The Annals of Applied Probability
1991, Vol. 1, No. 2, 219–227

PHASE-TYPE DISTRIBUTIONS AND MAJORIZATION[1]

BY COLM ART O'CINNEIDE

Louisiana State University and the University of Arkansas

Aldous and Shepp recently proved that the Erlang distribution of a given order is the least variable phase-type distribution of that order, in the sense of minimizing the coefficient of variation. Here we prove that it is also least variable in the sense of majorization. We give an example showing that the result does not extend in the obvious way to general distributions with rational transforms and this suggests that the inequality hinges on the Markov property.

1. Introduction. A phase-type (PH-) distribution is the distribution of a hitting time in a finite-state, time-homogeneous Markov chain. Except for a few remarks, we discuss only the continuous-time case. This family, which we denote by PH, was introduced by Neuts [9] as a tool for unifying a variety of stochastic models and for constructing new models that yield to algorithmic analysis. It represents the natural family to which Erlang's *method of stages* [4] extends. The basic idea is that, if a distribution for a time interval is needed in setting up a model and we choose a PH-distribution, then Markovian methods may be applicable. Various closure and approximation properties make this approach practicable. PH-distributions have rational Laplace–Stieltjes transforms and, at least formally, it seems that much that holds for the PH family should extend to all distributions with rational transforms; however, monotonicity, nonnegativity and properties of existence and uniqueness of solutions, on which the algorithms for PH-distributions are based, appear not to extend to distributions with rational transforms in general. The methodology related to PH-distributions has grown up around two basic structures: the GI/M/1 paradigm, explored in Neuts [10] and the M/G/1 paradigm, explored in Neuts [11].

The standard parametrization of PH-distributions is as follows. Let $Z = (Z_t,\ t \geq 0)$, be a continuous-time, time-homogenous Markov chain without instantaneous states on the state space $\{1, 2, \dots, n, n+1\}$ for which $\{1, 2, \dots, n\}$ is transient and $n+1$ is absorbing. Let S denote the generator of Z restricted to the transient states, so that S is of order n. The matrices S that arise in this way are called *PH-generators* and are characterized by the conditions that they have nonnegative off-diagonal entries, negative diagonal entries and are nonsingular. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ denote the initial distribution of Z on the transient states and write α_{n+1} for the probability that Z

Received January 1990; revised June 1990.

[1] Research supported in part by NSF Grant DMS-89-07630.

AMS 1980 *subject classifications.* Primary 60J27; secondary 60G42.

Key words and phrases. Hitting times, majorization, Markov chains, martingales, phase-type distributions.

is initialized in $n + 1$. Let T denote the time at which Z is absorbed in $n + 1$. The distribution of T is said to be phase type with *representation* (α, S). The integer n is called the *order of the representation*. The *order of a PH-distribution* is the smallest integer n for which a representation of order n is possible. The *degree* of a distribution with rational (Laplace–Stieltjes) transform is defined as the degree of the denominator of the transform when expressed as an irreducible ratio of polynomials.

Recently, the author has proved the following characterization of *PH*-distributions [13].

THEOREM 1. *A distribution on* $[0, \infty)$ *is phase type if and only if it is either the point mass at zero or*

(a) *it has a continuous positive density on* $(0, \infty)$ *and*

(b) *it has rational Laplace–Stieltjes transform with a unique pole of maximal real part.*

A discrete-time version of this is also proved, but that case is essentially due to Soittola [14] and Katayama, Okamoto and Enomoto [7]. A basic theme of ongoing research is the question: Given a *PH*-distribution, what can be deduced about its order or about its possible representations? The order of a *PH*-distribution is always at least as great as the degree and an example showing that these are not always equal is given in [13]. Another example arises in Section 3. A more specific goal is to establish that a *PH*-distribution that almost fails to satisfy condition (a) or (b) of Theorem 1 is of high order. To this end, it is established in O'Cinneide [12], using results of Dmitriev and Dynkin [3], that a *PH*-distribution whose transform has maximal pole at $-\lambda$, but also has a pole at $-\kappa \pm i\theta$, has order $n \geq \pi\theta/(\kappa - \lambda)$; as κ approaches λ, (b) almost fails and the order increases without bound. It remains to make precise the claim that a phase-type density $f(t)$ which almost fails to satisfy (a), in that it almost becomes zero at a positive argument, has large order. Theorem 1 gives a partial answer to the question of Dharmadhikari [2] and Heller [6]: When may a process be represented as a function of a finite-state Markov chain? The author continues to study this question.

