ON THE RATE OF CONVERGENCE OF WAITING TIMES

C. K. CHEONG and C. R. HEATHCOTE

(Received 29 October 1964)

1. Introduction

Let $K(y)$ be a known distribution function on $(-\infty, \infty)$ and let $\{F_n(y), n = 0, 1, \cdots\}$ be a sequence of unknown distribution functions related by

$$F_{n+1}(y) = \int_{-\infty}^{y} F_n(y-v)dK(v), \qquad y \geqq 0, \tag{1}$$
$$= 0, \qquad y < 0,$$

subject to the initial condition

$$F_0(y) = 1, \qquad y \geqq 0,$$
$$= 0, \qquad y < 0.$$

If the sequence $\{F_n(y)\}$ converges to a distribution function $F(y)$ then $F(y)$ satisfies the Wiener-Hopf equation

$$F(y) = \int_{-\infty}^{y} F(y-v)dK(v), \qquad y \geqq 0, \tag{2}$$
$$= 0, \qquad y < 0.$$

The problem considered here is the rate of convergence of $F_n(y)$ to its limit $F(y)$ in the case when $F_n(y)$ can be interpreted as the distribution function of the waiting time w_n of the nth arrival to a single server queue. This interpretation implies that $K(y)$ is the distribution function of the difference between two nonnegative random variables and also provides motivation for the simple probabilistic arguments used. In the terminology of queueing theory we have a sequence $\{t_n\}$ of independent and identically distributed interarrival times, with common distribution function $A(y)$, and a second sequence $\{s_n\}$ of independent and identically distributed service times whose distribution function is $B(y)$. The two sequences are independent so that the differences

$$u_n = s_n - t_n, \qquad n = 0, 1, \cdots,$$

constitute a sequence of independent variables identically distributed as u.

https://doi.org/10.1017/S1446788700027786 Published online by Cambridge University Press

The common distribution function is

(3) $$P(u \leqq y) = P(u_n \leqq y) = \int_{-\infty}^{y} B(y+v)dA(v).$$

The waiting time w_n satisfies the recurrence relation ([7], [9])

(4) $$w_{n+1} = (w_n+u_n)^+, \qquad n = 0, 1, 2, \cdots,$$

subject to an initial condition which we will take as $w_0 = 0$. Here $x^+ = \max(0, x)$. Thus the distribution functions $P(w_n \leqq y) = F_n(y)$ satisfy the system (1) with $P(u \leqq y) = K(y)$ and it is well known that $\{w_n\}$ converges in distribution to w only when $Eu = b_1 - a_1 < 0$ ([7]).

$$b_1 = \int_0^\infty ydB(y) \text{ and } a_1 = \int_0^\infty ydA(y)$$

are respectively the expected service and inter-arrival times. When this is the case $P(w \leqq y) = F(y)$ satisfies the Wiener-Hopf equation (2).

The present remarks are confined to the case when $P(w_n \leqq y)$ does converge to $P(w \leqq y)$ and the condition for this convergence to be exponentially fast (Theorems 1, 2, 3) relates only to the distribution function $B(x)$, or rather to its Laplace-Stieltjes transform,

$$B^*(s) = \int_0^\infty e^{-sy} dB(y).$$

This is equivalent to imposing exponentially fast convergence to zero of the upper tail $1-K(y)$ of the distribution function $K(y) = P(u \leqq y)$ appearing in (1), but the queueing problem is special in the sense that the form of $K(y)$ given by (3) makes it possible to give simple proofs. In the case when $b_1 - a_1 > 0$, $P(w_n \leqq y) \to 0$ for all finite y, and associated rate of convergence problems have been studied by Heathcote [4] by methods similar to those used here. In the third case, when $b_1 = a_1$, $P(w_n \leqq y)$ also approaches zero, but not exponentially fast. The study of this null-recurrent case is still incomplete (c.f. [6], [9]).

Busy periods will be referred to frequently and they are discussed in the next section. If $U_n = \sum_{i=0}^{n-1} u_i$ and

$$c_n = P(U_n < 0) = \int_0^\infty [1-A_n(y)]dB_n(y),$$

where $A_n(y)$ and $B_n(y)$ are the n fold iterated convolutions of $A(y)$ and $B(y)$ respectively, it is known ([3], see also [5]) that the probability γ_n that a busy period consists of n services is given by

(5) $$G(x) = \sum_1^\infty x^n \gamma_n = 1 - \exp\left\{-\sum_1^\infty n^{-1} x^n c_n\right\}.$$

This expression can also be obtained via the fundamental identity of Spitzer

((3.2) of [9]) and equation (7) below. Thus the Spitzer formula reduces to

$$W(x) = \sum_0^\infty x^n P(w_n = 0) = \exp\left\{\sum_1^\infty n^{-1} x^n c_n\right\},$$

and, using (7),

$$G(x) = \frac{W(x)-1}{W(x)},$$

which yields (5).

