An Algorithm for Estimating the Expected Number of Customers for a Class of Markovian Queueing Systems

Hung-Yuan Tu, Hillel Kumin

School of Industrial and Systems Engineering, University of Oklahoma, Norman, USA

Email: hkumin@ou.edu

Abstract

An algorithm is presented for estimating the expected number of customers for a class of Markovian queueing systems. The class is characterized by those systems whose transition matrix for the underlying customer arrival and departure process is finite, irreducible, and aperiodic. The algorithm does not depend on a closed-form solution for the limiting behavior of the queue. The expected number of customers is frequently used as a measure of effectiveness to describe the behavior of the system or to optimize its design or control. To calculate such a quantity one must usually obtain a closed-form expression for the steady-state probabilities. Unfortunately, of the myriad of Markovian queueing systems, only a few have known closed-form expressions for their steady-state probabilities. The most well-known, using Kendall's notation, are the M/M/1 and the M/M/c system. The algorithm described below estimates the expected number in the system under steady-state without a need for closed form steady-state probabilities. All that is needed is the transition matrix for the underlying Markov chain.

Keywords

Queueing Systems, Markov Chain, Expected Number in System

1. Preliminaries

Consider a Markovian queueing system in which the customer arrival and departure process is described by a finite, irreducible, and aperiodic Markov Chain $\{X_n : n = 0,1,\cdots\}$ whose $N \times N$ transition matrix is given by $A = (a_{ij})$, where $a_{ij} = pr\{X_n = j | X_{n+1} = i\}$.

It is well known that for a finite, irreducible, aperiodic Markov chain, there

exists a unique stationary probability vector $V = (v_0, \cdots, v_0)$ such that $VA = V$ [1].

Let L = Expected number of customers in the steady state system. The expected number of customers is frequently used as a measure of effectiveness to describe the behavior of the system or to optimize its design or control [2] [3] [4], and [5]. Now define the following

$$F = (0, 1, \cdots, N-1)^t$$

$$A^Z F = \left(w_0^{(z)}, \cdots, w_{N-1}^{(z)} \right)^t = W^{(z)}$$

$$\underline{w}^{(z)} = \min_i \left\{ w_i^{(z)} : i = 0, 1, \cdots, N-1 \right\}$$

$$\overline{w}^{(z)} = \max_i \left\{ w_i^{(z)} : i = 0, 1, \cdots, N-1 \right\}$$

$$\hat{L}^{(z)} = \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2}$$

<u>Theorem 1</u> For any positive integer z,

1)

$$\left| L - \hat{L}^{(z)} \right| \leq \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2}$$

2)

$$\lim_{z \to \infty} \hat{L}^{(z)} = L$$

Proof.

1)

$$L = VF = VA^Z F = \sum v_i w_i^{(z)}$$

$$\text{since} \quad w_i^{(z)} \leq \overline{w}^{(z)}$$

then,

$$\sum v_i w_i^{(z)} \leq \sum v_i \overline{w}^{(z)} = \overline{w}^{(z)} \sum v_i = \overline{w}^{(z)}$$

Similarly, since

$$w_i^{(z)} \geq \overline{w}^{(z)}$$

we have,

$$\sum v_i w_i^{(z)} \geq \sum v_i \underline{w}^{(z)} = \underline{w}^{(z)} \sum v_i = \underline{w}^{(z)}$$

thus

$$L - \hat{L}^{(z)} \leq \overline{w}^{(z)} - \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2} = \frac{\left(\overline{w}^{(z)} - \underline{w}^{(z)} \right)}{2}$$

and

$$L - \hat{L}^{(z)} \geq \overline{w}^{(z)} - \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2} = \frac{-\left(\overline{w}^{(z)} - \underline{w}^{(z)} \right)}{2}$$

It follows that

$$\left| L - \hat{L}^{(z)} \right| \leq \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2}$$

2) Let $A^z = \left(p_0^{(z)}, \cdots, p_{N-1}^{(z)} \right)^t$, where $p_i^{(z)}$ is the $(i+1)$ th row of A^z. Since we are concerned with finite, irreducible, aperiodic Markov chains, we have

$$\lim_{z \to \infty} p_i^{(z)} = V$$

For all i. It follows that

$$\lim_{n \to \infty} w_i^{(z)} = \lim_{z \to \infty} p_i^{(z)} F = VF = L$$

For all i. Let (s_n) be the sequence formed by combining the N sequences $\left(w_0^{(z)} \right), \cdots, \left(w_{N-1}^{(z)} \right)$ according to the ascendant order of z. Clearly,

$$\lim_{z \to \infty} s_n = L$$

Since $\left(\overline{w}^{(z)} \right)$ and $\left(\underline{w}^{(z)} \right)$ are both subsequence of (s_n), it is clear that

$$\lim_{z \to \infty} \overline{w}^{(z)} = \lim_{z \to \infty} \underline{w}^{(z)} = L$$

