STATIONARY WAITING-TIME DISTRIBUTIONS FOR SINGLE-SERVER QUEUES

BY R. M. LOYNES

University of Cambridge

1. Summary and introduction. In a single-server first-come-first-served queue the waiting-times of successive customers are related by the equation

$$w_{n+1} = [w_n + U_n]^+ \tag{1}$$

where

$$U_n = S_n - T_n . \tag{2}$$

Here S_n and T_n are the service-time of the nth customer, and the time between the arrivals of the nth and the $(n + 1)$th customers, respectively.

In the particular case of mutually independent identically distributed U_n the basic investigation of this equation was carried out by Lindley [1], who found a simple necessary and sufficient condition for the existence of a stationary waiting-time distribution and derived this distribution in certain special cases. The theory was developed by Smith [2], who under fairly weak conditions gave a systematic treatment of the Wiener-Hopf equation obtained by Lindley.

In the less restricted case when U_n is a strictly stationary process it has been shown by the author (Loynes [3]) that under a simple condition the existence of a unique stationary distribution is again ensured: it is the purpose of this paper to show how (when this condition is satisfied) the stationary distribution may sometimes be found, and to obtain some qualitative results. The theory will be developed in Sections 3 and 4, and some examples discussed in Section 5.

On the assumption that U_n is a metrically transitive sequence the condition just referred to is

$$E[U_n] < 0, \tag{3}$$

and we shall therefore suppose this satisfied throughout the paper. The assumption that U_n is metrically transitive does not affect most of the arguments in this paper, but it will be made for convenience.

For the existence of a *unique* stationary distribution (3) is both necessary and sufficient, but there is sometimes more than one stationary distribution when the inequality in (3) is replaced by equality. Such situations have been ignored here for several reasons: they are not very common, some of the arguments either break down or need adaptation, and in any case the problems are not then as difficult, for it follows from equation (7) of [3] that the function ψ occurring in (5) below is then identically equal to one, so that the only unknown

Received March 17, 1961; revised February 3, 1962.

is ϕ. There is never a stationary distribution when the inequality in (3) is reversed.

Throughout the entire investigation we shall suppose the queue has a structure which satisfies the following condition (H). This is actually no restriction at all when the queue is in the stationary state, since the variables z_n occurring in it may be taken as (W_{n+1}, T_{n+1}), where $\{W_n\}$ is the stationary waiting-time sequence. However, all the results will refer to the distribution of the waiting-time conditional on z_n, so that it will always be tacitly assumed that the distribution of z_n is known in order to allow deductions about the unconditional distribution of the waiting-time.

(H): There exists a sequence $\{z_n\}$ (defined on the same probability space as w_n) of random vectors—i.e., in finite-dimensional Euclidean space—with the following properties:

(i) $\{z_n, S_n, T_n\}$ is a strictly stationary process.

(ii) S_n, T_n, and w_n are conditionally independent given z_{n-1}, z_n.

(iii) w_n and z_n are conditionally independent given z_{n-1}.

The assumption that (H) is satisfied allows us to write down the equation satisfied by the stationary (conditional) waiting-time distribution function. We cannot, however, in such a general situation actually find the solution to the equation, and to discover appropriate further restrictions which may make this possible we may consider further the case treated by Lindley and Smith.

The general solution given by Smith was analytically complex and apparently only obtainable from the Wiener-Hopf technique or the equally specialized method due to Spitzer [4]. If, however, either the service-times or the inter-arrival times had a distribution with a rational characteristic function the solutions were much simpler in form, and in at least a few cases had been obtained previously by other methods. With this in mind we impose similar conditions here on the conditional characteristic functions; on that of the inter-arrival time in Section 3, and on that of the service-time in Section 4. The lack of success in treating the problem without these restrictions is probably no great loss in practice, in view of the complex nature of the Wiener-Hopf solution for the simple case.

The results obtained under the conditions imposed so far are in the nature of aids to the solution of the problem, rather than the solution itself. A simple situation for which the required distribution may then be obtained straightforwardly occurs when the following "finite-matrix" condition is satisfied:

(HF): Condition (H) holds, and z_n has (with probability one) only a finite number k of different possible values.

An interesting result obtained by Smith showed that, roughly speaking, the characteristic functions of the waiting-time and the service-time had the same number of poles, and a somewhat similar result related the distribution of the inter-arrival times to that of the waiting-time. Under (HF) it will be found that similar results are also true. Smith in fact gave explicit formulae of a simple type for the two special cases, but we shall content ourselves here with pointing out how corresponding formulae could be obtained, since it appears that they are no longer simple when k has a value other than one.

As we have already remarked, all simple queues are included in (H) when in the stationary state. Even under the condition that the distribution of z_n be known, a large class of queues satisfies (H), such as that in which S_n and T_n are independent stationary Markov processes. An example of a queue of considerable interest which apparently does not admit a specification of this type is that in which customers are due at regular intervals but are late independently with a lateness distribution which extends to ∞ (if the lateness distribution and service time distribution are both negative exponential, then (H) is satisfied, and a brief discussion of this example will be found in Section 5).

It will be seen that by taking $z_n = 1$ the classical case with independence is included under (HF); two other types of queue satisfying (HF) are of some interest, and are discussed in Section 5—one being a simple queue whose input is a mixture of two streams of customers, and the other a queue whose input consists of customers who have already been served in a first queue. In addition to any other queues which may fall into this class, it is presumably true, that any queue may be approximated by those of this "soluble" class, although there are of course very great difficulties in deciding how such an approximation should be carried out.

Finally we observe that since the content of a semi-infinite dam in discrete time is described by the same equation (1), the results obtained here may be applied in that context also, and in particular the finite-matrix case gives a class of dams with serially dependent inputs for which explicit solutions can be found.

