QUEUE INFERENCE ENGINE

Contributed by Richard C. Larson *Massachusetts Institute of Technology , Cambridge*

Imagine receiving your monthly bank statement and with it is your *personal probability distribution* of the times you spend waiting in bank queues. The queues could include both those involving human tellers and automatic teller machines (ATMs). With the technology of the Queue Inference Engine (QIE) such an innovation is now well within the realm of possibility.

BACKGROUND, MOTIVATION AND OVERVIEW

The idea of QIE was born in the late 1980s as a result of M.I.T.-based queueing research for Bay Banks, an eastern Massachusetts bank, under the auspices of a grant from the National Science Foundation. Bay-Banks had provided a large sample of transactional data from three of their ATM sites. Their question was, "Which, if any, of these sites is 'too congested' from a queueing point of view, thereby requiring additional ATM capacity at the site?" The transactional data consisted of the times of each ATM transaction by each customer over a period of up to a month.

The first approach to this problem was traditional: estimate arrival rates and service times from the data and then apply well known (steady state) queueing models, such as Erlang's results, or the M/G/1 model, etc. Examining the data set, it was realized that a substantial portion of the "sample path" of the queue had been preserved in the data set. That is, the data set contained a large subset of the information one would have if one "tracked" the actual queue with "clipboard and stopwatch." For instance, one could identify which customers had been delayed in queue (rather then enter service immediately) by noting the "signature" of a queued customer: a back to back service completion and service initiation at the same ATM, during a time when all N ATM's are busy with customers. The customer entering service in such a back-to-back situation was, with probability near one, delayed in queue. Moreover, by following this signature over time-adjacent customers, one could identify the entire set of customers who were delayed in queue during a single *congestion period*, a continuous period of time during which all N servers are continuously busy (excepting the small intervals during which a customer whose service is completed departs and the new [queued] customer enters service). We explored further the information content of the data set to see if it contained additional queue-related information.

Surprisingly, the partial information in the data set allowed a wide variety of queueing measures for each congestion period to be computed efficiently. Assuming Poisson arrivals, these measures include mean queue delay, mean queue length, probability distribution of the queue length and even the transient mean queue length over the course of the congestion period. Later research extended these first results in a number of important directions.

In this article, the focus is four-fold: (1) to illustrate the types of physical situations in which the QIE can be applied; (2) to describe one of three alternative analytical approaches to obtaining QIE results; (3) to guide the reader through the emerging literature in this new and exciting field; and (4) to discuss briefly several implementation experiences.

ILLUSTRATIVE QUEUE INFERENCING PROBLEMS

Retail Sales — With most "human server retail service systems," one has to collect the transactional data either from a modern POS (Point Of Sale) computer system that does the time marking or from some type of customer sensing device (e.g., pressure sensitive mats, infrared or ultrasonic sensors). For an ATM, the transactional data are recorded automatically, by time marking the moment that a customer inserts a bank card (corresponding to service initiation) and the moment that the ATM ejects the card (corresponding to service completion). The queue statistics generated by the QIE for ATMs may be used by bank managers to monitor the use of ATM sites, thereby providing an accurate method of identifying those sites requiring additional (or fewer) machines. With human servers in retail sales, at banks, post offices, fast food restaurants, etc., the manager would most likely use the results to (1) monitor service levels throughout the day and week, to assure that queue delays are within prescribed quality limits, and (2) to schedule servers optimally over the course of a day and week.

Invisible Queues in Telecommunication Systems — During periods of congestion, many finite capacity telecommunications systems have invisible queues of customers outside the system continuously trying to gain access to it. One example is a k-channel land mobile radio system. Whenever all k channels are simultaneously in use, potential users having a message to transmit (often in the field, in vehicles) continuously monitor channel use and at-tempt to acquire a channel as soon as any one of the current k communications is completed. If at any given time t there are $n(t)$ such potential users awaiting a channel, they constitute a spatially dispersed invisible queue, a queue in which one of the waiting customers enters service very shortly after another customer completes service. This queue can grow in size due to the Poisson arrivals of new potential users desiring

channel access. The user entering service next is the one who successfully "locks in" the channel very shortly after termination of a previous message. Service discipline is most likely not first come first served (FCFS). Within the context of the QIE the customer transaction times are the moments of gaining channel access (service initiation) and message termination (service completion). These times can be routinely monitored and recorded by technology, and thus the QIE can be used to deduce queueing behavior. The same argument, perhaps with minor modifications, can be applied to other telecommunications systems, including phone systems from airplanes, mobile cellular telephone systems, standard telephone systems and various digital communications networks.