It is convenient to rewrite (5) as

$$G(x) = 1-(1-x)\exp\left\{\sum_1^\infty n^{-1}x^n(1-c_n)\right\}, \tag{6}$$

the equality sign implying that both sides are finite or infinite together. It is shown in [3] and [5] that when $b_1 < a_1$ we have $G(1) = \sum_1^\infty \gamma_n = 1$ and

$$\mu_1 = \sum_1^\infty n\gamma_n = \exp\left\{\sum_1^\infty n^{-1}(1-c_n)\right\} < \infty.$$

2. Preliminary results

A feature of the study of many queueing systems is the use made of imbedded recurrent event processes. The only natural sequence of regeneration points of the queue we are concerned with is that formed by the epochs at which busy periods commence. This remains the case whether time is measured continuously or discretely in terms of numbers of customers. In the latter case a busy period is said to be of duration $N = 1, 2, \cdots$ if customers C_j and C_{j+N} do not have to wait and each of $C_{j+1}, C_{j+2}, \cdots, C_{j+N-1}$, do have to wait. N is therefore the recurrence time for the event ε of "no waiting", and the probability that ε occurs at the nth trial is $P(w_n = 0)$. The recurrence time distribution is $\{\gamma_r\}$, $r = 1, 2, \cdots$, where γ_r is the probability that a busy period consists of r customers served. Thus, with $\gamma_0 = 0$,

$$P(w_n = 0) = \gamma_n + \sum_{r=1}^n P(w_r = 0)\gamma_{n-r}, \qquad n = 1, 2, \cdots. \tag{7}$$

The initial condition assumed throughout is

$$P(w_0 = 0) = 1.$$

By the renewal theorem, (Chapter 13 of [2])

$$\lim_{n\to\infty} P(w_n = 0) = 1/\mu_1 \tag{8}$$

the limit being interpreted as zero if $\mu_1 = \infty$ or if $\sum_1^\infty \gamma_j < 1$.

https://doi.org/10.1017/S1446788700027786 Published online by Cambridge University Press

A similar renewal type argument can be used to derive a representation for $P(w_n \leqq y)$ when $y \geqq 0$. If customer C_n joins a non-empty queue his waiting time w_n is independent of the process prior to the initiation of the current busy period. Conditional on this busy period commencing at time $n-k$ it follows from the recurrence relation (4) that $w_n = \sum_{i=0}^{k-1} u_{n-k+i}$, so that w_n, conditionally, has the same distribution as U_n^+. Removing the conditioning we have

$$P(w_n \leqq y) = \sum_{k=1}^{n} P(U_k^+ \leqq y) P(w_{n-k} = 0, w_\nu > 0 \text{ for all } n-k < \nu < n).$$

The second probability on the right is simply the probability that the 'age' at time n of the recurrent event process defined in (7) is at least k. In terms of the tail probability $q_{k-1} = \sum_{j=k}^{\infty} \gamma_j$,

$$P(w_{n-k} = 0, w_\nu > 0 \text{ for all } n-k < \nu < n) = q_{k-1} P(w_{n-k} = 0).$$

These results can be summed up as follows:

LEMMA 1. *For* $y \geqq 0$

$$P(w_n \leqq y) = \sum_{k=1}^{n} P(U_k^+ \leqq y) q_{k-1} P(w_{n-k} = 0). \tag{9}$$

If $b_1 < a_1$, *so that the process is ergodic,*

$$P(w \leqq y) = \lim_{n\to\infty} P(w_n \leqq y) = \frac{1}{\mu_1} \sum_{k=1}^{\infty} q_{k-1} P(U_k^+ \leqq y). \tag{10}$$

Before proceeding to the main results one other Lemma is required, which is also of some independent interest.

LEMMA 2. *Suppose* $F(t)$ *is a distribution function on* $[0, \infty)$ *with finite mean* m_1. *Let* $K(z, t) = \sum_{n=1}^{\infty} z^n F_n(t)$, *where* $F_n(t)$ *is the nth iterated convolution of* $F(t)$. *If* $F(0) < 1$, *then there exists an* $x_0 > 1$ *such that, as* $t \to \infty$,

$$K(x_0, t) = O(e^{\varepsilon t}),$$

where ε *is the unique positive root in* s *of the equation* $x_0 F^*(s) = 1$.

