Phase-type distributions: open problems and a few properties

Colm Art O'Cinneide*
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907-1287

July 1997

Abstract

A phase-type distribution is the distribution of a killing time in a finite-state Markov chain. This paper describes some conjectures concerning these distributions. Some partial proofs and other evidence are offered in support of the conjectures. Some of the conjectures are structural in nature, and concern the geometry of the set of phase-type distributions, while others are quantitative, and concern the properties of their densities.

1 Introduction

A phase-type distribution is simply the distribution of a killing time in a finite-state Markov chain, and, as such, is a fundamental object in probability. It is surprising how little is known about them beyond elementary closure properties [5, 34, 39] and the simple characterization theorem given as Theorem 0 below [43]. Consider a phase-type distribution associated with an n-state *continuous time* Markov chain. It is fairly easy to see that its density has properties along the following lines (to be made precise later): (a) it cannot be too concentrated about its mean; (b) it cannot approach zero too closely at a positive argument; and (c) it cannot grow too steeply. Of these three properties, only one, (a), has been quantified precisely, by Aldous and Shepp [1], who showed that the coefficient of variation of such a distribution is no smaller than $1/\sqrt{n}$. (See also [45].)

One purpose of this paper is to set forth some conjectures quantifying properties like (b) and (c) (Sections 4 and 6). Other issues discussed here are more structural in nature. These include the question of "extremal" phase-type distributions (Section 5) and the existence of "sparse representations" (Section 7). Hopefully, this collection of open problems will provide a focus for ongoing research on phase-type distributions.

For some of the conjectures presented here, the supporting evidence is based on analyzing some special case, and is somewhat weak. The conjectures might be best taken in the spirit of the question "What is the correct generalization of what we have proved for such-and-such a special case?"

In Sections 2 and 3 below I review the basics of phase-type distributions and their role in stochastic modeling. Sections 4–8 concern the various conjectures, and Section 9 contains a few concluding remarks. Although I have tried to make this clear throughout, it may help to forewarn the reader that some of the arguments in this paper (especially in Section 6) are deductions from conjectures, and should be read with this in mind.

2 Preliminaries on phase-type distributions

This paper largely concerns *continuous* phase-type distributions. A few comments on the discrete case are made in Section 8. Let S be a *subgenerator* of order n, by which I mean an invertible $n \times n$ matrix with nonnegative off-diagonal entries and nonpositive row sums. With e denoting a column vector of 1's, and with $S^\circ \equiv -Se$ denoting the *killing-rate vector* (or the *exit-rate vector*), the condition that S be a subgenerator may be expressed by saying that the matrix

$$\begin{pmatrix} S & S^\circ \\ 0 & 0 \end{pmatrix}$$

is the generator of a Markov chain on $n+1$ states for which absorption in the last state is inevitable. With β denoting a substochastic n-vector, the *phase-type distribution* with

representation (β, S) is the distribution of killing time of the chain with subgenerator S and initial distribution β. The representation is said to be *irreducible* if $S + S^\circ\beta$ is irreducible. This means in essence that there are no superfluous states (that is, all states have a positive probability of being visited if the initial distribution is β).

The representation (β, S), with S an $n \times n$ matrix, is said to be of *order* n. We write PH_n for the set of all phase-type distributions that have an order-n representation. It is easy to verify that $\mathrm{PH}_n \subset \mathrm{PH}_m$ for $m > n$. The *order of a phase-type distribution* is the minimal order of all of its representations. A *minimal representation* of a given phase-type distribution is a representation whose order minimal.

The phase-type distribution with representation (β, S) assigns mass $1 - \beta e$ to 0. As a caution, note that some authors assume $\beta e = 1$ always; in this case the phase-type distribution is called "continuous" here. The density f of (the absolutely continuous part of) this distribution and its Laplace-Stieltjes transform ϕ are given by

$$f(t) = \beta e^{tS} S^\circ \text{ for } t \geq 0 \text{ and } \phi(s) = \beta(sI - Q)^{-1} e \text{ for } \Re(s) > 0.$$

(The latter would be a "Laplace transform" if there were no mass at zero. In future I will simply use the word "transform.") Note that ϕ is a rational function, and that $\phi(s) = p(s)/q(s)$ where p and q are polynomials of degree no more than n. It is convenient to define the *degree* of a distribution on $[0, \infty)$ with rational transform to be the degree of the denominator when the transform is expressed as an irredcuible ratio of polynomials. Thus the degree of a phase-type distribution is never greater than its order, but may be smaller [43]. For more on phase-type distributions, see [39, 44, 46]. A key result is

Theorem 0 (characterization of phase-type distributions). *A probability distribution on* $[0, \infty)$ *which is not the point mass at zero is phase type if and only if (a) it has a rational transform with a unique pole of maximal real part, and (b) the continuous density of its absolutely continuous part is positive everywhere on* $(0, \infty)$.

See [33] or [43] for proofs and Section 8 below for some further discussion. The maximal pole identified in (a) is of course real, and its negative is called the *decay parameter* of the phase-type distribution. That a *non-trivial* phase-type distribution (that is, one that is not the point mass at 0) has properties (a) and (b) is elementary; the converse is the interesting part of the theorem. Under the conditions of the theorem, we are assured that there exists a representation (α, S) for the given distribution, but the theorem provides no information about the nature of such representations. A general goal is to deduce properties of representations from properties of the distribution. Of particular interest is to place bounds on the order of a phase-type distribution based on knowledge of its density. Most of the conjectures presented here arose out of this question. See [44] and [45] for some discussion of what is known along these lines.

A subfamily of phase-type distributions of special interest is the *triangular phase-type distributions*, which are defined as those with *triangular representations* (β, S), S being an upper-triangular subgenerator. We write TPH_n for the family of phase-type

distributions with triangular representations of order n. The *triangular order* of such a distribution is the minimal number of states needed for a triangular representation. (The order of a triangular phase-type distribution may be smaller than its triangular order [42].) Much is known about this family [11, 12, 42, 46], and some of our conjectures may be resolved for this special case.

3 On the applications of phase-type distributions

The phase-type distributions are a generalization of the exponential distribution. Many stochastic models become tractable when the distributions of key times are assumed to be exponential. This is because exponentiality brings about Markovian properties. This tractability persists—at a cost—when these distributions are assumed to be phase type. The cost is a larger state space for the model, and the consequent increased complexity of the numerical solution. The benefit, of course, is in enlarging the class of distributions for which the model is solvable.

That this is a natural family for stochastic models is evidenced by the fact that, almost as soon as telephone traffic became the subject of engineering analysis, the first examples of phase-type distributions arose. I am referring to Erlang's [16] 1917 introduction of the distributions now bearing his name into the study of stochastic models. Erlang's "method of stages," used in its natural generality, leads to phase-type distributions. Over two decades ago, Marcel Neuts undertook a broad program of research whose goal was to explore the algorithmic implications of such methods [39, 40]. This program drew the interest of many students and researchers, with the result that phase-type distributions have become a basic paradigm of *matrix-analytic methods*, a substantial literature on stochastic models that are amenable to numerical solution. The main product of this literature is a collection of stochastic models for which stable matrix algorithms are available. Many of these models are based on phase-type distributions and their point-process analog, the *Markovian arrival process* ("MAP") [32, 50]. As an example of a particularly efficient algorithm, there is Latouche and Ramaswami's fast algorithm [30] for the $PH/PH/1$ queue. Another is Lucantoni's algorithm [26] for the $MAP/G/1$ queue (exploiting [52]). More recently, transform inversion algorithms have proved to be very effective in combination with matrix-analytic methods [27].

(A parallel development, in which distributions with rational transforms and their matrix-exponential representations play a key role, may be traced back to [9, 10], and results in extensions of matrix-analytic methods [3, 4, 7, 31]. In this more general setting, the natural extensions of the conjectures of this paper are easily resolved.)

