Basic elements for
Performance Evaluation of Computer Systems,
Distributed Systems and Communication Networks

COMPSCI 655 – Performance Evaluation, Fall 2018

Class: Tuesday, Thursday at 10am, Room CompSci 142

Professor: Philippe Nain
Senior Research Scientist at Inria

March 28, 2018

©August 2016 by Philippe Nain
# Contents

1 Foreword 4

2 Markov Chains 4
   2.1 Discrete-Time Markov Chain 4
      2.1.1 Example of a discrete-time Markov chain 8
      2.1.2 A communication line with error transmissions 10
   2.2 Continuous-Time Markov Chain 12
      2.2.1 A case study: Markov modulated arrivals 16

3 Birth and Death Process 17

4 Absorbing Markov Chains 18
   4.1 Discrete-time absorbing Markov chains 18
   4.2 Continuous-time absorbing Markov chains 21

5 Queueing Theory 22
   5.1 The M/M/1 queue 22
   5.2 The M/M/1/K queue 23
   5.3 The M/M/c queue 24
   5.4 The M/M/c/c queue 25
   5.5 The repairman model 25
   5.6 Little's formula 26
   5.7 Comparing different multiprocessor systems 27
   5.8 The M/G/1 FIFO queue 28
      5.8.1 Mean queue-length and mean response time 28

6 Priority Queues 30
   6.1 Non-preemptive priority 31
   6.2 Preemptive-resume priority 33

7 Single-Class Queueing Networks 34
   7.1 Networks of Markovian queues: open Jackson network 34
   7.2 Networks of Markovian queues: closed Jackson networks 39
      7.2.1 The convolution algorithm 40
      7.2.2 Performance measures from normalization constants 41
   7.3 The central server model 42

8 Multiclass Queueing Networks 43
   8.1 Multiclass open/closed/mixed Jackson queueing networks 43
   8.2 The multiclass PS queue 45
   8.3 The multiclass LIFO queue 48
   8.4 The multiclass IS server queue 49
   8.5 BCMP networks 50
   8.6 Kelly Networks 53

9 Queueing Theory (continued) 54
   9.1 The FIFO GI/GI/1 queue 54
   9.2 Application: effective bandwidth in multimedia networks 57
      9.2.1 Effective bandwidth for Poisson inputs 57
### A Probability Refresher

A.1 Sample-space, events and probability measure ......................................................... 60
A.2 Combinatorial analysis ................................................................................................. 61
A.3 Conditional probability ............................................................................................... 62
A.4 Random variables ....................................................................................................... 64
A.5 Parameters of a random variable ................................................................................ 66
A.6 Jointly distributed random variables ......................................................................... 67
A.7 Independent random variables .................................................................................. 68
A.8 Conditional expectation ............................................................................................... 68

### B Stochastic Process

B.1 Definitions ................................................................................................................. 69

### C Poisson Process

C ....................................................................................................................................... 70
1 Foreword

The material presented in Chapters 2, 3 and 5 comes mainly from reference [6] below. Chapter 4 has been borrowed from [8]. Chapter 6 and Section 5.8 come from [2]. Part of Chapters 7 and 8 come from [3] and [1], respectively. Section 8.6 is taken from [4]. Sections 9.1 and 9.2 are taken from [7] and [5], respectively.

References


2 Markov Chains

A Markov process\(^1\) is a stochastic process \((X(t), t \in T), X(t) \in E \subset \mathbb{R}\), such that

\[
P(X(t) \leq x | X(t_1) = x_1, \ldots, X(t_n) = x_n) = P(X(t) \leq x | X(t_n) = x_n)
\]

for all \(x_1, \ldots, x_n, x \in E, t_1, \ldots, t_n, t \in T\) with \(t_1 < t_2 < \cdots < t_n < t\).

Intuitively, (1) says that the probabilistic future of the process depends only on the current state and not upon the history of the process. In other words, the entire history of the process is summarized in the current state.

In the following we shall be mostly concerned with discrete-space Markov processes, commonly referred to as Markov chains.

We shall distinguish between discrete-time Markov chains and continuous-time Markov chains.

2.1 Discrete-Time Markov Chain

A discrete-time Markov Chain (abbreviated as D-MC) is a discrete-time (with index set \(N := \{0, 1, \ldots\}\)) discrete-space (with state-space \(E\)) stochastic process \((X(n), n \in N)\) such that for all \(n \geq 0\)

\[
P(X(n+1) = j | X(0) = i_0, X(1) = i_1, \ldots, X(n-1) = i_{n-1}, X(n) = i) = P(X(n+1) = j | X(n) = i)
\]

for all \(i_0, \ldots, i_{n-1}, i, j \in E\).

A D-MC is called a finite-state D-MC if the set \(E\) is finite.

A D-MC is homogeneous if \(P(X(n+1) = j | X(n) = i)\) does not depend on \(n\) for all \(i, j \in E\). If so, we shall write

\[
p_{i,j} = P(X(n+1) = j | X(n) = i) \quad \forall i, j \in E.
\]

\(^1\)A. A. Markov was a Russian mathematician.
$p_{i,j}$ is the one-step transition probability from state $i$ to state $j$, that is the probability to go from state $i$ to state $j$ in exactly one time-step.

Unless otherwise mentioned we shall only consider homogeneous D-MC’s.

Define $P = [p_{i,j}]$ to be the transition matrix of a D-MC, namely,

$$P = \begin{pmatrix}
    p_{0,0} & p_{0,1} & \cdots & p_{0,j} & \cdots \\
    p_{1,0} & p_{1,1} & \cdots & p_{1,j} & \cdots \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    p_{i,0} & p_{i,1} & \cdots & p_{i,j} & \cdots \\
    \vdots & \vdots & \ddots & \vdots & \vdots
\end{pmatrix}$$

(3)

We must have

$$p_{i,j} \geq 0 \quad \forall i, j \in \mathcal{E}$$

(4)

$$\sum_{j \in \mathcal{E}} p_{i,j} = 1 \quad \forall i \in \mathcal{E}.$$  

(5)

Equation (5) is a consequence of axiom (b) of a probability measure (see Section A.1) and says that the sum of the elements in each row is 1. A matrix satisfying (4) and (5) is called a stochastic matrix.

If the state-space $\mathcal{E}$ is finite (say, with $k$ states) then $P$ is a $k$-by-$k$ matrix; otherwise $P$ has infinite dimension.

**Example 1.** Consider a sequence of Bernoulli trials in which the probability of success (S) on each trial is $p$ and of failure (F) is $q$, where $p + q = 1$, $0 < p < 1$. Let the state of the process at trial $n$ (i.e., $X(n)$) be the number of uninterrupted successes that have been completed at this point. For instance, if the first 5 outcomes were SFSSF then $X(0) = 1$, $X(1) = 0$, $X(2) = 1$, $X(3) = 2$ and $X(4) = 0$. The transition matrix is given by

$$P = \begin{pmatrix}
    q & p & 0 & 0 & 0 & \cdots \\
    q & 0 & p & 0 & 0 & \cdots \\
    q & 0 & 0 & p & 0 & \cdots \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}$$

The state 0 can be reached in one transition from any state while the state $i + 1$, $i \geq 0$, can only be reached from the state $i$ (with the probability $p$). Observe that this D-MC is clearly homogeneous.

We now define the $n$-step transition probabilities $p_{i,j}^{(n)}$ by

$$p_{i,j}^{(n)} = P(X(n) = j \mid X(0) = i)$$

(6)

for all $i, j \in \mathcal{E}$, $n \geq 0$. $p_{ij}^{(n)}$ is the probability of going from state $i$ to state $j$ in $n$ steps.

**Proposition 1** (Chapman-Kolmogorov equation). For all $n \geq 0$, $m \geq 0$, $i, j \in \mathcal{E}$, we have

$$p_{i,j}^{(n+m)} = \sum_{k \in \mathcal{E}} p_{i,k}^{(n)} p_{k,j}^{(m)}$$

(7)

or, in matrix notation,

$$P^{(n+m)} = P^{(n)} P^{(m)}$$

(8)

where $P^{(n)} := [p_{i,j}^{(n)}]$. Therefore,

$$P^{(n)} = P^n \quad \forall n \geq 1$$

(9)

where $P^n$ is the $n$-th power of the matrix $P$. 
The Chapman-Kolmogorov equation merely says that if we are to travel from state \( i \) to state \( j \) in \( n + m \) steps then we must do so by first traveling from state \( i \) to some state \( k \) in \( n \) steps and then from state \( k \) to state \( j \) in \( m \) more steps.

The proof of Proposition 1 goes as follows. We have
\[
\pi^{(n+m)}_{i,j} = P(X(n + m) = j \mid X(0) = i)
\]
\[
= \sum_{k \in \mathcal{E}} P(X(n + m) = j, X(n) = k \mid X(0) = i)
\]
by axiom (b) of a probability measure
\[
= \sum_{k \in \mathcal{E}} P(X(n + m) = j \mid X(0) = i, X(n) = k)
\]
\[
\times P(X(n) = k \mid X(0) = i)
\]
by the generalized Bayes’ formula
\[
= \sum_{k \in \mathcal{E}} P(X(n + m) = j \mid X(n) = k) P(X(n) = k \mid X(0) = i)
\]
from Markov property (2)
\[
= \sum_{k \in \mathcal{E}} \pi^{(m)}_{k,j} \pi^{(n)}_{i,k}
\]

since the D-MC is homogeneous, which proves (7) and (8).

