < c$, where c is the capacity of the system, i.e., $c = sup\{\lambda$: *system with arrival rate* λ *is stable*$\}$. For example, $f(\lambda)$ might be the average sojourn time in the system or the variance of the queue length at some node or the 0.95^{th} quantile of the sojourn time distribution. Functions of this sort arise, for example, in models of computer and communication systems. The class of models for which analytic results are available is rather restrictive, so one is often confronted with the task of analyzing the model via simulation. One of the advantages of an analytic solution of a model is the ability to look at the entire function, $f(\lambda), 0 \le \lambda < c$. In this way many questions such as 'How big can λ be before a design requirement is violated?' can be answered easily. Typically, the same question can be difficult to answer via 'standard' simulation methodology, since several simulation experiments at different arrival rates will generally be required to provide the answer.

We propose a method that produces an estimate of $f(\lambda), 0 \le \lambda < c$ from the sample path associated with a single simulation experiment in which the arrival rate to the system is some $\lambda^*, 0 < \lambda^* < c$. In this paper we restrict ourselves to estimating mean sojourn times for regenerative systems, although higher moments and quantiles can be estimated as well.

The method can be outlined as follows. A 'normalized' version of $f(\lambda)$ (call it $g(\lambda)$) is chosen so that a heavy traffic limit for $g(\lambda)$ can be obtained. For our purposes (mean sojourn times) the normalized function will always have the form $g(\lambda) = (c-\lambda)f(\lambda)$, although other normalizations are possible. We choose λ^* and conduct a simulation experiment in which the arrival rate to the system is λ^*. Based on this single simulation experiment we construct estimates of

$g(0)$, $g'(0)$, $g(\lambda^*)$, $g'(\lambda^*)$ and $h = \lim_{\lambda \to c} g(\lambda)$, and an approximation to the large sample distribution of the estimates. All five estimates and the approximation to their large sample distribution are obtained at little additional computational cost above those associated with simply estimating $f(\lambda^*)$.

Using estimates of $g(0)$, $g'(0)$, $g(\lambda^*)$, $g'(\lambda^*)$, and h, and the approximation of the large sample distribution of the estimates, a polynomial interpolation, $\hat{g}(\lambda)$, is constructed that 'best' fits the estimates (in the least squares sense). We choose $\hat{g}(\lambda)$ to be the smallest degree polynomial that gives a reasonable fit. Once we have $\hat{g}(\lambda)$, the estimate, $\hat{f}(\lambda)$, of $f(\lambda)$ is obtained by reversing the normalization, i.e., $\hat{f}(\lambda) = \dfrac{\hat{g}(\lambda)}{(c-\lambda)}$.

A 'simterpolation' is an interpolation constructed from data generated from a simulation. In this paper we briefly describe the individual pieces of the simterpolation. In Reiman, Simon, and Willie (1987) the pieces are described in more detail and some experimental results will be presented.

In Sections 2, 3, and 4 we describe certain relationships that are key to the methodology for constructing estimates of $g(0)$, $g'(0)$, $g(\lambda^*)$, $g'(\lambda^*)$ and h. Section 5 briefly addresses the regenerative simulation methodology for constructing the estimates and presents a large sample approximation to the distribution of the estimates. In section 6 we describe the polynomial interpolation, $\hat{g}(\lambda)$, and the approximation, $\hat{f}(\lambda)$. We conclude with a short discussion in section 7.

In all that follows, we fix λ^*, $0 < \lambda^* < c$ and simulate the system with an arrival rate of λ^*. All probabilities and expectations are assumed to be with respect to λ^* unless stated otherwise. Also, we define $g(\lambda) = (c-\lambda)f(\lambda)$. The systems are assumed to be regenerative and the arrival processes are Poisson.