It is important to acknowledge the limitations of a methodology in order to make good use of it. A few words are in order to explain the limitations of phase-type distributions. The role of phase-type distributions is to give an exact Markovian analysis of a stochastic model. It is widely recognized that even in fairly simple models based on phase-type distributions an exact analysis may be practically impossible due to the size of the state space. For example, a 5 server queue with a 10-phase interarrival dis-

tribution, distinct 10-phase service-time distributions, and no waiting room demands a state space of a million elements. On the other hand, there are many interesting but more modest models for which the phase approach provides an exact analysis. These smaller models are often useful in getting an understanding of the behavior of a larger model. For example, in developing approximate "decomposition" analyses of queueing networks (along the lines of Whitt's QNA [56, 57]), matrix-analytic methods maybe used to test or refine approximations for individual stations; for a nice example of exploring an approximation in this way, see Ramaswami and Lucantoni [51].

In addition to the problem of unwieldy state spaces, another limitation of phase-type distributions is the fact that they are not well suited to approximating every distribution. Despite the fact that they are dense [2, 28], and so theoretically any distribution on the nonnegative reals may be approximated arbitrarily well by a phase-type distribution, the order of the approximating distribution may be disappointingly large. For example, suppose we wish to approximate a deterministic service time of one time unit by a phase-type distribution. With a phase-type distribution of order 100 (which may represent a substantial computational burden in a given model) the best approximation in the sense of mean squared error is the Erlang distribution of order 100 and mean 100/101 [1]. But the probability that this Erlang is within 10% of the target service time is only about 68% (by the central limit theorem). On the other hand, while the phase-type distributions are admittedly poorly suited to matching certain features, even low-order phase-type distributions do display a great variety of shapes. For example, there are *tri-modal* phase-type distributions of order only 5. This rich family may be used in place of exponential distributions in many models without destroying our ability to compute solutions.

Because of the role phase-type distributions play in matrix-analytic methods, the goal of understanding both their limitations and richness is an important one. This goal, which underlies the present paper, also seems to point to interesting mathematics.

4 A lower bound on a phase-type density

Condition (b) of Theorem 0 contains the elementary fact that, for f the density of a nontrivial phase-type distribution, we have $f(t) > 0$ for $t > 0$. This positivity condition, and a simple limit argument, may be used to establish that the infimum of $f(t)$ as f ranges over all densities of phase-type distributions with a given order, mean, and variance, but with t fixed, is actually positive. The goal of this section is to explore the implied lower bound on phase-type densities. Because of its important role here, we write $\mathcal{E}_{n,\lambda}$ for the Erlang(n, λ) density:

$$\mathcal{E}_{n,\lambda}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}, \quad \text{for } t \geq 0.$$

I propose

Conjecture 1: *Let f be the density of a phase-type distribution of order n with mean $\mu > 0$ and coefficient of variation c. Then with $\lambda \equiv n/\mu$*

$$f(t) \geq e^{-(nc^2-1)} \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} = e^{-(nc^2-1)} \mathcal{E}_{n,\lambda}(t), \quad \text{for } t \geq 0. \tag{1}$$

Note that (1) has a natural interpretation when $\mu = 0$ or ∞, in which the right side is taken to be zero, and its truth in these extreme cases is obvious. Inequality (1) says that the density f exceeds the Erlang density of the same order and mean, but scaled by the factor $\exp\{-(nc^2 - 1)\}$. If $c = 1/\sqrt{n}$, the scaling factor becomes 1, and, because both f and $\mathcal{E}_{n,\lambda}$ integrate to 1, the inequality forces $f = \mathcal{E}_{n,\lambda}$. Thus Conjecture 1 implies the result of Aldous and Shepp mentioned in the introduction, and so is in a sense a proposed strengthening of [1].

I have proved Conjecture 1 in a special case: Theorem 1 below assures us that (1) holds for triangular phase-type distributions of triangular order n. As evidence that (1) holds generally, and not just in the triangular case, I have also proved it for a non-triangular family, namely, the *feedback Erlang distributions*, to be defined in Section 5. The proof is tedious, ending with a case-by-case analysis for low orders, and is not presented here. I feel that the evidence in favor of (1) is fairly substantial. Section 7 of the paper [46] bears some relation to the issue at hand in that for a certain family of phase-type distributions it explores the relationship between a large order and a near-zero in the density.

Theorem 1. *Let f be the density of a triangular phase-type distribution of triangular order n with mean $\mu > 0$ and coefficient of variation c. Then (1) holds.*

We first prove (1) for f a mixture of two order-n Erlang distributions in Lemma 1 below. This easily generalizes to any mixture of order-n Erlang distributions. The proof is completed by showing that every order-n triangular phase-type distribution is in fact a mixture of order-n Erlang distributions.

Because the Erlangs of a fixed order form a "scale family," we may represent mixtures of them in the following simple way. Let U be a nonnegative random variable, whose distribution will be called the "mixing distribution," and let E denote an Erlang$(n, 1)$ random variable that is independent of U. Then the distribution of UE is a mixture of order-n Erlangs, and all mixtures of order-n Erlangs may be represented in this way. In the case of mixtures of *two* Erlangs, suppose T is a nonnegative random variable taking two values, which again is independent of E. To parametrize the distribution of T conveniently, suppose that

$$P(T = a(1 - q\epsilon)) = p, \text{ and } P(T = a(1 + p\epsilon)) = q, \tag{2}$$

where $a \geq 0, \epsilon > 0, 0 < p = 1 - q < 1$, and $q\epsilon \leq 1$. (If we allow equality in the strict

inequalities on p, q, or ϵ, T becomes degenerate at a.) Then

$$E(T) = a;\ \mathrm{Var}(T) = pqa^2\epsilon^2;\ c^2(T) \equiv \mathrm{Var}(T)/E(T)^2 = pq\epsilon^2.$$

($c^2(T)$ denotes the squared coefficient of variation of T.) Now we define $X \equiv TE$, so that the distribution of X is a general mixture of two Erlang distributions of order n. We find that

$$E(X) = na;\ \mathrm{Var}(X) = na^2(1 + pq\epsilon^2) + n^2a^2pq\epsilon^2;$$

$$c^2(X) = pq\epsilon^2\left(1 + \frac{1}{n}\right) + \frac{1}{n} \text{ so that } nc^2(X) - 1 = (n+1)pq\epsilon^2. \tag{3}$$

We use this parametrization of mixtures of two Erlang distributions in proving

Lemma 1. *Inequality (1) holds for f a mixture of two order-n Erlang densities.*

Proof: We take the mixing measure as specified in (2). Thus we may suppose that, for T as in (2) and E Erlang$(n, 1)$, $X \equiv TE$ has density f. We suppose that $q\epsilon < 1$; otherwise the result is easy. Then we must prove that, for $t \geq 0$,

$$p\frac{t^{n-1}}{(n-1)!a^n(1-q\epsilon)^n}e^{-t/(a(1-q\epsilon))} + q\frac{t^{n-1}}{(n-1)!a^n(1+p\epsilon)^n}e^{-t/(a(1+p\epsilon))}$$
$$\geq e^{-(n+1)pq\epsilon^2}\frac{t^{n-1}}{(n-1)!a^n}e^{-t/a}.$$

Note that the exponent in the first factor on the right is $nc^2(X) - 1$ by (3). Multiply each side by $(n-1)!a^nt^{-n+1}$, and then change the variable to $y = t/a$, to reduce the inequality to be proved to

$$p(1-q\epsilon)^{-n}e^{-y/(1-q\epsilon)} + q(1+p\epsilon)^{-n}e^{-y/(1+p\epsilon)} \geq e^{-(n+1)pq\epsilon^2}e^{-y},\ y \geq 0.$$

Multiply by e^y to reduce this to

$$p(1-q\epsilon)^{-n}e^{-yq\epsilon/(1-q\epsilon)} + q(1+p\epsilon)^{-n}e^{yp\epsilon/(1+p\epsilon)} \geq e^{-(n+1)pq\epsilon^2}. \tag{4}$$