Let us now establish (9). Since \( P^{(1)} = P \), we see from (8) that \( P^{(2)} = P^2 \), \( P^{(3)} = P^{(2)} P = P^3 \) and, more generally, that \( P^{(n)} = P^n \) for all \( n \geq 1 \). This concludes the proof.

Example 2. Consider a communication system that transmits the digits 0 and 1 through several stages. At each stage, the probability that the same digit will be received by the next stage is 0.75. What is the probability that a 0 that is entered at the first stage is received as a 0 at the fifth stage?

We want to find \( p^{(5)}_{0,0} \) for a D-MC with transition matrix \( P \) given by
\[
P = \begin{pmatrix}
0.75 & 0.25 \\
0.25 & 0.75
\end{pmatrix}.
\]

From Proposition 1 we know that \( p^{(5)}_{0,0} \) is the \((1,1)\)-entry of the matrix \( P^5 \). We find \( p^{(5)}_{0,0} = 0.515625 \) (compute \( P^2 \), then compute \( P^4 \) as the product of \( P^2 \) by itself, and finally compute \( P^5 \) as the product of matrices \( P^4 \) and \( P \)).

So far, we have only been dealing with conditional probabilities. For instance, \( p^{(n)}_{i,j} \) is the probability that the system is in state \( j \) at time \( n \) given it was in state \( i \) at time 0. We have shown in Proposition 1 that this probability is given by the \((i,j)\)-entry of the power matrix \( P^n \). What we would like to do now is to compute the unconditional probability that the system is in state \( j \) at time \( n \), namely, we would like to compute \( \pi_n(i) := P(X(n) = i) \).

This quantity can only be computed if we provide the initial distribution function of \( X(0) \), that is, if we provide \( \pi_0(i) = P(X(0) = i) \) for all \( i \in \mathcal{E} \), where of course \( \sum_{i \in \mathcal{E}} \pi_0(i) = 1 \).

In that case, we have from Bayes’ formula
\[
P(X(n) = j) = \sum_{i \in \mathcal{E}} P(X(n) = j \mid X(0) = i) \pi_0(i)
\]
\[
= \sum_{i \in \mathcal{E}} p^{(n)}_{i,j} \pi_0(i)
\]
from Proposition 1 or, equivalently, in matrix notation:

Proposition 2. For all \( n \geq 1 \),
\[
\pi_n = \pi_0 P^n.
\]
where \( \pi_m := (\pi_m(0), \pi_m(1), \ldots) \) for all \( m \geq 0 \). From (10) we deduce that (one can also obtain this result directly)
\[
\pi_{n+1} = \pi_n P \quad \forall n \geq 0.
\]
Assume that the limiting state distribution function \( \lim_{n \to \infty} \pi_n(i) \) exists for all \( i \in \mathcal{E} \). Call it \( \pi(i) \) and let \( \pi = (\pi(0), \pi(1), \ldots) \).

How can one compute \( \pi \)? Owing to (10) a natural answer is “by solving the system of linear equations defined by \( \pi = \pi P \)” to which one should add the normalization condition \( \sum_{i \in \mathcal{E}} \pi(i) = 1 \) or, in matrix notation, \( \pi \mathbf{1} = 1 \), where \( \mathbf{1} \) is the column vector where every component is 1.

We shall now give conditions under which the above results hold (i.e., \( (\pi_n(0), \pi_n(1), \ldots) \) has a limit as \( n \) goes to infinity and this limit solves the system of equations \( \pi = \pi P \) and \( \pi \mathbf{1} = 1 \)).

To do so, we need to introduce the notion of communication between the states. We shall say that a state \( j \) is reachable from a state \( i \) if \( p_{i,j}^{(n)} > 0 \) for some \( n \geq 1 \). If \( j \) is reachable from \( i \) and if \( i \) is reachable from \( j \) then we say that \( i \) and \( j \) communicate, and write \( i \leftrightarrow j \).

A D-MC is irreducible if \( i \leftrightarrow j \) for all \( i,j \in \mathcal{E} \).

For every state \( i \in \mathcal{E} \), define the integer \( d(i) \) as the largest common divisor of all integers \( n \) such that \( p_{i,i}^{(n)} > 0 \). If \( d(i) = 1 \) then the state \( i \) is aperiodic.

A D-MC chain is aperiodic if all states are aperiodic.

We have the following fundamental result of D-MC theory.

**Proposition 3** (Invariant probability measure of a D-MC). If a homogeneous D-MC with transition matrix \( P \) is irreducible and aperiodic, and if the system of equations

\[
\begin{align*}
\mathbf{x} &= \mathbf{xP} \\
\mathbf{x} \cdot \mathbf{1} &= 1
\end{align*}
\]

has a unique strictly positive solution (i.e., for all \( i \in \mathcal{E} \), \( x(i) \), the \( i \)th element of the row vector \( \mathbf{x} \), is strictly positive) denoted by \( \pi = (\pi(j), j \in \mathcal{E}) \), then

\[ \pi(j) = \lim_{n \to \infty} P(X(n) = j) \quad (12) \]

for all \( i \in \mathcal{E} \), regardless of the initial state \( X(0) \).

Observe that the convergence in (12) does not depend on the initial state \( i \). In other words, for any initial state the probability distribution of \( X(n) \) will converge to \( \pi(\cdot) \) as \( n \) goes to infinity.

We shall not prove this result. The equation \( \pi = \pi P \) is called the invariant equation and \( \pi \) is usually referred to as the invariant probability.

Let us briefly comment on the conditions that the D-MC has to be irreducible and aperiodic for Proposition 3 to hold.

**Irreducibility:** consider a three-state D-MC (with states 1, 2 and 3) with transition matrix

\[
P = \begin{pmatrix}
0.3 & 0.7 & 0 \\
0.1 & 0.9 & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

In words, if the chain is in state 3 at time 0 then it will stay in state 3 forever, while if at time 0 it is in state 1 or in state 2 then it will never leave these states. This D-MC is not irreducible since state 1 and 2 cannot be reached from state 3 and conversely. Therefore, it is obvious that the probability to find this D-MC in state 3 depends on whether or not the chain is in state 3 at time 0 (if then this probability is 1 and if not then it is 0).

**Aperiodicity:** consider a two-state D-MC (with state 1 and 2) with transition matrix

\[
P = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}.
\]

In words, the chain alternates between state 1 and 2. Therefore, if the chain is in state 1 at time 0, namely \( X(0) = 1 \), then \( P(X(n) = 2 | X(0) = 1) = 1 \) if \( n \) is odd and \( P(X(n) = 2 | X(0) = 1) = 0 \) if \( n \) is an even integer.
As a consequence \( \lim_{n \to \infty} P(X(2n + 1) = 2 | X(0) = 1) = 1 \) and \( \lim_{n \to \infty} P(X(2n) = 2 | X(0) = 1) = 0 \) thereby showing that \( P(X(n) = 2 | X(0) = 1) \) does not have a limit as \( n \) goes to infinity. These two examples show the necessity for the chain to be irreducible and aperiodic if one wants Proposition 3 to hold.

2.1.1 Example of a discrete-time Markov chain

Consider an homogeneous discrete-time Markov chain (D-MC) on the state-space \( \mathcal{E} = \{1, 2, 3\} \) with transition matrix \( \mathbf{P} \) given by

\[
\mathbf{P} = \begin{pmatrix}
0.2 & 0.2 & 0.6 \\
0 & 0.5 & 0.5 \\
1 & 0 & 0
\end{pmatrix}.
\] (13)

One often represents a D-MC through its transition diagram. In this case it is given by

![Transition Diagram](image)

The first type of question we may want to address is: *given that the chain is in state 1 at time \( t = 0 \), in what state will it be at time \( t = 4 \)?* There is no deterministic answer to this question since, by construction, at time \( t = 4 \) the chain can a priori be in any of the states 1, 2 and 3. A more relevant question is: *given that the chain is in state 1 at time \( t = 0 \), what is the probability that it will be in state \( j \) at time \( t = 4 \)?*

The answer to the latter question is obtained using (10) in Proposition 2, which will allow us to compute \( \pi_4 = (\pi_4(1), \pi_4(2), \pi_4(3)) \), where we recall that is \( \pi_n(i) \) the probability that the chain is in state \( i \in \mathcal{E} \) at time \( t = n \). Here, the initial probability distribution \( \pi_0 \) is given by

\( \pi_0 = (1, 0, 0) \)

since we have specified that the chain is initially in state 1. We find

\[
\pi_1 = (1, 0, 0) \begin{pmatrix}
0.2 & 0.2 & 0.6 \\
0 & 0.5 & 0.5 \\
1 & 0 & 0
\end{pmatrix} = (0.2, 0.2, 0.6)
\]

\[
\pi_2 = (0.2, 0.2, 0.6) \begin{pmatrix}
0.2 & 0.2 & 0.6 \\
0 & 0.5 & 0.5 \\
1 & 0 & 0
\end{pmatrix} = (0.64, 0.14, 0.22)
\]

\[
\pi_3 = (0.64, 0.14, 0.22) \begin{pmatrix}
0.2 & 0.2 & 0.6 \\
0 & 0.5 & 0.5 \\
1 & 0 & 0
\end{pmatrix} = (0.348, 0.198, 0.454)
\]
\[ \pi_4 = (0.348, 0.198, 0.454) \begin{pmatrix} 0.2 & 0.2 & 0.6 \\ 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \end{pmatrix} = (0.5236, 0.1686, 0.3078). \]

In summary, at time \( t = 4 \), the chain will be

- in state 1 with probability 0.5236
- in state 2 with probability 0.1686
- in state 3 with probability 0.3078.

