Inferring Balking Behavior and Queue Performance from Transactional Data

Lee K. Jones

OR 307-95 May 1995

Inferring Balking Behavior and Queue Performance From Transactional Data

Lee K. Jones
Department of Mathematical Sciences
University of Massachusetts Lowell
One University Ave.
Lowell, MA 01854

August 16, 1993

Revised March 28, 1994

Research supported in part by NSF Grant No. DMS-9202161

Abstract

Balking is the act of not joining a queue because the prospective arriving customer judges the queue to be too long. We analyze Poisson-arrival and more general queues in the presence of balking, using only the service start and stop data utilized in Larson's Queue Inference Engine (Q.I.E.). First we extend Larson's Queue Inference Engine to the case of an arbitrary given balking function and a general (partially unknown) customer arrival process. This yields new nonparametric estimates of the rate at which potential customers arrive. Second by parametrizing both the arrival process and balking function we present new maximum likelihood and Bayesian methods for inferring the arrival rate and balking parameters. The methodology is applicable to businesses that wish to estimate lost sales due to balking arising from queuing-type congestion. The techniques are applied to a small transactional data set for illustrative purposes.

I. INTRODUCTION

A customer who balks at joining a queue is a customer who does not purchase the associated good or service from the queue server(s). The corresponding lost revenues in various industries can be enormous. For instance, just a 1% balking rate in drive-thru windows of U.S.-based Quick Service Restaurants (QSR's) can reduce QSR revenues by over $100 million per year. Thus it is important to devise efficient methods to estimate lost sales due to balking from easily available data. With balking the task is especially difficult since each balker leaves no explicit natural entry in any data base. The balker comes and goes without ever entering a formal data collection system.

Larson [1] recently derived an algorithm, the "Queue Inference Engine" (Q.I.E.), to efficiently calculate various estimates of mean queue length for Poisson arrival queues from a set of transactional data. The transactional data are the times of service initiation and service completion for each customer served in an h-server system* with arbitrary service distributions. The main period of analysis of the Q.I.E. is a single congestion period, a continuous time interval during which all h servers are busy and all arriving customers must queue for service. The state of congestion is identified by the fact that a new customer will enter service *virtually immediately* following a departure of another customer from service. A congestion period commences the moment that all h servers become busy and ends the moment that one of the servers completes service and becomes idle.

Larson's analysis assumes no balking and (hence) his performance estimates do not depend on knowledge of the customer arrival rate. In many queuing applications (i.e. fast food restaurants, car washes, ticket outlets, and Automated Teller Machines (ATM's)) the only data available is transactional. However, customer behavior is characterized by both the arrival rate of potential customers and a (often unknown) balking probability sequence expressed as a function of the queue length found by an arriving potential customer. The balking probabilities are needed to estimate the amount of lost business and evaluate the service configuration. It is not immediately clear how congestion period information may be used to infer the arrival rate, balking function, or queue performance.

* Both Larson's methods and ours require that the number of servers be fixed.

Larson used order statistics to efficiently calculate probabilities of congestion and estimates of mean queue length for Poisson arrival queues. Subsequently Larson [2], Bertsimas and Servi [3], Daley and Servi [4], Jones and Larson [5], and Servi and Daley [6] have given improvements and extensions of the original algorithm using a variety of techniques.

We consider various models for arriving customers in a queuing system who balk with probability dependent on queue length at arrival. Given only transactional data the problem is to reconstruct the model. Nonparametric and Bayesian a posteriori probability approaches will be formulated. Unlike the case for many inverse problems (e.g. emission tomography) the number of unknowns to estimate is small but the difficulty lies in calculating the model likelihood of observing the sequence of congestion periods described by the data.

In this article we develop two new methods for analyzing transactional data: In order to efficiently calculate model likelihoods, first we extend Larson's Q.I.E. to the case of an arbitrary given balking function and a general (partially unknown) customer arrival process. In [6] this was done for homogeneous Poisson arrival queues only for the special case where there is balking just when the queue's state is beyond a threshold and then the balking probability is constant. The techniques do not generalize to other balking functions. (A more realistic balking function would increase to one in queue state. We introduce and motivate a family of balking functions for which this is the case.) To get Larson's probabilities of congestion for an arbitrary balking function (even in the homogeneous or inhomogeneous Poisson case) we need to use a novel combinatorial generating function with a sequence of approximating non-Poisson problems. As a byproduct a maximum likelihood method for queue inference in the more general than Poisson case is derived and used in one of the examples. Second, by parametrizing both the arrival process and balking function, we present new Bayesian methods for inferring the arrival rate and balking parameters. The techniques will be applied to a small transactional data set for illustrative purposes. Three important cases are treated: inferring arrival rate (expected potential customers/length of transaction interval) with known balking function for general queues; inferring a balking function for Poisson queues with known arrival rate; inferring both balking function and arrival rate for Poisson queues when both quantities are unknown. In practical applications all our algorithms will require $0(N^4)$ calculations per congestion period (where N is the number of customers serviced in a congestion period.)

II. PRELIMINARIES

Suppose the arrivals of potential customers (both those who queue and those who balk) are time points in some interval time domain Ω. The point process describing these arrivals will be of a very general nature; it generalizes the inhomogeneous (or homogeneous) Poisson process and seems most appropriate for queuing systems. We will call it an *order statistics process* (O) and it is defined as follows:

Definition Let $f(m)$ be a probability function for the non-negative integers and F(x) be a cumulative distribution function for a random variable taking values in Ω; then the *order statistics process* defined by $f(m)$ and F(x) is constructed as follows—first M potential customers decide to visit the facility during the interval Ω where $M \sim f(m)$; then their (ordered) arrival times are the (order statistics of the) i.i.d. sequence $X_1, X_2, \ldots X_M$ with each X_i having c.d.f. F(x).

In our applications we will always assume that F(x) is known but $f(m)$ may or may not be known. If it is unknown we will derive various estimates of it or its associated parameters and use them in estimating queue performance measures and balking probabilities. Order statistics processes include a very important non-Poisson case—that for which the number of potential customers is a (possibly unknown) constant.

Finally, order statistics processes have the following property which is easily verified:

Property Let $\Omega' \subset \Omega$ be a subdomain in which X_i lies with positive probability. Then the restriction of the order statistics process to include only points in Ω' is also an order statistics process. In fact, if we further restrict these points to be those occurring in Ω' when an additional condition on the points outside Ω' holds, then the resulting point process is still an order statistics process.

This property will be important in our algorithms since we will consider restrictions to congestion periods. The property follows by noting that the $\{X_i\}_1^M$ which are in Ω' are conditionally i.i.d.

We'll assume further that all X_i's in the arrival process are of absolutely continuous type with a Riemann integrable density. This will avoid simultaneous arrivals. Unless otherwise stated we assume a general service discipline with the only assumption that the service times be positive, independent of the arrival times, and have, for a given number of customers, a joint density which is a mixture of δ functions and Riemann integrable functions. In our applications only existence of these service densities will be assumed and not their explicit forms. The following probabilistic arguments can now be justified by making discrete approximations to the queuing system, performing the conditional probabilistic calculations ignoring the possibility of simultaneous arrivals, and then letting the discrete approximations get finer and finer (so that the probability of simultaneous arrivals becomes negligible).

Let a congestion period begin at $t_0=0$. (which may be assumed w.l.o.g. by a time shift.) This will coincide with the arrival of a customer who enters the system (which contains exactly one idle server) and forces all servers to be busy. These servers continue to be busy until $t_1>0$ at which time there is a service completion followed immediately by a service initiation. For the congestion to persist a customer had to have arrived, entered the system and queued at $X_{(1)}$, $0 < X_{(1)} \leq t_1$. The next service completion is at t_2 (possibly $t_2 = t_1$) followed immediately by a service initiation. Hence a second customer arrived at $X_{(2)}$, $X_{(1)} < X_{(2)} \leq t_2$, and queued for service. This process continues at $t_3 \leq t_4 \leq \ldots \leq t_N \leq t_{N+1}$. Customer i arrives and queues at $X_{(i)}$, $X_{(i-1)} < X_{(i)} \leq t_i$. (Recall that with probability one all $X_{(i)}$ are different.) Finally at $t_{N+1} \geq t_N$ congestion ends — a service completion not immediately followed by a service initiation, i.e. no more arrivals who queue in $[t_N, t_{N+1}]$. Now all other potential customers arriving during $(0, t_{N+1}]$* must have balked and not entered the system. We will assume each potential customer balks with probability $p(n)$ (n = 0,1,2,...) where n is the queue length he finds upon his arrival. If the above occurs for $t_0=0, t_1, t_2, \ldots t_{N+1}$ we call $(0, t_{N+1}]$ a congestion period. For the moment we will assume $p(n)$ is known. In Sections IV, VII and VIII we will assume that it is an unknown member of a parametric family $p(n;\alpha)$. The problem of estimating α (which may be a vector parameter) will then be discussed.

