Effective Minimization of Acyclic Phase-Type Representations

by
Reza Pulungan and Holger Hermanns
Effective Minimization of Acyclic Phase-Type Representations

Reza Pulungan and Holger Hermanns

Department of Computer Science, Saarland University
D-66123, Saarbrücken, Germany, Phone: +49(681)3025478, Fax: +49(681)3025636
{pulungan,hermanns}@cs.uni-sb.de

Abstract. Acyclic phase-type distributions are phase-type distributions with triangular matrix representations. They constitute a versatile modelling tool, since they (1) can serve as approximations to any continuous distributions, and (2) exhibit special properties and characteristics, which usually result in some ease in analysis. The size of the matrix representation has a strong influence on computational efforts needed when analyzing these distributions. This representation, however, is not unique, and two representations of the same distribution can differ drastically in size. This report proposes an effective procedure to aggregate the size of the matrix representation without altering the distribution.

Key words: Stochastic Models; Markov Models

1 Introduction

Phase-type (PH) distributions are enjoying increasing attention in various fields of computer and engineering sciences. They are frequently used as stochastic modelling aids in areas such as queueing theory [23, 2], computer network design [9, 21], and reliability analysis [8], for instance in the analysis of dynamic fault trees [22, 5].

In this report, we deal with continuous-time PH distributions. They are a versatile and tractable class of probability distributions, retaining the principal analytical tractability of exponential distributions in a more general setting: the class of PH distributions is topologically dense [20] on the support set \([0, \infty)\). Therefore they can be used to approximate other probability distributions or the trace of empirical distributions obtained from experimental observations.

Any PH distribution agrees with the distribution of the absorption time in some Markov chain with an absorbing state [23]. Such a Markov chain is usually the basis of numerical or analytical analysis for models involving that PH distribution, and it is therefore called the representation of that distribution. These representations are not unique: distinct absorbing Markov chains may represent the same distribution, and any PH distribution is represented by infinitely
many distinct absorbing Markov chains. The representations differ in particular with respect to their size, i.e. the number of states and transitions. Thus, for a given PH distribution, an obvious question to pose is what the minimal-size representation of that distribution may be—and how to construct it.

The nowadays standard approach for aggregating a Markov chain without altering its stochastic properties is based on lumpability, in various flavors [7, 19, 12, 4]. This aggregation is equally applicable to absorbing Markov chains, but the computed fix-point is not guaranteed to be the minimal representation in the above sense.

Therefore, the problem of identifying and constructing smaller-sized representations remains one of the most interesting theoretical research questions in the field of PH distributions. In this report, we focus on the class of acyclic PH (APH) distributions, namely distributions with upper triangular matrix representations. Like PH distributions, also APH distributions are topologically dense on the support set \([0, \infty)\) [20].

Practically, identifying smaller representations of APH distributions can have compelling computational impact. This is especially apparent in modelling formalisms that support compositionality, like PEPA [19], IMC [18], or dynamic fault trees [5]. In such formalisms, complex models are built from smaller and simpler components. The size of some complex model is roughly the product of the individual components’ sizes. In this setting, using smaller representations for the components can significantly alleviate the blow-up of the state space during composition. In some cases, this may turn an otherwise intractable model into one with tractable model size.

The systematic study of APH representations was pioneered by Cumani [11], and later by O’Cinneide [26, 27], who identified minimality conditions, but without algorithmic considerations. In fact, the quest for an algorithm to construct the minimal representation of any APH distribution has recently seen considerable advances: He and Zhang provided an algorithm for computing minimal representations of APH distributions [16]. This algorithm involves converting the given APH distribution to a representation that only contains the poles of the distribution. This representation is not necessarily an APH distribution, but a matrix-exponential distribution [14]. If this is the case, another state and its total outgoing rate are determined and added to the representation. This is performed one by one until an APH representation is obtained. This results in a representation of provably minimal size. This algorithm involves solving a system of nonlinear equations for each additional state. Since nonlinear programming is difficult, the practicality of this method for large models is not obvious, and has not been investigated so far.

The algorithm developed in this report addresses the very same problem, but in the opposite way. Instead of adding states to a representation until it becomes an APH representation, we eliminate states from a representation as we proceed, until no further elimination is possible. Each elimination of a state involves solving a system of linear equations. The resulting algorithm is of cubic complexity in the size of the original state space, and does only involve standard
Effective Minimization of Acyclic Phase-Type Representations

numerical steps. The algorithm is guaranteed to return a smaller representation. However, unlike the algorithm of He and Zhang, minimality of the final result is not guaranteed for arbitrary APH distributions.

In summary, this report contributes an efficient algorithm to minimize APH distributions. The algorithm is easy to implement and straightforward to parallelize. It only consists of vector-matrix multiplications and the solution of well-conditioned systems of linear equations. To illustrate the effectiveness of our approach, we study a dynamic fault tree scenario with a prototype implementation of the algorithm. As we will discuss, the results are promising.

This report extends [28] and contains details and proofs left out in [30].

2 Phase-Type Distributions

This section provides the preliminaries of PH distributions. This discussion is focused on APH distributions, their canonical forms and the transformation from arbitrary APH distribution to the canonical forms.