Let us now compute the stationary probabilities. First, note that the chain is clearly irreducible and aperiodic. Hence (cf. Proposition 3), if we find a unique strictly positive solution to the system of equations

\[
\begin{align*}
x_1 &= 0.2x_1 + 1.0x_3 \\
x_2 &= 0.2x_1 + 0.5x_2 \\
x_3 &= 0.6x_1 + 0.5x_3 \\
1 &= x_1 + x_2 + x_3
\end{align*}
\]

then this solution will be the stationary distribution of this Markov chain, that is, \( \lim_{n \to \infty} \pi_n(i) \) will exist and will be given by \( x_i \) for \( i \in E \).

Solving for this system of equations gives

\[ x = \left( \frac{5}{11}, \frac{2}{11}, \frac{4}{11} \right) \approx (0.4545, 0.1818, 0.3636). \]  

(14)

This solution being unique and strictly positive we conclude that this is the stationary probability of the chain, namely, \( \lim_{n \to \infty} \pi_n(1) = \frac{5}{11} \), \( \lim_{n \to \infty} \pi_n(2) = \frac{2}{11} \) and \( \lim_{n \to \infty} \pi_n(3) = \frac{4}{11} \).

In Figure 2 we have simulated the trajectory of the chain for 50 time units (starting in state 3 in red). The first bar is the state of the chain at \( t = 1 \), the second at \( t = 2 \), etc.

![Figure 2: Trajectory after 50 units of time](image)

Note that the chain does not stay more that 1 time unit in state 3 (in red) which is normal since the probability of looping back to state 3 is zero (i.e. \( p_{3,3} = 0 \)). More precisely, upon leaving state 3 (red) the chain always goes to state 1 (green) as \( p_{3,1} = 1 \) by definition of \( P \) in (13). We also observe that the state may stay several units of time in state 1 (green) and in state 2 (blue) but that it has the tendancy to stay more often in state 2 (blue). This is also normal since \( p_{2,2} \) is larger than \( p_{3,3} \).

If we count the number of times that the chain has visited each state we find

- 22 times in state 1, 8 times in state 2, 20 times in state 3

that is, in percentage of time it has stayed

- 44% in state 1, 16% in state 2, 40% in state 3

which is already close to the stationary probability found in (14).

If we observe the same trajectory for 100 units of time we obtain that is the chain has visited

- 45 times state 1, 18 times state 2, 37 times state 3

that is, in percentage of time it has stayed

- 45% in state 1, 18% in state 2, 37% in state 3
which is even closer from (14). There is actually a theoretical result saying that the transient probability (i.e. \( \pi_n \)) converges geometrically fast to the stationary probability (i.e. \( \pi \)) as \( n \to \infty \). This means that the convergence is fast \textit{in terms} of units of time but not necessarily in terms of real time (it may be long when the unit of time is one year and very short if the unit of time is 1 nanosecond).

It should be clear that another experiment will give totally different results in Figures 2 and 3. For instance, starting from state 1 the chain may stay in state 1 for 50 consecutive units of time with the probability \( \frac{50}{100} = 0.5 \). This is highly unlikely (!) but not impossible.

### 2.1.2 A communication line with error transmissions

We consider an infinite stream of packets that arrive at a gateway of a communication network, seeking admittance in the network. For the sake of simplicity we assume that the packets may only arrive in the intervals of time \((n, n + 1)\) for all \( n \geq 0 \) (i.e., we assume that packets do not arrive at times \( n = 0, 1, 2, \ldots \)). Upon arrival, a packet enters a buffer of infinite dimension. Let \( A(n) \in \{0, 1\} \) be the number of packets arriving in the interval of time \((n, n + 1)\). We assume that for each \( n \), \( A(n) \) follows a Bernoulli rv with parameter \( a \), \( 0 < a < 1 \).

We assume that one unit of time is needed to transmit a packet and that transmissions only start at times \( n = 0, 1, 2, \ldots \) provided that the buffer is nonempty (it does not matter which packet is transmitted).

We assume that transmission errors may occur. More precisely, a packet transmitted in any interval of time \([n, n + 1)\) is transmitted in error with the probability \( 0 \leq 1 - p < 1 \). When this happens, the packet is retransmitted in the next time-slot and the procedure is repeated until the transmission is a success eventually (which occurs with probability one since \( 1 - p < 1 \)).

We assume that the rvs \( \{A(n), n \geq 0\} \) are mutually independent rvs, that transmission errors are mutually independent, and further independent of the rvs \( \{A(n), n \geq 0\} \).

Let \( X(n) \) be the number of packets in the buffer at time \( n \). Our objective is to compute \( \pi(i) := \lim_{n \to \infty} P(X(n) = i) \) for all \( i \in \mathbb{N} \) when this limit exists.

We have the following \textit{evolution equation} for this system:

\[
X(n + 1) = A(n) \text{ if } X(n) = 0
\]

\[
X(n + 1) = X(n) + A(n) - D(n) \text{ if } X(n) > 0
\]

for all \( n \in \mathbb{N} \), where \( D(n) \in \{0, 1\} \) gives the number of packet transmitted with success in the interval of time \([n, n + 1)\).

Since the rvs \( A(n) \) and \( D(n) \) are independent of the rvs \( \{A(i), D(i), i = 0, 1, \ldots, n - 1\} \), and therefore of \( X(n) \) from (15)-(16), it should be clear from (15)-(16) that \( \{X(n), n \geq 0\} \) is a D-MC.

We shall however prove this result explicitly.

We have for \( i = 0 \) and for arbitrary \( i_0, \ldots, i_{n-1}, j \in \mathbb{N} \)

\[
P(X(n + 1) = j \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i) = P(A(n) = j \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i)
\]

which is equal to 0 if \( j \geq 2 \), is equal to \( a \) if \( j = 1 \), and is equal to \( 1 - a \) if \( j = 0 \). This shows that \( P(X(n + 1) = j \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i) \) is not a function of \( i_0, \ldots, i_{n-1} \).
Consider now the case when \( i \geq 1 \). We have
\[
P(X(n + 1) = j \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i) = P(A(n) - D(n) = j - i \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i)
\]
from (16)
\[
P(A(n) - D(n) = j - i)
\]
from the independence assumptions.

Clearly, \( P(A(n) - D(n) = j - i) \) in (17) is equal to 0 when \( |j - i| \geq 2 \). On the other hand, routine applications of Bayes formula together with the independence of \( A(n) \) and \( D(n) \) yield \( P(A(n) - D(n) = j - i) \) is equal to \( (1 - a)p \) when \( j = i - 1 \), is equal to \( ap + (1 - a)(1 - p) \) when \( i = j \), and is equal to \( (1 - p)a \) when \( j = i + 1 \).

Again, one observes that \( P(X(n + 1) = j \mid X(0) = i_0, \ldots, X(n - 1) = i_{n-1}, X(n) = i) \) does not depend on \( i_0, \ldots, i_{n-1} \) when \( i \geq 1 \). This proves that \( (X(n), n \geq 0) \) is a D-MC.

From the above we see that the transition matrix \( P = [p_{i,j}] \) of this D-MC is given by:
\[
\begin{align*}
p_{0,0} &= 1 - a \\
p_{0,1} &= a \\
p_{0,j} &= 0 \ \forall j \geq 2 \\
p_{i,i-1} &= (1 - a)p \ \forall i \geq 1 \\
p_{i,i} &= ap + (1 - a)(1 - p) \ \forall i \geq 1 \\
p_{i,i+1} &= a(1 - p) \ \forall i \geq 1 \\
p_{i,j} &= 0 \ \forall i \geq 1, j \notin \{i - 1, i, i + 1\}.
\end{align*}
\]

Let us check that this D-MC is irreducible and aperiodic. Since \( p_{i,i} > 0 \) for all \( i \in \mathbb{N} \) we see that this D-MC is aperiodic.

Let \( i, j \in \mathbb{N} \) be two arbitrary states. We first show that \( j \) is reachable from \( i \). We know that this is true if \( i = j \) since \( p_{i,i} > 0 \). If \( j > i \) then clearly \( p_{i,j}^{(j-i)} \geq a^{j-i}(1-p)^{j-i} > 0 \) if \( i > 0 \) and \( p_{i,j}^{(j-i)} \geq a^{j-i}(1-p)^{j-1} > 0 \) if \( i = 0 \); if \( 0 \leq j < i \) then clearly \( p_{i,j}^{(j-i)} \geq (1-a)^{i-j}p^{i-j} > 0 \). Hence, \( j \) is reachable from \( i \). Since \( i \) and \( j \) can be interchanged, we have in fact established that \( i \leftrightarrow j \). This shows that the D-MC is irreducible since \( i \) and \( j \) are arbitrary states.

We may therefore apply Proposition 3. This result says that we must find a strictly positive solution \( \pi = (\pi(0), \pi(1), \ldots) \) to the system of equations
\[
\begin{align*}
\pi(0) &= \pi(0)(1 - a) + \pi(1)(1 - a)p \\
\pi(1) &= \pi(0)a + \pi(1)(ap + (1 - a)(1 - p)) + \pi(2)(1 - a)p \\
\pi(j) &= \pi(j - 1)a(1 - p) + \pi(j)(ap + (1 - a)(1 - p)) + \pi(j + 1)(1 - a)p, \ \forall j \geq 2
\end{align*}
\]
such that
\[
\sum_{j=0}^{\infty} \pi(j) = 1.
\]

It is easily seen that
\[
\begin{align*}
\pi(1) &= \frac{a}{(1 - a)p} \pi(0) \\
\pi(j) &= \frac{1}{1 - p} \left( \frac{a(1 - p)}{p(1 - a)} \right)^j \pi(0), \ \forall j \geq 2
\end{align*}
\]
satisfies the system of equations (18)-(20) (hint: determine first \( \pi(1) \) and \( \pi(2) \) as functions of \( \pi(0) \) by using equations (18) and (19), then use equation (20) to recursively determine \( \pi(j) \) for \( j \geq 3 \).

