Chapter 5
Markov Chains with Special Structures

The previous chapter presented methods for analyzing stochastic models where some of the distributions were other than exponential. In these cases the analysis of the models is more complex than the analysis of Markov models. In this chapter we introduce a methodology to extend the set of models that can be analyzed by Markov models while the distributions can be other than exponential.

5.1 Phase Type Distributions

Combination of exponential distributions, such as convolution and probabilistic mixtures, was used for a long time to approximate nonexponential distributions such that the composed model remained a Markov model. The most general class of distributions fulfilling these requirement is the set of phase-type distributions (commonly abbreviated as PH distributions) [73, 74].

Definition 5.1. Time to absorption in a Markov chain with N transient and 1 absorbing state is *phase-type* distributed (cf. Fig. 5.1).

5.1.1 Continuous-Time PH Distributions

Definition 5.1 is valid for both CTMCs and DTMCs. In this section we focus on the case of CTMCs.

It is possible to define a PH distribution by defining the initial probability vector $\boldsymbol{p}$ and the generator matrix $\boldsymbol{Q}$ of a Markov chain with $N+1$ states. Let states of the Markov chain be numbered so that the first N states are transient and the $N+1$th is absorbing and let $X(t)$ be the state of the Markov chain at time t. The distributions of the time to absorption, T, is related to the transient probabilities of the Markov

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_5, © Springer Science+Business Media, LLC 2013

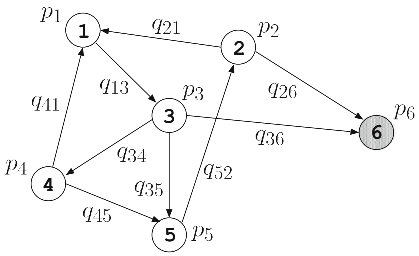


Fig. 5.1 Markov chain with five transient and an absorbing states defines a PH distribution

chain, which can be computed from the initial probability vector and the generator matrix as follows:

$$\mathbf{P}(T < t) = \mathbf{P}(X(t) = N+1) = \boldsymbol{p}\mathrm{e}^{\boldsymbol{Q}t}e_{N+1}^T,$$

where e_{N+1} is the row vector whose only nonzero element the $N+1$th is one. A multiplication of a row vector with e_{N+1}^T results in the $N+1$th element of the row vector.

Analysis of PH distributions based on this expression results in technical difficulties in more complex cases. A more convenient expression can be derived from the partitioned generator matrix, where the set of states is divided into transient states and an absorbing one

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A} & \mathbf{a} \\ \mathbf{0} & 0 \end{bmatrix},$$

where $\mathbf{a} = -\boldsymbol{A}\mathbb{1}$ and $\mathbb{1}$ is a column vector whose elements equal to one. The size of $\mathbb{1}$ is always assumed to be such that the multiplication is valid. A multiplication of a row vector by $\mathbb{1}$ results in the sum of the elements of the row vector. A column vector $\mathbf{a}$ that contains the transition rates to the absorbing state (Fig. 5.1) can be computed from $\boldsymbol{A}$ due to the fact that the row sum of $\boldsymbol{Q}$ is zero. The last row of $\boldsymbol{Q}$ is zero because the state $N+1$ is absorbing.

Matrix $\boldsymbol{A}$ is called a transient generator (or PH generator). It inherits its main properties from matrix $\boldsymbol{Q}$. The diagonal elements of $\boldsymbol{A}$ are negative, the nondiagonal elements are nonnegative, and the row sums of $\boldsymbol{A}$ are nonpositive. Due to the fact that the first N states are transient, matrix $\boldsymbol{A}$ is nonsingular, in contrast with matrix $\boldsymbol{Q}$, which is singular because $\boldsymbol{Q}\mathbb{1} = \mathbf{0}$.

In this book we restrict our attention to the case where a Markov chain starts from one of the transient states with probability one. In this case, the partitioned form of vector $\boldsymbol{p}$ is $\boldsymbol{p} = [\boldsymbol{\alpha} \mid 0]$. Based on the partitioned form of $\boldsymbol{p}$ and $\boldsymbol{Q}$, the CDF of the PH distribution is

$$\begin{aligned} F_T(t) &= \mathbf{P}(T \le t) = \mathbf{P}(T < t) = \mathbf{P}(X(t) = N+1) = 1 - \mathbf{P}(X(t) < N+1) \\ &= 1 - [\boldsymbol{\alpha} \mid 0]\,\mathrm{e}^{\boldsymbol{Q}t}\begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} = 1 - [\boldsymbol{\alpha} \mid 0]\sum_{i=0}^{\infty}\frac{t^i}{i!}\begin{bmatrix} \boldsymbol{A} & \mathbf{a} \\ \mathbf{0} & 0 \end{bmatrix}^i\begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&= 1 - [\boldsymbol{\alpha} \mid 0] \sum_{i=0}^{\infty} \frac{t^i}{i!} \begin{pmatrix} \boldsymbol{A}^i & \mathcal{I}_{\{i>0\}} \boldsymbol{A}^{i-1} \mathbf{a} \\ \mathbf{0} & 0 \end{pmatrix} \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} \\
&= 1 - [\boldsymbol{\alpha} \mid 0] \begin{bmatrix} \mathrm{e}^{\boldsymbol{A}t} & \bullet \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} = 1 - \boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t} \mathbb{1},
\end{aligned}
$$

where $\bullet$ denotes an irrelevant matrix block. Furthermore the PDF, the Laplace transform, and the moments of the PH distribution can be computed as

$$
\begin{aligned}
f_T(t) &= \frac{\mathrm{d}}{\mathrm{d}t} F_T(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^{\boldsymbol{A}t} \mathbb{1} = -\boldsymbol{\alpha} \sum_{i=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t} \frac{t^i}{i!} \boldsymbol{A}^i \mathbb{1} \\
&= -\boldsymbol{\alpha} \sum_{i=1}^{\infty} \frac{t^{i-1}}{(i-1)!} \boldsymbol{A}^{i-1} \boldsymbol{A} \mathbb{1} = -\boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t} \boldsymbol{A} \mathbb{1} = \boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t} \boldsymbol{a}, \\
f_T^*(s) &= \int_{t=0}^{\infty} \mathrm{e}^{-st} f_T(t) \mathrm{d}t = \boldsymbol{\alpha} \int_{t=0}^{\infty} \mathrm{e}^{-st} \mathrm{e}^{\boldsymbol{A}t} \mathrm{d}t\, \boldsymbol{a} \\
&= \boldsymbol{\alpha} \int_{t=0}^{\infty} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t\, \boldsymbol{a} = \boldsymbol{\alpha}\, (s\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{a}, \\
\mathbf{E}\left(T^n\right) &= \int_{t=0}^{\infty} t^n f_T(t) \mathrm{d}t = \boldsymbol{\alpha} \int_{t=0}^{\infty} t^n \mathrm{e}^{\boldsymbol{A}t} \mathrm{d}t\, \boldsymbol{a} = \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n-1} \boldsymbol{a} \\
&= \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n-1} (-\boldsymbol{A}) \mathbb{1} = \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n} \mathbb{1}.
\end{aligned}
$$

The infinite integrals of the preceding derivations are computed as follows:

$$
\begin{aligned}
\int_{t=0}^{\infty} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t &= \lim_{\tau \to \infty} \int_{t=0}^{\tau} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t = \lim_{\tau \to \infty} \int_{t=0}^{\tau} \sum_{i=0}^{\infty} \frac{t^i}{i!} (-s\boldsymbol{I} + \boldsymbol{A})^i \, \mathrm{d}t \\
&= \lim_{\tau \to \infty} \sum_{i=0}^{\infty} \frac{\tau^{i+1}}{(i+1)!} (-s\boldsymbol{I} + \boldsymbol{A})^{(i+1)} \; (-s\boldsymbol{I} + \boldsymbol{A})^{-1} \\
&= \lim_{\tau \to \infty} \left(\mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})\tau} - \boldsymbol{I} \right) (-s\boldsymbol{I} + \boldsymbol{A})^{-1} = (s\boldsymbol{I} - \boldsymbol{A})^{-1}, \qquad (5.1)
\end{aligned}
$$

where $\mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})\tau}$ vanishes in the convergence region of $f_T^*(s)$. The moments can also be computed from the Laplace transform

$$\begin{aligned}\mathbf{E}\left(T^n\right) &= (-1)^n \left.\frac{\mathrm{d}^n}{\mathrm{d}s^n} f_T^*(s)\right|_{s=0} = (-1)^n \boldsymbol{\alpha} \left.\frac{\mathrm{d}^n}{\mathrm{d}s^n}(s\boldsymbol{I}-\boldsymbol{A})^{-1}\right|_{s=0} \boldsymbol{a} \\ &= (-1)^n \boldsymbol{\alpha}\ (-1)^n n!(s\boldsymbol{I}-\boldsymbol{A})^{-n-1}\big|_{s=0} \boldsymbol{a} = \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n-1}\, \boldsymbol{a} \\ &= \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n} \mathbb{1}.\end{aligned}$$

