The Annals of Probability
1992, Vol. 20, No. 2, 772–789

PHASE-TYPE REPRESENTATIONS IN RANDOM WALK AND QUEUEING PROBLEMS

By Søren Asmussen

Aalborg University

The distributions of random walk quantities like ascending ladder heights and the maximum are shown to be phase-type provided that the generic random walk increment X has difference structure $X = U - T$ with U phase-type, or the one-sided assumption of X_+ being phase-type is imposed. As a corollary, it follows that the stationary waiting time in a GI/PH/1 queue with phase-type service times is again phase-type. The phase-type representations are characterized in terms of the intensity matrix $\mathbf{Q}$ of a certain Markov jump process associated with the random walk. From an algorithmic point of view, the fundamental step is the iterative solution of a fix-point problem $\mathbf{Q} = \psi(\mathbf{Q})$, and using a coupling argument it is shown that the iteration typically converges geometrically fast. Also, a variant of the classical approach based upon Rouché's theorem and root-finding in the complex plane is derived, and the relation between the approaches is shown to be that $\mathbf{Q}$ has the Rouché roots as its set of eigenvalues.

1. Introduction. Consider a random walk $S_n = X_1 + \cdots + X_n$ with increment distribution $F(x) = \mathbb{P}(X_n \le x)$ and let $\tau_- = \inf\{n \ge 1\colon S_n \le 0\}$ and $\tau_+ = \inf\{n \ge 1\colon S_n > 0\}$ be the ladder epochs. A fundamental problem in a number of applied probability areas like sequential analysis ([28]), queueing theory ([11], [10], [21] and [2]) and risk theory ([1], [4] and [2]) is then to compute quantities like the ladder height distributions $G_-(x) = \mathbb{P}(S_{\tau_-} \le x)$ and $G_+(x) = \mathbb{P}(S_{\tau_+} \le x)$, the distribution of the maximum $M = \max_{n=0,1,\dots} S_n$ (assuming $\mu = \mathbb{E}X < 0$) and (assuming $\mu \ge 0$) the distribution H of the stationary excess overshoot, defined as the weak limit as $u \to \infty$ of $S_{\tau(u)} - u$ where $\tau(u) = \inf\{n \ge 1\colon S_n > u\}$. In fact, in view of formulas like

$$\mathbb{P}(M \le x) = (1 - \|G_+\|) \sum_{n=0}^{\infty} G_+^{*n}(x), \tag{1.1}$$

$$\frac{dH}{dx}(x) = \frac{1 - G_+(x)}{\int_0^\infty (1 - G_+(y))\,dy}, \tag{1.2}$$

these problems are closely related.

Transform-free explicit expressions can be found in the literature only in a few special cases like one of the tails of F being exponential or hyperexponen-

Received September 1988; revised February 1991.

AMS 1980 *subject classifications.* Primary 60J15; secondary 60J05, 60K25.

Key words and phrases. Random walk, ladder height distribution, phase-type distribution, Wiener–Hopf factorization, Markov jump process, nonlinear matrix iteration, coupling, uniformization, PH/G/1 queue, GI/PH/1 queue.

tial (a convex combination of exponential distributions). In more general cases, at most the transforms are usually given and are traditionally determined, for example as follows. One first looks for a Wiener–Hopf factorization

$$1 - r\hat{F}[s] = H_{-}(r,s)H_{+}(r,s), \tag{1.3}$$

where $|r| < 1$, $\hat{F}[s]$ is the moment generating function (m.g.f.)

$$\hat{F}[s] = \int_{-\infty}^{\infty} e^{sx}F(dx), \tag{1.4}$$

which is defined at least when $\Re(s) = 0$, and $H_{-}(r,s)$ and $H_{+}(r,s)$ are functions that are bounded and analytical in the positive (resp., the negative) complex half-plane for fixed r. In the continuous case, this is typically possible if one tail of F has a rational function m.g.f. and $H_{-}(r,s)$ and $H_{+}(r,s)$ are then expressed in terms of the roots $\rho_1(r), \dots, \rho_d(r)$ of the equation $1 = r\hat{F}[\rho]$, which are counted and located by an application of Rouché's theorem. The desired quantities $\hat{G}_{-}(s)$ and $\hat{G}_{+}(s)$ are then obtained by normalization of $\lim_{r\uparrow 1} H_{-}(r,s)$ and $\lim_{r\uparrow 1} H_{+}(r,s)$. For examples of this approach, see [29], [16], Chapter 13, [9], [10], [21], [11] and [30]. Note, however, that the argument and the setting admit a number of variants; thus some of these references do not mention ladder heights at all, others determine the number of roots by means of contour integration and in much queueing literature one looks directly for the waiting time distribution without exploiting its Pollaczek–Khintchine representation (1.1).

A number of objections may, however, be raised (and have caused Wiener–Hopf methods to lose popularity in queueing theory):

1. Complex plane methods are always hard to interpret probabilistically and, in the present case, even more so since in the equation $1 = r\hat{F}[\rho]$ it is crucial that $\hat{F}[\rho]$ denotes the analytical continuation of the integral in (1.4) [which is thus in general not well defined when $s = \rho_i(r)$].
2. Complex variables also may cause problems from the computational point of view since complex arithmetic, in general, requires special software (typically not available, say, in the standard programming languages on microcomputers).
3. One would like expressions for the ladder height distributions themselves rather than the transforms. Even in examples where the form of G_{-} or G_{+} can be recognized immediately by a probabilistic argument (like phase-type distributions, see [2], page 216), it requires a cumbersome fractional expansion of the transform to find the parameters.
4. At least in some variants of the approach, there are problems with the limit $r \uparrow 1$. The difficulty is that though $\hat{G}_{-}(s)$ and $\hat{G}_{+}(s)$ can typically be expressed in terms of the roots $\rho_i = \lim_{r\uparrow 1} \rho_i(r)$ of the equation $1 = \hat{F}[\rho]$, then Rouché's theorem only applies when $|r| < 1$ and there is no guarantee that the ρ_i are distinct, which is crucial for the method. In our opinion, this problem has never been completely understood and is frequently simply neglected.

The purpose of the present paper is to present a different approach which circumvents some of these difficulties. The idea is to restrict the discussion to phase-type distributions (Neuts [19], or [2], Chapter III.6). This class of distributions is slightly less general than those having a rational m.g.f., but nevertheless comprises all standard examples of such distributions, is dense and has become the standard setup in modern applied probability. Within this setting, we obtain a solution of the random walk problems which is transform-free, avoids complex numbers (the computations are instead based on matrix manipulations) and has the appealing feature that many of the basic unknown distributions turn out to be again of phase-type.

Compared to the extensive queueing literature on matrix-geometric methods (see Neuts [19] and [20] for surveys and comprehensive bibliographies), the common features are the phase-type assumptions and the role of nonlinear matrix equations (fix-point problems) of the form $\mathbf{Q} = \psi(\mathbf{Q})$ (surveyed in Ramaswami [22]). What differs is that we deal directly with continuous distributions (the maximum/waiting time) instead of taking a detour via discrete ones (queue lengths). In particular, our main result for the GI/PH/1 queue (phase-type service times) states that the waiting time distribution is again phase-type. This result is much simpler than those of the matrix-geometric literature (Neuts [19]; Lucantoni and Ramaswami [17], [24] and [25]) and, in fact, it has come as a surprise leading to subsequent papers like [23], [27] and [6] on phase-type solutions of waiting time problems. Note, however, that in view of the simplicity of proof we do not consider this result to be deep (in fact, our study of the corresponding nonlinear matrix iteration scheme is technically much more difficult; see Section 3), and that, from a purely computational point of view, Sengupta's [26] solution of the GI/PH/1 queue is closely related.