The right side is now independent of y. We minimize the left side with respect to y, and show that the minimum exceeds the right side. The following positive parameters help make sense of the left side of (4).

$$r \equiv p(1-q\epsilon)^{-n};\ \alpha \equiv \frac{q\epsilon}{(1-q\epsilon)};\ s \equiv q(1+p\epsilon)^{-n};\ \beta \equiv \frac{p\epsilon}{(1+p\epsilon)}.$$

That left side may now be written as $re^{-\alpha y} + se^{\beta y}$. Its minimum with respect to y, $y > 0$, occurs at

$$y = \frac{1}{\alpha+\beta}\log\frac{\alpha r}{\beta s},$$

and the minimum value is

$$r\left(\frac{\alpha r}{\beta s}\right)^{-\alpha/(\alpha+\beta)} + s\left(\frac{\alpha r}{\beta s}\right)^{\beta/(\alpha+\beta)}.$$

Therefore the minimum of the left side of (4) above is

$$p(1-q\epsilon)^{-n}\left(\frac{1-q\epsilon}{1+p\epsilon}\right)^{(n+1)q(1+p\epsilon)} + \quad q(1+p\epsilon)^{-n}\left(\frac{1-q\epsilon}{1+p\epsilon}\right)^{-(n+1)p(1-q\epsilon)} =$$

$$p(1-q\epsilon)(1-q\epsilon)^{-(n+1)p(1-q\epsilon)}(1+p\epsilon)^{-(n+1)q(1+p\epsilon)} + q(1+p\epsilon)(1-q\epsilon)^{-(n+1)p(1-q\epsilon)}(1+p\epsilon)^{-(n+1)q(1+p\epsilon)}$$

$$= \left[(1-q\epsilon)^{-(1-q\epsilon)}\right]^{(n+1)p}\left[(1+p\epsilon)^{-(1+p\epsilon)}\right]^{(n+1)q}\ldots$$

Now we use the fact that $(1+x)^{-(1+x)} \geq \exp -\{x+x^2\}, x > -1$, an easy consequence of $\log(1+x) \leq x, x > 0$, to continue

$$\ldots \geq e^{-(n+1)p[-q\epsilon+q^2\epsilon^2]}e^{-(n+1)q[p\epsilon+p^2\epsilon^2]}$$

$$= e^{-(n+1)\epsilon^2(pq^2+qp^2)} = e^{-(n+1)pq\epsilon^2}.$$

This proves (4) and completes the proof of the lemma.

The elementary fact that *any distribution on* $[0,\infty)$ *with finite mean may be expressed as a mixture of two-point distributions on* $[0,\infty)$ *all having the same mean* is the basis of the following extension of Lemma 1.

Lemma 2. *Let f be the density of a mixture of Erlang distributions of order n. Then (1) holds.*

Proof. If the mixture has mean 0 or ∞, the result is obvious, and so we assume that its mean, a, satisfies $0 < a < \infty$. Let E be an Erlang $(n,1)$ random variable as before. Let U be a nonnegative random variable independent of E whose distribution is the mixing measure. Then $X \equiv UE$ has density f and $a \equiv E(U) = \mu/n$ where $\mu \equiv E(X)$. The distribution of U may be expressed as a mixture of two-point distributions all having mean a. Following the parametrization given in (2), we express these distributions in the form

$$p\delta_{a(1-q\epsilon)} + q\delta_{a(1+p\epsilon)},$$

δ_x denoting the unit mass as x, where $q \equiv 1-p$ and the parameters range over the set $\Omega = \{(p,\epsilon) \mid 0 \leq p \leq 1, \epsilon \geq 0, q\epsilon \leq 1\}$ as before. Let ν be the mixing measure on Ω. Then the measure

$$\int_\Omega \left(p\delta_{a(1-q\epsilon)} + q\delta_{a(1+p\epsilon)}\right)\, d\nu(p,\epsilon)$$

is the distribution of U. Applying (1) and (3) for the first inequality below and Jensen's inequality for the second we have

$$\begin{aligned} f(t) &= \int_\Omega \left(p\mathcal{E}_{n,1/(a(1-q\epsilon))}(t) + q\mathcal{E}_{n,1/(a(1+p\epsilon))}(t) \right) \ d\nu(p,\epsilon) \\ &\geq \int_\Omega e^{-(n+1)pq\epsilon^2} \mathcal{E}_{n,1/a}(t) \ \ d\nu(p,\epsilon) \\ &\geq \mathcal{E}_{n,1/a}(t) \exp\left\{ -(n+1) \int_\Omega pq\epsilon^2 \ d\nu(p,\epsilon) \right\} \\ &= \mathcal{E}_{n,1/a}(t) e^{-(nc^2-1)}. \end{aligned}$$

The final equality is from (3), using the fact that the two-point distributions all have the same mean $a = \mu/n$. This proves Lemma 2.

Here is the last step in proving Theorem 1. The approach we take is via Theorem 2 below, due to Cumani [11] and Dehon and Latouche [12], although the role of this interesting theorem is mainly to simplify the presentation.

Lemma 3. *A triangular phase-type distribution of triangular order n is a mixture of order-n Erlang distributions.*

Proof. By [11] (see also [12, 42]), any triangular phase-type distribution has a *bi-diagonal representation* of the same order. To explain, let us write $\Theta(\lambda_1, \lambda_2, \cdots, \lambda_n)$ for the *bi-diagonal subgenerator*

$$\begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda_{n-1} & \lambda_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_n \end{pmatrix}. \tag{5}$$

Then we have, following [42] for notation,

Theorem 2 (Cumani [11] and Dehon and Latouche [12]). *Let T be an order-n upper-triangular subgenerator. Then every phase-type distribution with representation (α, T) (for α substochastic) also has a bi-diagonal representation (β, Θ), where $\Theta \equiv \Theta(\lambda_1, \lambda_2, \cdots, \lambda_n)$ and $\lambda_1, \lambda_2, \cdots, \lambda_n$ are the diagonal entries of T in ascending order of magnitude.*

Expressed another way, this theorem says that a triangular phase-type distribution (α, T) may be represented as follows. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. exponential random variables with mean 1, and let $I_1, I_2, \ldots, I_n$ be disjoint indicator random variables, independent of the Y_i's, for which $P(I_i = 1) = \beta_i$. Then the random variable

$$X \equiv \sum_{i=1}^{n} I_i \left(\sum_{j=i}^{n} \frac{Y_j}{\lambda_j} \right)$$

has the distribution with representation (α, T). To relate these to the absorbing chain implied by the bi-diagonal representation (β, Θ), I_i is indicator of the event that the chain starts in state i and Y_i/λ_i is the exponential(λ_i)-distributed time spent by the chain in state i.

Let $E = Y_1 + Y_2 + \ldots + Y_n$. Then E has the Erlang$(n, 1)$ distribution and is independent of (the σ-field generated by)

$$I_1, I_2, \ldots, I_n, Y_1/E, Y_2/E, \ldots, Y_n/E.$$

In particular, E is independent of

$$U \equiv \frac{X}{E} = \sum_{i=1}^{n} I_i \sum_{j=1}^{n} \frac{Y_j}{\lambda_j E}.$$

(That the Y_i/E's are independent of E follows easily from Basu's theorem in statistics [25] applied to an expoential random sample to conclude that ratios of observations are collectively independent of their sum, because the sum is a complete sufficient statistic for the unknown mean whereas ratios of observations have a distribution that does not depend on the mean. This independence also follows readily from well-known properties of the Poisson process.) Thus the representation

$$X = UE$$

expresses X as a scale mixture of order-n Erlang distributions. This proves Lemma 3, and Theorem 1 follows in the manner described in the paragraph immediately following its statement above.