* Although the probability of arrival at any fixed time is 0, we will work with inclusive time intervals of the form (a,b] since in a computer program instructions must be given for all possible times.

Let us review the information at hand: we are given the transaction times $\{t_i\}_1^N$ in a congestion period $(0,t_{N+1}]$ and the fact that potential customers arrived according to a general order statistics point process in Ω, each potential customer balking with probability $p(n)$ where n= queue length he discovers upon arrival. Hence the order statistics vector $(X_{(1)}, X_{(2)},\ldots X_{(N)})$ for the arrival times in $(0,t_{N+1}]$ of the non-balking customers satisfies

$$(1) \qquad 0 < X_{(1)} \le t_1, X_{(2)} \le t_2, \ldots X_{(N)} \le t_N$$

Given a congestion period one may ask for a conditional estimate of probabilities of congestion or of some queue performance measure. In the first case multiplication over all congestion periods yields a key factor in the likelihood equations in Section IV. In the second case averaging over several congestion periods and combining with the performance measure in periods of non-congestion one may estimate the performance measure unconditionally. Consider a rectangular subregion of $[0,t_{N+1}]^N$ given by $(s_1, t_1]\ x\ (s_2, t_2]x \ldots x(s_N,t_N]$. Then let

$$(2)\ \Gamma(\vec{s},\vec{t}) = \Pr\left\{ \begin{array}{l} s_1 < X_{(1)} \le t_1,\ s_2 < X_{(2)} \le t_2, \ldots s_N < X_{(N)} \le t_N \\ \text{and no queuers in } \left(t_N,t_{N+1}\right]. \end{array} \middle| \text{congestion begins at 0.} \right\}$$

(Note $\Gamma(\vec{0},\vec{t})$ is just the probability that $(0,t_{N+1}]$ is a congestion period given congestion begins at 0.)

(We may assume w.l.o.g. that the s_i are nondecreasing.) In particular let $s_i^k = t_{\max\{i-k,0\}}$, $\tilde{s}_i^{\tau} = \max\{t_i - \tau, 0\}$. Then maximum experienced queue length (valid for any service discipline) and maximum delay (valid only for first-come first-served discipline) performance measures are given by:

$$(3)\ \Gamma(\vec{s}^k,\vec{t}) \,/\, \Gamma(\vec{0},\vec{t}) = \Pr\left\{ \begin{array}{l} \text{None of the N customers who enter} \\ \text{find a queue length} \ge k \end{array} \middle| \begin{array}{l} (0,t_{N+1}] \text{ is a congestion} \\ \text{period} \end{array} \right\},$$

$$(4)\ \Gamma(\vec{s}^{\tau},\vec{t}) \,/\, \Gamma(\vec{0},\vec{t}) = \Pr\left\{ \begin{array}{l} \text{None of the N customers} \\ \text{are delayed by} \ge \tau \text{ units} \end{array} \middle| \begin{array}{l} (0,t_{N+1}] \text{ is a congestion} \\ \text{period} \end{array} \right\}$$

Set $\vec{s}^{kj} = (\overbrace{0, 0, \ldots 0}^{j-1}, s_j^k\ ,\ s_j^k, \ldots s_j^k)$ and

$$\vec{s}^{\tau j} = (\overbrace{0, 0, \ldots 0}^{j-1}, \bar{s}_j^{\tau}, \bar{s}_{j,}^{\tau} \ldots \bar{s}_j^{\tau}).$$

Then average experienced queue length (valid for any service discipline) and average τ-delay (valid only for first-come first-served discipline) performance measures are given in (6) and (7) of the following sequence.

$$(5) \quad \beta_k = \frac{1}{N} \sum_{i=1}^{N} \Gamma(\vec{s}^{ki}, \vec{t}) \,/\, \Gamma(\vec{0}, \vec{t}) =$$

$$\Pr\left\{\begin{array}{l|l}\text{A randomly chosen (from N) customer} & (0,t_{N+1}] \text{ is a congestion} \\ \text{who entered finds queue length} < k & \text{period}\end{array}\right\}$$

$$(6) \quad L = \sum_{k=1}^{N-1} k\,(\beta_{k+1} - \beta_k) = \begin{array}{l}\text{Average experienced queue length expected in the congestion} \\ \text{period } (0,t_{N+1}] \text{ by the N entering customers}\end{array}$$

$$V = \sum_{k=1}^{N-1} (k - L)^2 (\beta_{k+1} - \beta_k) = \begin{array}{l}\text{Variance of a randomly chosen entering customer's} \\ \text{experienced queue length given congestion}\end{array}$$

$$(7) \quad D_\tau = \frac{1}{N} \sum_{i=1}^{N} \Gamma(\vec{s}^{\tau i}), \vec{t}) \,/\, \Gamma(\vec{0}, \vec{t}) =$$

$$\Pr\left\{\begin{array}{l|l}\text{a randomly chosen (from N)} & (0,t_{N+1}] \text{ is a congestion} \\ \text{customer is delayed} < \tau \text{ units} & \text{period}\end{array}\right\}$$

REMARKS

(a) Clearly the various quantities (2) - (7) may be calculated in principle if F(x) and f(m) are known explicitly by performing integration numerically in N and higher dimensions. The problem is then algorithmic — finding computationally efficient schemes to evaluate probabilities of congestion and estimate queue performance. For the case of homogeneous Poisson arrivals and no balking (p(n)=0), Larson [1] gave an $0(N^5)$ algorithm to evaluate (6) which did not require

knowledge of the arrival rate λ. For this same case first Bertsimas and Servi [3] and then Larson [2] gave $0(N^3)$ algorithms.

(b) For the general inhomogeneous Poisson arrival case without balking and with known time varying arrival rate, Daley and Servi [4] gave an $0(N^3)$ algorithm for (6) using the fact that the restriction of the queue to the congestion period is Markovian; Jones and Larson [5] gave $O(N^3)$ algorithms for (3), (4), and (7) using properties of order statistics. (It is unclear how these latter results may be obtained using the Markov technique without artificially increasing the number of states and/or the number of transitions in the model.)

(c) For the homogeneous Poisson case but with the special balking function $p(n) = \begin{cases} 0, & n \le m \\ p, & n > m \end{cases}$, Servi and Daley [6] give an $O(N^3)$ algorithm for (6) using the Markov technique. This seems very difficult to generalize to arbitrary $p(n)$. In the above case with general $p(n)$, we present a method which gives (6) and (7) in $O(N^4)$. Also, without the Poisson assumption, the queue may not be Markovian in the congestion period. For instance, when the total potential arrivals is a constant, the queue length at t_{i+1} depends on the number of customers who balked in $(0,t_i]$ and this, in turn, depends on the queue lengths at $t_1, t_2, \ldots t_i$. Finally the distribution $f(m)$ will in practice not be known even though F is.* For unknown $f(m)$ our methods yield estimates of the number of arrivals in each congestion period and these estimates will be used to estimate (3) - (7) and $f(m)$. Even if $f(m)$ is known for a non-Poisson case the conditional distribution of the number of arrivals during congestion given the congestion period $(0,t_{N+1}]$ may be infeasible to compute so that the algorithms for unknown $f(m)$ will still be used.

In the next six sections various algorithms for efficiently estimating the (distribution of the) number of potential customers, the balking probabilities, and queue performance will be developed. Classical statistical questions concerning the quality of these estimators (mean, variance, consistency, etc.) will not be addressed but should be the subject of future research. Quality of estimation in the Bayesian

* For example, given that a customer wants to use an A.T.M. (Automated Teller Machine), it may be known that he will arrive uniformly during a one-hour lunch break. The number of such people wanting to use the A.T.M. may be an unknown constant, however.

sense will be demonstrated by numerically intergrating posterior densities for the arrival and/or balking parameters in the Poisson arrival case. All methods proposed are applied to an illustrative case study.

III. THE FUNDAMENTAL RECURSION; EXTENSION OF THE QUEUE INFERENCE ENGINE

Here we give a forward recursion which may be used to calculate $\Gamma(\vec{s},\vec{t})$ and hence (3)-(7) in a brute force fashion. The arrays generated by the recursion yield several estimates of arrival rate when $f(m)$ is unknown and may used to estimate balking parameters. These arrays will be combined efficiently in the Appendix to evaluate (estimate) the performance measures (6), (7) N^2, resp. N times faster than by brute force.

We first consider the auxiliary function

$$S_g^{\omega} = S^{\omega}(p_1, p_2, \ldots, p_g) = \sum_{\substack{l_1+l_2+\ldots l_g = \omega \\ l_i \text{ non-negative integers}}} {p_1}^{l_1} {p_2}^{l_2} \ldots {p_g}^{l_g} \text{ defined for real } p_1, p_2, \ldots$$

and non-negative integer ω. In particular $S_g^0 = 1, S_g^1 = p_1 + p_2 + \ldots + p_g, S_g^2 = p_1^2 + p_2^2 + \ldots p_g^2 + \sum_{1 \le i < j \le g} p_i p_j$. S_g^{ω} may be computed using the recursion

$S_1^0 = 1$, $S_1^1 = p_1$, $S_1^2 = p_1^2$, ... $S_1^{\omega} = p_1^{\omega}$

. . .