2.1 Parameterization Definition

Let \( \{X(t) \in \mathcal{I} | t \in \mathbb{R}^+\} \) be a homogeneous Markov process defined on a discrete and finite state space \( \mathcal{I} = \{s_1, s_2, \ldots, s_n, s_{n+1}\} \) and with time parameter \( t \in \mathbb{R}^+ := [0, \infty) \). The Markov process is a finite continuous-time Markov chain (CTMC). We view the structure of such a CTMC as a tuple \( \mathcal{M} = (S, R) \) where \( S \) is a finite set of states, and \( R \) a rate matrix \( R : S \times S \rightarrow \mathbb{R}^+ \). The rate matrix \( R \) is related to the corresponding infinitesimal generator matrix by:

\[
Q(s, s') = R(s, s') \text{ if } s \neq s' \quad \text{else} \quad Q(s, s) = - \sum_{s' \neq s} R(s, s') \quad \text{for all } s, s' \in S.
\]

If state \( s_{n+1} \) is absorbing (i.e., \( Q(s_{n+1}, s_{n+1}) = 0 \)) and all other states \( s_i \) are transient (i.e., \( Q(s_i, s_i) < 0 \)), the infinitesimal generator matrix of the Markov chain can be written as

\[
Q = \begin{bmatrix}
A & \bar{A} \\
0 & 0
\end{bmatrix}.
\]

Matrix \( A \) is called a PH-generator and it is non-singular because the first \( n \) states in the Markov chain are transient. Vector \( \bar{A} \) is a column vector where its component \( \bar{A}_i \), for \( 1 \leq i \leq n \), represents the transition rate from state \( s_i \) to the absorbing state. Since \( Q \) is an infinitesimal generator matrix, \( \bar{A} = -\bar{A}e \) where \( e \) is an \( n \)-dimensional column vector whose components are all equal to 1. Let \( \pi^{(0)} = (\alpha, \alpha_{n+1}) \) be the initial probability distribution of the CTMC. Then the probability distribution of the time to absorption in the CTMC is called a phase-type distribution. The pair \( (\pi^{(0)}, A) \) is called the representation of the PH distribution and \( PH(\pi^{(0)}, A) \) is used to denote the PH distribution that has representation \( (\pi^{(0)}, A) \). The dimension of PH-generator \( A \) is called the order of the representation. All PH representations we are dealing with are assumed to be irreducible. A representation is irreducible if for the specified initial distribution all transient states are visited with non-zero probability.
The PH distribution is completely characterized by its (cumulative) distribution function and by its Laplace-Stieltjes transform (LST). The LST is a rational function. When expressed in irreducible ratio, the denominator of the LST has degree no more than \( n \). This degree of the denominator is called the degree of the distribution. The zeros of the denominator polynomial are called the poles of the distribution. It is known [23, 25] that a given PH distribution has more than one irreducible representation. The order of a minimal irreducible representation, namely a representation with the least possible number of states, is referred to as the order of the PH distribution. O’Cinneide in [26] showed that the order of a PH distribution may be different from but at least as great as its degree.

2.2 Acyclic Phase-Type Distributions

An interesting subset of the family of PH distributions is the family of acyclic PH distributions. The family can be identified by their triangular representations. A triangular representation is a representation whose PH-generator, under some permutation of its components, is an upper triangular matrix.

The family of acyclic PH distributions, in itself, is a very rich family of probability distributions. Among the members of the family are some important probability distributions which have been well-known and much-used in various models of random processes. These members include the exponential, hyperexponential, Erlang, hypoexponential, Cox distributions and their mixtures.

In [26], O’Cinneide proved the following theorem, which characterizes APH distributions in terms of the properties of their density functions and LSTs.

**Theorem 1 (Characterization Theorem [26]).** A probability distribution defined on \( \mathbb{R}^+ \), which is not the point mass at zero, is an acyclic PH distribution if and only if (1) its density function is strictly positive on \((0, \infty)\), and (2) its LST is rational and has only real poles.

Thus, any general PH representation—possibly containing cycles—represents an APH distribution (and hence has an acyclic representation) whenever all of the poles of its LST are real numbers.

**Example 1 (Fault Tolerant System).** Fault trees [17] are used to model fault-tolerant systems and to analyze their reliability. The fault tree in Fig. 1–(left) models a fault-tolerant system, which consists of two identical components. The redundancy of the components is introduced to increase the system’s reliability. The system fails if both components fail. Further, each component is comprised of three subcomponents, whose failure times are governed by exponential distributions with rate 1, 2 and 3, respectively. A component is operational if it has at least two functional subcomponents.

The fault tree gives rise to an APH distribution whose representation is depicted in Fig. 1–(right). The representation is obtained by applying a formalization of the dynamic fault tree semantics [22, 5]. Note that this representation is the result of a previous aggregation by weak bisimulation algorithm. This representation is not minimal. In the subsequent sections, we will provide the minimal representation.
2.3 Acyclic Canonical Forms

Let an infinitesimal generator matrix of the form

\[
\begin{bmatrix}
  -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\
  0 & -\lambda_2 & \lambda_2 & \cdots & 0 \\
  0 & 0 & -\lambda_3 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & 0 & \cdots & -\lambda_n
\end{bmatrix},
\]

be denoted by \( \text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n) \). A representation \((\beta, \text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n))\) is referred to as a bidiagonal representation.

Cumani in [11], presented canonical forms of APH representations. In particular, he proved that every APH representation has a bidiagonal representation of the same or less order. Aside from the bidiagonal representation, he also provided two other canonical forms and straightforward procedures to transform one to others.

A similar theorem was proved by O’Cinneide in [26]. Using the invariant polytopes method [25], he provided the characterization of APH distributions and proceeded to prove that such distributions are PH distributions with ordered bidiagonal representations. The following theorem provides the first of Cumani’s canonical forms (called the series canonical form).

**Theorem 2 ([11]).** Let \( PH(\alpha, A) \) be an APH representation and let the diagonal components of \(-A\) be \( \lambda_n \geq \lambda_{n-1} \geq \cdots \geq \lambda_1 \), so ordered. Then there is an ordered bidiagonal representation such that \( PH(\beta, \text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n)) = PH(\alpha, A) \).