5 Extremal phase-type distributions

Consider the family PH_n of all phase-type distributions of a given order n. Among these there is a naturally-defined family of extremal ones. An element of PH_n is called *extremal of order n* if it is not a nontrivial mixture of other elements of PH_n. (We omit the minor technicalities required to make the mixtures in question explicit.) It is easily verified that the Erlang distributions of order n are extremal of order n, as is the point mass at zero; to exclude the latter we again use the qualifier "non-trivial." If we define *extremal triangular phase-type distributions* analogously, then Lemma 3 easily implies that *the nontrivial extremal triangular phase-type distributions of order n are the order-n Erlang distributions.*

For $n = 1$ or 2, it is easy to show that all order-n phase-type distributions are also order-n *triangular* phase-type distributions, and so the extremal distributions are the order-n Erlang distributions in these cases. Next I establish that the Erlang distributions are not the only extremal phase-type distributions, by outlining an argument showing that *not all continuous phase-type distributions of order n are mixtures of*

order-n Erlang distributions. Towards this end, I introduce the *feedback-Erlang*(n, λ, p) distribution, $0 \le p < 1, \lambda > 0$, defined as the phase-type distributions with representation (e_1, S) where e_1 is the unit vector $(1, 0, \ldots, 0)$ and

$$S = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda & \lambda \\ p\lambda & 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix}. \tag{6}$$

This distribution is a geometric$(1-p)$ mixture of Erlang(nk, λ) distributions for $k = 1, 2, \ldots$. Consider the feedback Erlang distribution with parameters $\lambda = 1$, $p = 1/2$ and n large. Its density consists of a series of superimposed Erlang(nk,1) "humps" ($k = 1, 2, \ldots$), which are "bell-shaped" (by the central limit theorem) with means nk and coefficients of variation $1/\sqrt{nk}$. This means that the second hump ($k = 2$) is much less spread out than the Erlang of order n (in the relative sense; that is, in the sense of coefficient of variation), and so it cannot be approximated well by an Erlang-n mixture since mixing only increases the spread. It is routine but a little tedious to formalize this argument, and so I omit the details. (Moreover, there is an easy analytic way to establish the same thing: it is elementary that the transform of a mixture of Erlang distributions has an analytic continuation in the complex plane less the non-positive reals, and so no phase-type distribution whose transform has non-real poles can be a mixture of Erlangs.)

One hint that the feedback Erlang distributions may in fact be extremal is that they are extremal with respect to the inequality of Theorem 3.1 of [44] (which is a relative of inequalities of [14, 15]). To explain, consider a phase-type distribution of order n with decay parameter λ, and which also has a complex pair of poles $-\lambda_2 \pm \imath\theta$. Then of course $\lambda < \lambda_2$, and the inequality of [44] that we want is

$$\frac{\theta}{\lambda - \lambda_2} \le \cot \frac{\pi}{n}.$$

This inequality says that all the poles of the transform of this phase-type distribution lie in a wedge with vertex at $-\lambda$ extending horizontally in the negative-x direction, whose bounding rays make an angle of π/n with the vertical. *In the order-n feedback Erlang case (with $p > 0$), and only in this case, there is a pole at a boundary point of the wedge away from the vertex.* We omit the simple proof. Thus we propose

Conjecture 2: *The extremal phase-type distributions of order n are the order-n feedback Erlang distributions.*

If this were true, Conjecture 1 would follow at once from my (unpublished) proof for the feedback Erlang case alluded to in Section 4. Indeed, my interest in extremal phase-type distributions first arose from my attempts to prove Conjecture 1.

6 Steepest increase conjecture

Ultimately, all phase-type densities decay exponentially. An increase in the density is a local, transient effect, indicating that the inevitable asymptotic

$$f(t) \sim e^{-\lambda t} \text{ as } t \to \infty$$

(informally speaking), where λ is the decay parameter of f, has not yet gained supremacy. Intuitively, an increasing density is somewhat against the nature of a phase-type distribution. This statement requires refinement since an Erlang(n, n) distribution may have an arbitrarily steep increasing slope for large n. Here is the refinement. Consider the "relative increase" in f for a given "relative increase" in the argument, or, in other words, the derivative of $\log f(t)$ with respect to $\log t$. *I claim that this is bounded by* $n-1$ *for all phase-type distributions of a given order* n *and all* $t > 0$. This is the main conjecture of the present section. It may be written

$$\frac{d \log f(t)}{d \log t} = t \frac{f'(t)}{f(t)} \leq n - 1 \text{ for } t > 0. \tag{7}$$

For order $n = 1$, f is a mixture of an exponential distribution and a mass at 0, and therefore its density is always decreasing and the inequality is obvious. For order $n = 2$, all phase-type distributions are mixtures of Erlang distributions of order 2 (as was mentioned in the preceding section), and the bound need be verified only for these distributions. It is informative then to study the Erlang(n, λ) density, for which (7) holds as we have

$$\frac{d \log f(t)}{d \log t} = n - 1 - \lambda t \leq n - 1.$$

This inequality becomes tight as $t \to 0$.

So the conjecture is in essence that the Erlang distributions exhibit the extremal growth in the sense under consideration near $t = 0$. Inequality (7) is equivalent to

$$\frac{f'(t)}{f(t)} \leq \frac{n-1}{t}, \quad t > 0, \tag{8}$$

and if we integrate this with respect to t (or (7) with respect to $\log t$) over the interval $[u, v], 0 < u < v$, we find it is in turn equivalent to

$$\frac{f(v)}{f(u)} \leq \left(\frac{v}{u}\right)^{n-1},$$

and this leads to the following simple formal statement of the conjecture.

Conjecture 3: *For any phase-type density* f *of order* n, $f(t)/t^{n-1}$ *is nonincreasing for* $t > 0$.

There is a connection here with the discussion of Section 4. The claim that a density f is a mixture of order-n Erlang distributions is equivalent to the claim that $f(t)/t^{n-1}, t > 0$, is a *completely monotonic* function [49, 58]. If this were true of phase-type distributions of order n (which in fact it is not, because of the counterexample in Section 4), then the present conjecture would follow readily, but complete monotonicity is a much stronger condition than what is proposed in Conjecture 3. However, we have shown that *triangular* phase-type distributions are mixtures of Erlang distributions of the correct order (Theorem 2), and, from this, Conjecture 3 follows readily for the triangular phase-type case. This evidence, based on the triangular case, is weak; however, I am fairly certain that even if Conjecture 3 proves to be false, there is some increasing function of n to replace $n-1$ in (7) that yields a valid statement.

A natural place to look for a counterexample to this conjecture is the feedback Erlang distributions, because of their extremal character. These are superpositions of Erlang distributions of ever higher orders, $n, 2n, 3n, \ldots$, which have ever-steeper ascents. But I have not succeeded in finding a counterexample here.

Again, I have only succeeded in proving Conjecture 3 partially: in the triangular case as outlined above, and also *in the case of small* t. Before considering the latter result, we first replace (8) by an inequality for matrix exponentials which is of independent interest. *Suppose that the conjecture is true; that is, suppose that (8) holds for all order-n phase-type densities* f. Denote the representation of f by (α, S). Then (8) becomes

$$\frac{\alpha e^{tS} S S^\circ}{\alpha e^{tS} S^\circ} \leq \frac{n-1}{t},$$

or

$$\alpha((n-1)I - tS)e^{tS}S^\circ \geq 0.$$

This must be true for all α substochastic, and so we conclude that the conjecture is equivalent to

$$((n-1)I - tS)e^{tS}S^\circ \geq 0 \tag{9}$$

for all order-n subgenerators S. (This inequality is meant entrywise.) Upon noting that the matrix tS is in fact a "generic" subgenerator, this is equivalent to the simpler

$$((n-1)I - S)e^{S}S^\circ \geq 0. \tag{10}$$

We have ridden the inequality of α (the entrance vector) and t. Next we rid it of S°, the killing rate vector. To do this, first assume that S is *irreducible*. Then the vector $\delta = -S^{-1}e_1$ is positive and so the diagonal matrix $\Delta \equiv \mathrm{diag}(\delta)$ is invertible. Set $S_1 \equiv \Delta^{-1}S\Delta$. This is again an irreducible subgenerator: the key is that $S_1^\circ \equiv -S_1 e = \delta_1^{-1}e_1$. Inequality (10) applied to S_1 thus gives

$$((n-1)I - S_1)e^{S_1}e_1 \geq 0.$$

Premultiplication by Δ produces $((n-1)I - S)e^S)e_1 \geq 0$. A parallel argument with e_1 replaced by the other unit vector $e_2, e_3, \ldots, e_n$ leads to

$$((n-1)I - S)e^{S} \geq 0 \tag{11}$$

for all order-n irreducible subgenerators S (again entrywise). Now any subgenerator may be expressed as a limit of a sequence of irreducible subgenerators and so by taking limits, we see that (11) is true for all order-n subgenerators (rather than merely irreducible subgenerators). Furthermore, as every *generator* is also a limit of irreducible subgenerators, (11) is true for all generators of order n. *This all assumes the truth of Conjecture 3.*

