Acta Informatica 17, 43–61 (1982)

Acta Informatica

Information Theoretic Approximations for $M/G/1$ and $G/G/1$ Queuing Systems

John E. Shore

Naval Research Laboratory, Code 7591, Washington, D.C. 20375, USA

Summary. This paper presents new results concerning the use of information theoretic inference techniques in system modeling and concerning the widespread applicability of certain simple queuing theory formulas. For the case when an $M/G/1$ queue provides a reasonable system model but when information about the service time probability density is limited to knowledge of a few moments, entropy maximization and cross-entropy minimization are used to derive information theoretic approximations for various performance distributions such as queue length, waiting time, residence time, busy period, etc. Some of these approximations are shown to reduce to exact $M/M/1$ results when $G=M$. For the case when a $G/G/1$ queue provides a reasonable system model, but when information about the arrival and service distributions is limited to the average arrival and service rates, it is shown that various well known $M/M/1$ formulas are information theoretic approximations. These results not only provide a new method for approximating the performance distributions, but they help to explain the widespread applicability of the $M/M/1$ formulas.

I. Introduction

Performance modeling and analysis of computer systems have been important computer science problems for many years. Although queuing theory has provided the basis for remarkable success in solving these problems [1–3], this success has been somewhat puzzling because it is clear that computer systems often do not satisfy assumptions made by the stochastic process models that are used; it appears that queuing theory equations have wider applicability than is suggested by their classical derivations. One possible explanation for this is given by operational analysis [4]. Another is based on information-theoretic modeling techniques that exploit the principles of maximum entropy and minimum cross-entropy (a generalization) [5]. These principles provide methods for estimating probability distributions given information in the form of

0001-5903/82/0017/0043/$03.80

known expected values. The methods can be applied to system modeling because expected values of various distributions of interest are often known in terms of moments of the arrival and service time distributions. Because entropy maximization has been shown to be a uniquely correct, self-consistent method of inference about probability distributions [5], we refer to the resulting estimates of the performance distributions as information-theoretic approximations. The use of maximum entropy in system modeling problems has been studied by Ferdinand [6], Beneš [7], and Shore [8]. Recently, Bard [9] used entropy maximization in modeling an IBM System/370 I/0 subsystem.

In this paper, we present new results concerning the use of cross-entropy minimization in system modeling and concerning the applicability of certain simple queuing theory formulas. One set of results applies when an $M/G/1$ queue provides a reasonable system model but when information about the service time probability density is limited to knowledge of a few moments. We show how cross-entropy minimization yields information theoretic approximations for various "performance distributions" such as queue length, busy period length, number served during a busy period, waiting time, etc. We also show how some of these approximations reduce to exact $M/M/1$ results when $G=M$. A second set of results applies when a $G/G/1$ queue provides a reasonable system model but when information about the arrival and service distributions is limited to the average arrival and service rates. We show how the well known $M/M/1$ formulas are information theoretic approximations for the $G/G/1$ system given this limited information. That is, the $M/M/1$ formulas are the best hypotheses about the $G/G/1$ systems given only the mean arrival and service rates. This fact has nothing at all to do with the various assumptions that must be debated when considering the applicability of stochastic models, and it helps to explain why the $M/M/1$ formulas have been found to be so useful.

Section II of this paper summarizes the principles of maximum entropy and minimum cross-entropy, and discusses informally the sense in which these principles provide correct, general methods of inductive inference. Information theoretic approximations for $M/G/1$ performance distributions are discussed, with examples, in Sects. III–IV. In these applications we assume uniform distributions for estimates of the performance distributions available prior to learning the service time moments. Applications involving the use of non-uniform prior estimates are suggested in Sect. VII. Results for $G/G/1$ systems are derived in Sect. VIII. Discussion follows in Sect. IX.

II. Cross-Entropy Minimization and Entropy Maximization

A. General Statement of the Problem and the Minimum Cross-Entropy Solution

Let $\mathbf{x}$ denote a single state of some system that has a set $\mathbf{D}$ of possible system states and a probability density $q^{\dagger}(\mathbf{x})$ of states. Let $\mathscr{D}$ be the set of all probability densities q on $\mathbf{D}$ such that $q(\mathbf{x}) \geqq 0$ for $\mathbf{x} \in \mathbf{D}$ and

$$\int_{\mathbf{D}} d\mathbf{x}\, q(\mathbf{x}) = 1. \tag{1}$$

We assume that the existence of $q^\dagger \in \mathscr{D}$ is known but that $q^\dagger$ itself is unknown. The density $q^\dagger$ is sometimes known as a "true" density.

Suppose $p \in \mathscr{D}$ is a *prior* density that is our current estimate of $q^\dagger$, and suppose we gain new information about $q^\dagger$ in the form of a set of expected values

$$\int_{\mathbf{D}} d\mathbf{x}\, q^\dagger(\mathbf{x})\, g_r(\mathbf{x}) = \bar{g}_r, \tag{2}$$

for a known set of functions $g_r(\mathbf{x})$ and numbers $\bar{g}_r$, $r = 0, \ldots, M$. Now, because the *constraints* (1)–(2) do not determine $q^\dagger$ completely, they are satisfied not only by $q^\dagger$ but by some subset of densities $\mathscr{I} \subseteq \mathscr{D}$. Which single density should we choose from this subset to be our new estimate of $q^\dagger$, and how should we use the prior p and the new information (2) in making this choice?

The principle of minimum cross-entropy provides a general solution to this inference problem [5]. The principle states that, of all the distributions that satisfy the constraints, you should choose the posterior q with the least cross-entropy

$$H(q, p) = \int_{\mathbf{D}} d\mathbf{x}\, q(\mathbf{x}) \log(q(\mathbf{x})/p(\mathbf{x})) \tag{3}$$

with respect to the prior p. That is, the *posterior* density q satisfies

$$H(q, p) = \min_{q' \in \mathscr{I}} H(q', p),$$

where $\mathscr{I} \subseteq \mathscr{D}$ comprises all of the densities that satisfy the constraints (2). The principle of maximum entropy [10, 11] is a special case of cross-entropy minimization when the prior $p(\mathbf{x})$ is uniform on $\mathbf{D}$ [5].