Since all of the diagonal components of matrix \(-\text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n)\) are of a particular order (namely ascending), the first canonical form is also called the ordered bidiagonal representation.
Example 2 (Ordered Bidiagonal Representation). Fig. 2 depicts the ordered bidiagonal representation (first canonical form) of the triangular representation shown in Fig. 1–(right).

Cumani referred to his second canonical form as the canonical form A. We skip the discussion of this form, since we will not need it in the rest of the report. For the third canonical form (B), let an infinitesimal generator matrix of the form

\[
\begin{pmatrix}
-\lambda_1 & x_1 \lambda_1 & 0 & \cdots & 0 \\
0 & -\lambda_2 & x_2 \lambda_2 & \cdots & 0 \\
0 & 0 & -\lambda_3 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & -\lambda_n
\end{pmatrix},
\]

be denoted by \( \text{Cox}([\lambda_1, x_1], [\lambda_2, x_2], \ldots, [\lambda_n]) \). \( (\gamma, \text{Cox}([\lambda_1, x_1], [\lambda_2, x_2], \ldots, [\lambda_n])) \) is referred to as a Cox representation. The following theorem provides the transformation from the first to the third of Cumani’s canonical forms.

**Theorem 3 ([11]).** Let \( (\beta, \text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n)) \) be an ordered bidiagonal representation and let initial probability vector \( \gamma = [\beta e, 0, \cdots, 0] \).

\[
\text{PH}(\beta, \text{Bi}(\lambda_1, \lambda_2, \cdots, \lambda_n)) = \text{PH}(\gamma, \text{Cox}([\lambda_n, x_n], [\lambda_{n-1}, x_{n-1}], \cdots, [\lambda_1]))
\]

if and only if

\[
x_i = 1 - \beta \prod_{j=i+1}^{n} \frac{1}{x_j}, \quad \text{for } i = 2, \cdots, n. \quad (1)
\]

**2.4 Transformation Algorithms**

The proof of Theorem 2, provided by Cumani in [11], is constructive and can be turned into an algorithm to transform any acyclic representation to its ordered bidiagonal representation. However, the transformation algorithm is not polynomial in time, for it examines individual traces of the original acyclic representations.

In [24], O’Cinneide presented another similar algorithm which is based on the concept of PH-simplicity and PH-majoration. We will not treat the concepts in detail in this report, we just present the result instead.

For a generator matrix of a PH distribution \( \mathbf{B} \), let \( \text{PH}(\mathbf{B}) \) be the set of all distributions \( \text{PH}(\beta, \mathbf{B}) \) where \( \beta \) ranges over all sub-stochastic vectors of
appropriate dimension. The generator $B$ is called \textit{PH-simple} if all distributions in $PH(B)$ are pairwise distinct. A generator matrix $B$ \textit{PH-majorizes} $A$ if $PH(A) \subseteq PH(B)$. It is proven in [24] that if $B$ is PH-simple and $B$ PH-majorizes $A$ then there is exactly one matrix $P$ with unit row-sums such that $AP = PB$ and $P$ is of full rank if and only if $A$ is also PH-simple.

O’Cinneide’s algorithm proceeds by solving $P^{-1}A = BP^{-1}$ where $B$ is a bidiagonal representation matrix. However, this requires $P$ to be full rank and hence $A$ to be PH-simple, which restricts the use of the algorithm.

He and Zhang in [15] provided a better performing algorithm, which they called the \textit{spectral polynomial algorithm}, to obtain the bidiagonal representation of a given APH distribution. The algorithm does not make any assumptions on the PH-simplicity of the given generator matrix $A$ and may result in a bidiagonal representation of smaller order.

Let $PH(\alpha, A)$ be an APH distribution and let $\{-\lambda_1, -\lambda_2, \cdots, -\lambda_n\}$ be the eigenvalues of $A$. Denote the $i$-th column of matrix $P$ by $P_{*,i}$, then $AP = PBi(\lambda_1, \lambda_2, \cdots, \lambda_n)$ can be expressed by

$$AP_{*,i} = -\lambda_iP_{*,i} + \lambda_{i-1}P_{*,i-1}, \quad i = 1, \cdots, n,$$

with $P_{*,0} = 0$. The system of equations can be rewritten as

$$P_{*,i} = \frac{1}{\lambda_i}(A + \lambda_{i+1}I)P_{*,i+1}, \quad i = 1, \cdots, n-1.$$

For the matrix $P$ to be of unit row-sums, we have to set $P_{*,n} = -A\ell/\lambda_n$ (consult [15] for more details). If there exists a sub-stochastic vector such that $\beta = \alpha P$, then $PH(\alpha, A) = PH(\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))$. When the bidiagonal generator matrix is in canonical form, namely if $\lambda_n \geq \lambda_{n-1} \geq \cdots \geq \lambda_1$, then such sub-stochastic vector always exists because $Bi(\lambda_1, \lambda_2, \cdots, \lambda_n)$ PH-majorizes $A$.

The spectral polynomial algorithm has complexity $O(n^3)$ where $n$ is the order of the given acyclic representation.