$S_g^0 = 1$ ... $S_g^x = S_{g-1}^x + p_g S_{g-1}^{x-1} + p_g^2 S_{g-1}^{x-2} + \ldots + p_g^x$ . . .

We compute S_g^{ω} for $1 \le g \le N+1, 0 \le \omega \le \bar{m}$ based on the balking sequence $p_1 = p(0), p_2 = p(1), \ldots p_{N+1} = p(N)$. This involves an algorithm of complexity $\bar{m}N$. Call this array $S_g^{\omega}(0)$. Next we form the array $S_g^{\omega}(1)$ $(1 \le g \le N, 0 \le \omega \le \bar{m})$ using $p_1 = p(1)$, ... $p_N = p(N)$. Similarly we form the arrays $S_g^{\omega}(j)$ $(1 \le g \le N - j + 1, 0 \le \omega \le \bar{m})$ using $p_1 = p(j)$, ... $p_{N-j+1} = p(N)$. We stop at $j = N$. Clearly forming the arrays $S_g^{\omega}(j)$ requires

$O(\bar{m}N^2)$ computations. These will be stored for use in either the forward or backward (see Appendix) scheme. (The arrays will be computed with $\bar{m} = m_{max}$ of the main recursion.)

The *forward recursion* for calculating $\Gamma(\vec{s},\vec{t})$ can now be described as follows:

Let $0 = v_0, v_1, v_2, \ldots v_d = t_{N+1}$ be an increasing sequence which contains the s_i's and t_i's. Although it may be completely general, in this section this sequence is precisely the ordered merger of appropriate $\{s_i\}_1^N$ and $\{t_i\}_0^{N+1}$. In this case d would equal 2N+2 if the s_i's and t_i's are distinct. Some special other cases will also occur in the Appendix. Now for k=0, 1, 2, ... N; i = 0, 1, ... d; and $m \geq k$ we let

$$W_{ki}^m = \Pr\left\{ \begin{array}{l} \text{m arrivals in } (0,v_i] \text{ of which k are queuers, the } l\text{'th} \\ \text{arriving in } (s_l, \min\{t_l,v_i\}] \text{ for } l=1,2,\ldots k \text{ (with the } l\text{'th} \\ \text{automatically staying in queue until } t_l). \end{array} \middle| \begin{array}{l} \text{m arrivals} \\ \text{in } (0,t_{N+1}] \end{array} \right\}$$

$l_{ki} = \#\{t_l : t_l > v_{i-1}, s_l \leq v_{i-1}; l \leq k\}$

(l_{ki} is the maximum number of queuers among the first k who can arrive in $(v_{i-1}, v_i]$; it is also the number (among the first k) in queue just before v_i.)

$q(n) = 1 - p(n)$ = probability of "getting in queue" when a potential customer finds a queue of length n

$\Delta F_i = (F(v_i) - F(v_{i-1}))/(F(t_{N+1}) - F(0))$ = conditional probability of arriving in $(v_{i-1}, v_i]$ given arrival in $(0, t_{N+1}]$.

Pick some maximum value m_{max} for m. The recursion given by Theorem I is carried out in three nested loops

Do m = 1, m_{max}

Do k = 0 , min {m,N}

Do i = 1, v_d

with boundary conditions for W_{ki}^m given by $W_{0i}^0 = 1$, $W_{k0}^m = 0$ for $m > 0$.

Theorem I

$$W_{ki}^{m} = \begin{cases} \displaystyle\sum_{c=0}^{l_{ki}} \sum_{j=c}^{m-k+c} (\Delta F_i)^j \; W_{k-c\; i-1}^{m-j} \binom{m}{j} \left(\prod_{r=1}^{c} q(l_{ki}-r) \right) S_{c+1}^{j-c} \; (l_{ki} - c) & \text{For } k=0 \text{ or, if } s_k \le v_{i-1}, \text{ for } k > 0 \\ 0 & \text{For } k>0 \text{ if } s_k \ge v_i \end{cases}$$

Proof: If $k > 0$ and $s_k \ge v_i$ then clearly the first k queuing customers can not each arrive in $(0, v_i]$ and jointly arrive in the rectangle $(s_1, t_1] \times (s_2, t_2] \times \ldots \times (s_k, t_k]$. Hence $W_{ki}^{m} = 0$ in this case.

If $k = 0$ or $s_k \le v_{i-1}$ we may write

$$W_{ki}^{m} = \sum_{c=0}^{l_{ki}} \sum_{j=c}^{m-k+c} (\Delta F_i)^j \; W_{k-c\; i-1}^{m-j} \binom{m}{j} \sum_{l_1+l_2+\ldots+l_{c+1}=j-c} [p(l_{ki})]^{l_1} \; q(l_{ki}-1) \; [p(l_{ki}-1)]^{l_2} \; \cdots q(l_{ki}-c) \; [p(l_{ki}-c)]^{l_{c+1}}$$

This follows if we think of

- c as the number of arrivals in $(v_{i-1}, v_i]$ who become queuers
- j as the total number of arrivals in $(v_{i-1}, v_i]$
- $\binom{m}{j}$ as the number of ways to pick the particular set of j arrivals from m.
- $W_{k-c\; i-1}^{m-j}$ as the probability that the other m-j arrivals in $(0, v_{i-1}]$ behave accordingly
- The above inner-most Σ as the probability that the j arrivals in $(v_{i-1}, v_i]$ yield exactly c queuers.

The Theorem is now proven by factoring out the "q(n)" terms from the above innermost Σ and then noticing that $S_g^{\omega}(p_1, p_2, \ldots, p_g) = S_g^{\omega}(p_g, p_{g-1}, \ldots, p_1)$. (By convention the (empty) product of the "q(n)" terms is one for $c = 0$.)■

Let $\tilde{f}(m)$ be the conditional probability function for the number of arrivals in the congestion period (conditioned upon the fact that congestion starts at 0). Then

$$\Gamma(\vec{s},\vec{t}) = \lim_{m_{max} \to \infty} \sum_{N}^{m_{max}} \tilde{f}(m)\ W_{Nd}^{m}$$

so that $\Gamma(\vec{s},\vec{t})$ may be accurately approximated by summing the above to large enough m_{max}, assumed to be not larger than some fixed multiple of N. In this case the complexity for computing $\Gamma(\vec{s},\vec{t})$ is determined as follows: Since the products of the "q(n)" terms may be precalculated in $O(N^3)$ along with the S_g^{ω} (j) arrays, we need an additional $O(N^2)$ from the recursion in Theorem I times $O(N^3)$ from the Do loops for a total of $O(N^5)$ steps. For fixed k or τ, (3) or (4) may be evaluated in $0(N^5)$ by using the appropriate $\vec{s}$ sequence (i.e., $\vec{0}$, $\vec{s}^k$, or $\vec{s}^{\tau}$) with $\vec{t}$. Similarly for each k (5) may be computed by repeated recursion for i = 1, N in $0(N^6)$ and hence (6) can be evaluated in $0(N^7)$. Finally (7) may be likewise computed in $0(N^6)$ by brute force. If p(n)=1 beyond a fixed integer and we forbid c in Theorem I to exceed this integer, then the *forward recursion* for $\Gamma(\vec{s},\vec{t})$ requires only $0(N^4)$ steps and (3) - (7) require (with brute force) factors of N fewer.

A remark on the non-Poisson case is here in order: Even if f(m) is known it is not at all clear how to compute $\tilde{f}$(m) since the number of arrivals in the congestion period may depend on the number before congestion begins which in turn depends on the service distributions. Nevertheless the conditional distribution of the times of the arrivals in the congestion period has the simple form used in the algorithm. In the inhomogeneous Poisson case, however, the number of arrivals in the congestion period is independent of the number outside it and the conditional arrival probability function is

$$\tilde{f}(m) = \frac{\lambda^m [F(t_{N+1}) - F(0)]^m}{m!} \exp\{-\lambda[F(t_{N+1}) - F(0)]\}$$

where $\lambda F'(x)$ is the time varying arrival rate and $f(m) = \frac{\lambda^m}{m!} \exp\{-\lambda\}$. In the non-Poisson case we will show how to estimate $\tilde{f}$(m) in the next section.