2. Basic theory. For conditional probabilities we shall employ the usual notation; for example the distribution function of z_{n-1} conditional on z_n will be denoted by $\mathrm{pr}\ (z_{n-1} \leqq x \mid z_n)$. The lack of uniqueness of these distributions will be ignored, as for example in Theorem 1, but there is no difficulty in dealing with it. For details of this and of the properties of conditional characteristic functions reference may be made to Loève [5].

Suppose that the queue is in the stationary state and let $F(x, z_n) = \mathrm{pr}\ \{w_{n+1} \leqq x \mid z_n\}$; then $F(x, z_n) = 0$ for $x < 0$. By H(ii) and (iii), for $x \geqq 0$ we have

$$\begin{aligned}\mathrm{pr}\ \{w_{n+1} \leqq x \mid z_n, z_{n-1}\} &= \int \mathrm{pr}\ \{w_n \leqq y \mid z_n, z_{n-1}\}\, d_y\, \mathrm{pr}\ \{U_n \leqq x - y \mid z_n, z_{n-1}\} \\ &= \int \mathrm{pr}\ \{w_n \leqq y \mid z_{n-1}\}\, d_y\, \mathrm{pr}\ \{U_n \leqq x - y \mid z_n, z_{n-1}\};\end{aligned}$$

taking expectations conditional on z_n, and making use of the assumption of stationarity, we obtain

$$F(x, z_n) = \int \mathrm{pr}\ (dz_{n-1} \mid z_n) \int F(y, z_{n-1})\, d_y\, \mathrm{pr}\ \{U_n \leqq x - y \mid z_n, z_{n-1}\}, \quad x \geqq 0, \tag{4}$$

where $F(x, z_n) = 0$ for $x < 0$, as the equation satisfied by the stationary (conditional) distribution of waiting-times. The unconditional waiting-time distribution can of course be immediately obtained from $F(x, z_n)$ whenever we know the distribution of z_n. The necessary conditional distribution of U_n is obtained, according to H(ii), as the convolution of those of S_n and T_n. If U_n is a sequence of mutually independent random variables, then by putting $z_n \equiv 1$ we return to the usual Wiener-Hopf equation.

It is possible in principle to solve (4) by iteration: if we replace F under the integral sign by unity and use the resulting value of the right-hand side as the next approximation, then, continuing in the usual way, the sequence of functions obtained converges monotonically to the required solution, and gives in fact the waiting-time distributions of the successive customers when the first customer does not have to wait. This approach is unlikely to be often useful for finding explicit results.

Another possibility is to try to fit solutions of a particular type, such as sums of exponentials; while such an attempt will not in general be successful, this may very well be the simplest way of applying the results obtained in Section 4 (particularly Theorem 3), and it might succeed in other situations. If we know the form of the solution, we should like to be sure before we begin that at least in principle the unknowns are completely determined by (4). Conversely, if we have obtained a solution of (4) by some means we have to decide whether this is the one we require, for although we know there is only one solution which is a distribution function in x for all z_n, there may be other solutions not of this type. We therefore give the following uniqueness theorem, which may settle these questions. This theorem is closely connected with Lindley's demonstration that his solution for deterministic arrivals and χ^2 service-time is a distribution function.

THEOREM 1. *If, for all* z_n, z_{n-1}, $\text{pr}\ (U_n \leqq x \mid z_n, z_{n-1}) < 1$ *for all finite* x, *then the required solution is the only solution of* (4) *vanishing for negative* x *and tending to unity as* x *tends to* ∞, *such that* $E[\sup_{x \geqq 0} |F(x, z_n)|] < \infty$.

PROOF. It is evident that the required solution is of such a type. Suppose there are two such solutions: then their difference is another, tending however to zero at ∞, and we shall show that this implies that it vanishes everywhere, and the theorem is proved.

Let this difference be $D(x, z_n)$, and let $A(z_n) = \sup_{x \geqq 0} |D(x, z_n)|$. Then directly from (4) we find that

$$|D(x, z_n)| \leqq \int \text{pr}\ (dz_{n-1} \mid z_n) A(z_{n-1})\ \text{pr}\ (U_n \leqq x \mid z_n, z_{n-1}).$$

There exists a monotone subsequence x_m, with a limit point $x(z_n)$, such that $|D(x_m, z_n)|$ tends to $A(z_n)$, and if we apply the dominated convergence theorem to the right hand side of the equation above using this subsequence we find that

$$A(z_n) \leqq \int \text{pr}\ (dz_{n-1} \mid z_n) A(z_{n-1})\ \text{pr}\ (U_n \leqq x(z_n) \mid z_n, z_{n-1}).$$

Hence

$$E[A(z_{n-1})] = E[A(z_n)] \leqq E[A(z_{n-1}) \text{ pr } (U_n \leqq x(z_n) \mid z_n , z_{n-1})],$$

which is only possible if

$$[1 - \text{pr } (U_n \leqq x(z_n) \mid z_n , z_{n-1})]A(z_{n-1}) = 0,$$

with probability 1. Thus either $A(z_{n-1}) = 0$, in which case the proof is complete, or $x(z_n)$ is infinite; since, however, $D(x, z_n)$ tends to zero at ∞, the latter would imply that $D(x, z_n) = 0$ everywhere, and the stated conclusion follows.

Although in some circumstances the solution may be found by trying solutions of special forms, the only systematic method of investigating (4) seems to be by taking Laplace (or Fourier) transforms. These (Laplace-Stieltjes) transforms exist on the imaginary axis, but may not do so elsewhere. On the imaginary axis, then, directly from (1) and (H),

$$(5)\quad \phi(s, z_n) + \psi(s, z_n) - 1 = E[\phi(s, z_{n-1})H(s, z_n , z_{n-1})G(-s, z_n , z_{n-1}) \mid z_n],$$

where $\phi(s, z_n) = \int e^{-sx} d_xF(x, z_n)$ is, for $s = -i\tau$, the characteristic function of w_{n+1} conditional on z_n. Similarly $\psi(s, z_n)$ refers to $-[w_n + U_n]^- = w_n + U_n - w_{n+1}$, and $H(s, z_n , z_{n-1})$ and $G(s, z_n , z_{n-1})$ refer to S_n and T_n respectively, conditional on z_n and z_{n-1}. In addition to existing on the imaginary axis, ϕ, G and H, being transforms of the distribution functions of non-negative random variables, also exist and are analytic (for fixed z_n , z_{n-1}) in $Rs > 0$, and ψ exists and is regular in $Rs < 0$. Furthermore, in the half-planes in which their existence is guaranteed, these functions are uniformly bounded by unity.