3 Reducing the Order of Representations

In this section, we propose a procedure to reduce the order of acyclic representations. Roughly the procedure is as follows: (1) Given an APH distribution with representation $(\alpha, A)$, we transform the representation into an ordered bidiagonal representation $(\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))$ by using spectral polynomial algorithm of He and Zhang. The order of the new representation is at most the same as the order of the original one. (2) Once an ordered bidiagonal representation is obtained, some of its states can be removed without affecting its distribution function. In the rest of the section, we will show that a smaller representation can be obtained by removing some unnecessary states from the ordered bidiagonal representation. A method for identifying and removing those unnecessary states will be provided. The resulting representation is also an ordered bidiagonal representation of fewer number of states.
3.1 The $L$-terms

Let $L(\lambda) = \frac{s + \lambda}{s + \lambda}$ be the reciprocal of the LST of an exponential distribution with rate $-\lambda$. We refer to a single expression of $L(\cdot)$ as an $L$-term. The LST of an ordered bidiagonal representation $(\beta, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_n))$ can be written as

$$\tilde{f}(s) = \frac{\beta_1 L(\lambda_1) \cdots L(\lambda_n)}{L(\lambda_1) \cdots L(\lambda_n)} + \frac{\beta_2 L(\lambda_2) \cdots L(\lambda_n)}{L(\lambda_2) \cdots L(\lambda_n)} + \cdots + \frac{\beta_n L(\lambda_n)}{L(\lambda_n)}.$$ \hspace{1cm} (2)

Note that the LST expression in Equation (2) may not be in irreducible ratio form. The LST is produced in such a way that the denominator polynomial corresponds exactly to the sequence of the transition rates of the ordered bidiagonal representation. Hence, the degree of the denominator polynomial is the same as the order of the ordered bidiagonal representation.

3.2 Reducing the Order of Bidiagonal Representations

Observing Equation (2), if we are to remove a state from the ordered bidiagonal representation, we will have to find a common $L$-term in both of the numerator and denominator polynomials of Equation (2). Removing such a common $L$-term from the numerator and denominator polynomials means removing the corresponding state from the representation.

However, such removal of a state can only be carried out if after the removal, the initial probability distribution $\beta$ is redistributed in a correct manner, namely the new initial probability distribution, say $\gamma$, is a sub-stochastic vector. This procedure of identifying and properly removing a state from an ordered bidiagonal representation is described formally in the following lemma.

**Lemma 1.** Let $(\beta, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_n))$ be an ordered bidiagonal representation of order $n$. If for some $1 \leq i \leq n$, $\beta_1 + \beta_2 L(\lambda_1) + \cdots + \beta_i L(\lambda_1) L(\lambda_2) \cdots L(\lambda_{i-1})$ is divisible by $L(\lambda_i)$ then

$$PH(\beta, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_n)) = PH(\gamma, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_n))$$

if $\gamma$ is a sub-stochastic vector. $(\gamma, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_n))$ is an ordered bidiagonal representation of order $n - 1$.

**Proof.** Assume that the conditions are true. Then we can find a common $L$-term in both the numerator and denominator polynomials of the LST of representation $(\beta, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_n))$ because the rest of the numerator polynomial is divisible by $L(\lambda_i)$. Removing this common $L$-term reduces the order by 1 and hence we obtain a new representation $(\gamma, \text{Bi}(\lambda_1, \lambda_2, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_n))$.

Since the new representation is constructed from the same LST (albeit simplified), it constitutes the same PH distribution. \qed
The first observation we can gain from inspecting the form of Equation (2) and Lemma 1 is that 

\[ L(\lambda_1) \] cannot divide the numerator polynomial if \( \beta_1 \neq 0 \). Hence, Erlang distributions and the convex combination of Erlang distributions with the same parameter are always irreducible.

To reduce the order of an ordered bidiagonal representation, we need to check two conditions in Lemma 1, namely the divisibility of the numerator polynomial and the sub-stochasticity of the resulting initial probability vector. For this, let

\[ R(s) = \beta_1 + \beta_2 L(\lambda_1) + \cdots + \beta_i L(\lambda_1)L(\lambda_2)\cdots L(\lambda_{i-1}). \]  

(3)

By ordinary algebra, we can check whether \( R(s) \) is divisible by \( L(\lambda_i) \) simply by checking whether \( R(-\lambda_i) = 0 \).

The sub-stochasticity of the resulting initial probability vector, on the other hand, can be checked while computing it. Referring back to Lemma 1, let us denote the generator matrix \( B_{1}(\cdot) := B_{1}(\lambda_{1}, \lambda_{2}, \cdots, \lambda_{n}) \) and the generator matrix \( B_{2}(\cdot) := B_{2}(\lambda_{1}, \lambda_{2}, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_{n}) \). The two ordered bidiagonal representations represent the same PH distribution, hence

\[ PH(\beta, B_{1}(\cdot)) = PH(\gamma, B_{2}(\cdot)), \]  

(4)

where \( e_n \) is a column vector of dimension \( n \) whose components are all equal to 1. To compute vector \( \gamma \) from vector \( \beta \), we need \( n - 1 \) equations relating their components. We can evaluate Equation (4) at \( n - 1 \) different \( t \) values to obtain such required system of equations. However, such function evaluations can be costly. To avoid this, we proceed differently.

For a PH representation \( (\alpha, A) \), the \( i \)-th derivative, for \( i > 0 \), of its distribution function is given by

\[ F^{(i)}(t) = -\alpha A^i \exp(At)e. \]

Evaluating these derivatives at \( t = 0 \) allows us to avoid computing the exponential of matrices. Hence, the components of vector \( \gamma \) can be computed by solving the following system of equations

\[ \beta B_{1}(\cdot)^i e_n = \gamma B_{2}(\cdot)^i e_{n-1}, \quad i = 0, \cdots, n - 2. \]  

(5)

Once Equation (5) is solved, the sub-stochasticity of \( \gamma \) can be determined simply by checking that all of its components are nonnegative real numbers. The equation and the sub-stochasticity of \( \gamma \) are the sufficient conditions for \( B_{1}(\cdot) = B_{2}(\cdot) \). For the necessity proof we refer to Lemma 2 in Section 3.5.