USING ORDER STATISTICS TO DERIVE QIE PERFORMANCE MEASURES

The analysis of the queue inferencing problem is rooted in order statistics. Suppose we consider a homogeneous Poisson process with rate parameter $\lambda > 0$. Over a fixed time interval $[0,T]$, we are told that precisely N Poisson events (e.g., "queue system arrivals") occur. The N *ordered* arrival times are $0 \le X_{(1)} \le X_{(2)} \le \cdots \le X_{(N)} \le T$ (by implication $X_{(N+1)} > T$). The N *unordered* arrival times are $X_1, X_2, \ldots, X_N, 0 \le X_i \le T (i = 1, 2, \ldots, N)$. Since the Poisson process is time homogeneous, it is well known that the $\{X_i\}$ are independent and uniformly distributed over $[0, T]$. If the Poisson process is non-homogeneous, that is, having time varying rate parameter $\lambda(t)$, then the N unordered arrival times are independent identically distributed (iid) over $[0, T]$, with a pdf (probability density function) proportional to $\lambda(t)$. For simplicity in this discussion, we focus on homogeneous processes.

A Pedestrian Queueing Example — To illustrate queue inference, consider a signalized pedestrian cross walk having fixed cycle time T. Poisson-arriving pedestrians queue at curbside waiting to cross the street during a time interval of length T, and all such queuers are served in "bulk" fashion when the light changes at time T allowing them safely to cross the street. The number N of queued arrivals in any particular light cycle is Poisson distributed with mean λT. Given N, $X_{(i)}$ is the arrival time during $[0, T]$ of the ith queued pedestrian. Here X_i could be viewed as the arrival time of a *random* queued pedestrian, selected from, say, a photograph of all queued pedestrians taken just before the light changed at time T. The Poisson arrival assumption is usually thought of as evolving sequentially over time, with customer interarrival times selected in an iid manner from a negative exponential pdf with mean λ^{-1}.

An equivalent way to conduct the pedestrian cross walk experiment is to first select N from the Poisson distribution, and then, for each of the N queuers, to select the arrival time over $[0, T]$ independently from a uniform pdf. This experiment is probabilistically identical to the sequential Poisson arrival realization of the experiment. Suppose now at some intermediate time T, we focus on the total number of queued pedestrians $N(t)$, defined as the total number of arrivals (at curbside) during the interval $[0, t)$. The following results, derived from the second model of the process, are well known for $N(t)$:

$$
\begin{aligned}
&E[N(t)] = (t/T)T \\
&\mathrm{var}[N(t)] \equiv \sigma^2 N(t) = [N(t)/T]y[(T-t)/T] \\
&\Pr\{N(t) = k\} = \binom{N}{k}\left(\frac{t}{T}\right)^k\left(\frac{T-t}{T}\right)^{N-k}
\end{aligned}
\qquad (1)
$$

Here the *transactional* data are N, the total number of queuers, and T, the time until bulk service. From these data we have found transient values of conditional mean, variance and probability distribution of the queue length. Similar logic can be applied to find other performance measures, such as mean delay in queue, that in this case is trivially equal to $T/2$. This is one of the simplest examples of queue inferencing.

Queue Inference in More General Queues —In most queues, customers usually leave one-at-a-time. Their service completion times within a congestion period, recorded as part of the transactional data set, impose a set of inequality constraints on the arrival times of customers who waited in queue. It is this set of inequality constraints that produces precise conditioning information within the general context of order statistics, conditioning information that we use to deduce queue behavior.