We must now compute \( \pi(0) \) such that (21) hold. Introducing the values of \( \pi(j) \) obtained above in equation (21) gives after trivial algebra
\[
\pi(0) \left( 1 + \frac{a}{p(1 - a)} \sum_{j=0}^{\infty} \left( \frac{a(1 - p)}{p(1 - a)} \right)^j \right) = 1.
\]
Fix $r \geq 0$. Recall that the power series $\sum_{j=0}^{\infty} r^j$ converges if and only if $r < 1$. If $r < 1$ then $\sum_{j=0}^{\infty} r^j = 1/(1 - r)$.

Therefore, we see that the factor of $\pi(0)$ in (24) is finite if and only if $a (1 - p)/p (1 - a) < 1$, or equivalently, if and only if $a < p$. If $a < p$ then

$$\pi(0) = \frac{p - a}{p}.$$  

(25)

In summary, we have found a strictly positive solution to the system of equations (18)-(21) if $a < p$. We may therefore conclude from Proposition 3 that $\lim_{n\to\infty} P(X(n) = i)$ exists for all $i \in \mathbb{N}$ if $a < p$, is independent of the initial state, and is given by $\pi(i)$ for all $i \in \mathbb{N}$.

The condition $a < p$ is not surprising, since it simply says that the system is stable if the mean number of arrivals in any interval of time $(n, n+1)$ is strictly smaller than the probability of having a successful transmission in this interval.

We will assume from now on that $a < p$. When the system is stable, clearly the input rate $a$ must be equal to the throughput. Let us check this intuitive result.

Since a packet may leave the system (with probability $p$) only when the queue is not empty (which occurs with probability $1 - \pi(0)$), the throughput $T$ is given by

$$T = p (1 - \pi(0)) = \frac{a}{p}.$$  

from (25), which is the expected result.

Let $Q$ be the number of packets in the waiting room in steady-state, including the packet being transmitted, if any. We now want to do some flow control. More precisely, we want to determine the input rate $a$ such that $P(Q > k) < \beta$ where $k \in \mathbb{N}$ and $\beta \in (0,1)$ are arbitrary numbers.

Let us compute $P(Q \leq k)$. We have

$$P(Q \leq k) = \sum_{j=0}^{k} \pi(j).$$

After elementary algebra we finally obtain

$$P(Q \leq k) = 1 - \frac{a}{p} \left( \frac{a (1 - p)}{p (1 - a)} \right)^k.$$  

(Observe from the above result that $\lim_{k \to \infty} P(Q \leq k) = 1$ under the stability condition $a < p$, which simply says that the number of packets in the waiting room is finite with probability one in steady-state.)

In conclusion, if we want $P(Q > k) < \beta$, we must choose $a$ such that $0 \leq a < p$ and

$$\frac{a}{p} \left( \frac{a (1 - p)}{p (1 - a)} \right)^k < \beta.$$  

Such a result is useful, for instance, for dimensioning the size of the buffer so that the probability of loosing a packet is below a given threshold.

Other interesting performance measures for this system are $E[Q]$, $\text{Var}(Q)$, $P(X(n) = j)$ given the distribution function of $X(0)$ is known (use Proposition 2), etc.

### 2.2 Continuous-Time Markov Chain

A continuous-time Markov chain (abbreviated as C-MC) is a continuous-time (with index set $[0, \infty)$), discrete-space (with state-space $\mathcal{E}$) stochastic process $(X(t), t \geq 0)$ such that

$$P(X(t) = j \mid X(s_1) = i_1, \ldots, X(s_{n-1}) = i_{n-1}, X(s) = i) = P(X(t) = j \mid X(s) = i)$$  

(26)

for all $i_1, \ldots, i_{n-1}, i, j \in \mathcal{E}$, $0 \leq s_1 < \ldots < s_{n-1} < s < t$.

---

\[2\text{Here we use the well-known identity } \sum_{j=0}^{k} r^j = (1 - r^{k+1})/(1 - r) \text{ for all } k \geq 0 \text{ if } 0 \leq r < 1.\]
A C-MC is homogenous if 

\[ P(X(t+u) = j \mid X(s+u) = i) = P(X(t) = j \mid X(s) = i) := p_{i,j}(t-s) \]

for all \( i, j \in \mathcal{E}, \quad 0 \leq s < t, \quad u \geq 0. \)

From now on we shall only consider homogeneous C-MC’s even if we simply speak of a C-MC.

We have the following analog to Proposition 1:

**Proposition 4** (Chapman-Kolmogorov equation for C-MC’s). For all \( t > 0, s > 0, i, j \in \mathcal{E} \),

\[ p_{i,j}(t+s) = \sum_{k \in \mathcal{E}} p_{i,k}(t)p_{k,j}(s). \quad (27) \]

The proof is the same as the proof of 1 and is therefore omitted.

Define

\[ q_{i,i} := \lim_{h \to 0} \frac{p_{i,i}(h) - 1}{h} \leq 0 \quad \text{(28)} \]

\[ q_{i,j} := \lim_{h \to 0} \frac{p_{i,j}(h)}{h} \geq 0 \quad \text{(29)} \]

and let \( Q \) be the matrix \( Q = [q_{i,j}] \) (we will assume that the limits in (28)-(29) always exist. They will exist in all cases to be considered in this course).

The matrix \( Q \) is called the **infinitesimal generator** of the C-MC. If \( \mathcal{E} = \mathbb{N} \), then

\[
Q = 
\begin{pmatrix}
-q_{0,0} & 0 & \cdots & \cdots & \cdots & \cdots \\
q_{1,0} & -q_{1,1} & q_{1,2} & \cdots & \cdots & \cdots \\
& \ddots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
q_{i,0} & \cdots & \cdots & \cdots & q_{i,i-1} & \sum_{j \neq i} q_{i,j} \\
& \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix}
\]

In contrast with (5) note that

\[ \sum_{j \in \mathcal{E}} q_{i,j} = 0, \quad \forall i \in \mathcal{E}. \quad (30) \]

The quantity \( q_{i,j} \) has the following interpretation: when the system is in state \( i \) then the rate at which it departs state \( i \) is \( -q_{i,i} \), and the rate at which it moves from state \( i \) to state \( j, j \neq i \), is \( q_{i,j} \). In the literature \( q_{i,j} \) is referred to as the **transition rate** from state \( i \) to state \( j \neq i \) and \( -q_{i,i} \) is the transition rate out of state \( i \).

What is the probability \( p(i,j) \) that the C-MC goes to state \( j \) after leaving state \( i \neq j \)? From the interpretation of \( q_{i,j} \) given in the previous paragraph it should be clear that

\[ p(i,j) = \frac{q_{i,j}}{-q_{i,i}}, \quad i, j \in \mathcal{E}, \quad i \neq j. \quad (31) \]

What is the sojourn time distribution of the C-MC in state \( i \)? It can be shown that it is exponentially distributed with rate \( -q_{i,i} \) [6, Section 2.4]. In particular, the expected sojourn time in state \( i \) is \( -1/q_{i,i} \).

In summary, a homogeneous C-MC on \( \mathcal{E} \) is a stochastic process which sojourns in each state it visits for an exponentially distributed amount of time (with rate \( -q_{i,i} \) for state \( i \)) independent of the past evolution of the process, and upon leaving state \( i \) instantaneously joins state \( j \neq i \) with the probability \( p(i,j) \) given in (31). These features of a homogeneous C-MC will be used to establish the construction rule (see below) that we will use to check if a stochastic process is a C-MC.

Observe that the infinitesimal generator \( Q \) fully characterizes the behavior of a homogeneous continuous-time Markov chain, alike the behavior of a discrete-time Markov chain that is fully characterized by the transition matrix \( P \) (see Section 2.1).
Let us now focus on the transient probability distribution of a C-MC. Define the row vector \( \pi(t) = (\pi_i(t), i \in \mathcal{E}) \), where \( \pi_i(t):=P(X(t) = i) \) is the probability that the C-MC is in state \( i \) at time \( t \).

For all \( j \in \mathcal{E} \) we have (Hint: use Bayes’ formula)

\[
\pi_j(t+h) = P(X(t+h) = j \mid X(t) = j)\pi_j(t) + \sum_{i \neq j} P(X(t+h) = j \mid X(t) = i)\pi_i(t)
\]

so that

\[
\frac{\pi_j(t+h) - \pi_j(t)}{h} = \frac{p_{j,j}(h)\pi_j(t)}{h} + \sum_{i \neq j} \frac{p_{i,j}(h)}{h} \pi_i(t), \quad t \geq 0.
\]

Letting \( h \to 0 \) in the latter equation and using (28)-(29) gives

\[
\frac{d\pi_j(t)}{dt} = q_{j,j}\pi_j(t) + \sum_{i \neq j} q_{i,j}\pi_i(t), \quad j \in \mathcal{E}, \ t \geq 0
\]

or, in matrix form,

\[
\frac{d}{dt} \pi(t) = \pi(t)Q, \quad t \geq 0.
\]

Once \( \pi(0) \) is specified this first-order differential equation has a unique solution, given by\(^4\)

\[
\pi(t) = \pi(0)e^{Qt}, \quad t \geq 0.
\] (32)

In practice this result is not very useful since it is in general very difficult to calculate \( e^{Qt} \). This is especially true when \( \mathcal{E} \) has infinite dimension (infinite state-space).