3.3 The Algorithm

Let such a state which corresponds to the \( L \)-term \( L(\lambda_i) \) in Lemma 1 be called a removable state. Thus, a state is removable in an ordered bidiagonal representation, if the numerator polynomial of the LST is divisible by the \( L \)-term
corresponding to the state and if removing the state results in a valid initial probability distribution.

Lemma 1 can be turned into an algorithm that reduces the order of a given APH representation.

**Algorithm 1 (Reducing Acyclic Representation)**

1: function REDACYCRep(β, Bi)
2:     n ← ORDERof(β, Bi)
3:     i ← 2
4:    while i ≤ n do
5:        if REMOVABLE(i, (β, Bi)) then
6:            (β, Bi′) ← REMOVE(i, (β, Bi))
7:            n ← n − 1
8:        else
9:            i ← i + 1
10:    end if
11:    (β, Bi) ← (β, Bi′)
12: end while
13: return (β, Bi)
14: end function

Algorithm 1 inputs an ordered bidiagonal representation and outputs its reduced ordered bidiagonal representation. Function ORDERof(·) returns the order of the given representation. Function REMOVABLE(·) returns True if state $s_i$ is removable from the given representation. Removing it means redistributing the initial probability distribution, whose computation was shown in Equation (5). This is basically what function REMOVE(·) does, namely removing state $s_i$ from the given representation.

The algorithm proceeds by checking each state whether it is removable or not. If it is removable then it is eliminated. Further, the algorithm terminates once all states have been checked and the removable ones have been eliminated. Hence, the algorithm does what it is supposed to do and terminates.

Let $n$ be the order of the given APH representation. The spectral polynomial algorithm of He and Zhang entails $n$ matrix-vector multiplications, thus amounts to $O(n^3)$. Checking whether state $s_i$ is removable needs an evaluation of Equation (3), which requires $(i^2 − i)$ multiplications and $(i^2 + i − 1)$ additions. At worst this costs $O(n^2)$.

If state $s_i$ is removable, eliminating it entails solving Equation (5). We observe that for any bidiagonal generator $Bi$ of dimension $d$, $Bi(s_i, s_i) = −Bi(s_i, s_{i+1})$ for $0 < i < d$. Now, since both matrices in the system of equations are bidiagonal generators, this observation allows us to show—by induction on $n$—that Equation (5) can be transformed into

\[ \underline{b} = A \underline{\gamma}^T \]  

where $A$ is an upper triangular matrix and $\underline{b}$ is a column vector which is obtained by evaluating the left hand side of Equation (5) $n − 1$ times. A detailed analysis
Effective Minimization of Acyclic Phase-Type Representations

shows that this transformation requires \( \frac{3n^2 + 5n}{2} \) multiplications and \( (n^2) \) additions, and is thus of order \( O(n^2) \) time. Equation (6), on the other hand, can be solved in \( O(n^2) \) time, because \( A \) is upper triangular, which means we only need to apply backward substitutions.

Since the procedure for checking whether a state is removable and then removing it is actually carried out at most \( n \) times, hence the overall time complexity of the algorithm is \( O(n^3 + n(n^2 + n^2 + n^2)) = O(n^3) \).

**Example 3 (Reduced Acyclic Representation).** Inputting the model depicted in Fig. 1 into Algorithm 1, it is first transformed into an ordered bidiagonal representation. The ordered bidiagonal representation has the same number of states as the original representation. It is depicted in Fig. 2.

In the following, we describe the elimination of state 4 from the resulting ordered bidiagonal representation. First, as we have observed before, state 1 cannot be eliminated, because the numerator polynomial cannot be divisible by \( L(3) \). Next, state 2 is not removable because \( \frac{7}{5280} + \frac{599}{142560} L(3) = \frac{7}{142560} L(3) \) is not divisible by \( L(4) \). State 3 also, since \( \frac{7}{5280} + \frac{1267}{142560} L(3) + \frac{149}{9504} L(3) \) is not divisible by \( L(5) \). For state 4, on the other hand, \( \frac{7}{5280} + \frac{1267}{142560} L(3) + \frac{149}{9504} L(3) L(4) + \frac{149}{9504} L(3) L(4) L(5) \), is divisible by \( L(6) \).

Now we have to compute the new initial probability distribution. Note, however, that the initial probabilities of states 1, 2, 3 and 4 are redistributed only to states 1, 2 and 3 after the removal of state 4. This means that instead of using the whole matrix, we need only to compute the new initial probabilities of states 1, 2, 3 and 4. The system of equations to be solved, then, is

\[
\begin{bmatrix}
\frac{7}{5280}, & \frac{599}{142560}, & \frac{1267}{142560}, & \frac{149}{9504}
\end{bmatrix} \overline{Bi(3, 4, 5, 6)} \rightarrow \begin{bmatrix}
\gamma_1, & \gamma_2, & \gamma_3
\end{bmatrix} \overline{Bi(3, 4, 5)}
\]

for \( i = 0, 1, 2 \). The solution is \( \gamma = \begin{bmatrix}
\frac{7}{2040}, & \frac{41}{4752}, & \frac{149}{7920}
\end{bmatrix} \).

![Fig. 3. A Minimal Acyclic Representation of Fig. 2](image)

The ordered bidiagonal representation after all its removable states are eliminated is shown in Fig. 3. It is a minimal representation, because its order is the same as its algebraic degree.