Given a positive prior probability density p, if there exists a posterior that minimizes the cross-entropy (3) and satisfies the constraints (1)–(2), then it has the form

$$q(\mathbf{x}) = p(\mathbf{x}) \exp\left(-\lambda - \sum_{k=0}^{M} \beta_k\, g_k(\mathbf{x})\right), \tag{4}$$

with the possible exception of a set of points on which the constraints imply that q vanishes [12, p. 38]–[14]. In (4), λ and β_k are Lagrangian multipliers whose values are determined by the constraints (1) and (2), respectively. Conversely, if one can find values for λ and β_k in (4) such that the constraints are satisfied, then the solution exists and is given by (4). Conditions for the existence of solutions are discussed by Csiszàr [13]. Now, the normalization constraint (1) requires that

$$\lambda = \log \int_{\mathbf{D}} d\mathbf{x}\, p(\mathbf{x}) \exp\left(-\sum_{k=0}^{M} \beta_k g_k(\mathbf{x})\right). \tag{5}$$

If the integral in (5) can be performed, one can sometimes find values for the β_k from the relations

$$-\frac{\partial \lambda}{\partial \beta_k} = \bar{g}_k. \tag{6}$$

It unfortunately is usually impossible to solve (6) for the β_k explicitly, in order to obtain a closed-form solution expressed directly in terms of the known expected values $\bar{g}_k$ rather than in terms of the Lagrangian multipliers. Computational methods for finding approximate solutions are, however, available ([14, Appendix A], [15]).

When **D** is a discrete state space, the integrals in (1)–(3) and (5) are replaced by sums in the usual way. Solutions for maximum entropy are the same as (4)–(6) with the prior deleted.

As a general method of statistical inference, cross-entropy minimization was first introduced by Kullback [12] and has been advocated in various forms by others [16]–[19]. The name cross-entropy is due to Good [16]. Other names include expected weight of evidence [20, p. 72], directed divergence [12, p. 7], discrimination information [12, p. 37], and relative entropy [21, p. 19]. There is a substantial history of applications of cross-entropy minimization in various fields [5]. Recent successful applications include spectral analysis [22], speech coding [23], and pattern classification [24]. General properties of cross-entropy minimization are discussed by Shore and Johnson [14].

B. Justification of Cross-Entropy Minimization

One could imagine using a procedure that chooses the posterior estimate of $q^\dagger$ by minimizing some function of q and p other than $H(q,p)$. In what sense does minimizing cross entropy yield the best estimate q of $q^\dagger$? One answer to this question is provided by recent work of Shore and Johnson [5] that characterizes cross-entropy minimization as an inference procedure by means of certain consistency axioms. In describing this work, it is usful to view an inference procedure as an operator $\circ$ that takes two arguments, a prior probability density p and new constraint information I of the form (2), and yields a posterior probability density $p \circ I$. It is assumed in [5] that $\circ$ is implemented by minimization of some well behaved functional $H'(q,p)$ – that is, that $q = p \circ I$ is defined as that density, among all the densities that satisfy the constraints I, for which $H'(q,p)$ is least. It is further assumed that the operator $\circ$ satisfies consistency axioms that, informally, require different ways of taking information I into account (for example, in different coordinate systems) to lead to equivalent results. It is then shown to follow from the assumptions that $p \circ I$ equals the result of minimizing the cross entropy $H(q,p)$. The axioms do not imply that H' must be H – for instance a monotonic function of H would do just as well – but they do uniquely characterize the result $p \circ I$ of the minimization: cross-entropy minimization is uniquely correct in the sense that minimization of any other functional either gives the same result or leads to a contradiction with one of the axioms.

Other justifications for the use of cross-entropy minimization can be based on cross entropy's properties as an information measure [12, 13–15, 19, 25]. For instance, $H(q,p)$, informally speaking, measures the distortion, "information dissimilarity," or "information divergence" of q from p. $H(q,p)$ can be

interpreted as the amount of information needed to change a prior p into the posterior q or to determine q given p [14]; indeed,

$$H(q, p) = H(q^\dagger, p) - H(q^\dagger, q) \tag{7}$$

holds when $q = p \circ I$ is defined by cross-entropy minimization [13, 14]. In these terms the minimum-cross-entropy principle is justified intuitively as the choice of posterior q that introduces the least distortion, least additional information, or fewest unjustified assumtpions consistent with the given constraints. From (7) it follows that $H(q^\dagger, q) \leqq H(q^\dagger, p)$. Thus the posterior q is closer to $q^\dagger$ in the cross-entropy sense than is the prior p.

Yet another justification for cross-entropy minimization is provided by the "expectation-matching" property [14], which states that for an arbitrary fixed density q^* and densities q of the form (4), $H(q^*, q)$ is least when the expectations of q match those of q^*. In particular, it follows that $q = p \circ I$ is not only the density satisfying (2) that minimizes $H(q, p)$, but also the density of the form (4) that minimizes $H(q^\dagger, q)$. Hence $p \circ I$ is not only closer to $q^\dagger$ than is p, but it is the closest possible density of the form (4). The expectation-matching property is a generalization of a property of orthogonal polynomials [26, p. 12] that, in the case of speech analysis [27], is well-known as the "correlation-matching property" [28, Chap. 2].

III. $M/G/1$ Queue Length Distribution

In Sects. III–VI, we consider $M/G/1$ systems: customers arrive with independent, exponentially distributed interarrival times from an infinite customer pool, wait in an infinite capacity queue, are served independently by a single server with a general service time distribution, and return to the customer pool. The performance of such systems depends on the details of the service time distribution and is characterized by performance distributions such as queue length, busy period length, etc. In principle, given the service time probability density $s(t)$, one can compute the performance distributions using standard techniques [29–31]. But suppose, instead of $s(t)$, one knows only its first n moments

$$s_m = \int dt\, t^m s(t), \qquad m = 1, \ldots, n.$$

What is the best way to use this information in estimating the performance distributions?

Our approach exploits the fact that moments of the performance distributions are themselves determined by the service time moments s_m and the average arrival time (a sufficient statistic of the exponential interarrival time density). Thus, knowledge of the service time moments is equivalent to knowledge of moments of the performance distributions. Given these moments, we use the principle of maximum entropy to estimate the performance distributions themselves. Because entropy maximization has been shown to be a uniquely correct, self-consistent method of inference about probability distri-

butions [5], we refer to the resulting estimates of the performance distributions as information-theoretic approximations.