The elements of $(-\boldsymbol{A})^{-1}$ have an important stochastic interpretation. Let T_{ij} be the time spent in state j before moving to the absorbing state when the process starts in state i:

$$\begin{aligned}\mathbf{E}\left(T_{ij}\right) &= \int_{t=0}^{\infty} \mathbf{E}\left(\mathcal{I}_{\{X(t)=j\,|\,X(0)=i\}}\right) \mathrm{d}t = \int_{t=0}^{\infty} \mathbf{P}\left(X(t)=j\,|\,X(0)=i\right) \mathrm{d}t \\ &= \int_{t=0}^{\infty} \left(\mathrm{e}^{\boldsymbol{A}t}\right)_{ij} \mathrm{d}t = \left(\int_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{A}t} \mathrm{d}t\right)_{ij} = \left((-\boldsymbol{A})^{-1}\right)_{ij}. \end{aligned} \tag{5.2}$$

Consequently, $(-\boldsymbol{A})^{-1}$ is nonnegative. Some characteristics of PH distributions can be seen from these expressions. From

$$f^*(s) = \boldsymbol{\alpha}(s\boldsymbol{I}-\boldsymbol{A})^{-1}\mathbf{a} = \boldsymbol{\alpha}\left[\frac{\det(s\boldsymbol{I}-Ay)_{ji}}{\det(s\boldsymbol{I}-\boldsymbol{A})}\right]\mathbf{a}$$

we have that the Laplace transform is a rational function of s where the degree of the polynomial in the numerator is at most $N-1$ and in the denominator it is at most N, where N is the number of transient states and $\det(s\boldsymbol{I}-\boldsymbol{A})_{ji}$ denotes the subdeterminant associated with element i, j. The related properties of PH distributions in the time domain can be obtained from the spectral decomposition of $\boldsymbol{A}$. Let η be the number of eigenvalues of $\boldsymbol{A}$ and λ_i the ith eigenvalue whose multiplicity is η_i. In this case

$$f_T(t) = \boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t}\boldsymbol{a} = \sum_{i=1}^{\eta}\sum_{j=1}^{\eta_i} a_{ij} t^{j-1} \mathrm{e}^{\lambda_i t}.$$

This means that in the case of distinct eigenvalues ($\eta = N$, $\eta_i = 1$) $f_T(t)$ is a combination of exponential functions with possibly negative coefficients, and in the case of multiple eigenvalues $f_T(t)$ is a combination of exponential polynomial functions. As a consequence, as t goes to infinity, the exponential function associated with the eigenvalue with maximal real part dominates the density, meaning that PH distributions have asymptotically exponentially decaying tail behavior.

A wide range of positive distributions can be approximated with PH distributions of size N. A set of PH distributions approximating different positive distributions are depicted in Fig. 5.2. The exponentially decaying tail behavior is not visible in the figure, but there is another significant limitation of PH distributions of size N.

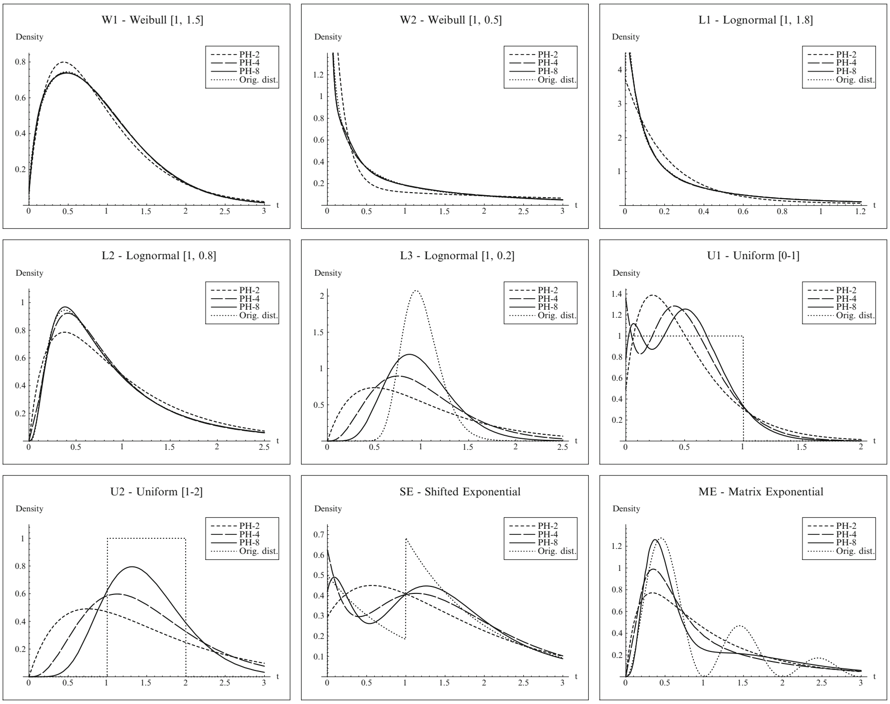


Fig. 5.2 Approximation of different positive distributions with $N = 2, 4, 8$ (figure copied from [13])

Theorem 5.2 ([3]). *The squared coefficient of variation of T ($cv^2(\tau) = \mathbf{E}\left(T^2\right)/\mathbf{E}\left(T\right)^2$) satisfies*

$$cv^2(\tau) \geq \frac{1}{N},$$

and the only CPH distribution that satisfies the equality is the Erlang (N) distribution:

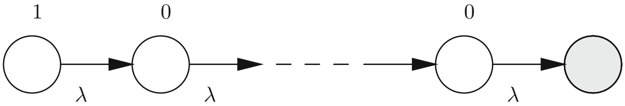


Figure 5.2 shows several distributions with low coefficient of variation whose approximation is poor due to this bound of the coefficient of variation. It is visible that PH distributions with larger N approximate these distributions significantly better. Theoretical results prove that as N tends to infinity, any positive distribution can be approximated arbitrarily closely.

5.1.2 Discrete-Time PH Distributions

The majority of the analysis steps and the properties of discrete-time PH distributions are similar to those of continuous-time PH distributions. Using a similar approach as for continuous-time PH distributions, the state-transition probability matrix can be partitioned as $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B} & \mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}$, where $\mathbf{b} = \mathbb{1} - \boldsymbol{B}\mathbb{1}$ and the initial probability vector $\boldsymbol{p}$ as $\boldsymbol{p} = [\boldsymbol{\alpha} \mid 0]$. $\boldsymbol{B}$ is a sub-stochastic matrix, whose elements are nonnegative and row sums are not greater than one. The probability that the chain moves to the absorbing state in the kth step is

$$r_k = Pr(T = k) = \boldsymbol{\alpha}\boldsymbol{B}^{k-1}\mathbf{b},$$

which defines the probability mass function (PMF) of T. The CDF can be obtained as

$$F(k) = Pr(T \leq k) = Pr(X_k = N+1) = 1 - Pr(X_k < N+1) = 1 - \boldsymbol{\alpha}\boldsymbol{B}^k\mathbb{1},$$

and the z-transform or generator function of T is

$$\mathcal{F}(z) = \mathbf{E}\left(z^T\right) = \sum_{k=0}^{\infty} z^k r_k = z\,\boldsymbol{\alpha}(\boldsymbol{I} - z\boldsymbol{B})^{-1}\mathbf{b}.$$

The factorial moments are

$$\gamma_n = \mathbf{E}\left(T(T-1)\ldots(T-n+1)\right) = \frac{\mathrm{d}^n}{\mathrm{d}z^n}\mathcal{F}(z)|_{z=1} = n!\,\boldsymbol{\alpha}(\boldsymbol{I} - \boldsymbol{B})^{-n}\boldsymbol{B}^{n-1}\mathbb{1}.$$

Like the continuous-time case the z-transform is a rational function of z

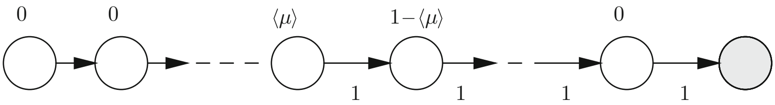
$$\mathcal{F}(z) = \mathbf{E}\left(z^T\right) = z\,\boldsymbol{\alpha}(\boldsymbol{I} - z\boldsymbol{B})^{-1}\mathbf{b} = z\,\boldsymbol{\alpha}\left[\frac{\det(\boldsymbol{I} - z\boldsymbol{B})_{ji}}{\det(\boldsymbol{I} - z\boldsymbol{B})}\right]\mathbf{b},$$

and based on the spectral decomposition of $\boldsymbol{B}$, the PMF is a combination of geometric series. The coefficient of variation of discrete PH (DPH) distributions is also bounded from below, but one of the most significant differences between the continuous and discrete PH distributions is that the bound in this case also depends on the mean of the distribution, $\mu = \mathbf{E}(T)$.