### 3.4 Non-minimal Representation

In the previous examples, the algorithm reduces the acyclic representation to a minimal representation. This can be verified by the fact that the order of the representation depicted in Fig. 3 is the same as its degree. The degree is precisely
the number of $L$-terms which are not the divisor of the numerator polynomial, which in this case is the same as the number of states in the representation.

However, to arrive at a minimal result is not always possible. To shed some light on this, consider the representations depicted in Fig. 4. Both representations have the same distribution. Given representation in Fig. 4–(left), our algorithm cannot produce smaller representations, because both states 4 and 5 are not removable. However, the representation in Fig. 4–(right) is a minimal representation, and it is smaller than what the algorithm returns.

![Fig. 4. Non-Minimal (left) and Minimal (right) Acyclic Representations](image)

The reason behind this deficiency lies in the fact that the algorithm is bound to the set of present total outgoing rates (as $L$-terms), while in reality the representation depends on the interplay of total outgoing rates and initial probability distribution. These dependencies are in full generality difficult to detect, because we are then left with the problem of finding matches over a continuous domain of candidates, akin to the nonlinearity of the problem encountered by He and Zhang [16].

### 3.5 The Use of Order Reduction

Let $Red(\alpha, A)$ be the reduced representation of APH representation $(\alpha, A)$ obtained by the application of Algorithm 1.

**Lemma 2.** Let $(\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))$ be an ordered bidiagonal representation. For an arbitrary $\lambda_i \geq \lambda_i$, for $1 \leq i \leq n$, we can find an ordered bidiagonal representation $(\beta', Bi(\lambda_1, \cdots, \lambda_i, \lambda_{i+1}, \cdots, \lambda_n))$ that has the same distribution.

**Proof.** The LST of the original ordered bidiagonal representation expressed in $L$-terms is

$$\frac{\beta_1 + \cdots + \beta_n L(\lambda_1) L(\lambda_2) \cdots L(\lambda_{n-1})}{L(\lambda_1) L(\lambda_2) \cdots L(\lambda_n)}.$$

which can be rewritten as

$$\frac{\beta_1}{L(\lambda_1) L(\lambda_2) \cdots L(\lambda_n)} + \frac{\beta_1 L(\lambda_1) + \cdots + \beta_n L(\lambda_1) L(\lambda_2) \cdots L(\lambda_{n-1})}{L(\lambda_1) L(\lambda_2) \cdots L(\lambda_n)}.$$  \hspace{1cm} (7)

Using identity

$$\frac{1}{L(\lambda)} = \frac{p}{L(\mu)} + \frac{(1-p)}{L(\lambda) L(\mu)},$$
where \( \lambda \leq \mu \) and \( p = \frac{\lambda}{\mu} \), the first term of Equation (7) can be written
\[
\left( \frac{\beta_1 (1 - w)}{L(\lambda_1) L(\lambda_x)} + \frac{\beta_2 w}{L(\lambda_2) \cdots L(\lambda_n)} \right) \frac{1}{L(\lambda_2) \cdots L(\lambda_n)} + \frac{\beta_1 L(\lambda_1) + \cdots + \beta_n L(\lambda_1) L(\lambda_2) \cdots L(\lambda_{n-1})}{L(\lambda_1) L(\lambda_2) \cdots L(\lambda_n)}. \tag{8}
\]
where \( w = \frac{\lambda_1}{\lambda_x} \). It is straightforward to construct an APH representation of Equation (8). Using [29, Lemma 3], we obtain an ordered bidiagonal representation \((\beta', Bi(\lambda_1, \cdots, \lambda_1, \lambda_x, \lambda_{i+1}, \cdots, \lambda_n))\) of order \( n+1 \) for some sub-stochastic vector \( \beta' \).

Since throughout the procedure, the LST remains the same, the two ordered bidiagonal representations have the same distribution. \(\Box\)

**Lemma 3.** Let \((\alpha, A)\) be an APH representation. Furthermore, let the algebraic degree and the order of \(PH(\alpha, A)\) be the same. Then the order of \(Red(\alpha, A)\) is the same as the order of \(PH(\alpha, A)\).

**Proof.** Let \( n \) be the dimension of the generator matrix \( A \) and let the ordered bidiagonal representation of \((\alpha, A)\) be \((\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))\). The LST of the ordered bidiagonal representation expressed in \(L\)-terms is
\[
\frac{P(s)}{Q(s)} = \frac{\beta_1 + \cdots + \beta L(\lambda_1) L(\lambda_2) \cdots L(\lambda_{n-1})}{L(\lambda_1) L(\lambda_2) \cdots L(\lambda_n)}. \]
This LST may be reducible.

Suppose \( m \leq n \) be the order of \(PH(\alpha, A)\). Let \( \mathcal{P} \subset \{-\lambda_1, \cdots, -\lambda_n\} \) be the multiset of the \( m \) poles of the distribution and \( \overline{\mathcal{P}} \) be its complement. For each \( \lambda \in \overline{\mathcal{P}}, P(s) \) must be divisible by \( L(\lambda) \). Furthermore, Lemma 2 assures us that the state associated with \( L(\lambda) \) is removable, namely the stochasticity of the resulting initial distribution vector after the state is removed is preserved (second condition of Lemma 1). The uniqueness of the way to incorporate the state associated with \( L(\lambda) \) implies that the resulting initial distribution vector is also unique. Therefore we can utilize Algorithm 1 to remove all states associated with exit rates in \( \overline{\mathcal{P}} \). Hence the order of \( Red(\alpha, A)\) is the same as the order of \(PH(\alpha, A)\). \(\Box\)

Lemma 3 basically establishes that applying Algorithm 1 to any APH representation whose algebraic degree is equal to the order of the distribution is
certain to result in a minimal representation. The algorithm can also be applied to APH representations whose algebraic degree is different from order of the distribution, although in this case it is not guaranteed to result in a minimal representation.