The difficulty of (5) lies in the presence of two unknown functions, essentially determined only by their descriptions as transforms. This can be overcome in certain cases, and further investigation will begin with a rearranged version of (5).

$$(6)\quad 1 - \psi(s, z_n) = \phi(s, z_n) - E[\phi(s, z_{n-1})H(s, z_n , z_{n-1})G(-s, z_n , z_{n-1}) \mid z_n].$$

In Section 3 we shall put restrictions on G, and in Section 4 on H.

3. Particular types of inter-arrival time distribution. The left hand side of (6) is analytic in $Rs < 0$; we shall now suppose that G satisfies a condition that will allow us to make the right hand side analytic in $Rs > 0$, and shall then be able to determine the form of ψ.

THEOREM 2. *If* (i) $G(-s, z_n , z_{n-1}) = N(s, z_n , z_{n-1})/D(s)$, *where* N *is everywhere analytic, and* D *is a polynomial of degree* m; *and* (ii) *there is a constant* A *such that whenever* $|s| \geqq A$ *and* $Rs \geqq 0$, $|GH| < 1$ *for all* z_n , z_{n-1}, *then* $1 - \psi(s, z_n) = sR(z_n , s)/D(s)$, *where as a function of* s, R *is a polynomial of degree* $m - 1$.

The expression of G as a quotient need not be in its lowest terms. When the theorem is applicable, the problem of finding ψ, initially a function of two variables, is reduced to that of finding m functions of the single variable z_n, which must be such as to make ϕ be regular in the right half plane and take the value unity at $s = 0$. If it is possible to solve (6) for ϕ as an explicit function of ψ (a condition which is in any case necessary if this approach is to be useful), then these properties of ψ may very well be enough to completely determine it, and hence finally ϕ. This procedure is equivalent to substituting in (6) those values of s for which the linear operator acting on ϕ fails to have an inverse, and this may occasionally be done directly very simply—for an instance, see example (ii) of Section 5.

Proof. In any finite region of the right half plane HN is bounded, uniformly in z_n, z_{n-1}: for we can enclose the region in a semicircle standing on the imaginary axis on the circumference and diameter of which $|HN| \leqq |D|$, and the maximum modulus principle then gives the result.

If (6) is multiplied by $D(s)$, the left hand side remains analytic in $Rs < 0$, and the right hand side now becomes analytic in $Rs > 0$. The proof that the expectation term is analytic is the only non-trivial step, and this can be carried out by showing the term to be differentiable. The function whose expectation is being evaluated is clearly analytic, and by the Cauchy integral formula, its derivative is bounded uniformly in z_n, z_{n-1}, so that application of the bounded convergence theorem to the real and imaginary parts is possible, and the result then follows.

Now according to the principle of analytic continuation, this means that the two sides of the equation together define a function which is analytic everywhere, and since, moreover, ψ is bounded by unity in the left half plane, and GH by unity sufficiently far out in the right half plane, we may apply Liouville's theorem to show that this analytic function is in fact a polynomial in s of degree not greater than that of $D(s)$. Since $\psi(0, z_n) = 1$, this polynomial must have a factor s, and the theorem is proved.

Under (HF) we can continue the discussion and see that with mild restrictions there are in general enough conditions to determine $R(z_n, s)$ exactly, and in that case ϕ can certainly be found by solving (6), which is then essentially a set of k linear equations in k unknowns.

To see that this is so, let us multiply (6) by $D(s)$ and consider it as a relationship between certain vectors and matrices. With no loss of generality we can suppose the possible values of z_n to be $1, 2, \cdots, k$, and then we can write, for instance, $H(s, z_n, z_{n-1}) = H_{ij}(s)$ when $z_n = i$, $z_{n-1} = j$. The (backward) transition matrix of the z_n will be denoted by $P = [p_{ij}]$, where $p_{ij} = \text{pr}\,[z_{n-1} = j \mid z_n = i]$. The result of multiplication by $D(s)$ can now be written, using the conclusion of Theorem 2, as

$$sR(s) = [D(s)I - Q]\phi(s) \tag{7}$$

where $R(s)$ and $\phi(s)$ are column vectors having $R(z_n, s)$ and $\phi(s, z_n)$ respectively as components, I is the unit matrix, and

$$Q = [q_{ij}(s)] = [p_{ij}H_{ij}(s)N_{ij}(s)]. \tag{8}$$

According to Theorem 2, $R(s)$ is a polynomial in s of degree $m - 1$, so that the unknown coefficients are in fact km in number. Suppose now that each H_{ij} is analytic at the origin, and that the matrix P is irreducible: then according to Theorem 5b of the appendix, the matrix $D(s)I - Q$ is singular for $km - 1$ values of s inside the right half plane. As each term of (7) is regular there, this implies that, for these values of s, $R(s)$ is orthogonal to the corresponding left eigenvector of the matrix, and thus we obtain $km - 1$ equations relating the coefficients of $R(s)$, and another k equations are obtained from the fact that $\phi(0, z_n) = 1$. With more than km equations relating km unknowns, we have in general enough to completely determine the solution, for inconsistency is clearly impossible. It is easy to construct examples in which there are not km independent equations among those given by the above approach, but in all such cases that have been tried ϕ has been uniquely determined by the fact that it is regular in the right half plane. It has not been found possible to prove that this is necessarily so, but it would be rather surprising if it were not.