In the rest of this section we shall only be concerned with the steady-state probability distribution of a C-MC, namely with \( \pi := \lim_{t \to \infty} \pi(t) \) whenever this limit exists and does not depend on the initial state. In direct analogy with the definition given for a D-MC, we shall say that an homogeneous C-MC is irreducible if for every state \( i \in \mathcal{E} \) there exists \( s > 0 \) such that \( p_{i,s}(s) > 0 \).

Before stating the main result of this section, let us make a guess on what the limiting distribution \( \pi \), when it exists, is. Assume that \( \pi := (\pi(i), i \in \mathcal{E}) \) is the stationary probability distribution.

Alike for D-MCs, we first develop an argument to find \( \pi \).

For any \( 0 < h < t \), we have from (32)

\[
\pi(t) = \pi(0)e^{Q(t-h)}e^{Qh} = \pi(t-h)e^{Qh} = \pi(t-h)(I + Qh + o(h))
\] (33)

with \( o(h) \) denotes any function such that \( \lim_{h \to 0} o(h)/h = 0 \). Letting \( t \to \infty \) in (33) gives

\[
\pi Qh = o(h).
\] (34)

Dividing now both sides of (34) by \( h \) and letting \( h \to 0 \) gives

\[
\pi Q = 0.
\] (35)

We have shown that if the steady-state exists (in the sense that \( \pi = \lim_{t \to \infty} \pi(t) \) exists and does not depend on the initial state) then this limit satisfies the linear equation (35).

The following result gives necessary conditions for the existence of the steady-state of a C-MC.

\(^3\)When the cardinality of \( \mathcal{E} \) is infinite the identity \( \lim_{h \downarrow 0} \sum_{i \in \mathcal{E}} (p_{i,j}(h)/h)\pi_i(t) = \sum_{i \in \mathcal{E}} \lim_{h \downarrow 0} (p_{i,j}(h)/h)\pi_i(t) \) is justified by the bounded convergence theorem.

\(^4\)\( e^A = \sum_{k=0}^{\infty} A^k/k! \) for any matrix \( A \).
Proposition 5 (Limiting distribution function of a C-MC). If a homogeneous C-MC with infinitesimal generator $Q$ is irreducible, and if the system of equations

\[
\begin{align*}
x Q &= 0 \\
x \cdot 1 &= 1
\end{align*}
\]

has a strictly positive solution (i.e., for all $i \in E$, $x(i)$, the $i$-th component of the row vector $x$, is strictly positive) then

\[
x(i) = \lim_{t \to \infty} P(X(t) = i)
\]

for all $i \in E$, regardless of the initial state $X(0)$.

We shall not prove this result. We may compare this result with our earlier equation for a D-MC namely, $\pi = \pi P$ (see Proposition 3); here $P$ is the transition matrix, whereas the infinitesimal generator $Q$ is a matrix of transition rates.

The equation $\pi Q = 0$ in Proposition 5 can be rewritten as

\[
\sum_{j \in E} \pi(j) q_{j,i} = 0
\]

for all $i \in E$, or equivalently, cf. (30),

\[
\left( \sum_{j \neq i} q_{i,j} \right) \pi(i) = \sum_{j \neq i} \pi(j) q_{j,i}.
\]

The above equation expresses the fact that for any state

the probability flow out of a state is equal to the probability flow into that state

when the system is at equilibrium. Equations (37) are called the balance equations and the sentence in italic below (37) is referred to as the principle of flow conservation.

Many examples of continuous-time Markov chains will be discussed in the forthcoming sections.

It remains one important question to address: how does one check that a stochastic process is a C-MC? In the discrete-time setting it is in general easy to check whether or not definition (2) in Section 2.1 is satisfied.

The one-step transition matrix $P$ in (3) is usually easy to find.

In the continuous-time setting checking that definition (26) holds is much more difficult since it has to be checked for a non-countable number of points in time. Below we give a construction rule that can be used to check whether or not a stochastic process is a C-MC and, if yes, to identify its infinitesimal generator.

Construction rule. Let $\{X(t), t \geq 0\}$ be a continuous-time stochastic process with the following properties.

When the process enters state $i$ at time $t$ then

- for each $j \neq i$ a rv $Y_{ij}$ with an exponential distribution with parameter $\mu_{ij}$ is generated. These rvs are mutually independent and are independent of the past of the process. One may have $\mu_{ij} = 0$ in which case $Y_{ij} = +\infty$ (corresponding to the case where state $j$ cannot be reached from state $i$ in one transition).

- assume that $Y_{ik}$ is the smallest among all the rvs $\{Y_{ij}\}$. It is well known that $Y_{ik}$ as an exponential distribution with parameter $\sum_j \mu_{ij}$. At time $t + Y_{ik}$ the process instantaneously jumps into state $k$.

If the above holds for all states then $\{X(t), t \geq 0\}$ is a C-MC, with infinitesimal generator $Q = [q_{i,j}]$ given by

\[
q_{i,j} = \begin{cases} 
\mu_{ij} & \text{if } i \neq j \\
-\sum_{j \neq i} \mu_{ij} & \text{if } i = j.
\end{cases}
\]
2.2.1 A case study: Markov modulated arrivals

Consider a process \( Y := \{Y(t), t \geq 0\} \) that alternates between two states \( H \) (for High) and \( L \) (for Low), such that \( Y(t) = H \) (resp. \( Y(t) = L \)) if the process is in state \( H \) (resp. \( L \)) at time \( t \). The sojourn time of \( Y \) in state \( H \) (resp. \( L \)) is random and exponentially distributed with mean \( 1/\mu_H \) (resp. \( 1/\mu_L \)). We assume that all sojourn times are mutually independent rvs and that \( \mu_H > 0 \) and \( \mu_L > 0 \).

In steady-state the probability \( p \) of finding \( Y \) in state \( H \) is

\[
p = \frac{1/\mu_H}{1/\mu_H + 1/\mu_L} = \frac{\mu_L}{\mu_H + \mu_L}
\]

and the probability \( 1 - p \) of finding it in state \( L \) is

\[
1 - p = \frac{\mu_H}{\mu_H + \mu_L}.
\]

Note that \( p \) is simply the ratio of time spent by \( Y \) on average in state \( H \) (given by \( 1/\mu_H \)) over the average duration of a cycle (given by \( 1/\mu_H + 1/\mu_L \)), where a cycle is defined as the time that elapses before two consecutive entries of \( Y \) in state \( H \) or in state \( L \).

To prove (39) rigorously, observe that under the above assumptions \( Y \) is an irreducible C-MC (Hint: apply rule #1) with infinitesimal generator

\[
Q = \begin{pmatrix} -\mu_H & \mu_H \\ \mu_L & -\mu_L \end{pmatrix}.
\]

Solving for the equations (i) \( \pi Q = 0 \) and (ii) \( \pi \cdot 1 = 1 \) with \( \pi = (\pi_H, \pi_L) \), gives

\[
\pi_H = \frac{\mu_L}{\mu_H + \mu_L} \quad \text{and} \quad \pi_L = \frac{\mu_H}{\mu_H + \mu_L}.
\]

Since \( \pi = (\pi_H, \pi_L) \) is the unique strictly positive solution of eqns (i)-(ii) we conclude from Proposition 5 that \( \pi \) is the steady-state distribution of \( Y \), regardless the initial state of \( Y(0) \). Hence, \( p = \pi_H \) and \( 1 - p = \pi_L \) as announced.

Assume that during periods of type \( H \) (resp. \( L \)) objects (customers, packets, jobs, etc.) are generated according to a Poisson process with rate \( \lambda_H > 0 \) (resp. \( \lambda_L > 0 \)). We further assume that the number of objects generated in different periods are independent rvs. We say that an object is of type \( H \) (resp. \( L \)) if it has been generated in a period of type \( H \) (resp. \( L \)).

Define \( q_{b,a} = P(X_{n+1} = b \mid X_n = a) \), \( a, b \in \{H, L\} \), the probability that the \((n+1)\)-st object is of type \( b \) given that the \( n \)-th object is of type \( a \). Our object is to compute \( q_{b,a} \) for \( a, b \in \{H, L\} \).

Let \( F_n \) be the event that the \( n \)-th and the \((n+1)\)-st object arrive in the same period. We have

\[
P(F_n \mid X_n = a) = \lambda_a/\lambda_a + \mu_a \quad \text{for} \quad a \in \{H, L\}. 
\]

Indeed, when \( Y \) is in state \( a \), the next object will be generated after an exponential duration \( A \) with rate \( \lambda_a \) and \( Y \) will jump to state \( b \neq a \) after an exponential duration \( B \) with rate \( \mu_a \). Hence, \( P(F \mid X_n = a) = P(A < B) = \lambda_a/\lambda_a + \mu_a \), since \( A \) and \( B \) are independent exponential rvs. To see this, condition on (for instance) \( A \) to get

\[
P(A < B) = \int_0^\infty P(x < B)\lambda_a\exp(-\lambda_a x)\,dx = \lambda_a/\lambda_a + \mu_a.
\]

By the law of total probability (see Proposition 33)

\[
q_{L,L} = P(X_{n+1} = L \mid X_n = L, F_n)P(F_n \mid X_n = L) + P(X_{n+1} = L \mid X_n = L, F_n^c)P(F_n^c \mid X_n = L) = \frac{1}{\lambda_L + \mu_L} + P(X_{n+1} = L \mid X_n = L, F_n^c)\frac{\mu_L}{\lambda_L + \mu_L} = \frac{\lambda_L}{\lambda_L + \mu_L} + q_{L,H} + q_{L,L} \frac{\mu_L}{\lambda_L + \mu_L}.
\]