The paper is organized as follows. The main results are stated in Section 2 and proved to the extent that this provides a reasonably simple intuitive explanation. The idea is to define an imbedding of the random walk in the continuous component $\{Y_t\}$ of a certain bivariate Markov process $\{(J_t, Y_t)\}_{t \geq 0}$ with a discrete first component $\{J_t\}$ closely related to the phase representation. Observing $\{J_t\}$ only when $\{Y_t\}$ is at its maximum then leads to a new Markov process $\{m_x\}_{x \geq 0}$, whose intensity matrix $\mathbf{Q}$ (say) turns out to be the basic unknown quantity determining the solution of the random walk problems. The fundamental equation $\mathbf{Q} = \psi(\mathbf{Q})$ is derived and we show how the ladder height distributions and the maximum can be expressed in terms of $\mathbf{Q}$. In Section 3, we study the equation $\mathbf{Q} = \psi(\mathbf{Q})$ in more detail by means of a coupling argument which is nonstandard in this setting. Section 4 contains a discussion of the relation of the present approach to the classical Rouché root algorithm. In this setting, our point of view is then that the Rouché roots $\rho_1, \dots, \rho_d$ are important, not per se, but simply because the $-\rho_i$ are the eigenvalues of $\mathbf{Q}$ and $\mathbf{Q}$ can be determined once they are known (this observation also leads to a variant of the classical approach by passing from the ρ_i not to $\hat{G}_-$ and $\hat{G}_+$, but rather via $\mathbf{Q}$ to transform-free expressions for G_- and G_+). Finally, some extensions of the results and some concluding discussion are given in Section 5.

For the sake of completeness, we mention finally that the set of problems of the present paper has been attacked by at least one additional completely different approach, where formulas for ladder height functionals are given in terms of numerical integration of the characteristic function of F; see, for example [8], [28] and [31].

2. Main results. We assume for the main part of the paper (see, however, Section 5) that F is a difference distribution, $X = U - T$ with U, T independent with distribution functions B (resp., A) (in queueing theory, A is the interarrival distribution and B is the service time distribution). We further assume that the distribution B of U is of phase-type with representation, say, $(\pi, \mathbf{T}, d)$. This means that $\mathbf{T}$ is the restriction to $\{1, \dots, d\}$ of the intensity matrix

$$\left(\begin{array}{c|c} 0 & \mathbf{0} \\ \hline t_0 & \mathbf{T} \end{array}\right) = \left(\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \hline t_{10} & t_{11} & t_{12} & \cdots & t_{1d} \\ t_{20} & t_{21} & t_{22} & \cdots & t_{2d} \\ t_{d0} & t_{d1} & t_{d2} & \cdots & t_{dd} \end{array}\right)$$

of a Markov jump process on $\{0, \dots, d\}$ with 0 as absorbing state, and that U is distributed as the time to absorption in 0 given the initial distribution π (written as a row vector in the following). In particular, the vector t_0 of exit rates is $t_0 = -\mathbf{T}e$, where e is the column vector of ones, and

$$B(x) = \int_0^\infty B(dx) = \pi e^{\mathbf{T}x} e, \qquad b(x) = B'(x) = xe^{\mathbf{T}x} t_0, \tag{2.1}$$

$$\hat{B}[s] = \int_0^\infty e^{sx} B(dx) \tag{2.2}$$

$$= 1 - \pi\left(\mathbf{I} + \frac{\mathbf{T}}{s}\right)^{-1} e = \pi(-s\mathbf{I} - \mathbf{T})^{-1} t_0, \tag{2.3}$$

$$\mu_B^{(n)} = \int_0^\infty x^n B(dx) = (-1)^n n! \pi \mathbf{T}^{-n} e. \tag{2.4}$$

See [19] or [2], Chapter III.6 for more detail. Since d is fixed in the following, we omit the qualification in the specification $(\pi, \mathbf{T})$ of a phase-type distribution. Note that by analytic continuation we can interpret $\hat{B}[s]$ as defined and given by (2.3) [but not (2.2)] at least for $s \notin \mathrm{sp}(\mathbf{T})$ (a similar remark applies to $\hat{F}[s] = \hat{A}[-s]\hat{B}[s]$ when $\Re(s) > 0$). This fact plays a role in the traditional analytical approach [with the exponential set sp(**T**) usually characterized as the set of poles for $\hat{B}[s]$] and will be needed in Section 4. We write $\mu_B = \mu_B^{(1)}$ and use similar notation at other places so that, for example,

$$\mu = \mu_F = \mu_B - \mu_A = -\pi \mathbf{T}^{-1} e - \mu_A. \tag{2.5}$$

We now define what we call a $(\pi, \mathbf{T})$ *phase process*, a concept which will play a fundamental role in the following. This is a Markov process with state space $\{1, \dots, d\}$ and jump rates given by the intensity matrix $\mathbf{T} + t_0\pi = \mathbf{T} + t_0 \otimes \pi$.

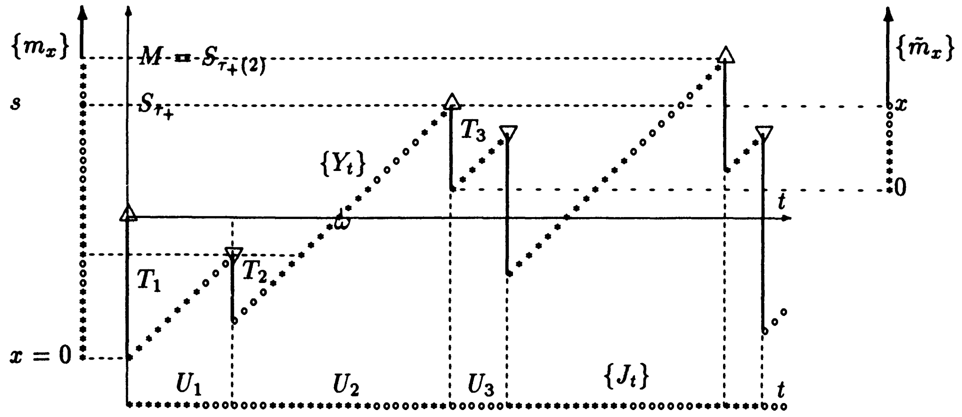

FIG. 1.

We say that jumps according to the $\mathbf{T}$ component are of the *first type* and those governed by $t_0\pi$ are of the *second type*. The initial condition is a jump of the second type at (or rather just before) $t = 0$ so that J_0 has distribution π. Thus, jumps of the second type form a phase-type renewal process and we shall identify the interarrival times with the U_i (alternatively, one may, as in [19], visualize jumps of the second type as visits to the instantaneous state 0). In particular, from standard Markov process theory it follows that the renewal measure $U_B(dx)$ is given by an atom at zero and the density

$$u_B(x) = \pi \exp((\mathbf{T} + t_0\pi)x)t_0 \tag{2.6}$$

on $(0, \infty)$ [we do not claim priority for (2.6), but the formula seems certainly not as well known as it deserves to be].

The phase process $\{J_t\}$ can be extended to an imbedding of the random walk in a bivariate continuous time process $\{(J_t, Y_t)\}_{t \geq 0}$ in the way indicated on Figure 1. Here $\{Y_t\}$ starts at zero ($Y_{0-} = 0$) but has an instantaneous downward jump of size T_1 ($Y_0 = -T_1$), a further downward jump of size T_2 at time U_1, the next downward jump of size T_3 at time $U_1 + U_2$ and so on, and in between jumps $\{Y_t\}$ increases linearly at a unit rate. The random walk is then obtained by observing $\{Y_t\}$ just before jumps, that is, $S_n = Y_{U_1 + \cdots + U_n -}$. The values are marked with the symbols ▽ and △ on Figure 1, the △ representing ladder points (on the figure, there are two ladder points following zero and thus $M = S_{\tau_+(2)}$). We let $\{m_x\}$ be the process obtained by observing $\{J_t\}$ only when $\{Y_t\}$ is at its maximum, that is,

$$m_x = J_{\delta^{-1}(x)} \quad \text{where} \quad \delta(T) = \int_0^T I(Y_t \geq Y_s, 0 \leq s \leq t)\, dt.$$

Note that $Y_t \to -\infty$ a.s. when $\mu < 0$ and then $\{m_x\}$ is nonconservative (terminating). The following result is basic for the paper:

PROPOSITION 2.1. *$\{m_x\}$ is a Markov jump process on $\{1,\dots,d\}$ and the intensity matrix $\mathbf{Q}$ (say) solves the nonlinear matrix equation*

$$\mathbf{Q} = \psi(\mathbf{Q}) \quad \text{where} \quad \psi(\mathbf{S}) = \mathbf{T} + t_0\pi\hat{A}[\mathbf{S}]. \tag{2.7}$$

Furthermore, $\mathbf{Q}$ is conservative, $\mathbf{Q}e = 0$, precisely when $\mu \geq 0$, whereas otherwise $\mathbf{Q}e \leq 0$ with at least one component strictly negative.