PROOF. Note firstly that by Theorem 1 of Belyaev and Maksimov [1], $K(z, t)$ is analytic inside a circle of radius $[F(0)]^{-1}$ for all finite t. For a fixed x_0 in $(1, [F(0)]^{-1})$ the convexity of $F^*(s)$ and the condition $F^*(0) = 1$ establishes that the equation $x_0 F^*(s) = 1$ has a unique root in s, say $s = \varepsilon$. From the definition of $K(z, t)$ it is easily verified that

$$e^{-\varepsilon t} K(x_0, t) = x_0 e^{-\varepsilon t} F(t) + \int_0^t e^{-\varepsilon(t-y)} K(x_0, t-y) x_0 e^{-\varepsilon y} dF(y).$$

The function $G(t) = \int_0^t x_0 e^{-\varepsilon y} dF(y)$ is positive, non-decreasing, with total variation unity, and so is an honest distribution function. Further, $G(t)$ has a finite mean since

$$\int_0^\infty t dG(t) = \int_0^\infty t x_0 e^{-\varepsilon t} dF(t) \leqq x_0 m_1 < \infty.$$

Thus for fixed x_0 (and hence fixed ε) the function $H(x_0, t) = \int_0^t e^{-\varepsilon y} K(x_0, y) dy$ satisfies the renewal type equation

$$H(x_0, t) = \int_0^t x_0 e^{-\varepsilon y} F(y) dy + \int_0^t H(x_0, t-y) dG(y).$$

By the elementary renewal theorem (page 246 of [8])

$$\lim_{t\to\infty} \frac{1}{t} H(x_0, t) = \frac{\int_0^\infty x_0 e^{-\varepsilon t} F(t) dt}{\int_0^\infty t dG(t)} < \infty.$$

It follows that $e^{-\varepsilon t} K(x_0, t) = O(1)$ and the proof is complete.

3. The main results

It is assumed throughout this section that $b_1 < a_1 < \infty$.

THEOREM 1. *$G(x) = \sum_1^\infty x^n \gamma_n$ has a radius of convergence greater than unity if and only if there exists an $\varepsilon > 0$ such that $B^*(-\varepsilon) < \infty$.*

PROOF. If $B^*(-\varepsilon)$ converges for an $\varepsilon > 0$ then

$$\phi(\theta) = \int_{-\infty}^\infty e^{\theta y} dP(u \leqq y) = A^*(\theta) B^*(-\theta)$$

is defined and continuous in the interval $[0, \varepsilon]$. Since $\phi'(0) = b_1 - a_1 < 0$ it follows that there exists a δ in $[0, \varepsilon]$ such that $\phi(\delta) < 1$. Further,

$$B^*(-\delta) \geqq e^{\delta y}[1-B(y)],$$

and

$$[B^*(-\delta)]^n \geqq e^{\delta y}[1-B_n(y)]$$

for $y \geqq 0$ and a δ in $[0, \varepsilon]$. Thus

$$\begin{aligned} \sum_1^\infty n^{-1} x^n (1-c_n) &= \sum_1^\infty n^{-1} x^n \int_0^\infty [1-B_n(y)] dA_n(y) \\ &\leqq \sum_1^\infty n^{-1} x^n \int_0^\infty e^{-\delta y} [B^*(-\delta)]^n dA_n(y) \\ &= -\log[1 - xB^*(-\delta)A^*(\delta)]. \end{aligned}$$

The series on the left has a radius of convergence greater than unity since $B^*(-\delta)A^*(\delta) < 1$ for an appropriate δ in $[0, \varepsilon]$. From (6) it follows that the series $G(x)$ also converges for an $x > 1$.

Conversely, if $G(x)$, and hence $\sum_1^\infty n^{-1}x^n(1-c_n)$, has a radius of convergence $r > 1$, then the series of derivatives $\sum_1^\infty x^n(1-c_n)$ also converges for $|x| < r$. Thus there exists an $x_0 > 1$ such that

$$\begin{aligned}\infty > \sum_1^\infty x_0^n(1-c_n) &= \sum_1^\infty x_0^n \int_0^\infty [1-B_n(y)]dA_n(y) \\ &\geqq \sum_1^\infty x_0^n \int_0^\infty [1-B(y)]dA_n(y) \\ &= \int_0^\infty \sum_1^\infty x_0^n A_n(y)dB(y).\end{aligned}$$

By the theorem of Belyaev and Maksimov [1], there exists a $y_0 > 1$ such that $\sum_1^\infty y_0^n A_n(t) < \infty$ for all finite t. Taking $z_0 = \min(x_0, y_0)$ we have

$$\infty > \int_0^\infty \sum_1^\infty z_0^n A_n(y)dB(y),$$

and the result of Lemma 2 implies that the integrals $\int_0^\infty \sum_1^\infty z_0^n A_n(y)dB(y)$ and $\int_0^\infty e^{\varepsilon y}dB(y)$ converge or diverge together, for an $\varepsilon > 0$. In fact, ε is the root in s of the equation $z_0A^*(s) = 1$.