The manipulations with the eigenvectors and the subsequent solution for ϕ can be combined without too much difficulty into a single explicit formula. As the result is neither particularly informative nor particularly elegant, it has not been thought worthwhile to reproduce it here, but even without it the connection between the solution of (7) and the corresponding result of Smith is clear, for ϕ can obviously be expressed as a fraction with a known denominator and a numerator known except for certain constants.

4. Particular types of service time distribution. In Section 3 G and H were supposed to satisfy conditions which made it possible to continue ψ analytically into the right half plane. Now we shall impose conditions which will instead make it possible to extend ϕ into the left half plane. Throughout this section it will therefore be supposed that, for fixed z_n, z_{n-1}, H can be continued analytically into the left half plane, the result being a single-valued function, analytic everywhere except for certain isolated singularities.

If at some point s_0 H is defined for all z_n, z_{n-1}, then (6) is a well-defined equation for ϕ at this point, and may thus have a solution there (which will be expressed, of course, as a function of the unknown value of $1 - \psi$). Any function $\beta(s, z_n)$, which is obtained by assigning to each such point s_0 the corresponding ϕ, will be called a solution of (6) in the left half plane. There may of course be more than one solution, and in any case the solutions will not normally be everywhere defined.

The method of approach will be on the one hand to find as nearly as possible the form of β, and on the other to relate β with ϕ, the transform we are seeking. For these purposes the following theorem is basic.

THEOREM 3. *Suppose that* (6) *has a solution* β *with the following properties:*

(i) $\beta(s, z_n)$ *is, for fixed* z_n, *the analytic continuation of* $\phi(s, z_n)$;

(ii) *there is some* $a \geqq 0$ *(possibly depending on* z_n*) such that, for fixed* z_n, $e^{as}\beta(s, z_n)/s$ *tends to zero as* s *tends to* ∞ *(in the left half plane)*;

(iii) *for fixed* z_n *the analytic function composed of* $\phi(s, z_n)$ *and* $\beta(s, z_n)$ *is regular everywhere except for poles.*

Then for $x \geqq a$, $F(x, z_n) - 1$ *is a finite sum of terms of the form* $\sum_{r=0}^{k-1} g_r(z_n)x^r e^{-bx}$, *where* $-b$ *is a pole of* β *of order* k. *The poles* b *may depend on* z_n, *but in any case* $Rb > 0$.

It would be expected, by analogy with the solution of matrix equations, that the values of b and k occurring in the theorem depend only on G and H and can therefore often be found directly, but that the functions $g_r(z_n)$ depend on the unknown ψ and must thus be found in a different way, although it is possible in certain cases to obtain some information about them directly (see example (i) Section 5).

The form of $F(x, z_n)$ guaranteed by the conclusion of the theorem is of a simple and familiar type; it is therefore desirable to find fairly simple conditions under which the premises of the theorem are satisfied. Although a detailed examination of the structure of the equation is the only way in which the applicability or otherwise of condition (iii) may be determined, the following theorem shows that it is sometimes possible to conclude directly that conditions (i) and (ii) are satisfied.

THEOREM 4. *Suppose there are positive constants* $c < 1$ *and* A *such that, when* $|s| > A$ *and* $Rs \leqq 0$, $H(s, z_n, z_{n-1})$ *is regular and*

$$|H(s, z_n, z_{n-1})G(-s, z_n, z_{n-1})| \leqq c$$

for all z_n, z_{n-1}. *Then in the region* $Rs < 0$, $|s| > A$, (6) *has a solution* β *which is regular and bounded (so that condition* (ii) *of Theorem* 3 *is satisfied with* $a = 0$) *and is the analytic continuation of* ϕ.

An obvious and interesting situation in which the conditions of this theorem are satisfied is that in which the service-times form a completely independent sequence and have a rational characteristic function, for then H is independent of z_n, z_{n-1}, $|H|$ tends to some constant less than one as s tends to ∞, and $|GH| \leqq |H|$ in the left half plane.

PROOF OF THEOREM 3. We first observe that by (i) β and ϕ are essentially the same function, and we shall for convenience refer to the complete function as ϕ. This function ϕ has according to (iii) no singularities other than poles, and is, furthermore, regular in $Rs > 0$.

Next we remark that ϕ can not have a pole at the origin (since it takes there the value unity), and therefore, again by (iii), is regular at that point. According to Theorem 5b (p. 58) of Widder [6], the axis of convergence of ϕ passes to the left of the origin; by Widder's Theorem 7.6b (p. 70), it follows that, for suitable $c < 0$,

$$F(x, z_n) - 1 = \lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \frac{\phi(s, z_n)}{s} e^{sx}\, ds, \qquad x > 0. \tag{9}$$

From this we obtain, by an application of Jordan's Lemma and (ii), that

$$F(x, z_n) - 1 = \frac{1}{2\pi i} \int_C \frac{\phi(s, z_n)}{s} e^{sx}\, ds, \qquad x > a, \tag{10}$$

where C is any contour which can be obtained from the infinite semi-circle standing on and to the left of the line of integration in (9) without leaving the region in which ϕ is regular.

Since, because of (iii), we can evaluate the integral occurring in (10) in terms of the residues of the integrand at its poles, the theorem follows; there can only be a finite number of poles because they are, by (ii), confined to a finite part of the plane and can have no limit point.

In passing we note that (iii) was not used to prove (10) except in order to show that ϕ is regular at the origin, although it is rather unlikely that (10) would be of use unless ϕ has no singularities other than poles.

PROOF OF THEOREM 4. We construct a solution $\beta(s, z_n)$ of (6) in $Rs \leqq 0$, $|s| > A$, by iteration. Let $\beta_0(s, z_n) = 1 - \psi(s, z_n)$ and

$$\beta_{k+1}(s, z_n) = 1 - \psi(s, z_n) + E[\beta_k(s, z_{n-1})K(s, z_n, z_{n-1}) \mid z_n] \tag{11}$$

for $k \geqq 0$, where $K = GH$.