Conversely, suppose we know (11) for all order-n *generators* S. We shall see that this implies that it is true for all order-n *subgenerators* also. To see this, let S be an *irreducible* subgenerator with right Perron-Frobenius eigenvector v and Perron-Frobenius eigenvalue $-\lambda$. Now define $Q \equiv \Delta^{-1} S \Delta + \lambda I$, where $\Delta = \mathrm{diag}(v)$. Then $Qe = 0$ and Q is now a *generator* so that (11) holds for this Q by assumption. Upon pre- and post-multiplication by the appropriate diagonal matrices, (11) applied to this Q gives, after simplification,

$$((n-1-\lambda)I - S)e^S \geq 0$$

which is a *stronger* statement than (11) since $\lambda > 0$ and $e^S \geq 0$. The steepest-increase conjecture has been shown to be equivalent to

Alternative form of Conjecture 3: *For all order-n* generators Q*,*

$$Qe^Q \leq (n-1)e^Q. \tag{12}$$

Note that, for Q irreducible, $e^{tQ} \to e\pi$ as $t \to \infty$, where π is the stationary distribution of Q. Thus $Qe^{tQ} \to Qe\pi = 0$. The inequality (12) may thus be viewed as placing an upper bound of $(n-1)e^{tQ}/t$ on the quantities Qe^{tQ} which controls how far they may *exceed* their limit of 0 as t increases from 0 to ∞.

The bound (12) is reminiscent of some inequalities in [6], Section 6.3, that are based on evaluating a competely monotonic function at a matrix argument. However, the dependency on the order n in (7)–(11), which does not arise in the inequalities of [6], indicates that Conjecture 3 is of a different nature.

A bound on the derivative of e^{tQ} with respect to t tells us something about the sensitivity of exponentials of generators to small relative perturbations (see [36]). This leads to a whole range of questions related to extending what is known in the limit as $t \to \infty$ (in which case the question is the perturbation theory of steady-state distributions, about which a great deal is known [20, 47, 48, 54, 55]) to finite t where little is know.

We now turn to a special case of the conjecture for which a proof is available. It is the case in which we take $Q = P - I$ in (12), with P stochastic, which, after a little simplification, becomes (13) below.

Theorem 3. *For P an $n \times n$ stochastic matrix,*

$$Pe^P \leq ne^P. \tag{13}$$

By retracing the steps that lead from Conjecture 3 to (11), we see that Theorem 3 implies that $f(t)/t^{n-1}$ *is nonincreasing for all sufficiently small* t as we claimed earlier. In fact, supposing that f has representation (α, Q), then this theorem implies that $f(t)/t^{n-1}$ is nonincreasing for $0 < t \leq 1/||Q||$ where $||Q||$ is the *maximal rate* of Q, which is the maximum of the absolute values of the entries of Q. In what follows, we write $p_{ij}^{(m)}$ for the (i,j)-entry of the m-step transition matrix P^m. The proof of Theorem 3 is based on the

Proposition. *For P an $n \times n$ stochastic matrix and $k \geq n$, we have*

$$P^k \leq \sum_{c=0}^{n-1} \binom{k}{c} P^c.$$

The proof is developed through the following two lemmas. In these lemmas, a "state" is an element of the set $\mathcal{X} \equiv \{1, 2, \ldots, n\}$ and i and j denote arbitrary fixed states.

Lemma 4. *Let $k \geq n$ and consider a sequence of states $i = i_0, i_1, i_2, \ldots, i_k = j$ in $\mathcal{X}$. Then there is a subset $J = \{a_1, a_2, \ldots, a_c\}$, $a_1 < a_2 < \ldots < a_c$, of the index set $I \equiv \{1, 2, \ldots, k\}$ with cardinality $c < n$, such that*

$$i_0 = i_{a_1} - 1,\ i_{a_1} = i_{a_2} - 1,\ i_{a_2} = i_{a_3} - 1,\ \ldots,\ i_{a_{c-1}} = i_{a_c} - 1,\ i_{a_c} = j.$$

Proof: Since $k \geq n$ there is at least one repeated state among the $k+1$ states $i_\ell, 0 \leq \ell \leq k$. Suppose $i_a = i_b$ for some $0 \leq a < b \leq k$. Now remove the indices $a+1, a+2, \ldots, b$ from I to produce a smaller index set I_1. Note that the first and last states associated with indices in I_1 are still i and j. If $\#I_1 > n$, the process may be repeated, producing a decreasing sequence of index sets $I_2, I_3 \ldots$, until we reach one with fewer than n indices and this is a set J with the desired properties. Lemma 4 follows.

For a set $J = \{a_1, a_2, \ldots, a_c\}$, $a_0 \equiv 0 < a_1 < a_2 < \ldots < a_c$, of integer "times," we define ξ_J as the event

$$(X_{a_0} = X_{a_1} - 1, X_{a_1} = X_{a_2} - 1, \ldots, X_{a_{c-1}} = X_{a_c} - 1, X_{a_c} = j).$$

Then we have

Lemma 5.

$$P(\xi_J \mid X_0 = i) \leq p_{ij}^{(c)}.$$

Proof: The left side may be expressed as follows, where $a_0 = 0$, $i_0 = i$, $i_c = j$, and the i_ℓ's, $0 < \ell < c$, range over the whole state space in the summation.

$$\sum_{i_1,i_2,\ldots,i_{c-1}} \prod_{p=1}^{c} \Big[P(X_{a_p} - 1 = i_{p-1} \mid X_{a_{p-1}} = i_{p-1}) P(X_{a_p} = i_p \mid X_{a_p} - 1 = i_{p-1}) \Big]$$

$$\leq \sum_{i_1,i_2,\ldots,i_{c-1}} \prod_{p=1}^{c} P(X_{a_p} = i_p \mid X_{a_p} - 1 = i_{p-1})$$

$$= \sum_{i_1,i_2,\ldots,i_{c-1}} p_{ii_1} p_{i_1 i_2} \cdots p_{i_{c-1} j} = p_{ij}^{(c)},$$

recalling that $i_0 = i$ and $i_c = j$ on the next-to-last step, as required.

Proof of the Proposition: By Lemma 4,

$$(X_0 = i, X_k = j) \subset \bigcup_{\substack{J \subset I \\ \#J < n}} \xi_J,$$

and so

$$p_{ij}^{(k)} \leq \sum_{\substack{J \subset I \\ \#J < n}} P(\xi_J \mid X_0 = i) \leq \sum_{c=0}^{n-1} \sum_{\substack{J \subset I \\ \#J = c}} p_{ij}^{(c)} = \sum_{c=0}^{n-1} \binom{k}{c} p_{ij}^{(c)}.$$

The first inequality is due to the set inclusion just established and the second is from Lemma 5. The last equality is because $\binom{k}{c}$ is the number of subsets J of size c in I.