PROOF. It is probabilistically obvious that $\{m_x\}$ is Markov with piecewise constant paths. In particular, $\{m_x\}$ is uniquely determined by the intensity matrix $\mathbf{Q}$. To compute $\mathbf{Q}$, note that if $\{m_x\}$ jumps from i to $j \neq i$ at time s (say), then $\{J_t\}$ must jump out of state i at time $\delta^{-1}(s)$. If the jump is of the first type, it must be to j and, hence, the contribution to q_{ij} from jumps of the first type is t_{ij}. Suppose, on the other hand, that the jump is of the second type, say $\{J_t\}$ is reset to state k. Such jumps occur at rate $t_{i0}\pi_k$, and if x is the size of the corresponding jump of $\{Y_t\}$, the excursion of $\{Y_t\}$ below the level just before the jump gives rise to a new process $\{\tilde{m}_t\}$ distributed as $\{m_t\}$ and with $\tilde{m}_0 = k$, such that $m_s = \tilde{m}_x$ (see Figure 1); that is, m_s is distributed according to the kth row of $e^{\mathbf{Q}x}$. Hence jumps of the second type contribute the amount

$$\sum_{k=1}^{d} t_{i0}\pi_k \int_0^\infty (e^{\mathbf{Q}x})_{kj} A(dx)$$

to q_{ij}. An easy modification shows that this is correct also when $j = i$, proving (2.7). The last statement follows by standard random walk results. □

Here is our first main result:

THEOREM 2.1. *Consider a random walk with $X = U - T$, with U phase-type with representation $(\pi, \mathbf{T})$ and $A(x) = \mathbb{P}(T \leq x)$ arbitrary. Then:*

(a) *The ascending ladder height distribution G_+ is phase-type with representation $(\pi_+, \mathbf{T})$ where π_+ is related to the matrix $\mathbf{Q}$ of Proposition 2.1 by means of*

$$\mathbf{Q} = \mathbf{T} + t_0\pi_+. \tag{2.8}$$

Also, π_+ satisfies

$$\pi_+ = \pi\hat{A}[\mathbf{T} + t_0\pi_+]. \tag{2.9}$$

Further, $\pi_+ e = 1$ if and only if $\mu = \mathbb{E}X \geq 0$ whereas $\pi_+ e < 1$ when $\mu < 0$.

(b) *If $\mu < 0$, then the maximum M has distribution given by an atom of size $1 - \pi_+ e$ at zero and a (defective) phase-type distribution with representation $(\pi_+, \mathbf{Q})$ on $(0, \infty)$.*

PROOF. That G_+ is phase-type is noted in [2], pages 216–217. This is also seen directly from Figure 1: If $\omega = \inf\{t > 0\colon Y_t = 0\}$, then obviously the statement holds true with π_+ being the distribution of $J_\omega = m_{T_1}$. That is, since m_0 has distribution π, we have

$$\pi_+ = \pi\int_0^\infty e^{\mathbf{Q}x}A(dx) = \pi\hat{A}[\mathbf{Q}], \tag{2.10}$$

proving (2.9). Equation (2.8) now follows from (2.7). The rest of (a) is easy.

For (b), we note that M is simply the lifetime of the post-T_1 segment of the Markov process $\{m_x\}$. The result follows by noting that the initial distribution (which is defective) is π_+ and that the intensity matrix is $\mathbf{Q}$. □

From (2.8), we get:

Corollary 2.1. *Given* $\mathbf{Q}$, π_+ *can be computed as* $\pi_+ = \nu(\mathbf{Q} - \mathbf{T})/\nu t_0$ *where* ν *is any column vector* (*say* e *or* π) *with* $\nu t_0 \neq 0$.

Thus, combining (2.8), (2.9) and Corollary 2.1, it is seen that the main computational problem is to compute either $\mathbf{Q}$ or π_+. In our opinion, the most natural approach is to solve (2.9) iteratively for π_+, and we have:

Theorem 2.2. *Let* $\pi_+^{(0)} \geq 0$ *be some given vector and define* $\pi_+^{(n+1)} = \pi\hat{A}[\mathbf{Q} + t_0\pi_+^{(n)}]$, $n = 0, 1, \dots$. *Then* $\pi_+^{(n)} \to \pi_+$, $n \to \infty$, *provided*: (a) $\mu \geq 0$ *and* $\pi_+^{(0)}$ *is a subprobability vector* (*i.e.*, $\pi_+^{(0)}e \leq 1$); *or* (b) $\mu < 0$ *and* $\pi_+^{(0)}h \leq 1$ *where* $h = (-\gamma\mathbf{I} - \mathbf{T})^{-1}t_0$ *with* γ *the unique solution greater than* 0 *of the equation* $\hat{A}[-s]\hat{B}[s] = 1$. *In particular, the solution of* (2.9) *is unique in the class of nonnegative vectors satisfying the given constraints.*

The proof is more technical than those of the present section and is deferred to Section 3 (error bounds are given there also; some discussion of computational aspects is in Section 5). An alternative computational procedure based on the Rouché roots is developed in Section 4.

Corollary 2.2. *Consider the GI/PH/1 queue with interarrival distribution A and service times U which are of phase-type with representation* $(\pi, \mathbf{T})$, *assume* $\rho = \mathbb{E}U/\mathbb{E}T < 1$, *and let* $\pi_+, \mathbf{Q}$ *be as above. Then*:

(a) *The distribution of the steady-state actual waiting time W is given by an atom of size* $1 - \pi_+e$ *at zero and a* (*defective*) *phase-type distribution with representation* $(\pi_+, \mathbf{Q})$ *on* $(0, \infty)$.

(b) *The distribution of the steady-state virtual waiting time V is given by an atom of size* $1 - \rho$ *at zero and a* (*defective*) *phase-type distribution with representation* $(\rho\nu, \mathbf{Q})$ *on* $(0, \infty)$. *Here* ν *is the stationary distribution of the phase process* $\{J_t\}$; *that is*, $\nu(\mathbf{T} + t_0\pi) = \nu$, $\nu e = 1$.

(c) *The distribution of the steady-state sojourn time* W^* *is phase-type with representation* $(\pi, \mathbf{Q})$ *on* $(0, \infty)$.

Proof. Part (a) is immediate from $W =_{\mathscr{D}} M$. In (b), $\mathbb{P}(V = 0) = 1 - \rho$ is standard. Also ([2], page 189), the conditional distribution of V given $V > 0$ is that of $T^* + W$ where T^* is independent of W and has the limiting stationary excess distribution B_0 corresponding to B. Here inspection of the phase process $\{J_t\}$ shows immediately that B_0 is phase-type with representation $(\nu, \mathbf{T})$. Thus, we may visualize $T^* + W$ as the lifetime of a particle which initially (the T^* phase) starts according to ν, moves according to $\mathbf{T}$ and exits according to t_0 to be restarted (the W phase) according to π_+ and then goes on moving in this way until the final exit. This describes, however, just a phase-type distribution with representation $(\nu, \mathbf{T} + t_0\pi_+) = (\nu, \mathbf{Q})$. Part (c) follows in a similar manner from $W^* = U + W$ where U has the service time distribution and is independent of W. □

2.1. *Descending ladder heights. PH/G/1.* We shall need the Wiener–Hopf factorization identity ([2], Chapter VII.4; see also [3] for a slightly simpler proof).