Let the $M/G/1$ system have an average arrival rate λ and a service time density $s(t)$ with moments s_i. Let $q_c(k)$ be the steady state probability that k customers are in the system (queued or being served), and let c_m be the moments

$$c_m = \sum_{k=0}^{\infty} k^m q_c(k).$$

In this Section, we use the Pollaczeck-Khinchin formula to express the expected number of customers c_1 in terms of the first two service time moments, and we derive a maximum entropy estimate of $q_c(k)$ given c_1. We then derive a formula expressing c_2 in terms of s_1, s_2, and s_3, and we compute maximum entropy estimates of q_c given c_1 and c_2. As examples, we consider $M/M/1$, $M/H_2/1$, and $M/D/1$ systems.

The result of maximizing the entropy of q_c subject to the single known constraint c_1 and the normalization constraint $\sum_k q_c(k) = 1$ is

$$q_c(k) = Z^{-1} e^{-\beta k}, \tag{8}$$

where,

$$Z = e^{\lambda} = \sum_{k=0}^{\infty} e^{-\beta k} = (1 - e^{-\beta})^{-1} \tag{9}$$

(see (4)-(5)). We apply (6) in order to express the multiplier β in terms of the constraints c_1,

$$c_1 = -\frac{\partial}{\partial \beta} \log(Z) = (e^{\beta} - 1)^{-1}.$$

Solving this for β enables us to eliminate β from (8):

$$q_c(k) = \frac{1}{1 + c_1} \left(\frac{c_1}{1 + c_1} \right)^k \tag{10}$$

This expression gives the maximum entropy estimate of q_c directly in terms of the known information c_1.

Now, knowledge of s_1 and s_2 yields knowledge of c_1 by the Pollaczeck-Khinchen formula [31, p. 187]

$$\begin{aligned} c_1 &= \rho + \rho^2 \frac{(1 + C^2)}{2(1 - \rho)}, \\ &= \rho + \frac{\lambda^2 s_2}{2(1 - \rho)}, \end{aligned} \tag{11}$$

where $\rho = \lambda s_1$, and C is the coefficient of variation $C = (s_2 - s_1^2)^{\frac{1}{2}} / s_1$. Thus, (10) and (11) provide an information theoretic approximation to q_c for an $M/G/1$ system given the average arrival rate and the first two moments of the service time density.

As an example application, we consider an $M/H_2/1$ system solved exactly by Kleinrock [31, pp. 195-96]. The service time distribution is

$$s(t)=\tfrac{1}{4}\lambda e^{-\lambda t}+\tfrac{3}{4}(2\lambda)e^{-2\lambda t} \tag{12}$$

for which $\rho=\lambda s_1=5/8$ and $C^2=31/25$. Substituting these values into (11) yields $c_1=1.79166$. The information theoretic approximation (10) then becomes

$$q_c(k)=0.358209(0.641791)^k. \tag{13}$$

The exact solution for q_c is [31, p. 196]

$$q_c(k)=\tfrac{3}{32}(\tfrac{2}{5})^k+\tfrac{9}{32}(\tfrac{2}{3})^k. \tag{14}$$

We compare the one-moment approximation (13) with the exact solution (14) in the first three columns of Table 1. The close agreement arises because the exact solution (14) is the sum of two similar geometric terms, which can be approximated closely by a single geometric term (13). In general, the single-moment result (10) can be thought of as providing the geometric distribution that is the best information theoretic approximation to q_c.

Table 1. Comparison of exact and approximate solutions for $M/H_2/1$ queue length distribution

k	$q_c(k)$ (exact)	$q_c(k)$ (1 moment approx.)	$q_c(k)$ (2 moment approx.)
0	0.375	0.358	0.367
1	0.225	0.230	0.229
2	0.140	0.148	0.144
3	0.0893	0.0947	0.0914
4	0.0580	0.0608	0.0583
5	0.0380	0.0390	0.0375
6	0.0251	0.0250	0.0243
7	0.0166	0.0161	0.0158
8	0.0110	0.0103	0.0104
9	0.00734	0.00662	0.00688
10	0.00489	0.00425	0.00458

In the exact solution itself happens to be geometric, then the approximation (10) will be the same as the exact solution. For example, suppose that the service time distribution is exponential $s(t)=\mu e^{-\mu t}$. Then (11) reduces to $c_1=\rho/(1-\rho)$, with $\rho=\lambda/\mu$, and the approximation (10) becomes $q_c(k)=(1-\rho)\rho^k$, which is the exact solution for the $M/M/1$ system [31, p. 96].

If other moments besides c_1 are known, the maximum entropy estimate of q_c will no longer in general be geometric. In order to illustrate muiti-moment approximations, we begin by expressing c_2 in terms of the service time moments s_m. Our starting point is the relation

$$c_2=c_1+\lambda^2 r_2, \tag{15}$$

where r_2 is the second moment of the system residence time probability density [31, p. 240]. The moments r_m are related to the s_m and to the moments w_m of the waiting time probability density by

$$r_r = \sum_{i=0}^{k} \binom{k}{i} w_{k-i} s_i, \tag{16}$$

where $w_0 = s_0 \equiv 1$ [31, p. 202], and the w_m are in turn related to the s_m by the Takacs recurrance formula [31, p. 201]

$$w_k = \frac{\lambda}{1-\rho} \sum_{i=1}^{k} \binom{k}{i} \frac{s_{i+1}}{i+1} w_{k-i}. \tag{17}$$

By combing (11), (16), and (17) with (15), we obtain

$$c_2 = \rho + \frac{\lambda^2 s_2}{2(1-\rho)} + \frac{\lambda^4 s_2}{2(1-\rho)^2} + \frac{\lambda^3 s_3}{3(1-\rho)} + \frac{\lambda^3 s_1 s_2}{(1-\rho)} + \lambda^2 s_2. \tag{18}$$

Now the maximum entropy solution for $q_c(k)$ given c_1 and c_2 cannot be expressed analytically in terms of the moments s_m, so we resort to numerical techniques. We use an APL function written by Johnson [15] that computes maximum entropy distributions given arbitrary expected value constraints. For the $M/H_2/1$ example, we have $c_1 = 1.79166$ from before. The moment s_3 is easily obtained from (12), and $c_2 = 8.68055$ follows from (18). Using the APL function to find the maximum entropy approximations for q_c given c_1 and c_2, we obtain the results shown in the fourth column in Table 1. This approximation, which was computed for 50 points, required 1.5 CPU seconds on a DEC PDP-10 KI processor. It is worth noting that single-moment results from the APL function agreed with the analytic expression (10) up to eight digits.