Recall that the *Erlang* (n, λ) distribution is the distribution of $E_1 + E_2 + \cdots + E_n$, where the E_i's are independent and identically distributed (i.i.d.) exponential random variables of rate λ. This is a *PH*-distribution of order n for $\lambda < \infty$; if $\lambda = \infty$, we interpret the Erlang as the point mass at zero, denoted by δ_0. The quantity $1/\lambda$ plays the role of a scale parameter. Aldous and Shepp [1] have proved the following theorem.

THEOREM 2. *The coefficient of variation of an order n PH-distribution is at least* $1/\sqrt{n}$ *and the order n Erlang distribution is the only one to attain the bound.*

As these authors point out, a basic question is how well a given distribution can be approximated by a *PH*-distribution of a specified order. This theorem tells us that the Erlang distribution is, in a sense, the best approximation to a point mass at a positive constant. The method of proof was to study the quadratic variation of the martingale formed by conditioning the absorption time T on the natural filtration associated with the chain Z. Orthogonality of increments allows the coefficient of variation to be analyzed. We prove the following strengthening of Theorem 2.

THEOREM 3. *A PH-distribution with an order n representation majorizes the order n Erlang distribution of the same mean.*

Majorization, introduced by Hardy, Littlewood and Polya [5], is a stochastic ordering for comparing variability. A more recent treatment may be found in Marshall and Olkin [8]. Section 2 contains a discussion of majorization and some elementary results. It also contains the *decoupling* idea (Theorem 5) on which the proof of Theorem 3 is based. That proof is given in Section 3, with an example showing that Aldous and Shepp's result hinges on the Markov property and on the *order* of a *PH*-distribution, rather that the more elementary quantity, its *degree*.

We remark that the discrete-time analogues of Theorems 2 and 3, in which a point mass at a fixed integer m is to be approximated by a discrete *PH*-distribution of order $n \leq m$, also holds true, where the Erlang is replaced by its discrete analogue, being a sum of i.i.d. geometric random variables on the positive integers.

2. Majorization. Let μ and ν be two probability measures on R^n with finite means. We say that ν *majorizes* μ if $\int f d\nu \geq \int f d\mu$ for all convex functions f; this is written $\mu \prec \nu$. It implies, in particular, that the means of μ and ν are equal. The following is a well-known description of majorization, due to Strassen [15].

THEOREM 4. *Let μ and ν be probability measures on R^n having finite means. Then ν majorizes μ if and only if there exists a pair of R^n-valued random variables (X, Y) for which $E(Y|X) = X$ and the distributions of X and Y are, respectively, μ and ν.*

The condition of the theorem is that the two measures are the marginals of a martingale and this and Jensen's inequality immediately imply majorization. The proof of the converse is more difficult. We write $\mathscr{L}(X)$ for the distribution of the random variable X. All random variables appearing later are real.

Here are some basic lemmas that will be needed. Some proofs are outlined, although they are elementary.

LEMMA 1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with finite mean and suppose that $a_1, a_2, \ldots, a_n$ are real constants and $\bar{a}$ is their mean. Then*

$$\mathscr{L}\left(\bar{a}\sum_{i=1}^{n} X_i\right) \prec \mathscr{L}\left(\sum_{i=1}^{n} a_i X_i\right).$$

PROOF. Let σ denote a random permutation of $1, 2, \ldots, n$, distributed uniformly on the set of all such permutations, which is independent of the X_i's. Then

$$E\left(\sum_{i=1}^{n} a_{\sigma(i)} X_i \middle| X_1, X_2, \ldots, X_n\right) = \bar{a} \sum_{i=1}^{n} X_i,$$

so that

$$E\left(\sum_{i=1}^{n} a_{\sigma(i)} X_i \middle| \bar{a} \sum_{i=1}^{n} X_i\right) = \bar{a} \sum_{i=1}^{n} X_i.$$

The lemma now follows from Theorem 4, noting that

$$\mathscr{L}\left(\sum_{i=1}^{n} a_i X_i\right) = \mathscr{L}\left(\sum_{i=1}^{n} a_{\sigma(i)} X_i\right)$$

by exchangeability of the X_i's. □

Lemma 2 says, in essence, that a scale mixture of a distribution is more variable than that distribution. Lemma 3 says that majorization behaves as expected under convolution and mixture.

