J. Austral. Math. Soc. Ser. B 31(1989), 135–149

SENSITIVITY BOUNDS ON A $GI/M/n/n$ QUEUEING SYSTEM

ANDREW COYLE[1]

(Received December 1988; revised January 1989)

Abstract

A method for determining the upper and lower bounds for performance measures for certain types of Generalised Semi-Markov Processes has been described in Taylor and Coyle [8]. A brief description of this method and its use in finding an upper bound for the time congestion of a $GI/M/n/n$ queueing system will be given. This bound turns out to have a simple form which is quickly calculated and easy to use in practice.

1. Introduction

Much work has gone into the analysis of teletraffic systems in which the arrival traffic streams are Poisson. With the introduction of a heterogeneous ISDN telecommunications system much of the offered traffic to the system is no longer Poisson. The investigation of systems where the arrival traffic streams are non-Poisson is therefore of much importance.

Often the exact nature of an arrival stream is unknown; the only data we have about it being the mean interarrival time. Because of this it is of interest to determine bounds on the variation of performance measures as the interarrival distributions vary over a set of distributions with fixed mean.

In this paper we study the time congestion in the $GI/M/n/n$ queue. To do this we model this queue as a Generalised Semi-Markov Process (see for example König and Jansen [4]) and use this formulation to find sensitivity bounds for the time congestion. A GSMP is insensitive when the steady state probabilities of the system depends only on the mean values of the generally distributed lifetimes. That is, no matter what the actual distribution of the

[1] Department of Applied Mathematics, The University of Adelaide, South Australia 5001.

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

general lifetimes the steady state probability distribution of the system can be determined when the mean values of these lifetimes are known.

Many GSMPs turn out to be insensitive. Others, including the $GI/M/n/n$ queue, are not insensitive and so the value of the steady state probabilities of the system do depend on the actual form that the generally distributed lifetimes take, not only on the mean value of the distributions. In a non insensitive GSMP there must exist at least two different lifetime distributions with a fixed mean value which give rise to different steady state probability distributions. The supremum and infimum of a performance measure as the lifetime distribution range over the set of distributions with fixed mean form the sensitivity bounds for this performance measure.

In Taylor and Coyle [8] a method for finding the sensitivity bounds for GSMPs in which there is only one generally distributed lifetime is presented. This method is described in Section 2 of this paper for the case when the lifetime is always active. In Section 3 we apply this method to the $GI/M/n/n$ system. In Section 4 the method of Lagrange multipliers is briefly described and in Section 5 an upper bound for the time congestion in a $GI/M/n/n$ system is found using the method of Lagrange multipliers. The proof of this upper bound when $n = 2$ is given in this paper; the full proof can be found in Coyle [2].

2. The method

In Taylor and Coyle [8] a method for finding the upper and lower values of a performance measure for a GSMP with only a single general lifetime is presented. In the situation described in that paper the single general lifetime may or may not be active at any one time. In this paper a brief description of this result when the single general distribution is always active is given. This situation often holds when the general distribution is an interarrival distribution in a renewal input stream. For the case when the general lifetime is not always active there are two sets of possible states of the system, one set corresponding to when the general lifetime is active and the other corresponding to when the general lifetime is not active. In this paper only the states corresponding to when the general lifetime is active are used and so the result presented is a simplified version of the full result given in Taylor and Coyle [8].

We start with an irreducible Markov Process $\mathscr{M}$ on a set of possible states A. The process $\mathscr{M}$ has one active lifetime which is generally distributed with distribution function $G(\cdot)$. Without loss of generality it can be assumed that

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

this lifetime has unit mean. Denote the set of all possible general distributions with unit mean by $\mathcal{G}$.

Two types of transitions may occur in this process, those that occur when a general lifetime finishes, with rates $q^E(x,x')$ from state $x \in A$ to state $x' \in A$, and those that occur whilst the general lifetime is still active, with rates $q^I(x,x')$ from $x \in A$ to $x' \in A$. The general lifetime will be worked off at a rate $c(x)$ when the process is in state $x \in A$, where $c(x) = \sum_{x' \in A} q^E(x,x')$. The transition rate matrices are,

$$\begin{aligned}
[Q_1]_{x,x'} &= \begin{cases} q^I(x,x'), & x \neq x' \\ -\sum_{z \in A} q^I(x,z), & x = x' \end{cases} \\
[Q^E]_{x,x'} &= q^E(x,x'), \\
[C]_{x,x'} &= \begin{cases} 0 & x \neq x' \\ c(x), & x = x' \end{cases}
\end{aligned} \tag{2.1}$$

where $x, x' \in A$, and we define

$$Q_2 = -C + Q^E.$$

A property of the matrix $Q_1 C^{-1}$ is that there are no eigenvalues with positive real part and the dominant eigenvalue α_0 is always equal to zero; see Taylor [7]. Here we assume that the eigenvalues are distinct. If this is not the case then the following method must be altered accordingly. Assume that A contains $n + 1$ states. Thus the eigenvalues of $Q_1 C^{-1}$ can be written in descending order as $-\alpha_i$, $i = 0, \ldots, n$ and the respective left eigenvectors as $\mathbf{w}_i$.

