1 The limiting behavior of a reducible DTMC

As was illustrated in Example 2.18, the limiting behavior of a reducible DTMC depends on the initial distribution. In this section, we will show how for these chains the limiting distribution, if it exists, can be determined. The procedure to calculate the limiting distribution consists of three steps:

1. In the first step we divide the state space S of the DTMC in a number of *end classes* $E_1, \ldots, E_m$ and a set of transient states C such that $S = C \cup E_1 \cup \cdots \cup E_m$. Here, an end class is a set of *connected* states such that once the DTMC is in this set of states, it will never leave it anymore. With a set of connected states we mean that from each state in the set you can reach, in one or more steps, all other states in the set. Transient states are those states that do not belong to one of the end classes.

2. In the second step we determine for each state $i \in S$ the probabilities q_{i,E_j} that the DTMC, when it starts in i will eventually end up in the end class E_j, $j = 1, \ldots, m$.

3. In the third step we determine for each end class the limiting distribution of the DTMC given that it entered the end class. Remark that once the DTMC enters an end class, it behaves like an irreducible DTMC with the end class as state space. Hence, we know how to find this limiting distribution if it exists (see Theorem 2.5).

In the following example we will show how a combination of these three steps leads to the limiting distribution of a reducible DTMC.

Example: Consider a DTMC on the state space $S = \{1, 2, 3, 4, 5, 6, 7\}$ with transition probability matrix

$$P = \begin{pmatrix} 1/3 & 1/6 & 1/3 & 0 & 0 & 0 & 1/6 \\ 3/5 & 0 & 0 & 1/5 & 1/5 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/4 & 3/4 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix}.$$

Step 1: For this DTMC, there are two end classes $E_1 = \{3, 4\}$ and $E_2 = \{5, 6, 7\}$ and the set of transient states is $C = \{1, 2\}$.

Step 2: Clearly, for $i \in \{3, 4\}$ we have $q_{i,E_1} = 1$ and $q_{i,E_2} = 0$, and similarly for $i \in \{5, 6, 7\}$ we have $q_{i,E_1} = 0$ and $q_{i,E_2} = 1$. For $i \in \{1, 2\}$, q_{i,E_1} can be determined from the following system of equations

$$\begin{aligned} q_{1,E_1} &= \frac{1}{3} + \frac{1}{3} q_{1,E_1} + \frac{1}{6} q_{2,E_1}, \\ q_{2,E_1} &= \frac{1}{5} + \frac{3}{5} q_{1,E_1}, \end{aligned}$$

and similarly q_{i,E_2}, $i \in \{1,2\}$, can be determined from the system of equations

$$\begin{aligned} q_{1,E_2} &= \frac{1}{6} + \frac{1}{3} q_{1,E_2} + \frac{1}{6} q_{2,E_2}, \\ q_{2,E_2} &= \frac{1}{5} + \frac{3}{5} q_{1,E_2}. \end{aligned}$$

These systems of equations follow from looking at what happens at the first time step. For example, starting in state 1, with probability 1/3 you immediately jump into E_1, with probability 1/6 you immediately jump into E_2, and with probabilities 1/3 and 1/6 you jump to the transient states 1 and 2, respectively. Using the Markov property, this leads to the equations for q_{1,E_1} and q_{1,E_2}. Starting in state 2, similar reasoning leads to the equations for q_{2,E_1} and q_{2,E_2}. Check that the solutions of the systems of equations are given by $q_{1,E_1} = 11/17$, $q_{2,E_1} = 10/17$ and $q_{1,E_2} = 6/17$, $q_{2,E_2} = 7/17$. Furthermore, remark that of course we have for all $i \in S$ that $q_{i,E_1} + q_{i,E_2} = 1$.

Step 3: For end class E_1, the limiting distribution is given by the unique solution to

$$[\pi_3 \ \pi_4] = [\pi_3 \ \pi_4] * \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

such that $\pi_3 + \pi_4 = 1$. This yields $\pi_3 = 1/3, \pi_4 = 2/3$.

Similarly, for end class E_2, the limiting distribution is given by the unique solution to

$$[\pi_5 \ \pi_6 \ \pi_7] = [\pi_5 \ \pi_6 \ \pi_7] * \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

such that $\pi_5 + \pi_6 + \pi_7 = 1$. This yields $\pi_5 = 3/13, \pi_6 = 4/13$ and $\pi_7 = 6/13$.

Depending on the initial distribution, the limiting distribution for the reducible DTMC can now be calculated by combining steps 2 and 3. For example, if $P(X_0 = 1) = 1$, then the limiting distribution will be

$$(0,\, 0,\, \frac{11}{17} \cdot \frac{1}{3},\, \frac{11}{17} \cdot \frac{2}{3},\, \frac{6}{17} \cdot \frac{3}{13},\, \frac{6}{17} \cdot \frac{4}{13},\, \frac{6}{17} \cdot \frac{6}{13}).$$

Questions:

1. What is the limiting distribution if $P(X_0 = 3) = 1$?
2. And what if $P(X_0 = 1) = P(X_0 = 2) = \frac{1}{2}$?
3. What can you say about $\lim_{n\to\infty} P^n$ in this example?

Exercises:

1. Consider a DTMC on the state space $S = \{1, 2, 3, 4, 5, 6, 7\}$ with transition probability matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/2 & 1/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 3/4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 2/3 & 1/12 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

   (a) Construct the transition diagram of the DTMC, determine the three end classes E_1, E_2, E_3 and the set C of transient states.

   (b) The initial distribution is $a^{(0)} = (0, 0, 1, 0, 0, 0, 0)$. What is the limiting distribution?

   (c) Determine $P^\infty = \lim_{n\to\infty} P^n$.