As an additional example, we consider a system with constant ("deterministic") service time $1/\mu$ - i.e., $M/D/1$. The service time probability density is $s(t) = \delta(t - 1/\mu)$, with moments

$$s_m = 1/\mu^m. \tag{19}$$

We use (19), (11) and (10) to obtain a single-moment approximation for q_c, and we use (19), (18), (11), and the APL function to obtain a two-moment approximation. For $\lambda = 1$ and $\mu = 2$, the results are shown in Table 2 together with simulation results. The simulation result $q_c(k)$ is the relative amount of time the system had k customers present during an overall period covering 5,000 arrivals. The two-moment approximation in Table 2 required 1.6 CPU seconds.

Approximations involving more moments can be computed similarly since c_m can in general be expressed as a function of $s_1, \ldots, s_{m+1}$ - one method is to differentiate the Pollaczek-Khinchin transform equation [31, p. 194]. But the accuracy of the two-moment approximation for the $M/H_2/1$ and $M/D/1$ examples, which have radically different service time densities, and the reduction of the one-moment approximation to the exact result in the $M/M/1$ case, together

Table 2. Comparison of information theoretic approximations and simulation results for $M/D/1$ queue length distribution. ($\lambda=1$ and $\mu=2$)

k	$q_c(k)$ (simulation)	$q_c(k)$ (1 moment approx.)	$q_c(k)$ (2 moment approx.)
0	0.50	0.57	0.51
1	0.33	0.24	0.30
2	0.12	0.10	0.13
3	0.038	0.045	0.044
4	0.0093	0.019	0.011
5	0.0025	0.0083	0.0022
6	0.00047	0.0035	0.00033
7	0.0000081	0.0015	0.000037
8	0	0.00065	0.0000032

suggest that the two-moment approximation will in general be quite accurate for $M/G/1$ systems. This is only a conjecture, however, and more detailed studies are needed.

IV. Number Served in a $M/G/1$ Busy Period

If the system is empty and a customer arrives at time t_1, and if t_2 is the next time at which the system is empty, then the period between t_1 and t_2 is called a busy period. Let $q_n(k)$ be the probability that the number of customers served in a busy period is k, and let n_m be the moments of q_n.

In general, n_m can be expressed as a function of λ and the service time moments $s_1, \ldots, s_m$. For $n_1, \ldots, n_4$, explicit formulas are given in [30, p. 158]. For example, we have

$$n_1 = \frac{1}{1-\rho} \tag{20}$$

where $\rho=\lambda s_1$. Given the mean number served during a busy period (n_1), the maximum entropy distribution is $q_n(k)=Z^{-1}e^{-\beta k}$, where

$$Z=\sum_{k=1}^{\infty} e^{-\beta k}=(e^{\beta}-1)^{-1}.$$

Applying (6), we eliminate β and express q_n directly in terms of the known constraint n_1:

$$q_n(k)=\frac{1}{n_1-1}\left(\frac{n_1-1}{n_1}\right)^k. \tag{21}$$

This result differs from (10) because the domain of $q_n(k)$ is $k=1, \ldots, \infty$ instead of $k=0, \ldots, \infty$. Combining (20) and (21) yields

$$q_n(k)=(1-\rho)\,\rho^{k-1}, \tag{22}$$

where $\rho=\lambda s_1$. Equation (22) provides an information theoretic approximation to the number served in a busy period for an $M/G/1$ system given the mean arrival rate and the mean service time.

Now, unlike the case for the distribution q_c, the distribution q_n for an $M/M/1$ system is not geometric. In fact, the exact result is [31, p. 218]

$$q_n(k)=\frac{1}{k}\binom{2k-2}{k-1}\rho^{k-1}(1+\rho)^{1-2k}. \tag{23}$$

This gives an opportunity to show how knowledge of higher moments than s_1 can be used to provide better approximations than (22). Now, for an $M/M/1$ system with $s(t)=\mu e^{-\mu t}$, the moments s_m are

$$s_m=m!/\mu^m. \tag{24}$$

For a given λ and μ, we use (24) and the formulas in [30, p. 158] to compute the moments n_m, and we use the APL function to compute the maximum entropy distribution $q_n(k)$ given the n_m. In Table 3, for $\lambda=2$ and $\mu=8$, we compare the exact solution (23) with the single moment approximation (22) and the four moment approximation computed by the APL function. (Approximations based on two and three moments fall between the approximations shown.) In Table 4 we present the same comparison for $\lambda=1$ and $\mu=2$.

As another example, we again consider the $M/D/1$ system. As in the $M/M/1$ case, the exact result for q_n is known, namely [31, p. 219]

$$q_n(k)=\frac{(k\rho)^{k-1}}{k!}e^{-k\rho}. \tag{25}$$

For a given λ and μ, we use (19) and the formulas in [30, p. 158] to compute the n_m and then the APL function to compute maximum entropy approxi-

Table 3. Comparison of exact and approximate solutions for distribution of number served in an $M/M/1$ busy period. ($\lambda=2$ and $\mu=8$)

k	$q_n(k)$ (exact)	$q_n(k)$ (1-moment approx.)	$q_n(k)$ (4-moment approx.)
1	0.800	0.750	0.793
2	0.128	0.187	0.142
3	0.0410	0.0469	0.0372
4	0.0164	0.0117	0.0133
5	0.00734	0.00293	0.00611
6	0.00352	0.000732	0.00334
7	0.00177	0.000183	0.00205
8	0.000921	0.0000458	0.00134
9	0.000491	0.0000114	0.000888
10	0.000267	0.00000286	0.000567

Table 4. Comparison of exact and approximate solutions for distribution of number served in an $M/M/1$ busy period. ($\lambda=1$ and $\mu=2$)

k	$q_n(k)$ (exact)	$q_n(k)$ (1 moment approx.)	$q_n(k)$ (4 moment approx.)
1	0.666	0.500	0.629
2	0.148	0.250	0.195
3	0.0658	0.125	0.0737
4	0.0365	0.0625	0.0332
5	0.0227	0.0312	0.0174
6	0.0152	0.0156	0.0104
7	0.0106	0.00781	0.00696
8	0.00765	0.00391	0.00511
9	0.00567	0.00195	0.00404
10	0.00428	0.000977	0.00337