2. LIGHT TRAFFIC

In Reiman and Simon (1988a), a general method is presented for analytically determining the n^{th} derivative of $f(\lambda)$ at $\lambda = 0$ for a large class of functions and systems. If $f(\lambda)$ is the average sojourn time then $f(\lambda) = E_\lambda(\psi)$, where $\psi(\omega)$ is the sojourn time of a 'tagged' customer placed in the system at time zero, for the sample path, ω. Light traffic derivatives of order zero and one can be obtained analytically by the following formulas:

$$f(0) = \hat{\psi}(\{\varnothing\}) ,$$

$$f'(0) = \int_{-\infty}^{\infty} [\hat{\psi}(\{t\}) - \hat{\psi}(\{\varnothing\})]dt ,$$

where $\hat{\psi}(\{\varnothing\}) = E(\psi \mid$ *no arrivals in all of time* $)$, $\hat{\psi}(\{t\}) = E(\psi \mid$ *one arrival in all of time; at time t*$)$. Expressions for higher order derivatives are given in Reiman and Simon (1988a). Although $f(0)$ and $f'(0)$ can be explicitly computed this way for fairly general systems (e.g. Markovian networks of priority queues), there is no obvious way to evaluate or estimate $\hat{\psi}(\{t\})$ during the course of a simulation. We therefore seek an alternate approach for computing light traffic limits.

Clearly, $f(0)$ is the expected sojourn time of a customer in an empty system. One can determine the sojourn time that each customer would have experienced in an empty system by summing all its service times. Thus, if v_{ij} is the total service time needed by the j^{th} customer in the i^{th} busy period, then the v_{ij}'s are iid and $E(v_{ij}) = f(0)$. Also, if $v_i = \sum_j v_{ij}$ and N_i is the number of customers in the i^{th} busy period then

$$f(0) = \frac{E(v_i)}{E(N_i)} ,$$

and

$$g(0) = cf(0) .$$

Let

$$V_i^{(1)} = (\lambda^*)^{1-N_i} e^{\lambda^* T_i} 1_{\{N_i \le 2\}} (-T_i)^{2-N_i} W_i ,$$

$$V_i^{(2)} = e^{\lambda^* T_i} 1_{\{N_i = 1\}} W_i ,$$

and

$$V_i^{(3)} = \frac{1}{\lambda^*} e^{\lambda^* T_i} 1_{\{N_i = 2\}} ,$$

where T_i is the length of the i^{th} busy period and $W_i = \sum_j W_{ij}$, where W_{ij} is the sojourn time of the j^{th} customer of the i^{th} busy period. In Reiman and Weiss (1987b), the following formula is derived:

$$f'(0) = E(V_i^{(1)}) - E(V_i^{(2)})\,E(V_i^{(3)}) .$$

It follows that

$$g'(0) = c\,f'(0) - f(0).$$

For each busy period in the single simulation experiment at λ^*, it is clear that all the random variables needed to evaluate $V_i^{(1)}$, $V_i^{(2)}$, and $V_i^{(3)}$ are readily available. Analogous expressions exist for higher order light traffic limits (see Reiman and Weiss (1987b)), although the variance of the estimates will increase rapidly with the order.

3. HEAVY TRAFFIC

In Reiman (1987) an expression is given for the heavy traffic limit of the sojourn time distribution of networks of priority queues with a single bottleneck node. From this expression, the heavy traffic limit of the average sojourn time is easily seen to be

$$h = \frac{M\,\Theta}{\Gamma} ,$$

Where M is the expected number of visits a customer makes to the bottleneck node at lowest priority, Θ is a constant times the second moment of the total service time a customer needs at the bottleneck node, and Γ is the expected value of a certain weighted sum of individual service times at the bottleneck node. If the routing of the customers in the network is Markovian then h can be computed explicitly (see Simon (1987)). In those cases, and in even more general cases (non-Markov routing), h can be estimated from the simulation. The j^{th} customer of the i^{th} busy period generates random variables, M_{ij}, Θ_{ij} and Γ_{ij} which are easily evaluated and are iid, with $E(M_{ij}) = M$, $E(\Theta_{ij}) = \Theta$ and $E(\Gamma_{ij}) = \Gamma$. In particular,

$$M = \frac{E(M_i)}{E(N_i)} , \qquad \Theta = \frac{E(\Theta_i)}{E(N_i)} ,$$

and

$$\Gamma = \frac{E(\Gamma_i)}{E(N_i)} ,$$

where $M_i = \sum_j M_{ij}$, $\Theta_i = \sum_j \Theta_{ij}$ and $\Gamma_i = \sum_j \Gamma_{ij}$.