2. Consider a DTMC on the state space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ with transition probability matrix

$$P = \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/4 & 1/8 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 3/4 \end{pmatrix}.$$

   (a) Construct the transition diagram of the DTMC, determine the three end classes E_1, E_2, E_3 and the set C of transient states.

   (b) Determine the limiting distribution of the DTMC if the initial distribution is given by $a^{(0)} = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$.

   (c) Determine $P^\infty = \lim_{n\to\infty} P^n$.

3. A training consists of three successive courses. A course has to be finished with success before the next course can be started. Each course takes a week and has a final exam at the end. All participating students take this exam. Fifty percent of the students, who do not pass the final exam of a course at the first time they do it, leaves the

training and the other fifty percent repeats the course and the exam in the next week. A student who does not pass the exam at the second trial, also leaves the training. The probability of passing the exam is given below:

	trial 1	trial 2
course 1	0.7	0.5
course 2	0.8	0.6
course 3	0.9	0.7

(a) Give a DTMC that models the development of the training of an arbitrary student: state space S, matrix of transition probabilities P, transition diagram. Leaving the training by a student, with or without diploma, has to be described also in the model.

(b) Calculate the probability that a student leaves the training without a diploma at the end of the fourth week.

(c) Calculate the probabilities that a student leaves the training with and without diploma.

4. A company has for a certain job a salary scale with four salary levels 1, 2, 3 and 4. During a calendar year employees can move over from this job to another job with another salary scale or can leave the company. At the end of each calendar year the company determines for each employee in these jobs the salary level for the next year. In a model for the development of the salary of an employee in this job in the time, the simplifying assumption is made that moving over to another job and leaving the company take also place at the end of a calendar year.
Experience from the past has lead to the following assumptions in the model. From the group of employees on one of the four salary levels at the end of a year, $p\%$ is leaving the company, $q\%$ is moving over to another job within the company, $r\%$ is staying at the same salary level in the next year and $s\%$ is promoted to the next salary level in the scale. The values of p, q, r, s are:

	p	q	r	s
salary level 1	20	0	0	80
salary level 2	20	0	30	50
salary level 3	20	10	40	30
salary level 4	10	40	50	0

A new employee in this job starts at salary level 1.

(a) Give a DTMC that models the development of the salary of an employee in this job: state space S, matrix of transition probabilities P, transition diagram. Leaving the salary scale by an employee has to be described also in the model.

(b) Calculate the probability that an employee has another job within the company at the beginning of the fifth year.

(c) Calculate the probability that an employee leaves the company. What percentage of the employees in these jobs gets another job within the company?

2 Cohort models

Discrete time Markov chains are often used in the study of the behavior of a group of persons or objects. These systems are often called *Cohort models*. An example of a cohort model already appeared in the manpower example (see Examples 2.6, 2.28 and 2.31). In this example we assumed that each time an employee leaves the company, he or she is instantaneously replaced by a new employee. In this way the total number of employees is kept constant. In order to calculate now for example the fraction of employees in the different grades, we then have to calculate the unique occupancy distribution of an irreducible DTMC as was illustrated in Example 2.28 on page 39.

However, sometimes in applications it is not realistic to assume that the number of persons in the group is constant over time. For example, if we model the group of persons that have a car insurance at a certain insurance company, then the number of persons ending their insurance in a given period may be different from the number of persons starting an insurance in that period. Hence, clearly the total number of persons having a car insurance varies over time. In the sequel we will show how in this case we can calculate quantities like the expected number of persons in the group at a certain time instant, the long-run expected number of persons in the group, the division of the persons in the group over the different levels (here different levels could for example be different premium categories in a bonus-malus system) and so on.

Assume we have a group of persons, where the behavior of each person can be described by a DTMC with state space $S = \{0, 1, 2, \ldots, N\}$ and transition probability matrix P. Furthermore, assume that state 0 represents the situation that the person has left the system (cf. Example 2.31) and denote with Q that part of the transition probability matrix corresponding to transitions from states $\{1, 2, \ldots, N\}$ to states $\{1, 2, \ldots, N\}$. Here, Q is a *sub*-stochastic matrix, i.e., a matrix with $q_{i,j} \geq 0$ for all i and j and $\sum_{j=1}^{N} q_{i,j} \leq 1$ for all i.

Now let us introduce the following notation:

- $r_i^{(n)}$: the expected number of *new* persons, called *recruits*, entering the group from outside at time n in state i.
- $s_i^{(n)}$: the expected total number of persons in the group at time n in state i.

If we denote with $r^{(n)}$ and $s^{(n)}$ the transient vectors

$$r^{(n)} = [r_1^{(n)}, r_2^{(n)}, \ldots, r_N^{(n)}], \; s^{(n)} = [s_1^{(n)}, s_2^{(n)}, \ldots, s_N^{(n)}],$$

then we have

$$s^{(n)} = r^{(n)} + s^{(n-1)} \cdot Q \tag{1}$$

So, once we know the initial expected group size $s^{(0)}$ and the successive sizes of the expected number of recruits $r^{(1)}, r^{(2)}, r^{(3)}, \ldots$, we can calculate $s^{(1)}, s^{(2)}, s^{(3)}, \ldots$.

In the case that the expected number of recruits is constant over time, i.e., $r^{(n)} = r$ for all n, we can also determine the long-run expected number of persons in the group in the different states. ¿From (1) we see that, if $s = \lim_{n\to\infty} s^{(n)}$ exists, it satisfies

$$s = r + s \cdot Q,$$

and hence

$$s = r \cdot (I - Q)^{-1}$$

where I is the identity matrix.