Let $p_x(y, G)$ be the stationary probability density that the process is in state x with a spent lifetime y for $x \in A$ when the general lifetime is distributed according to $G(\cdot)$. Define $\mathbf{P}(y, G) = (p_x(y, G), x \in A)$ as the vector containing these densities. So $\int_0^\infty \mathbf{P}(y, G)\, dy$ gives the probability densities that the process is in the discrete states x. It is shown in Taylor [7] that $\mathbf{P}(y, G)$ satisfies

$$\left[h(y)\mathbf{P}(y,G) + \frac{d}{dy}\mathbf{P}(y,G)\right] C - \mathbf{P}(y,G)Q_1 = 0 \tag{2.2a}$$

and

$$\int_0^\infty \mathbf{P}(y,G)h(y)\, dy Q^E = \mathbf{P}(0,G)C \tag{2.2b}$$

where $h(y)$ is the hazard function associated with the distribution $G(\cdot)$.

Simple manipulations of (2.2) will give us

$$\int_0^\infty \mathbf{P}(y,G)\, dy Q_1 + \int_0^\infty \mathbf{P}(y,G)h(y)\, dy Q_2 = 0. \tag{2.3}$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

Solving the differential equation (2.2a) using the spectral representation of Q_1C^{-1} (Taylor [7]) we find that

$$\int_0^\infty \mathbf{P}(y,G)\,dy = A_0^{(G)}\mathbf{w}_0 + \sum_{i=1}^{n}(A_i^{(G)}/\alpha_i)[1-\hat{G}(\alpha_i)]\mathbf{w}_i \tag{2.4a}$$

and

$$\int_0^\infty \mathbf{P}(y,G)h(y)\,dy = A_0^{(G)}\mathbf{w}_0 + \sum_{i=1}^{n} A_i^{(G)}\hat{G}(\alpha_i)\mathbf{w}_i \tag{2.4b}$$

where $\hat{G}(\alpha_i) = \int_0^\infty \exp(-\alpha_i y)\,dG(y)$ and the constants $A_i^{(G)}$ depend on $G(\cdot)$.

THEOREM 1. *Let $F(\int_0^\infty \mathbf{P}(y,G)\,dy) \equiv H(G)$ be a bounded function of the stationary distribution of $\mathcal{M}$. The supremum of $H(G)$ is less than or equal to the solution of the constrained optimisation problem $\mathcal{P}$, in the variables A_i, $i = 0,\dots,n$ and X_i, $i = 1,\dots,n$ defined by*

$$\max F\left(A_0\mathbf{w}_0 + \sum_{i=1}^{n}(A_i/\alpha_i)[1-X_i]\mathbf{w}_i\right) \tag{2.5}$$

subject to the constraints

$$e^{-\alpha_i} \le X_i \quad \textit{for} \;\; 2 \le i \le n, \tag{2.6}$$

$$X_1 \le 1, \tag{2.7a}$$

$$X_i \le X_{i-1} \quad \textit{for} \;\; 2 < i \le n, \tag{2.7b}$$

$$(X_1\alpha_2 + \alpha_1 - \alpha_2)/\alpha_1 \le X_2, \tag{2.8a}$$

$$[X_i(\alpha_{i+1}-\alpha_{i-1}) + X_{i-1}(\alpha_i - \alpha_{i+1})]/(\alpha_i - \alpha_{i-1}) \le X_{i+1} \quad \textit{for}\; 2 < i < n-1 \tag{2.8b}$$

$$\left[A_0\mathbf{w}_0 + \sum_{i=1}^{n}(A_i/\alpha_i)[1-X_i]\mathbf{w}_i\right]Q_1 + \left[A_0\mathbf{w}_0 + \sum_{i=1}^{n}A_iX\mathbf{w}_i\right]Q_2 = 0 \tag{2.9}$$

and

$$\left[A_0\mathbf{w}_0 + \sum_{i=1}^{n}(A_i/\alpha_i)[1-X_i]\mathbf{w}_i\right]\mathbf{e} = 1 \tag{2.10}$$

where $\mathbf{e}$ *is a vector of* 1*s*.

