Chapter 2
Phase-Type Distributions

2.1 Basic Definitions

Continuous-time Markov chains (CTMCs) are a class of stochastic processes with a discrete state space in which the time between transitions follows an exponential distribution. In this section, we first provide the basic definitions for CTMCs and notations associated with this model. We then proceed with an explanation of the basic concepts for phase-type distributions (PHDs) and the analysis of such models. For theoretical details about CTMCs and related stochastic processes we refer to the literature [151].

2.1.1 Markov Chains

Let $\mathcal{S}$ denote a countable set of states, and let $\{X(t)\}_{t\geq 0}^{\infty}$ be a stochastic process with state space $\mathcal{S}$. Let $n \in \mathbb{N} \cup \{\infty\}$ be the size of the state space. Since the state space is isomorphic to (a subset of) $\mathbb{N}$, we denote states by their numbers.

$\{X(t)\}_{t\geq 0}^{\infty}$ is a continuous-time Markov chain, if it is characterized by the *Markov property* [151]:

$$Prob(X(t_{k+1}) = x_{k+1}|X(t_k) = x_k, \ldots, X(t_0) = x_0) = Prob(X(t_{k+1}) = x_{k+1}|X(t_k) = x_k), \tag{2.1}$$

for any $0 \leq t_0 \leq t_1 \leq \ldots \leq t_k \leq t_{k+1}$ and $x_l \in \mathcal{S}$. Thus, indicating that given the current state x_k and the time t_k, the value taken by $X(t_{k+1})$ depends only on x_k and on t_k but not on the past of the process.

The process is time homogeneous if for all $u \geq 0$

$$Prob(X(t+u) = j|X(u) = i) = Prob(X(t) = j|X(0) = i) = p_t(i, j). \tag{2.2}$$

P. Buchholz et al., *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*, SpringerBriefs in Mathematics,
DOI 10.1007/978-3-319-06674-5_2, © Peter Buchholz, Jan Kriege, Iryna Felko 2014

The *homogeneity* in the definition of homogeneous Markov chains is due to the fact that the transition probabilities only depend on the difference t between u and $t+u$ and not on the actual times $(u, t+u)$. The values $p_t(i,j)$ define a matrix with transition probabilities $\mathbf{P}_t$.

The state probabilities at time t are denoted by $p_t(j) = Prob(X(t) = j),\ j \in \mathcal{S}$ with $\sum_j p_t(j) = 1$. Consequently, the vector $\boldsymbol{\pi}(0) = [p_0(1), p_0(2), \ldots]$ is the initial probability vector of the CTMC. Moreover, the random times between state transitions are exponentially distributed random variables $V_1, V_2, \ldots$ with parameter $\lambda(i)$ for state i. That is, each V_i describes the exponential holding time in state i and has distribution function

$$Prob(V_i \leq t) = 1 - e^{-\lambda(i)t},\ \ t \geq 0. \tag{2.3}$$

Such a process evolves as follows: at any time t, $X(t) = i$, the process remains in state i for a period of time determined by an exponential distribution with parameter $\lambda(i)$, $0 \leq \lambda(i) < \infty$, and then jumps to a state j with probability $p(i,j) = \lambda(i,j)/\lambda(i)$. Therefore, $\lambda(i,j)$ is the rate at which a state transition from i to j occurs. Moreover, $\lambda(i)$ represents the total event rate characterizing state i. We now summarize the probabilistic behavior of the continuous time Markov chain in terms of its infinitesimal generator [151]. The infinitesimal generator is a $n \times n$ matrix $\mathbf{Q}$ with components

$$\mathbf{Q}(i,j) = \begin{cases} -\lambda(i) & \text{if } i = j, \\ \lambda(i,j) & \text{if } i \neq j. \end{cases} \tag{2.4}$$

$\mathbf{Q}$ is also denoted as the transition rate or generator matrix. Since $\lambda(i) \geq 0$, it follows $\mathbf{Q}(i,i) \leq 0$ indicating that all diagonal elements of the matrix $\mathbf{Q}$ are non-positive. If the transition to some state j is feasible in state i then $\mathbf{Q}(i,j) > 0$, otherwise, $\mathbf{Q}(i,j) = 0$. Thus, all non-diagonal elements must be non-negative. One can now see, that from the definition of the rates it follows that

$$\sum_j \mathbf{Q}(i,j) = 0. \tag{2.5}$$

The embedded process $\{X_r\}_{r \in \mathbb{N}_0}$, with $X_0 = X(0)$, is a discrete-time Markov chain with transition probability matrix $\mathbf{P}$. It applies for the elements $\mathbf{P}(i,j) = Prob(X(r+1) = j | X(r) = i)$ which can be expressed in terms of $\mathbf{Q}$

$$\mathbf{P}(i,j) = \frac{\mathbf{Q}(i,j)}{-\mathbf{Q}(i,i)},\ \text{ for } j \neq i, \mathbf{Q}(i,i) \neq 0, \tag{2.6}$$

and $\mathbf{P}(i,i) = 0$ in this case. For $\mathbf{Q}(i,i) = 0$ also $\mathbf{Q}(i,j) = 0$ for all $j \neq i$. In this case we define $\mathbf{P}(i,i) = 1$ and $\mathbf{P}(i,j) = 0$ for $i \neq j$. States with $\mathbf{Q}(i,i) = 0$ are denoted as absorbing states. Note that summing over all j results in $\sum_j \mathbf{P}(i,j) = 1$.

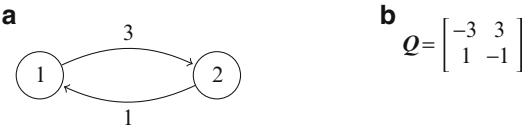


Fig. 2.1 State transition diagram and generator matrix of a CTMC. (**a**) A state transition diagram of the CTMC. (**b**) The infinitesimal generator **Q**

Markov chains can be equivalently described by the generator matrix **Q** or by a state transition diagram as shown in Fig. 2.1. The edge connecting states i and j is weighted with the transition rate from i to j, i.e., with $\mathbf{Q}(i, j)$.

2.1.2 Absorbing Markov Chains and Phase-Type Distributions

Next we consider Markov chains with absorbing states. We will introduce several definitions and classifications of the states of a Markov chain in ways that turn out to be particularly convenient for our purposes, namely the study of the behavior of the chain up to the moment that it enters an absorbing state. The states of the Markov chain can be classified according to whether it is possible to move from a given state to another given state. Clearly, if a state j is reachable from a state i we have $p_t(i, j) = Prob(X(t+u) = j|X(u) = i) > 0$ for some t, i.e., the process can move from state i to state j after some amount of time t. A complete treatment of this classification can be found in [93, 151]. We will take a few concepts from that description.

Definition 2.1. States i and j can communicate with each other if i is reachable from j and vice versa.

Let $\mathcal{C}$ be a subset of the state space $\mathcal{S}$. If all states in set $\mathcal{C}$ communicate, we call it a communicating set. If there no feasible transition from any state in $\mathcal{C}$ to any state outside $\mathcal{C}$, then $\mathcal{C}$ forms a closed set:

Definition 2.2. A subset $\mathcal{C}$ of the state space $\mathcal{S}$ is said to be closed if $\mathbf{P}(i, j) = 0$ for any $i \in \mathcal{C}$, $j \notin \mathcal{C}$.

If $\mathcal{C}$ consists of a single state, say i, then i is said to be an absorbing state.

A closed set $\mathcal{C}$ where all members communicate is a closed communicating set.

Clearly, if i is an absorbing state it holds that $\mathbf{P}(i, i) = 1$. A process can never leave a closed set after entering it. The following two classifications indicate whether and when a process returns to a state after leaving it.

Definition 2.3. A state $i \in \mathcal{S}$ is a transient state, if the probability of returning to i after leaving it is less than 1.

A state $i \in \mathcal{S}$ is a recurrent state, if the probability of returning to i after leaving it is 1. If the mean time to return to i is finite, then the state is positive recurrent otherwise null recurrent.

All states in a finite closed communicating set are positive recurrent. Every state space $\mathcal{S}$ can be partitioned into closed communicating subsets $\mathcal{C}_I \subset \mathcal{S}\,(I = 1,\ldots,N)$ and the remaining states collected in subset $\mathcal{O} = \mathcal{S} \setminus \{\cup_{I=1}^{N} \mathcal{C}_I\}$. If $\mathcal{O}$ is non-empty, then it contains states i that cannot be entered from a state in one of the closed communicating subsets $\mathcal{C}_I$ but there is a non-zero probability to enter at least one of the subsets $\mathcal{C}_I$ starting in i which implies that all states in $\mathcal{O}$ are transient. This also implies that in a CTMC with a finite state space every state is positive recurrent or transient. The situation is more complex for infinite state spaces because in a closed communicating set of an infinite size, states can be transient, null recurrent or positive recurrent.

If every state in a Markov chain is either absorbing or transient, then the Markov chain is called *absorbing Markov chain*. A particularly interesting case of absorbing Markov chains is one consisting of a single absorbing state. These chains will be central in this book.