Example: Consider a DTMC on the state space $S = \{0, 1, 2, 3, 4, \}$ with transition probability matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 & 0 \\ 0.05 & 0 & 0.7 & 0.25 & 0 \\ 0.1 & 0 & 0 & 0.7 & 0.2 \\ 0.10 & 0 & 0 & 0 & 0.9 \end{pmatrix}.$$

Furthermore, assume $s^{(0)} = [10, 10, 10, 10]$ and $r^{(n)} = [10, 0, 0, 0]$ for all n. Then we will have

$$\begin{aligned} s^{(1)} &= [10,0,0,0] + [10,10,10,10] \cdot \begin{pmatrix} 0.6 & 0.2 & 0 & 0 \\ 0 & 0.7 & 0.25 & 0 \\ 0 & 0 & 0.7 & 0.2 \\ 0 & 0 & 0 & 0.9 \end{pmatrix} = [16, 9, 9.5, 11], \\ s^{(2)} &= [10,0,0,0] + [16,9,9.5,11] \cdot \begin{pmatrix} 0.6 & 0.2 & 0 & 0 \\ 0 & 0.7 & 0.25 & 0 \\ 0 & 0 & 0.7 & 0.2 \\ 0 & 0 & 0 & 0.9 \end{pmatrix} = [19.6, 9.5, 8.9, 11.8], \\ s^{(3)} &= [10,0,0,0] + [19.6,9.5,8.9,11.8] \cdot \begin{pmatrix} 0.6 & 0.2 & 0 & 0 \\ 0 & 0.7 & 0.25 & 0 \\ 0 & 0 & 0.7 & 0.2 \\ 0 & 0 & 0 & 0.9 \end{pmatrix} = [21.76, 10.57, 8.61, 12.40] \end{aligned}$$

and

$$\begin{aligned} s = \lim_{n\to\infty} s^{(n)} &= [10,0,0,0] \cdot \begin{pmatrix} 0.4 & -0.2 & 0 & 0 \\ 0 & 0.3 & -0.25 & 0 \\ 0 & 0 & 0.3 & -0.2 \\ 0 & 0 & 0 & 0.1 \end{pmatrix}^{-1} \\ &= [25, 16.67, 13.89, 27, 78] \approx [25, 17, 14, 28]. \end{aligned}$$

Exercises:

1. A car company has a number of mechanics that maintain the cars of the customers of the company. Within this group of mechanics three positions can be distinguished: 1 = trainee, 2 = junior mechanic, 3 = senior mechanic. At the end of each year the management looks at the structure of the group of mechanics.
   It appears after the assessment of the management that next year: 40% of the trainees, present in the last year, is staying another year on the trainee post, that 10% of the present trainees will be working in the car company as junior mechanic and that 50% (fore some reason) will not be working anymore for the car company.
   It appears furthermore that next year: 50% of the junior mechanics is staying another year in the same job, that 10% of the present junior mechanics will be working in the car company as senior mechanic and that 40% (fore some reason) will not be working anymore for the car company.
   Finally appears that 80% of the present senior mechanics is staying in the car company next year and that 20% is leaving.

   (a) The management has the following recruitment policy: every time someone leaves the group of mechanics, a new trainee from outside is hired. The total number of mechanics is 39. Show that the expected long-run numbers of mechanics in the three positions 1, 2 and 3 are 30, 6 and 3.

   (b) The management changes the recruitment policy; every year a certain number of new employees may be recruted from outside for each of the three positions in the group of mechanics.

      i. At the beginning of a certain year 15 trainees, 3 junior mechanics and 1 senior mechanic are working for the car company. During that year the management intends to recruit 7 new trainees, 2 new junior mechanics and 1 new senior mechanic. Determine the expected number of trainees, junior and senior mechanics at the beginning of the next year.

      ii. The coaching of the trainees demands a lot of time, therefore the management decides to employ some extra junior and senior mechanics and aims at a long-run expected number of 30 trainees, 8 junior mechanics and 4 senior mechanics. Calculate for each of the three positions in the group of mechanics the average number of new employees that will have to be recruted every year.

2. An insurance company has a insurance on body-work for cars. The nominal premium depends on several factors. Premium is always paid for a period of 12 months. Every owner of an insured car can get a discount or a surcharge on the nominal premium. The height of the discount or surcharge depends on the insurance claim history. Such a bonus-malus system has a large number of levels.
   Some simplifying assumptions have to be made here. Four premium levels will be distinguished in a bonus-malus system: 1 = 10% surcharge on the nominal premium, 2 = 0% discount on the nominal premium, 3 = 25% discount on the nominal premium, 4 = 50% discount on the nominal premium. Level 1 is called the "lowest" level, level 4 is called the "highest" level for the premium. If an owner of an insured car has a

premium on a certain level in a year and has no insurance claim for the company in that year, the premium in the next year will be on nearest higher level or stays on the highest level. If the owner does have a insurance claim in tha year, the premium in the next year will be on the nearest lower level or stays on the lowest level. The probability of two or more insuranceclaims in one year is neglectible.
For all insured cars the 12 months premium period coincides with a calendaryear.
At the end of a year the composition of the total group of insured cars is considered and the premium level for next year for each of the insured cars is determined. Cars on a certain premium level in a year can stay on that level or go to a nearest level. Owners can leave the insurance company and go to another. The company has collected a huge amount of statistical data concerning the insurance claims and the changes in premium level in the next year that are caused through these. From these data the following percentages are calculated.
Owners/cars on level 1 at the end of a year: 20% leaves the company, 50% stays on level 1, 30% goes to level 2. Owners/cars on level 2 at the end of a year: 10% leaves the company, 40% goes to level 1, 50% goes to level 3. Owners/cars on level 3 at the end of a year: 5% leaves the company, 25% goes to level 2, 70% goes to level 4. Owners/cars on level 4 at the end of a year: 2.5% leaves the company, $12,5$% goes to level 3, 85% stays on level 4.

(a) Model this car insurance with a DTMC. Give the state space, the matrix of transition probabilities, the transition diagram.