Proof of Theorem 3: Expanding both sides of (13) in powers of P and reorganizing terms we find that we are required to prove

$$\sum_{\ell=n+1}^{\infty} \frac{P^\ell}{\ell!}(\ell - n) \leq \sum_{\ell=0}^{n-1} \frac{P^\ell}{\ell!}(n - \ell).$$

Working on the left, using the proposition on each term in turn, we find

$$\sum_{\ell=n+1}^{\infty} \frac{P^\ell}{\ell!}(\ell - n) \leq \sum_{c=0}^{n-1} P^c \sum_{\ell=n+1}^{\infty} \binom{\ell}{c} \frac{\ell - n}{\ell!}$$

$$= \sum_{c=0}^{n-1} P^c \sum_{\ell=n+1}^{\infty} \frac{\ell - n}{c!(\ell - c)!}.$$

Comparing coefficients of P^c, it is sufficient to prove that

$$\sum_{\ell=n+1}^{\infty} \frac{\ell - n}{c!(\ell - c)!} \leq \frac{n - c}{c!} \text{ or } \sum_{\ell=n+1}^{\infty} \frac{\ell - n}{(\ell - c)!} \leq n - c.$$

The right side of the second form of the inequality here is decreasing in c while the left side is increasing in c. Thus it suffices to prove the inequality for $c = n - 1$, in which case the right side is 1 while the left side is

$$\sum_{\ell=n+1}^{\infty} \frac{\ell - n}{(\ell - n - 1)!} = \sum_{j=1}^{\infty} \frac{j}{(j+1)!} = \sum_{j=1}^{\infty} \left(\frac{1}{j!} - \frac{1}{(j+1)!} \right) = 1,$$

the last series "telescoping." This completes the proof.

7 The unicyclic conjecture

Theorem 2 states that triangular phase-type distributions of order n have bi-diagonal representations of order n. But not all phase-type distributions of order n are triangular (of the same or any other order), because the latter have transforms with only real poles whereas the former may have transforms with complex poles. The "holy grail" in the study of phase-type distributions is to prove an analog of Theorem 2 for general (non-triangular) phase-type distributions. One way of looking at the question is that we wish to find *sparse representations* of phase-type distributions. The reason that this is a reasonable goal is that order-n phase-type distributions are naturally parametrized by only $2n$ independent parameters through their transforms (expressed as ratios of polynomials of degree n), whereas their representations involve $n^2 + n$ independent parameters. Some numerical experience with fitting phase-type distributions to data suggests that sparse representations are at least quite common. This experimentation has not been carefully documented but has been observed by the author and others, in particular, Mary A. Johnson [29], Malcolm Faddy, and Sören Asmussen (private communications). The natural place to look for such a result is among subgenerators that are "small modifications" of bi-diagonal subgenerators. The simplest such modification would appear to be

$$\begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda_{n-1} & \lambda_{n-1} \\ \mu & 0 & 0 & \cdots & 0 & -\lambda_n \end{pmatrix}.$$

Note that feedback Erlang distributions have representations of this form. Phase-type distributions with such representations, or the representations themselves, will be described as "restricted unicyclic" in what follows. The word "unicyclic" refers to the fact that the states in the graph of the chain are connected in a single "cycle" $1 \to 2 \to \ldots \to n \to 1$ or "killed". This family has $2n + 1$ parameters, which is one more than the bi-diagonal representations (5), and in fact this is an overparametrization of the intrinsically $2n$-dimensional space of order-n phase-type distributions. Unfortunately, not all phase-type distributions have such a simple minimal representation. We identify

a counterexample among the *unicyclic representations*, defined by generators of the form

$$\Theta = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda_{n-1} & \lambda_{n-1} \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{n-1} & -\lambda_n \end{pmatrix}.$$

Here, the states are again connected "cyclically": $1 \to 2 \to \ldots \to n \to$ "any". The transform of the unicyclic phase-type distribution with representation (α, Θ) takes the form

$$\frac{P(s)}{L_n(s) - p_1 - p_2 L_1(s) - p_3 L_2(s) - \ldots - p_{n-1} L_{n-2}(s)}, \tag{14}$$

where P is a polynomial of degree no more than n, $p_i \equiv \mu_i/\lambda_n$, and

$$L_k(s) \equiv (1 + \ell_1 s)(1 + \ell_2 s) \ldots (1 + \ell_k s), \ k = 0, 1, \ldots, n,$$

ℓ_i being $1/\lambda_i$. For the strictly unicyclic case, the transform takes the special form

$$\frac{P(s)}{L_n(s) - p}, \tag{15}$$

where $p \equiv \mu/\lambda_n$. To see that (15) is truly more restrictive than (14) we consider an order 4 example with $p_1 = p_3 = 0$ but with $p_2 \neq 0$. The denominator in (14) then takes the form

$$d(s) \equiv (1 + \ell_1 s)[(1 + \ell_2 s)(1 + \ell_3 s)(1 + \ell_4 s) - p_2].$$

First take $\ell_2 = 1, \ell_3 = 3$ and $\ell_4 = 4$. Then it may be verified that the cubic factor of the denominator d has exactly one real root $\xi < 0$, and so upon taking $\ell_1 = 1/\xi$ we find that the d is always nonnegative and has a single real root of order 2. This cannot arise for the restricted case (15) because for the polynomial $L_4(s) - p = (1 + \ell_1 s)(1 + \ell_2 s)(1 + \ell_3 s)(1 + \ell_4 s) - p$ to have a single real root we would have to have $\ell_1 = \ell_2 = \ell_3 = \ell_4$ and $p = 0$, but then it would have a real root of degree 4 rather than the established degree of 2. This is the desired counterexample.

In the light of this counterexample, an optimistic but not-yet-eliminated conjecture on sparse representations appears to be

Conjecture 4 (the unicyclic conjecture): *Every phase-type distribution of order n has a unicyclic representation of order n.*

G. Latouche and V. Ramaswami proposed a conjecture along these lines to me in 1989 in Tucson. Kohno [24] considered this family for other reasons. Phase-type distributions of order $n = 1$ or 2 are TPH_n, and so the conjecture is true in these cases. I believe I have a proof of the conjecture for $n = 3$ also, but the details are tedious and

are not presented.

The unicyclic conjecture is somewhat poorly motivated. Its basis is that it proposes a fairly simple, sparse form which has not yet been excluded by counterexample as a candidate for representing all phase-type distributions of a given order. An obvious relaxation of the conjecture is to take the same non-zero entries as the unicyclic generator Θ but not to require that the first $n-1$ rows have zero sum, *but in fact this may be shown by Theorem 2 to be equivalent to the original conjecture.* At a minimum, it seems likely that every order-n phase-type distribution has a highly sparse order-n representation, and that this representation may be taken to be of "a few" specific forms. Conjecture 4 proposes to take what is known for low orders ($n \leq 3$) and extend it to higher orders; unfortunately, the mathematics literature is full of difficult counterexamples to conjectures of this type. A nice example of this is the conjecture that *an n-dimensional polytope such that every pair of vertices is connected by an edge must be a simplex.* It is true in 2 or 3 dimensions, but in 4 dimensions the *cyclic polytopes* [35] provides a counterexample. Another is the conjecture that *if the angle between every pair of vectors in a set of n n-dimensional vectors is less than 90^o, then the entire set may be rotated into the positive quadrant.* True up to dimension 4, but false in dimension 5 [6, 18].