Then $|\beta_0| \leqq 2$, and by induction

$$|\beta_{k+1} - \beta_k| \leqq 2c^{k+1}, \tag{12}$$

and hence, if we define β by

$$\beta(s, z_n) = \beta_0(s, z_n) + \sum_0^\infty [\beta_{k+1}(s, z_n) - \beta_k(s, z_n)], \tag{13}$$

the series converges absolutely and uniformly. The function β defined by (13) is clearly a bounded solution of (6) in the region under consideration, which is regular in the interior of the region.

We have now only to prove that β is the analytic continuation of ϕ. Both β and ϕ exist on the imaginary axis for $|s| > A$ and satisfy (6), and on combining these two versions of (6) we find

$$\sup_{z_n} |\beta - \phi| \leqq c \sup_{z_n} |\beta - \phi|; \tag{14}$$

since β and ϕ are bounded this implies that $\beta = \phi$, and an application of the principle of analytic continuation completes the proof of the theorem.

It is rather clear that here again much more precision is obtainable under (HF), and in fact the following corollary is true.

COROLLARY. *Under* (HF), *if, for fixed z_n, z_{n-1}, $H(s, z_n, z_{n-1})$ is meromorphic with a finite number of poles and there exist c and A such that when $|s| > A$ and $Rs \leqq 0$ $|H(s, z_n, z_{n-1})G(-s, z_n, z_{n-1})| \leqq c < 1$, then each component of ϕ is a rational function of s. If in addition each $G(-s, z_n, z_{n-1})$ is regular at the origin and the backward transition matrix of z_n is irreducible, the degree of the denominator of any component of ϕ is (not greater than) mk, where m is the degree of the least common multiple of the denominators of H. In any case if for each z_n, z_{n-1}, H has at least one pole, there is only one solution of* (6) *with rational characteristic function.*

PROOF. Since there are only a finite number of pairs z_n, z_{n-1}, we may suppose that the constants c and A are independent of z_n and z_{n-1}, and hence Theorem 4 is applicable, showing that conditions (i) and (ii) of Theorem 3 are satisfied.

As we have supposed $H(s, z_n, z_{n-1})$ meromorphic with a finite number of poles we can write it as the quotient of an everywhere regular function by a polynomial. Let us now write $D(s)$ for the least common multiple of the denominators of H, and use matrix notation as in Section 3. Then we have, for instance, $H(s, z_n, z_{n-1}) = N_{ij}(s)/D(s)$ when $z_n = i$, $z_{n-1} = j$, where $N_{ij}(s)$ is everywhere regular, and $D(s)$ is a polynomial. On multiplying (6) by $D(s)$ the result is

$$D(s)\gamma = [D(s)I - Q]\phi, \tag{15}$$

where $\gamma_i(s) = 1 - \psi_i(s)$, and

$$Q = [q_{ij}(s)] = [p_{ij}N_{ij}(s)G_{ij}(-s)]. \tag{16}$$

Now in the region $Rs < 0$ the functions $D(s)$, $\gamma_i(s)$, and $q_{ij}(s)$ are analytic; consequently, provided the determinant of $D(s)I - Q$ does not vanish everywhere, the solution of (15) for ϕ can be expressed as the quotient by this determinant of a function regular in this region and hence has no singularities there but poles. The fact that there is a unique solution to (6) and (15), when s is sufficiently far from the origin, was proved in Theorem 4 and implies that $D(s)I - Q$ does not have a determinant vanishing everywhere. It only remains to show that ϕ has no singularity on the imaginary axis, for condition (iii) of Theorem 3 to be verified, and then the first statement of the corollary will be proved.

Suppose for the moment that each $G_{ij}(-s)$ is regular at the origin. Then again by Widder's Theorem 5b, for s with sufficiently small real part each $G_{ij}(-s)$ exists and is regular, and hence, applying analytic continuation to the two sides of (15), each $\gamma_i(s)$ is regular at the origin, and by the argument already used $\phi_i(s)$ can have no singularity there other than a pole. Such a pole being however impossible, since ϕ is finite, it follows that ϕ is regular at the origin and thus on the entire imaginary axis.

Suppose now that $G_{ij}(-s)$ is not necessarily regular at the origin. Then we truncate the variables T_n at some finite value, large enough to ensure that (3) is still satisfied when U_n is the difference between S_n and the new variables T_n^1. A stationary waiting-time distribution exists for the new queue, which satisfies (HF) with the same z_n, and as G^1 is regular at the origin it follows by what we have already proved that ϕ^1 is also regular at the origin. Since T_n^1 is smaller than T_n, it is clear from (1) and (2) that the new waiting-time sequence is larger than the old one, and then from the existence of ϕ^1 for small negative values of s we deduce that the axis of convergence of ϕ lies in the left half plane, so that ϕ also is not singular at the origin.

The above argument, showing essentially that (for some purposes) the inter-arrival time characteristic function may be assumed regular at the origin, does not depend on the finite-matrix hypothesis and may be useful in other circumstances.

The second statement of the corollary follows directly from Theorem 5a of the appendix, and the third from Theorem 1, since if H has a pole the condition there is obviously met.

When the conditions of Theorem 5a are satisfied, it is possible to count equations, just as in Section 3, to see that there are in general enough to determine the constants. We have, in fact, shown above that then ϕ is rational with denominator $T(s)$, say, where $T(s)$ is a polynominal of degree km whose zeroes are those of the determinant of $D(s)I - Q$. Suppose that $R(s)$ is the numerator of ϕ, so that $R(s)$ is also a polynomial of degree km, and substitute this expression for ϕ into (15).

$$T(s)D(s)\gamma = [D(s)I - Q]R(s). \tag{17}$$

There are $km + 1$ unknown functions of z_n in R, or equivalently $k(km + 1)$ unknown constants. Whenever $T(s)$ vanishes $R(s)$ must be a right eigenvector of the matrix, and the ratios of its components are determined, giving a total of $km(k - 1)$ equations. Whenever $D(s)$ vanishes $R(s)$ must also vanish, giving a further km equations, and a final k derive from the fact that each component of ϕ is unity at $s = 0$. This enumeration may of course break down in special cases.