(b) At the beginning of a certain year there are 15.000 insured cars on level 1, 60.000 on level 2, 40.000 on level 3 and 35.000 on level 4. During that year new car-insurances are effected: 0 on level 1, 10.000 on level 2, 5.000 on level 3 and 1.000 on level 4. Calculate for each level the expected number of insured cars at the beginning of next year.

(c) Determine the long-run expected number of insured cars on the four premium levels, if yearly an average number of 0, 10.000, 5.000 and 1.000 new car-insurances is effected on level 1, 2, 3 and 4.

3. Consider problem 3 in section 1 of this "hand-out" with the theory of the limiting behavior of a reducible DTMC. This DTMC can be interpreted as a cohort model.

(a) If weekly an long-run expected number of 10 persons has to finish the training, how many new students on the average will have to start the training every week?

(b) Consider the situation of the previous question. Calculate the expected number of students that has left without passing any exam, that has left only with a good result for course 1, that has left only without a good result for course 3.

4. Consider problem 4 in section 1 of this "hand-out" with the theory of the limiting behavior of a reducible DTMC. This DTMC can be interpreted as a cohort model.
At the end of a certain year the number of employees on the four salary levels is: 13, 17, 14, 12. The management aims at a long-run expected number of 77 employees in total for this job. Determine the expected number of new employees for this job that has to be recruted every year from now on.

3 Queues with general inter-arrival and service times

So far, we looked at queueing systems in which either the inter-arrival times or the service times are exponentially distributed random variables. In this section we shortly discuss queueing systems in which both the inter-arrival times and the service times are generally distributed. We present some approximative results, first for the single server $G/G/1$ queue and after that for the multi server $G/G/s$ queue.

3.1 The $G/G/1$ queue

For the $G/G/1$ queue only approximations exist for performance measures like the expected waiting time and the expected number in the system. The simplest approximations assume that the randomness in the inter-arrival times has more or less the same effect on the expected waiting time in the queue as the randomness in the service times.

In the sequel, we denote with $\mathrm{E}(A)$ the mean inter-arrival time and with $\mathrm{Var}(A)$ the variance of the inter-arrival time. Similarly, we use $\mathrm{E}(B)$ and $\mathrm{Var}(B)$ for the mean and variance of the service times. Finally, with

$$c_A^2 = \frac{\mathrm{Var}(A)}{(\mathrm{E}(A))^2}, \text{ and } c_B^2 = \frac{\mathrm{Var}(B)}{(\mathrm{E}(B))^2}$$

we denote the squared coefficient of variation of the inter-arrival times and the service times, respectively. The squared coefficient of variation is a dimensionless measure of the variation of a random variable. Remark that in the case of exponentially distributed inter-arrival times (resp. service times) we have that $c_A^2 = 1$ (resp. $c_B^2 = 1$). An often used approximation for the mean waiting time in the queue in the $G/G/1$ queue is given by (see, e.g., [2, 3])

$$W_q \approx \frac{\rho}{1-\rho} \cdot \frac{c_A^2 + c_B^2}{2} \cdot E(B), \tag{2}$$

where $\rho = \mathrm{E}(B)/\mathrm{E(A)}$ is the traffic intensity. Of course, once an approximation for W_q is obtained, we can also approximate the mean time in the system $W = W_q + E(B)$ and via Little's Law also the expected number in the queue and in the system, respectively.

Example ($M/M/1$ queue) For the $M/M/1$ queue we had (see Section 6.4.1)

$$W_q = W - \frac{1}{\mu} = \frac{1}{\mu} \cdot \frac{1}{1-\rho} - \frac{1}{\mu} = \frac{\rho}{1-\rho} \cdot \mathrm{E}(B),$$

and so (2) is exact in this case.

Example ($M/G/1$ queue) For the $M/G/1$ queue we had (see Section 6.5)

$$W_q = \frac{\lambda s^2}{2(1-\rho)} = \frac{\lambda(\sigma^2 + \tau^2)}{2(1-\rho)} = \frac{\rho}{1-\rho} \cdot \frac{1 + c_B^2}{2} \cdot \mathrm{E}(B),$$

where we used $\rho = \lambda\tau, \mathrm{E}(B) = \tau$ and $c_B^2 = \sigma^2/\tau^2$, and so (2) is also exact in this case.

In the case that the inter-arrival times are not exponentially distributed, and hence the arrival process is not a Poisson process, (2) is really an approximation. To illustrate this, in the next table we compare the approximation with exact results in the case of Erlang

	c_A^2		
	1/4	1/3	1/2
(2)	2.250	2.625	3.375
exact	2.076	2.466	3.250

Table 1: Comparison of the approximation of the mean waiting time in the queue with exact results

distributed interarrival and processing times. We have chosen $\rho = 0.9$, $E(B) = 1$, $c_B^2 = 1/4$ and $c_A^2 = 1/4$, 1/3 and 1/2, respectively.

Example In a workstation jobs are delivered at a rate of one job every 8 hours. The standard deviation of the time between successice delivery times is 4 hours (as past records indicate). The average production time of a job is 6 hours with a standard deviation of 2 hours. Using (2), the prediction for the waiting time W of a job is roughly 9.25 hours. What would be the reduction in the waiting time if the deliveries could be made more regular, for instance with a standard deviation of only one hour? In this case the waiting time is reduced to approximately 7 hours, which is an improvement of almost 25%. Hence, simple formulas like (2) may be used to quickly determine rough-cut answers for the mean waiting time.