4. THE DERIVATIVE AT λ^*

In recent years there has been quite a bit of work aimed at obtaining sensitivity measures (i.e., derivatives of the function being estimated) from a simulation. The method we choose to use is the likelihood ratio method (Reiman and Weiss (1987a)). Other alternatives are infinitesimal perturbation analysis (Suri (1983)) and the light traffic perturbation method (Simon (1987)). The scope of infinitesimal perturbation analysis is too restrictive for our purposes at this time, and the light traffic perturbation method appears to be too difficult to implement in a general setting.

The likelihood ratio method is easy to implement and is widely applicable. For the i^{th} busy period, let

$$D_i = \left[\frac{N_i}{\lambda^*} - T_i\right] W_i ,$$

and

$$N_i^* = \left[\frac{N_i}{\lambda^*} - T_i\right] N_i .$$

Then,

$$f'(\lambda^*) = \frac{E(D_i)}{E(N_i)} - \frac{E(W_i)}{E(N_i)}\,\frac{E(N_i^*)}{E(N_i)} ,$$

and

$$g'(\lambda^*) = (c-\lambda^*)f'(\lambda^*) - f(\lambda^*).$$

Again, the quantities needed to estimate $f'(\lambda^*)$ are easily obtained during the course of the simulation. Analogous expressions exist for higher order derivatives (see Reiman and Weiss (1987a)), although the variance of the estimates will increase rapidly with the order.

5. LARGE SAMPLE THEORY

In this section we present a large sample approximation to the joint distribution of the estimates of $g(0)$, $g'(0)$, $g(\lambda^*)$, $g'(\lambda^*)$ and h based on a single simulation experiment in which the arrival rate to the system is λ^*. Suppose that the sample path from the simulation experiment is a realization of a regenerative stochastic process and that data for (exactly) n busy periods is generated. Let $g_n(\lambda^*) = (c - \lambda^*) f_n(\lambda^*)$, where $f_n(\lambda^*)$ is the estimate of $f(\lambda^*)$ based on n busy periods and constructed via standard regenerative simulation methodology (see, for example Iglehart and Shedler (1980)), i.e.,

$$f_n(\lambda^*) = \frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} N_i} .$$

Similarly, let $g_n(0)$, $g_n'(0)$, $g_n'(\lambda^*)$ and h_n denote the estimates of $g(0)$, $g'(0)$, $g'(\lambda^*)$ and h, respectively, constructed from the relationships described in sections 2, 3 and 4.

Based on the theory of regenerative stochastic processes, it can be shown that under quite general conditions, as $n \rightarrow \infty$ the asymptotic distribution of the vector

$$\sqrt{n} \begin{bmatrix} g_n(0) - g(0) \\ g_n'(0) - g'(0) \\ g_n(\lambda^*) - g_n(\lambda^*) \\ g_n'(\lambda^*) - g'(\lambda^*) \\ h_n - h \end{bmatrix}$$

is 5-variate normal with mean vector zero and covariance matrix $\mathbf{C} = (\sigma_{ij})$. Furthermore, consistent estimates of the elements of $\mathbf{C}$ can be constructed from the simulation experiment. Details of the large sample theory will be presented in Reiman, Simon, and Willie (1987).

6. INTERPOLATION METHODOLOGY

In this section we describe the construction of the estimate of the function $f(\lambda)$, $0 \leq \lambda < c$, based on the estimates $g_n(0)$, $g_n(0)$, $g_n(\lambda^*)$, $g_n'(\lambda^*)$, and h_n of the previous section. The methodology described here is, in a sense, a generalization of the interpolation methodologies described in Reiman and Simon (1988b) and Simon, and Willie (1986).

Suppose that for $0 \leq \lambda < c$,

$$g(\lambda) \approx \sum_{k=0}^{d} b_k \lambda^k$$

for some integer, d, and coefficients $b_0, \cdots, b_d$. In other words, we assume that the normalized $f(\lambda)$ can be sufficiently well approximated by a polynomial of (hopefully low) order, d. Empirical evidence suggests that a properly normalized $f(\lambda)$ can be well approximated by quadratic or cubic polynomial in many cases of interest (see Reiman and Simon (1988b), Simon, and Willie (1986)). Thus, the problem of estimating $f(\lambda)$, $0 \leq \lambda < c$ is one of determining the order, d, and coefficients, $b_0, \ldots, b_d$ of the polynomial approximation to $g(\lambda)$ and then performing a reverse normalization.