We assume here that the state space $\mathcal{S}$ of the continuous time absorbing Markov process $\{X(t)\}_{t\geq 0}^{\infty}$ is finite and contains the set of transient states $\mathcal{S}_T = \{1,\ldots,n\}$ and a single absorbing state $n+1$. We order the states of the CTMC such that the n transient states occur first and can write the infinitesimal generator matrix $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{array}{c} \overbrace{\phantom{\mathbf{D}_0}}^{n} \quad \overbrace{\phantom{\mathbf{d}_1}}^{1} \\ \left[\begin{array}{c|c} \mathbf{D}_0 & \mathbf{d}_1 \\ \hline \mathbf{0} & 0 \end{array}\right] \end{array} \begin{array}{l} \left.\phantom{\begin{array}{c} a \\ a \end{array}}\right\} n \\ \left.\phantom{\begin{array}{c} a \\ a \end{array}}\right\} 1 \end{array} \tag{2.7}$$

Combining all transient states together we obtain a $n \times n$ submatrix $\mathbf{D}_0$ describing only transitions between transient states. The $n \times 1$ vector $\mathbf{d}_1$ contains transition intensities from transient states to the absorbing state. The row vector $\mathbf{0}$ consists entirely of 0's since no transition from the absorbing state to transient states can occur. The remaining element of the matrix $\mathbf{Q}$ is 0 and gives the transition rate out off the absorbing state. Consider the absorbing chain with transition matrix $\mathbf{Q}$ in Fig. 2.2. State 4 is absorbing, hence the transition rates to other states are 0. The regions of the matrix $\mathbf{Q}$, namely $\mathbf{D}_0$, $\mathbf{d}_1$, and $\mathbf{0}$ are marked.

Since the states are transient, matrix $\mathbf{D}_0$ is nonsingular [103, Theorem 2.4.3], i.e. $\lim_{t\to\infty} Prob(X(t) < n+1) = 0$. The stated theorem establishes that the absorption occurs with probability 1. Furthermore, matrix $(-\mathbf{D}_0)^{-1}$ is the fundamental matrix of the absorbing continuous time Markov chain as defined in [93]. The value $(-\mathbf{D}_0)^{-1}(i,j)$ gives the expected total time spent in state j before absorption given that the initial state is i.

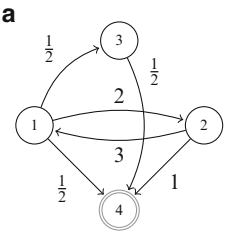


Fig. 2.2 An absorbing CTMC with absorbing state 4. Hence $\mathbf{D}_0$ is 3×3 and $\mathbf{d}_1$ is 3×1 in this example. (**a**) The state transition diagram for the absorbing CTMC. (**b**) The matrix for the absorbing CTMC

Now we are able to introduce the concept of the phase-type distribution (PHD) along the lines of Neuts [124].

Definition 2.4. A phase-type distribution (PHD) is defined as the distribution of the lifetime X, i.e., the time to enter an absorbing state from the set of transient states $\mathcal{S}_T$ of an absorbing continuous time Markov process $\{X(t)\}_{t\geq 0}^{\infty}$.

The finite state space $\mathcal{S}$ of the continuous time absorbing Markov process $\{X(t)\}_{t\geq 0}^{\infty}$ contains the set of transient states $\mathcal{S}_T = \{1,\ldots,n\}$ and the set of absorbing states $\mathcal{S}_A = \{n+1\}$. The transient states are called *phases*. A PHD with n transient states is said to have *order* n. Furthermore the background CTMC $\{X(t)\}_{t\geq 0}^{\infty}$ has an initial probability vector $[\boldsymbol{\pi}, \boldsymbol{\pi}(n+1)]$ and the infinitesimal generator $\mathbf{Q}$ given in Eq. (2.7). In particular, $\boldsymbol{\pi}$ is a row vector of size n, $\boldsymbol{\pi}(n+1)$ is the initial probability for the absorbing state $n+1$. Note that $\mathbf{Q}$ is the intensity matrix thus indicating that the rows sum to zero as shown in Eq. (2.5). Therefore,

$$\mathbf{D}_0\mathbf{1} + \mathbf{d}_1 = \mathbf{0}, \tag{2.8}$$

where $\mathbf{1}$ is the column n-vector of 1's and $\mathbf{0}$ is the column n-vector of 0's. The sum 2.8 in matrix notation can be rewritten as $\mathbf{d}_1 = -\mathbf{D}_0\mathbf{1}$. In particular, the matrix $\mathbf{D}_0$ is a subintensity and we have that

$$\mathbf{D}_0(i,i) \leq 0,\ \mathbf{D}_0(i,j) \geq 0 \text{ for } i \neq j,\ \mathbf{d}_1(i) \geq 0 \text{ and } \sum_{j\in\mathcal{S}_T} \mathbf{D}_0(i,j) \leq 0. \tag{2.9}$$

The Markov process starts in an arbitrary state from $\mathcal{S} = \mathcal{S}_T \cup \mathcal{S}_A$. The vector $\boldsymbol{\pi} = [\pi(1),\ldots,\pi(n)]$ describes the initial probabilities for transient states and $\pi(n+1)$ gives the probability for a direct start in the absorbing state, called point mass at zero. Thus, it holds that $\boldsymbol{\pi}\mathbf{1} + \pi(n+1) = 1$. In most cases it is assumed that $\pi(n+1) = 0$ and there is no start in the absorbing state $n+1$. Then, $\boldsymbol{\pi}\mathbf{1} = 1$, so that X is strictly positive, which we will assume in the following.

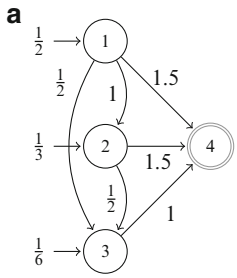


Fig. 2.3 Symbolic representation of the PHD of Example 2.1. (**a**) A state transition diagram of a PHD. (**b**) The sub generator matrix $\mathbf{D}_0$, the exit-rate vector $\mathbf{d}_1$ and the initial probabilities $\boldsymbol{\pi}$ of the PHD

Since the background Markov process $\{X(t)\}_{t\geq 0}^{\infty}$ is a CTMC, the holding time of each phase i, $1 \leq i \leq n$, is exponentially distributed with parameter $-\mathbf{D}_0(i,i)$ as explained in Eq. (2.3). We obtain for diagonal elements of the subintensity $\mathbf{D}_0$

$$\mathbf{D}_0(i,i) = -\left(\sum_{j\neq i} \mathbf{D}_0(i,j) + \mathbf{d}_1(i)\right). \tag{2.10}$$

The column vector $\mathbf{d}_1$ has the interpretation of the exit rates, i.e. $\mathbf{d}_1(i)$ gives the intensity in state i for leaving $\mathcal{S}_T$ and going to the absorbing state $n+1$.

We now say that the random variable X describing the time till absorption is of phase-type with representation $(\boldsymbol{\pi}, \mathbf{D}_0)$. The vector $\mathbf{d}_1$ and the value $\pi(n+1)$ are implicitly given by the matrix $\mathbf{D}_0$ and vector $\boldsymbol{\pi}$, which is the reason for excluding them in the representation.

Example 2.1. Consider a PHD with the subgenerator $\mathbf{D}_0$ given in Fig. 2.3b. Here all states are entry states, since $\pi(i) \neq 0$ is satisfied for all transient states i. Furthermore all transient states i are exit states. This is the case if it is possible to reach an absorbing state directly from the state i.

2.1.3 Analysis of Phase-Type Distributions

We shall now give the basic analytic properties of PHDs. First of all we recall that the matrix exponential $e^{\mathbf{Q}}$ is defined by the standard series expansion $\sum_{k\geq 0} \frac{1}{k!}\mathbf{Q}^k$ (see, e.g. [103]). The distribution function of a phase-type distributed variable with representation $(\boldsymbol{\pi}, \mathbf{D}_0)$ is given by

$$F(x) = 1 - \boldsymbol{\pi} e^{\mathbf{D}_0 x}\mathbf{1} \text{ for } x \geq 0 \tag{2.11}$$

and its associated density function is given by

$$f(x) = \boldsymbol{\pi} e^{\mathbf{D}_0 x} \mathbf{d}_1 \text{ for } x \geq 0. \tag{2.12}$$

Let us assume that a Markov process $\{X(t)\}_{t\geq 0}^{\infty}$ with an infinitesimal generator $\mathbf{Q}$ given in Eq. (2.7) is associated with a random variable X. The transition matrix $\mathbf{P}_t$ contains elements $\mathbf{P}_t(i,j) = Prob(X(t) = j | X(0) = i)$ which is the probability of being in phase j at time t, given that the initial phase is i. These probabilities are given by $\mathbf{P}_t = e^{\mathbf{Q}t}$, where

$$e^{\mathbf{Q}t} = \begin{bmatrix} e^{\mathbf{D}_0 t} & \mathbf{1} - e^{\mathbf{D}_0 t}\mathbf{1} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{2.13}$$

For the distribution of the time till absorption we have that

$$\begin{aligned} F(t) &= Prob(X(t) = n+1) \\ &= \sum_{1\leq i\leq n+1} Prob(X(0) = i)\, Prob(X(t) = n+1 | X(0) = i) \\ &= \sum_{1\leq i\leq n+1} \boldsymbol{\pi}(i) \mathbf{P}_t(i, n+1) \\ &\overset{(2.13)}{=} \boldsymbol{\pi}\mathbf{1} - e^{\mathbf{D}_0 t}\mathbf{1} = 1 - \boldsymbol{\pi} e^{\mathbf{D}_0 t}\mathbf{1}, \end{aligned}$$

holds, which proves Eq. (2.11). Equation 2.12 then follows from $F'(t) = -\boldsymbol{\pi}\frac{d}{dt}\mathbf{P}_t\mathbf{1} = -\boldsymbol{\pi} e^{\mathbf{D}_0 t}\mathbf{D}_0\mathbf{1} = \boldsymbol{\pi} e^{\mathbf{D}_0 t}\mathbf{d}_1$ using $\mathbf{d}_1 = -\mathbf{D}_0\mathbf{1}$.