To evaluate (6) and (7) conditioned on the entire transactional period*, instead of one congestion period, let L^i, D_{τ}^i be the average experienced queue length, average probability of "delay < τ " computed for the congestion periods i = 1, 2, ... P

* The transactional period, which consists of the congestion periods and the intervals between congestion, might be a proper subinterval of Ω.

with each involving N_i customers. If N_0 is the number of customers in the transactional period who arrive outside of congestion (who are not delayed), then the appropriate average queue length and average probability of "delay < τ " are

$$\bar{L} = \frac{N_1 L^1 + N_2 L^2 + \ldots + N_p L^P}{N_0 + N_1 + \ldots N_P}$$

$$\bar{D}_\tau = \frac{N_1 D_\tau^1 + N_2 D_\tau^2 + \ldots + D_\tau^P}{N_0 + N_1 + \ldots N_P}$$

IV. THE POSTERIOR LIKELIHOOD FUNCTION

A. Known Balking Function, Unknown (not necessarily Poisson) Arrival Process

In most practical non-Poisson problems, f or $\tilde{f}$ is unknown and must be estimated. One maximum likelihood estimater for $\tilde{f}$ is given as follows: Compute the array $\bar{W}_{ki}^m$ based on $\vec{s} = \vec{0}$ and $\vec{t}$. Let $\hat{m}$ be an m which maximizes $\bar{W}_{Nd}^m$ for $m \geq N$. Then use

$$\tilde{f}(m) = \begin{cases} 1 & m = \hat{m} \\ 0 & m \neq \hat{m} \end{cases}$$

$\hat{m}$ is the most likely number of potential customers to have yielded congestion given the transactional data in the congestion period $(0, t_{N+1}]$. This method provides a different arrival estimate for each congestion period. Let $(w_i, z_i]$ be the congestion intervals, $(x_j, y_j]$ the intervals in the partition of the uncongested time domain generated by the arrival times when there is no congestion* , and ${}^i\bar{W}$ be the array corresponding to $\vec{0}, \vec{t}$ for $(w_i, z_i]$. The associated nonparametric likelihood function, valid for any O-process, is

$$L(m_1, m_2, \ldots m_P) = \prod_{i=1}^{P} {}^i W_{N i d i}^{m_i}$$

* Note that each w_i is such an arrival time.

If $\hat{m}_i$ denotes an m which maximizes ${}^{i}W^{m}_{N_i d_i}$, then we call $(\hat{m}_1, \hat{m}_2, \ldots \hat{m}_P)$ a nonparametric maximum likelihood arrival estimate. The associated estimates of arrival rate (expected potential customers/length of transaction interval) are the non-congestion average

$$\hat{\lambda}_1 = N_o / \Sigma(y_j - x_j),$$

the congestion average

$$\hat{\lambda}_2 = \sum_{i=1}^{P} \hat{m}_i \; / \sum_{i=1}^{P} (z_i - w_i) = \frac{\text{potential arrivals in congestion}}{\text{time of congestion}}$$

and the combined average

$$\hat{\lambda}_c = \left(\sum_{i=1}^{P} \hat{m}_i + N_o \right) / \text{ length of transaction interval.}$$

B. Poisson arrivals with unknown rate and known balking function

In the inhomogeneous Poisson case with unknown λ (but known $p(n)$ and $F(x)$) we derive the maximum aposteriori probability density for λ using all of the transactional data. If $g(\lambda)$ is a prior density on λ, by the Bayes principle the posterior density $g_1(\lambda)$, based on the non-congestion data, is proportional to the product of $g(\lambda)$ and the density of the uncongested arrival times given λ and the service times. Some elementary probability and calculus yields:

$$(8a)\quad g_1(\lambda) \propto g(\lambda) \cdot \left[\prod_{j=1}^{N_0} [F'(y_j)] \right] \cdot \lambda^{N_0} \cdot \exp\left\{ -\lambda \sum_{1}^{N_0} [F(y_j) - F(x_j)] \right\};$$

Similarly the posterior density $g_2(\lambda)$, based on the congestion data, is proportional to $g(\lambda)$ times the product of the congestion period probabilities

$$(8b)\quad g_2(\lambda) \propto g(\lambda) \cdot \prod_{i=1}^{P} \left(\sum_{m \geq N_i} \frac{\exp\{-\lambda[F(z_i)-F(w_i)]\}(\lambda[F(z_i)-F(w_i)])^m \; {}^{i}\bar{W}^{m}_{N_i d_i}}{m!} \right);$$

And finally the posterior based on all data, $g_T(\lambda)$, is

(8) $g_T(\lambda) \propto g_1(\lambda)\, g_2(\lambda)/g(\lambda)$

$g_T(\lambda)$ may be computed numerically and the maximizing $\hat{\lambda}$ is the maximum a posteriori probability (M.A.P.) estimate. For the improper prior, $g(\lambda)$ = constant, $\hat{\lambda}$ is the maximum likelihood estimate. Using numerical integration the mean square Bayes estimate (the mean of $g_T(\lambda)$) and Bayesian confidence regions may also be calculated. (In practice the index m in (8b) will only vary from N_i to m_{max} for the i'th congestion period.)

C. Poisson Arrivals with unknown λ and unknown balking function

For the case of unknown λ and unknown $p(n) = p(n,\alpha)$ (i.e. unknown α) (8) may be rewritten as

$$(9)\quad g_T(\lambda,\alpha) \propto g(\lambda,\alpha) \cdot \left[\prod_{j=1}^{N_0} [F'(y_j)]\right] \cdot \lambda^{N_0} \cdot \exp\left\{-\lambda \sum_{1}^{N_0} [F(y_j) - F(x_j)]\right\} \cdot \prod_{i=1}^{P} c_i \quad (\lambda,\alpha)$$

where $g(\lambda,\alpha)$ = prior density of λ,α; $g_T(\lambda,\alpha)$ = posterior density of λ,α; and $c_i(\lambda,\alpha)$ is the ith factor in the product in (8b) where the ${}^{i}\bar{W}$ array is computed for each α by using $p(n) = p(n;\alpha)$. $c_i(\lambda,\alpha)$ is just the probability of making period *i* a congestion period given λ,α, and the times defining the period. Again numerical methods lead to M.A.P. estimates $\hat{\lambda},\hat{\alpha}$ (maximum likelihood when taking $g(\lambda,\alpha)$ = constant) and various other Bayesian estimates and confidence regions. In the special case $g(\lambda,\alpha) = \delta(\lambda-\lambda_o)\, g(\alpha)$ we get the posterior density for inferring α with known $\lambda=\lambda_o$. This is discussed in Section VII.

One particular model for the functional dependence of λ and α in (9), which is examined for the constant service time example in Sec. VIII, is the following *shift model*: $c_i(\lambda,\alpha) = h_i(\lambda - d(\alpha))$ where each h_i is unimodal with maximum value $h_i(0)$. The motivation for this is that $c_i(\lambda,\alpha)$ would be, to a first approximation, a unimodal function of some linear combination of λ and a reparametrization of α (w.l.o.g. of form λ-$d(\alpha)$). In such a case one reaches the following interesting conclusion: If $p(n)$ is known (α known) then the maximum likelihood estimate $\hat{\lambda}$ will depend on both non-congestion data and congestion data. On the other hand if α is unknown and the range of $d(\alpha)$ contains the maximum likelihood estimate $\hat{\lambda}'$, based on non-congestion data (i.e. setting $g(\lambda,\alpha)$ =

constant and ignoring the last product in (9)), then the maximum likelihood estimate of λ,α based on all data is $\hat{\lambda} = \hat{\lambda}'$, $\hat{\alpha}$ s.t. $d(\hat{\alpha}) = \hat{\lambda}'$. Hence the estimated customer arrival rate will not depend on congestion data. Although the shift model may be inaccurate the above conclusion that $\hat{\lambda} = \hat{\lambda}'$ may still hold. This we call the *shift conclusion* and will be elucidated further in Sec. VIII. It should be a subject of future research to determine under what conditions the shift conclusion is approximately valid.

Finally, if $p(n,\alpha)$ is one beyond a fixed integer n, then (9) may be calculated in

$$\sum_{1}^{P} O(N_i^4) + L \sum_{1}^{P} O(N_i) \text{ for fixed } \alpha \text{ and L values of } \lambda.$$

V. THE EXPONENTIAL BALKING MODEL; AN ILLUSTRATIVE DATA SET

Although our algorithms will compute congestion probabilities and estimate queue performance for any balking function, there is a very natural parametric family which should be appropriate in many applications. This we call the exponential balking family and write as follows:

$$p(n;\alpha) = \begin{cases} 0 & n = 0 \\ 1 - \exp\{-\frac{1}{2} n \alpha\} & 1 \le n < r \\ 1 & n \ge r \end{cases}$$

where $\alpha \ge 0$ is the balking parameter and r is the waiting room size. The rationale for assuming this family is as follows: Potential customers balk if they cannot gain access to the waiting room. Otherwise an arriving potential customer perceives the size of the queue sequentially, balking with probability p after registering each waiting customer visually. The probability of not balking after registering all n waiting customers is then $(1-p)^n$. Setting $p = 1 - \exp\{-\frac{1}{2}\alpha\}$ yields the above $p(n;\alpha)$. Since the waiting room size is usually known the inference problem is to determine α. Of course there are many other families and our methods will apply to those.