5. Examples. In this section we shall consider some examples, chosen either for their simplicity or their interest, on which our results throw some light.

(i) Winsten, [7], has considered a queue in which the nth customer, due to arrive at time n, actually arrives at $n + l_n$, where $\{l_n\}$ is a sequence of positive, bounded, mutually independent, (and identically distributed) random variables. The service-times are independent and negative exponential.

Such a queue can be treated by the methods of the present paper, although the complexity which occurs when the customers can arrive so late as to be out of the order in which they are due is, if anything, rather greater than in Winsten's approach. We should in fact take $(l_{n-p+2}, l_{n-p+3}, l_{n-p+4}, \cdots, l_{n+p})$ for z_n when l_n is bounded by the integer p, in which case T_n is a (complicated) function of z_n and z_{n-1}. Then the integral equation (6) for ϕ has a degenerate kernel.

This example is a convenient one, at least when $0 \leqq l_n < 1$, for showing the application of our results, but in order to avoid mere duplication of Winsten's results we shall suppose the service-times to be the sum of a constant d and a negative exponential variable with parameter α. The structure of the arrival process will not however be changed, since it will suggest two generalizations.

Thus, assuming $0 \leqq l_n < 1$ with probability one, we have

$$T_n = 1 + l_{n+1} - l_n, \tag{18}$$

and taking l_{n+1} as a suitable z_n,

$$H(s, z_n, z_{n-1})G(-s, z_n, z_{n-1}) = [\alpha e^{-ds}/(\alpha + s)]e^{s(1+l_{n+1}-l_n)}. \tag{19}$$

It must be supposed that $d < 1$, since otherwise inequality (3) cannot be satisfied.

From (6), which becomes now

$$(20)\quad 1 - \psi(s, l_{n+1}) = \phi(s, l_{n+1}) - \int d\text{pr}(l_n)\phi(s, l_n)[\alpha e^{-ds}/(\alpha + s)]e^{s(1+l_{n+1}-l_n)},$$

it follows that ψ can be extended to be analytic everywhere, in particular at the origin, and on solving (20) for ϕ it is immediate that the conditions of Theorem 3 are met with $a = d - l_{n+1}$.

The solution is

$$(21)\quad \begin{aligned} \phi(s, l_{n+1}) &= 1 - \psi(s, l_{n+1}) \\ &\quad + \frac{\alpha e^{(1+l_{n+1}-d)s}}{\alpha + s - \alpha e^{(1-d)s}} \int [1 - \psi(s, l_n)]e^{-sl_n}\, d\text{pr}(l_n), \end{aligned}$$

so that the only pole of ϕ occurs at the unique root b in the left half plane of the equation

$$(22)\quad \alpha + b = \alpha e^{(1-d)b}.$$

The application of Theorem 3, with at the same time a more careful investigation of the contour integral in (10), leads now to the conclusion that

$$(23)\quad F(x, l_{n+1}) = 1 - ce^{b(x+l_{n+1})}$$

when $x + l_{n+1} \geqq d$, where c is a constant as yet unknown.

To obtain $F(x, l_{n+1})$ when $x + l_{n+1} < d$, and to find c, we use the integral equation (4) directly. We see immediately that, provided $x \geqq 0$, $F(x, l_{n+1})$ is a function only of $y \equiv x + l_{n+1}$, say $f(y)$, and after an integration by parts find the following equation for $g(y) = f(y)e^{\alpha y}$:

$$(24)\quad g(y) = \alpha e^{-\alpha(1-d)} \int_0^{y+1-d} g(t)P(t)\, dt,$$

where $P(t) = \text{pr}\,[l_n \leqq t]$.

Now the values of $g(y)$ for $y \geqq d$ are known in terms of c, so that by the aid of the equation

$$(25)\quad g(d) - g(y) = \alpha e^{-\alpha(1-d)} \int_{y+1-d}^{1} g(t)P(t)\, dt,$$

deduced straightforwardly from (24), $g(y)$ can be found successively in the intervals $y \geqq 2d - 1$, $y \geqq 3d - 2$, $\cdots$, always in terms of c. Since $d < 1$, only a finite number of these intervals need be considered to give $g(y)$ for all non-negative y, and then a linear equation for c is obtained from (24) by setting $y = 0$.

The fact that c is determined uniquely by (24) follows from Theorem 1, so that the linear equation for c cannot be an identity.

(ii) The example of paragraph (i) is easy to solve largely because the integral equation (20) has a degenerate kernel. A more general form for T_n which also gives rise (in conjunction with suitable service-time structure) to a degenerate kernel is

$$T_n = f_0(l_{n+1}) + f_1(l_n) + \cdots + f_{k+1}(l_{n-k}), \tag{26}$$

where $\{l_n\}$ is a sequence of mutually independent identically distributed random variables, and the functions f_i are unrestricted except for the limitation that the resulting T_n must be positive. The appropriate choice for z_n would be $(l_{n+1}, l_n, \cdots, l_{n-k+1})$.

Similarly, when S_n has this generalized moving average form and T_n has a characteristic function of such a type that Theorem 2 is applicable, the solution is straightforward, for the way in which the unknown coefficients of $R(z_n, s)$ depend on z_n may be found by setting s equal to the various zeroes of $D(s)$.

It is of some interest that the queue whose arrival process structure is given by (26) and whose service-times are mutually independent and have a rational characteristic function can be investigated in a completely different manner. Supposing for simplicity that the service-times are negative exponential with parameter α, then without any restriction on T_n it follows from (4) that $F(x, z_n)$ is differentiable with respect to x (except at $x = 0$) and satisfies

$$[\partial F(x, z_n)/\partial x] + \alpha F(x, z_n) = \alpha E[F(x + T_n, z_{n-1}) \mid z_n]. \tag{27}$$

When T_n is given by (26) it happens that (27) can easily be manipulated to give the solution, and although in this case the results are more directly obtained via Theorems 3 and 4, it is conceivable that there are queues for which the opposite is true.