LEMMA 2. *Let X and Y be independent with finite means and suppose that $E(Y) = a$. Then $\mathscr{L}(aX) \prec \mathscr{L}(XY)$.*

PROOF. Clearly, $E(XY|X) = aX$ and the result follows from Theorem 4. □

LEMMA 3. *Suppose we have probability measures satisfying $\mu_i \prec \nu_i$, $i = 1, 2$, and suppose that $0 \le p = 1 - q \le 1$. Then with $*$ denoting convolution we have*

$$\mu_1 * \mu_2 \prec \nu_1 * \nu_2 \quad \textit{and} \quad p\mu_1 + q\mu_2 \prec p\nu_1 + q\nu_2.$$

The next result is the key to the proof of Theorem 3.

THEOREM 5. *Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ and N_1, N_2 be three independent i.i.d. sequences of nonnegative random variables having finite means, the N_i's being integer valued. Then*

$$\mathscr{L}\left(\sum_{i=1}^{N_1} X_i + \sum_{i=1}^{N_2} Y_i\right) \prec \mathscr{L}\left(\sum_{i=1}^{N_1} (X_i + Y_i)\right). \tag{1}$$

REMARK. A superficial insight into this result is that decoupling the sum of the X_i's from the sum of the Y_i's allows variability-reducing cancellation.

PROOF. We construct a martingale (W_1, W_2) such that the left-hand side of (1) is the marginal distribution of W_1 and the right-hand side is the marginal distribution of W_2. Without loss of generality, assume that the underlying probability space supports a random variable U, uniformly distributed on $[0,1]$, which is independent of the random variables appearing in the statement of the theorem. We define W_1 by

$$W_1 = \sum_{i=1}^{N_1} X_i + \sum_{i=1}^{N_2} Y_i,$$

so that it clearly has the desired distribution. Before introducing W_2, we define an integer-valued random variable N (which we later show has the same distribution as N_1) as follows. First let

$$C = \frac{\sum_{L+1}^{M} X_i}{\sum_{L+1}^{M} X_i + Y_i},$$

where $L = \min(N_1, N_2)$ and $M = \max(N_1, N_2)$. If the denominator here is ever 0, we define C arbitrarily to be $\frac{1}{2}$. Now define N by:

$$\text{If } N_1 = N_2, \text{ then set } N = N_1;$$

$$\text{(2)} \qquad \text{If } N_1 > N_2, \text{ then } \begin{cases} \text{if } U \leq C, & \text{set } N = N_1; \\ \text{if } U > C, & \text{set } N = N_2; \end{cases}$$

$$\text{If } N_1 < N_2, \text{ then } \begin{cases} \text{if } U \leq 1 - C, & \text{set } N = N_2; \\ \text{if } U > 1 - C, & \text{set } N = N_1. \end{cases}$$

With N so defined, we define W_2 by

$$W_2 = \sum_{i=1}^{N} (X_i + Y_i).$$

We now prove:

(i) $E(W_2 | W_1) = W_1$;

(ii) N is independent of the X_i's and Y_i's and has the same distribution as N_1.

Part (i) establishes the martingale property, while part (ii) establishes that the right-hand side of (1) is the distribution of W_2. Theorem 5 is an immediate consequence of these facts and Theorem 4.

We write $\mathscr{F}$ for the σ-field generated by X_i, Y_i, $i = 1, 2, \ldots$. Let us prove (i). Equation (2) defines a regular conditional probability for N given $\mathscr{F}$, N_1 and N_2, from which we have:

For $n_1 = n_2 = n$:

$$E(W_2 | \mathscr{F}, N_1 = n_1, N_2 = n_2) = \sum_{i=1}^{n} X_i + Y_i.$$

For $n_1 > n_2$:

$$\begin{aligned} E(W_2 | \mathscr{F}, N_1 = n_1, N_2 = n_2) &= C \sum_{i=1}^{n_1} (X_i + Y_i) + (1 - C) \sum_{i=1}^{n_2} (X_i + Y_i) \\ &= \sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i. \end{aligned}$$

For $n_1 < n_2$:

$$E(W_2|\mathscr{F}, N_1 = n_1, N_2 = n_2) = (1 - C)\sum_{i=1}^{n_2} (X_i + Y_i) + C\sum_{i=1}^{n_1} (X_i + Y_i)$$
$$= \sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i.$$

Together, the three statements imply that

$$E(W_2|\mathscr{F}, N_1, N_2) = \sum_{i=1}^{N_1} X_i + \sum_{i=1}^{N_2} Y_i = W_1.$$

As this is a function of W_1 only, (i) follows.