Lemma 2.1. (a) $1 - \hat{F} = (1 - \hat{G}_-)(1 - \hat{G}_+)$.

(b) *Let* $U_+ = \sum_0^\infty G_+^{*n}$ *be the renewal measure associated with* G_+. *Then* G_- *is the restriction of* $U_+ * F$ *to* $(-\infty, 0)$.

Corollary 2.3. *Under the assumptions of Theorem* 2.1 *and with* $\pi_+, \mathbf{Q}$ *as defined there, the descending ladder height distribution is given by the density*

$$g_-(x) = \int_{-x}^{\infty} \pi e^{\mathbf{Q}(u+x)} t_0 A(du), \qquad x < 0. \tag{2.11}$$

If $\mu = \mathbb{E}X > 0$, *then furthermore*

$$1 - \|G_-\| = \frac{\mu}{-\pi_+ \mathbf{T}^{-1} e}. \tag{2.12}$$

Proof. A similar argument as in the remarks leading to (2.6) shows that U_+ is given by an atom at zero and the density $\pi_+ e^{\mathbf{Q}y} t_0$, $y > 0$. Arguing as in the proof of Corollary 2.2 it follows that $U_+ * B$ is given by the density $\pi e^{\mathbf{Q}y} t_0$, $y > 0$. Let C be the distribution of $-T$ so that $F = B * C$. Then, according to Lemma 2.1(b), the density of G_- is obtained by convolution of the density of $U_+ * B$ with C which immediately leads to (2.11). Finally ([2], page 169, and Wald's identity)

$$1 - \|G_-\| = \frac{1}{\mathbb{E}\tau_+} = \frac{\mathbb{E}X}{\mathbb{E}S_{\tau_+}} = \frac{\mu_F}{-\pi_+ \mathbf{T}^{-1} e}. \qquad \square$$

Corollary 2.4. *Consider the PH/G/1 queue with service time distribution* B^* *and interarrival time distribution* A^*, *say, and assume that* A^* *is phase-type with representation* $(\pi, \mathbf{T})$ *and that* $\rho = \mu_{B^*}/\mu_{A^*} < 1$. *Then the solution* π_- *to*

$$\pi_- = \pi \hat{B}^*[\mathbf{T} + t_0 \pi_-] \tag{2.13}$$

exists, is unique and can be computed by iteration starting from any subproba-

bility vector $\pi_-^{(0)}$. *Let further* $\mathbf{Q} = \mathbf{T} + t_0\pi_-$ *and let* D, γ *be given by*

$$\text{(2.14)}\qquad \begin{aligned} \frac{dD}{dx}(x) &= \int_x^\infty \pi e^{\mathbf{Q}(u-x)} t_0 B^*(du), \qquad x > 0,\\ \gamma = \|D\| &= 1 - \frac{\mu_{B^*} - \mu_{A^*}}{-\pi_- \mathbf{T}^{-1} e}. \end{aligned}$$

Then the distribution of the steady-state waiting time W is the normalized renewal measure

$$\text{(2.15)}\qquad (1-\gamma)\sum_{n=0}^{\infty} D^{*n}.$$

Furthermore, the idle period I is phase-type with representation $(\pi_-, \mathbf{T})$.

PROOF. In the setting of Theorems 2.1 and 2.2, let $A = B^*$ and $B = A^*$. Then W is distributed as $-\min\{S_n\}$, and from this the first part of the corollary follows by sign reversion of (2.11), noting that D is the distribution of $-S_{\tau_-}$ and invoking Corollary 2.3 and the Pollaczek–Khintchine formula (1.1). The result on I follows from a well-known ladder height representation of the idle period (e.g., [2], page 182). □

One may note the similarity of (2.14) and (2.15) to the standard solution of the M/G/1 queue. Here A^* is just exponential, with rate β say, so that the renewal density $\pi e^{\mathbf{Q}y} t_0$ reduces to β and D to the stationary excess distribution for B^* multiplied by ρ. The distribution of I is useful, for example, for deriving moments of W; see [2], pages 185–186 (the approach is due to Marshall in the queueing setting but essentially the same comes out by relating the moments of X, M, S_{τ_+} and S_{τ_-}). This was carried out independently in the first version of this paper and in Neuts [20], but we omit the details here since the formulas can be found in [20].

3. Nonlinear matrix iteration.

The proof of results like Theorem 2.2 on the iteration scheme

$$\text{(3.1)}\qquad \pi_+^{(n+1)} = \pi\hat{A}\big[\mathbf{T} + t_0\pi_+^{(n)}\big], \qquad n = 0, 1, \dots,$$

is frequently carried out by arguments involving fix-points and contractions. So far this has, however, not proven to be useful in the matrix-geometric setting where the standard approach instead is by monotonicity (starting from $\pi_+^{(0)} = 0$); see Neuts [19], [20] and Ramaswami [22]. We shall use a third approach, namely, coupling, which has not yet been implemented in this setting before but turns out to produce results which are stronger than those of the matrix-geometric literature by allowing a general (nonzero) $\pi_+^{(0)}$ and also containing error estimates (see Corollaries 3.1, 3.2 and 3.3).

Recall that a $(\pi, \mathbf{T})$ phase process is a Markov process with jump rates given by $\mathbf{T} + t_0\pi$ and started according to the distribution π of J_0. If instead J_0 has distribution ν, we talk about a ν-delayed $(\pi, \mathbf{T})$ process. As on Figure 1, ω

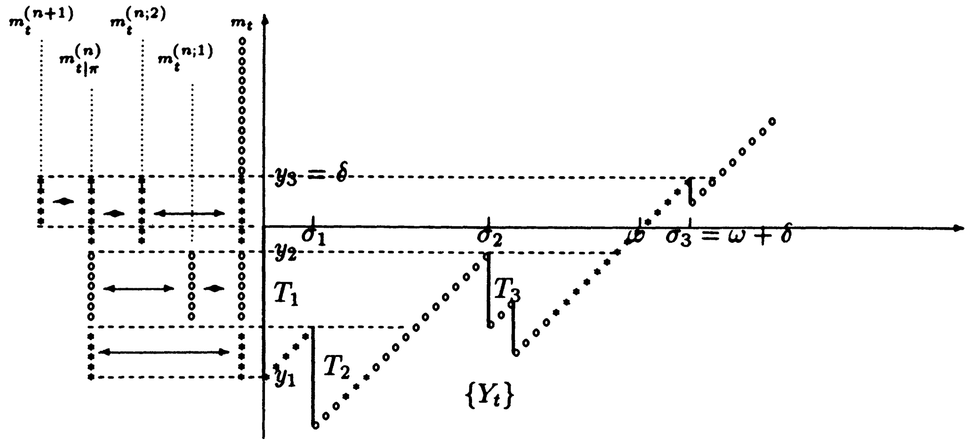

FIG. 2.

denotes the time of the first up-crossing of level zero of $\{Y_t\}$, and we let δ be the time until the next jump of the second type (see Figure 2). The key step in our approach is the following:

LEMMA 3.1. *It is possible to enlarge the probability space on which* $\{J_t, Y_t\}$ *and* $\{m_x\}$ *are defined, such that for any* $n = 1, 2, \dots,$ *there are defined two additional processes* $\{m_t^{(n)}\}$ *and* $\{m_{t|\pi}^{(n)}\}$ *which are independent of* T_1 *and have the following properties*:

(a) $\{m_t^{(n)}\}$ *is a* $(\pi_+^{(n)}, \mathbf{T})$ *phase process and if* D_n *is the set on which* $m_0^{(n)}$ *is defined, then*

$$(3.2) \qquad \{m_t^{(n)}\}_{t<\delta} = \{m_{T_1+t}\}_{t<\delta} \quad \textit{on } D_n \cap \{\tau_+ \le n\}.$$