(iii) A generalization of (18) in a rather different direction results from supposing that

$$T_n = V_n + L(z_n) - L(z_{n-1}), \tag{28}$$

where V_n is a sequence of mutually independent (identically distributed) random variables, $\{z_n\}$ a stationary Markov sequence, and L an arbitrary function, all restricted by the condition that T_n must be positive.

If the service times form a completely independent sequence, (6) can be written as

$$e^{-sL(z_n)}[1 - \psi(s, z_n)] = \Phi(s, z_n) - H(s)G(-s)E[\Phi(s, z_{n-1}) \mid z_n], \tag{29}$$

where $\Phi(s, z_n) = e^{-sL(z_n)}\phi(s, z_n)$, and $G(s)$ is the transform of the distribution of V_n, rather than of T_n.

For fixed s (29) is a linear integral equation of the second kind, with parameter $H(s)G(-s)$, and the well-developed theory of such equations may make it possible to discuss the behaviour of Φ, and hence of ϕ, with relative ease. If, for instance, H is such that the conditions of Theorem 4 are satisfied, then all the singularities of ϕ are clearly confined to a bounded part of the plane: at least one such singularity is to be expected for every eigenvalue, so that if there is an infinite number of eigenvalues, there is likely to be a non-isolated singularity of Φ. This will usually be the case unless (HF) holds, or the kernel is degenerate. If (HF) holds, a somewhat different approach has of course already been given in Section 4.

A queue which happens to have just this structure is one (apparently due to Kendall) mentioned by Wishart in the discussion on Winsten's paper: the customers are due to arrive at unit intervals, but are late independently with a negative exponential lateness distribution, and the service is also independent and negative exponential. It is readily verified that

$$T_n = 1 + m_{n+1} + x_{n+1} - m_n - x_n , \tag{30}$$

where m_n is the number of customers overdue at the instant of arrival of the nth customer, and x_n is the time that has then elapsed since the last arrival-due point. Furthermore (m_n, x_n) is clearly a Markov process (and consequently, since m_n is integral and $0 \leqq x_n < 1$, $m_n + x_n$ is also Markov). Unfortunately the transition probabilities are so complicated that detailed progress seems unlikely, and, in addition, according to the remarks made in the previous paragraph ϕ has probably a non-isolated singularity.

The structure assumed may also arise when customers arrive late in a different way. If they are due at intervals V_n, but are late an amount $L(z_n)$, then provided these quantities are restricted to ensure that customers arrive in order, T_n is clearly described by (28). If in particular the customers were not much late, they would normally arrive in order, in which case (28) could be considered to represent small perturbations of the expected arrival pattern; these perturbations need not be supposed independent of each other, since $\{z_n\}$ is a Markov process.

(iv) In addition to the queue treated by Lindley, Smith, and others, there are as we have already observed at least two others with a direct physical meaning that fall in effect under (HF). In both we use the Erlang device, representing a variable whose distribution is of Erlang type E_l as the sum of l independent Poisson variables.

Suppose that a queue with independent inter-arrival-times and service-times, having distributions E_P and E_Q respectively, has a finite waiting-room of size M, and that the customers join a second queue immediately on completing their service in the first. On applying the Erlang device we obtain at any instant two phases p, q with $1 \leqq p \leqq P$, $1 \leqq q \leqq Q$ and we construct a Markov chain z by writing $z = (m, p, q)$ where m is the number of customers present in the first queue. One customer actually arrives at the second queue whenever m decreases by one, but the waiting-time will be unaffected if instead we suppose that a customer arrives each time z changes its state in any way, provided that the service-times for these extra customers are set equal to zero. If now the (real) service-times form a sequence of mutually independent identically distributed random variables, (HF) is clearly satisfied for the modified queue, taking for z_n the state of z immediately after the arrival of the nth customer; it is in fact the situation of Section 3 that concerns us here, and aside from purely practical difficulties the problem may be considered solved. Some care must be taken in the interpretation of the results, since only those transitions at which real customers arrive are of interest. There is, however, no difficulty, for the answer we obtain, pr $[w_n \leqq x \mid z_{n-1}]$, is by H(iii) equal to pr $[w_n \leqq x \mid z_{n-1}, z_n]$ and we can therefore ignore other transitions, and merely evaluate all quantities conditional on a real customer having arrived.

It would, of course, be rather more interesting to deal with the case $M = \infty$, but this seems to be impossible unless $P = Q = 1$. In fact, except in this latter case, it seems that even with negative exponential service-times the characteristic function of the waiting-time may not be meromorphic.

For the second example we suppose that the stream of arriving customers is formed by mixing several independent streams, each of which has independent Erlang inter-arrival times. On representing each of these component streams by means of the Erlang device, a Markov chain $z = (p_1, p_2, \cdots)$ is formed, where $p_1, p_2, \cdots$ are the various phases. As in the previous example, with suitable service-time structure the queue is made to satisfy (HF) by adding fictitious customers with zero service-time.

6. Acknowledgments. I am indebted to Mr. A. M. Walker for various helpful discussions and to the Department of Scientific and Industrial Research for a Research Studentship.

APPENDIX

Our goal is the following theorem on the number of singularities of the operators acting on ϕ, on the right hand sides of (7) and (15) respectively. The theorem applies in the finite matrix case and is obtained at the cost of imposing further restrictions. The number of possible values of z_n will again be denoted by k, and the (backward) transition matrix of the z_n by $P \equiv [p_{ij}]$, where $p_{ij} =$ pr $[z_{n-1} = j \mid z_n = i]$.