4 Weak Bisimulation

One may wonder how likely it is to encounter a model which is reducible by our algorithm—or by lumping. Actually, Commault and Mocanu in [10] showed that for any pre-specified structured PH representation of order \( n \), the set of parameter values producing PH distributions of algebraic degree less than \( n \) has measure zero. In other words, the chance to find a reducible Markov chain appears very low. This is indeed true for models obtained from parameter estimation mechanism such as fitting methods.

However, this measure based interpretation is often misleading. Many cases have been reported that show the (sometimes tremendous) reduction obtained by lumping a model. The question thus is: why are aggregation techniques for Markov chains such as lumping successful after all? They are indeed successful for representations which are constructed out of a structured behavioral representation. The reason lies in the way models are constructed which usually results in very specific and particular parameter values:

Constructive mechanisms used in building Markovian models from smaller components usually involve convolution, choice and composition operations which correspond to stochastic operations: sum, minimum (or convex combination) and maximum, respectively. These operations often produce models with repeatable, symmetric, and similar sub-structures, which enable reduction by the aggregation techniques. In the context of APH, for instance, the idea of core series in [29] identifies these sub-structures in the set of paths to the absorbing state.

To shed some light on the relation of our reduction algorithm to lumpability, we here consider weak bisimilarity. The notion of weak bisimilarity is formalized in the following definition, which is a variation of the original definition [6, 4], accounting for the case of absorbing states (and otherwise treating the chain as unlabelled). For \( C \subseteq S \), let \( R(s, C) = \sum_{s' \in C} R(s, s') \). If \( R \) is an equivalence relation on \( S \), \( S/R \) is the partition of \( S \) induced by \( R \) and for \( s \in S \), \([s]_R \) is the class which contains \( s \).

**Definition 1 (Weak Bisimulation).** For \( M = (S, R) \) let \( R \) be an equivalence relation on \( S \). \( R \) is a weak bisimulation on \( M \) if for all \( s_1 R s_2 \): \( R(s_1, C) = R(s_2, C) \) for all \( C \in S/R \) with \( C \neq [s_1]_R \) and if absorbing states are only related to absorbing states. States \( s_1 \) and \( s_2 \) are weakly bisimilar, denoted \( s_1 \approx s_2 \) if and only if there exists a weak bisimulation \( R \) on \( M \) such that \( s_1 R s_2 \).

Weak bisimulation differs from the more prominent notion of strong bisimulation (or ordinary lumpability [7]), in that transitions which do not cross class boundaries are not considered. This can be seen as the ‘class generalization’ of
the fact that in a CTMC the values of $R(s, s)$ (i.e., loops at states) are irrelevant for the diagonal entries of $Q$.

States that are weakly bisimilar in a CTMC model can be lumped by moving to the quotient induced by this equivalence, thus producing an aggregated CTMC model. An algorithm for computing the weak bisimulation quotient is at hand [3]. It has cubic complexity, in the number of states of the original model.

Weak bisimulation and lumpability play an important role in CTMC modelling and analysis. Weakly bisimilar states and models possess the same probabilistic reachability properties [4], which means that their probability distributions of reaching certain subsets of the state space are the same. Hence weakly bisimilar models are exchangeable so far as their reachability properties are concerned. Weak bisimulation can be used to identify bisimilar states in a model, and by lumping them, to reduce the order of the model. Weak bisimulation can also be applied to PH representations, without affecting the distribution:

**Lemma 4.** Let $(\alpha, A)$ be a PH representation. If $(\beta, B)$ is obtained by lumping weak bisimilar states in $(\alpha, A)$, then $PH(\alpha, A) = \overline{PH}(\beta, B)$ and the order of $(\beta, B)$ is at most the order $(\alpha, A)$.

The proof is a straightforward consequence of the fact that weak bisimulation does not alter transient probabilities of state classes [4], together with the particular role played by absorbing states in our variation of weak bisimulation. We now make precise the relation between our order reduction and weak bisimulation.

**Lemma 5.** Let $(\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))$ be an ordered bidiagonal representation of order $n$. If

1. $\beta_i \neq 0$, and
2. $\overline{\beta_2} + \beta_i L(\lambda_1) = (\overline{\beta_2} + \beta_i)L(\lambda_i), i \in \{2, \cdots, n\}$ and $\overline{\beta_j} = 0$ for all $2 < j \leq i$,

then both $(\beta, Bi(\lambda_1, \lambda_2, \cdots, \lambda_n))$ and $(\gamma, Bi(\lambda_1, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_n))$ represent the same PH distribution, for some sub-stochastic vector $\gamma$ of dimension $n-1$. Moreover, there is a weak bisimulation relating their respective Cox representations.

**Proof.** Assume now the conditions of the lemma to be true. The divisibility of the numerator polynomial by $L(\lambda_i)$ is straightforward. For the sub-stochasticity of vector $\gamma$, observe that by the given conditions: $\gamma_i = \overline{\beta_i} + \overline{\beta_j}$ and $\gamma_j = \overline{\beta_j + 1}$ for all $2 \leq j \leq n - 1$. Hence state $s_i$ is removable and both representations are of the same PH distribution.

By Theorem 3, the original representation has Cox representation

$$(x, Cox([\lambda_n, x_n], \cdots, [\lambda_i, 1], \cdots, [\lambda_2, x_2], \lambda_1])$$

where $x = [0, 0, \cdots, 0]$. By Equation (1) in the same theorem, $x_j = 1$ for all $2 < j \leq i$. Therefore all states $s_k$’s, for $2 \leq k \leq i$, can be reordered and we have $(x, Cox([\lambda_n, x_n], \cdots, [\lambda_{i-1}, 1], \cdots, [\lambda_2, 1], [\lambda_1, x_2], \lambda_1])$. 