The expected total time spent in phase j before absorption, given that the initial phase is i equals $-\mathbf{D}_0^{-1}(i,j)$. The ith moment of a PHD is derived from the moment matrix $\mathbf{M} = -\mathbf{D}_0^{-1}$:

$$\mu_i = E[X^i] = i!\boldsymbol{\pi}\mathbf{M}^i\mathbf{1}. \tag{2.14}$$

The event rate is obtained as

$$\lambda = \frac{1}{E[X]} = \frac{1}{\boldsymbol{\pi}\mathbf{M}\mathbf{1}}. \tag{2.15}$$

The squared coefficient of variation equals

$$C^2 = \frac{E[X^2]}{(E[X])^2} - 1 = \frac{2\boldsymbol{\pi}(\mathbf{M})^2\mathbf{1}}{(\boldsymbol{\pi}\mathbf{M}\mathbf{1})^2} - 1. \tag{2.16}$$

Continuous PHDs are dense in the class of distributions on $\mathbb{R}_{\geq 0}$. That is, any distribution with a strictly positive density in $(0,\infty)$ can be approximated arbitrarily close by a PHD (see, e.g. [132]).

2.2 Similarity and Equivalence

It is well known that the matrix representation of a PHD is not unique. Since a PHD of order n is determined by at most $2n-1$ independent parameters [130, 132], but the matrix representation $(\boldsymbol{\pi}, \mathbf{D}_0)$ has n^2+n-1 parameters,[1] this representation is highly redundant, implying that in general, a PH distribution has infinitely many representations with the same order and identical cdf (distribution of the time till absorption) but with different representations of matrix $\mathbf{D}_0$ and vector $\boldsymbol{\pi}$ [129]. Moreover, PHDs $(\boldsymbol{\pi}', \mathbf{D}_0')$ with order $m \neq n$ exist that describe the same distribution as $(\boldsymbol{\pi}, \mathbf{D}_0)$.

Example 2.2. Consider the three PHDs PH_A, PH_B and PH_C defined as

$$\boldsymbol{\pi}^{(A)} = (0.3, 0.3, 0.4) \qquad \mathbf{D}_0^{(A)} = \begin{bmatrix} -4 & 1 & 2 \\ 1 & -4 & 2 \\ 0 & 0 & -1 \end{bmatrix},$$

$$\boldsymbol{\pi}^{(B)} = (0.21, 0.39, 0.4) \qquad \mathbf{D}_0^{(B)} = \begin{bmatrix} -4.3 & 1.3 & 2 \\ 0.7 & -3.7 & 2 \\ 0 & 0 & -1 \end{bmatrix},$$

$$\boldsymbol{\pi}^{(C)} = (0.6, 0.4) \qquad \mathbf{D}_0^{(C)} = \begin{bmatrix} -3 & 2 \\ 0 & -1 \end{bmatrix}.$$

All three representations describe the same PHD. We will continue this example throughout this section and show how one representation can be transformed into another.

Depending on the purpose the PHD is used for, different representations are favorable. If the parameters of a PHD should be estimated (this will be treated in Chap. 3) one is usually interested in canonical forms with only the minimal numbers of parameters required to describe the PHD. Canonical forms are only known for a subclass of PHDs and will be introduced in Sect. 2.3.3. As we will see later (cf. Chap. 5) several approaches for estimating the parameters of a Markovian Arrival Process (MAP) start with a PHD that is expanded into a MAP and the representation of this PHD has a large impact on the flexibility and quality of the estimation approach for the MAP. When the PHD is part of a larger model that should be analyzed numerically, the state space explosion which makes models intractable because of their size, becomes a major problem. In these cases, one is

[1] As mentioned in Sect. 2.1.2 we assume that the point mass at zero, i.e., the probability of starting in the absorbing state is 0. If the absorbing state may have an initial probability greater than zero the number of independent parameters increases to $2n$ and the matrix representation has n^2+n parameters.

usually interested in finding a smaller representation of the PHD to diminish the problem. From these examples it is obvious, that there is a need to transform the representation of a PHD into an equivalent representation (either of the same or a smaller order) to be able to find the most adequate representation for a specific application. Several approaches have been proposed for this task. In Sect. 2.2.1 we will give a brief overview on transformations that preserve the order of the PHD, while Sect. 2.2.2 deals with techniques to reduce the number of states. The transformations presented here do not make any assumptions on the structure of $(\boldsymbol{\pi}, \mathbf{D}_0)$ and can be applied to any PHD. For certain subclasses with specific restrictions on $(\boldsymbol{\pi}, \mathbf{D}_0)$ (in particular acyclic PHDs with an upper triangular matrix $\mathbf{D}_0$) further transformations exist and will be treated in Sect. 2.3.3.

2.2.1 Similarity Transformations

From linear algebra it is well known that two matrices $\mathbf{C}$ and $\mathbf{D}$ are similar if there exists a non-singular matrix $\mathbf{B}$ such that $\mathbf{C} = \mathbf{BDB}^{-1}$ (or equivalently $\mathbf{CB} = \mathbf{BD}$) [119]. Similar matrices share many properties like, e.g. eigenvalues or the characteristic polynomial. This concept of similarity can be used to define equivalence for PH distributions [129, 155].

Let $(\boldsymbol{\pi}, \mathbf{D}_0)$ and $(\boldsymbol{\pi}', \mathbf{D}_0')$ be two PHDs of the same order with cumulative distribution functions $F(x) = 1 - \boldsymbol{\pi} e^{\mathbf{D}_0 x}\mathbf{1}$ and $G(x) = 1 - \boldsymbol{\pi}' e^{\mathbf{D}_0' x}\mathbf{1}$, respectively. Let $\mathbf{B}$ be a non-singular matrix with the additional requirement $\mathbf{B}^{-1}\mathbf{1} = \mathbf{1}$. It was shown in [155] that the two PHDs are equivalent if $\boldsymbol{\pi}' = \boldsymbol{\pi}\mathbf{B}$ and $\mathbf{D}_0' = \mathbf{B}^{-1}\mathbf{D}_0\mathbf{B}$ holds for a matrix $\mathbf{B}$ that fulfills the requirements above. The equivalence follows immediately from

$$G(x) = 1 - \boldsymbol{\pi}' e^{\mathbf{D}_0' x}\mathbf{1} = 1 - \boldsymbol{\pi}\mathbf{B} e^{\mathbf{B}^{-1}\mathbf{D}_0\mathbf{B}x}\mathbf{B}^{-1}\mathbf{1} = 1 - \boldsymbol{\pi}\mathbf{B}\mathbf{B}^{-1} e^{\mathbf{D}_0 x}\mathbf{B}\mathbf{B}^{-1}\mathbf{1} = 1 - \boldsymbol{\pi} e^{\mathbf{D}_0 x}\mathbf{1} = F(x).$$

It is obvious that for a given PHD $(\boldsymbol{\pi}, \mathbf{D}_0)$ the similarity transformation described above does not result in a valid PHD $(\boldsymbol{\pi}', \mathbf{D}_0')$ for all matrices $\mathbf{B}$. However, if two PHDs are given, it is easy to verify if the distributions are equivalent by finding a matrix $\mathbf{B}$ that transforms one representation into the other.

Example 2.3. We will use the similarity transformation with a matrix $\mathbf{B}$ to show that PH_A and PH_B describe the same distribution, i.e. $\boldsymbol{\pi}^{(B)} = \boldsymbol{\pi}^{(A)}\mathbf{B}$ and $\mathbf{D}_0^{(B)} = \mathbf{B}^{-1}\mathbf{D}_0^{(A)}\mathbf{B}$ has to hold. We can solve the equations and obtain

$$\mathbf{B} = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which proves that PH_A and PH_B are indeed equivalent.