Table 5. Comparison of exact and approximate solutions for distribution of number served in an $M/D/1$ busy period. ($\lambda=2$ and $\mu=8$)

k	$q_n(k)$ (exact)	$q_n(k)$ (1 moment approx.)	$q_n(k)$ (4 moment approx.)
1	0.779	0.750	0.767
2	0.151	0.187	0.169
3	0.0443	0.0469	0.0433
4	0.0153	0.0117	0.0127
5	0.00583	0.00293	0.00426
6	0.00235	0.000732	0.00169
7	0.000990	0.000183	0.000682
8	0.000430	0.0000458	0.000321
9	0.000191	0.0000114	0.000167
10	0.0000863	0.00000286	0.0000949

Table 6. Comparison of exact and approximate solutions for distribution of number served in an $M/D/1$ busy period. ($\lambda=1$ and $\mu=2$)

k	$q_n(k)$ (exact)	$q_n(k)$ (1 moment approx.)	$q_n(k)$ (4 moment approx.)
1	0.606	0.500	0.589
2	0.184	0.250	0.208
3	0.0837	0.125	0.0868
4	0.0451	0.0625	0.0420
5	0.0267	0.0312	0.0230
6	0.0168	0.0156	0.0140
7	0.0110	0.00781	0.00927
8	0.00744	0.00391	0.00657
9	0.00515	0.00195	0.00490
10	0.00363	0.000977	0.00378

Table 7. Comparison of exact and approximate solutions for distribution of number served in an $M/D/1$ busy period. ($\lambda=7$ and $\mu=10$)

k	$q_n(k)$ (exact)	$q_n(k)$ (1 moment approx.)	$q_n(k)$ (4 moment approx.)
1	0.497	0.300	0.432
2	0.173	0.210	0.218
3	0.0900	0.147	0.118
4	0.0556	0.103	0.0673
5	0.0378	0.0720	0.0407
6	0.0272	0.0504	0.0260
7	0.0204	0.0353	0.0174
8	0.0158	0.0247	0.0122
9	0.0126	0.0173	0.00892
10	0.0101	0.0121	0.00678

mations to q_n. Results comparing the exact solution (25) with one- and four-moment approximations are given in Tables 5–7. The values for λ and μ in Tables 5 and 6 are the same as those for the $M/M/1$ examples in Table 3 and 4. Table 7 is for $\lambda=7$ and $\mu=10$. The four-moment approximations in Tables 5–7 required about 1.5 CPU seconds each.

V. $M/G/1$ Busy Period Length

We now consider the probability density $q_b(t)$ for the length of the busy period. In general, the moments $b_m=\int dt\, t^m q_b(t)$ can be expressed in terms of λ and the service time moments $s_1, \ldots, s_m$. For $b_1, \ldots, b_4$, explicit formulas are given in

[31, p. 213–14]. For example, we have

$$b_1 = \frac{s_1}{1-\rho} \tag{26}$$

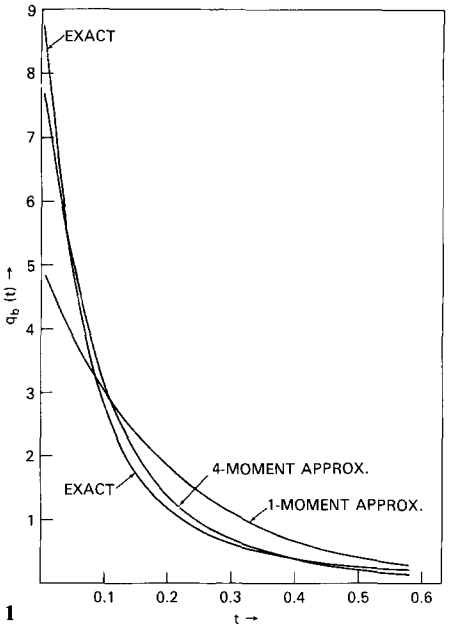
where, as usual, $\rho = \lambda s_1$. If only s_1 is known, then only b_1 is determined. The resulting maximum entropy solution for q_b is $q_b(t) = (1/b_1)\exp(-t/b_1)$. (We omit the standard derivation, which is just the continuous analog of the derivation of (10).) Combining this solution with (26) yields

$$q_b(t) = (\mu' - \lambda)\, e^{-(\mu' - \lambda)t} \tag{27}$$

where $\mu' = 1/s_1$. Equation (27) provides an information theoretic approximation to the busy period probability density for an $M/G/1$ system given the mean arrival rate and the mean service time.

If higher moments than s_1 are known, then better approximations can be obtained using the formulas in [31, pp. 213–14] and numerical techniques. As in the previous section, the exact solution for an $M/M/1$ system is known:

$$q_b(t) = \frac{1}{t\sqrt{\lambda/\mu}}\, e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu}), \tag{28}$$



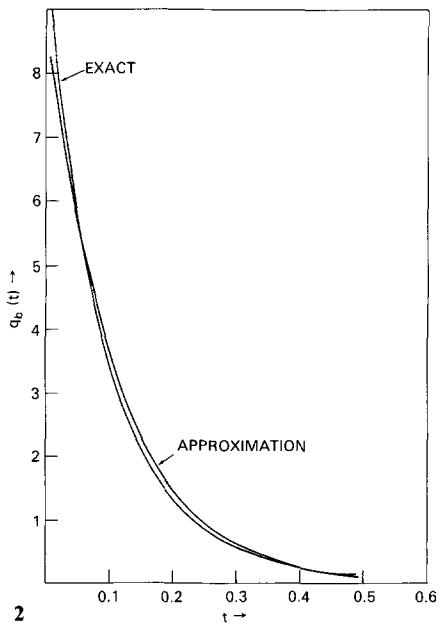


Fig. 1. Exact and approximate $M/M/1$ busy period probability densities ($\lambda = 5$, $\mu = 10$)

Fig. 2. Exact and 1-moment approximation for $M/M/1$ busy period probability density ($\lambda = 1$, $\mu = 10$)

Table 8. Comparison of exact and four-moment approximation for probability density of $M/M/1$ busy period length ($\lambda=1$ and $\mu=10$)