The last equality comes from the fact that the probability \( P(X_{n+1} = L \mid X_n = L, F_n^c) \) is the same as the probability \( q_{L,H} \) due to the memoryless property of the (Poisson) arrival process of objects of type \( L \). Similarly

\[
q_{L,H} = P(X_{n+1} = L \mid X_n = H, F_n)P(F_n \mid X_n = H) + P(X_{n+1} = L \mid X_n = H, F_n^c)P(F_n^c \mid X_n = H)
\]

\[
q_{H,L} = P(X_{n+1} = H \mid X_n = L, F_n)P(F_n \mid X_n = L) + P(X_{n+1} = H \mid X_n = L, F_n^c)P(F_n^c \mid X_n = L) = \frac{1}{\lambda_H + \mu_L} + P(X_{n+1} = H \mid X_n = L, F_n^c)\frac{\mu_L}{\lambda_H + \mu_L}.
\]

\[
q_{H,H} = P(X_{n+1} = H \mid X_n = L, F_n)P(F_n \mid X_n = L) + P(X_{n+1} = H \mid X_n = L, F_n^c)P(F_n^c \mid X_n = L) = \frac{1}{\lambda_H + \mu_L} + P(X_{n+1} = H \mid X_n = L, F_n^c)\frac{\mu_L}{\lambda_H + \mu_L}. 
\]
\[ = 0 \times \frac{\lambda_H}{\lambda_H + \mu_H} + P(X_{n+1} = L \mid X_n = H, F_n^c) \frac{\mu_H}{\lambda_H + \mu_H} \]
\[ = q_{L,H} \frac{\mu_H}{\lambda_H + \mu_H}. \]

Substituting the above value of \( q_{L,H} \) in (40) gives
\[ q_{L,L} = \frac{\lambda_L}{\lambda_L + \mu_L} + q_{L,L} \frac{\mu_H}{\lambda_H + \mu_H} \frac{\mu_L}{\lambda_L + \mu_L}, \]
so that
\[ q_{L,L} = \frac{\lambda_L(\lambda_H + \mu_H)}{\lambda_L(\lambda_H + \mu_H) + \lambda_H \mu_L}, \quad \text{and therefore} \quad q_{L,H} = \frac{\lambda_L \mu_H}{\lambda_L(\lambda_H + \mu_H) + \lambda_H \mu_L}. \]

By symmetry
\[ q_{H,H} = \frac{\lambda_H(\lambda_L + \mu_L)}{\lambda_H(\lambda_L + \mu_L) + \lambda_L \mu_H}, \quad q_{H,L} = \frac{\lambda_H \mu_L}{\lambda_H(\lambda_L + \mu_L) + \lambda_L \mu_H}. \]
Assume that \( \lambda_H = \lambda_L = \lambda \) (all objects arrive according to a Poisson process with rate \( \lambda \)). Then,
\[ q_{L,L} = \frac{\lambda + \mu_H}{\lambda + \mu_H + \mu_L}, \quad q_{H,L} = \frac{\mu_L}{\lambda + \mu_L + \mu_H}, \]
\[ q_{H,H} = \frac{\lambda + \mu_L}{\lambda + \mu_L + \mu_H}, \quad q_{L,H} = \frac{\mu_H}{\lambda + \mu_H + \mu_L}. \]

Is it possible to select the parameters in (43)-(44) so that an arriving object is equally likely to be of either type given the type of the previous arrival is known? In other words, when do we have \( q_{L,L} = q_{H,L} = q_{H,H} = q_{L,H} = 1/2 \). It is easy to see that this is possible if and only if \( \lambda = 0 \) and \( \mu_H = \mu_L \), a case which is irrelevant since there is no arrival with probability 1.

On the other hand, if \( \mu_H \to 0 \) (resp. \( \mu_L \to 0 \)) then \( q_{L,L} \to 1 \) and \( q_{H,H} \to 0 \) (resp. \( q_{H,H} \to 1 \) and \( q_{L,H} \to 0 \)), namely all arrivals tend to be of type \( H \) (resp. type \( L \)).

### 3 Birth and Death Process

A birth and death process \( \{X(t), t \geq 0\} \) is a continuous-time discrete-space (with state-space which is \( \mathbb{N} \) or which is a subset, finite or infinite, of \( \mathbb{N} \)) Markov chain (or Markov process) with infinitesimal generator \( \mathbf{Q} = \{q_{i,j}\} \) such that \( q_{i,j} = 0 \) for \( j \notin \{i-1, i, i+1\} \) for \( i \geq 1 \) and \( j \notin \{i, i+1\} \) for \( i = 0 \).

In other words, the infinitesimal generator \( \mathbf{Q} \) is a tridiagonal matrix of the following form
\[
\mathbf{Q} = \begin{pmatrix}
-\lambda_0 & \lambda_0 & 0 & 0 \\
\mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 \\
0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 \\
0 & 0 & \mu_3 & -\lambda_3 - \mu_3 & \lambda_3 \\
& & & \ddots & \ddots \\
& & & & \ddots & \ddots \\
\end{pmatrix}
\]

where we have set \( q_{i,i+1} = \lambda_i \) for \( i = 0, 1, \ldots \) and \( q_{i,i-1} = \mu_i \) for \( i = 1, 2, \ldots \).

The random variable (rv) \( X(t) \) may be interpreted as the size of the population at time \( t \). In that case, \( \lambda_i \) gives the birth-rate when the size of the population is \( i \) and \( \mu_i \) gives the death-rate when the size of the population is \( i \) with \( i \geq 1 \) (by convention \( \mu_0 = 0 \)).

Assume that the Markov process \( \{X(t), t \geq 0\} \) is irreducible (if \( \lambda_i, \mu_i > 0 \) for all \( i \in \mathbb{N} \) then it will be irreducible on \( \mathbb{N} \); if \( \lambda_j, \mu_j > 0 \) for \( j = 0, 1, \ldots, i \) and \( \lambda_{i+1} = 0 \) then it will be irreducible on \( \{0, 1, \ldots, i\} \)). Then, from Proposition 5 in Section 2.2 we know that if the equation \( \pi \mathbf{Q} = 0 \) (with \( \pi = (\pi(0), \pi(1), \ldots) \)) has a solution strictly positive (i.e., \( \pi(i) > 0 \) for all \( i = 0, 1, \ldots \)) such that \( \sum_{i=0}^{\infty} \pi(i) = 1 \) then \( \lim_{t \to \infty} P(X(t) = i) = \pi(i) \) for all \( i = 0, 1, \ldots \).
Proposition 6 (Balance equations of a birth and death process).

The equation $\pi Q = 0$ also writes

$$
\lambda_0 \pi(0) = \mu_1 \pi(1)
$$

(45)

$$(\lambda_i + \mu_i) \pi(i) = \lambda_{i-1} \pi(i-1) + \mu_{i+1} \pi(i+1) \quad i = 1, 2, \ldots.
$$

(46)

Equations (45)-(46) are the balance equations - also called the equilibrium equations - of a birth and death process (see Section 2.2 for the balance equations of an arbitrary continuous-time Markov chain). They express the property that in steady-state the flow of probability out of a state is equal to the flow of probability into that state. This observation, of a physical nature, will allow us in most cases to easily compute the equilibrium equations of a birth and death process.

From Proposition 6 we find that

Proposition 7 (Stationary probabilities of a birth and death process).

Assume that the series

$$
C := 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots + \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} + \cdots
$$

(47)

converges (i.e., $C < \infty$). Then, for each $n = 1, 2, \ldots,$

$$
\pi(n) = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \pi(0)
$$

(48)

where $\pi(0) = 1/C$.

This result is obtained by direct substitution of (48) into (45)-(46). The computation of $\pi(0)$ relies on the fact that $\sum_{n=0}^{\infty} \pi(n) = 1$.

The condition (47) is called the stability condition of a birth and death process.

4 Absorbing Markov Chains

4.1 Discrete-time absorbing Markov chains

Let $\{X(n), n = 0, 1, \ldots\}$ be a homogeneous discrete-time Markov chain (D-MC), taking values in the state-space $E := \{1, 2, \ldots, N, 1^*, 2^*, \ldots, M^*\}$. Let $P = [p_{i,j}]_{i,j \in E}$ be its transition probability matrix, namely,

$$
p_{i,j} = P(X(n+1) = j \mid X(n) = i)
$$

for all $i, j \in E$.

States $1, 2, \ldots, N$ are all transient and states $1^*, 2^*, \ldots, M^*$ are all absorbing. This means that starting at time $n = 0$ from any transient state the Markov chain will eventually reach an absorbing state and will remain forever in this absorbing state. Starting from an absorbing state the Markov chain will remain in this absorbing state.

With this definition, the transition matrix $P$ takes the form

$$
P = \begin{pmatrix}
\mathbf{A} & \mathbf{R} \\
0 & \mathbf{I}
\end{pmatrix}
$$

where $\mathbf{A} = [a_{i,j}]_{1 \leq i,j \leq N}$ and $\mathbf{R} = [r_{i,j}]_{1 \leq i \leq M, 1^* \leq j \leq M^*}$ are $N$-by-$N$ and $N$-by-$M$ matrices, respectively, and $\mathbf{I}$ is the $M$-by-$M$ identity matrix. $a_{i,j}$ is the probability of going from transient state $i$ to transient state $j$ in one step, and $r_{i,j}$ is the probability of going from transient state $i$ to absorbing state $j$ in one step.