(b) $\{m_{t|\pi}^{(n)}\}$ *is a* π*-delayed* $(\pi_+^{(n)}, \mathbf{T})$ *phase process and*

$$(3.3) \qquad \{m_{t|\pi}^{(n)}\}_{t<T_1+\delta} = \{m_t\}_{t<T_1+\delta} \quad \textit{on } \{\tau_+ \le n+1\}.$$

The proof is by induction and is carried out in the following three parts:

Construction of $\{m_t^{(1)}\}$. Define $\theta_n = \|\pi_+^{(n)}\| = \mathbb{P}D_n$. Obviously,

$$(3.4) \qquad \{\tau_+ \le 1\} = \{\tau_+ = 1\} = \{T_1 < U_1\},$$

$$(3.5) \qquad \pi_+^{(1)} = \pi\hat{A}\left[\mathbf{T} + t_0\pi_+^{(0)}\right] \ge \pi\hat{A}[\mathbf{T}],$$

$$(3.6) \qquad \theta_1 \ge \pi\hat{A}[\mathbf{T}]e = \mathbb{P}(T_1 < U_1) = \tilde{\theta}_1 \quad \text{(say)},$$

noting in (3.6) that $\pi\hat{A}[\mathbf{T}]$ is precisely the distribution of m_{T_1} restricted to the set (3.4). Thus on (3.4), we can let $m_t^{(1)} = m_{T_1+t}$, when $t < \delta = U_1 - T_1$, and continue $\{m_t^{(1)}\}$ after δ in some arbitrary manner according to requirements for a $(\pi_+^{(1)}, \mathbf{T})$ phase process. On an additional set of probability $\theta_1 - \tilde{\theta}_1$, we then need to start $\{m_t^{(1)}\}$ according to $\pi_+^{(1)} - \pi\hat{A}[\mathbf{T}]$.

Construction of $\{m^{(n)}_{t|\pi}\}$ *from* $\{m^{(n)}_t\}$. This is the key step, and the argument is illustrated on Figure 2. Here $*$ and $\circ$ in the graph of $\{Y_t\}$ indicate the corresponding states of $\{J_t\}$ in the same manner as in Figure 1, and $\sigma_1, \sigma_2, \dots$ are random times defined recursively by $\sigma_1 = U_1$ and by letting σ_{k+1} be the time of the next downward jump of $\{Y_t\}$ following the first up-crossing after σ_k of level $y_k = Y_{\sigma_k -}$. We then obviously have

$$(3.7) \qquad \big\{\big(J_{\sigma_k+t}, Y_{\sigma_k+t} - y_k\big)\big\}_{t<\sigma_{k+1}-\sigma_k} =_{\mathscr{D}} \{(J_t, Y_t)\}_{t<\delta}.$$

Thus, we may construct versions $\{m^{(n;1)}_t\}, \{m^{(n;2)}_t\}, \dots$ of $\{m^{(n)}_t\}$, such that $\{m^{(n;k)}_t\}$ is associated with the l.h.s. of (3.7) in the same way as $\{m^{(n)}_t\}$ is associated with $\{(J_t, Y_t)\}$ (thus, on the figure, $\{m^{(n;2)}_t\}$ is associated with $T_3, U_3, T_4, U_4, \dots$ in the same way as $\{m^{(n)}_t\}$ with $T_1, U_1, T_2, U_2, \dots$). Up to the first jump of the second type (i.e., for $t < \sigma_1$), we just let $m^{(n)}_{t|\pi} = m_t = J_t$. Up to the second jump of the second type, we substitute the values of $m^{(n;1)}_t$; then the values $m^{(n;2)}_t$ are used, and so on (on Figure 2, segments of processes matching by construction are connected with double arrows; other segments are dotted, so that the states are not marked). It can then readily be checked that $\{m^{(n)}_{t|\pi}\}$ has the required jump rates. That (3.3) holds follows from the induction assumption since on $\{\tau_+ \le n+1\}$ any of the excursions determining $\{m^{(n)}_{t|\pi}\}$ up to time $T_1 + \delta$ can at most correspond to n random walk steps, and hence by (3.2) they are coupled to $\{m_t\}$ in the desired manner.

Construction of $\{m^{(n+1)}_t\}$ *from* $\{m^{(n)}_{t|\pi}\}$. We just let $m^{(n+1)}_t = m^{(n)}_{T_1+t|\pi}$, $t < \delta$, and continue in the required way after δ in some arbitrary manner. To see that this provides the required properties, we only need to check that $m^{(n+1)}_0 = m^{(n)}_{T_1|\pi}$ has the correct distribution; that is, $\pi^{(n+1)}_+$. But the distribution of $m^{(n)}_{T_1|\pi}$ is

$$(3.8) \qquad \int_0^\infty \pi \exp\big((\mathbf{T} + t_0\pi^{(n)}_+)t\big)A(dt) = \pi\hat{A}\big[\mathbf{T} + t_0\pi^{(n)}_+\big] = \pi^{(n+1)}_+. \qquad \square$$

COROLLARY 3.1. $\|\pi^{(n)}_+ - \pi_+\| \le \mathbb{P}(\tau_+ > n)$ [*here* $\|\cdot\|$ *means the supremum* (*total variation*) *norm*].

PROOF. This follows immediately from (3.2) since m_{T_1} and $M^{(n)}_0$ have distributions $\pi\hat{A}[\mathbf{T} + t_0\pi] = \pi_+$ (resp., $\pi^{(n)}_+$). $\square$

When $\mu \ge 0$, Corollary 3.1 completes the proof of Theorem 2.2 since then $\tau_+ < \infty$ a.s. so that $\mathbb{P}(\tau_+ > n)$ converges to zero. Also convergence rates and explicit error bounds are obtained with conditions and constants which are familiar from large deviations theory (e.g., [7] or [2], Chapter XII.2):

COROLLARY 3.2. *Suppose* $\mu > 0$ *and that* $\hat{F}[s] = \hat{A}[-s]\hat{B}[s]$ *has a minimum at* $\eta < 0$. *Then* $\|\pi^{(n)}_+ - \pi_+\| \le C\delta^{n+1}$ *where* C^{-1} *is the minimal component of the vector* $(-\eta\mathbf{I} - \mathbf{T})^{-1}t_0$ *and* $\delta = \hat{F}[\eta]$.

PROOF. It follows from a version of Wald's fundamental identity ([2], page 267) that $\mathbb{E}e^{\eta S_{\tau_+}}/\delta^{\tau_+} = 1$. Obviously, the conditional distribution of S_{τ_+} given $\tau_+ = k$ is phase-type $(\pi_+(k), \mathbf{T})$ for some proper distribution $\pi_+(k)$; hence,

$$\mathbb{E}\big[e^{\eta S_{\tau_+}} | \tau_+ = k\big] = \pi_+(k)(-\eta\mathbf{I} - \mathbf{T})^{-1}t_0 \geq C^{-1}$$

so that $\mathbb{E}\delta^{-\tau_+} \leq C$, which gives the desired bound on $\mathbb{P}(\tau_+ > n)$. □

As a comparison, one may note the asymptotic formula $\mathbb{P}(\tau_+ > n) \approx \tilde{C}\delta^n/n^{3/2}$, $n \to \infty$, shown by Iglehart [14] (here $\tilde{C}$ is a constant). One might conjecture from Corollary 3.2 that the mapping defined by (3.1) is a contraction, but we have no proof of this. Inspection of the graph of $\hat{F}[s]$ shows that if μ is close to zero, then so is γ and hence δ is close to one; that is, the speed of convergence deteriorates as $\mu \to 0$ (heavy traffic). If $\mu = 0$, Corollary 3.2 does not provide any information on convergence rates, and we have personal experience from a related setting that the iteration may converge exceedingly slowly. Note finally also that even without the existence of exponential moments of A required when $\mu > 0$, we can still get convergence rates. For example, if A has a finite pth moment, it follows from Corollary 3.1 and results of Gut ([12] and [13]) that $\|\pi_+^{(n)} - \pi_+\| = O(n^{-p})$.