We close with a related conjecture which is of special interest not only because of its simplicity but also because it may be tested numerically with ease. If the unicyclic conjecture were true then this conjecture would also be true:

Conjecture 5 (the inverse characteristic polynomial conjecture): *Let S be a subgenerator and let $\phi(s) = d(0)/d(s)$, where d is the characteristic polynomial $d(s) = |sI - S|$ of S. Then ϕ is the transform of a phase-type distribution of order n.*

To explain how this is a consequence of Conjecture 4, but stopping short of the attention-to-detail needed for a proof, suppose (α, S) is a phase-type distribution of order n and of *degree* n. (See Section 2 for *degree*. Care is needed here as it is possible that there is no appropriate α [8, 42]; but this can be side-stepped with a diagonal similarity transformation along the lines of the argument leading to the "Alternative form of Conjecture 3.") Thus its transform may be expressed as a ratio of polynomials $p(s)/q(s)$ that cannot be reduced, with q is of degree n. Then the denominator polynomial q may be taken to be the characteristic polynomial d of S. Now, assuming Conjecture 4 is true, this phase-type distribution has a representation (β, Θ) for a unicyclic subgenerator Θ of order n. But then it is clear that the characteristic polynomial of Θ is proportional to that of S, and therefore the phase-type distribution (e_1, Θ) has transform $d(0)/d(s)$, as required. To see this, note that the numerator polynomial P in (14) is constant for the representation (e_1, Θ). In fact, Conjecture 5 would follow from a lower-Hessenberg [17] relaxation of Conjecture 4.

Conjecture 5 is easy to study for low orders ($n \leq 4$) with commonly available symbolic mathematics software packages such as MAPLE. Starting with S, its charac-

teristic polynomial d is readily computed, and then the inverse transform of $1/d$ may be computed and graphed. If this graph crosses the x-axis, or touches it at a positive argument, then we have a counterexample. In my unsystematic numerical experiments along these lines, I did not find a counterexample.

8 Maier's proof of the characterization theorem and the maximal rate conjecture

In this paper I have focused on continuous-time phase-type distributions. In fact, the quantitative conjectures 1 and 3 do not appear to have natural discrete-time analogs. On the other hand, the structural conjectures 2, 4, and 5 translate naturally to discrete time.

I proved Theorem 0 while visiting the University of Arizona in 1989. Robert S. Maier of the department of mathematics there took an interest in this result because he recognized its connection with results in theoretical computer science, in particular, Soittola's theorem characterizing generating functions of regular languages [33, 34]. Soittola's result is in essence Theorem 0 in discrete time. (See also Katayama, Okamoto and Enomoto [21], who proved the result independently a little later than Soittola, and the Section 5 of [43].) The continuous case has a simplicity that is not shared completely by the discrete case because of the possibility of periodicity in the latter. Maier discovered another, shorter proof of Theorem 0, taking Soittola's theorem as its starting point and using the following key lemma (Lemma 5.1 of [33]) to translate this into continuous time. We state this lemma in a slightly modified form.

Maier's Lemma. *Let f be a probability density function satisfying conditions (a) and (b) of Theorem 0. Then, for a sufficiently large positive c, the function $g(t) = e^{ct}f(t)$ and all its derivatives are nonnegative for all $t > 0$.*

With this lemma on hand, it may be verified from the discrete case (Soittola's theorem) that, for ϕ the transform of a density satisfying conditions (a) and (b) of Theorem 0 and with c as in Maier's lemma, the function $G(z) \equiv \phi(c(1-z)/z)$ is the probability generating function of a discrete-time phase-type distribution. From this it follows readily that ϕ itself is the transform of a continuous phase-type distribution. This is because for any G that is the probability generating function of a discrete phase-type distribution, $\phi(s) \equiv G(c/(c+s))$ is the Laplace-Steiltjes transform of the continuous phase-type distribution that results from giving each state in the implied discrete-time chain an exponential(c) holding time. (Note that $z = c/(c+s)$ inverts to give $s = c(1-z)/z$.) So Maier's lemma allows one to move easily between continuous and discrete time and allows us to view Theorem 0 as a corollary to Soittola's theorem.

Knowing in advance that f is phase-type, it is easy to identify a suitable c for Maier's lemma: simply take c to be the maximal rate $||Q||$ of Q. Then $g(t) = e^{ct}f(t) =$

$\alpha e^{t(Q+cI)}Q^\circ$, and because $Q+cI$ is now a nonnegative matrix, the monotonicity property claimed in Maier's lemma follows easily. It is natural to wonder whether there is always a minimal representation (α, Q) for f that keeps this maximal rate $||Q||$ moderate in size. To formulate a precise conjecture along these lines, let $c_{\min}(f)$ denote the infimum of the c's given by Maier's lemma for a given phase-type density f.

Conjecture 6: *There is a universal constant Δ such that for any phase-type density f, of any order, f has a minimal representation (α, Q) satisfying*

$$||Q|| \leq \Delta c_{\min}(f).$$

Even in the triangular case this is not immediately obvious. There is a somewhat tentative connection between this conjecture and Kendall's *Markov group conjecture* [22]. We give Kendall's conjecture in a form due to Kingman [23]. To state it, for a subgenerator Q let $\langle Q \rangle$ denote the maximum of the magnitudes of the real parts of the eigenvalues of Q.

Markov group conjecture: *There is a universal constant $\delta > 0$ such that for any generator matrix Q of any order,*

$$\langle Q \rangle \geq \delta ||Q||.$$

Simply and imprecisely stated, this says that a generator with large rates must have eigenvalues with large real parts. To make the connection to Conjecture 6, suppose f is a phase-type density of order n, and suppose again that f also has degree n. Let (α, Q) be an order-n representation of f. Then the poles of the transform of f coincide with the eigenvalues of Q. It is readily verified that $c \geq \langle Q \rangle / 2$ (for otherwise the eigenvalues whose powers dominate the growth of the derivatives of g at zero in Maier's lemma are not positive). The *Markov group conjecture* would then imply that $||Q|| \leq 2c/\delta$, so that $||Q||$ *cannot* be "too big." This would trivially imply the conclusion of Conjecture 6 for this phase-type distribution, taking $\Delta = 2/\delta$, which is simply that we may *choose* a Q with $||Q||$ "not too big." The link here is that both conjectures give information relating eigenvalues of Q to the maximal rate $||Q||$. If we relax the condition that f be of degree n, then the link between the two conjectures vanishes.

9 Concluding remarks

I have outlined a collection of conjectures concerning phase-type distributions and have given some partial arguments. The conjectures are intentionally optimistic in that they generally propose simple extensions of what is known. Even if these conjectures prove to be false, they may help to suggest what to look for, or hope for, in developing

a deeper understanding of phase-type distributions. Some further conjectures were given in [46]. There is a possibility that the truth underlying these conjectures may be unpleasantly complex and that they may therefore be an unrewarding subject for investigation. The author offers a prize of $25 for the first published proof or disproof of each of Conjectures 1–6. (If you are interested in easy money, I suggest starting by trying to disprove Conjecture 4! For lots of money, try proving Conjecture 2, because Conjectures 1 and 3 will probably follow easily.)

Many of the conjectures here may be thought of as statements about nonnegative matrices, M-matrices, or exponentials of M-matrices; however, they arise most naturally from probabilistic considerations. See the "supplement" at the end of [6] for some related conjectures in the context of nonnegative matrices.

I have referred from time to time above to some proofs of special cases that I have in my notes. These proofs have not been reviewed by other researchers, and so the possibility of an error cannot be excluded.

Acknowledgements An early version of this paper was presented to the Department of Mathematics at the University of Queensland in December 1995. The author is most grateful to Phil Pollett for arranging my visit, to the department, for its hospitality, and to the participants at the most enjoyable "stochastic coffee mornings," Nigel Bean, Laird Breyer, Andrew Hart, Malcolm Faddy, and organizer Phil Pollett. for many stimulating discussions.

References

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

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

[3] S. Asmussen and M. Bladt. Renewal theory and queueing algorithms for matrix-exponential distributions. *Proceedings of the First International Conference on Matrix Analytic Methods in Stochastic Models* (A.S. Alfa, & S. Chakravarthy, eds.) Marcel Dekker, New York. 1996.

[4] Asmussen, S., and O'Cinneide C.A. Matrix-exponential distributions. To appear in *Encyclopaedia of Statistical Sciences* (Supplement). 1997.