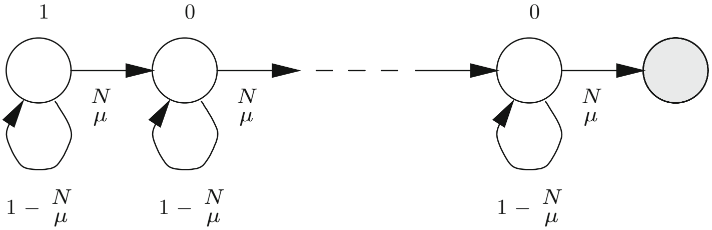
Theorem 5.3 ([92]). *The squared coefficient of variation of T satisfies the inequality*

$$cv^2(\tau) \geq \begin{cases} \dfrac{\langle\mu\rangle(1 - \langle\mu\rangle)}{\mu^2} & \text{if } \mu < N, \\[2ex] \dfrac{1}{N} - \dfrac{1}{\mu} & \text{if } \mu \geq N, \end{cases} \tag{5.3}$$

where $\langle x \rangle$ denotes the fraction part of x ($x = \lfloor x \rfloor + \langle x \rangle$). For $\mu \leq N$, $CV_{\min}$ is provided by the mixture of two deterministic distributions. Its DPH representation is



For $\mu > N$, $CV_{\min}$ is provided by the discrete Erlang distribution, whose DPH representation is



5.1.3 Special PH Classes

The set of PH distributions with N transient states is often too complex for particular practical applications (e.g., derivations by hand). There are special subclasses with restricted flexibility whose application is often more convenient. The most often used subclasses are

- Acyclic PH distributions,
- Hyper-Erlang distributions,
- Hyperexponential distributions ("parallel," "$cv > 1$").

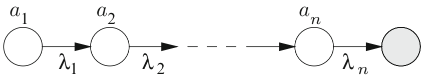
Acyclic PH Distributions

Definition 5.4. *Acyclic PH distributions* are PH distributions whose generator is an upper triangular matrix.

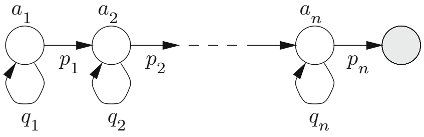
A direct consequence of the structural property of acyclic PH distributions is that the eigenvalues are explicitly given in the diagonal of the generator.

The practical applicability of acyclic PH distributions is due to the following result.

Theorem 5.5 ([25]). *Any acyclic PH distribution can be transformed into the following canonical form. In the case of continuous-time acyclic PH distributions:*



in the case of discrete-time acyclic PH distributions:



where the transition rates and probabilities are ordered such that $\lambda_i \leq \lambda_{i+1}$ *and* $p_i \leq p_{i+1}$.

This essential result allows one to consider only these canonical forms with $2N$ parameters to represent the whole acyclic PH class with N transient states.

Hyper-Erlang Distributions

Definition 5.6. A *hyper-Erlang distribution* is a probabilistic mixture of Erlang distributions.

Hyper-Erlang distributions are special acyclic PH distributions, and even fewer than $2N$ parameters can define them. Let ϑ be the number of Erlang branches, p_i the probability of taking branch i, and λ_i and n_i the parameters of the ith Erlang branch. These 3ϑ parameters completely define the hyper-Erlang distribution

$$f(t) = \sum_{i=1}^{\vartheta} p_i \, \frac{\lambda_i^{n_i} t^{n_i-1} \mathrm{e}^{-\lambda_i t}}{(n_i - 1)!}.$$

Hyperexponential Distributions

Definition 5.7. A *hyperexponential distribution* is a probabilistic mixture of exponential distributions.

Hyperexponential distributions are special hyper-Erlang distributions where the order parameter of the Erlang distribution is one ($n_i = 1$). The PDF of hyperexponential distributions

$$f(t) = \sum_{i=1}^{\vartheta} p_i \lambda_i \mathrm{e}^{-\lambda_i t}$$

is monotonically decreasing due to the fact that it is the mixture of monotonically decreasing exponential density functions.

5.1.4 Fitting with PH Distributions

As was mentioned in the introduction of this chapter, PH distributions are often used to approximate experimental or exactly given but nonexponential positive distributions in order to analyze the obtained system behavior with discrete-state Markov chains. The engineering description of the fitting procedure is rather straightforward: given a nonnegative distribution or a set of experimental data, find a "similar" PH distribution, but for the practical implementation of this approach we need to answer several underlying questions. First we formalize the problem as an optimization problem:

$$\min_{\text{PHparameters}} \left\{ \text{Distance}(F_{PH}(t), \hat{F}_{\text{Original}}(t)) \right\},$$

that is, we optimize the parameters of the PH distribution such that the distance between the original distribution and the PH distribution is minimal. The two main technical problems are finding a proper distance measure and solving the optimization problem. Several solutions to these problems have been proposed in the literature, but there is room for further improvement. Some of the typical distance measures are

- Squared CDF difference: $\int_0^\infty (F(t) - \hat{F}(t))^2 \mathrm{d}t$;
- Density difference: $\int_0^\infty |f(t) - \hat{f}(t)| \mathrm{d}t$;
- Relative entropy: $\int_0^\infty f(t) \log \left(\frac{f(t)}{\hat{f}(t)} \right) \mathrm{d}t$.

The optimization problems according to these distance measures are typically nonlinear and numerically difficult. The close relation of the relative entropy measure with commonly applied statistical parameters (likelihood) makes this measure the most popular one in practice. It is worth mentioning that the complexity of the optimization procedures largely depends on the number of parameters of the PH distributions. That is why we discussed the number of parameters of the aforementioned special PH subclasses. A few implemented fitting procedures are available on the Internet. One fitting procedure that uses acyclic PH distributions is PhFit [43], and one using hyper-Erlang distributions is G-fit [93]. The literature of PH fitting is rather extended. Several other heuristic fitting approaches exist, e.g., combined with moment matching, that are left to the ambitions of interested readers.

5.2 Markov Arrival Process

A continuous-time Markov arrival process (MAP) is a generalization of a Poisson process such that the interarrival times are PH distributed and can be dependent. One of the simplest interpretations of MAPs considers a CTMC, $J(t)$, with N states and

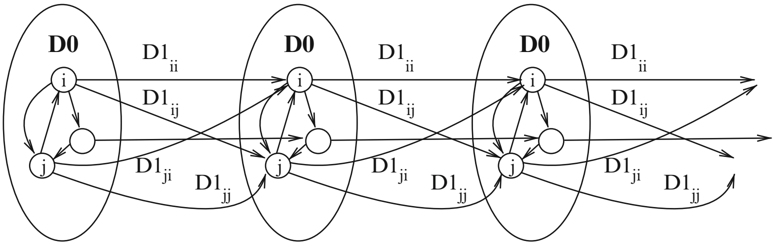


Fig. 5.3 Structure of Markov chain describing arrivals of a MAP

with generator $\boldsymbol{D}$, which determines the arrivals in the following way. While the Markov chain remains in state i, it generates arrivals according to a Poisson process at rate λ_i. When the Markov chain experiences a state transition from state i to j, then an arrival occurs with probability p_{ij} and does not occur with probability $1-p_{ij}$. Based on generator $\boldsymbol{D}$, rates λ_i $(i = 1, \ldots, N)$, and probabilities p_{ij} $(i, j = 1, \ldots, N, i \neq j)$, one can easily simulate the behavior of the MAP. Due to technical convenience MAPs are most commonly defined by a pair of matrices $\boldsymbol{D_0}$, $\boldsymbol{D_1}$, which are obtained from the previously introduced parameters in the following way:

$$\boldsymbol{D}_{0_{ij}} = \begin{cases} \boldsymbol{D}_{ij}(1-p_{ij}) & \text{if } i \neq j, \\ \boldsymbol{D}_{ii} - \lambda_i & \text{if } i = j, \end{cases} \qquad \boldsymbol{D}_{1_{ij}} = \begin{cases} \boldsymbol{D}_{ij}\, p_{ij} & \text{if } i \neq j, \\ \lambda_i & \text{if } i = j. \end{cases}$$

In this description, matrix $\boldsymbol{D}_0$ is associated with events that do not result in an arrival, and matrix $\boldsymbol{D}_1$ is associated with events that result in arrivals. By these definitions we have $\boldsymbol{D}_0 + \boldsymbol{D}_1 = \boldsymbol{D}$.

Based on these two matrices, we can investigate the counting process of arrivals. Let $N(t)$ be the number of arrivals of a MAP and $J(t)$ the state of the background Markov chain at time t. The $(N(t), J(t))$ $(N(t) \in \mathbb{N}, J(t) \in \{1, \ldots, N\})$ process is a CTMC. The transition structure of this Markov chain is depicted in Fig. 5.3. The set of states where the number of arrivals is n is commonly referred to as level n, and the state of the background Markov chain $(J(t))$ is commonly referred to as phase.