The definition of an absorbing discrete-time Markov chain also implies that

$$
P(X(n) = j \mid X(0) = i) \to 0 \quad \text{as } n \to \infty
$$

for any pair of transient states $i$ and $j$, since eventually each transient state will reach an absorbing state.
Raising the matrix $P$ to the power $n$ gives

$$P^n = \left( \begin{array}{cc} A^n & R \sum_{k=0}^{n-1} A^i \\ 0 & I \end{array} \right).$$

We know that the $(i,j)$-entry of the matrix $P^n$ gives $P(X(n) = j \mid X(0) = i)$, the probability of going from state $i$ to state $j$ in $n$ steps (see Section 2.1, Proposition 1). Therefore, the $(i,j)$-entry of the matrix $A^n$ – denoted by $a^{(n)}_{i,j}$ – gives the probability of going from transient state $i$ to transient state $j$ in $n$ steps. Since starting from any transient state the Markov chain will eventually reach an absorbing state, this implies that all entries of the matrix $A^n$ will converge to 0 as $n \to \infty$, that is,

$$A^n \to 0 \quad \text{as} \quad n \to \infty. \quad (49)$$

Let $T(i)$ be the expected time before absorption starting from state $i$ at time $n = 0$, namely,

$$T(i) = E[\inf\{n \geq 0 : X(n) \in \{1^*, \ldots, M^*\} \mid X(0) = i\}].$$

Clearly $T(i) = 0$ if $i$ is an absorbing state. From now on we will only consider $T(i)$ when $i$ is a transient state.

Let $n_{i,j}$ be the expected number of visits to state transient $j$ starting from transient state $i$, namely,

$$n_{i,j} = E\left[ \sum_{n \geq 0} 1(X(n) = j) \mid X(0) = i \right], \quad i,j \in \{1,2,\ldots,N\}. \quad (50)$$

As usual $1(X(n) = j)$ is equal to 1 if $X(n) = j$ and is equal to 0 otherwise. Note that if (for instance) $X(n) = j$, $X(n+1) = j$ and $X(n+2) = j$ then the number of visits to state $j$ in $[n,n+2]$ is three (and not one). Therefore,

$$T(i) = \sum_{j=1}^{N} n_{i,j}, \quad i \in \{1,2,\ldots,N\}. \quad (50)$$

Our objective is to find $T(i)$ for all $i \in \{1,2,\ldots,N\}$.

Define the column vector $T = (T(1), \ldots, T(N))^T$ of dimension $N$ and the $N$-by-$N$ matrix $N = [n_{i,j}]_{1 \leq i,j \leq N}$. In matrix notation (50) writes

$$T = N.1 \quad (51)$$

where here $1$ is the column vector of dimension $N$ which has all its entries equal to 1.

Below is the main result:

**Proposition 8 (Mean number of visits and expected absorption time).**

The matrix $N$ is given by

$$N = (I - A)^{-1} \quad (52)$$

and

$$T = (I - A)^{-1}.1. \quad (53)$$

**Proof.** Let $i$ and $j$ be arbitrary transient states. Define the binary random variable (rv) $X_j^{(n)} = 1(X(n) = j)$, which is equal to 1 if $X(n) = j$ and to 0 otherwise. We have

$$E[X_j^{(0)} + X_j^{(1)} + \cdots + X_j^{(n)} \mid X(0) = i] = \sum_{k=0}^{n} E[X_j^{(k)} \mid X(0) = i] = \sum_{k=0}^{n} E[1(X(k) = j) \mid X(0) = i] \quad (54)$$

$$= \sum_{k=0}^{n} P(X(k) = j \mid X(0) = i)$$

$$= \sum_{k=0}^{n} P(X(k) = j \mid X(0) = i)$$
where we recall that \( a_{i,j}^{(k)} \) is the \((i,j)\)-entry of the matrix \( A^k \).

Note that the left-hand side of (54) is nothing but the expected number of visits to state \( j \) in \([0,n]\) starting from state \( i \). On the other hand,

\[
\lim_{n \to \infty} E[X_j(0) + X_j(1) + \cdots + X_j(n) \mid X(0) = i] = n_{i,j},
\]

where the first equality holds because the rvs \( \{X_j^{(k)}\}_k \) are all nonnegative (Bepo-Levi theorem), and the second equality is nothing but the definition of \( N_{i,j} \).

Combining (55) and (56) yields

\[
n_{i,j} = \sum_{k \geq 0} a_{i,j}^{(k)}
\]

or, in matrix notation,

\[
N = \sum_{k \geq 0} A^k.
\]

It remains to evaluate the matrix \( \sum_{k \geq 0} A^k \). This can be done as follows. We have

\[
(I - A) \sum_{k=0}^{n} A^k = I - A^{n+1}.
\]

Letting \( n \to \infty \) in the above identity, then using (58) and (49) gives

\[
(I - A)N = I
\]

which in turn implies that \( N \) is the inverse of the matrix \( I - A \), i.e. \( N = (I - A)^{-1} \), which proves (52). The proof is then concluded by using (51).

Let \( B \) the \( N \)-by-\( M^* \) matrix whose \((i,j)\)-entry \( B_{i,j} \) gives the probability that the chain is absorbed in state \( j \in \{1^*, 2^*, \ldots, M^*\} \) given that it was initially in transient state \( i \in \{1, 2, \ldots, N\} \) (i.e. \( X(0) = i \)).

The following holds:

**Proposition 9** (Absorption probabilities).

\[
B = (I - A)^{-1} R.
\]

**Proof.** Let \( \tau(i) \) be the time before absorption given that \( X(0) = i \) with \( i \in \{1, 2, \ldots, N\} \).

We have for \( i \in \{1, 2, \ldots, N\} \), \( j \in \{1^*, 2^*, \ldots, M^*\} \),

\[
B_{i,j} = P(X(\tau(i)) = j \mid X(0) = i)
\]

\[
= \sum_{k=1}^{N} \sum_{n \geq 0} P(X(\tau(i)) = j, X(\tau(i)) = k, \tau(i) = n + 1 \mid X(0) = i)
\]

\[
= \sum_{k=1}^{N} \sum_{n \geq 0} P(X(n + 1) = j, X(n) = k, \tau(i) = n + 1 \mid X(0) = i)
\]

\[
= \sum_{k=1}^{N} \sum_{n \geq 0} P(X(n + 1) = j, \tau(i) = n + 1 \mid X(n) = k, X(0) = i) \times P(X(n) = k \mid X(0) = i)
\]

by using Bayes’ formula.
Let us observe the Markov chain \( \{X(n)\}_{n \geq 0} \)

In matrix form (62) writes

\[
X \text{ is transient state } i\] 

Define \( T_{i,j} \) transient state \( i \) in this setting, the \( (i,j) \) if \( i \neq j \)

For future use, let us introduce the \( N \)-by-\( N \) matrix \( \tilde{Q} \) defined by

\[
\tilde{Q} = [q_{i,j}]_{1 \leq i,j \leq N}. \tag{60}
\]

Let us observe the Markov chain \( \{X(t), t \geq 0\} \) at jump times. This is an homogeneous, absorbing, discrete-time Markov chain, whose the \( (i,j) \)-entry of its transition matrix \( P = [p_{i,j}]_{i,j \in \mathcal{E}} \) is given by \( p_{i,j} = -q_{i,j}/q_{i,i} \) if \( i \neq j \), and \( p_{i,i} = 0 \).

In this setting, the \( (i,j) \)-entry of the matrix \( A = [a_{i,j}]_{1 \leq i,j \leq N} \) defined in Section 4.1 is

\[
a_{i,j} = \begin{cases} 
-\frac{q_{i,j}}{q_{i,i}} & \text{if } i \neq j \\
0 & \text{if } i = j.
\end{cases} \tag{61}
\]

Hence, we know by Proposition 8 that \( n_{i,j} \), the expected number of visits to transient state \( j \) starting from transient state \( i \), is given by the \( (i,j) \)-entry of the inverse of the matrix \( I - A \) defined in (61).

Let \( T(i) \) be the expected time before absorption starting from transient state \( i \). Since the expected time in state \( j \) follows an exponential distribution with mean \( 1/q_{j,j} \), we deduce that

\[
T(i) = -\sum_{j=1}^{N} (I - A)_{i,j}^{-1} \times \frac{1}{q_{j,j}}, \quad i = 1, 2, \ldots, N, \tag{62}
\]

where \( (I - A)^{-1} \) denotes the \( (i,j) \)-entry of the matrix \( (I - A)^{-1} \).

Define \( T = (T(1), \ldots, T(N))^T \) and let \( \text{diag}(x_1, \ldots, x_N) \) denote a \( N \)-by-\( N \) diagonal matrix whose \( (i,i) \)-entry is \( x_i \).

In matrix form (62) writes

\[
T = -(I - A)^{-1} \text{D.1} \tag{63}
\]

with \( \text{D} = \text{diag}(1/q_{1,1}, \ldots, 1/q_{N,N}) \).

It is easily seen from (61) and the definition of the matrix \( \tilde{Q} \) that \( \text{D}^{-1}(I - A) = \tilde{Q} \).

We have then shown the following result:

4.2 Continuous-time absorbing Markov chains

This is the same framework as in the previous section but now we consider an absorbing, homogenous, continuous-time Markov chain (C-MC) \( \{X(t), t \geq 0\} \) on the finite state-space

\[
\mathcal{E} = \{1, 2, \ldots, N, 1^*, 2^*, \ldots, M^*\}.
\]

States 1, 2, \ldots, \( N \) are transient and states 1\( ^* \), 2\( ^* \), \ldots, \( M^* \) are absorbant.