Time	$q_b(t)$ (exact)	$q_b(t)$ (4 moment approx.)
0.01	8.96	9.05
0.03	7.22	7.27
0.05	5.84	5.87
0.07	4.74	4.75
0.09	3.87	3.87
0.11	3.17	3.16
0.13	2.60	2.59
0.15	2.14	2.14
0.25	0.861	0.854
0.35	0.373	0.369
0.45	0.171	0.171

where I_1 is the modified Bessel function of the first kind (order one) [31, p. 215]. We therefore assume $s(t)$ to be exponential, compute various approximations based on (24) and the formulas in [31, pp. 213–14], and compare the results with (28). Results for the case $\lambda=5$ and $\mu=10$ are shown in Fig. 1. Results for the case $\lambda=1$ and $\mu=10$ are shown in Fig. 2 for the one-moment approximation and in Table 8 for the four-moment approximation at selected points. The single moment approximations, which were computed by the APL function, agree in both cases with (27). (Note that, since $q_b(t)$ is a continuous probability density, appropriate care must be taken when using the APL function to compute approximations. For details, see [32].)

The results in Fig. 2 suggest that (27) might be a good light-load approximation for the $M/M/1$ busy period density (28), a conjecture that has been supported by further studies: For $\rho \lesssim 0.1$, (27) is accurate to within 5–10% in the range where the cumulative probability distribution of $q_b(t)$ is as large as about 0.95 [33]. The conjecture is supported further by the following argument, which is due to A.E. Ephremides [34]: Equation (27) is identical to the exact $M/M/1$ residence time probability density [31, p. 202]. Since most busy periods will consist of single customer residences under light load conditions, it makes sense that the busy period should tend to (27).

VI. $M/G/1$ Residence Time and Waiting Time

Residence time is the total time a customer spends in the system. Waiting time is the interval between the arrival time and the time at which service begins. Moments r_m of the residence time probability density $q_r(t)$ can be expressed in

terms of the service time moments s_m by using (16) and (17). For example, we have

$$r_1 = \frac{\lambda s_2}{2(1-\rho)} + \frac{\rho}{\lambda}. \tag{29}$$

where $\rho = \lambda s_1$. The maximum entropy density $q_r(t)$ given r_1 is just

$$q_r(t) = (1/r_1) \exp(-t/r_1). \tag{30}$$

Equations (29)-(30) provide an information theoretic approximation to the residence time probability density for an $M/G/1$ system given the mean arrival rate and the first two moments of the service time density. If higher moments than s_2 are known, then better approximations for q_r can be obtained by using (16), (17), and the computational methods discussed earlier.

For an $M/M/1$ system, (29) reduces to $r_1 = \rho/\lambda(1-\rho)$ and (30) becomes $g_r(t) = \mu(1-\rho)\exp(-\mu(1-\rho)t)$, where $\mu = 1/s_1$, which is the exact $M/M/1$ solution [31, p. 202]. This behavior is similar to that of the one-moment approximation for $q_c(k)$ discussed in Sect. III. The similarity arises from (30) being the continuous analog of (10) and from Little's result.

The situation for waiting times is somewhat more complicated. Let $q_w(t)$ be the waiting time probability density with moments w_m. The w_m can be expressed in terms of the s_m using (17); for example,

$$w_1 = \frac{\lambda s_2}{2(1-\rho)} \tag{31}$$

where $\rho = \lambda s_1$. The maximum entropy solution given just w_1 is

$$q_w(t) = (1/w_1) \exp(-t/w_1). \tag{32}$$

In the $M/M/1$ case, (31) becomes $w_1 = \rho/\mu(1-\rho)$ and (32) becomes

$$q_w(t) = (\mu/\rho)(1-\rho)\exp(-\mu(1-\rho)t/\rho), \tag{33}$$

in contrast to the exact $M/M/1$ result [31, p. 203]

$$q_w(t) = (1-\rho)\delta(t) + \lambda(1-\rho)\exp(-\mu(1-\rho)t). \tag{34}$$

Equations (33) and (34) have the same mean w_1, but (33) lacks the impulse term at $t=0$ that results from the finite probability $q_c(0)$ that the system is empty when a customer arrives. We can, however, improve on (33) by noting that s_1 and s_2 provide information about $q_c(0)$. In particular, we have

$$q_c(0) = (1+c_1)^{-1} = (1+\rho+\lambda w_1)^{-1}$$

from (10) and (31). Now the total probability in $q_w(t)$ that is concentrated at $t=0$ must equal $q_c(0)$. We express this fact as

$$\lim_{\varepsilon \to 0} \int dt\, u_\varepsilon(t)\, q_w(t) = q_c(0) = (1+\rho+\lambda w_1)^{-1}, \tag{35}$$

where

$$u_\varepsilon(t)=\begin{cases}1, & t\leqq\varepsilon\\ 0, & t>\varepsilon.\end{cases}$$

But the integral in (35) is just a constraint (2) that we can impose in addition to the moment constraint $\int dt\, t\, q_w(t)=w_1$. The maximum entropy density that satisfies both of these constraints is

$$q_w(t)=(\lambda w_1+\rho+1)^{-1}\,\delta(t)+w_1 B^2\exp(-Bt), \tag{36}$$

where

$$B=\frac{\rho+\lambda w_1}{w_1(1+\rho+\lambda w_1)}. \tag{37}$$

Equations (36), (37), and (31) provide an information theoretic approximation to the waiting time probability density for an $M/G/1$ system given λ, s_1, and s_2. Unlike (32), (36) reduces to (34) in the $M/M/1$ case $w_1=\rho/\mu(1-\rho)$.

VII. Some $G/G/1$ Results

We consider a $G/G/1$ queue that has a probability density of interarrival times $a(t)$ with moments a_m and a probability density of service times $s(t)$ with moments s_m. We discuss approximations for the case in which only a_1 and s_1 are known.