PROOF. Represent $\hat{G}(\alpha_i)$ and $A_i^{(G)}$ by the variables X_i and A_i respectively. The form of the objective function (2.5) follows from (2.4a). Substitution of (2.4a) and (2.4b) back into (2.3) gives the equality constraint (2.9). Since the sum of all the probabilities must be 1 the constraint (2.10) can be found by summing the probability density vector given by (2.4a). From Jensen's inequality we know that

$$\exp(-\alpha_i) \leq X_i$$

and so the inequality constraint (2.6) must be satisfied. Constraints (2.7) and (2.8) follow from the fact that the function $\hat{G}(s)$ is completely monotone (i.e. $(-1)^m d^m\hat{G}(s)/ds^m \geq 0 \ \forall m \geq 0$; see Feller [3]); a proof of this result is given in Taylor and Coyle [8].

Any feasible solution to the problem $\mathscr{P}$ must satisfy these constraints. Therefore a solution that maximises (2.5) and satisfies the constraints must be greater than or equal to the maximum feasible solution.

3. The $GI/M/n/n$ queue

We will now look at a specific GSMP, the $GI/M/n/n$ queue with arrival rate a. A GI arrival distribution is one for which the interarrival lifetimes are generally and independently distributed. For this system the arrivals are offered to n servers each with negative exponentially distributed service times. Since in this situation the general lifetime is always active the above version of the full theorem can be used. The states of this GSMP correspond to the number of busy servers in the system and so there are $n+1$ possible states. If $c(i) = a$, $\forall\ 0 \leq i \leq n$ the general lifetime is being worked off at rate a. This is equivalent to an arrival rate a. When a general lifetime has been worked off an arrival occurs and another general lifetime begins. If there is a spare server this server will service the new arrival, and if no spare exists the call is lost. Without loss of generality assume that the service times have mean 1 and therefore the rate at which a transition from state $i, i \in \{1, \dots, n\}$ to state $i-1$ takes place is i. So we have

$$q^E(i,j) = \begin{cases} a, & j = i+1, \\ 0, & \text{otherwise} \end{cases} \quad i,j = 0,\dots,n,$$

$$q^I(i,j) = \begin{cases} i, & j = i-1, \\ -i, & j = i, \\ 0, & \text{otherwise} \end{cases} \quad i,j = 0,\dots,n,$$

$$C(i,j) = \begin{cases} a, & j = i, \\ 0 & \text{otherwise}, \end{cases} \quad i,j = 0,\dots,n$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

and so it can be shown that

$$Q_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 2 & -2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 3 & -3 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & 2-n & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & n-1 & 1-n & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & n & -n \end{pmatrix}$$

and

$$Q_2 = \begin{pmatrix} -a & a & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -a & a & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & -a & a & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -a & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & -a & a & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -a & a \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of Q_1C^{-1} are

$$-\alpha_i = -i/a, \qquad i = 0, \dots,$$

and the corresponding left eigenvectors are $\mathbf{w}_i$ with coordinates

$$(\mathbf{w}_i)_j = \begin{cases} (-1)^j \binom{i}{j}, & 0 \le j \le i \le n, \\ 0, & \text{otherwise.} \end{cases}$$

Given these results and from (2.8a) and (2.8b) we get

$$-1 \le X_2 - 2X_1, \tag{3.1a}$$

and

$$0 \le X_{i+1} - 2X_i + X_{i-1}, \qquad i = 2, \dots, n-1. \tag{3.1b}$$

The dot product of the right and left eigenvectors of a matrix corresponding to different eigenvalues is 0. Therefore since $\mathbf{e}$ is the right eigenvector of Q_1C^{-1} corresponding to the eigenvalue 0,

$$\mathbf{w}_i.\mathbf{e} = \begin{cases} 1, & i = 0, \\ 0, & i = 1, \dots, n \end{cases}$$

and so from (2.10)

$$A_0 = 1. \tag{3.2}$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

Also

$$\mathbf{w}_i Q_1 = (-a)\alpha_i \mathbf{w}_i$$

and

$$\mathbf{w}_i Q_2 = \begin{cases} (-a)\mathbf{w}_{i+1}, & i = 0, \dots, n-1, \\ (-a)\hat{\mathbf{w}}, & i = n \end{cases}$$

where

$$(\hat{\mathbf{w}})_k = \begin{cases} (-1)^k \binom{n+1}{k}, & k = 0, \dots, n-1, \\ (-1)^n \binom{n}{n-1}, & k = n. \end{cases}$$