If the states are numbered in lexicographical order $((0,1), \ldots, (0,N), (1,1), \ldots, (1,N), \ldots)$, then the generator matrix has the form

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|}\hline \boldsymbol{D_0} & \boldsymbol{D_1} & & & \\ \hline & \boldsymbol{D_0} & \boldsymbol{D_1} & & \\ \hline & & \boldsymbol{D_0} & \boldsymbol{D_1} & \\ \hline & & & \boldsymbol{D_0} & \boldsymbol{D_1} \\ \hline & & & & \ddots \\ \hline \end{array},$$

where the matrix blocks are of size N. Comparing this with the CTMC describing the number of arrivals of the Poisson process in Eq. (3.18), we have conspicuous similarities: only the diagonal elements/blocks and the first subdiagonal elements/blocks are nonzero, and the transition structure of the arrival process is independent of the number of arrivals.

It is commonly assumed that $N(0) = 0$, and thus the initial probability is 0 for all states (n, j) where $n > 0$. Let vector $\boldsymbol{\pi}_0$ be the initial probability for states with $n = 0$. The arrival instants are determined by $N(t)$ as follows: $\Theta_n = \min(t : N(t) = n)$, and the nth interarrival time is $T_n = \Theta_n - \Theta_{n-1}$. Based on the simple block structure of the CTMC, we can analyze the properties of $N(t)$ and T_n. For example, the distribution of T_1 is

$$\begin{aligned}\mathbf{P}(T_1 \le t) &= 1 - \mathbf{P}(T_1 > t) = 1 - \mathbf{P}(N(t) = 0)\\ &= 1 - \sum_{i=1}^{N} \mathbf{P}(N(t) = 0, J(t) = i) = 1 - \pi_0 \mathrm{e}^{\boldsymbol{D}_0 t}\mathbb{1},\end{aligned}$$

that is, T_1 is PH distributed with initial vector $\boldsymbol{\pi}$ and generator $\boldsymbol{D}_0$. For the analysis of the nth interarrival time we introduce the phase distributions vector after the $n-1$th arrivals, $\boldsymbol{\pi}_{n-1}$. The ith element of this vector is the probability that after the $n-1$th arrivals the background Markov chain is in state i, that is, $(\boldsymbol{\pi}_{n-1})_i = \mathbf{P}(J(\Theta_{n-1}) = i)$. Based on $\boldsymbol{\pi}_{n-1}$ the distribution of T_n is

$$\begin{aligned}\mathbf{P}(T_n \le t) &= 1 - \mathbf{P}(T_n > t) = 1 - \mathbf{P}(N(t + \Theta_{n-1}) = n - 1)\\ &= 1 - \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{P}(J(\Theta_{n-1}) = i)\\ \mathbf{P}(N(t + \Theta_{n-1}) &= n - 1, J(t + \Theta_{n-1}) = j \,|\, J(\Theta_{n-1}) = i) = 1 - \text{ß}_{n-1}\mathrm{e}^{\boldsymbol{D}_0 t}\mathbb{1},\end{aligned}$$

that is, T_n is PH distributed with initial vector $\boldsymbol{\pi}_{n-1}$ and generator $\boldsymbol{D}_0$. The $\boldsymbol{\pi}_n$ vectors can be computed recursively. The ith element of $\boldsymbol{\pi}_1$ has the following stochastic interpretation:

$$\begin{aligned}(\boldsymbol{\pi}_1)_i &= \lim_{\Delta\to 0}\sum_{n=0}^{\infty}\sum_{j=1}^{N} \mathbf{P}(J(n\Delta) = j, T_1 > n\Delta)\\ &\quad \times \mathbf{P}(J((n+1)\Delta) = i, n\Delta < T_1 \le (n+1)\Delta)\\ &= \lim_{\Delta\to 0}\sum_{n=0}^{\infty}\sum_{j=1}^{N} \left(\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 n\Delta}\right)_j \left(\boldsymbol{D}_{1_{j,i}}\Delta + \sigma(\Delta)\right)\\ &= \int_{t=0}^{\infty}\sum_{j=1}^{N} \left(\text{ß}\mathrm{e}^{\boldsymbol{D}_0 t}\right)_j \boldsymbol{D}_{1_{j,i}}\,\mathrm{d}t\,,\end{aligned}$$

where the first term on the right-hand side of the first row is the probability that there is no arrival up to time $n\Delta$ and the background Markov chain is in state j at time

$n\Delta$, and the second term on the right-hand side of the first row is the probability that there is an arrival between $n\Delta$ and $(n+1)\Delta$ such that the background Markov chain is in state i at time $(n+1)\Delta$. Using Eq. (5.1), we further have

$$\text{ß}_1 = \text{ß} \int\limits_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{D}_0 t}\, \mathrm{d}t\ \boldsymbol{D}_1 = \text{ß}(-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1. \tag{5.4}$$

According to Eq. (5.4) we can compute the phase distribution after the first arrival from the initial distribution and the phase-transition probability matrix $\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1$. $\boldsymbol{P}$ is a stochastic matrix because from $(\boldsymbol{D}_0 + \boldsymbol{D}_1)\mathbb{1} = 0$ we have $-\boldsymbol{D}_0\mathbb{1} = \boldsymbol{D}_1\mathbb{1}$, from which $\boldsymbol{P}\mathbb{1} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1\mathbb{1} = (-\boldsymbol{D}_0)^{-1}(-\boldsymbol{D}_0)\mathbb{1} = \mathbb{1}$, and $(-\boldsymbol{D}_0)^{-1}$ is nonnegative according to Eq. (5.2). Applying the same analysis for the nth interval starting with initial phase distribution $\boldsymbol{\pi}_{n-1}$ we have $\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\boldsymbol{P}$.

5.2.1 Properties of Markov Arrival Processes

The basic properties of MAPs or the $(N(t), J(t))$ CTMC [with level process $N(t) \in \mathbb{N}$ and phase process $J(t) \in \{1, \dots, N\}$] are as follows.

- The phase distribution at arrival instants form a DTMC with transition probability matrix $\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1$. As a consequence, the phase distributions might be correlated at consecutive arrivals.
- The interarrival times are PH distributed with representation $(\boldsymbol{\pi}_0, \boldsymbol{D}_0)$, $(\boldsymbol{\pi}_1, \boldsymbol{D}_0)$, $(\boldsymbol{\pi}_2, \boldsymbol{D}_0)$, ... The interarrival times can be correlated due to the correlation of the initial phases.
- The phase process $(J(t))$ is a CTMC with generator $\mathbf{D} = \boldsymbol{D}_0 + \boldsymbol{D}_1$, which means that some properties of the phase process can be analyzed independent of the level process.
- The (time) stationary phase distribution $\boldsymbol{\alpha}$ is the solution of $\boldsymbol{\alpha}\mathbf{D} = 0, \boldsymbol{\alpha}\mathbb{1} = 1$.
- The (embedded) stationary phase distribution right after an arrival $\boldsymbol{\pi}$ is the solution of $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}, \boldsymbol{\pi}\mathbb{1} = 1$.
- These stationary distributions are closely related. On the one hand, the row vector of the mean time spent in the different phases during the stationary interarrival interval is $\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}$ [cf. Eq. (5.2)], from which the portion of time spent in the phases is

  $$\boldsymbol{\alpha} = \frac{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}}{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\mathbb{1}}.$$

  On the other hand, when the phase process is (time) stationary, the arrival intensities resulting in different initial phases for the next interarrival period are given by $\boldsymbol{\alpha}\boldsymbol{D}_1$, and after normalizing the result we have

  $$\boldsymbol{\pi} = \frac{\boldsymbol{\alpha}\boldsymbol{D}_1}{\boldsymbol{\alpha}\boldsymbol{D}_1\mathbb{1}}.$$

- The stationary interarrival time (T) is PH distributed with representation $(\boldsymbol{\pi}, \boldsymbol{D}_0)$, and its nth moment is $\mathbf{E}(T^n) = n!\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-n}\mathbb{1}$.
- The stationary arrival intensity can be computed both from $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ as follows:

$$\lambda = \boldsymbol{\alpha}\boldsymbol{D}_1\mathbb{1} = \frac{1}{\mathbf{E}(X)} = \frac{1}{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\mathbb{1}}.$$

The first equality is based on the arrival intensities in the (time) stationary phase process. The second equality is based on the mean stationary interarrival time.