It therefore follows that

$$\lim_{z \to \infty} \hat{L}^{(z)} = \lim_{z \to \infty} \frac{\left(\overline{w}^{(z)} + \underline{w}^{(z)} \right)}{2} = \frac{\left(\lim_{z \to \infty} \overline{w}^{(z)} + \lim_{z \to \infty} \underline{w}^{(z)} \right)}{2} = \frac{L+L}{2} = L.$$

2. Algorithm for Estimating the Expected Number of Customers

The relationship between $\hat{L}$ and L as stated in Theorem 1 can be used to develop an iterative algorithm for estimating the expected number of customers from the system's transition matrix.

Definition 1

The absolute error of an estimation is the absolute value of the difference between the estimation and the true value. When the true value is unknown, the largest absolute error that may occur is called the maximal absolute error.

Definition 2

The relative error of an estimation is the ratio of the absolute error of the estimation to the true value. When the true value is unknown, the largest relative error that may occur is called the maximal relative error.

The algorithm can now be stated as follows:

Step 1. Determine the allowable maximal absolute error a.

Step 2. Set $z = 0$ and $W^{(0)} = (0, \cdots, N-l)^t$.

Step 3. Compute $W^{(z+1)} = AW^{(z)}$.

Step 4. If $\dfrac{\left(\overline{w}^{(z+1)} - \underline{w}^{(z+1)} \right)}{2} \le a$, go to step 5; otherwise, increase z by 1 then go to step 3.

Step 5. The desired accuracy has been reached. Let $L = \hat{L}^{(z+1)} = \dfrac{\left(\overline{w}^{(z+1)} + \underline{w}^{(z+1)} \right)}{2}$.

Terminate

Theorem 2 Let $\hat{L}^{(n)}$ be the estimation of the expected number of customers obtained from the Algorithm above. Using an allowable maximal absolute error

of a, we then have that

1) The maximal absolute error of $\hat{L}^{(n)}$ is a,

2) The maximal relative of $\hat{L}^{(n)}$ is $\dfrac{a}{\left(\hat{L}^{(n)}-a\right)}$

Proof. 1) Step 4 of the algorithm implies $\dfrac{\left(\overline{w}^{(n)}-\underline{w}^{(n)}\right)}{2} \leq a$. Thus, by Theorem 1, $\left|L-\hat{L}^{(n)}\right| \leq \dfrac{\left(\overline{w}^{(n)}-\underline{w}^{(n)}\right)}{2} \leq a$. Hence, the maximal absolute error of $L^{(n)}$ is a.

2) The maximal relative error occurs when $L = \hat{L}^{(n)} - a$. Thus, the maximal relative error $= \dfrac{\left|\hat{L}^{(n)}-L\right|}{L} = \dfrac{\left|\hat{L}^{(n)}-\left(\hat{L}^{(n)}-a\right)\right|}{\left(\hat{L}^{(n)}-a\right)} = \dfrac{a}{\left(\hat{L}^{(n)}-a\right)}$

The allowable error associated with the algorithm above is specified in terms of the maximal absolute error, and is thus independent of the magnitude of the expected number of customers. If one needs to associate the error of the estimation with the magnitude of the expected number of customers, then the algorithm may be modified as follows:

Step 1. Determine the allowable maximal relative error r.

Step 2. Set $z = 0$ and $W^{(0)} = \left(0, \cdots, N-l\right)^{t}$

Step 3. Compute $W^{(z+1)} = AW^{(z)}$

Step 4. If $\dfrac{\left(\overline{w}^{(z+1)}-\underline{w}^{(z+1)}\right)}{2} \leq \dfrac{r\hat{L}^{(z+1)}}{\left(1+r\right)}$, go to step 5; otherwise, increase z by 1 then go to step 3.

Step 5. The desired accuracy has been reached. Let $L = \hat{L}^{(z+1)} = \dfrac{\left(\overline{w}^{(z+1)}+\underline{w}^{(z+1)}\right)}{2}$. Terminate.

<u>Theorem 3</u> let $\hat{L}^{(n)}$ be the estimation of the expected number of customers obtained from the modified algorithm above, using an allowable maximal relative error of r, then

1) The maximal absolute error of the $\hat{L}^{(n)}$ is $\dfrac{r\hat{L}^{(n)}}{\left(1+r\right)}$

2) The maximal relative error of $\hat{L}^{(n)}$ is r

Proof. 1) Step 4 of the modified algorithm 2 implies $\dfrac{\left(\overline{w}^{(n)}-\underline{w}^{(n)}\right)}{2} \leq \dfrac{r\hat{L}^{(n)}}{\left(1+r\right)}$.