Rearranging $\beta_1 + \beta_2 L(\lambda_1) = (\beta_1 + \beta_2)L(\lambda_i)$ we obtain

$$\lambda_1 = \frac{\beta_1}{\beta_1 + \beta_2} \lambda_i.$$  \hspace{1cm} (9)

Inspecting the Cox representation or evaluating Equation (1), we can infer that the following relations hold

$$x_n x_{n-1} \cdots x_3 x_2 = \beta_1,$$ \hspace{0.5cm} and

$$x_n x_{n-1} \cdots x_3 (1-x_2) = \beta_2.$$ \hspace{1cm} (10)

From Equations (9) and (10): $\lambda_1 = (1-x_2)\lambda_i$, which means states $s_1$ and $s_i$ are weakly bisimilar. Lumping the two states, we obtain

$$(\kappa, Cox(\{\lambda_n, x_n\}, \cdots, \{\lambda_i-1, 1\}, \cdots, \{\lambda_2, 1\}, \lambda_1))$$

with appropriately sized $\kappa$. This representation is the Cox representation of $(\gamma, Bi(\lambda_1, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_n))$. Hence both representations are related by a weak bisimulation. \hfill \Box

The lemma indicates that once an acyclic model is transformed into its ordered bidiagonal representation, it may be possible to reduce the state space by weak bisimulation. In the circumstances described in the lemma, our order reduction coincides with weak bisimulation aggregation on the Cox representation of the APH distribution.

## 5 Implementation

We have implemented the reduction procedure as described in Algorithm 1 in C++. In order to handle the sensitivity of the initial probability distribution of the ordered bidiagonal representation in relation to the distribution it represents, we have to resort to using rational numbers in the implementation. For this purpose we use GMP library [1]. In this section we report the result of the implementation.

Fig. 5 depicts a model of fault-tolerant parallel processors (FTPP) [13]. The FTPP consists of four network elements (NE), which are fully connected to each others and four groups of processors. In each of these groups, one processor is originally powered down and is used as a spare. Each NE is attached to four processors, one from each group. Overall in the system, there are as many processors in a single group as the number of NEs. At least two of these processors are required to be operational for the whole system to be considered operational, otherwise it is considered failed. Each processor and each NE may experience failures which are governed by some exponential distributions. Furthermore, a failed NE brings down the four processors connected to it. Based on the fault tree of the FTPP model, we obtain the absorbing CTMC model associated with
the distribution of the time to system’s failure. A more detailed discussion of
this reliability model can be found in [5].

The resulting absorbing CTMC model is an APH representation. We use
Algorithm 1 to reduce the state space (order) of the representation. Table 1
summarizes the result of the experiments. We have three FTPP models where
we vary the number of NEs and thus the number of processors in each group.
For each model, the table provides the size of the state space before and after
the reduction. Note that the state spaces before the reduction have been minimized
according to a stochastic weak bisimulation algorithm. In general, the reduction
algorithm produces state spaces which are orders of magnitude smaller than
the original ones. The computation time (in seconds) of the reduction (RED)
procedure is shown in the last column of the table. The numbers in the brackets
are the computation times required to the transformation (TRA) to the ordered
bidiagonal representations by the spectral polynomial algorithm.

The models prior to the reduction are obtained by composing smaller components. Dur-
ing the composition, the size of the representation blows up fast, because it is basically a
cross-product operation over the smaller representations. However, when many of the smaller
components are stochastically similar, the composition alters the representation drastically
while the stochastic behavior remains the same. We expect the reduction algorithm to be most
useful in such circumstances. The FTPP examples we provide above demonstrates this.

Moreover, for instance in [29], it is shown that the minimal representation of
the compositions of several Erlang distributions is much smaller than the original
representation.

<table>
<thead>
<tr>
<th>#NE</th>
<th>States Before Reduction</th>
<th>Trans.</th>
<th>States After Reduction</th>
<th>Trans.</th>
<th>Time (sec.)</th>
<th>RED (TRA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>65</td>
<td>304</td>
<td>39</td>
<td>&lt;1</td>
<td></td>
<td>1 (1)</td>
</tr>
<tr>
<td>5</td>
<td>391</td>
<td>2956</td>
<td>76</td>
<td>13</td>
<td>944 (18062)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1728</td>
<td>17211</td>
<td>64</td>
<td>121</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Conclusions

This report has introduced an algorithm to reduce acyclic representations of PH
distributions (viz. acyclic absorbing CTMCs). In each iteration, the algorithm
requires quadratic time (in the current number of states), to reduce the representation by at least one state, as long as no further reduction is possible. We have highlighted that the resulting CTMC is not necessarily minimal, and have clarified the relation to weak bisimulation lumping.

In practice, one may wonder whether it is beneficial to run an overall cubic algorithm to reduce the matrix representation of a PH distribution. This, of course, depends on the application context. If one intends to numerically compute the absorption probability at many time points (or at a single large time point) it might very well be worthwhile to run the suggested algorithm as a preprocessing step. This step reduces the dimensions of the involved matrices and vectors, and hence speeds up the subsequent (usually uniformization-based) iterations.

Further, if the CTMC representation is used in a concurrency context, then a one-state saving in a single component—prior to exploring the crossproduct with other components—saves states in the order of the entirety of all other components. This can in general lead to an exponential saving in size of the overall system.

References