[5] Assaf, D. and Levikson, B. Closure of phase-type distributions under operations arising in reliability theory. *Ann. Probab.* 10, 265–269. 1982.

[6] Berman, A., and Plemmons, R.J. *Nonnegative matrices in the mathematical sciences.* SIAM Classics in Applied Mathematics series. (First published by Academic Press, 1979.) 1994.

[7] Carrol, J.L., van de Liefvoort, A., and Lipsky, L. Solutions of M/G/1/N-type loops with extensions to M/G/1 and GI/M/1 queues. *Operations Research* 30(3), 490–513. 1982.

[8] Commault, C., and Chelma, J.P. On dual and minimal phase-type representations. *Laboratoire d'Automatique de Grenoble,* Technical Report. November 1991.

[9] Cox, D.R. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Camb. Phil. Soc.* 51, 433–441. 1955.

[10] Cox, D.R. On the use of complex probabilities in the theory of stochastic processes. *Proc. Camb. Phil. Soc.* 51, 313–319. 1955.

[11] Cumani, A. "On the canonical representation of Markov processes modelling failure time distributions." *Microelectronics and Reliability* 22(3), 583–602. 1982.

[12] Dehon, M., and G. Latouche. A geometric interpretation of the relations between the exponential and generalized Erlang distributions. *Advances in Applied Probability* 14, 885–897. 1982.

[13] Dharmadhikari, S.W. Sufficient conditions for a Markov process to be a function of a finite Markov chain. *Ann. Math. Statist.* 34, 1033–1041. 1963.

[14] Dmitriev, N., and Dynkin, E.B. On the characteristic numbers of a stochastic matrix. *C.R. (Doklady) Acad. Sci. URSS (N.S.)* 49, 159–162. 1945.

[15] Dmitriev, N., and Dynkin, E.B. On the characteristic numbers of stochastic matrices. *Bull. Acad. Sci. URSS. Ser. Math. [Izvestia Akad. Nauk SSSR]* 10, 167–184. 1945. (in Russian with English summary).

[16] Erlang, A.K. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineer's Journal* 10, 189–197. 1917-18.

[17] Golub, G.H., and Van Loan, C.F. *Matrix Computations.* Johns Hopkins University Press, Baltimore. 1989.

[18] Hall, M., and Newman, M. Copositive and completely positive quadratic forms. *Proceedings of the Cambridge Philosophical Society* 59, 329–339. 1963.

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

[20] Ipsen, I., and Meyer, C.D. Uniform stability of Markov chains. *SIAM J. Matrix Analysis and Applications.*

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

[22] Kendall, D.G. On Markov groups. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* 47-72. University of California Press. 1973.

[23] Kingman, J.F.C. Three unsolved problems in discrete Markov theory. *Stochastic Analysis: A Tribute to the Memory of Rollo Davidson* 180-191. John Wiley, London. 1973.

[24] Kohno, S. Tokyo Institute of Technology. Private communication. 1992.

[25] Lehmann, E.L. *Theory of point estimation.* Wiley, New York. 1983.

[26] Lucantoni, D.M. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models* 7(1), 1–46. 1991.

[27] Lucantoni, D.M., Choudhury, G.L., and Whitt, W. The transient *BMAP/G/1* queue. *Stochastic Models* 10(1), 145–182. 1994.

[28] M.A. Johnson, and M.R. Taaffe. The Denseness of Phase Distributions. Research Memorandum No. 88-20, School of Industrial Engineering, Purdue University. 1988.

[29] Johnson, M.A. Private communication. 1992.

[30] Latouche, G., and Ramaswami, V. A Logarithmic Reduction Algorithm for Quasi-Birth-Death Processes. *J. Appl. Prob.* 30, 650-674. 1993.

[31] Lipsky, L. *Queueing Theory: A Linear Algebraic Approach.* Macmillan, New York. 1992.

[32] Lucantoni, D.M., Meier-Hellstern, K.S., and Neuts, M.F. A Single-Server Queue with Server Vacations and a Class of Non-renewal Arrival Processes. *Advances in Applied Probability* 22, 676-705. 1990.

[33] Maier, R.S. The algebraic construction of phase-type distributions. *Stochastic Models* 7(4), 573–602. 1991.

[34] Maier, R.S., and O'Cinneide, C.A. A closure characterization of phase-type distributions. *Journal of Applied Probability* 29, 92–103. 1992.

[35] McMullen, P., and Shephard, G.C. *Convex Polytopes and the Upper Bound Conjecture.* London Mathematical Society Lecture Notes Series #3. Cambridge University Press, Cambridge. 1971.

[36] C. Moler, and C. Van Loan, Nineteen Dubious Ways to Compute the Exponential of a Matrix. *SIAM Review,* 20, No. 4, 801-836. 1978.

[37] M.F. Neuts. Probability Distributions of Phase Type. In *Liber Amicorum prof. Emeritus H. Florin,* Department of Mathematics, University of Louvain, Belgium, 173-206. 1975.

[38] Neuts, M.F. A versatile Markovian point process. *Journal of Applied Probability* 16, 764–779. 1979.

[39] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach.* The Johns Hopkins University Press, Baltimore. 1981.

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

[41] Neuts, M.F. Phase-type distributions: A bibliography. *Department of Systems and Industrial Engineering, University of Arizona,* Working Paper 89-005. 1989.

[42] O'Cinneide, C.A. On non-uniqueness of representations of phase-type distributions. *Stochastic Models* 5, No. 2, 247-259. 1989.

[43] O'Cinneide, C.A. Characterization of phase-type distributions. *Stochastic Models* 6, No. 1, 1-57. 1990.

[44] O'Cinneide, C.A. Phase-type distributions and invariant polytopes. *Advances in Applied Probability* 23, 515–535. 1991.

[45] O'Cinneide, C.A. Phase-type distributions and majorization. *Annals of Applied Probability* 1(2), 219–227. 1991.

[46] O'Cinneide, C.A. Triangular order of triangular phase-type distributions. *Stochastic Models* 9, No. 4, 507-529. 1993.

[47] O'Cinneide, C.A. Entrywise perturbation theory and error analysis for Markov chains. *Numerische Mathematik* 65, 109–120. 1993.

[48] O'Cinneide, C.A. Relative-error bounds for the LU decomposition via the GTH algorithm. *Numerische Mathematik* 73, 507–519. 1996.

[49] Phelps, R.R. *Lectures on Choquet's theorem.* Van Nostrand Mathematical Studies No. 7. Van Nostrand. Princeton, N.J. 1966.

[50] Ramaswami, V. The N/G/1 Queue and its Detailed Analysis. *Advances in Applied Probability* 12, 222-261. 1980.

[51] Ramaswami, V., and Lucantoni, D.M. On the merits of an approximation to the busy period of the $G/G/1$ queue. *Management Science* 25, 285–289. 1979.

[52] Sengupta, B. Markov processes whose steady state distribution is matrix-exponential with an application to the $GI/PH/1$ queue. *Advances in Applied Probability* 21(1), 159–180. 1989.

[53] Soittola, M. Positive rational sequences. *Theoretical Computer Science* 2, 317–322. 1976.

[54] Takahashi, Y. On the effects of small deviations in the transition matrix of a finite Markov chain. *Journal of the Operations Research Society of Japan* 16, 104–129. 1973.

[55] Tweedie, R.L. Perturbations of countable Markov chains and processes. *Ann. Inst. Statistical Math.* 32, 283–90. 1980.

[56] Whitt, W. The Queueing Network Analyzer. *Bell Sys. Tech. J.* 62, No. 9, 2279-2815. 1983.

[57] W. Whitt. Performance of the Queueing Network Analyzer. *Bell Sys. Tech. J.,* 62, No. 9, 2817-2843. 1983.

[58] Widder, D.V. *The Laplace Transform.* Princeton University Press. 1946.