To prove (ii), let π_i denote $P(N_1 = i)$. Then we have

$$\begin{aligned} P(N = n|\mathscr{F}) &= P(N = n|\mathscr{F}, N_1 = N_2 = n)\pi_n^2 \\ &\quad + \sum_{a=0}^{n-1} \big(P(N = n|\mathscr{F}, N_1 = n, N_2 = a) \\ &\qquad + P(N = n|\mathscr{F}, N_1 = a, N_2 = n)\big)\pi_n\pi_a \\ &\quad + \sum_{b=n+1}^{\infty} \big(P(N = n|\mathscr{F}, N_1 = b, N_2 = n) \\ &\qquad + P(N = n|\mathscr{F}, N_1 = n, N_2 = b)\big)\pi_b\pi_n \\ &= \pi_n^2 + \sum_{a=0}^{n-1} (C + 1 - C)\pi_n\pi_a \\ &\quad + \sum_{b=n+1}^{\infty} (1 - C + C)\pi_b\pi_n \qquad \text{[by (2)]} \\ &= \pi_n. \end{aligned}$$

This proves (ii) and completes the proof of the theorem. □

3. Proof of the main theorem and an example. We are now ready to prove the main theorem. The proof is by induction on representation order. A *PH*-distribution with a representation of order 1 is a mixture of the point mass at zero and an exponential distribution. By Lemma 2, this majorizes the exponential distribution of the same mean, which is an Erlang distribution of order 1. So the result holds for $n = 1$.

Suppose we have established the result for *PH*-distributions of order $n - 1$. Let μ denote a *PH*-distribution with an order n representation (α, S). Let μ_i denote the *PH*-distribution with the order n representation (e_i, S), e_i being the ith unit vector, $i = 1, 2, \dots, n$. Then $\mu = \sum_{i=1}^{n}\alpha_i\mu_i + \alpha_{n+1}\delta_0$. If we can show that each μ_i majorizes the order n Erlang distribution of the same mean, which we call ν_i, then Lemma 3 implies that μ majorizes $\sum_{i=1}^{n}\alpha_i\nu_i + \alpha_{n+1}\delta_0$. But the latter is a scale mixture of Erlangs of order n and so by

Lemma 2 it too majorizes an order n Erlang. From this it follows that μ majorizes the order n Erlang of the same mean, as required. By relabeling states, we see that it suffices to prove that μ_1 majorizes an order n Erlang distribution. This we do as follows.

Let Z be an absorbing Markov chain with generator S on its transient states $\{1, 2, \dots, n\}$ which is initialized in state 1 and let T denote its time to absorption. Thus T has distribution μ_1. Let T_1 denote the time Z spends in state 1 and $T_{>1}$ the time Z spends in states $\{2, 3, \dots, n\}$, before absorption. Of course $T_1 + T_{>1} = T$, but T_1 and $T_{>1}$ are not independent in general. Let N denote the number of visits Z makes to state 1, not counting the last visit. Let X_i denote the time Z spends in state 1 on its ith visit and let Y_i denote the time Z spends in states $2, 3, \dots, n$ between the ith and i + 1st visit to state 1 for $1 \le i \le N$. To clarify, if $N \ge 1$, then X_1 is the time Z first leaves state 1, whereas if $N = 0$, then X_1 is not defined. We write U for the length of the final visit to state 1 and V for the time from the last exit from state 1 until absorption. By the strong Markov property, given $N = k \ge 0$, $X_1, X_2, \dots, X_k$ are i.i.d., $Y_1, Y_2, \dots, Y_k$ are i.i.d. and the X_i's, Y_i's, U and V are all independent, with distributions not depending on k. Let us extend the sequence X_i to an infinite i.i.d. sequence, independent of the Y_i's, U, V and N, enlarging the underlying probability space as needed. We have

$$T = \sum_{i=1}^{N} (X_i + Y_i) + U + V.$$

N is independent of the X_i's and Y_i's and U and V are independent of all of these. Therefore, Theorem 5 may be applied to the summation on the right-hand side (using Lemma 3 in the process), to conclude that μ_1 is majorized by

$$\mathscr{L}\left(\sum_{i=1}^{N'} X_i + \sum_{i=1}^{N} Y_i + U + V \right),$$

where we have introduced a random variable N' which is independent of all variables introduced up to now and which has the same distribution as N. This distribution is the convolution

$$\mathscr{L}\left(\sum_{i=1}^{N} Y_i + V \right) * \mathscr{L}\left(\sum_{i=1}^{N'} X_i + U \right)$$

by independence. Now the expression

$$\sum_{i=1}^{N} Y_i + V$$

is precisely the time $T_{>1}$ that Z spends in states other than state 1. Similarly, the distribution of