Hence from (2.9)

$$\sum_{i=1}^{n} A_i [X_\iota - 1]\mathbf{w}_i - A_0 \mathbf{w}_1 - \sum_{i=1}^{n-1} A_i X_i \mathbf{w}_{i+1} - A_n X_n \hat{\mathbf{w}} = \mathbf{0}. \tag{3.3}$$

In this case we want to look at the maximum possible time congestion in the system, and so the objective function we are looking at is the probability that n servers are occupied. Since

$$\int_0^\infty \mathbf{P}(y, G)\, dy = A_0 \mathbf{w}_0 + \sum_{i=1}^{n} (A_i/\alpha_\iota)[1 - X_i]\mathbf{w}_i$$

is the probability density vector, the nth component of this vector is the probability that n servers are occupied. The nth component of all the eigenvectors apart from the nth one is zero. The nth component of the nth eigenvector is $(-1)^n$ and so the probability that n servers are busy in given by

$$P_n = (-1)^n (aA_n/n)[1 - X_n]. \tag{3.4}$$

4. The method of Lagrange multipliers

To solve a nonlinear bounded optimisation problem with nonlinear equality and inequality constraints, the method of Lagrange multipliers can be used (see for example Avriel [1]). If the optimisation problem is

$$\text{Max } f(\mathbf{x}), \tag{4.1}$$

$$\text{such that} \quad h_j(\mathbf{x}) = 0, \qquad j = 1, \dots, p, \tag{4.2}$$

$$\text{and} \quad g_i(\mathbf{x}) \geq 0, \qquad i = 1, \dots, m, \tag{4.3}$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

then if $\mathbf{x}^*$ is a feasible solution to (4.1), (4.2) and (4.3), and there exist vectors $\mathbf{\Lambda}^* = (\lambda_i^*, i = 1, \ldots, m)$ and $\mathbf{\Phi}^* = (\phi_j^*, j = 1, \ldots, p)$ satisfying

$$\nabla_x L(\mathbf{x}^*, \mathbf{\Lambda}^*, \mathbf{\Phi}^*) \equiv \nabla_x f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \phi_j^* \nabla_x h_j(\mathbf{x}^*) = 0, \tag{4.4}$$

$$\lambda_i^* g_i(\mathbf{x}^*) = 0, \qquad i = 1, \ldots, m, \tag{4.5}$$

$$\lambda_i^* \geq 0, \qquad i = 1, \ldots, m \tag{4.6}$$

and for every $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{z} \in Z(\mathbf{x}^*)$ it follows that

$$\mathbf{z}^T \left[\nabla_x^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x^2 g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \phi_j^* \nabla_x^2 h_j(\mathbf{x}^*) \right] \mathbf{z} < 0 \tag{4.7}$$

where

$$\begin{aligned} Z(\mathbf{x}^*) = \{\mathbf{z}\colon\ & \mathbf{z}^T \nabla_x g_i(\mathbf{x}^*) = 0,\ i \in I'(\mathbf{x}^*),\ \mathbf{z}^T \nabla_x g_i(\mathbf{x}^*) \geq 0, \\ & i \in I(\mathbf{x}^*),\ \mathbf{z}^T \nabla_x h_j(\mathbf{x}^*) = 0,\ j = 1, \ldots, p\}, \end{aligned} \tag{4.8}$$

where $I(\mathbf{x}^*)$ is the set of indices for which $g_i(\mathbf{x}^*) = 0$ and $I'(\mathbf{x}^*)$ is the set of indices for which $g_i(\mathbf{x}^*) = 0$ and $\lambda_i^* > 0$, then $\mathbf{x}^*$ is a strict local maximum of $f(\mathbf{x})$.

Note that in general a point that is a local maximum of a problem is not also a global maximum. To prove that a point satisfying the above is in fact a global maximum is often very difficult, if not impossible. In the problem that follows it can be shown that the local maximum presented is also a global maximum.