2.2.2 Lumping and General Equivalence

Lumping is a technique to reduce the size of state space of a Markov chain [29]. The basic idea is to divide the state space into partitions and to represent each partition by a single state. More formally, we divide the state space $\mathcal{S} = \{1, \cdots, n\}$ into partitions $\Omega = \{\Omega_1, \cdots, \Omega_N\}$ such that

$$\Omega_I \subseteq \mathcal{S}, \qquad \Omega_I \neq \emptyset, \qquad \Omega_I \cap \Omega_J = \emptyset, \qquad \cup_{I=1}^{N} \Omega_I = \mathcal{S}.$$

If we assume that the states belonging to each Ω_I are grouped together we may write the generator matrix $\mathbf{Q}$ of the CTMC as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{1,1} & \cdots & \mathbf{Q}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{Q}_{N,1} & \cdots & \mathbf{Q}_{N,N} \end{bmatrix}$$

where submatrix $\mathbf{Q}_{I,J}$ contains the transitions from Ω_I to Ω_J. Additionally, $\boldsymbol{\pi}_I$ is a subvector of $\boldsymbol{\pi}$ that contains the initial probabilities of all states belonging to Ω_I. The partition Ω can be used to construct an aggregated Markov chain by substituting each partition group Ω_I by a single state. Let $q_{i,\Omega_J} = \sum_{j \in \Omega_J} \mathbf{Q}(i,j)$ be the sum of the transition rates from state i into any of the states from Ω_J. If for every pair of partition groups Ω_I and Ω_J, q_{i,Ω_J} is equal for all $i \in \Omega_I$ the Markov chain is lumpable [93] and we may construct the aggregated Markov chain as follows. The matrix $\hat{\mathbf{Q}}$ of the aggregated chain is constructed using a collector matrix $\mathbf{V}$ and a distributor matrix $\mathbf{W}$, i.e. $\hat{\mathbf{Q}} = \mathbf{WQV}$, where $\mathbf{V}$ is a $n \times N$ matrix with $\mathbf{V}(i,I) = 1$ if $i \in \Omega_I$ and 0 otherwise and $\mathbf{W}$ is a non-negative $N \times n$ matrix with unit row sums that contains in row i equal probabilities for states in Ω_I and 0 elsewhere. It is easy to show that $\mathbf{WV} = \mathbf{I}$ holds.

Example 2.4. Consider PHD PH_A with $\boldsymbol{\pi}^{(A)} = (0.3, 0.3, 0.4)$ and

$$\mathbf{D}_0^{(A)} = \begin{bmatrix} -4 & 1 & 2 \\ 1 & -4 & 2 \\ 0 & 0 & -1 \end{bmatrix}.$$

The complete generator matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \begin{bmatrix} -4 & 1 & 2 & 1 \\ 1 & -4 & 2 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We partition the states into three groups, i.e. $\Omega_1 = \{1,2\}, \Omega_2 = \{3\}$ and $\Omega_3 = \{4\}$ contains the absorbing state. It is easy to verify that the q_{i,Ω_J} are equal for the partition groups. The matrices $\mathbf{V}$ and $\mathbf{W}$ are then given by

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The generator matrix of the aggregated chain is obtained from

$$\hat{\mathbf{Q}} = \mathbf{WQV} = \begin{bmatrix} -3 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

which is the generator matrix of PH_C. The initial probability vector of the aggregate can be obtained by adding all initial probabilities of the states in each Ω_I, which is the same as computing $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}\mathbf{V}$.

In most cases PHDs are not lumpable at all or only very few states can be saved by this aggregation approach.

Note, that from an algebraic point of view the similarity transformation from the previous section and lumping are defined in a similar way, i.e. in both cases the (sub-)generator matrix of the PHD is multiplied by two matrices with row sums equal to one.

Consequently, the most general definition of equivalence between PHDs that has been proposed in [36–38] generalizes the similarity transformation from [155] and lumpability and introduces a description of equivalence between PHDs of the same and different orders. Let $(\boldsymbol{\pi}, \mathbf{D}_0)$ and $(\boldsymbol{\pi}', \mathbf{D}_0')$ be two PHDs of order m and n ($m \geq n$), respectively. Then the two representations are equivalent if there exists a $m \times n$ matrix $\mathbf{V}$ such that $\mathbf{V1} = \mathbf{1}$, $\mathbf{D}_0\mathbf{V} = \mathbf{VD}_0'$ and $\boldsymbol{\pi}\mathbf{V} = \boldsymbol{\pi}'$. The same holds for a $n \times m$ matrix $\mathbf{W}$ with $\mathbf{W1} = \mathbf{1}$, $\mathbf{WD}_0 = \mathbf{D}_0'\mathbf{W}$ and $\boldsymbol{\pi} = \mathbf{W}\boldsymbol{\pi}'$. This definition is not only valid for PHDs but also holds for ME distributions [111], which we will not cover throughout this work.

2.3 Acyclic Phase-Type Distributions

We introduce some basic properties of acyclic PHDs.

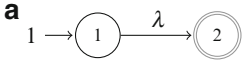
Definition 2.5. If the transition rate matrix $\mathbf{D}_0$ can be transformed into an upper (or lower) triangular matrix by symmetric permutations of rows and columns the PHD is called an acyclic phase-type distribution (APHD).

Since the matrix $\mathbf{D}_0$ is of an upper triangular form, phase i can only be connected with phase j if $j > i$. Consequently, each transient phase is visited at most once before absorption. The matrix representation $(\boldsymbol{\pi}, \mathbf{D}_0)$ has $(n^2 + n)/2$ parameters for the upper triangular matrix and $n - 1$ free parameters for the initial distribution vector. APHDs are the largest subclass of PHDs for which canonical representations exist (cf. Sect. 2.3.3).

APHDs can be divided into various subclasses depending on the structure of $\mathbf{D}_0$ and $\boldsymbol{\pi}$. In the following we will give an overview of these distributions.

2.3.1 Erlang and Hyper-Erlang Distributions

Since PHDs can be considered as a natural generalization of the exponential and Erlang distributions we introduce them in this section. The exponential distribution is completely characterized by its rate parameter λ and is the simplest case of a PHD with a single transient state as shown in Fig. 2.4a. By the fact that $\boldsymbol{\pi} = [1]$, the only transient phase is visited before absorption. The corresponding infinitesimal generator matrix includes $\mathbf{D}_0 = [-\lambda]$ and the exit vector $\mathbf{d}_1 = [\lambda]$ as shown in Eq. (2.17) in Fig. 2.4b.



(2.17)

Fig. 2.4 Markovian representation of the exponential distribution. (**a**) An exponential distribution with parameter λ, and 2 being an absorbing state. (**b**) The infinitesimal generator $\mathbf{Q}$

The exponential distribution has density

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0, \tag{2.18}$$

and its distribution function is given by

$$F(x) = 1 - e^{-\lambda x} \text{ for } x \geq 0. \tag{2.19}$$

Foremost it is the only continuous distribution with the memoryless property

$$Prob(X > t + s | X > t) = Prob(X > s) \text{ for all } t, s \geq 0.$$

The mean is given by $E[X] = \frac{1}{\lambda}$ and the variance is $VAR[X] = \frac{1}{\lambda^2}$.

Erlang introduced in [53] the representation of distributions as a sum of n exponential phases with the same intensity λ. Consider n mutually independent, exponentially distributed random variables X_i with parameter $\lambda > 0$, $1 \leq i \leq n$. If we define a random variable Y as $Y = \sum_{1 \leq i \leq n} X_i$, then it has an Erlang distribution denoted by $E(n, \lambda)$, and its density is given by

$$f(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \text{ for } x \geq 0. \tag{2.20}$$

The distribution function is defined by

$$F(x) = 1 - \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} e^{-\lambda x} \text{ for } x \geq 0. \tag{2.21}$$

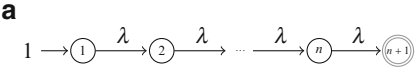
The ith moment of the Erlang distributed random variable Y is given by

$$E[Y^i] = \frac{(n+i-1)!}{(n-1)!} \frac{1}{\lambda^i}. \tag{2.22}$$

Thus, the mean of Y is $E[Y] = \frac{n}{\lambda}$ and the variance equals $VAR[Y] = \frac{n}{\lambda^2}$.

The underlying Markov process can be described by the infinitesimal generator matrix given in Eq. (2.23), where the initial probability vector is $\boldsymbol{\pi} = [1, 0, \ldots, 0]$ as shown in Fig. 2.5a.

a



b

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda & \lambda & \ldots & 0 & 0 \\ 0 & -\lambda & \ldots & 0 & 0 \\ & & \ddots & & \\ \ldots & \ldots & & \ldots & \ldots \\ 0 & 0 & \ldots & -\lambda & \lambda \\ 0 & 0 & \ldots & 0 & -\lambda \end{bmatrix} \tag{2.23}$$

Fig. 2.5 Erlang representation of a PHD. (**a**) A graphical representation of the Erlang(n,λ) PHD. (**b**) The sub generator $\mathbf{D}_0$

The Markov process must start in phase 1 and traverses through the successive states until it reaches the absorbing state $n+1$. Thus, the time to absorption described by Y is the summation of all holding times which are identically exponentially distributed with parameter λ. The Erlang distribution $E(n, \lambda)$ has a squared coefficient of variation of $C^2 = n^{-1}$ which is less than one for $n > 1$. Distributions with a coefficient of variation greater than one can be modeled as finite mixtures of exponential distributions.