Equations (10) is the maximum entropy distribution of queue length q_c given the first moment c_1. The probability that the system is empty is therefore

$$q_c(0)=(1+c_1)^{-1}. \tag{38}$$

Now, if the $G/G/1$ system is in equilibrium, $(1-q_c(0))/s_1=1/a_1$ must hold. Solving for $q_c(0)$ and substituting the result into (38) yields

$$c_1=\frac{s_1/a_1}{(1-s_1/a_1)}. \tag{39}$$

Equation (10) then yields

$$q_c(k)=(1-\rho)\,\rho^k, \tag{40}$$

where $\rho=s_1/a_1$. This is an information theoretic approximation for the $G/G/1$ queue length given the first moments of the arrival and service time densities. As was the case for the $M/G/1$ approximation (10)–(11), Eq. (40) yields the exact $M/M/1$ result when $a(t)$ and $s(t)$ are exponential. Stated differently, (40) shows that the $M/M/1$ result is also the proper information theoretic approximation for $G/G/1$ systems given only a_1 and s_1.

Next we consider the residence time density q_r. Equation (39) and Little's result $c_1=r_1/a_1$ yield $r_1=s_1/(1-s_1/a_1)$. The maximum entropy density q_r given r_1 is then

$$q_r(t)=\mu(1-\rho)\exp(-\mu(1-\rho)t), \tag{41}$$

where $\rho = s_1/a_1$ and $\mu = 1/s_1$. This is an informatic theoretic approximation for the $G/G/1$ queue length given the first moments of the arrival and service time densities. Like the $M/G/1$ approximation (29)-(30), (41) yields the exact $M/M/1$ result when $a(t)$ and $s(t)$ are exponential. This also shows that the $M/M/1$ result is the proper information theoretic approximation for $G/G/1$ systems given only a_1 and s_1.

Similar arguments based on results from Section VI apply in the case of the waiting time density w_t. In this case, the $G/G/1$ approximation given a_1 and s_1 is

$$q_w(t) = (1-\rho)\,\delta(t) + \lambda(1-\rho)\exp(-\mu(1-\rho)\,t), \tag{42}$$

where $\rho = s_1/a_1$ and $\mu = 1/s_1$.

VIII. Using non-Uniform Priors

Since entropy maximization is equivalent to cross-entropy minimization with a uniform prior (see Sect. II), the information theoretic approximations discussed in Sects. III-VII are properly thought of as being based on uniform prior estimates of the performance distributions. If information about the performance distributions in addition to the s_m is available and can be expressed as non-uniform prior estimates, better approximations can result. For example, if it is suspected that the service time density $s(t)$ is nearly exponential, it is reasonable to use $M/M/1$ performance distributions as prior estimates of $M/G/1$ distributions. Note that the prior distribution reflects one's belief about the variable being measured - e.g., the busy period - and not one's belief about parameters of the variable's distribution as might be the case in a Bayesian context.

As a specific example, suppose we wish to estimate the busy period density q_b based on measurements of λ, s_1 and s_2. As a prior estimate, we use the exact $M/M/1$ result (28) with $\mu = 1/s_1$, and we compute the moments b_1 and b_2 from the formulas in [31, pp. 231-14]. We obtain a posterior approximation by minimizing cross-entropy with respect to the prior subject to the constraints involving b_1 and b_2. If s_2 happens to satisfy $s_2 = 2s_1^2$, which would always be the case if $s(t)$ were exponential, then the posterior would be unchanged from the prior since the $M/M/1$ prior itself satisfies the contraints b_1 and b_2 [14]. If the $M/M/1$ prior does not satisfy the constraints b_1 and b_2, the posterior will be different. In an information theoretic sense, however, it will be the closest density that satisfies the constraints. Figure 3 shows an example in which two-moment approximations for q_b were computed using both uniform and $M/M/1$ priors. The parameters in both cases were $\lambda = 5$, $s_1 = 0.1$ and $s_2 = 0.04$. The second moment is larger than it would be if $s(t)$ were exponential - the coefficient of variation is 1.74 instead of one. Since $\lambda = 5$ and $1/s_1 = 10$, the non-uniform prior used in computing the result in Fig. 3 is the same as the $M/M/1$ curve shown in Fig. 1. The results in Fig. 3 were obtained using an APL function that finds a minimum cross-entropy posterior given an arbitrary prior

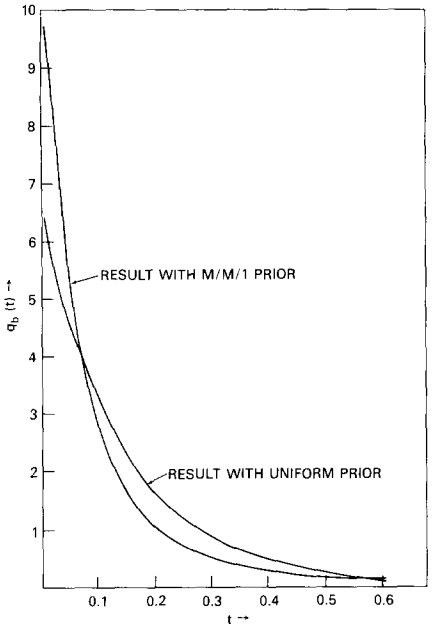


Fig. 3. Two-moment approximations for $M/G/1$ busy period probability density using uniform and $M/M/1$ priors ($\lambda = 5$, $s_1 = 0.1$, $s_2 = 0.04$)

and an arbitrary constraint matrix [15]. Note that, although the prior used in this example happens to be Markovian, any form of prior can be used.

IX. Discussion

We have presented a variety of results concerning the use of information theoretic methods in system modeling applications. We have shown that many well-known $M/M/1$ formulas are also information theoretic approximations for $G/G/1$ systems in cases when only the mean arrival and service rates are known. That is, the $M/M/1$ formulas are the best hypotheses about the $G/G/1$ systems given only the mean arrival and service rates. This fact has nothing at all to do with the various assumptions that must be debated when considering the applicability of stochastic models, and it helps to explain why the $M/M/1$ formulas have been found to be so useful.

Beyond this rather general conclusion, our results can be used in three specific ways. First, the techniques presented could be used as a general method of computing the performance distributions in cases where all of the service density moments are available, i.e., when the density $s(t)$ is known exactly. Second, the analytic approximations – (10) and (11), (22), (27), (29) and (30), (31) and (36), (40–(42) – could be useful in various studies whenever explicit forms for the performance distributions are required. Third, the tech-

niques provide a means of estimating the performance distributions when only the first few moments of $s(t)$ are known and $s(t)$ itself is not known. If the first few moments are estimated rather than known exactly, then it is important that unbiased estimators be used and that the resulting set of estimates be consistent.