Suppose for a M/D/1 system, we examine a congestion period having precisely $N = 2$ queued customers. For simplicity the service time is one minute per customer and the server's congestion period starts at time zero. Then since $N = 2$, we know that precisely 2 customers queued during this congestion period and after their service the server was again idle. The busy period for the server is *3* minutes in length, the time to serve 3 customers, the two who queued and the first arrival who initiated the congestion period. From the transactional data, we know that zero customers arrived during service of the last customer, the third in the congestion period and the second to queue (assuming FCFS queueing). We know that at least one of the 2 queued

customers must have arrived in [0,1], else there would be no queued customer to select for service commencement at time $T = 1^+$. Similarly, the second queued customer must have arrived by time $T = 2$.

Without the ordering information, the conditional arrival times for the two queued customers are independent uniformly distributed over [0,2]. In the joint sample space of random variables X_1 and X_2, this corresponds to X_1 and X_2 uniformly distributed over the square of size *2* in the positive quadrant. We can split the sample space into four equal subsquares, (1) $0 \le X_1 \le 1$, $0 \le X_2 \le 1$; (2)$1 \le X_1 \le 2$, $0 \le X_2 \le 1$; (3) $0 \le X_1 \le 1$, $1 \le X_2 \le 2$; (4) $1 \le X_1 \le 2$, $1 \le X_2 \le 2$. Without the additional conditioning information regarding service completion times, the outcome of the experiment is equally likely to be within each of the four subsquares, and conditional on being in a subsquare the r.v.'s X_1 and X_2 are condition-ally uniformly distributed over that subsquare. But the additional conditioning information from the transactional data imposes the constraints: $X_{(1)} \le 1$, $X_{(2)} \le 2$, thereby eliminating subsquare (4). The a priori probability of this event, called the *master probability*, is 3/4. *For any number of queued customers N, the master probability is the a priori probability that the order statistics will obey the ordered inequalities imposed by the transactional data*. Once we can efficiently calculate the master probability, most other quantities of interest are easy to compute.

Continuing with the $N = 2$ example, if we know that X_1 and X_2 fall in subsquare (1), then these two arrival times are uniform identically distributed over [0,1]. If one falls in [0,1] and the other falls in [1,2], that is, subsquare (2) or (3), then the minimum is uniformly distributed over [0,1] and the maximum is uniformly independently distributed over [1,2]. This property generalizes: *once we know that n_1 of N arrival times are contained in subinterval $[t_k, t_{k+1})$, where t_k and t_{k+1} are the entry into service times of queued customers k and k_{+1}, respectively, during the congestion period, then the n_1 arrival times are conditionally uniform and independently distributed over $[t_k, t_{k+1})$* (Larson, 1990). These facts allow us to obtain many useful performance characteristics of the queueing system, conditioned on the transactional data.

A simple application of the above observation yields for the pdf of the arrival time *A* of a randomly queued customer the step-wise decreasing pdf shown in Figure 1a. The form of this pdf generalizes to arbitrary *N*: *the marginal pdf for the arrival time of a random queued customer has a step-wise de creasing pdf over the duration of the congestion period, with each step occurring at an end-of-service time t_i* (Hall, 1992; Larson, 1990).