Theorem 1 in conjunction with Lemma 1 is useful in studying the rate of convergence of $P(w_n \leqq y)$ to its limit. It is convenient to proceed in two stages.

THEOREM 2. *There exists an* $\alpha < 1$ *such that*

$$\lim_{n\to\infty} \frac{P(w_n = 0)-P(w = 0)}{\alpha^n} = 0$$

if and only if there exists an $\varepsilon > 0$ *such that* $B^*(-\varepsilon) < \infty$.

PROOF. From (7)—(10), for $n = 1, 2, \cdots$,

$$P(w_n = 0)-P(w = 0) = \sum_{r=0}^{n} [P(w_r = 0)-P(w = 0)]\gamma_{n-r}-\mu_1^{-1} \sum_{r=n+1}^{\infty} \gamma_r.$$

Let $Q(x) = \sum_{n=0}^\infty x^n[P(w_n = 0)-P(w = 0)]$. A simple calculation, using $P(w_0 = 0) = 1$ and $\gamma_0 = 0$, shows that

$$(11)\qquad \begin{aligned}Q(x) &= \frac{1}{1-G(x)} - \frac{1}{\mu_1(1-x)} \\ &= \frac{\mu_1 \exp[-\sum_1^\infty n^{-1}x^n(1-c_n)]-1}{\mu_1(1-x)}.\end{aligned}$$

Suppose now that $B^*(-\varepsilon) < \infty$ for an $\varepsilon > 0$. By the previous theorem this implies that the radius of convergence r of $\sum_1^\infty n^{-1}x^n(1-c_n)$ is greater than unity and, since $Q(1) = \mu_1^{-1}\sum_1^\infty (1-c_n) < \infty$, $Q(x)$ has a radius of convergence $r > 1$. Choosing α so that $r^{-1} < \alpha < 1$ we have also the convergence of the series $Q(1/\alpha)$, implying that

$$\alpha^{-n}[P(w_n = 0)-P(w = 0)] \to 0 \text{ as } n \to \infty.$$

Conversely, $P(w_n = 0)-P(w = 0) = o(\alpha^n)$ implies that $Q(x)$, and hence $\sum_1^\infty n^{-1}x^n(1-c_n)$, converges for an $x > 1$, and an appeal to Theorem 1 completes the proof.

THEOREM 3. *There exists an* $\alpha < 1$ *such that*

$$\lim_{n\to\infty} \frac{P(w_n \leqq y)-P(w \leqq y)}{\alpha^n} = 0$$

if and only if there exists an $\varepsilon > 0$ *such that* $B^*(-\varepsilon) < \infty$. *When the asserted result holds, it is true uniformly for* $y \geqq 0$.

PROOF. Let $R(x, y) = \sum_{n=1}^\infty x^n q_{n-1} P(U_n^+ \leqq y)$. From Lemma 1,

$$P(w_n \leqq y)-P(w \leqq y) = \sum_{k=1}^{n} q_{k-1}P(U_k^+ \leqq y)\{P(w_{n-k} = 0)-P(w = 0)\}$$
$$-\mu_1^{-1} \sum_{k=n+1}^{\infty} q_{k-1}P(U_k^+ \leqq y),$$

and hence

$$(12) \quad \sum_{n=1}^{\infty} x^n\{P(w_n \leqq y)-P(w \leqq y)\} = R(x, y)Q(x)-\frac{\{xR(1, y)-R(x, y)\}}{\mu_1(1-x)}.$$

Further, using (6),

$$R(x, y) \leqq \sum_1^\infty x^n q_{n-1} = x \exp\left\{\sum_1^\infty n^{-1}x^n(1-c_n)\right\}.$$

If we now suppose that $B^*(-\varepsilon)$ is convergent for an $\varepsilon > 0$ then, by Theorem 1 and (6), $R(x, y)$ has a radius of convergence greater than unity for all $y \geqq 0$, and by Theorem 2, so has $Q(x)$. Thus the radius of convergence R of the series on the left side of (12) is greater than unity. Selecting an α in the interval $(R^{-1}, 1)$ we have that $\sum_1^\infty \alpha^{-1}\{P(w_n \leqq y)-P(w \leqq y)\}$ converges uniformly in $y \geqq 0$ and hence $\alpha^{-1}\{P(w_n \leqq y)-P(w \leqq y)\} \to 0$ as asserted.