How accurate are these information theoretic approximations? Unfortunately, about all that can be said in general is that the approximations are the least-biased choices given the information available. To use the language of statistics [12, 16], the approximations are the hypotheses best supported by the information available. Depending on the actual performance distribution and the number of moments considered, an information theoretic approximation may or may not be a good approximation in the mean-squared-error sense, although it is true that the mean-squared-error can always be made sufficiently small by taking sufficiently many moments into account. On the other hand, it is not generally known what kind of error measure is best for judging the accuracy of performance distribution approximations. It may well be that measures such as mean-squared-error are less important than information theoretic measures such as cross-entropy. More can be said about the queue length distribution q_c and the busy period density q_b, because, although an explicit proof is lacking, it seems clear that these must be monotonically decreasing functions for a wide class of $M/G/1$ systems. If so, then q_c and q_b don't have basic structure that would be seen in approximations based on many moments but not seen in approximations based on only a few moments. This in turn means that the basic shape will be revealed by approximations based on the first few moments, and suggests that a large number of moments will not in general be required in order to achieve low mean-squared-error. In the case of the queue length distribution, the diverse examples discussed in Sect. III suggest that a two-moment approximation may in general be quite good.

Acknowledgements. I thank R.W. Johnson and H. Vantilborgh for helpful discussions, H. Vantilborgh for suggesting that (32) might be a good light-load approximation. I also thank D. Baker, A.E. Ephemides, and a referee for their comments on an earlier version of this paper.

References

1. Muntz, R.R.: Analytic Modeling of Interactive Systems. Proc. IEEE **63**, 946–953 (1975)
2. Chen, P.P.: Queueing Network Model of Interactive Computing Systems. Proc. IEEE **63**, 954–957 (1975)
3. Kleinrock, L.: Queueing Systems, Vol. **II**: Computer Applications. New York: John Wiley 1976
4. Denning, P.J., Buzen, F.P.: The Operational Analysis of Queueing Network Models. Computing Surveys **10**, 225–261 (1978)
5. Shore, J.E., Johnson, R.W.: Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. IEEE Trans. Information Theory IT-**26**, 26–37 (1980)
6. Ferdinand, A.E.: A Statistical Mechanics Approach to Systems Analysis. IBM J. Res. Dev. 539–547 (1970)

7. Beneš, V.E.: Mathematical Theory of Connecting Networks and Telephone Traffic. New York: Academic Press 1965
8. Shore, J.E.: Derivation of Equilibrium and Time-Dependent Solutions ot $M/M/\infty//N$ and $M/M/\infty$ Queueing Systems Using Entropy Maximization. Proceedings 1978 National Computer Conference, AFIPS, 1978, pp. 483–487
9. Bard, Y.: A Model of Shared DASD and Multipathing. CACM **23**, 564–572 (1980)
10. Jaynes, E.T.: Information Theory and Statistical Mechanics I. Phys. Rev. **106**, 171–190 (1957)
11. Elsasser, W.M.: On Quantum Measurements and the Role of the Uncertainty Relations in Statistical Mechanics. Phys. Rev. **52**, 987–999 (1937)
12. Kullback, S.: Information Theory and Statistics. New York: Wiley 1959
13. Csiszàr, I.: I-Divergence Geometry of Probability Distributions and Minimization Problems. Ann. Prob. **3**, 146–58 (1975)
14. Shore, J.E., Johnson, R.W.: Properties of Cross-Entropy Minimization. IEEE Trans. Inf. Theory IT-**27**, 472–482 (1980)
15. Johnson, R.W.: Determining Probability Distributions by Maximum Entropy and Minimum Cross-Entropy. Proceedings APL79, (ACM 0-89791-005), May 1979, pp. 24–29
16. Good, I.J.: Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables. Annals Math. Stat. **34**, 911–934 (1963)
17. Jaynes, E.T.: Prior Probabilities. IEEE Trans. on Systems Science and Cybernetics SSC-**4**, 227–241 (1968)
18. Hobson, A., Cheng, B.: A Comparison of the Shannon and Kullback Information Measures. J. Stat. Phys. **7**, 301–310 (1973)
19. Johnson, R.W.: Axiomatic Characterization of the Directed Divergences and Their Linear Combinations. IEEE Trans. Information Theory IT-**25**, 709–716 (1979)
20. Good, I.J.: Probability and the Weighing of Evidence. London: Charles Griffen 1950
21. Pinsker, M.S.: Information and Information Stability of Random Variables and Processes. Holden-Day: San Francisco 1964 (1981)
22. Shore, J.E.: Minimum Cross-Entropy Spectral Analysis. IEEE Trans. Acoustics, Speech, & Signal Proc. ASSP-**29**, 230–237 (1981)
23. Gray, R.M., Gray, A.H., Jr., Rebolledo, G., Shore, J.E.: Rate-Distortion Speech Coding With a Minimum Discrimination Information Distortion Measure. IEEE Trans. Inf. Theory IT-**27**, November 1981
24. Shore, J.E., Gray, R.M.: Minimum Cross-Entropy Pattern Classification and Cluster Analysis. IEEE Trans. Pattern Anal. & Mach. Intell., January 1982
25. Hobson, A.: A New Theorem of Information Theory. J. Stat. Phys. **1**, 383–391 (1969)
26. Geronimus, L.: Orthogonal Polynomials. New York: Consultants bureau 1961
27. Gray, R.M., Buzo, A., Gray, A.H., Jr., Matsuyama, Y.: Distortion Measures for Speech Processing. IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-**28**, 367–376 (1980)
28. Markel, J.D., Gray, A.H., Jr.: Linear Prediction of Speech. New York: Springer 1976
29. Cohen, J.W.: The Single Server Queue. Amsterdam: North-Holland 1969
30. Cox, D.R., Smith, W.L.: Queues. London: Chapman and Hall 1961
31. Kleinrock, L.: Queueing Systems, Vol I: Theory. New York: John Wiley 1975
32. Shore, J.E.: Information Theoretic Approximations for $M/G/1$ and $G/G/1$ Queueing Systems. Naval Research Laboratory Memorandum Report 4047, July 1979
33. Shore, J.E.: Accuracy of an Information-Theoretic, Light-Load Approximation for the $M/M/1$ Busy Period Probability Density. Naval Research Laboratory Memorandum Report 4037, July 1979
34. Ephremides, A.E.: private communication

Received March 22, 1981