Further properties of stationary MAPs can be computed from the joint density functions of consecutive interarrivals:

$$f_{T_0,T_1,\ldots,T_k}(x_0,\ldots,x_k) = \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\mathrm{e}^{\boldsymbol{D}_0 x_1}\boldsymbol{D}_1\ldots\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1}.$$

This joint density function describes the probability density that the process starts in phase i with probability $\boldsymbol{\pi}_i$ at time 0, it does not generate an arrival until time x_0, and an arrival occurs at x_0 according to the arrival intensities in $\boldsymbol{D}_1$. This arrival results in the second interarrival period's starting in phase j, and so on. If the MAP starts from a different initial phase distribution, e.g., γ, then the stationary embedded phase distribution vector $\boldsymbol{\pi}$ needs to be replaced by γ and the same joint density function applies. For example, we can compute the joint pdf of T_0 and T_k as

$$\begin{aligned} f_{T_0,T_k}(x_0,x_k) &= \int\limits_{x_1}\ldots\int\limits_{x_{k-1}} f_{T_0,T_1,\ldots,T_k}(x_0,\ldots,x_k)\,\mathrm{d}x_{k-1}\ldots\mathrm{d}x_1 \\ &= \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\boldsymbol{P}^{k-1}\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1}, \end{aligned}$$

where we used that $\int_x \mathrm{e}^{\boldsymbol{D}_0 x}\mathrm{d}x = (-\boldsymbol{D}_0)^{-1}$ according to Eq. (5.1). This expression indicates that T_0 and T_k are dependent due to their dependent initial phases. It is also visible that as k tends to infinity, this dependency vanishes according to the speed at which the Markov chain of the initial vectors with transition probability matrix $\boldsymbol{P}$ converges to its stationary distribution $\boldsymbol{\pi}$.

The lag-k correlation of a MAP can be computed based on $f_{T_0,T_k}(x_0,x_k)$ as follows:

$$\begin{aligned} \mathbf{E}(T_0T_k) &= \int\limits_{t=0}^{\infty}\int\limits_{\tau=0}^{\infty} t\ \tau\ \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 t}\boldsymbol{D}_1\boldsymbol{P}^{k-1}\mathrm{e}^{\boldsymbol{D}_0\tau}\boldsymbol{D}_1\mathbb{1}\ \mathrm{d}\tau\ \mathrm{d}t \\ &= \boldsymbol{\pi}(-\boldsymbol{D}_0)^{-2}\boldsymbol{D}_1\boldsymbol{P}^{k-1}(-\boldsymbol{D}_0)^{-2}\underbrace{\boldsymbol{D}_1\mathbb{1}}_{-\boldsymbol{D}_0\mathbb{1}} \\ &= \boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\boldsymbol{P}^{k}(-\boldsymbol{D}_0)^{-1}\mathbb{1} = \frac{1}{\lambda}\boldsymbol{\alpha}\boldsymbol{P}^{k}(-\boldsymbol{D}_0)^{-1}\mathbb{1}, \end{aligned}$$

since

$$\int_{t=0}^{\infty} t\ \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t = \underbrace{\left[t\ (\boldsymbol{D}_0)^{-1}\mathrm{e}^{\boldsymbol{D}_0 t}\right]_0^{\infty}}_{0} - \int_{t=0}^{\infty} (\boldsymbol{D}_0)^{-1}\ \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t$$

and

$$\int_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t = \lim_{T\to\infty}\textstyle\sum_{i=0}^{\infty}\frac{\boldsymbol{D}_0^i}{i!}\int_0^T t^i\mathrm{d}t = \displaystyle\lim_{T\to\infty}\sum_{i=0}^{\infty}\frac{\boldsymbol{D}_0^i}{i!}\frac{T^{i+1}}{i+1}$$

$$= \lim_{T\to\infty}(\boldsymbol{D}_0)^{-1}\left(\underbrace{\mathrm{e}^{\boldsymbol{D}_0 T}}_{\to 0} - I\right) = (-\boldsymbol{D}_0)^{-1}\ .$$

Based on $\mathbf{E}\left(T_0T_k\right)$ the covariance is

$$\mathrm{Cov}(T_0, T_k) = \mathbf{E}\left(T_0T_k\right) - \mathbf{E}\left(T\right)^2 = \frac{1}{\lambda}\ \boldsymbol{\alpha}\boldsymbol{P}^k(-\boldsymbol{D}_0)^{-1}\mathbb{1} - \frac{1}{\lambda^2},$$

and the coefficient of correlation is

$$\mathrm{Corr}(T_0, T_k) = \frac{\mathrm{Cov}(T_0, T_k)}{\mathbf{E}\left(T^2\right) - \mathbf{E}\left(T\right)^2} = \frac{\frac{\mathbf{E}(T_0T_k)}{\mathbf{E}(T)^2} - 1}{\frac{\mathbf{E}(T^2)}{\mathbf{E}(T)^2} - 1} = \frac{\lambda\ \boldsymbol{\alpha}\boldsymbol{P}^k(-\boldsymbol{D}_0)^{-1}\mathbb{1} - 1}{2\lambda\ \boldsymbol{\alpha}(-\boldsymbol{D}_0)^{-1}\mathbb{1} - 1}.$$

Starting from the joint density function of consecutive interarrivals we compute any joint moment for arbitrary series of interarrivals in a similar way as the lag-k correlation. For the interarrival series $a_0 = 0 < a_1 < a_2 < \ldots < a_k$ we have

$$f_{T_{a_0},T_{a_1},\ldots,T_{a_k}}(x_0, x_1, \ldots, x_k)$$
$$= \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\boldsymbol{P}^{a_1-a_0-1}\mathrm{e}^{\boldsymbol{D}_0 x_1}\boldsymbol{D}_1\boldsymbol{P}^{a_2-a_1-1}\ldots\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1},$$

and from that the joint moment $\mathbf{E}\left(T_{a_0}^{i_0}, T_{a_1}^{i_0}, \ldots, T_{a_k}^{i_0}\right)$ is

$$\mathbf{E}\left(T_{a_0}^{i_0}, T_{a_1}^{i_0}, \ldots, T_{a_k}^{i_0}\right)$$
$$= \boldsymbol{\pi} i_0!(-\boldsymbol{D}_0)^{-i_0}\boldsymbol{P}^{a_1-a_0}i_1!(-\boldsymbol{D}_0)^{-i_1}\boldsymbol{P}^{a_2-a_1}\ldots i_k!(-\boldsymbol{D}_0)^{-i_k}\mathbb{1}.$$

5.2.2 Examples of Simple Markov Arrival Processes

In this section we describe some basic arrival processes with MAP notations.

- PH renewal process: Consider an arrival process whose interarrival times are independent PH distributed with representation $(\boldsymbol{\alpha}, \boldsymbol{A})$. This is a special MAP characterized by $\boldsymbol{D}_0 = \boldsymbol{A}$, $\boldsymbol{D}_1 = \mathbf{a}\boldsymbol{\alpha}$.

- Interrupted Poisson process (IPP): Consider an arrival process determined by a background CTMC with two states, ON and OFF. The transition rate from ON to OFF is α and from OFF to ON it is β. There is no arrival in state OFF, and customers arrive according to a Poisson process at rate λ in state ON. The MAP description of the process is

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline -\alpha-\lambda & \alpha \\ \hline 0 & -\beta \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline \lambda & 0 \\ \hline 0 & 0 \\ \hline \end{array}.$$

- Markov modulated Poisson process (MMPP): Consider the arrival process determined by a background CTMC with generator $\boldsymbol{Q}$. While the CTMC is in state i, arrivals occur according to a Poisson process at rate λ_i. Let $\boldsymbol{\lambda}$ be the vector of arrival rates. This is a special MAP with representation $\boldsymbol{D}_0 = \boldsymbol{Q} - \text{diag}\langle\boldsymbol{\lambda}\rangle$, $\boldsymbol{D}_1 = \text{diag}\langle\boldsymbol{\lambda}\rangle$.
- Filtered MAP: Consider a MAP with representation $\hat{\boldsymbol{D}}_{\mathbf{0}}$, $\hat{\boldsymbol{D}}_{\mathbf{1}}$. The arrivals of this MAP are discarded with probability p. The obtained process is a MAP with representation $\boldsymbol{D}_0 = \hat{\boldsymbol{D}}_{\mathbf{0}} + p\hat{\boldsymbol{D}}_{\mathbf{1}}$, $\boldsymbol{D}_1 = (1-p)\hat{\boldsymbol{D}}_{\mathbf{1}}$.
- Cyclically filtered MAP: In the previous example, every MAP arrival is discarded with probability p. Now we consider the same MAP such that only every second arrival is discarded with probability p. It requires that we keep track of odd and even arrivals of the original MAP. It can be done by duplicating the phases such that the first half of them represents odd arrivals of the original MAP and the second half of them the even arrivals of the original MAP. The obtained process is a MAP with representation

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline \hat{\boldsymbol{D}}_{\mathbf{0}} & 0 \\ \hline p\hat{\boldsymbol{D}}_{\mathbf{1}} & \hat{\boldsymbol{D}}_{\mathbf{0}} \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline 0 & \hat{\boldsymbol{D}}_{\mathbf{1}} \\ \hline (1-p)\hat{\boldsymbol{D}}_{\mathbf{1}} & 0 \\ \hline \end{array}.$$