Thus, by Theorem 1, $\left|L-\hat{L}^{(n)}\right| \leq \dfrac{\left(\overline{w}^{(n)}-\underline{w}^{(n)}\right)}{2} \leq \dfrac{r\hat{L}^{(n)}}{\left(1+r\right)}$. Hence the maximal absolute error of $\hat{L}^{(n)}$ is $\dfrac{r\hat{L}^{(n)}}{\left(1+r\right)}$.

2) The maximal relative error occurs when $L = \hat{L}^{(n)} - \dfrac{r\hat{L}^{(n)}}{\left(1+r\right)}$. Thus the max-

imal relative error $= \dfrac{\left|\hat{L}^{(n)} - L\right|}{L} = \dfrac{\left|\hat{L}^{(n)} - \left(\hat{L}^{(n)} - \dfrac{r\hat{L}^{(n)}}{(1+r)}\right)\right|}{\left(\hat{L}^{(n)} - \dfrac{r\hat{L}^{(n)}}{(1+r)}\right)} = r$

Thus, the initial algorithm will have a maximal absolute error of a, and a maximal relative error of $\dfrac{a}{\left(\hat{L}^{(n)} - a\right)}$. The modified algorithm will have a maximal absolute error of $\dfrac{r\hat{L}^{(n)}}{(1+r)}$ and a maximal relative error of r.

3. Numerical Example (Table 1)

Consider the M/M/c/c queueing system with finite waiting space. Assume that customers arrive according to a Poisson process with rate λ, and that service times are exponential with mean $1/\mu$. Although the process that describe the number of customers in the system at time is a process that take place in continues time, several relevant Markov chains may be defined for this system in discrete time. For example, consider the Markov chain whose transition is given by

$$A = a_{ij} = \begin{bmatrix} 1-\lambda & \lambda & & & & & \\ \mu & 1-\lambda-\mu & \lambda & & & & \\ & 2\mu & 1-\lambda-2\mu & \lambda & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ & & & & (c-1)\mu & 1-\lambda-(c-1)\mu & \lambda \\ & & & & & c\mu & 1-c\mu \end{bmatrix}$$

where λ and μ have been scaled such that A is a legitimate transition matrix, a is set at 0.01, and L is computed using the following formulas for the M/M/c/c (Erlang B) system [6].

Let

$$B(c,r) = p_c = \dfrac{\dfrac{r^c}{c!}}{\sum_{i=0}^{c} \dfrac{r^i}{i!}}$$

where $r = \dfrac{\lambda}{\mu}$ and let

$$C(c,r) = \dfrac{cB(c,r)}{c-r+rB(c,r)}$$

then $L_q = \dfrac{r}{c-r} C(c,r)$ and $L = L_q + r(1-p_c)$ where L_q is the expected number in the queue.

Table 1. M/M/c/c system.

Arrival rate	No. of servers	Service rate	Expected number of customers	
			L	Algorithm
0.2	3	0.1	2.4678	2.4694
0.2	4	0.1	1.9834	1.9842
0.2	5	0.1	1.9664	1.9683
0.2	3	0.08	5.3058	5.3076
0.2	4	0.08	2.6583	2.6596
0.2	5	0.08	2.4560	2.4579
0.15	2	0.1	2.9630	2.9644
0.15	3	0.1	1.5353	1.5373
0.15	4	0.1	1.4728	1.4741
0.3	4	0.1	3.9100	3.9115
0.3	5	0.1	3.0241	3.0253
0.3	6	0.1	2.9427	2.9442

These computations were made on a 64-bit Intel® Core™ m3-6Y30 CPU @ 0.90 GHZ laptop computer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cinlar, E. (1975) Introduction to Stochastic Processes. Prentice-Hall, New Jersey.

[2] Brigham, G. (1955) On a Congestion Problem in an Aircraft Factory. *Operations Research*, **3**, 412-428. https://doi.org/10.1287/opre.3.4.412

[3] Kumin, H. (1973) On Characterizing the Minimum of a Function of Two Variables, One of Which Is Discrete. *Management Science*, **20**, 126-129. https://doi.org/10.1287/mnsc.20.1.126

[4] Smith, J.M. (2004) Optimal Design and Performance Modelling of M/G/1/K Queueing Systems. *Mathematical and Computer Modelling*, **39**, 1049-1081. https://doi.org/10.1016/S0895-7177(04)90534-1

[5] Stidham, S. (2009) Optimal Design of Queueing Systems. CRC Press, New York. https://doi.org/10.1201/9781420010008

[6] Gross, D., Shortle, J., Thompson, H. and Harris, C. (2008) Fundamentals of Queueing Theory. John Wiley & Sons, New York. https://doi.org/10.1002/9781118625651