THEOREM 5. *Under* (HF): (a) *If $H(s, z_n, z_{n-1})$ can be expressed as $N(s, z_n, z_{n-1})/D(s)$, where N is analytic everywhere and $D(s)$ is a polynomial of degree m, if G is analytic at the origin, if P is irreducible, and if sufficiently far from the origin in the left half plane $|GH| < 1$, then the operator $D(s)I - Q$ occurring in* (15) *has singularities for km values of s inside the left half plane.* (b) *If $G(-s, z_n, z_{n-1})$ can be expressed as $N(s, z_n, z_{n-1})/D(s)$, where N is analytic everywhere and $D(s)$ is a polynomial of degree m, if H is analytic at the origin, if P is irreducible, and if sufficiently far from the origin in the right half plane $|GH| < 1$, then the operator $D(s)I - Q$ occurring in* (7) *has a singularity for $km - 1$ values of s inside the right half plane.*

We treat the situation considered in Section 4 and Theorem 5a. The matrix operator in which we are interested is then

$$D(s)I - Q \tag{A1}$$

where I is the $k \times k$ unit matrix,

$$Q = [q_{ij}] = [p_{ij}N_{ij}(s)G_{ij}(-s)], \tag{A2}$$

and $N_{ij}(s) = D(s)H_{ij}(s)$ is an analytic function of s.

This operator is singular whenever $D(s)$ is equal to one of the eigenvalues of the matrix Q. Inside the left half plane these eigenvalues are the roots of a polynomial equation of degree k, in which the coefficients are analytic functions of s. It is not difficult to show that the eigenvalues may be combined to form a number of analytic multiple-valued functions, with isolated branch points. To achieve this we may merely adapt the discussion of algebraic functions given in Knopp [8], taking care of the number and position of the branch points with the following remark. Either the discriminant of the characteristic equation is analytic and not identically zero, or it vanishes everywhere, in which case the equation can be factored into two (or more) polynomials whose discriminants are analytic and not identically zero. The zeroes of the discriminants, which give the branch points of the multiple-valued functions induced, are therefore isolated. In the whole half plane there may of course be an infinite number of branch points.

Since the eigenvalues generate multiple-valued functions, we may represent them on a Riemann surface; it is not difficult to see that Rouché's theorem may be applied to a contour on a Riemann surface, except that a zero occurring at a branch point is counted only once, no matter how many sheets join there, and we shall apply Rouché's theorem in this way to the two functions $D(s)$ and $\Lambda(s)$. $\Lambda(s)$ is, of course, the eigenvalue function. The contour we shall use will be that induced on every sheet by the semi-circle in the left half plane standing on the imaginary axis on and outside of which $|G_{ij}(-s)H_{ij}(s)| < 1$. This contour will clearly enclose all the zeroes of $D(s)$, and since there are just k sheets, it will encircle them k times.

To be able to use this contour, G will be assumed analytic at the origin. It may be that this is not necessary.

For given s on this contour we know that

$$|N_{ij}(s)G_{ij}(-s)| < |D(s)| \tag{A3}$$

except at the origin, where equality holds, and if $\lambda(s)$ is an eigenvalue of Q, there is some vector $c_i(s)$ such that

$$\lambda(s)c_i(s) = \sum_j p_{ij}N_{ij}(s)G_{ij}(-s)c_j(s). \tag{A4}$$

If M is the maximum value of $|c_i(s)|$ as i varies then $|\lambda(s)|\, M \leqq |D(s)|\, M$, with strict inequality everywhere except possibly at the origin. Hence on every sheet, everywhere on the contour except possibly at the origin, $|D(s)| > |\Lambda(s)|$. By considering the integrals used in the proof of Rouché's theorem, it will be seen that this exceptional point does not matter unless $D(0) = \Lambda(0)$.

From inspection of (A4), $\Lambda(0)/D(0)$ are just the eigenvalues of the matrix P. If we now assume that P is irreducible, it follows that the sheet of the Riemann surface on which $\Lambda(0) = D(0)$ has no branch point at the origin. Thus this branch of $\Lambda(s)$ is differentiable at $s = 0$, and consequently so is the corresponding $c_i(s)$. Dividing by $D(s)$, differentiating, and setting $s = 0$, we find, since $c_i(0) = 1$

$$\text{(A5)} \qquad [\lambda(s)/D(s)]'_{s=0} + c_i'(0) = \sum_j p_{ij}c_j'(0) + \sum_j p_{ij}[E_{ij}T - E_{ij}S],$$

where $E_{ij}T = E\{T \mid z_n = i, z_{n-i} = j\}$.

Multiplying (A5) by p_i and summing over i, where $\{p_i\}$ is the unique stationary distribution of the z_n, we find $[\lambda(0)/D(0)]' = E(T) - E(S)$ which is positive because of the condition necessary for a stationary distribution to exist.

Hence if the contour on this sheet is deformed to pass just to the left of the origin, then everywhere on the contour $|D(s)| > |\Lambda(s)|$, and Rouché's theorem can be applied to the function $D(s) - \Lambda(s)$. The number of zeroes of $D(s)$ enclosed by the contour is km.

The case when G is restricted is treated similarly, and we have completed the proof of the theorem.

REFERENCES

[1] Lindley, D. V. (1952). The theory of queues with a single server. *Proc. Cambridge Philos. Soc.* **48** 277–289.

[2] Smith, W. L. (1953). On the distribution of queueing times. *Proc. Cambridge Philos. Soc.* **49** 449–461.

[3] Loynes, R. M. (1962). The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambridge Philos. Soc.* **58** 497–520.

[4] Spitzer, F. (1957). The Wiener-Hopf equation whose kernel is a probability density. *Duke Math. J.* **24** 327–343.

[5] Loève, M. (1955). *Probability Theory.* Van Nostrand, New York.

[6] Widder, D. V. (1946). *The Laplace Transform.* Princeton Univ. Press.

[7] Winsten, C. B. (1959). Geometric distributions in the theory of queues. *J. Roy. Statist Soc., Ser. B.* **21** 1–22.

[8] Knopp, K. (1947). *Theory of Functions,* Part II. Dover Pub., New York.