- Superposition of MAPs: Consider two MAPs with representation $\hat{\boldsymbol{D}}_{\mathbf{0}}$, $\hat{\boldsymbol{D}}_{\mathbf{1}}$ and $\tilde{\boldsymbol{D}}_{\mathbf{0}}$, $\tilde{\boldsymbol{D}}_{\mathbf{1}}$. The superposition of their arrival processes is a MAP with

$$\mathbf{D_0} = \hat{\boldsymbol{D}}_{\mathbf{0}} \bigoplus \tilde{\boldsymbol{D}}_{\mathbf{0}}, \text{ and } \mathbf{D_1} = \hat{\boldsymbol{D}}_{\mathbf{1}} \bigoplus \tilde{\boldsymbol{D}}_{\mathbf{1}},$$

where the Kronecker product is defined as $\boldsymbol{A} \bigotimes \boldsymbol{B} = \begin{array}{|c|}\hline A_{11}\boldsymbol{B} \ \dots \ A_{1n}\boldsymbol{B} \\ \vdots \qquad\qquad \vdots \\ A_{n1}\boldsymbol{B} \ \dots \ A_{nn}\boldsymbol{B} \\ \hline \end{array}$ and the Kronecker sum as $\boldsymbol{A} \bigoplus \boldsymbol{B} = \boldsymbol{A} \bigotimes \mathbf{I_B} + \mathbf{I_A} \bigotimes \boldsymbol{B}$. This example indicates one advantage of the $\mathbf{D_0}$, $\mathbf{D_1}$ description of MAPs. Using these matrices the description of the superposed process inherits the related property of the Cartesian product of independent Markov chains.
- Consider an arrival process where the interarrival time is either exponentially distributed with parameter λ_1 or with parameter λ_2 ($\lambda_1 \neq \lambda_2$). The arrivals are correlated such that an interarrival period with parameter λ_1 is followed by one

with parameter λ_1 with probability p or one with parameter λ_2 with probability $1 - p$. The interarrival periods with parameter λ_2 follow the same behavior. The obtained process is a MAP with

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline -\lambda_1 & 0 \\ \hline 0 & -\lambda 2 \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline p\lambda_1 & (1-p)\lambda_1 \\ \hline (1-p)\lambda_2 & p\lambda_2 \\ \hline \end{array}.$$

Probability p has a very intuitive meaning in this model. If $p \to 1$, then the correlation of the consecutive interarrivals is increasing and vice versa.

5.2.3 Batch Markov Arrival Process

A batch Markov arrival process (BMAP) is an extension of MAP with batch arrivals. It has an interpretation similar to that of a MAP.

A CTMC with generator $\boldsymbol{D}$ determines arrivals in the following way. While the Markov chain stays in state i, arrivals of batch size k occur according to a Poisson process at rate $\lambda_i^{(k)}$. When the Markov chain experiences a state transition from state i to j, arrivals of batch size k occur with probability $p_{ij}^{(k)}$ and no arrival occurs with probability $1 - \sum_k p_{ij}^{(k)}$. Generator $\boldsymbol{D}$, rates $\lambda_i^{(k)}$ $(i = 1, \ldots, N)$, and probabilities $p_{ij}^{(k)}$ $(i, j = 1, \ldots, N, i \neq j)$ determine the stationary behavior of BMAPs. Additionally, the initial distribution of the CTMC is needed for the analysis of the transient behavior. A BMAP is commonly described by matrices $\boldsymbol{D_k}$, which are obtained from the previously introduced parameters in the following way:

$$\boldsymbol{D}_{\boldsymbol{0}ij} = \begin{cases} \boldsymbol{D}_{ij}(1 - \sum_k p_{ij}^{(k)}) & \text{if } i \neq j, \\ \boldsymbol{D}_{ii} - \sum_k \lambda_i^{(k)} & \text{if } i = j, \end{cases} \qquad \boldsymbol{D}_{\boldsymbol{k}ij} = \begin{cases} \boldsymbol{D}_{ij}\, p_{ij}^{(k)} & \text{if } i \neq j, \\ \lambda_i^{(k)} & \text{if } i = j. \end{cases}$$

Based on this description the $(N(t), J(t))$ $(N(t) \in \mathbb{N}, J(t) \in \{1, \ldots, N\})$ process is a CTMC with transition structure depicted in Fig. 5.4. If the states are numbered in lexicographical order $((0, 1), \ldots, (0, N), (1, 1), \ldots, (1, N), \ldots)$, then the generator matrix has the form

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|}\hline \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 & D_3 & D_4 \\ \hline & \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 & D_3 \\ \hline & & \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 \\ \hline & & & \boldsymbol{D}_0 & \boldsymbol{D}_1 \\ \hline & & & & \ddots \\ \hline \end{array},$$

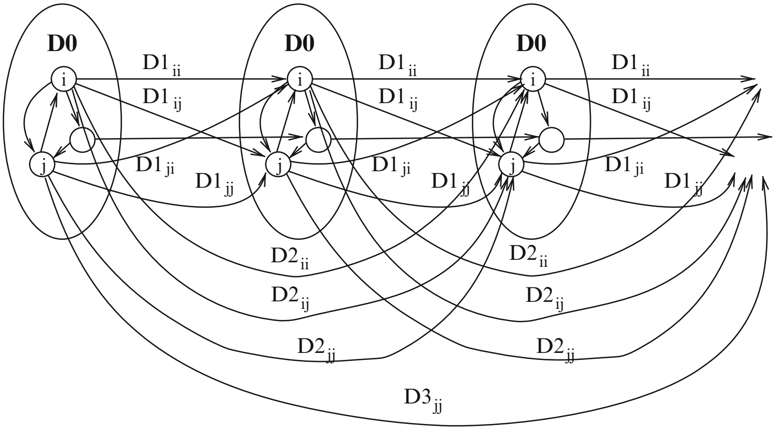


Fig. 5.4 Structure of Markov chain describing arrivals of a BMAP

To avoid complex cases it is commonly assumed that the considered BMAPs are regular:

- The phase process ($\mathbf{D}$) is irreducible.
- The mean interarrival time is positive and finite ($\boldsymbol{D}_0$ nonsingular).
- The mean arrival rate, $\mathbf{d} = \sum_{k=0}^{\infty} k\mathbf{D_k}\mathbb{1}$, is finite.

BMAP properties are similar to MAP properties. We refer the reader to [62] for further details.

5.3 Quasi-Birth-Death Process

There are very few Markov chain structures that ensure solutions with convenient analytical properties. One of these few Markov chain structures is the quasi-birth-death (QBD) process.

Definition 5.8. A CTMC $\{N(t), J(t)\}$ with state space $\{n, j\}$ ($n \in \mathbb{N}$, $j \in \{1, \ldots, J\}$) is a *QBD process* if transitions are restricted to modify n by at most one and the transitions are homogeneous for different n values for $n \geq 1$, i.e., the transition rate from $\{n, j\}$ to $\{n', j'\}$ is zero if $|n - n'| \geq 2$ and the transition rate from $\{n, j\}$ to $\{n', j'\}$ equals the transition rate from $\{1, j\}$ to $\{n' - n + 1, j'\}$ (cf. Fig. 5.5).

These structural descriptions are relaxed subsequently by considering various versions of this basic regular QBD model. Similar to the case of MAPs, $N(t)$ is commonly referred to as a *level* process (it represents, e.g., the number of customers in a queue), and $J(t)$ is commonly referred to as a *phase* process (it represents,

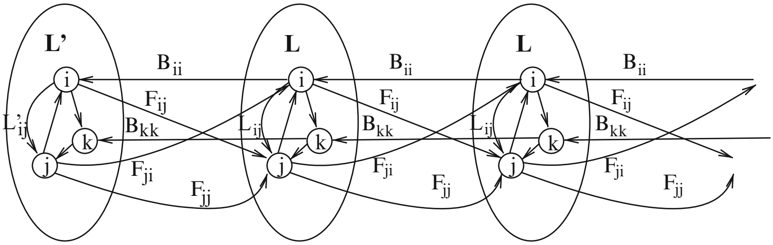


Fig. 5.5 Transition structure of QBD processes

e.g., the state of a randomly changing environment). Henceforth we assume that the considered QBD processes are irreducible with irreducible phase processes at the $n \geq 1$ levels (as detailed below).

Due to the structural properties of QBD processes, their state transitions can be classified as forward ($n \to n+1$), local ($n \to n$), and backward ($n \to n-1$). We apply the following notations:

- Matrix $\boldsymbol{F}$ of size $J \times J$ contains the rates of the forward transitions. The i, j element of $\boldsymbol{F}$ is the transition rate from $\{n, i\}$ to $\{n+1, j\}$ ($n \geq 0$).
- Matrix $\boldsymbol{L}$ of size $J \times J$ contains the rates of the local transitions for $n \geq 1$.
- Matrix $\boldsymbol{L}'$ of size $J \times J$ contains the rates of the local transitions for $n = 0$. Level 0 is irregular because there is no backward transition from level 0.
- Matrix $\boldsymbol{B}$ of size $J \times J$ contains the rates of the backward transitions. The i, j element of $\boldsymbol{F}$ is the transition rate from $\{n+1, i\}$ to $\{n, j\}$ ($n \geq 0$).