4.1 The $GI/M/n/n$ queue

If we let the vector $\mathbf{x}$ correspond to $(A_i,\ i = 0, \ldots, n,\ X_i,\ i = 1, \ldots, n)$, then using the above results we find that the problem of finding an upper bound for the time congestion in a $GI/M/n/n$ queue using (2.6), (2.7), (3.1), (3.2) and (3.3) is equivalent to

$$\text{Max } f(\mathbf{x}) = P_n = (-1)^n (aA_n/n)[1 - X_n] \tag{4.9}$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

such that

$$
\begin{aligned}
h_j(\mathbf{x}) \equiv \sum_{i=1}^{n} A_i[X_i - 1](\mathbf{w}_i)_j - A_0(\mathbf{w}_1)_j& \\
- \sum_{i=1}^{n-1} A_i X_i(\mathbf{w}_{i+1})_j - A_n X_n(\hat{\mathbf{w}})_j = 0,& \qquad j = 0, \dots, n, \\
h_{n+1}(\mathbf{x}) \equiv A_0 - 1 = 0,& \\
g_1(\mathbf{x}) \equiv X_2 - 2X_1 + 1 \geq 0,& \\
g_i(\mathbf{x}) \equiv X_{i+1} - 2X_i + X_{i-1} \geq 0,& \qquad i = 2, \dots, n-1, \\
g_n(\mathbf{x}) \equiv 1 - X_1 \geq 0,& \\
g_{i+n-1}(\mathbf{x}) \equiv X_{i-1} - X_i \geq 0,& \qquad i = 2, \dots, n, \\
g_{i+2n-1}(\mathbf{x}) \equiv X_i - \exp(-i/a) \geq 0,& \qquad i = 1, \dots, n
\end{aligned}
\tag{4.10}
$$

and the method of Lagrange multipliers can be used.

5. The solution

THEOREM 2. *The solution to the problem formulated in* (4.9) *and* (14.10) *is*

$$A_i^* = (-1)^i \binom{n}{i}, \qquad i = 0, \dots, n \tag{5.1}$$

and

$$X_i^* = \left(\frac{e^{-n/a} - 1}{n}\right) i + 1, \qquad i = 1, \dots, n \tag{5.2}$$

so

$$P_n^* = (-1)^n (aA_n^*/n)[1 - X_n^*] = (a/n)(1 - e^{-n/a}). \tag{5.3}$$

PROOF. The proposed solution given by (5.1), (5.2) and (5.3) can be shown to satisfy the feasible solution conditions (4.1), (4.2) and (4.3) and also satisfy the necessary optimality conditions (4.4) to (4.8) when we have

$$
\begin{aligned}
&\phi_0^* = -P_n^* \left(1 + \sum_{k=1}^{n-1} \frac{1}{X_k^*} \right), \\
&\phi_j^* = -P_n^* \sum_{k=j}^{n-1} \frac{1}{X_k^*}, \qquad j = 1, \dots, n-1, \\
&\phi_n^* = 0, \qquad \phi_{n+1}^* = -P_n^*, \\
&\lambda_1^* = -\phi_1^*, \\
&\lambda_i^* = -i\phi_1^* + \sum_{j=1}^{i-1} A_j^*(i-j) \left(\sum_{k=0}^{j} \binom{j}{k} (-1)^k \phi_{k+1}^* \right), \qquad i = 2, \dots, n-1, \\
&\lambda_n^* = 0, \qquad \lambda_{i+n-1}^* = 0, \qquad i = 2, \dots, n \\
&\lambda_{i+2n-1}^* = 0, \qquad i = 1, \dots, n-1, \qquad \lambda_{2n-1}^* = (-1)^n \frac{aA_n^*}{n}.
\end{aligned}
\tag{5.4}
$$

The proof of this result is achieved by showing that (4.2) to (4.8) are satisfied by (5.4) when (4.9) defines the objective and (4.10) define the constraints. The algebra to show this result for the general case is messy and is given in Coyle [2]. A proof of this result when there are 2 servers is given here; the case when there is only one server is relatively simple.

5.1 The case $n = 2$

The vector of variables for this optimisation problem, $\mathbf{x}$, is $(A_0, A_1, A_2, X_1, X_2)$, the eigenvalues of $Q_1 C^{-1}$ are $-\alpha_0 = 0, -\alpha_1 = -1/a, -\alpha_2 = -2/a$ and the respective eigenvectors are $\mathbf{w}_0 = (1, 0, 0), \mathbf{w}_1 = (1, -1, 0)$ and $\mathbf{w}_2 = (1, -2, 1)$, also $\mathbf{w}_2 Q_2 = -a\hat{\mathbf{w}} = -a(1, -3, 2)$. The optimisation problem as formulated by (4.9) and (4.10) can be written as