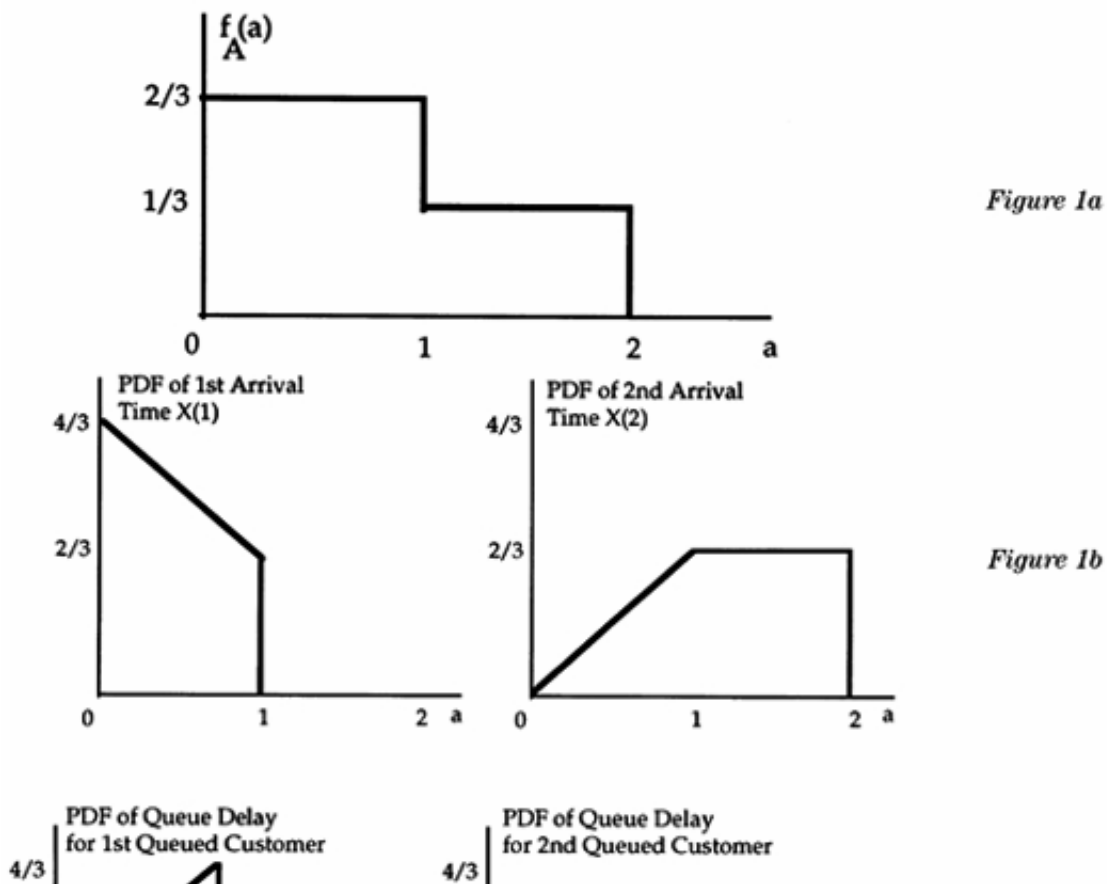

Figure 1a

Figure 1b

Figure 1c

Figure 1
Probability density functions of arrival times and queueing delays of queued customers.

As a second illustration, the conditional arrival time $X_{(1)}$ of the first queued customer is either the minimum of two uniform independent r.v.'s over [0,1] *or* simply uniform over [0,1], with the former situation applying only if the experimental outcome is within subsquare (1). Likewise the conditional arrival time $X_{(2)}$ of the second queued customer is either the maximum of two uniform independent r.v.'s over [0,1] or simply uniform over [1,2], the former applying again only within subsquare (1). Recalling that such minimum and maximum r.v.'s have triangular pdf's and combining results appropriately, we immediately obtain the pdf's for the arrival times of the two respective customers, as shown in Figure 1b. Finally, assuming a FCFS queueing discipline, the queueing delay for the first queued customer is $1 - X_{(1)}$ and the queueing delay of the second is $2 - X_{(2)}$. The corresponding queue delay pdf 's are inverted forms of those in Figure 1b, as shown in Figure 1c. If a bank knows that you were the second customer in this congestion period, it then has the required information to begin to build your personal pdf for bank queueing. To obtain the monthly pdf, the bank simply has to add together such conditional pdf's for each banking service session that you had during the month.

A General Result in Order Statistics and Application to Queue Inference — Suppose that the service end/start time transactional data are given by the vector $t = \{t_i;\ i = 1,\ldots, N\}$. In a queue inferencing setting, t_i has two definitions: (1) it is the observed time of departure of the *i*th departing customer to leave the system during the congestion period; (2) it is also the observed time for the *i*th customer from the queue to enter service, not necessarily in a FCFS manner. The two sets of individuals comprising the set of arriving customers and the set of departing customers during a congestion period are never identical and may be disjoint. The number of servers *M* does not enter into the analysis, nor do any distributional properties of the service times (e.g., there is no requirement for *iid* service times). We do assume that service times are independent of arrival times. For any given congestion period, the QIE computations may occur any time after completion of the congestion period.