With these notations the structure of the generator matrix of a QBD process is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L}' & \boldsymbol{F} & & & \\ \hline \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & & \\ \hline & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & \\ \hline & & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \ddots & \ddots \\ \hline \end{array} .$$

The name QBD process comes from the fact that on the matrix block level the generator matrix has a birth–death structure.

Condition of Stability

The phase process of a QBD process in the regular part ($n > 1$) is a CTMC with generator matrix $\boldsymbol{A} = \boldsymbol{F} + \boldsymbol{L} + \boldsymbol{B}$. Let $\boldsymbol{A}$ be irreducible with stationary distribution

$\boldsymbol{\alpha}$ (that is, $\boldsymbol{\alpha A} = \mathbf{0}, \boldsymbol{\alpha}\mathbb{1} = 1$). The drift associated with the stationary distribution of the regular phase process is $d = \boldsymbol{\alpha F}\mathbb{1} - \boldsymbol{\alpha B}\mathbb{1}$. The sign of this drift indicates whether the average tendency of the level process is increasing in the regular part. If $d < 0$, then the QBD process is positive recurrent [74]. That is, the condition of stability of QBD processes is $d = \boldsymbol{\alpha F}\mathbb{1} - \boldsymbol{\alpha B}\mathbb{1} < 0$, where $\boldsymbol{\alpha}$ is the solution of $\boldsymbol{\alpha}(\boldsymbol{F} + \boldsymbol{L} + \boldsymbol{B}) = \mathbf{0}, \boldsymbol{\alpha}\mathbb{1} = 1$.

5.3.1 Matrix-Geometric Distribution

The stationary solution of a QBD process with generator $\boldsymbol{Q}$ is the solution of the linear system of equations $\boldsymbol{\pi Q} = \mathbf{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$, where $\boldsymbol{\pi}$ is the row vector of stationary probabilities. To utilize the regular structure of matrix $\boldsymbol{Q}$, we partition vector $\boldsymbol{\pi}$ according to the levels of the QBD process: $\boldsymbol{\pi} = \{\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots\}$. Using this partitioning the linear system of equations takes the following form:

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_1 \boldsymbol{B} = \mathbf{0}, \tag{5.5}$$

$$\boldsymbol{\pi}_{n-1} \boldsymbol{F} + \boldsymbol{\pi}_n \boldsymbol{L} + \boldsymbol{\pi}_{n+1} \boldsymbol{B} = \mathbf{0} \quad \forall n \geq 1, \tag{5.6}$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_n \mathbb{1} = 1. \tag{5.7}$$

Theorem 5.9. *The solution of Eqs.* (5.5)–(5.7) *in the case of a stable QBD process is $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$, where matrix $\boldsymbol{R}$ is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{RL} + \boldsymbol{R}^2 \boldsymbol{B} = \mathbf{0},$$

whose eigenvalues are inside the unit disk, and vector π_0 is the solution of a linear system of size J

$$\boldsymbol{\pi}_0(\boldsymbol{L}' + \boldsymbol{RB}) = \mathbf{0}$$

with normalizing condition

$$\boldsymbol{\pi}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\mathbb{1} = 1.$$

Proof. In the case of stable irreducible CTMCs, the solution of the linear system $\boldsymbol{\pi Q} = \mathbf{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$ is unique and identical with the stationary distribution of the CTMC. In this proof we only show that $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ satisfies the linear system and do not discuss the properties of the solutions of the quadratic matrix equations. The details of the spectral properties of the solutions are discussed, for example, in [62]. Substituting the $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ solution into Eq. (5.6) gives

$$\boldsymbol{\pi}_0 \boldsymbol{R}^{n-1} \boldsymbol{F} + \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{L} + \boldsymbol{\pi}_0 \boldsymbol{R}^{n+1} \boldsymbol{B} = \boldsymbol{\pi}_0 \boldsymbol{R}^{n-1}(\boldsymbol{F} + \boldsymbol{RL} + \boldsymbol{R}^2 \boldsymbol{B}) = \mathbf{0} \quad \forall n \geq 1,$$

which holds according to the definition of $\boldsymbol{R}$. Due to the fact that the eigenvalues of $\boldsymbol{R}$ are inside the unit disk, the infinite sum $\sum_{n=0}^{\infty} \boldsymbol{R}^n$ is finite, and we have $\sum_{n=0}^{\infty} \boldsymbol{R}^n = (\boldsymbol{I} - \boldsymbol{R})^{-1}$. Using this and substituting the $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ solution into Eqs. (5.5) and (5.7) gives

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_0 \boldsymbol{R} \boldsymbol{B} = \boldsymbol{0},$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \mathbb{1} = \boldsymbol{\pi}_0 (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbb{1} = 1,$$

which is the linear system defining $\boldsymbol{\pi}_0$. □

The stationary distribution of the form $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ are commonly referred to as matrix geometric distributions. This terminology refers also to the relation of homogeneous birth and death processes and QBD processes since the stationary distribution of homogeneous birth and death processes is geometric. Similar to the relation of Poisson processes and MAPs, QBD processes can be interpreted as an extension of birth and death processes such that their generator matrices have the same structure on the level of matrix blocks.

An extensive literature exists that deals with the properties of QBD processes and the efficient computation of matrix $\boldsymbol{R}$; therefore, we present here only two computational methods for matrix $\boldsymbol{R}$ and refer interested readers to [12] and references therein.

Linear algorithm

```
$\boldsymbol{R} := \boldsymbol{0}$;
REPEAT
  $\boldsymbol{R}_{old} := \boldsymbol{R}$;
  $\boldsymbol{R} := \boldsymbol{F}\,(-\boldsymbol{L} - \boldsymbol{R}\boldsymbol{B})^{-1}$;
UNTIL$||\boldsymbol{R} - \boldsymbol{R}_{old}|| \leq \epsilon$
```

Logarithmic algorithm

```
$\mathbf{H} := \boldsymbol{F}\,(-\boldsymbol{L})^{-1}$;
$\mathbf{K} := \boldsymbol{B}\,(-\boldsymbol{L})^{-1}$;
$\boldsymbol{R} := \mathbf{H}$;
$\mathbf{T} := \mathbf{K}$;
REPEAT
  $\boldsymbol{R}_{old} := \boldsymbol{R}$;
  $\boldsymbol{U} := \mathbf{HK} + \mathbf{KH}$;
  $\mathbf{H} := \mathbf{H}^2\,(\boldsymbol{I} - \boldsymbol{U})^{-1}$;
  $\mathbf{K} := \mathbf{K}^2\,(\boldsymbol{I} - \boldsymbol{U})^{-1}$;
  $\boldsymbol{R} := \boldsymbol{R} + \mathbf{HT}$;
  $\mathbf{T} := \mathbf{KT}$;
UNTIL$||\boldsymbol{R} - \boldsymbol{R}_{old}|| \leq \epsilon$
```

The input data of these algorithms are matrices $\boldsymbol{F}$, $\boldsymbol{L}$, $\boldsymbol{B}$, and a predefined accuracy parameter ϵ. The main differences between the algorithms are that the linear algorithm has a simpler iteration step and is more sensitive to drift d. When the drift is close to 0, the linear algorithm performs a huge number of iterations. The properties of the logarithmic algorithm are different. It has a more complex iteration step, but the number of iterations is tolerable also for drift values close to 0.

The following sections present different QBD variants whose stationary distributions are different variants of the matrix geometric distribution.

5.3.2 Quasi-Birth-and-Death Process with Irregular Level **0**

Many practical examples exist where the system has a regular behavior when it is in normal operation mode in some sense, but it has a different behavior (e.g., a different state transition structure or rates or even a different number of phases) when it is idle in some sense. Additionally, any CTMC that exhibits a regular QBD structure from a given point on can be considered a QBD process with irregular level 0, where level 0 is defined such that it contains the whole irregular part of the state space.

In general, a QBD process with irregular level 0 has the following block structure

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L}' & \boldsymbol{F}' & & & \\ \hline \boldsymbol{B}' & \boldsymbol{L} & \boldsymbol{F} & & \\ \hline & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & \\ \hline & & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \ddots & \ddots \\ \hline \end{array},$$

where the sizes of the blocks are identical for levels $1, 2, \ldots$, but the sizes of the blocks at level 0 can be different from the regular block size. If J is the regular block size and J_0 the block size at level 0, then matrices $\boldsymbol{F}$, $\boldsymbol{L}$, and $\boldsymbol{B}$ are of size $J \times J$, matrix $\boldsymbol{F'}$ is of size $J_0 \times J$, matrix $\boldsymbol{L'}$ is of size $J_0 \times J_0$, and matrix $\boldsymbol{B'}$ is of size $J \times J_0$.