Remark 2.1. The Erlang distribution with mean m has variance $\frac{m^2}{n}$ and thus tends to be deterministic for $n \to \infty$. In this case the squared coefficient of variation is close to zero. This coefficient is used to express the variance of the random variable relative to its average value. Consequently, the Erlang distribution can be used to approximate deterministic distributions.

As a next example we consider the hypo-exponential distribution which is a generalized Erlang distribution. Consider a set of exponential distributions $F_i(\cdot)$ with

$$F_i(x) = 1 - e^{-\lambda(i)x} \text{ for } x \geq 0,\ 1 \leq i \leq n,$$

and where rates $\lambda(1),\ldots,\lambda(n)$ are not necessarily identical. Consequently we have

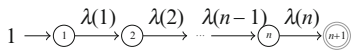
$$f_i(x) = \lambda(i)e^{-\lambda(i)x} \text{ for } x \geq 0.$$

The hypo-exponential distribution is then characterized by the number n and the set of parameters $\lambda(i)$. Its density function is given by

$$f(x) = \sum_{i=1}^{n}\left(\prod_{j=1,j\neq i}^{n}\frac{\lambda(j)}{\lambda(j)-\lambda(i)}\right)f_i(x) \text{ for } x \geq 0,\ \lambda(i) \neq \lambda(j) \text{ for } i \neq j. \tag{2.24}$$

We obtain the mean of the hypo-exponentially distributed random variable as $E[X] = \sum_{i=1}^{n}\frac{1}{\lambda(i)}$ and the variance as $VAR[X] = \sum_{i=1}^{n}\frac{1}{\lambda(i)^2}$. If all parameters $\lambda(i)$ are equal, we obtain the Erlang distribution since it is the convolution of n identical exponentials. The sub generator of the hypo-exponential distribution is shown in Eq. (2.25) and Fig. 2.6b. The initial distribution vector is $\boldsymbol{\pi} = [1,0,\ldots,0]$ as shown in Fig. 2.6a.

a



b

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda(1) & \lambda(1) & \ldots & 0 & 0 \\ 0 & -\lambda(2) & \ldots & 0 & 0 \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ 0 & 0 & \ldots & -\lambda(n-1) & \lambda(n-1) \\ 0 & 0 & \ldots & 0 & -\lambda(n) \end{bmatrix} \tag{2.25}$$

Fig. 2.6 The hypo-exponential distribution. (**a**) Markovian representation of the hypo-exponential distribution. In particular $\lambda(1),\ldots,\lambda(n)$ are not necessarily identical. (**b**) The sub generator $\mathbf{D}_0$

The hyper-exponential distribution is a convex mixture of n exponential distributions and is visualized in Fig. 2.7. The density is given by

$$f(x) = \sum_{i=1}^{n}\pi(i)\lambda(i)e^{-\lambda(i)x} \text{ for } x \geq 0, \tag{2.26}$$

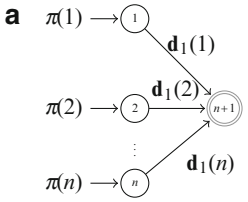
where $\pi(i) > 0$ for all phases i and $\sum_{i=1}^{n}\pi(i) = 1$. The distribution function of a hyper-exponentially distributed random variable X is defined as

$$F(x) = \sum_{i=1}^{n}\pi(i)(1-e^{-\lambda(i)x}) \text{ for } x \geq 0. \tag{2.27}$$

The first moment is obtained as $E[X] = \sum_{i=1}^{n}\frac{\pi(i)}{\lambda(i)}$ and its variance is given by

$$VAR[X] = 2\sum_{i=1}^{n}\frac{\pi(i)}{\lambda(i)^2} - \left(\sum_{i=1}^{n}\frac{\pi(i)}{\lambda(i)}\right)^2. \tag{2.28}$$

The intensity matrix of the hyper-exponential distribution has a form shown in Eq. (2.29) shown in Fig. 2.7b, and the initial probability vector is $\boldsymbol{\pi} = [\pi(1), \pi(2), \ldots, \pi(n)]$. A graphical representation of the corresponding Markovian process is



$$\mathbf{D}_0 = \begin{bmatrix} -\lambda(1) & 0 & \ldots & 0 & 0 \\ 0 & -\lambda(2) & \ldots & 0 & 0 \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ 0 & 0 & \ldots & -\lambda(n-1) & 0 \\ 0 & 0 & \ldots & 0 & -\lambda(n) \end{bmatrix} \tag{2.29}$$

Fig. 2.7 The hyper-exponential distribution. (**a**) A graphical representation of the hyper-exponential distribution. (**b**) The sub generator $\mathbf{D}_0$

visualized in Fig. 2.7a. By the fact, that the Markov process can start in each phase an additional dispersion is introduced which leads to the squared coefficient of variation greater than one; and equal to one for the exponential case $n = 1$.

$$C^2 = \frac{E[Y^2]}{(E[Y])^2} - 1 = 2\frac{\sum_{i=1}^{n} \dfrac{\pi(i)}{\lambda(i)^2}}{\left(\sum_{i=1}^{n} \dfrac{\pi(i)}{\lambda(i)}\right)^2} - 1 \tag{2.30}$$

A hyper-Erlang distribution denoted as HErD [57], is a mixture of m mutually independent Erlang distributions weighted with the initial probabilities $\pi(1), \ldots, \pi(m)$, where $\pi(i) \geq 0$, and the vector $\boldsymbol{\pi}$ is stochastic, i.e. $\sum_{i=1}^{m} \pi(i) = 1$. Let s_i denote the number of phases of the ith Erlang distribution. Then the density is

$$f(x) = \sum_{i=1}^{m} \pi(i) \frac{(\lambda(i)x)^{s_i - 1}}{(s_i - 1)!} \lambda(i) e^{-\lambda(i)x} \text{ for } x \geq 0, \tag{2.31}$$

and its distribution function is given by

$$F(x) = 1 - \sum_{i=1}^{m} \pi(i) \sum_{j=0}^{s_i - 1} \frac{(\lambda(i)x)^j}{j!} e^{-\lambda(i)x} \text{ for } x \geq 0. \tag{2.32}$$

The ith moment can be obtained as

$$E[Y^i] = \sum_{j=1}^{m} \pi(i) \frac{(s_j + i - 1)!}{(s_j - 1)!} \frac{1}{\lambda(j)^i}. \tag{2.33}$$

The state space of HErD consists of $\sum_{i=1}^{m} s_i$ transient and one absorbing state. For $m = 1$ we have a single Erlang distribution $E(s_1, \lambda(1))$ and the case that

$s_i = 1$ for all $1 \leq i \leq m$ represents a hyper-exponential distribution. The underlying Markov chain of the HErD can be described by the infinitesimal generator given in Eq. (2.34) as shown in Fig. 2.8b, which has the matrices $\mathbf{Q}_i$ on its diagonal. Matrix $\mathbf{Q}_i$ represents the infinitesimal generator of the ith Erlang branch and its form is given in Eq. (2.23). The initial distribution vector is $\boldsymbol{\pi} = [\pi(1), 0, \ldots, 0, \pi(2), 0, \ldots, \pi(m), 0, \ldots, 0]$.

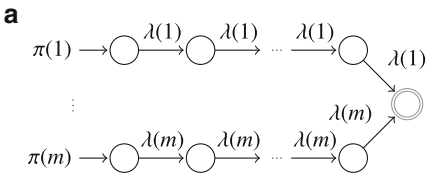


Fig. 2.8 Symbolic representation of the HErD. (**a**) A graphical representation of the HErD. (**b**) The infinitesimal generator $\mathbf{Q}$

Recall that an Erlang distribution with s phases is defined as the sum of s independent identically distributed random variables. Therefore, a HErD is constructed from a mixture of sums of exponential distributions.

2.3.2 Coxian Distributions

Coxian distributions can be considered as a mixture of hypo- and hyper-exponential distributions [45, 157]. The initial distribution vector is given by $\boldsymbol{\pi} = [1, 0, \ldots, 0]$. After starting in phase 1 the process traverses through the n successive phases with possibly different rates $\lambda(i)$. From the phase i the transition to the next phase $i+1$ can occur with probability g_i or the absorbing state is reached with the complementary probability $1 - g_i$.