Before proceeding further, we introduce some notation: Let

$$\mathbf{Y}_n = \left[g_n(0), g_n'(0), g_n(\lambda^*), g_n'(\lambda^*), h_n \right]^T ,$$

$$\mathbf{b} = (b_0, b_1, \cdots, b_d)^T ,$$

and let $\mathbf{C}_n$ denote the estimate of the asymptotic covariance matrix of $\mathbf{Y}_n$ from the previous section (i.e., $\mathbf{C}_n$ is the 5×5 matrix with elements $\hat{\sigma}_{ij} / n$, where $\hat{\sigma}_{ij}$ is a consistent estimate of σ_{ij} based on n busy periods). The matrix $\mathbf{X}$ will denote the $5 \times (d+1)$ matrix, which for $d \leq 4$ is given by the first $d+1$ columns of the matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & \lambda^* & (\lambda^*)^2 & (\lambda^*)^3 & (\lambda^*)^4 \\ 0 & 1 & 2\lambda^* & 3(\lambda^*)^2 & 4(\lambda^*)^3 \\ 1 & c & c^2 & c^3 & c^4 \end{bmatrix} .$$

Suppose for the moment that the order, d, of the polynomial approximation of $g(\lambda)$, $0 \leq \lambda \leq c$ is known. Here, the results presented in section 5 strongly suggest (see Lewis and Odell (1971)) that for sufficiently large n, the coefficient vector be chosen so that

$$(\mathbf{Y}_n - \mathbf{X}\mathbf{b})\, \mathbf{C}_n^{-1} (\mathbf{Y}_n - \mathbf{X}\mathbf{b})^T$$

is minimized. From the classical form of the Gauss-Markov Theorem (again, see Lewis and Odell (1971)), the minimum is attained when $\mathbf{b} = \hat{\mathbf{b}}$, where

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{C}_n^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}_n^{-1} \mathbf{Y}_n \, .$$

Our estimate of the function $f(\lambda)$, $0 \le \lambda < c$ is obtained by a reverse normalization of the fitted polynomial, i.e.,

$$\hat{f}(\lambda) = \frac{\mathbf{b}^T \mathbf{\Lambda}}{c - \lambda} \, ,$$

where $\mathbf{\Lambda} = (1, \lambda, \lambda^2, \ldots, \lambda^d)^T$. The variance of $\hat{f}(\lambda)$ is given by

$$\frac{\mathbf{\Lambda}^T (\mathbf{X}^T \mathbf{C}_n^{-1} \mathbf{X}) \mathbf{\Lambda}}{(c - \lambda)^2} \, .$$

Unfortunately, the order of the polynomial, $g(\lambda)$, is generally unknown (except for a few simple systems) and must be inferred from the data. Techniques for choosing d will be discussed in Reiman, Simon, and Willie (1987).

In many cases, some of the quantities being estimated in the simulation can be determined exactly by analytic means. For example, in the cases we are considering here (average sojourn times), $f(0)$ can often be determined by inspection. As was pointed out in sections 2 and 3, if the system is Markovian, it might be possible to compute $f'(0)$, and if the customer routing is not too complex, it might be possible to compute h. When some of the data is known exactly, the problem of estimating $g(\lambda)$, $0 \le \lambda \le c$ becomes a constrained least squares problem. The constrained least squares problem is treated in Lewis and Odell (1971). A few of the most common cases (e.g. when $f(0)$ and h are known exactly) will be worked out in detail in Reiman, Simon, and Willie (1987).

7. DISCUSSION

Simterpolations can be constructed for any function of any system for which it is possible to simultaneously estimate $f(0)$, $f'(0)$, $f(\lambda^*)$, $f'(\lambda^*)$, h and $\mathbf{C}$ from a single simulation. As we have shown, average sojourn times in networks of priority queues with a single bottleneck node can be handled easily. The jump to estimating higher moments is straightforward.

Initially, one might think that since the simulation is run at $\lambda = \lambda^*$, the simterpolation's accuracy will deteriorate as λ moves away from λ^*. The estimate of $f(0)$ and h typically have lower variance than the estimates of $f(\lambda^*)$, so the simterpolation is in fact more accurate near the endpoints than it is in the middle. This will obviously be the case when $f(0)$ and h can be determined exactly.