$$\text{Max } f(\mathbf{x}) = P_2 = a/2A_2(1 - X_2) \tag{5.5}$$

such that

$$\begin{aligned}&h_0(\mathbf{x}) \equiv -A_0 - A_1 - A_2 = 0, \quad h_1(\mathbf{x}) \equiv A_0 + A_1 + 2A_2 + A_1X_1 + A_2X_2 = 0,\\ &h_2(\mathbf{x}) \equiv -A_2 - A_1X_1 - A_2X_2 = 0, \quad h_3(\mathbf{x}) \equiv A_0 - 1 = 0,\\ &g_1(\mathbf{x}) \equiv X_2 - 2X_1 + 1 \geq 0, \quad g_2(\mathbf{x}) \equiv 1 - X_1 \geq 0, \quad g_3(\mathbf{x}) \equiv X_1 - X_2 \geq 0,\\ &g_4(\mathbf{x}) \equiv X_1 - e^{-1/a} \geq 0 \quad g_5(\mathbf{x}) \equiv X_2 - e^{-2/a} \geq 0.\end{aligned} \tag{5.6}$$

The proposed solution to this problem given by (5.1) and (5.2) is

$$A_0^* = 1, \quad A_1^* = -2, \quad A_2^* = 1, \quad X_1^* = (e^{-2/a} + 1)/2, \quad X_2^* = e^{-2/a},$$

and so

$$P_2^* = (a/2)(1 - e^{-2/a}). \tag{5.7}$$

The solution given by (5.7) satisfies all the constraints in (5.6). The Lagrangian given by (4.4) for this case is equivalent to

$$\begin{aligned}\partial L/\partial A_0 &= -\phi_0 + \phi_1 + \phi_3 = 0,\\ \partial L/\partial A_1 &= -\phi_0 + \phi_1(X_1 + 1) - \phi_2 X_1 = 0,\\ \partial L/\partial A_2 &= -\phi_0 + \phi_1(X_2 + 2) - \phi_2(X_2 + 1) + \frac{a}{2}(1 - X_2) = 0,\\ \partial L/\partial X_1 &= A_1\phi_1 - A_1\phi_2 - 2\lambda_1 - \lambda_2 + \lambda_3 + \lambda_4 = 0\\ \partial L/\partial X_2 &= A_2\phi_1 - A_2\phi_2 + \lambda_1 - \lambda_3 + \lambda_5 - a\frac{A_2}{2} = 0.\end{aligned} \tag{5.8}$$

These equations are satisfied when we choose

$$\begin{aligned}&\phi_0^* = -P_2^*(1 + (1/X_1^*)), \quad \phi_1^* = -P_2^*(1/X_1^*), \quad \phi_2^* = 0, \quad \phi_3^* = -P_2^*;\\ &\lambda_1^* = -\phi_1^*, \quad \lambda_2^* = 0, \quad \lambda_3^* = 0, \quad \lambda_4^* = 0 \quad \lambda_5^* = a(A_2^*/2)\end{aligned} \tag{5.9}$$

as proposed by (5.4). These values also satisfy (4.5) and (4.6). To show that (4.7) is satisfied the vectors $\mathbf{z} \in Z(\mathbf{x}^*)$ must be found. These are given by (4.8). The components of $\mathbf{z}$ correspond to the variables in this problem so let $\mathbf{z}^T = (z_{A_0}, z_{A_1}, z_{A_2}, z_{X_1}, z_{X_2})$. The vector $\mathbf{z}$ must firstly satisfy $\mathbf{z}^T \nabla_x g_i(\mathbf{x}^*) = 0$ for the inequality constraints where $g_i(\mathbf{x}^*) = 0$. The inequality constraints g_1 and g_5 are equal to zero so

$$\mathbf{z}^T \nabla_x g_1(\mathbf{x}^*) = \mathbf{z}^T(0, 0, 0, -2, 1)^T = 0 \quad \Rightarrow -2z_{X_1} + z_{X_2} = 0$$

and

$$\mathbf{z}^T \nabla_x g_5(\mathbf{x}^*) = \mathbf{z}^T(0, 0, 0, 0, 1)^T = 0 \quad \Rightarrow z_{X_2} = 0$$

and so $z_{X_1} = z_{X_2} = 0$. It is also required that $\mathbf{z}^T \nabla_x h_j(\mathbf{x}^*) = 0$, $j = 0, \ldots, 3$ and so

$$\begin{aligned}\mathbf{z}^T \nabla_x h_0(\mathbf{x}^*) &= \mathbf{z}^T(-1, -1, -1, 0, 0)^T = 0,\\ \mathbf{z}^T \nabla_x h_1(\mathbf{x}^*) &= \mathbf{z}^T(1, 1 + X_1, 2 + X_2, A_1, A_2)^T = 0,\\ \mathbf{z}^T \nabla_x h_2(\mathbf{x}^*) &= \mathbf{z}^T(0, -X_1, -(1 + X_2), -A_1, -A_2)^T = 0\end{aligned}$$