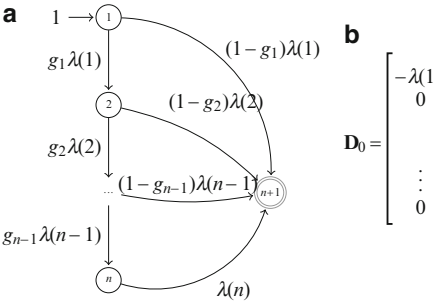


Fig. 2.9 Symbolical representation of the Coxian distribution. (**a**) A graphical representation of the Coxian distribution. (**b**) The sub generator $\mathbf{D}_0$

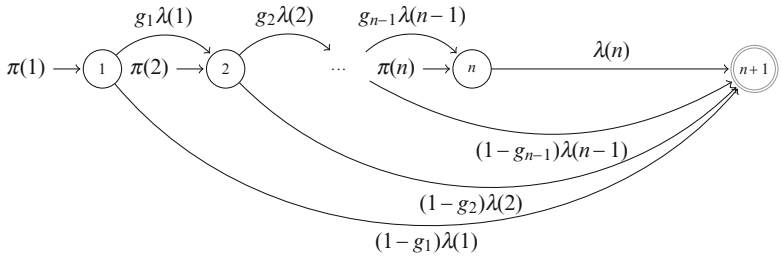


Fig. 2.10 The generalized Coxian PHD

The matrix representation of the Coxian distribution is given by Eq. (2.35) presented in Fig. 2.9b. A generalized Coxian distribution extends Coxian distribution with a random initial vector $\boldsymbol{\pi}$, so that each state can be an entry state. Thus exhibiting low or high variability. A CTMC representation of the Coxian PHD is visualized in Fig. 2.9a, and of the generalized Coxian PHD in Fig. 2.10.

2.3.3 Canonical Representations

We now consider distinct canonical representations of an APHD, i.e. a triangular matrix $\mathbf{D}_0$ with nonzero diagonal elements. The analysis of APHDs in [47] has revealed that the cdf of n-phase APHDs has at most $2n-1$ degrees of freedom. On the other hand, the number of parameters needed to specify an APHD with an upper triangular matrix representation is $(n^2+n)/2$ for the matrix $\mathbf{D}_0$ and $n-1$ parameters for the initial distribution vector. Since the representation of the cdf by a tuple $(\boldsymbol{\pi}, \mathbf{D}_0)$ is highly redundant, minimal and cdf retentive representations have been developed. We will now pose the canonical representations with $2n-1$ free parameters regarding the APHD (see Sect. 2.3). The key idea of the approach from [47] is to express an APHD in terms of its elementary series, i.e. paths from an initial to the absorbing state.

Definition 2.6. For an APHD of order n an elementary series of order $m \leq n$ is defined as a series of the form

$$ES = < \lambda(i_1)\lambda(i_2)\ldots\lambda(i_m) >,$$

where $i_1, i_2, \ldots, i_{m-1}, i_m$ is a sequence of states along the acyclic path from an initial state to the absorbing state such that

$$\mathbf{D}_0(i_k, i_{k+1}) \neq 0,\ k = 1, 2, \ldots, m.$$

and $\mathbf{d}_1(i_m) > 0$, i.e. a transition from the last state of the series i_m to the absorbing state is possible.

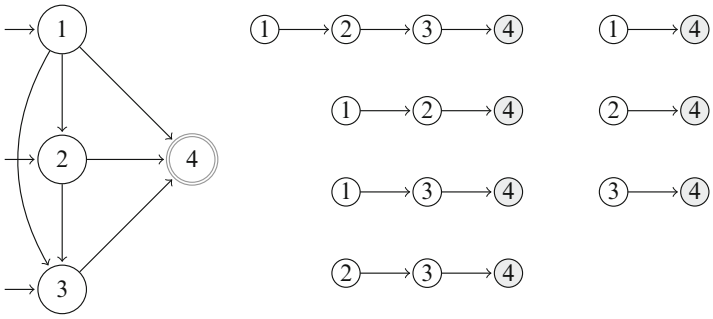


Fig. 2.11 An acyclic 3-phase PHD and its elementary series

The rate between two states of an elementary series is given by the total event rate of the first of the two states, i.e. by the corresponding diagonal entry of $-\mathbf{D}_0$. One can also introduce a path as a sequence of connected states by inspection of the transition diagram corresponding to $(\boldsymbol{\pi}, \mathbf{D}_0)$. The number of possible elementary series in a n-phase APHD is bounded by $2^n - 1$.

Figure 2.11 shows a 3-phase APHD with its elementary series.

Now consider the following identity: Given two positive real numbers λ and μ, with $\lambda \leq \mu$, one obtains

$$\frac{\lambda}{s+\lambda} = \tau \frac{\mu}{s+\mu} + (1-\tau)\frac{\lambda\mu}{(s+\lambda)(s+\mu)}, \tag{2.36}$$

where $\tau = \frac{\lambda}{\mu} \in (0, 1]$ represents the probability for the path till absorption containing only a phase with the transition rate μ. Consequently, $(1-\tau)$ is the remaining probability for the elementary series containing rates λ and μ, i.e. there are successive phases with transition rates λ and μ. Whenever it holds that $\lambda \leq \mu$, we obtain an ascending ordering of the transition rates. Accordingly, we indicate that the cdf of an elementary series has the Laplace transform

$$F(s) = \frac{\lambda(i_1)\lambda(i_2)\ldots\lambda(i_{m-1})}{s(s+\lambda(i_1))(s+\lambda(i_2))\ldots(s+\lambda(i_{m-1}))}, \tag{2.37}$$

e.g. the cdf of the elementary series consisting only of the rate λ has the Laplace transform $F(s) = \frac{\lambda}{s+\lambda}$. With these observations in mind, we can now retain that an elementary series for some phase with transition rate λ can be substituted by a mixture of two elementary series, one containing a phase with transition rate $\mu > \lambda$, and the other containing both phases with the rates λ and μ. This substitution step is visualized in Fig. 2.12 and by repeated use of Eq. (2.36) the elementary series can be transformed to a mixture of basic series, which can be specified now.

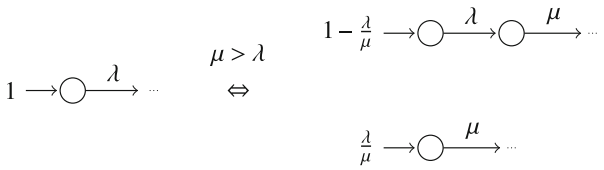


Fig. 2.12 Substitution step for the exponential distribution

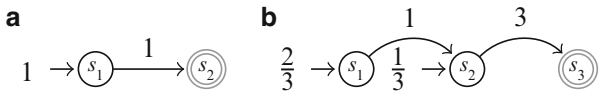


Fig. 2.13 Equivalent representation of the exponential distribution (**a**) Series for an exponential distribution. (**b**) Series for a 2 phase APHD

Example 2.5. The 3-phase PHD shown in Fig. 2.13b has the Laplace transform $F(s) = \frac{2}{3}\frac{1\cdot 3}{(s+1)(s+3)} + \frac{1}{3}\frac{3}{(s+3)} = \frac{2}{(s+1)(s+3)} + \frac{1}{s+3} = \frac{1}{s+1}$. Thus, this PHD is cdf-equivalent to the former PHD shown in Fig. 2.13a since it contains the elementary series $<\mu>$ and $<\lambda\mu>$ with $\lambda = 1$ and $\mu = 3$.

Definition 2.7. For n positive real numbers $0 < \lambda(1) \leq \lambda(2) \leq \ldots \leq \lambda(n)$ the basic series (BS) are defined as the tuples of $i,\ldots,n$ transient states that determine the acyclic path till absorption [47]. Each basic series can be represented by the notation,

$$BS_i =< \lambda(i)\ldots\lambda(n-1)\lambda(n) >.$$

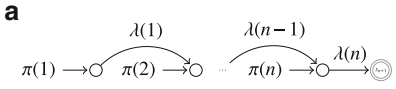
The Markov process can start in a phase with transition rate $\lambda(i)$ and follows the path till absorption. Thus, the exit rate of each path is $\lambda(n)$ and according to the substitution step the canonical ordering of $\lambda(i)$'s is incorporated.