We now turn to the case $\mu < 0$. The problem which arises in the proof given above for $\mu \geq 0$ is infinite excursions, that is, taking into account what happens if $\{(J_t, Y_t)\}$ never returns to the level just before a jump of the second type. In fact, the results cannot be the same for the two cases since, for example, the equation (2.13) can be seen to have at least two solutions in the class of subprobability measures when $\mu < 0$, namely, π_+ which has $\pi_+ e < 1$ and an additional vector $\tilde{\pi}_+$ with $\tilde{\pi}_+ e = 1$ (take $\pi_+^{(0)} e = 1$; note that then $\pi_+^{(n)} e = 1$ for all n and use a compactness argument). Using a familiar transformation from random walk theory (the "associated random walk" of Feller [11], Chapter XII or the "Lundberg conjugate" of [2], Chapter XII), we shall, however, quite easily reduce the proof for $\mu < 0$ to what has already been proved for $\mu > 0$. Let γ, h be as in Theorem 2.2(b), let Δ be the diagonal matrix with the elements of h on the diagonal and let $\tilde{A}$, $\tilde{B}$ be the distributions with densities $e^{-\gamma t}/\hat{A}[-\gamma]$ and $e^{\gamma u}/\hat{B}[\gamma]$ w.r.t. A (resp., B). We shall need the following result from [5]:

LEMMA 3.2. *$\tilde{B}$ is phase-type with representation $(\tilde{\pi}, \tilde{\mathbf{T}})$ where $\tilde{\mathbf{T}} = \Delta^{-1}\mathbf{T}\Delta + \gamma\mathbf{I}$ and $\tilde{\pi} = \pi\Delta/\hat{B}[\gamma]$. The corresponding exit rate vector is $\tilde{t}_0 = \Delta^{-1}t_0$.*

PROOF OF THEOREM 2.2 WHEN $\mu < 0$. Let $\pi_+^{(0)}$ be given, define $\pi_+^{(1)}, \pi_+^{(2)}, \dots$ by (3.1) and let $\tilde{\pi}_+^{(n)} = \pi_+^{(n)}\Delta$. Then, using Lemma 3.2,

$$\begin{aligned}\pi_+^{(n+1)} &= \pi\hat{A}\big[\mathbf{T} + t_0\pi_+^{(n)}\big] \\ &= \tilde{\pi}\Delta^{-1}\hat{B}[\gamma] \cdot \hat{A}[-\gamma]\hat{\tilde{A}}\big[\mathbf{T} + t_0\pi_+^{(n)} + \gamma\mathbf{I}\big] \\ &= \tilde{\pi}\hat{\tilde{A}}\big[\Delta^{-1}\mathbf{T}\Delta + \Delta^{-1}t_0\pi_+^{(n)}\Delta + \gamma\mathbf{I}\big]\Delta^{-1} \\ &= \tilde{\pi}\hat{\tilde{A}}\Big[\tilde{\mathbf{T}} + \tilde{t}_0\tilde{\pi}_+^{(n)}\Big]\Delta^{-1},\end{aligned} \tag{3.9}$$

which shows that the $\tilde{\pi}_+^{(n)}$ are again given by the recursion (3.1), only with A replaced by $\tilde{A}$, $\mathbf{T}$ by $\tilde{\mathbf{T}}$ and so on. Thus, $\tilde{\pi}_+^{(n)} \to \tilde{\pi}_+$ whenever $\tilde{\pi}_+^{(0)}e \le 1$, where $\tilde{\pi}_+$ is the initial vector in the phase-type representation of $\tilde{G}_+$. This means that $\pi_+^{(n)} \to \tilde{\pi}_+\Delta^{-1}$ whenever $\pi_+^{(n)}h \le 1$, and it only remains to show that $\pi_+ = \tilde{\pi}_+\Delta^{-1}$. This can be done, for example, by applying Wald's fundamental identity for Markov additive processes to $\{(J_t, Y_t)\}$. We omit the details which follow [4], Section 4, closely. □

Applying Corollary 3.2 to the $\tilde{\pi}_+^{(n)}$ and performing some translation, we get:

COROLLARY 3.3. *Suppose* $\mu < 0$. *Then* $\hat{F}[s] = \hat{A}[-s]\hat{B}[s]$ *has a minimum at a unique* $\eta > 0$, *and* $\|\pi_+^{(n)} - \pi_+\| \le C\delta^{n+1}$, *where* $C = C_1C_2$ *with* C_1^{-1} *the minimal component of the vector* $(-\eta\mathbf{I} - \mathbf{T})^{-1}t_0$, $C_2 = \max_i h_i/\min_i h_i$ *and* $\delta = \hat{F}[\eta]$.

4. Starting from the Rouché roots.

We first provide a probabilistic interpretation of the Rouché roots:

THEOREM 4.1. *Let s be some complex number with* $\Re(s) > 0$, $-s \notin \mathrm{sp}(\mathbf{T})$. *Then* $-s$ *is an eigenvalue of* $\mathbf{Q}$ *if and only if* $1 = \hat{F}[s] = \hat{A}[-s]\hat{B}[s]$, *with* $\hat{B}[s], \hat{F}[s]$ *being interpreted in the sense of the analytical continuation of the m.g.f. In that case, the corresponding eigenvector may be taken as* $(-s\mathbf{I} - \mathbf{T})^{-1}t_0$.

PROOF. According to (2.3), $1 = \hat{F}[s]$ means

$$1 = \pi(-s\mathbf{I} - \mathbf{T})^{-1}t_0\hat{A}[-s]. \tag{4.1}$$

Suppose first $\mathbf{Q}h = -sh$. Then $e^{\mathbf{Q}x}h = e^{-sx}h$ and hence

$$-sh = \mathbf{Q}h = \big(\mathbf{T} + t_0\pi\hat{A}[\mathbf{Q}]\big)h = \mathbf{T}h + \big(\pi h\hat{A}[-s]\big)t_0. \tag{4.2}$$

Since $-s \notin \mathrm{sp}(\mathbf{T})$, this implies that $\pi h\hat{A}[-s] \ne 0$, and hence we may assume that h has been normalized such that $\pi h\hat{A}[-s] = 1$. Then (4.2) yields $h = (-s\mathbf{I} - \mathbf{T})^{-1}t_0$. Thus the normalization is equivalent to (4.1), and hence $\hat{F}[s] = 1$.

Suppose next $\hat{F}(s) = 1$. Since $\Re(s) > 0$ and G_- is concentrated on $(-\infty, 0)$, we have $|\hat{G}_-(s)| < 1$, and hence by the Wiener–Hopf factorization identity [Lemma 2.1(a)] $\hat{G}_+(s) = 1$, which according to Theorem 2.1(a) means that $\pi_+(-s\mathbf{I} - \mathbf{T})^{-1}t_0 = 1$. Hence with $h = (-s\mathbf{I} - \mathbf{T})^{-1}t_0$ we get

$$\mathbf{Q}h = (\mathbf{T} + t_0\pi_+)h = \mathbf{T}(-s\mathbf{I} - \mathbf{T})^{-1}t_0 + t_0 = -s(-s\mathbf{I} - \mathbf{T})^{-1}t_0 = -sh. \quad \square$$

Theorem 4.1 further substantiates that the problem of double roots mentioned in the introduction may be a delicate one: There is no a priori indication whether $\mathbf{Q}$ may or may not have multiple eigenvalues. It is clear from (2.3) that $\hat{F}[s]$ will typically not be defined when $s \in \mathrm{sp}(\mathbf{T})$, and thus it might be conjectured that by imposing some regularity conditions it may be possible to omit the condition $s \notin \mathrm{sp}(\mathbf{T})$ in Theorem 4.1. It is our feeling that the complete resolution of this problem has to do with uniqueness and minimality of phase-type representations (which is not settled at present). This is based upon the discussion of [2], Chapter IX.6, where the restrictions on the model are chosen precisely with the purpose of avoiding Rouché roots which are eigenvalues of $\mathbf{T}$. A further substantiation is given by the following example:

EXAMPLE 4.1. Suppose $\lambda > \beta > 0$ and

$$\mathbf{T} = \begin{pmatrix} -\lambda & \lambda - \beta \\ 0 & -\beta \end{pmatrix}, \qquad t_0 = \begin{pmatrix} \beta \\ \beta \end{pmatrix}$$

so that

$$\mathbf{Q} = \begin{pmatrix} -\lambda + \beta\pi_+(1) & \lambda - \beta + \beta\pi_+(1) \\ \beta\pi_+(2) & -\beta + \beta\pi_+(2) \end{pmatrix}.$$

This corresponds to a nonminimal representation of an exponential B with rate β, and we shall see that indeed $\mathbf{Q}$ has the eigenvalue $-\lambda \in \mathrm{sp}(\mathbf{T}) = \{-\lambda, -\beta\}$. If $\mu < 0$, then $0 < \theta = \|G_+\| = \pi_+(1) + \pi_+(2) < 1$, and it is a standard fact ([2], page 203) that θ solves $\hat{F}[\beta(\theta - 1)] = 1$. Obviously, $\beta(\theta - 1) \notin \mathrm{sp}(\mathbf{T})$; thus $\rho_1 = \beta(\theta - 1) \in \mathrm{sp}(\mathbf{Q})$ and the other eigenvalue of $\mathbf{Q}$ is $\rho_2 = \mathrm{tr}(\mathbf{Q}) - \rho_1 = -\lambda$. If $\mu > 0$, then $\theta = \|G_+\| = \pi_+(1) + \pi_+(2) = 1$, $\rho_1 = 0 \in \mathrm{sp}(\mathbf{Q})$ corresponding to the eigenvector e, and the other eigenvalue is $\rho_2 = \mathrm{tr}(\mathbf{Q}) - \rho_1 = -\lambda$.

COROLLARY 4.1. *Suppose $\mu < 0$, that the equation $\hat{F}(s) = 1$ has d distinct roots $\rho_1, \dots, \rho_d$ in the domain $\Re(s) > 0$ and define $h_i = (-\rho_i\mathbf{I} - \mathbf{T})^{-1}t_0$, $\mathbf{Q} = \mathbf{C}\mathbf{D}^{-1}$, where $\mathbf{C}$ is the matrix with columns $h_1, \dots, h_d$ and $\mathbf{D}$ is the matrix with columns $-\rho_1 h_1, \dots, -\rho_d h_d$. Then G_+ is phase-type with representation $(\pi_+, \mathbf{T})$ with $\pi_+ = \pi(\mathbf{Q} - \mathbf{T})/\pi t_0$. Further, letting ν_i be the left eigenvector of $\mathbf{Q}$ corresponding to $-\rho_i$ and normalized by $\nu_i h_i = 1$, $\mathbf{Q}$ has diagonal form*

$$\mathbf{Q} = -\sum_{i=1}^{d} h_i \otimes \nu_i = -\sum_{i=1}^{d} h_i \nu_i \tag{4.3}$$

and G_- is given by the density

$$g_-(x) = \sum_{i=1}^{d} c_i e^{-\rho_i x} \int_{-x}^{\infty} e^{-\rho_i u} A(du), \qquad x < 0, \text{ where } c_i = \pi_+ h_i \nu_i t_0. \tag{4.4}$$

PROOF. Appealing to Theorem 4.1, the matrix $\mathbf{Q}$ has the d distinct eigenvalues $-\rho_1, \dots, -\rho_d$ with corresponding eigenvectors $h_1, \dots, h_d$. This immediately implies that $\mathbf{Q}$ has the form $\mathbf{C}\mathbf{D}^{-1}$, and thus the result on G_+ is immediate from Theorem 2.1(a). For G_-, combining (4.3) and (2.11) we get

$$g_-(x) = \int_{-x}^{\infty} \pi_+ e^{\mathbf{Q}(x+u)} A(du) = \int_{-x}^{\infty} \pi_+ \sum_{i=1}^{d} e^{-\rho_i(x+u)} h_i \nu_i t_0 A(du),$$

which is the same as (4.4). □

In a similar way we get:

COROLLARY 4.2. *Suppose $\mu \geq 0$, that the equation $\hat{F}(s) = 1$ has $d-1$ distinct roots $\rho_2, \dots, \rho_d$ in the domain $\Re(s) > 0$ and define $\rho_1 = 0$, $h_1 = e$, $h_i = (-\rho_i \mathbf{I} - \mathbf{T})^{-1} t_0$, $i \geq 2$, and $\mathbf{Q} = \mathbf{C}\mathbf{D}^{-1}$, where $\mathbf{C}$ is the matrix with columns $h_1, \dots, h_d$ and $\mathbf{D}$ is the matrix with columns $-\rho_1 h_1, \dots, -\rho_d h_d$. Then all results of Corollary* 4.1 *on G_-, G_+ hold true.*

5. Concluding discussion. In much of the literature (e.g., [16] or [11]), the basic setup is not a difference structure $X = U - T$ with the relevant conditions imposed on, say, U, but rather one assumes a specific form of one tail of X. That is, one writes

$$F = pF_1 + qF_2, \tag{5.1}$$

where $p + q = 1$, F_1, F_2 are concentrated on $(0, \infty)$ [resp. $(-\infty, 0)$] and F_1 is assumed to belong to some given class. However, from the point of view of the present paper this setup can easily be reduced to difference structure. We state and prove some of the main results in that setting and omit the translation of more specialized topics like convergence rates.

COROLLARY 5.1. *Consider a random walk with increment distribution of the form* (5.1) *and F_1 phase-type with representation $(\pi, \mathbf{T})$. Then*:

(a) *The ascending ladder height distribution G_+ is phase-type with representation $(\pi_+, \mathbf{T})$ where π_+ is the solution of*

$$\pi_+ = \pi p \left(\mathbf{I} - q\hat{F}_2[-\mathbf{T} - t_0 \pi_+]\right)^{-1}. \tag{5.2}$$

(b) *If $\mu < 0$, then the maximum M has distribution given by an atom of size $1 - \pi_+ e$ at zero and a (defective) phase-type distribution with representation $(\pi_+, \mathbf{Q})$ on $(0, \infty)$ where $\mathbf{Q} = \mathbf{T} + t_0 \pi_+$.*

(c) *The descending ladder height distribution G_- is given by*

$$G_-(x) = qF_2(x) + \int_{-x}^{\infty} \pi_+ e^{\mathbf{Q}u} t_0 q F_2(x-u)\, du, \qquad x < 0. \tag{5.3}$$

(d) *The equation* (5.2) *can be solved by iteration, starting from any $\pi_+^{(0)}$ satisfying $\pi_+^{(0)} e \leq 1$ when $\mu \geq 0$ and $\pi_+^{(0)} h \leq 1$ when $\mu < 0$. Here $h = (-\gamma \mathbf{I} - \mathbf{T})^{-1} t_0$ with γ the unique solution greater than* 0 *of the equation $\hat{F}[s] = 1$. In particular, the solution of* (5.2) *is unique in the class of nonnegative vectors satisfying the given constraints.*

PROOF. Consider a random walk with increments distributed as $U - T$ where U is the first positive increment of the given random walk and $-T$ is the sum of all preceding negative increments. Obviously, U, T are independent, U follows the given the phase-type distribution F_1 and T has distribution A (say) given by $\hat{A}[s] = \Sigma_0^\infty pq^n \hat{F}_2[-s]^n$. Further, this random walk has the same ladder height distribution as the given one, and thus parts (a) and (b) are immediate from Theorem 2.1 by noting that $\hat{A}[\mathbf{Q}] = (\mathbf{I} - q\hat{F}_2[-\mathbf{Q}])^{-1}$. Part (c) follows by inserting (5.1) in Lemma 2.1(b), and part (d) is just a translation of Theorem 2.2. □

Obviously, the setup is sometimes discrete (i.e., the random walk is concentrated on a lattice) and the typical assumption of the literature is then that one tail has a finite support. With some notational changes, the methods of the present paper can be applied to the more general case where one tail is instead discrete phase-type (the lifetime of a Markov chain in discrete time), but we shall not give the details here.