To illustrate our methods we constructed a small data set. Practical application to large data sets often reduces to breaking the data into smaller groups corresponding to varying conditions and then obtaining different estimates of some of the parameters for each subset. For instance a chain of fast food restaurants in a

given region may have differing physical layouts for their stores. First the arrival rate may be estimated using, say, all non-congestion transactional data; then balking parameters would be separately estimated for each layout by using only the congestion data for that layout assuming the arrival rate to equal that previously estimated. This will be illustrated in Section VII.

We considered a single server facility with constant service time of .01 hr. Constant service times in the experiment should keep noise at a minimum so that estimates based on a small sample may be best examined. Potential customers arrived during a one hour time interval Ω. The balking probabilities were assumed to be $p(0)=0$, $p(1) = 1 - \exp\{-.5\alpha\}$, $p(2) = 1 - \exp\{-1.0\alpha\}$, $p(3) = 1 - \exp\{-1.5\alpha\}$, $p(4) = 1 - \exp\{-2.0\alpha\}$, $p(5) = p(6) = \ldots = 1$. Thus the waiting room size, r, was 5. We took $\alpha=1.0$. We independently let 150 potential customers arrive, each with the uniform distribution in Ω. Balking was applied via a random number generator for Bernouli events to the ordered arrivals taking queue statistics into account. We recorded the transactional data until the end of the last congestion period lying totally within Ω. We then applied techniques for both the non-Poisson and Poisson case to this sample. We used $m_{max} = 3N_i$ in the Q.I.E. for the i'th congestion period. The results are summarized in Table 1.

VI. INFERRING ARRIVAL RATE AND QUEUE PERFORMANCE WITH KNOWN BALKING BEHAVIOR

Here we fixed $\alpha = 1$ in all calculations. This corresponds to analyzing a limited data set for a facility for which we can predict balking behavior. See Table 1.

Note that the non-parametric estimate of arrival rate

$$\hat{\lambda}_2 = \sum_1^P \hat{m}_i \Big/ \sum_1^P |z_i - w_i| \ ,$$

based on the congestion periods, is much more accurate for this sample than the non-parametric estimate

$$\hat{\lambda}_1 = N_o \Big/ \sum_1^{No} |y_j - x_j| \ ,$$

based on the intervals outside congestion. We also used the nonparametric maximum likelihood arrival estimate to estimate average experienced queue length. See Table 1. We also examined the posterior densities in the homogeneous Poisson analysis: Let $g(\lambda)$ be a constant improper prior. Then the posterior density based on congestion data alone has the form

$$g_2(\lambda) \propto \prod_{i=1}^{P} \left(\sum_{m \geq N_i} \frac{\exp\{-\lambda(z_i - w_i)\}[\lambda(z_i - w_i)]^m \; {}_i\bar{W}^m_{N_i d_i}}{m!} \right)$$

while the posterior based only on the intervals outside congestion has the form

$$g_1(\lambda) \propto \lambda^{N_o} \exp\left\{-\lambda \sum_{1}^{N_o} (y_j - x_j)\right\}.$$

Finally

$$g_T(\lambda) \propto g_1(\lambda)\, g_2(\lambda) \; .$$

$g_1(\lambda)$ is a gamma and has a standard deviation $\sqrt{12}/(.094) = 36.9$ and mean $12/(.094) = 127.7$ for our data. (Note that the mean of $g_1(\lambda)$ differs from $\hat{\lambda}_1$ by a one in the numerator; see Table 1.) If we expand each factor in $g_2(\lambda)$ (in the order of decreasing $\bar{W}$)starting at $m = \hat{m}_i$ we get a leading term in $g_2(\lambda)$ of the form

$$\propto \lambda^{\sum_{1}^{P} \hat{m}_i} \exp\left\{-\lambda \sum_{1}^{P} (z_i - w_i)\right\}$$

If this dominated, the mean and standard deviation for our sample would be $109/(.77) = 142.6$ and $\sqrt{109}/(.77) = 13.6$. (Again note the mean of the leading term differs from $\hat{\lambda}_2$ by a one in the numerator.) This term is not totally dominant as is seen in Fig. 1 but the standard deviation for $g_2(\lambda)$ is half that of $g_1(\lambda)$. Hence the congestion data is four times as informative as the non-congestion data! (Using inverse variance as a measure of information content.) Finally note that $g_T(\lambda)$ is much closer to g_2 than g_1.

VII. INFERRING BALKING PROBABILITIES WITH KNOWN ARRIVAL RATE

Here we assume $\lambda = 150$ in the Poisson model and we want to estimate α. This might be appropriate if non-congestion data from many facilities gives an accurate arrival rate whereas a particular given facility has a unique waiting room layout necessitating use of only its data to determine its balking parameter.

In the first Bayesian analysis we let $g(\alpha)$ be a uniform prior for $0 \leq \alpha \leq 1.6$. This corresponds to assuming that balking is at most moderate but could be negligible. The highest balking probability when one person is in queue is .55. See Table 2 and Figures 2 and 3. Note from Figure 3 that the Bayes mean is reasonably accurate and that the upper 90% confidence region is rather narrow. (Assuming α was originally a uniform random variable on [0,1.6], 90% of the time the balking probabilities are in the shaded region.)

For the second Bayesian analysis we used $p = 1 - \exp\{-\frac{1}{2}\alpha\} = p(1;\alpha)$ as parameter, taking the prior to be uniform on [0,1]. Then the posterior is proportional to $g_T(150, -21n(1-p))$ which is shown in Figure 4 together with the 10% lower confidence limit. Note that there is about a 90% chance that $p > \frac{1}{3}$.

VIII. INFERENCE FROM A SAMPLE WHEN BOTH λ AND α ARE UNKNOWN

For the Poisson case assume $g(\lambda,\alpha)$ = constant. First we show $c_5(\lambda,\alpha)$ and $c_6(\lambda,\alpha)$ (corresponding to congestion periods 5 and 6 of Table 1) for various values of α in Fig. 5. The *shift model* appears to be a somewhat course approximation, but the shift conclusion nearly holds. Indeed when we inspect the maximum likelihood estimates $\hat{\lambda},\hat{\alpha}$ we see that $\hat{\lambda}$ is very close to the mode $\hat{\lambda}'$ of $g_1(\lambda)$ (which is the likelihood of λ based on non-congestion data.) See Fig. 6. Since the standard deviation of $\hat{\lambda}'$ will vary inversely with the square root of N_o, we conclude that large samples are needed to estimate both λ and α. λ should be estimated from the non-congestion data and then the methods of VII should be implemented to infer α assuming the estimated value for λ.

IX. ACKNOWLEDGMENTS

Special thanks are due Mr. James MacDougall for programming the algorithm and Prof. Richard Larson for his advice and encouragement during this project.

APPENDIX

MORE EFFICIENT EVALUATION OF QUEUE PERFORMANCE MEASURES

This section describes computational improvements of the algorithms for estimating performance measures.

The *backward recursion* for calculating $\Gamma(\vec{s},\vec{t})$ is described as follows: As in the forward recursion we let $0 = v_0, v_1, \ldots v_d = t_{N+1}$ be a refinement of the s_i's and t_i's. For $k = 0, 1, 2, \ldots N$; $i = 0, 1, \ldots d$; and $m \geq k$ we let

$$H_{ki}^{m} = \Pr\left\{ \begin{array}{l} \text{m arrivals in } (v_i, t_{N+1}] \text{ of which k are queuers arriving in } (\max\{s_l, , v_i\}, t_l] \\ \text{for } l\text{=N-k+1,...N (with each queuer automatically staying in queue until } t_l). \end{array} \middle| \begin{array}{l} \text{m arrivals in } (0, t_{N+1}]; \text{ N-k other queuers assumed to arrive in } (0, v_i] \text{ (with each leaving at } t_l \text{ for } l\text{=1,2,... N-k)} \end{array} \right\}$$

q(n), ΔF_i be as in the forward recursion

$\tau_{ki} = \#\{s_l : s_l < v_{i+1} \text{ and } t_l \geq v_{i+1} ; l = N\text{-}k+1, \ldots N\}$

(τ_{ki} = maximum number of the k queuers among the m arrivals who may arrive in $(v_i, v_{i+1}]$)

$\rho_{ki} = \#\{t_l : t_l \geq v_{i+1} ; l = 1, 2, \ldots N\text{-}k\}$

(ρ_{ki} = number of the N - k conditional queuers who leave at or after v_{i+1} (who arrived by the conditional assumption before or at v_i))

$\tau_{ki} + \rho_{ki}$ = maximum number (conditional and unconditional) in queue just before v_{i+1}

Pick some value m_{max} for m. The backward recursion (Theorem II) proceeds in three nested loops

Do m = 1, m_{max}

Do k = 0, min {m,N}

Do $i = d - 1, 0$

with boundary conditions for H^m_{ki} given by $H^o_{oi} = 1$, $H^m_{kd} = 0$ for $m > 0$.