It is now easy to see, that each basic series BS_i describes a hypo-exponential distribution (cf. Eq. (2.25) visualized in Fig. 2.6b) with the corresponding distribution of the time till absorption, i.e., it is the convolution of $(n-i+1)$ exponentials with parameters $\lambda(i),\ldots,\lambda(n-1),\lambda(n)$. The results show that for an APH, the cdf of each of its elementary series is a mixture of the cdfs of its basic series. Furthermore, the mixture of the cdfs of all elementary series determines the distribution of the time till absorption for the underlying APH, where each elementary series is weighted proportional to its probability. The probability of each elementary series is computed as the product of the transition rates along the corresponding path and the initial probability of the first state of the path. Let $i_1, i_2, \ldots, i_m$ be m states from the jth elementary series of a PHD. Then, the probability of the ES_j is given by

$$\tau_j = \pi(i_1)\frac{\mathbf{D}_0(i_1,i_2)}{-\mathbf{D}_0(i_1,i_1)}\frac{\mathbf{D}_0(i_2,i_3)}{-\mathbf{D}_0(i_2,i_2)}\cdots\frac{\mathbf{D}_0(i_{m-1},i_m)}{-\mathbf{D}_0(i_{m-1},i_{m-1})}\frac{\mathbf{d}_1(i_m)}{-\mathbf{D}_0(i_m,i_m)}, \tag{2.38}$$

where the term $\frac{\mathbf{D}_0(i_k,i_{k+1})}{-\mathbf{D}_0(i_k,i_k)}$, for any $k = 1,\ldots,m-1$, represents the transition probability from the state i_k to the state i_{k+1}. Basic series together with appropriate initial probabilities yield the series canonical form.

Definition 2.8. The series canonical form is defined as a mixture of basic series of an APHD with transition rates in ascending order, i.e. $\lambda(i) \leq \lambda(i+1) \leq \ldots \leq \lambda(n)$. Transitions are only possible from phase i to the neighbor phase $i+1$. There is only one exit state, but all states may be entry states satisfying $\boldsymbol{\pi}(i) \geq 0$, for all $i = 1, \ldots, n$. The matrix representation of the series canonical form is given by Eq. (2.39) in Fig. 2.14a.



$$\begin{bmatrix} -\lambda(1) & \lambda(1) & 0 & \ldots & 0 \\ 0 & -\lambda(2) & \lambda(2) & & \\ & & \ddots & \ddots & \vdots \\ \vdots & & & -\lambda(n-1) & \lambda(n-1) \\ 0 & & & & -\lambda(n) \end{bmatrix} \tag{2.39}$$

Fig. 2.14 PHD in series canonical form. (**a**) The series canonical form, in particular $0 < \lambda(1) \leq \lambda(2) \leq \ldots \leq \lambda(n-1) \leq \lambda(n)$. (**b**) The sub generator $\mathbf{D}_0$

The series canonical form is visualized in Fig. 2.14a and has $2n-1$ degrees of freedom, namely n transition rates and $n-1$ independent initial probabilities [47].

For general PHDs canonical forms are only known for the cases $n = 2$ or $n = 3$. For the case $n = 2$ it has been shown that APHDs and general PHDs are equivalent in the sense that every distribution that is represented as a PHD also has a representation as an APHD [45]. For $n = 3$ a canonical form for general PHDs has been developed in [78] based on earlier results in [66]. The canonical representation is generated by symbolically performing similarity transformations on matrix $\mathbf{D}_0$ and vector $\boldsymbol{\pi}$ using the equivalence of PHDs defined in Sect. 2.2. However, it can be shown that this approach cannot be applied for $n > 3$.

2.4 Properties

Since the class of PHDs is closed under certain operations such as convolutions and finite mixtures, we summarize below some of the basic properties without proof. The presentation is based on results discussed in [103, 127]. Further closure properties of PHDs can be found in [116].

First we consider the sum of two independent random variables of the phase-type. Let $PH_A = (\boldsymbol{\pi}^{(A)}, \mathbf{D}_0^{(A)})$ be of order n, and $PH_B = (\boldsymbol{\pi}^{(B)}, \mathbf{D}_0^{(B)})$ be of order m. Furthermore PH_A is the distribution of the random variable $X^{(A)}$, and PH_B is the distribution of the random variable $X^{(B)}$. Then the sum $X^{(C)} = X^{(A)} + X^{(B)}$ is phase-type distributed. The underlying Markov process can be described by the

sub generator matrix given in Eq. (2.40) where the initial probability vector is $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)}, \boldsymbol{\pi}^{(A)}(n+1)\boldsymbol{\pi}^{(B)}]$.[2]

$$\mathbf{D}_0^{(C)} = \begin{bmatrix} \mathbf{D}_0^{(A)} & \mathbf{d}_1^{(A)}\boldsymbol{\pi}^{(B)} \\ \mathbf{0} & \mathbf{D}_0^{(B)} \end{bmatrix}. \tag{2.40}$$

In the underlying Markov chain $\{X^{(C)}(t)\}_{t\geq 0}^{\infty}$ the paths of the Markov chains associated with PH_A and PH_B are concatenated such that after traversing the paths of the chain $\{X^{(A)}(t)\}_{t\geq 0}^{\infty}$ the process moves along the paths of the chain $\{X^{(B)}(t)\}_{t\geq 0}^{\infty}$. The resulting number of transient phases in $\{X^{(C)}(t)\}_{t\geq 0}^{\infty}$ is $n+m$. The initial phase is selected according to $\boldsymbol{\pi}^{(A)}$. Then the process moves along the paths of the $\{X^{(A)}(t)\}_{t\geq 0}^{\infty}$ until some entry state of the chain $\{X^{(B)}(t)\}_{t\geq 0}^{\infty}$ is reached. By the fact that $\mathbf{d}_1^{(A)}\boldsymbol{\pi}^{(B)}$ is the rate of entering some initial state of the second chain, the initial distribution of that process remains unchanged. Thus, the absorbing time of the constructed Markov process $\{X^{(C)}(t)\}_{t\geq 0}^{\infty}$ is the sum of the absorbing time of the first and the second Markov chains.

As next operation we consider the convex mixture of PHDs. Let again $PH_A = (\boldsymbol{\pi}^{(A)}, \mathbf{D}_0^{(A)})$ be of order n, and $PH_B = (\boldsymbol{\pi}^{(B)}, \mathbf{D}_0^{(B)})$ be of order m. Furthermore let $F^{(A)}(\cdot)$ and $F^{(B)}(\cdot)$ be the corresponding probability distribution functions. Then let the PHD PH_C be a convex mixture of these distribution functions which is defined as $\alpha F^{(A)}(\cdot) + (1-\alpha)F^{(B)}(\cdot)$, for $0 \geq \alpha \leq 1$. The underlying Markov process has $n+m$ transient phases, the initial probability vector equals $\boldsymbol{\pi}^{(C)} = [\alpha\boldsymbol{\pi}^{(A)}, (1-\alpha)\boldsymbol{\pi}^{(B)}]$, and the sub generator matrix is given in Eq. (2.41).

$$\mathbf{D}_0^{(C)} = \begin{bmatrix} \mathbf{D}_0^{(A)} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_0^{(B)} \end{bmatrix}. \tag{2.41}$$

Since the states of PH_A, PH_B are disjoint in the process associated with PH_C, the set of the passed states corresponds either to the PH_A or to the PH_B. Thus, the time until absorption is distributed either according to PH_A with probability α, or it is distributed like PH_B with complementary probability.

All order statistics of a finite number of independent PHD random variables, e.g. the kth smallest of a set of random variables, minima, or maxima, are PHDs [103].

We focus on the distribution of the smallest and the largest of two independent random variables $X^{(A)}$, $X^{(B)}$ of the phase-type. The corresponding representation of the PHDs is $PH_A = (\boldsymbol{\pi}^{(A)}, \mathbf{D}_0^{(A)})$ of order n, and $PH_B = (\boldsymbol{\pi}^{(B)}, \mathbf{D}_0^{(B)})$ of order m, respectively. Then the random variable $X^{(C)} = min(X^{(A)}, X^{(B)})$ has a PHD with representation $(\boldsymbol{\pi}^{(C)}, \mathbf{D}_0^{(C)})$. The initial distribution is defined as $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)} \otimes \boldsymbol{\pi}^{(B)}]$, and the corresponding sub generator matrix is given in Eq. (2.42).

[2] If the case that $\boldsymbol{\pi}^{(A)}(n+1) = 0$, i.e., there is no start in the absorbing state, the random variable $X^{(A)}$ is strictly positive. Then the initial probability vector is given by $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)}, \mathbf{0}]$ where $\mathbf{0}$ is the row m-vector of 0's.

$$\mathbf{D}_0^{(C)} = \mathbf{D}_0^{(A)} \otimes \mathbf{I}^{(B)} + \mathbf{I}^{(A)} \otimes \mathbf{D}_0^{(B)} = \mathbf{D}_0^{(A)} \oplus \mathbf{D}_0^{(B)}, \tag{2.42}$$

where $\mathbf{I}^{(A)}$, $\mathbf{I}^{(B)}$ are identity matrices of order n and m, respectively. $\otimes$ and $\oplus$ denote the Kronecker product and Kronecker sum [49, 112] which are defined for two square matrices $\mathbf{A}$ and $\mathbf{B}$ of order n^a and n^b as

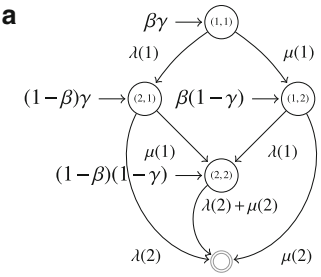
$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}(1,1)\mathbf{B} & \cdots & \mathbf{A}(1,n^a)\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}(n^a,1)\mathbf{B} & \cdots & \mathbf{A}(n^a,n^a)\mathbf{B} \end{bmatrix} \text{ and } \mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}^{n^a} + \mathbf{I}^{n^b} \otimes \mathbf{B}. \tag{2.43}$$

In the resulting two-dimensional Markov process of order nm the state space is given by pairs of phases $\{(i,j) : i \in \mathcal{S}_T^{(A)}, j \in \mathcal{S}_T^{(B)}\}$ and the single absorbing state $nm+1$. The underlying chain models the concurrent behavior of the two original processes $\{X^{(A)}(t)\}_{t\geq 0}^{\infty}$, $\{X^{(B)}(t)\}_{t\geq 0}^{\infty}$. Strictly, the expanded chain has been obtained through sequencing of concurrent state transitions in the original chains. For example, it is possible that either the first chain or the second chain alone evolves through the state space till absorption. In one of that cases only one component in the tuple (i,j) changes. If it gets absorbed, this means that the minimum of both PHDs is determined.