There are numerous interesting problems that come up in the implementation of the simterpolation method that we have not fully addressed. Among them are:

1. What is the best λ^* to simulate at?
2. How many derivatives (in light traffic and at λ^*) should be estimated?
3. How should one normalize $f(\lambda)$?
4. How does one choose the order of the polynomial approximation of $g(\lambda)$?

REFERENCES

Iglehart, D. L. and Shedler, G. S. (1980), Regenerative Simulation of Response Times in Networks of Queues, Lecture Notes in Control and Information Sciences, Springer-Verlag, New York.

Lewis, T. O. and Odell, P. L. (1971), Estimation in Linear Models, Prentice Hall, Inc., Englewood Cliffs, NJ.

Reiman, M. I. (1987), A Network of Priority Queues in Heavy Traffic: One Bottleneck Node, draft.

Reiman, M. I. and Simon, B. (1988a), Open Queueing Systems in Light Traffic, *Math. Oper. Res.*, to appear.

Reiman, M. I. and Simon, B. (1988b), An Interpolation Approximation for Queueing Systems with Poisson Input, *Oper. Res.*, to appear.

Reiman, M. I., Simon, B. and Willie, J. S. (1987), Simterpolations, in preparation.

Reiman, M. I. and Weiss, A. (1987a), Sensitivity Analysis for Simulations via Likelihood Ratios, draft.

Reiman, M. I. and Weiss, A. (1987b), Light

Traffic Derivatives via Likelihood Ratios, draft.

Simon, B. (1987), Computing Heavy Traffic Limits for Networks of Priority Queues with Nested Markov Routing, draft.

Simon, B. (1987), Light Traffic Perturbations of Queueing Systems, draft.

Simon, B. and Willie, J. S. (1986), Estimation of Response Time Characteristics in Priority Queueing Networks via an Interpolation Methodology Based on Simulation and Heavy Traffic Limits, *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, T.J. Boardman, Ed. American Statistical Association, Wash. D.C., 251-256.

Suri, R. (1983), Infinitesimal Perturbation Analysis of Discrete Event Dynamic Systems: A General Theory, IEEE Decision and Control Conference, San Antonio, TX.

AUTHORS' BIOGRAPHIES

MARTIN I. REIMAN received the B.A. degree in Physics and Mathematics from Cornell University in 1974 and the Ph.D. degree in Operations Research from Stanford University in 1977. Since 1977 he has been with AT&T Bell Laboratories. From 1977 to 1980 he was in the Data Communications Laboratory in Holmdel, New Jersey. Since 1980 he has been in the Mathematical Sciences Research Center in Murray Hill, New Jersey. His current research interests are in the analysis and control of stochastic systems.

Martin I. Reiman
AT&T Bell Laboratories
600 Mountain Ave
Murray Hill, NJ 07974
(201) 582-2368

BURTON SIMON is an assistant professor in the Department of Mathematics at the University of Colorado at Denver. He received a Ph.D. in Mathematics and Operations Research from the University of Michigan in 1980. From 1980 to 1986 he was at Bell Laboratories, Holmdel, New Jersey, and AT&T Information Systems, Denver, Colorado, where he was involved in computer system modeling and research in queueing theory. He has been at the University of Colorado at Denver since 1986. His current research interests include the analysis of functions of queueing systems and stochastic system modeling.

Burton Simon
Department of Mathematics
University of Colorado at Denver
1100 14th st.
Denver, CO 80202
(303) 556-8444

STAN WILLIE is presently a member of the Advanced Systems Research Group of US West Advanced Technologies. From 1979 to 1986 he worked at Bell Laboratories in Holmdel, New Jersey and Denver, Colorado where he was involved in the design and development of advanced systems. From 1973 to 1975 Stan worked in a biomathematics group at the University of California, San Francisco. He received Ph.D. and M.A. degrees in statistics from the University of California, Berkeley in 1979 and 1973 respectively and a B.S. in mathematics from the University of Utah in 1972. Stan is a member of the American Statistical Association, the Biometric Society, and the IEEE. His current interests include the analysis of time dependent random phenomena and computer and communication system modeling and analysis.

Stan Willie
US West Advanced Technologies
6200 South Quebec St.
Englewood, CO 80111
(303) 889-6038