From a computational point of view, the main problem in the iterative solution of $\mathbf{Q} = \psi(\mathbf{Q})$ is obviously to compute $\hat{A}[\mathbf{Q}]$. The standard series expansion of the matrix-exponential function leads to the formula

$$(5.4) \qquad \hat{A}[\mathbf{Q}] = \sum_{n=0}^{\infty} \frac{\mu_A^{(n)}}{n!} \mathbf{Q}^n,$$

where $\mu_A^{(n)}$ is the nth moment of A. This formula may be convenient in some cases, for example when A is discrete. It requires all moments of A to be finite and, to be computationally useful, that the $\mu_A^{(n)}$ do not increase too rapidly so that (5.4) can be truncated to a reasonable number of terms.

The currently most widely adopted method for computing matrix exponentials is, however, uniformization. Implemented in the present context, this means that we choose an η such that η is an upper bound on the absolute value of the entries of $\mathbf{Q}$. Then

$$(5.5) \quad \hat{A}[\mathbf{Q}] = \int_0^\infty \sum_{n=0}^{\infty} \left(\mathbf{I} + \frac{\mathbf{Q}}{\eta}\right)^n e^{-\eta x} \frac{(\eta x)^n}{n!} B(dx) = \sum_{n=0}^{\infty} c_n \left(\mathbf{I} + \frac{\mathbf{Q}}{\eta}\right)^n,$$

where

$$c_n = \frac{\eta^n}{n!} \mathbb{E} U^n e^{-\eta U} = \frac{\eta^n}{n!} \hat{A}^{(n)}[-\eta].$$

The two series in (5.5) converge without conditions and are convenient whenever simple expressions for the $\hat{A}^{(n)}[-\eta]$ are available (to this end, it may be useful to note that the $\hat{A}^{(n)}[-\eta]$ can be expressed in terms of moments in the exponential family generated by B). For the iteration scheme in Theorems 4.1 and 4.2 one can use $\eta = \max_{1 \le i \le d}(-t_{ii})$ for all $\mathbf{Q}^{(n)}$.

If A itself is of phase-type or, more generally, has a rational m.g.f. $\hat{A}[s] = q(s)/r(s)$ with q, r polynomials, one further alternative is available which seems largely unnoticed. This consists in computing $r(\mathbf{Q})^{-1}$ which exists whenever $\mathbf{Q}$ is a subintensity like all $\mathbf{Q}^{(n)}$ [this is so because the set $\mathrm{sp}(r(\mathbf{Q})) = r(\mathrm{sp}(\mathbf{Q}))$ cannot contain zero since all eigenvalues of $\mathbf{Q}$ have a nonpositive real part], and noting that $\hat{A}[\mathbf{Q}] = q(\mathbf{Q})r(\mathbf{Q})^{-1}$.

For some further aspects of the computation of matrix-valued m.g.f.'s, see [26] and also [17] in connection with (5.5) (a standard general reference on matrix exponentials is [18]). Even if the present author finds nonlinear matrix iteration more appealing in most cases, it should be noted that we do not insist that the method is universally superior. For example, the Rouché root algorithms provided by Corollaries 4.1 and 4.2 do not appear unappealing when there is a priori knowledge that the roots are distinct and real (like for $d = 2$ or B hyperexponential, [2], pages 219 and 221–222) and also when $\mu = 0$ or μ is close to 0, and where, as mentioned earlier, the matrix iteration may converge exceedingly slowly (examples where $\mu = 0$ occur, for example, in the computation of corrected diffusion approximations [28]). Note, however, that when two or more ρ_i are close, then computations based on diagonal forms like (4.3) and (4.4) tend to be numerically unstable.

REFERENCES

[1] Asmussen, S. (1984). Approximations for the probability of ruin within finite time. *Scand. Actvar. J.* **67** 31–57 [Correction: (1985) **68** 64.]

[2] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, New York.

[3] Asmussen, S. (1989a). Aspects of matrix Wiener–Hopf factorisation in applied probability. *Math. Sci.* **14** 101–116.

[4] Asmussen, S. (1989b). Risk theory in a Markovian environment. *Scand. Actvar. J.* **72** 69–100.

[5] Asmussen, S. (1989c). Exponential families generated by phase-type distributions and other Markov lifetimes. *Scand. J. Statist.* **16** 319–334.

[6] Asmussen, S. (1991). Ladder heights and the Markov-modulated M/G/1 queue. *Stochastic Process Appl.* **37** 313–326.

[7] Bahadur, R. R. and Ranga Rao, R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.

[8] Borovkov, A. A. (1962). New limit theorems in boundary problems for sums of independent random variables. *Siberian Math. J.* **3** 645–694 (in Russian).

[9] Borovkov, A. A. (1970). Factorization identities and properties of the distribution of the supremum of sequential sums. *Theory Probab. Appl.* **15** 359–402.

[10] Borovkov, A. A. (1976). *Stochastic Processes in Queueing Theory*. Springer, New York.

[11] Feller, W. (1984). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

[12] Gut, A. (1974). On the moments and limit distributions of some first passage times. *Ann. Probab.* **2** 277–308.

[13] Gut, A. (1988). *Stopped Random Walks. Limit Theorems and Applications*. Springer, New York.

[14] Iglehart, D. L. (1974). Random walks with negative drift conditioned to stay positive. *J. Appl. Probab.* **11** 742–751.

[15] Keilson, J. (1979). *Markov Chains-Rarity and Exponentiality*. Springer, New York.

[16] Kemperman, J. H. B. (1961). *The Passage Problem for a Markov Chain*. Univ. Chicago Press.

[17] Lucantoni, D. M. and Ramaswami, V. (1985). Efficient algorithms for solving the nonlinear matrix equation arising in phase-type queues. *Stochastic Models* **1** 29–51.

[18] Moler, C. and Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.* **20** 801–836.

[19] NEUTS, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins Univ. Press.

[20] NEUTS, M. F. (1989). *Structured Stochastic Matrices of the M/G/1 Type and Their Applications.* Dekker, New York.

[21] PRABHU, N. U. (1980). *Stochastic Storage Processes. Queues, Insurance Risk and Dams.* Springer, New York.

[22] RAMASWAMI, V. (1988). Non-linear matrix equations in applied probability: Solution techniques and open problems. *SIAM Rev.* **30** 256–263.

[23] RAMASWAMI, V. (1990). From the matrix-geometric to the matrix-exponential. *Queueing Systems Theory Appl.* **6** 229–260.

[24] RAMASWAMI, V. and LUCANTONI, D. (1985). Stationary waiting time distributions in queues with phase-type service and in quasi-birth-and-death processes. *Stochastic Models* **1** 125–136.

[25] RAMASWAMI, V. and LUCANTONI, D. (1988). Moments of the stationary waiting time distribution in the GI/PH/1 queue. *J. Appl. Probab.* **25** 636–641.

[26] SENGUPTA, B. (1989). Markov processes whose steady-state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Adv. in Appl. Probab.* **21** 159–180.

[27] SENGUPTA, B. (1990). The semi-Markov queue. Theory and Applications. *Stochastic Models* **6** 383–413.

[28] SIEGMUND, D. (1985). *Sequential Analysis.* Springer, New York.

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

[30] TAKACS, L. (1962). *Introduction to the Theory of Queues.* Oxford University Press.

[31] WOODROOFE, M. (1979). Repeated likelihood ratio tests. *Biometrika* **66** 453–463.

INSTITUTE OF ELECTRONIC SYSTEMS
AALBORG UNIVERSITY
FR. BAJERSV. 7
DK-9220 AALBORG Ø
DENMARK