Departure process of the $G/G/1$ system

The $G/G/1$ system forms the building block in the development of approximation techniques for the analysis of production networks. In a network the departures from one machine are arrivals to another machine. Hence, it is useful to be able to characterize the departure process of the $G/G/1$ system. In general, the interdeparture times may not be independent, but as an approximation, we will act as if they are. By flow conservation, the departure rate is equal to the arrival rate, so the mean of the inter-departure time equals the mean of the inter-arrival time. A simple approximation for the squared coefficient of variation of the inter-departure time, c_D^2, is given by (see, e.g., [2])

$$c_D^2 \approx (1 - \rho^2)c_A^2 + \rho^2 c_B^2.$$

This approximation is intuitively appealing: under light load conditions (ρ close to 0), c_D^2 is approximately equal to c_A^2 and under heavy load conditions (ρ close to 1) it is approximately equal to c_B^2.

3.2 The $G/G/s$ queue

For the mean waiting time in the queue in the $G/G/1$ system we have that

$$W_q^{G/G/1} \approx \frac{c_A^2 + c_B^2}{2} \cdot W_q^{M/M/1},$$

where we compare the $G/G/1$ queue and the $M/M/1$ queue with the same mean inter-arrival time and the same mean service time. Similarly, for the $G/G/s$ queue one can use the approximation

$$E(W_q^{G/G/s}) \approx \frac{c_A^2 + c_B^2}{2} \cdot E(W_q^{M/M/s}).$$

If we substitute the expression for $E(W_q^{M/M/s})$ from Section 6.4.2 we get

$$E(W_q^{G/G/s}) \approx \frac{p_W}{1-\rho} \cdot \frac{c_A^2 + c_B^2}{2} \cdot \frac{E(B)}{s},$$

where $\rho = \lambda E(B)/s$ and p_W the probability of waiting in the corresponding $M/M/s$ queue.

Departure process of the $G/G/s$ queue

In approximating the departure process of the $G/G/s$ system we again act as if the inter-departures times are independent. The mean interdeparture time is equal to the mean inter-arrival time (by conservation of flow) and the squared coefficient of variation c_D^2 is approximated by

$$c_D^2 \approx 1 + (1-\rho^2)(c_A^2 - 1) + \frac{\rho^2(c_B^2 - 1)}{\sqrt{s}}.$$

For $s = 1$ this approximation is the same as the one proposed for the $G/G/1$ system, and for the $M/M/s$ it yields $c_D^2 = 1$ (which agrees with the property that the output process of the $M/M/s$ queue is again a Poisson process).

Exercises:

1. The Exponential distribution with parameter λ, the Erlang distribution with parameters k and λ and the Hyperexponential distribution with parameters k, λ and p play an important role in queueing models (see Appendix B.2 of Kulkarni's book, where several commonly used continuous random variables with their distribution are introduced). In this exercise you have to show that the coefficient of variation of these three distributions is equal to one, smaller than one and greater than one, respectively.

   (a) Let X be an Exp(λ) random variable. Show that $c_X = 1$.

   (b) Let X be an Erl(k, λ) random variable. Show that $c_X < 1$.

   (c) Let X be a Hex(k, λ, p) random variable, with $k = 2, \lambda = [\lambda_1, \lambda_2]$ and $p = [p_1, p_2]$. Show that $c_X > 1$.
   Hints for part (c):
   - Compute the expectations E(X) and E(X^2).
   - Show that Var(X) - (E(X))$^2 > 0$, using Var(X) = E(X^2) - (E(X))^2.
   - Show that $c_X > 1$, using the previous result.

2. In a $G/G/1$ queue the inter-arrival time is uniformly distributed (see Appendix B.2 of Kulkarni's book) on the interval $[1, 5]$ days. The service time is uniformly distributed on the interval $[1, 3]$ days. Compute the exact value or an approximation for:

   (a) the *throughput* of this queueing system (see the remark at the end of this exercise),

   (b) the mean service time τ and the traffic intensity ρ,

   (c) the expected time in the queue W_q and the expected time in the system W,

   (d) the mean number of customers at the server, in the queue and in the system,

   (e) the mean inter-departure time and the coefficient of variation of the interdeparture-time.

   Remark: The *throughput* or *departure rate* is the long-run average number of customers that leaves the queueing system per unit of time. This concept is analogous to the *arrival rate* of a queueing system, the long-run average number of customers that enter the system per unit of time.

3. In a $G/G/s$ queueing system the inter-arrival time is uniformly distributed on the interval $[0, \frac{1}{2}]$ days. The service time T (in days) has probability density function $f_T(t) = 6t(1-t),\ 0 \leq t \leq 1$. The number of servers $s = 3$. Compute the exact value or an approximation for:

   (a) the *throughput* of this queueing system,

   (b) the mean service time τ and the traffic intensity ρ,

   (c) the expected time in the queue W_q and the expected time in the system W,

   (d) the mean number of customers at the server, in the queue and in the system,

   (e) the mean inter-departure time and the coefficient of variation of the interdeparture-time.

4 Closed networks of queues

In Section 6.7 we looked at queueing networks with *free inflow of jobs*. Arrival processes of jobs were modelled by Poisson processes, independent of the state of the network. In this section we will look at networks for which the nature of the input process is different: there is only new input when a finished job leaves the network. For example, one may think of a production system where jobs are transported through the system on *pallets*. Typically, since pallets are expensive, the number of available pallets is limited. When the production of a job is completely finished, the job is removed from the pallet and a new job is attached to the pallet and released in the network. In this way, the number of circulating jobs in the network remains constant over time. Networks with a fixed number of jobs in the system are called *closed networks*. In the following subsection we first look at a closed network model consisting of only two single-server stations.

4.1 Closed network with two single-server stations

Consider the following model:

- The network consists of two single-server stations, station 1 and station 2.
- Service times of customers at station i are iid $\text{Exp}(\mu_i)$ random variables, $i = 1, 2$.
- Customers finishing service at station i move to station j with probability $p_{i,j}$, independently of each other. The routing matrix

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} \\ p_{2,1} & p_{2,2} \end{pmatrix}$$

  is a stochastic matrix, i.e., $p_{i,j} \geq 0$ for all i and j and $p_{i,1} + p_{i,2} = 1$ for all i. In contrary to an open network it is not possible that customers leave the network.
- The total number of customers in the network is fixed and equal to K.