Let $X_1, X_2, =, X_{N(1)}$ be an *iid* sequence of r.v.'s with values in [0,1], where the sequence length $N(1)$ is an independent random integer. We seek a computationally efficient algorithm to calculate the probability of an order statistics vector lying in a given *N*-rectangle,

$$\Gamma(\underline{s},\ \underline{t}) \equiv \Pr\{s_1 < X_{(1)} \le t_1,\ s_2 < X_{(2)} \le t_2,\ \ldots,\ s_N < X_{(N)} \le t_N \mid N(1) = N\}, \tag{2}$$

where s $\equiv$ $(s_1,s_2,\ldots s_n)$,t$(t_1,t_2,\ldots t_n)$ and without loss of generality the sequences $\{s_i\}$ and $\{t_i\}$ are increasing. Using the fact that the *N* unordered Poisson arrival times during any fixed time interval (0, *T*] are *iid* and now (for convenience) scaling the congestion period to (0,1], then in our notation, $\Lambda(0,t)$ is the a priori probability that the (unobserved) arrival times $X_{(1)}, X_{(2)},\ldots, X_{(N)}$ obey the inequalities $X_{(i)} \le t_i$ for all $i = 1, 2, \ldots, N$, that is, it is the "master probability" discussed above. That is, "$X_{(i)} \le t_i$" simply says that the *i*th arriving queued customer must arrive (and enter the queue) before completion of service of the *i*th departing customer from service.

If the Poisson arrival process is homogeneous, then the unordered arrival times are *iid* uniform and the rate parameter of the process *does not* enter the analysis. If the arrival process is nonhomogeneous, then the time-dependent arrival rate parameter $\lambda(t)$ must be known up to a positive multiplicative constant for use in computing the cdf $F(x)$, that is,

$$F(x) = \frac{\int_0^x \lambda(t)dt}{\int_0^1 \lambda(t)dt} \qquad 0 \ (\leq x \leq 1).$$

For simplicity, we assume that $F(x)$ is strictly monotone nondecreasing continuous. Jones and Larson (1995) have derived an $O(N^3)$ algorithm for finding $\Lambda(s,t)$. We next discuss briefly several of its queue inference applications.

The Maximum Experienced Queue Delay — Assume we have a FCFS queue. Suppose we consider a congestion period having N customers with observed departure time vector t, and we are interested in the maximum time that any of the N customers was delayed in queue, given t. More precisely, we are interested in the cdf of the maximum of N nonindependent r.v.'s, the in-queue waiting times of the N queued customers, given t.

Define $D(\tau/t)$ as the conditional probability that none of the N customers waited τ or more time units, given the observed departure time data. Set $s_i = \max\{t_i - \tau, 0\}$ for all $i = 1, 2, \ldots, N$. Then $\Lambda(t\text{-}\tau,t)$ is the a priori probability that the observed departure time inequalities will be obeyed and that no arrival waits τ or more time units in queue. Clearly,

$$D(\tau|\underline{t}) = \Gamma(\underline{t} - \tau, \underline{t})/\Gamma(0, \underline{t}). \qquad (3)$$

Maximum Queue Length — Without any assumption regarding queue discipline, suppose we define $s=t^{*k}$ such that $s^* \ t_{(i\text{-}K}$ for all $i = 1, 2,\ldots, N$; $K = 1, 2, \ldots, N$, where a non-positive subscript on t implies a value of zero. These values for s imply that each arriving customer i has to arrive after the departure time of departing customer $i = K$ during the congestion period. Now we can compute the conditional probability that the queue length did not exceed K during the congestion period, given t:

Pr{Q≤k|t}=Pr{queue length did not exceed K during the congestion period | observed departure time data}, or

$$\Pr\{Q \leq K|\underline{t}\} = \Gamma(\underline{s}^{*K}, \underline{t})/\Gamma(0, \underline{t}). \qquad (4)$$

Probability Distribution of Queue Length — Following the same arguments as in Larson (1990), we can use the $O(N^3)$ computational algorithm to determine for any queue discipline the probability distribution of queue length at departure epochs, and by a balance of flow argument, this distribution is also the queue length distribution experienced by arriving customers.