$$\sum_{i=1}^{N'} X_i + U$$

is the same as that of T_1, the time spent by Z in state 1. We conclude that

$$\mu_1 \succ \mathscr{L}(T_1) * \mathscr{L}(T_{>1}). \tag{3}$$

It is elementary that the time Z spends in a given set of k transient states is phase type, with an order k representation: this is because when we excise the time spent by Z outside of the given k states, the result is again a Markov chain but on k transient states. Thus, $\mathscr{L}(T_1)$ is a *PH*-distribution with an order 1 representation, while $\mathscr{L}(T_{>1})$ is a *PH*-distribution with an order $n-1$ representation. It follows from (3) that μ_1 majorizes the convolution of a *PH*-distribution with an order 1 representation and one with an order $n-1$ representation. By Lemma 3 and the induction hypothesis, this majorizes the convolution of an exponential distribution and an Erlang of order $n-1$. But Lemma 1 implies that this in turn majorizes the order n Erlang of the same mean. In sum, μ_1 majorizes an order n Erlang. This establishes, as we saw before, that μ itself majorizes an Erlang of order n and the proof is complete by induction.

Is the Erlang the least variable distribution with rational transform of a given *degree*? We answer this question in the negative, as follows. Consider the probability distribution μ with density function proportional to

$$(x^2 - 2\varepsilon x + 2\varepsilon^2)e^{-x}, \qquad x \geq 0,$$

where ε is a small positive number. By computing the transform of μ, we find that it has degree 3. Since its density is positive for $x > 0$ and its transform has only one pole, being -1, μ is a *PH*-distribution by Theorem 1. So, if Aldous and Shepp's result extends as proposed, its coefficient of variation should be at least that of the order 3 Erlang distribution, which is $1/\sqrt{3}$. Elementary computations lead to the following expression for the squared coefficient of variation:

$$\frac{\text{Variance}}{\text{mean}^2} = \frac{(2 - 2\varepsilon + 2\varepsilon^2)(24 - 12\varepsilon + 4\varepsilon^2)}{(6 - 4\varepsilon + 2\varepsilon^2)^2} - 1 = \frac{1}{3} - \frac{2}{9}\varepsilon + o(\varepsilon).$$

Thus the conjecture fails for ε sufficiently small and positive. We remark that the order of μ must exceed its degree, by Aldous and Shepp's inequality.

Acknowledgment. This work was carried out in part at the Quantitative Business Analysis Department of Louisiana State University and the author is grateful for a most enjoyable visit there.

REFERENCES

[1] ALDOUS, D. and SHEPP, L. (1987). The least variable phase-type distribution is Erlang. *Commun. Statist. Stochastic Models* **3** 467–473.

[2] DHARMADHIKARI, S. W. (1963). Sufficient conditions for a stationary process to be a function of a finite markov chain. *Ann. Math. Statist.* **34** 1033–1041.

[3] DMITRIEV, N. and DYNKIN, E. B. (1945). On the characteristic numbers of a stochastic matrix. *C.R. (Dokl.) Akad. Sci. URSS (N.S.)* **49** 159–162.

[4] Erlang, A. K. (1917 / 1918). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Engineer's Journal* **10** 189–197.

[5] Hardy, G. H., Littlewood, J. E. and Polya, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press.

[6] Heller, A. (1965). On stochastic processes derived from Markov chains. *Ann. Math. Statist.* **36** 1286–1291.

[7] Katayama, T., Okamoto, M. and Enomoto, H. (1978). Characterization of the structure-generating functions of regular sets and the DOL growth functions. *Inform. and Control* **36** 85–101.

[8] Marshall, A. W. and Olkin, I. (1979). *Inequalities: The Theory of Majorization and its Applications*. Academic, New York.

[9] Neuts, M. F. (1975). Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*. 173–206. Dept. Mathematics, Univ. Louvain, Belgium.

[10] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins Univ. Press, Baltimore.

[11] Neuts, M. F. (1989). *Structured Stochastic Matrices of M/G/1 Type and their Applications*. Dekker, New York.

[12] O'Cinneide, C. A. (1989). Phase-type distributions and invariant polytopes. Systems and Industrial Engineering (Laboratory for Algorithmic Research), Univ. Arizona Working Paper 89-025.

[13] O'Cinneide, C. A. (1990). Characterization of phase-type distributions. *Comm. Statist. Stochastic Models* **6** 1–57.

[14] Soittola, M. (1976). Positive rational sequences. *Theoretical Comp. Sci.* **2** 317–322.

[15] Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–439.

Department of Mathematical Sciences
SCEN 301
University of Arkansas
Fayetteville, Arkansas 72701