First of all, remark that we do not have to bother about stability in this case because the number of customers in the system is kept constant. Therefore, we directly study the limiting behavior of the network. Denoting with $X_i(t)$ the number of customers in station i at time t, then $(X_1(t), X_2(t))$ is a CTMC with state space $S = \{(k_1, k_2) : k_1 \geq 0, k_2 \geq 0, k_1 + k_2 = K\}$.

Theorem 4.1 *The limiting distribution of the CTMC $(X_1(t), X_2(t))$ is given by*

$$\begin{aligned} p(k_1, k_2) &= \lim_{t\to\infty} P(X_1(t) = k_1, X_2(t) = k_2) \\ &= C \cdot \left(\frac{1/p_{1,2}}{\mu_1}\right)^{k_1} \cdot \left(\frac{1/p_{2,1}}{\mu_2}\right)^{k_2} \end{aligned}$$

for $(k_1, k_2) \in S$. Here, C is a normalization constant and the factors $v_1 = 1/p_{1,2}$ and $v_2 = 1/p_{2,1}$ are relative visiting frequencies of station 1 and station 2, respectively. Relative visiting frequencies $v = (v_1, v_2)$ are non-unique solutions of the set of traffic equations $v = vP$.

Proof: Balancing the flow between the set of states $\{(0,K),\ldots,(k_1,K-k_1)\}$ and the set of states $\{(k_1+1,K-k_1-1),\ldots,(K,0)\}$ yields

$$p_{1,2}\mu_1 p(k_1+1,K-k_1-1) = p_{2,1}\mu_2 p(k_1,K-k_1),$$

for $k_1 = 0,1,2,\ldots,K-1$. Hence, we have

$$\begin{aligned} p(1,K-1) &= \left(\frac{p_{2,1}\mu_2}{p_{1,2}\mu_1}\right) p(0,K), \\ p(2,K-2) &= \left(\frac{p_{2,1}\mu_2}{p_{1,2}\mu_1}\right)^2 p(0,K), \end{aligned}$$

and in general

$$p(k_1,K-k_1) = \left(\frac{p_{2,1}\mu_2}{p_{1,2}\mu_1}\right)^{k_1} p(0,K).$$

We conclude that

$$\begin{aligned} p(k_1,K-k_1) &= \left(\frac{p_{2,1}\mu_2}{p_{1,2}\mu_1}\right)^{k_1} p(0,K) \\ &= p(0,K)(p_{2,1}\mu_2)^K \cdot \left(\frac{1/p_{1,2}}{\mu_1}\right)^{k_1} \cdot \left(\frac{1/p_{2,1}}{\mu_2}\right)^{K-k_1}, \end{aligned}$$

i.e., we have

$$p(k_1,k_2) = C \cdot \left(\frac{1/p_{1,2}}{\mu_1}\right)^{k_1} \cdot \left(\frac{1/p_{2,1}}{\mu_2}\right)^{k_2}$$

for $(k_1,k_2) \in S$ and with C a normalization constant. Denoting with v_j the visiting frequency (i.e. the average number of visits per time unit) at station j, the traffic equations (cf. formula (6.45)) in this case are given by

$$\begin{aligned} v_1 &= v_1 p_{1,1} + v_2 p_{2,1}, \\ v_2 &= v_1 p_{1,2} + v_2 p_{2,2}. \end{aligned}$$

Using $p_{1,1} + p_{1,2} = 1$ (resp. $p_{2,1} + p_{2,2} = 1$), both equations reduce to the same equation

$$v_1 p_{1,2} = v_2 p_{2,1}.$$

So clearly, $v_1 = 1/p_{1,2}$, $v_2 = 1/p_{2,1}$ is a solution of the traffic equations (but $v_1 = 1$, $v_2 = p_{1,2}/p_{2,1}$ would be another solution). Because of the non-uniqueness of the solution of the traffic equations, v_1 and v_2 are called *relative visiting frequencies*.

4.2 Closed network with N single-server stations

The result of Theorem 4.1 can be extended to networks consisting of an arbitrary number, say N, single-server stations. Let the service times at station i be iid $\text{Exp}(\mu_i)$ random variables, $i = 1, 2, \ldots, N$ and denote the routing matrix by

$$P = \begin{pmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & & \vdots \\ p_{N,1} & \cdots & p_{N,N} \end{pmatrix}.$$

Furthermore, let $X_i(t)$ again be the number of customers in station i at time t, and, finally, denote with S the set of possible states, i.e.,

$$S = \{(k_1, \ldots, k_N) : k_1 \geq 0, \ldots, k_N \geq 0, k_1 + \cdots + k_N = K\}.$$

Without proof we state the following theorem.

Theorem 4.2 *The limiting distribution of the CTMC $(X_1(t), \ldots, X_N(t))$ is given by*

$$\begin{aligned} p(k_1, \ldots, k_N) &= \lim_{t\to\infty} P(X_1(t) = k_1, \ldots, X_N(t) = k_N) \\ &= C \left(\frac{v_1}{\mu_1}\right)^{k_1} \cdots \left(\frac{v_N}{\mu_N}\right)^{k_N} \end{aligned}$$

for $(k_1, \ldots, k_N) \in S$. Here, C is a normalization constant and the factors $v_1, \ldots, v_N$ are relative visiting frequencies of the different stations. Relative visiting frequencies $v = (v_1, \ldots, v_N)$ are non-unique solutions of the set of traffic equations $v = vP$.