Theorem II

$$H^m_{ki} = \begin{cases} \displaystyle\sum_{c=0}^{\tau_{ki}} \sum_{j=c}^{m-k+c} (\Delta F_{i+1})^j \, H^{m-j}_{k-c\, i+1} \binom{m}{j} \left(\prod_{r=1}^{c} q(\tau_{ki}+\rho_{ki}-r) \right) S^{j-c}_{c+1} \quad (\tau_{ki} + \rho_{ki} - c) & \text{If } t_{N-k+1} > v_i \\ 0 & \text{If } t_{N-k+1} \le v_i \end{cases}$$

Proof: Similar to that of Theorem I.■

We now compute the average probability of "delay < τ ". Perform the backward recursion with $\vec{s} = \vec{0}$ and $\vec{t}$ given by the congestion data. Let the v_i's be the ordered merger of $\vec{s}^{\tau}, \vec{t}$ and 0. Call this array $\tilde{H}^m_{ki}$. Perform the forward recursion with $\vec{s} = \vec{0}$ and $\vec{t}$ and the same v_i's as in the backward recursion. Call this array $\tilde{W}^m_{ki}$. Finally define i(j) as the value of i such that $v_i = \bar{s}^{\tau}_j$ (v_i=max $\{t_j-\tau, 0\}$.) Now the average probability of {"delay<τ " and maintaining congestion | m arrivals} is given by

$$D^m_N = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{j} \sum_{\substack{k-1 \le a \le m \\ m-a \ge N-k+1}} \binom{m}{a} \tilde{W}^{a}_{k-1\, i(j)} \tilde{H}^{m-a}_{N-k+1\, i(j)}$$

To justify this it is enough to show that the inner double sum represents the probability that customer j is delayed < τ and that congestion is maintained given m arrivals. Now this event is the disjoint union of events of the form "a arrive in $(0, v_{i(j)}]$ of which k-1 are queuers ($k \le j$), m-a arrive in $(v_{i(j)}, t_{N+1}]$ of which N-k+1 are queuers, and congestion is maintained." The inner double sum is just the sum of the probabilities of these disjoint events and hence the result follows.

Now (7) is given (accurately approximated) by

$$D_\tau = \sum_{m=N}^{m_{max}} \tilde{f}(m)\ D_N^m \ \Big/ \ \sum_{m=N}^{m_{max}} \tilde{f}(m)\ \tilde{W}_{Nd}^m .$$

Assuming m_{max} is not larger than some fixed multiple of N, the arrays $\tilde{H}_{ki}^m$ and $\tilde{W}_{ki}^m$ are computed in $O(N^5)$ and then the D_N^m $(N \le m \le m_{max})$ are generated in $O(N^4)$. Hence (7) is evaluated in $O(N^5)$ which is N times faster than brute force. If $p(n)$ is one beyond a fixed integer the arrays and (7) are evaluated in $O(N^4)$ (by restricting c in Theorem II).

To get the average experienced queue length we first calculate β_k^m = Pr {a randomly chosen customer finds queue length < k and congestion is maintained | m arrivals}. Perform the backward recursion with $\vec{s} = \vec{0}$ and $\vec{t}$ given by the congestion data. Let the v_i's be the t_i's and 0. Call this array $\tilde{\tilde{W}}_{ki}^m$. Perform the forward recursion with the same parameters as the backward recursion. Call the array $\tilde{\tilde{H}}_{ki}^m$. Define i'(j) as the value of i such that $v_i = s_j^k$. ($v_i = t_{max\{j-k,0\}}$ so $i'(j) = max\ \{j-k,0\}$.) By an argument nearly identical to that for D_N^m

$$\beta_k^m = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{j} \sum_{\substack{k-1 \le a \le m \\ m-a \ge N-k+1}}^{N} \binom{m}{a} \tilde{\tilde{W}}_{k-1\ i'(j)}^{\ a} \tilde{\tilde{H}}_{N-k+1\ i'(j)}^{\ m-a}$$

So β_k is given (to a high degree of accuracy) by

$$\beta_k = \sum_{m=N}^{m_{max}} \tilde{f}(m)\ \beta_k^m \ \Big/ \ \sum_{m=N}^{m_{max}} \tilde{f}(m)\ \tilde{\tilde{W}}_{Nd}^m$$

If m_{max} is not larger than a fixed multiple on N, the β_k may be calculated for k=1, ... N in $0(N^4)$ from the $\tilde{\tilde{W}}$ and $\tilde{\tilde{H}}$, which are both computed in $0(N^5)$. Clearly (6) can now be obtained in $0(N^5)$ which is N^2 times faster than brute force. If p(n) is one beyond a fixed integer the arrays and (6) are computed in $O(N^4)$.

REFERENCES

[1] R.C. Larson, 1990. "The Queue Inference Engine: Deducing Queue Statistics from Transactional Data," *Management Science* 36:5, 586-601.

[2] R.C. Larson, 1991. "The Queue Inference Engine: Addendum." *Management Science* 37:8, 1062.

[3] Bertsimas, Dimitris and Les D. Servi, 1991 "Deducing Queuing from Transactional Data: The Queue Inference Engine Revisited," to appear in *Operations Research.*

[4] Daley, D.J. and Les D. Servi, 1992. "Exploiting Markov Chains to Infer Queue Length from Transactional Data," *Journal of Applied Probability,* 29.

[5] Jones, L.K. and R.C. Larson. "Efficient Computation of Probabilities of Events Described by Order Statistics and Applications to Queue Inference," ORSA Journal of Computation (in press).

[6] Servi, L.D. and D.J. Daley, "A Two Point Markov Chain Boundary Value Problem," to appear in *Advances in Applied Probability.*

Table 1. Q.I.E. with balking applied to transactional data.

Interval Outside Congestion	Number of Arrivals	Length of Interval
1	1	.002
2	1	.015
3	1	.005
4	1	.013
5	1	.013
6	1	.001
7	1	.009
8	1	.015
9	1	.0005
10	1	.014
11	1	.007
Totals	11	.094

Estimate 1 of Arrivals/hr. = $\hat{\lambda}_1$ = Total arrivals/total length = 11/.094 = 117

Congestion Period	N_i	$\hat{m}_i$	Length of Congestion Period	Average experienced queue length based on maximum likelihood $\hat{m}_i$
1	0	0	.01	0.00
2	2	2	.03	0.12
3	3	3	.04	0.20
4	0	0	.01	0.00
5	11	17	.12	0.72
6	17	30	.18	0.94
7	0	0	.01	0.00
8	22	41	.23	1.03
9	2	2	.03	0.12
10	9	13	.10	0.62
11	0	0	.01	0.00
Totals	--	108	.77	---

Estimate 2 of Arrivals/hr. = $\hat{\lambda}_2$ = Total potential arrivals/total length = 108/.77 = 140

Average queue length = $\bar{L}$ = .69

$\hat{\lambda}_C$ = combined average = Combined potential arrivals/combined length = 119/.864 = 138