The Cumulative Distribution of Queue Delay — The algorithm allows computation of points on the conditional in-queue waiting time distribution, given the observed departure data. Again assume we have a FCFS queue. Suppose we define $\beta_j(\tau|t) \equiv \equiv$ Pr{jth customer to arrive during the congestion period waited less than τ time units|observed departure time data}. Then if we set $s=s^j$, defined so that

$$\begin{aligned} s_i^j &= 0 & i &= 1, 2, \ldots, j-1 \\ s_i^j &= \text{Max}\{t_j - \tau, 0\} & i &= j, j+1, \ldots, N \end{aligned}$$

we can write

$$\beta_j(\tau|\underline{t}) = \Gamma(\underline{s}^j, \underline{t})/\Gamma(0, \underline{t}). \qquad (5)$$

This result allows us to determine for any congestion period the probability that a random customer waited less than τ time units, given the observed departure data. We simply compute Equation (5) for each value of j and average the results. Jones and Larson (1995) have developed a separate algorithm that allows $O(N^3)$ computation of this average probability of queue delay exceeding some threshold.

RESEARCH LITERATURE

Research in queue inferencing is rather extensive. For $O(N)$ algorithms for queue performance estimation, see Bertsimas and Servi (1992), Larson (1990), Daley and Servi (1992, 1993); for personnel queue delay pdf, see Hall (1992); for balking, see Larson (1990), Daley and Servi (1993), Jones (1994, 1999). Applications of QIE are discussed in Gawlick (1990), and Chandrs and Jones (1994). QIE concepts have been incorporated into a commercial software product Queue Management System (QMS) and has been used by banks, an air-line, and the United States Postal Service.

See Queueing theory; Retailing.

References

Bertsimas, D.J. and Servi, L.D. (1992). "Deducing Queues from Transactional Data: The Queue Inference Engine Revisited," Operations Research, 40(S2), 217–228.

Chandrs, K. and Jones, L.K. (1994). "Transactional Data Inference for Telecommunication Models," presentation at First Annual Technical Conference on Telecommunications R & D in Massachusetts, University of Massachusetts, Lowell, Massachusetts.

Daley, D.J. and Servi, L.D. (1992). "Exploiting Markov Chains to Infer Queue-Length from Transactional Data," Jl. Applied Probability, 29, 713–732.

Daley, D.J. and Servi, L.D. (1993). "A Two-Point Markov Chain Boundary-Value Problem," Adv. Applied Probability, 25, 607–630.

Gawlick, R. (1990). "Estimating Disperse Network Queues: The Queue Inference Engine," Computer Communication Review, 20, 111–118.

Hall, S.A. (1992). "New Directions in Queue Inference for Management Implementations," Ph.D. dissertation in Operations Research, Massachusetts Institute of Technology, available as Technical Report No. 200, Operations Research Center, M.I.T., Cambridge.

Jones, L.K. (1994). "Inferring Balking Behavior and Queue Performance From Transactional Data," Technical report, Operations Research Center, M.I.T., Cambridge.

Jones, L.K. (1999). "Inferring Balking Behavior From Transactional Data," Operations Research, 47, 778–784.

Jones, L.K. and Larson, R.C. (1995). "Efficient Computation of Probabilities of Events Described by Order Statistics and Applications to Queue Inference," INFORMS Jl. Computing, 7, 89–100.

Larson, R.C. (1990). "The Queue Inference Engine: Deducing Queue Statistics From Transactional Data," Management Science, 36, 586–601. Addendum, 37, 1062, 1991.

This text originally appeared in Encyclopedia of Operations Research and Management Science - ISBN 079237827x