In this case, the partitioned form of the linear system $\boldsymbol{\pi Q} = \boldsymbol{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$ is

$$\boldsymbol{\pi}_0\boldsymbol{L'} + \boldsymbol{\pi}_1\boldsymbol{B'} = \boldsymbol{0}, \tag{5.8}$$

$$\boldsymbol{\pi}_0\boldsymbol{F'} + \boldsymbol{\pi}_1\boldsymbol{L} + \boldsymbol{\pi}_2\boldsymbol{B} = \boldsymbol{0}, \tag{5.9}$$

$$\boldsymbol{\pi}_{n-1}\boldsymbol{F} + \boldsymbol{\pi}_n\boldsymbol{L} + \boldsymbol{\pi}_{n+1}\boldsymbol{B} = \boldsymbol{0} \quad \forall n \geq 2, \tag{5.10}$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_n\mathbb{1} = 1. \tag{5.11}$$

Theorem 5.10. *The solution of Eqs.* (5.8)–(5.11) *in the case of a stable QBD process is* $\boldsymbol{\pi}_0$ *and* $\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 \boldsymbol{R}^{n-1}$ $(n \geq 1)$*, where matrix* $\boldsymbol{R}$ *is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L} + \boldsymbol{R}^2\boldsymbol{B} = \boldsymbol{0}$$

whose eigenvalues are inside the unit disk and vectors $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$ *come from the solution of the linear system of size* $J_0 + J$

$$\boldsymbol{\pi}_0\boldsymbol{L'} + \boldsymbol{\pi}_1\boldsymbol{B'} = \boldsymbol{0},$$

$$\boldsymbol{\pi}_0\boldsymbol{F'} + \boldsymbol{\pi}_1(\boldsymbol{L'} + \boldsymbol{R}\boldsymbol{B}) = \boldsymbol{0},$$

with normalizing condition

$$\boldsymbol{\pi}_0\mathbb{1} + \boldsymbol{\pi}_1(\boldsymbol{I} - \boldsymbol{R})^{-1}\mathbb{1} = 1.$$

Proof. The proof follows the same pattern as that of Theorem 5.9. Substituting the matrix-geometric solution into the partitioned form of the stationary equations indicates that the solution satisfies the stationary equations. □

The linear system for $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be rewritten into the matrix form

$$[\boldsymbol{\pi}_0|\boldsymbol{\pi}_1]\begin{array}{|c|c|}\hline \boldsymbol{L'} & \mathbf{F}' \\ \hline \mathbf{B}' & \boldsymbol{L} + \boldsymbol{R}\boldsymbol{B} \\ \hline \end{array} = [\,\boldsymbol{0}\,|\,\boldsymbol{0}\,].$$

5.3.3 Finite Quasi-Birth-and-Death Process

Another frequently applied variant of QBD processes is the case where the level process has an upper limit. When the upper limit is at level m, the generator matrix takes the form

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|}\hline \boldsymbol{L'} & \boldsymbol{F} & & & \\ \hline \boldsymbol{B} & \boldsymbol{L} & \ddots & & \\ \hline & \boldsymbol{B} & \ddots & \boldsymbol{F} & \\ \hline & & \ddots & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \boldsymbol{B} & \boldsymbol{L''} \\ \hline \end{array}\,,$$

and the partitioned form of the stationary equation is

$$\boldsymbol{\pi}_0\boldsymbol{L}' + \boldsymbol{\pi}_1\boldsymbol{B} = \boldsymbol{0}, \tag{5.12}$$

$$\boldsymbol{\pi}_{n-1}\boldsymbol{F} + \boldsymbol{\pi}_n\boldsymbol{L} + \boldsymbol{\pi}_{n+1}\boldsymbol{B} = \boldsymbol{0} \quad 1 \le n \le m-1, \tag{5.13}$$

$$\boldsymbol{\pi}_{m-1}\boldsymbol{F} + \boldsymbol{\pi}_m\boldsymbol{L}” = \boldsymbol{0}, \tag{5.14}$$

$$\sum_{n=0}^{m}\boldsymbol{\pi}_n\mathbb{1} = 1. \tag{5.15}$$

Theorem 5.11. *The solution of Eqs.* (5.12)–(5.15) *in the case of a finite QBD process with* $d < 0$ *is* $\boldsymbol{\pi}_n = \boldsymbol{\alpha}\boldsymbol{R}^n + \boldsymbol{\beta}\boldsymbol{S}^{m-n}$ *(*$0 \le n \le m$*), where matrix* $\boldsymbol{R}$ *is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L} + \boldsymbol{R}^2\boldsymbol{B} = \boldsymbol{0}$$

whose eigenvalues are inside the open unit disk, matrix $\boldsymbol{S}$ *is the only solution of the quadratic matrix equation*

$$\boldsymbol{B} + \boldsymbol{S}\boldsymbol{L} + \boldsymbol{S}^2\boldsymbol{F} = \boldsymbol{0}$$

whose eigenvalues are on the closed unit disk, and vectors $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *are the solution of the size* $2J$ *linear system*

$$\boldsymbol{\alpha}\left(\boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B}\right) + \boldsymbol{\beta}\boldsymbol{S}^{m-1}\left(\boldsymbol{S}\boldsymbol{L}' + \boldsymbol{B}\right) = \boldsymbol{0},$$

$$\boldsymbol{\alpha}\boldsymbol{R}^{m-1}\left(\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L}”\right) + \boldsymbol{\beta}\left(\boldsymbol{S}\boldsymbol{F} + \boldsymbol{L}”\right) = \boldsymbol{0},$$

with normalizing condition

$$\boldsymbol{\alpha}\sum_{n=0}^{m}\boldsymbol{R}^n\mathbb{1} + \boldsymbol{\beta}\sum_{n=0}^{m}\boldsymbol{S}^n\mathbb{1} = 1.$$

Proof. The proof follows the same pattern as that of Theorem 5.9. Substituting the solution into the partitioned form of the stationary equations indicates that the solution satisfies the stationary equations. □

The matrix form of the linear system for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is

$$[\boldsymbol{\alpha}|\boldsymbol{\beta}]\begin{array}{|c|c|}\hline \boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B} & \boldsymbol{R}^{m-1}\left(\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L}”\right) \\ \hline \boldsymbol{S}^{m-1}\left(\boldsymbol{S}\boldsymbol{L}' + \boldsymbol{B}\right) & \boldsymbol{S}\boldsymbol{F} + \boldsymbol{L}” \\ \hline \end{array} = [\,0 \mid 0\,]\,.$$

Matrix $\boldsymbol{S}$ can be computed by the same linear or logarithmic procedures as matrix $\boldsymbol{R}$. If the drift is positive ($d > 0$) in a finite QBD process, then the numbering of the levels needs to be inverted ($0 \to m, 1 \to m-1, \dots, m \to 0$), and we obtain a new finite QBD process whose drift is negative. It is worth mentioning that due to the fact that $d < 0$, matrix $\boldsymbol{S}$ has an eigenvalue on the unit circle, and consequently $\sum_{n=0}^{\infty} \boldsymbol{S}^n$ does not converge. Fortunately, this does not affect the applicability of Theorem 5.11 because we need to compute only the finite sum $\sum_{n=0}^{m} \boldsymbol{S}^n$.

5.4 Exercises

Exercise 5.1. X and Y are independent continuous-time PH distributed random variables with representations $(\boldsymbol{\alpha}, \boldsymbol{A})$ and $(\boldsymbol{\beta}, \boldsymbol{B})$, respectively. Define the distribution of the following random variables:

- $Z_1 = c_1 X$;
- Z_2 equals X with probability p and to Y with probability $1 - p$;
- $Z_3 = c_1 X + c_2 Y$;
- $Z_4 = \text{Min}(X, Y)$;
- $Z_5 = \text{Max}(X, Y)$.

Exercise 5.2. X and Y are independent discrete-time PH distributed random variables with representations $(\boldsymbol{\alpha}, \boldsymbol{A})$ and $(\boldsymbol{\beta}, \boldsymbol{B})$, respectively. Define the distribution of the following random variables:

- $Z_1 = c_1 X$;
- Z_2 equals to X with probability p and to Y with probability $1 - p$;
- $Z_3 = c_1 X + c_2 Y$;
- $Z_4 = \text{Min}(X, Y)$;
- $Z_5 = \text{Max}(X, Y)$.

Exercise 5.3. There are two machines, A and B, at a production site. Their failure times are exponentially distributed with parameters λ_A and λ_B, respectively. Their repair times are also exponentially distributed with parameters μ_A and μ_B, respectively. A lone repairman can work on only one machine at a time. At a given time, both machines work. Compute the distribution and the moments of the time to the first complete breakdown when both machines fail.