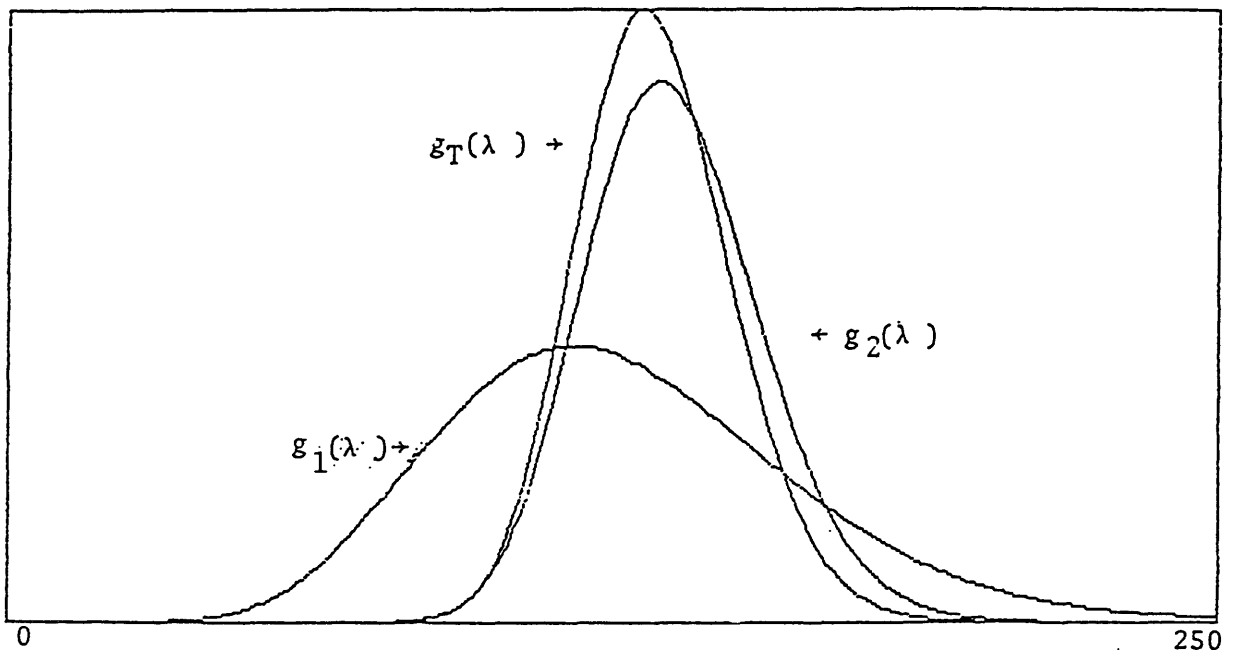
Table 2

Balking Probability Estimates with Known λ

True	0.0	0.39	0.63	0.78	0.86	1.0
Bayes Mean	0.0	0.44	0.69	0.82	0.90	1.0
Upper Bayes Limit	0.0	0.55	0.80	0.91	0.96	1.0
Lower Bayes Limit	0.0	0.0	0.0	0.0	0.0	1.0
One-Sided 10% confid.1.	0.0	0.31	0.53	0.68	0.78	1.0
	0	1	2	3	4	≥ 5

STATE OF SYSTEM

FIGURE 1.



MOMENTS

	mean	standard deviation	mode
$g_2(\lambda)$	137.2	18.3	134.5
$g_T(\lambda)$	133.2	16.2	131.0
$g_1(\lambda)$	127.7	36.9	116.5