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

and

$$\mathbf{z}^T \nabla_x h_3(\mathbf{x}^*) = \mathbf{z}^T (1,0,0,0,0)^T = 0.$$

For all of these to be satisfied it is necessary for $z_{A_0} = z_{A_1} = z_{A_2} = 0$. Therefore the only possible vector $\mathbf{z} \in Z(\mathbf{x}^*)$ is the zero vector and (4.7) is trivially satisfied. The necessary conditions (4.2) through to (4.8) have been shown to be satisfied by the proposed solution which is therefore a strict local maximum of $f(\mathbf{x})$. So P_n^* is a local maximum for the time congestion in the $GI/M/n/n$ queue when $n = 2$.

To prove that this local maximum is also a global maximum it must be shown that no other local maximum exists that is larger than the presented result. The equality constraints h_0, h_1, h_2 and h_3 can be reduced to the single constraint $h(\mathbf{x}) \equiv -A_2[1+X_2-X_1]+X_1 = 0$. The problem is now in the three variables X_1, X_2, and A_2. Rearranging the above gives $A_2 = X_1(1 - X_1 + X_2)$ and so for any feasible solution A_2 must be positive.

The partial derivatives of the Lagrangian are given by

$$\begin{aligned}
\partial L/\partial A_2 &= \phi(1 - X_1 + X_2) + \frac{a}{2}(1 - X_2) = 0, \\
\partial L/\partial X_1 &= -\phi(A_2 + 1) - 2\lambda_1 - \lambda_2 + \lambda_3 + \lambda_4 = 0, \\
\partial L/\partial X_2 &= \phi A_2 + \lambda_1 - \lambda_3 + \lambda_5 - \frac{a}{2}A_2 = 0.
\end{aligned}$$

Rearranging the first of these partial derivatives gives

$$\phi = -a(1 - X_2)/[2(1 - X_1 + X_2)]$$

and so ϕ must be negative or zero and so it is found from the other two partial derivatives that

$$2\lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 \geq 0 \tag{5.11}$$

and

$$\lambda_1 - \lambda_3 + \lambda_5 > 0. \tag{5.12}$$

Any local maximum, and so the global maximum, of the optimisation problem must satisfy both the above two equations and the inequalities of (4.10). Note that from (4.6) the λs must be positive or zero and from (4.5) an inequality constraint and its corresponding Lagrange Multiplier cannot both be nonzero.

If (5.11) is zero then $\phi = 0$, $X_2 = 1$, $X_1 = 1$ and $P_2 = 0$. Whilst this is a feasible solution it is obviously not the global maximum of the problem.

Equation (5.11) must be positive and so either λ_1 or λ_2 or both must be positive and so $X_1 = 1$ or $X_2 - 2X_1 + 1 = 0$ or both. The case when $X_1 = 1$ has already been shown to be infeasible so $X_2 - 2X_1 + 1 = 0$ and $X_1 < 1$. Since $X_1 < 1$ it follows that $\lambda_2 = 0$. Using this and manipulating (5.11) and (5.12) it is found that $-3\lambda_3 - \lambda_4 + 2\lambda_5 > 0$ and so λ_5 must also be positive