Remark 4.3 *The state space of the CTMC $(X_1(t), \ldots, X_N(t))$ is given by the set S. You can show that the total number of states in S is given by $\binom{K+N-1}{N-1}$, so it becomes very big for already moderate values of K and N (e.g. for $K = 10$ and $N = 5$ we have $\binom{K+N-1}{N-1} = \binom{14}{4} = 1001$). As a consequence, it is not so easy to compute the normalization constant C. Simply adding the products $(\frac{v_1}{\mu_1})^{k_1} \cdots (\frac{v_N}{\mu_N})^{k_N}$ over all possible states $(k_1, \ldots, k_N) \in S$ will lead to numerical complications when the state space is large. However, efficient algorithms for the computation of the normalization constant have been developed (e.g. Buzen's convolution algorithm [1]).*

References

[1] J.P. Buzen, *Computational algorithms for closed queueing networks with exponential servers*, Communications of the ACM, 16 (1973), 527–531.

[2] P.J. Kuehn, *Approximate analysis of general queueing networks by decomposition*, IEEE Trans. Comm., 27 (1979), 113–126.

[3] J.G. Shanthikumar, J.A. Buzacott, *On the approximations to the single-server queue*, Internat. J. Prod. Res., 18 (1980), pp. 761–773.

Exercises:

1. A hospital has equipment for making MRI-scans. Before an MRI-scan of a patient can be made, the patient has to go through several preliminary investigations and tests. The hospital wants to make optimal use of the expensive MRI-equipment and has decided to work at any time with a constant total number of K patients in the process.
   A patient, who starts with the preliminary investigations and tests, never needs to wait; for each patient an employee is available. After these investigations and tests are done, patients go to the room with the scan-equipment. Here the patients will have to await their turn, because they are treated one by one in order of arrival.
   The hospital wants to get insight in the traffic intensity at the scan-equipment, in the mean waiting time of a patient, in the mean total time a patient spends in the system, etcetera. The process is modelled as a closed network (a constant number of K patients/customers) with two stations 1 and 2. Station 1 is the phase of the preliminary investigations and tests, station 2 is the phase of waiting for and actually making of the MRI-scan. The number of "servers" in station 1 is always equal to the number of patients in the station, while station 2 is a single-server station. The service times in the stations 1 and 2 are the times needed for the investigations and tests and the actually making of the MRI-scan. The service times in station i are iid $\text{Exp}(\mu_i)(i = 1, 2)$ random variables. Patients always start in station 1, go to station 2 after finishing the investigations, and leave the system after making the MRI-scan. Each time a patient leaves the system, a new patient is admitted (there are always patients available).
   Define $X_i(t)$ as the number of patients in station i at time t, then $(X_1(t), X_2(t))$ is a CTMC with state space $S = \{(k_1, k_2) : k_1 \geq 0, k_2 \geq 0, k_1 + k_2 = K\}$.

   (a) Give the rate diagram of this network.

   (b) Determine for $k_1 = 0, 1, \ \ldots \ , K - 1$ the balance equation that describes the balance of the flow between the set of states $\{(0, K), \ \ldots \ , (k_1, K - k_1)\}$ and the set of states $\{(k_1 + 1, K - k_1 - 1), \ \ldots \ , (K, 0)\}$.

   The limiting distribution of this CTMC $(X_1(t), X_2(t))$ is given by

$$\begin{aligned} p(k_1, k_2) &= \lim_{t\to\infty} P(X_1(t) = k_1, X_2(t) = k_2) \\ &= C \cdot \left(\frac{1}{k_1!}\right) \cdot \left(\frac{v_1}{\mu_1}\right)^{k_1} \cdot \left(\frac{v_2}{\mu_2}\right)^{k_2} \end{aligned}$$

   for $(k_1, k_2) \in S$. Here, C is a normalization constant and the factors v_1 and v_2 are relative visiting frequencies of station 1 and station 2, respectively. Relative visiting frequencies $v = (v_1, v_2)$ are non-unique solutions of the set of traffic equations $v = vP$.

   (c) Show that $v_1 \ : \ v_2 = 1 \ : \ 1$ and that the formula for $p(k_1, k_2)$ with these values for v_1, v_2 satisfies the balance equations.

Let $K = 5, \dfrac{1}{\mu_1} = 2.5$ hours, $\dfrac{1}{\mu_2} = 0.5$ hours, then the normalization constant $C = \dfrac{384}{1097}$.

(d) Compute the probability $p_1(k_1)$ of k_1 ($k_1 = 0, 1, \ldots, K$) patients in station 1.

(e) Determine for station 1 and station 2

- the expected number of patients,
- the traffic intensity per employee and at the scan equipment,
- the throughput of the station,
- the mean waiting time of a patient in the queue.

(f) Compute the throughput of the whole system and the mean time spent in the whole system by an arbitrary patient.

2. A car company has divided the activities for maintenance and reparation of cars into three categories; for each category there is a separate workshop. The company has a "quick-service station" (QS) for small maintenance activities (refresh oil, etc.) and small reparations (new tyres, new exhaust pipe, etc.). Next, there is a repair shop (RS) for all great maintenance and repair activities. Finally, there is a sheet metal and paint workshop (SP) for damages at the car body. Car owners, who bring their car to the car company for maintenance or reparation, check in at the reception (RE). The employees in the reception discuss the problems with the customers and decide to which workshop a car will go. They also do the office work.
The management of the car company wants to get a better insight in the traffic intensity in each of the workshops, in the total mean time that a car is in the garage for maintenance or reparation, etc. Therefore a model, based on a closed network of queues, is formulated. ¿From experience it is known that of all cars, arriving for maintenance or reparation at the reception of the company, initially 50% goes to QS, 40% goes to RS and 10% goes to SP. During the treatment of cars in the QS it turns out in 20% of the cases that the technical problems with the car are greater than initially thought and in such cases the car is still brought to th RS; the other 80% of the cars in the QS leaves the company after the treatment. From all cars in the RS 40% goes (after some preparing activities) to the SP, 60% leaves the company after the treatment. All cars in the SP go after the treatment in the SP to the RS. At any time the number of cars in the garage is kept constant and equal to $K = 6$. Each time a car leaves the garage, a new car is admitted (there are always cars available for maintenance or reparation). The RE, QS, RS, SP are considered as single-server stations 1, 2, 3 and 4. The service times in station i are iid $\text{Exp}(\mu_i)(i = 1, 2, 3, 4)$ random variables. The mean service times are 0.25, 0.5, 1 and 2 hours, respectively. The service discipline in each station is FCFS.