Figure 2. $g_T(\lambda, \alpha)$ **for** $\lambda = 150$

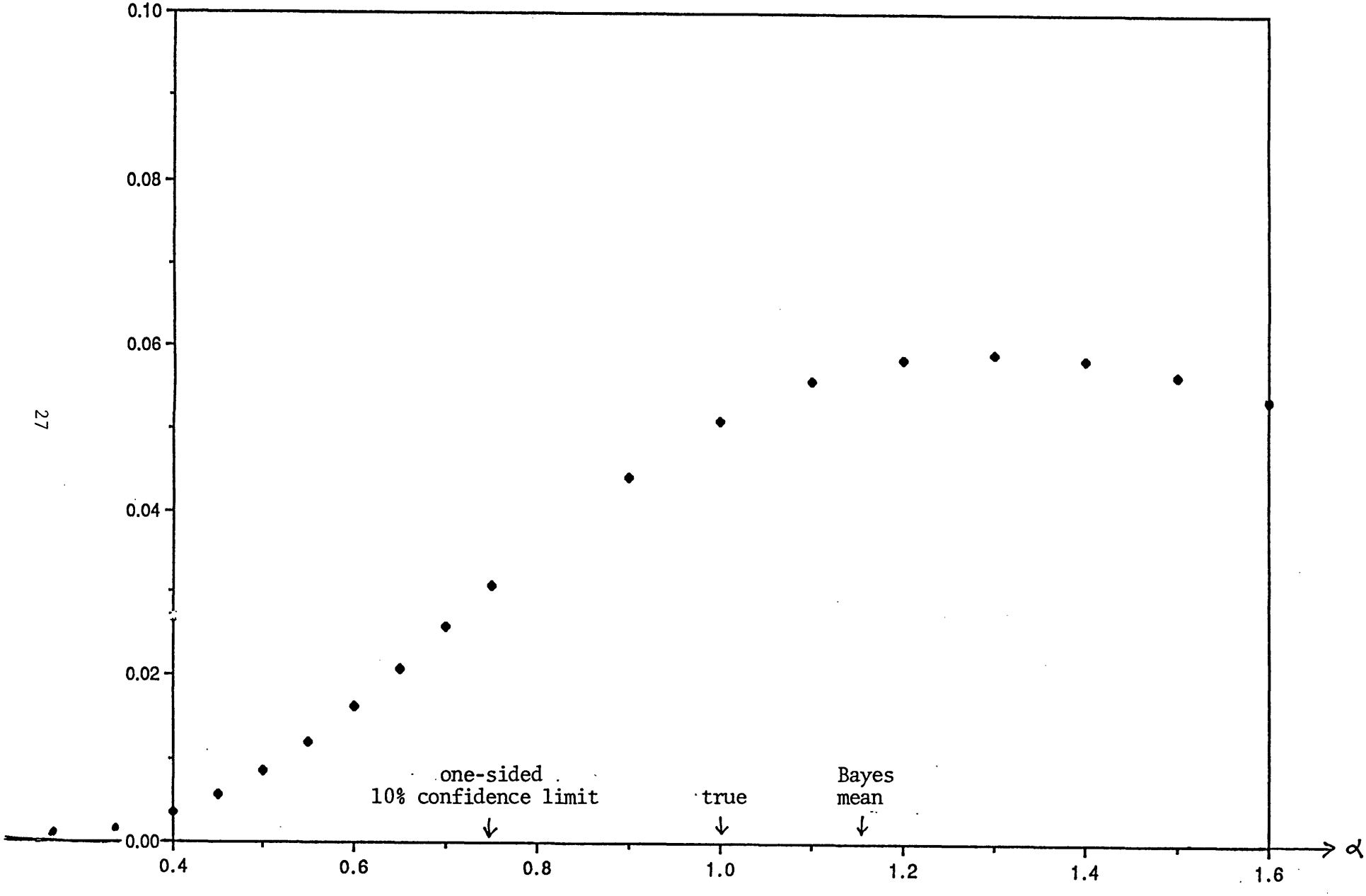


FIGURE 3

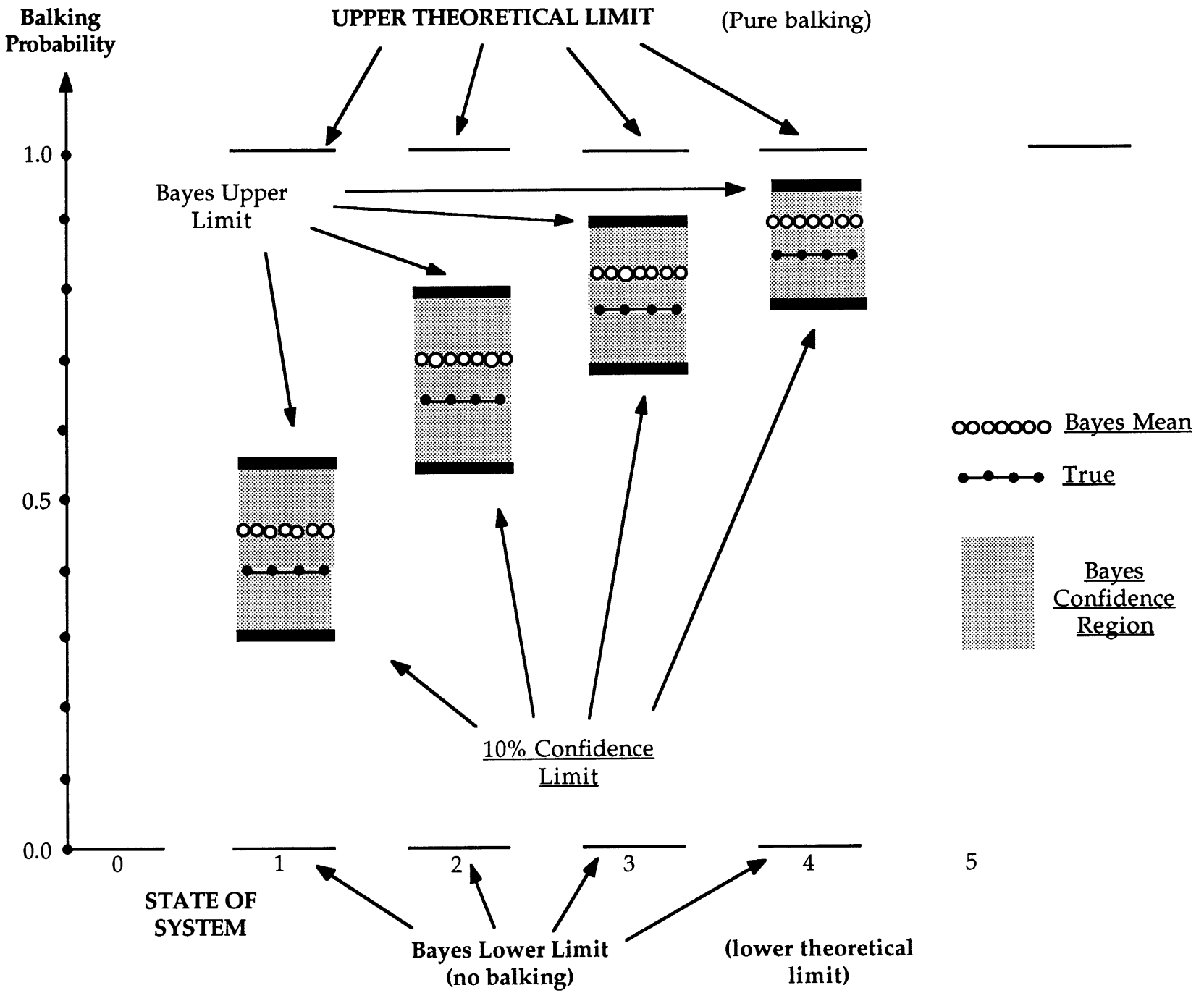
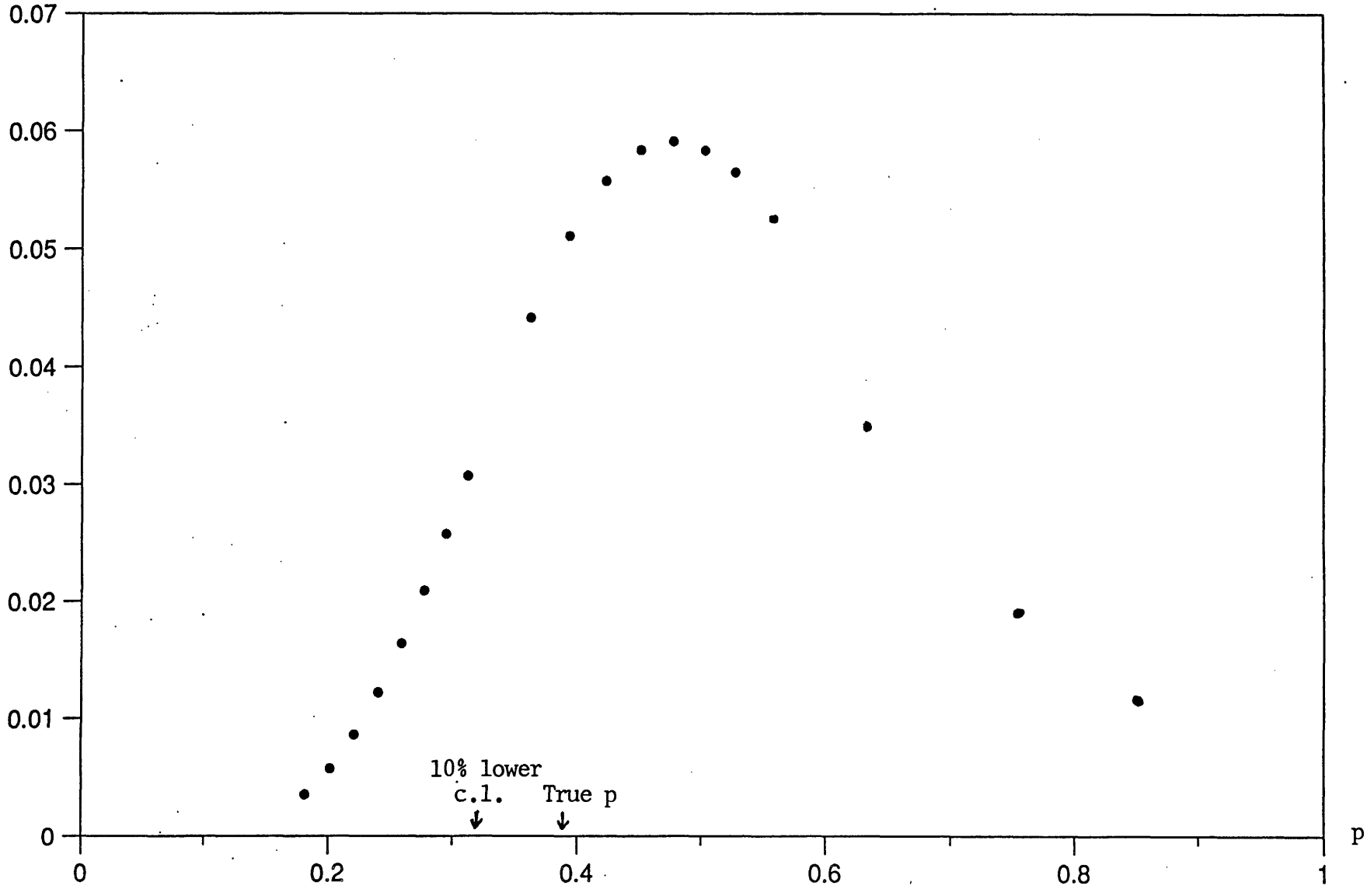

Balking Probability
UPPER THEORETICAL LIMIT
(Pure balking)
1.0
Bayes Upper Limit
0.5
Bayes Mean
True
Bayes Confidence Region
10% Confidence Limit
0.0
0
1
2
3
4
5
STATE OF SYSTEM
Bayes Lower Limit (no balking)
(lower theoretical limit)

Figure 4. $g_T(\lambda; -2\ln(1-p))$ for $\lambda = 150$



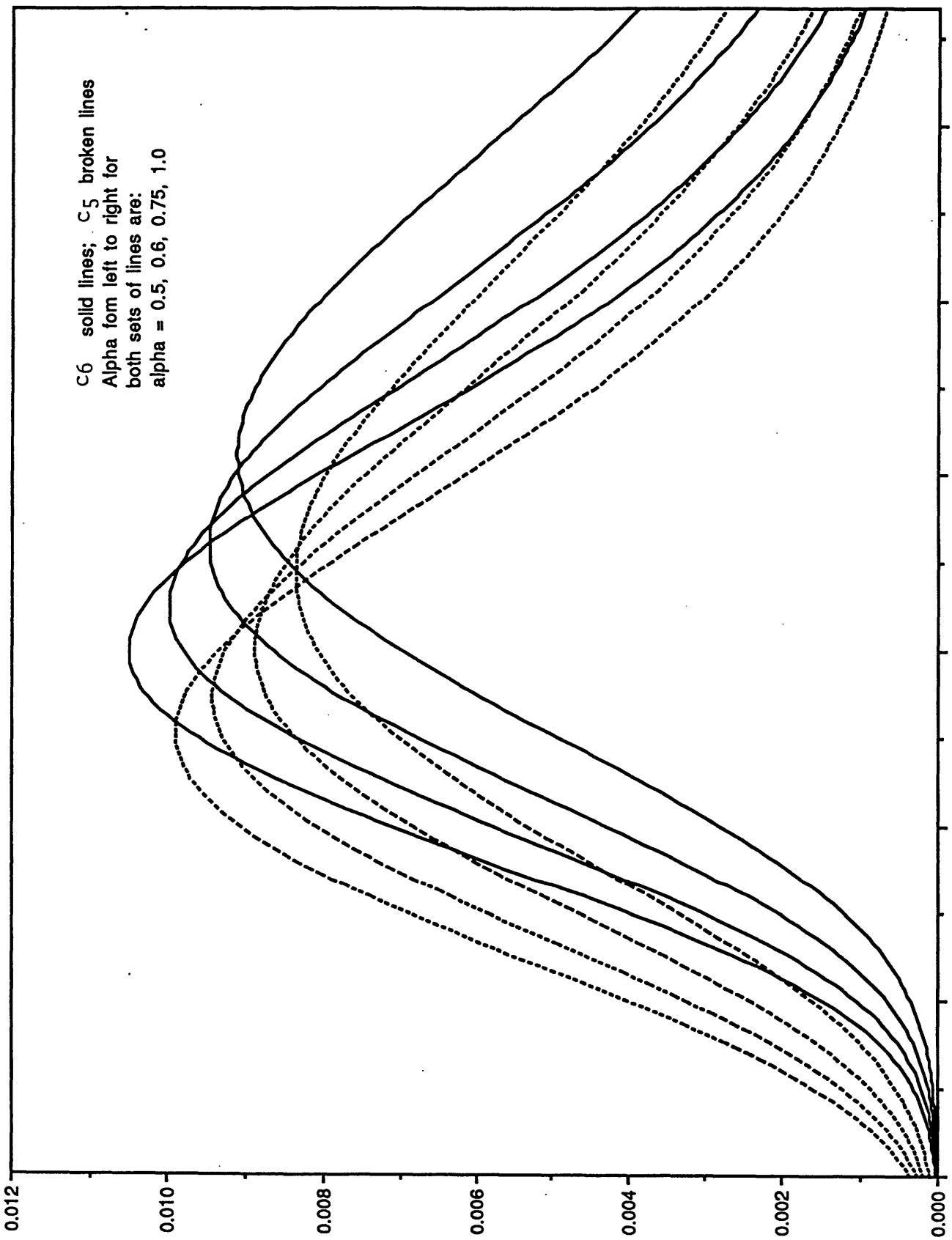


Figure 5. $c_5(\lambda, \alpha)$ and $c_6(\lambda, \alpha)$

Figure 6. $g_T(\lambda, \alpha)$ for various α