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

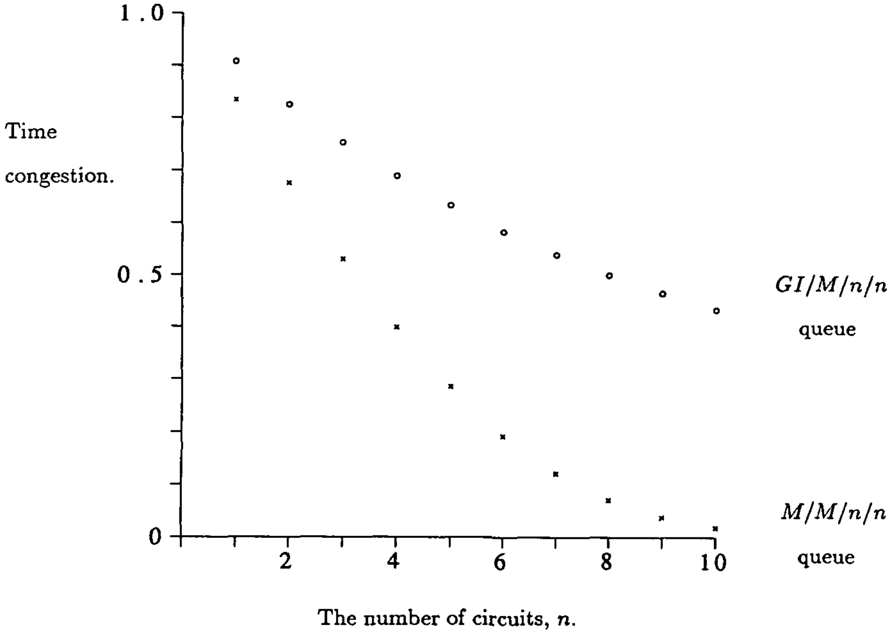

FIGURE 1. A comparison of the maximum time congestion in the $GI/M/n/n$ queue with the time congestion in the $M/M/n/n$ queue when the arrival rate is 5.0

which means that $X_2 = \exp(-2/a)$. The only feasible solution is therefore $X_1 = (1+\exp(-2/a))/2$, $A_2 = 1$ and so P_2^* is the same as the solution given by (5.3). It has been shown that for the case when $n = 2$ the value of the time congestion in a $GI/M/n/n$ queue given by (5.3) is the maximum possible value.

Work by Kuczura [5] has shown that the call congestion in a $GI/M/1/1$ system is minimised when the call arrival distribution is deterministic and conjectured this was also true for cases when there was more than one server. For the single server case, (5.2) gives $X_1^* = \exp(-1/a)$. It is shown in Pearce [6] that this implies that the interarrival lifetime is deterministic. Thus it has been shown that the time congestion is maximised for the single server case when there is a deterministic interarrival distribution; however for situations where there is more than one server this is no longer true.

An obvious lower bound for the time congestion in the $GI/M/n/n$ queue is 0 for all values of a and n. The results that are obtained using the above method for a lower bound of the $GI/M/n/n$ correspond to this lower bound of 0. Unlike the upper bound the lower bound is achieved by a known

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

distribution given by

$$G(\cdot) = \begin{cases} 1-\varepsilon, & 0 \le y \le 1/\varepsilon, \\ 1, & y > 1/\varepsilon \end{cases}$$

for the case where $\varepsilon \to 0$. This distribution is known to be achieved when $X_i = 1$ for all i (see Taylor and Coyle [8]). These two results give bounds on the sensitivity of the time congestion in the $GI/M/n/n$ system.

A comparison between the maximum time congestion in the $GI/M/n/n$ queue and the time congestion in a $M/M/n/n$ queue is shown in Figure 1. The maximum possible time congestion in the $GI/M/n/n$ queue is considerably higher than that of the $M/M/n/n$ queue for certain values of n.

6. Conclusion

In this paper a method for determining sensitivity bounds for GSMPs in which there is only one general distribution has been presented. Using this method sensitivity bounds for the time congestion in a $GI/M/n/n$ system have been produced which are quickly calculated and easy to use. For this system an analytic result has been produced, although the full proof is long and complicated. For other systems numerical results have been calculated which are much easier to obtain. These numerical results often lead to interesting analytic results as was true in the case presented here. A few examples of systems investigated using the above method can be found in [8]. In general the main problem with the method at present is that there is still a certain degree of uncertainty as to whether the results obtained are global maxima and minima for the problem presented or just local maxima and minima. As with many problems of this type it may be impossible to ever prove whether a local optimal point is also a global optimal point.

References

[1] M.Avriel, *Nonlinear programming: analysis and methods* (Prentice-Hall, New Jersey, 1976).

[2] A. J. Coyle, "An upper bound on the time congestion in a $GI/M/n/n$ system", to appear.

[3] W. Feller, *An introduction to probability theory and its applications* Vol. II (John Wiley, New York, 1966).

[4] D. König and U. Jansen, "Stochastic processes and properties of invariance for queueing systems and speeds and temporary interruption" *Trans. 7th Prague Conference Inf. Th., Stat. Dec. Fns. and Rand. Proc.* (1974) 335–343.

[5] A. Kuczura, "Queues with mixed renewal and Poisson inputs" *Bell System Tech. J.* **51** (1972) 1305–1326.

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press

[6] C. E. M. Pearce, "On the peakedness of primary and secondary processes" *Australian Telecommunications Research* **12**, 2 (1978) 18–24.

[7] P. G. Taylor, "Aspects of insensitivity in stochastic processes", Ph.D. Thesis, The University of Adelaide, 1987.

[8] P. G. Taylor and A. J. Coyle, "Bounds on the sensitivity of generalised semi-Markov processes with a single generally distributed lifetime", submitted.

https://doi.org/10.1017/S033427000000655X Published online by Cambridge University Press