(a) Show that the relative visiting frequencies $v = (v_1, v_2, v_3, v_4)$ of the stations 1, 2, 3 and 4 satisfy $v_1 : v_2 : v_3 : v_4 = 4 : 2 : 4 : 2$.

(b) Determine the limiting distribution $p(k_1, k_2, k_3, k_4)$, where k_i is the number of cars in station i with $k_i \geq 0, (i = 1, 2, 3, 4)$ and $k_1 + k_2 + k_3 + k_4 = 6$.
Remark: using the values for v_1, v_2, v_3, v_4 from (a) and using hours as unit of time for the mean service times, the normalization constant $C = 1/46119$.

(c) The probability of $p_i(k_i)$ cars in station i $(i = 1, 2, 3, 4)$ is given by

k_i	$p_1(k_1)$	$p_2(k_2)$	$p_3(k_3)$	$p_4(k_4)$
0	0.7895	0.7895	0.1578	0.1578
1	0.1678	0.1678	0.1577	0.1577
2	0.0345	0.0345	0.1572	0.1572
3	0.0068	0.0068	0.1554	0.1554
4	0.0012	0.0012	0.1499	0.1499
5	0.0002	0.0002	0.1332	0.1332
6	0.0000	0.0000	0.0888	0.0888

Explain why the probabilities are the same for, respectively, stations 1 and 2 and for stations 3 and 4.

(d) Compute for each station

- the traffic intensity,
- the throughput (repairs/hour).

(e) The mean number of cars in station 1 and 2 is 0.2633. Determine the mean number of cars in the RS and SP.

(f) Compute for each station

- the mean waiting time of a car in the queue,
- the mean time of a car in the station.

(g) Compute the throughput of the whole garage system and the mean time spent in that system by an arbitrary car.

3. Consider a closed network of queues with a constant number of $K = 3$ customers and with four single-server stations 1, 2, 3, 4. Let the service times at station i be iid $\text{Exp}(\mu_i)$ random variables, $i = 1, 2, 3, 4$ and denote the routing matrix by

$$P = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{3}{5} & 0 \\ \frac{3}{5} & 0 & 0 & \frac{2}{5} \\ \frac{2}{5} & 0 & 0 & \frac{3}{5} \\ \frac{3}{5} & 0 & 0 & \frac{2}{5} \end{pmatrix}.$$

The mean service times in stations 1, 2, 3 and 4 are 33, 165, 55 and 45 minutes, respectively. The service discipline is FCFS.

(a) Show that the limiting probability $p(k_1, k_2, k_3, k_4) = \dfrac{1}{20}$, where k_i is the number of customers in station i with $k_i \geq 0, (i = 1, 2, 3, 4)$ and $k_1 + k_2 + k_3 + k_4 = 3$.

(b) Determine the probability $p_1(k_1)$ of k_1 $(k_1 = 0, 1, 2, 3)$ customers in station 1.

(c) What is the traffic intensity ρ_i in station i $(i = 1, 2, 3, 4)$?

4. Consider a closed network of queues with four single-server stations 1, 2, 3 and 4. The total number of customers in the network is kept constant, $K = 5$. A new customer enters the network at station 1 with probability 4/9 and at station 2 with probability 5/9. After a service in station 1, a customer is sent back to station 1 with probability 1/3 and is sent on to station 2 with probability 2/3. After a visit at station 2 a customer goes with probability 1 to station 3. A customer who is served at station 3, is sent back to station 3 with probability 1/4 and is sent on to station 4 with probability 3/4. After a visit at station 4 a customer leaves the system and a new customer is admitted tot the system. The service times at station i are iid $\text{Exp}(\mu_i)$ random variables, $i = 1, 2, 3, 4$. The mean service times are 8, 5, 3 and 4 minutes, respectively. The service discipline is FCFS.

   (a) Let v_1, v_2, v_3, v_4 be the relative visiting frequencies. Show that the *relative traffic intensities* $v_1/\mu_1, v_2/\mu_2, v_3/\mu_3, v_4/\mu_4$ are in the proportion of $4/3 : 5/4 : 1 : 1$.

   The probability that all $K = 5$ customers are in station 4 is equal to 0.009.

   (b) Determine the limiting probability $p(k_1, k_2, k_3, k_4)$, where k_i is the number of customers in station i with $k_i \geq 0, (i = 1, 2, 3, 4)$ and $k_1 + k_2 + k_3 + k_4 = 5$.

   The probability of $p_i(k_i)$ customers in station i $(i = 1, 2, 3, 4)$ is given by

k_i	$p_1(k_1)$	$p_2(k_2)$	$p_3(k_3)$	$p_4(k_4)$
0	0.282	0.327	0.462	0.462
1	0.245	0.258	0.272	0.272
2	0.199	0.190	0.151	0.151
3	0.146	0.127	0.075	0.075
4	0.090	0.071	0.031	0.031
5	0.037	0.027	0.009	0.009

   The mean numbers of customers in station 1, 2, 3 and 4 are equal to 1.626, 1.438, 0.968, and 0.968, respectively.

   (c) Compute the throughputs of the four stations.

   (d) Determine the total mean time an arbitrary customer spends in the network.

   (e) A customer visits twice station 1 and twice station 3. What is the total mean time of this customer in the network?