The converse follows immediately since if the right hand side of (12), and hence $Q(x)$, converges for an $x > 1$ then by Theorem 2, $B^*(-\varepsilon)$ converges for an $\varepsilon > 0$.

When the distribution $B(x)$ has only a finite number of moments these theorems fail and the rate of convergence of $P(w_n \leqq y)$ is of the order of some power of n, the power depending on the number of moments that are finite. This situation can be investigated by methods similar to those used above and we state here (without proof) only the following.

THEOREM 4. *If* $b_r = \int_0^\infty x^r dB(x) < \infty$ (and, as before, $b_1 < a_1$) *then as* $n \to \infty$

$$P(w_n \leqq y) - P(w \leqq y) = o(n^{-r+2}).$$

4. Concluding remarks

When the input process is Poisson and the service times are exponentially distributed, the busy period is given by

$$G(x) = \frac{(1+\rho)}{2\rho}\left\{1 - \sqrt{1 - \frac{4\rho x}{(1+\rho)^2}}\right\}$$

where $\rho = b_1/a_1$ is the traffic intensity [10].

The radius of convergence r is $(1+\rho)^2(4\rho)^{-1} > 1$ for $\rho \neq 1$. $G(r) = 1+\rho/2\rho$ and when $\rho < 1$, $G(1) = 1$. In the interval $[0, r]$, $G(x)$ is strictly increasing and attains its maximum value at $x = r$. By (11),

$$Q(r) = \sum_0^\infty \left\{\frac{(1+\rho)^2}{4\rho}\right\}^n \{P(w_n = 0) - P(w = 0)\} = \frac{2\rho}{1-\rho}$$

which is convergent for all $\rho < 1$. Thus, the best possible value of α in theorem 2 (and, by a similar argument, in Theorem 3) is

$$\alpha = \frac{1}{r} = \frac{4\rho}{(1+\rho)^2}.$$

This simple example leads one to the following result.

THEOREM 5. *If* $b_1 < a_1$, $r > 1$ *and* $G(r)$ *is convergent, then there does not exist a number* β *in* $[1/r, 1)$ *such that*

$$\frac{P(w_n = 0) - P(w = 0)}{\beta^n} \to l > 0.$$

PROOF. Suppose $\dfrac{P(w_n = 0) - P(w = 0)}{\beta^n} \to l > 0$ for some β in $[r^{-1}, 1)$. Then by (11),

$$0 < l = \lim_{x \to 1-} (1-x) \left[\frac{\mu_1(1-x/\beta) - \{1 - G(x/\beta)\}}{\mu_1(1-x/\beta)\{1 - G(x/\beta)\}}\right],$$

which implies that $G(1/\beta) = 1$. But this is impossible since $G(1) = 1$ and $G(1/\beta) > 1$ whenever $1 < \beta^{-1} \leqq r$.

In fact, under the hypotheses of the theorem, the result,

$$\frac{P(w_n = 0) - P(w = 0)}{\gamma^n} \to 0,$$

is the best possible.

It can be similarly proved that whenever $G(r)$ is convergent, Theorem 3 cannot be improved upon.

References

[1] Belyaev, Ya. K., and Maksimov, V. M., Analytical properties of a generating function for the number of renewals, Theory of Probability (SIAM Translation), 8 (1963), 102—104.

[2] Feller, W., An introduction to Probability theory and its applications, Vol. I (2nd ed.), Wiley, New York (1957).

[3] Finch, P. D., On the busy period in the queueing system GI/G/1, J. Austral. Math. Soc., 2 (1961), 217—228.

[4] Heathcote, C. R., Divergent single server queues. To appear in Proceedings of the Symposium on Congestion Theory, August, 1964, University of North Carolina Press.

[5] Heyde, C. C., Two probability theorems and their application to some first passage problems, J. Austral. Math. Soc., 4 (1964), 214—222.

[6] Kingman, J. F. C., On queues in heavy traffic, J. Roy. Stat. Soc. (B), 24 (1962), 383—392.

[7] Lindley, D. V., Theory of queues with a single server, Proc. Camb. Phil. Soc., 48 (1952), 277—289.

[8] Smith, W. L., Renewal Theory and its ramifications, J. Roy. Stat. Soc. (B), 20 (1958) 243—302.

[9] Spitzer, F., The Wiener-Hopf equation whose kernel is a probability density, Duke Math. J., Vol. 24 (1957) 327—344.

[10] Takács, L., Introduction to the theory of queues, Oxford University Press, New York (1962).

The Australian National University