We conclude with results for the largest of the two independent random variables $X^{(A)}$, $X^{(B)}$ of the phase-type. The random variable $X^{(C)} = max(X^{(A)}, X^{(B)})$ has a PHD with representation $(\boldsymbol{\pi}^{(C)}, \mathbf{D}_0^{(C)})$, where the initial probability vector is given by $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)} \otimes \boldsymbol{\pi}^{(B)}, \boldsymbol{\pi}^{(A)}\boldsymbol{\pi}^{(B)}(m+1), \boldsymbol{\pi}^{(A)}(n+1)\boldsymbol{\pi}^{(B)}]$. The sub generator of the associated Markov chain is given in Eq. (2.44)

$$\mathbf{D}_0^{(C)} = \begin{bmatrix} \mathbf{D}_0^{(A)} \oplus \mathbf{D}_0^{(B)} & \mathbf{I}^{(A)} \otimes \mathbf{d}_1^{(B)} & \mathbf{d}_1^{(A)} \otimes \mathbf{I}^{(B)} \\ \mathbf{0} & \mathbf{D}_0^{(A)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0^{(B)} \end{bmatrix}. \tag{2.44}$$

The expanded Markov chain corresponding to the distribution with representation $(\boldsymbol{\pi}^{(C)}, \mathbf{D}_0^{(C)})$ has been formed as follows. Its state space consists of $nm+n+m$ transient states. The submatrix $\mathbf{D}_0^{(A)} \otimes \mathbf{I}^{(B)} + \mathbf{I}^{(A)} \otimes \mathbf{D}_0^{(B)}$ describes the part where the original processes evolve simultaneously until one of them gets absorbed. This means that the minimum of both PHDs is known and we shall retain the absorbed process. Thus additional states $(n+1,\cdot)$ and $(\cdot, m+1)$ corresponds to the absorption of one of the involved processes. Reaching one of the states $(n+1,\cdot)$ or $(\cdot,m+1)$ the chain evolves according to the remaining process which has been not absorbed yet. The state space contains pairs of phases $\{(i,j) \mid i \in \mathcal{S}_T^{(A)},\ j \in \mathcal{S}_T^{(B)}\} \cup \{(n+1,j) \mid n+1 \in \mathcal{S}^{(A)} \setminus \mathcal{S}_T^{(A)},\ j \in \mathcal{S}_T^{(B)}\} \cup \{(i,m+1) \mid , i \in \mathcal{S}_T^{(A)},\ m+1 \in \mathcal{S}^{(B)} \setminus \mathcal{S}_T^{(B)}\}$, and the absorbing state. Hence all combinations of transitions until absorption of both Markov chains are considered.



b

$$\mathbf{D}_0^{(C)} = \begin{bmatrix} -(\lambda(1)+\mu(1)) & \mu(1) & \lambda(1) & 0 \\ \vdots & -(\lambda(1)+\mu(2)) & 0 & \lambda(1) \\ & & -(\lambda(2)+\mu(1)) & \mu(1) \\ 0 & \cdots & & -(\lambda(2)+\mu(2)) \end{bmatrix}$$

Fig. 2.15 PHD of the minimum of two generalized Erlang PHDs. (**a**) Diagrammatic representation of the PHD of the minimum. (**b**) The sub generator $\mathbf{D}_0^{(C)}$, in particular $\boldsymbol{\pi}^{(C)} = [\beta\gamma, \beta(1-\gamma), (1-\beta)\gamma, (1-\beta)(1-\gamma)$

Example 2.6. Let us consider concrete examples of the last two operations. We obtain the random variable $X^{(C)}$ which is the minimum of two generalized Erlang distributed random variables $X^{(A)}$, $X^{(B)}$. The representation of the corresponding PHDs is given by

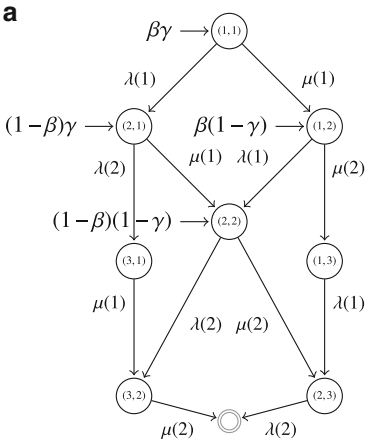
$$\boldsymbol{\pi}^{(A)} = [\beta, 1-\beta],\ \mathbf{D}_0^{(A)} = \begin{bmatrix} -\lambda(1) & \lambda(1) \\ 0 & -\lambda(2) \end{bmatrix}, \qquad \boldsymbol{\pi}^{(B)} = [\gamma, 1-\gamma],\ \mathbf{D}_0^{(B)} = \begin{bmatrix} -\mu(1) & \mu(1) \\ 0 & -\mu(2) \end{bmatrix}.$$

The transition rate matrix of the expanded process representing the distribution of the minimum is given in Fig. 2.15.

Now the case where the random variable $X^{(D)}$ is the maximum of both defined random variables $X^{(A)}$, $X^{(B)}$ is treated. The PHD of the process describing the $max(X^{(A)}, X^{(B)})$ is given in Fig. 2.16. Note that the submatrix $\mathbf{D}_0^{(C)}$ of $\mathbf{D}_0^{(D)}$ is defined in Fig. 2.15b. In particular, the initial distribution vector is given by $\boldsymbol{\pi}^{(D)} = [\beta\gamma, \beta(1-\gamma), (1-\beta)\gamma, (1-\beta)(1-\gamma), 0, 0, 0, 0]$.

2.5 Concluding Remarks

PHDs have a long history dating back to the early work of Erlang [53]. Since then numerous papers have appeared on the subject such that any introduction of the topic must be incomplete which has already been mentioned by Neuts who gave an introduction of PHDs in his famous book on matrix geometric solutions [125]. Like Neuts we tried to introduce PHDs in a form that allows one to use them in computational methods. Later, in Sect. 6, we present models where PHDs are used as building blocks to model event times, and we also outline how these models are analyzed numerically or by simulation. The representation of PHDs using matrices and vectors is also useful for computational algorithms for the parameterization of PHDs which are presented in the following chapter.



b

$$\mathbf{D}_0^{(D)} = \begin{bmatrix} \mathbf{D}_0^{(C)} & \mathbf{I}^{(A)} \otimes \mathbf{d}_1^{(B)} & \mathbf{I}^{(B)} \otimes \mathbf{d}_1^{(A)} \\ \mathbf{0} & \mathbf{D}_0^{(A)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0^{(B)} \end{bmatrix}$$

Fig. 2.16 PHD of the maximum of two generalized Erlang PHDs. (**a**) Diagrammatic representation of the PHD of the maximum. (**b**) The sub generator $\mathbf{D}_0^{(D)}$

Several additional aspects of PHDs are handled in the literature. The eigenvalues of the matrix $\mathbf{D}_0$ and the Laplace transform play an important role in the characterization of PHDs and in the finding of canonical and minimal representations. Fundamental work of Cox [45] and many successors [66, 121, 129, 131] present results in this directions. However, a canonical representation for general PHDs is missing and also the question whether a given PHD has a representation with less states is still unanswered.

We consider here only PHDs in continuous time. As already mentioned, it is obvious that also discrete time PHDs based on DTMCs rather than CTMCs can be defined. Several results can be transferred from the continuous time to the discrete time area, but there are also some specific aspects that need to be considered, details can be found in [20, 103]. Additionally, one can neglect the probabilistic interpretation and interpret the vector matrix pair describing a PHD in a linear algebraic context. The resulting distributions are denoted as matrix exponential distributions [54, 67, 111] but will not be considered here since the corresponding theory and the practical applicability is less advanced.

http://www.springer.com/978-3-319-06673-8

Input Modeling with Phase-Type Distributions and Markov Models
Theory and Applications
Buchholz, P.; Kriege, J.; Felko, I.
2014, XII, 127 p. 42 illus., 35 illus. in color., Softcover
ISBN: 978-3-319-06673-8