Let \( q_{i,j} \) be the transition rate from state \( i \) to state \( j \) for \( i \neq j \) and denote by \( -q_{i,i} \) the transition rate out of state \( i \). We know from the section on C-MC that \( q_{i,i} = -\sum_{j \in E_j \neq i} q_{i,j} \) for all \( i \in \mathcal{E} \). We also know that \( -q_{i,j}/q_{i,i} \) is the probability that the C-MC enters state \( j \neq i \) when leaving state \( i \).

Note that \( q_{i,j} = 0 \) for any state \( j \in \mathcal{E} \), whenever if \( i \) is an absorbing state.

We have then shown the following result:
Proposition 10 (Expected absorption time for a C-MC).
Let \( T = (T(1), \ldots, T(N))^T \), with \( T(i) \) the expected time before absorption starting from state \( i \) at time \( t = 0 \). Then,
\[
T = -\tilde{Q}^{-1}1
\]
where \( \tilde{Q} \) is defined in (60).

5 Queueing Theory

5.1 The M/M/1 queue

In this queueing system the customers arrive according to a Poisson process with rate \( \lambda \). The time it takes to serve every customer is an exponential rv with parameter \( \mu \). We say that the customers have exponential service times. The service times are supposed to be mutually independent and further independent of the interarrival times.

When a customer enters an empty system his service starts at once; if the system is nonempty the incoming customer joins the queue. When a service completion occurs, a customer from the queue (we do not need to specify which one for the time being), if any, enters the service facility at once to get served.

Let \( X(t) \) be the number of customers in the system at time \( t \).

Proposition 11 (Birth and death process for an M/M/1 queue).

The process \( \{X(t), t \geq 0\} \) is a birth and death process with birth rate \( \lambda_i = \lambda \) for all \( i \geq 0 \) and with death rate \( \mu_i = \mu \) for all \( i \geq 1 \).

This result follows from the construction rule (see Section 2.2). Indeed, in any state \( i = 1, 2, \ldots \), there are two events that compete to drive the system out of \( i \): an arrival of a new customer (birth) in which case the system will jump to state \( i + 1 \) or a departure of a customer (death) in which case the system will jump to state \( i - 1 \); if \( i = 0 \) then only an arrival will trigger a jump to state 1).

It the system is in state \( i \) (\( i = 1, 2, \ldots \)) at some time \( t \), then the time to go until it leaves that state will be the minimum of two exponential rvs: the 1st rv gives the time to go until the next arrival after time \( t \) (this time is indeed distributed like an exponential rv since the interarrival times have been assumed to be exponentially distributed and that the exponential distribution is memoryless ), the 2nd random variable gives the time to go until the next departure after time \( t \) (this time is also distributed like an exponential rv since we have assumed that the service times are exponentially distributed and because of the memoryless property of the exponential distribution). Since the minimum of two independent exponential rvs is an exponential rv, we see that the construction rule applies to state \( i = 1, 2, \ldots \). If the system is in state \( i = 0 \) at some time \( t \) then the time before the next event (here, necessarily an arrival) is exponentially distributed (again because the exponential distribution is memoryless and that the arrival process is Poisson).

We may therefore conclude from the construction rule that the process \( \{X(t), t \geq 0\} \) is a Markov process. Since the only possible transitions out of state \( i \) are to enter state \( i - 1 \) or state \( i + 1 \) if \( i = 1, 2, \ldots \) (state 1 if the system is in state 0) then we see that this Markov process is actually a birth and death process with state-space \( \{0, 1, 2, \ldots \} \).

Let \( \pi(i), i \geq 0 \), be the probability distribution of the number of customers in the system in steady-state. The balance equations for this birth and death process read
\[
\begin{align*}
\lambda \pi(0) & = \mu \pi(1) \\
(\lambda + \mu) \pi(i) & = \lambda \pi(i - 1) + \mu \pi(i + 1) \quad \forall i \geq 1.
\end{align*}
\]
Define
\[
\rho = \frac{\lambda}{\mu}.
\]
(64)
The quantity \( \rho \) is referred to as the traffic intensity since it gives the mean quantity of work brought to the system per unit of time.
A direct application of Proposition 7 yields:

\[\text{From now on } \rho \text{ will always be defined as } \lambda/\mu \text{ unless otherwise mention.}\]
Proposition 12 (Stationary queue-length distribution function of an M/M/1 queue).

If \( \rho < 1 \) then

\[
\pi(i) = (1 - \rho) \rho^i
\]

for all \( i \geq 0 \).

Therefore, the stability condition \( \rho < 1 \) simply says that the system is stable if the work that is brought to the system per unit of time is strictly smaller than the processing rate (which is 1 here since there is only one server).

Proposition 12 therefore says that the distribution function of the queue-length in steady-state is a geometric distribution.

From (65) we can compute (in particular) the mean number of customers \( E[X] \) (still in steady-state). We find

\[
E[X] = \frac{\rho}{1 - \rho}.
\]

Observe that \( E[X] \to \infty \) when \( \rho \to 1 \), so that, in practice if the system is not stable, then the queue will explode. It is also worth observing that the queue will empty infinitely many times when the system is stable since \( \pi(0) = 1 - \rho > 0 \).

We may also be interested in the probability that the queue exceeds, say, \( K \) customers, in steady-state. From (65) we have

\[
P(X \geq K) = \rho^K.
\]

What is the throughput \( T \) of an M/M/1 in equilibrium? The answer should be \( T = \lambda \). Let us check this guess.

We have

\[
T = (1 - \pi(0)) \mu.
\]

Since \( \pi(0) = 1 - \rho \) from (65) we see that \( T = \lambda \) by definition of \( \rho \).

5.2 The M/M/1/K queue

In practice, queues are always finite. In that case, a new customer is lost when he finds the system full (e.g., telephone calls).

The M/M/1/K may accommodate at most \( K \) customers, including the customer in the service facility, if any. Let \( \lambda \) and \( \mu \) be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively.

Let \( \pi(i), i = 0, 1, \ldots, K \), be the distribution function of the queue-length in steady-state. The balance equations for this birth and death process read

\[
\begin{align*}
\lambda \pi(0) &= \mu \pi(1) \\
(\lambda + \mu) \pi(i) &= \lambda \pi(i-1) + \mu \pi(i+1) \quad \text{for } i = 1, 2, \ldots, K-1 \\
\lambda \pi(K-1) &= \mu \pi(K).
\end{align*}
\]

Proposition 13 (Stationary queue-length distribution function in an M/M/1/K queue).

If \( \rho \neq 1 \) then

\[
\pi(i) = \frac{(1 - \rho) \rho^i}{1 - \rho^{K+1}}
\]

for \( i = 0, 1, \ldots, K \), and \( \pi(i) = 0 \) for \( i > K \).

If \( \rho = 1 \) then

\[
\pi(i) = \frac{1}{K+1}
\]

for \( i = 0, 1, \ldots, K \), and \( \pi(i) = 0 \) for \( i > K \).
Here again the proof of Proposition 13 relies on the fact that \( \{X(t), t \geq 0\} \), where \( X(t) \in \{0,1,\ldots,K\} \) is the number of customers in the system at time \( t \), can be modeled as a birth and death process with birth rates \( \lambda_i = \lambda \) for \( i = 0,1,\ldots,K-1 \) and \( \lambda_i = 0 \) for \( i \geq K \). The proof that \( \{X(t), t \geq 0\} \) is Markov process is analogous to the proof given above for the M/M/1 queue (the only difference is that here we only need to focus on states \( i = 0,1,\ldots,K \); note that from state \( i = K \) the system may only go to state \( K-1 \)).

In particular, the probability that an incoming customer is rejected is \( \pi(K) \).

### 5.3 The M/M/c queue

There are \( c \geq 1 \) servers and the waiting room has infinite capacity. If more than one server is available when a new customer arrives (which necessarily implies that the waiting room is empty) then the incoming customer may enter any of the free servers.

Let \( \lambda \) and \( \mu \) be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively.

Here again the process \( \{X(t), t \geq 0\} \) of the number of customers in the system can be modeled as a birth and death process (use the construction rule). The birth rate is \( \lambda_i = \lambda \) when \( i \geq 0 \). The death rate is given by

\[
\mu_i = \begin{cases} 
  i\mu & \text{for } i = 1,2,\ldots,c-1 \\
  c\mu & \text{for } i \geq c
\end{cases}
\]

which can be also written as \( \mu_i = \mu \min(i,c) \) for all \( i \geq 1 \).

Using these values of \( \lambda_i \) and \( \mu_i \) in Proposition 7 yields

**Proposition 14** (Stationary queue-length distribution function in an M/M/c queue).

If \( \rho < c \) then

\[
\pi(i) = \begin{cases} 
  \pi(0) \frac{\rho^i}{i!} & \text{if } i = 0,1,\ldots,c \\
  \pi(0) \frac{\rho^i}{c!} e^{c-i} & \text{if } i \geq c
\end{cases} \quad (70)
\]

where

\[
\pi(0) = \left[ \sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \left( \frac{\rho^c}{c!} \right) \left( \frac{1}{1-\rho/c} \right) \right]^{-1}. \quad (71)
\]

The probability that an arriving customer is forced to join the queue is given by

\[
P(\text{queueing}) = \sum_{i=c}^{\infty} \pi(i) = \sum_{i=c}^{\infty} \pi(0) \frac{\rho^i}{c!} e^{c-i}. \]

Thus,

\[
P(\text{queueing}) = \frac{\left( \frac{\rho^c}{c!} \right) \left( \frac{1}{1-\rho/c} \right)}{\sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \left( \frac{\rho^c}{c!} \right) \left( \frac{1}{1-\rho/c} \right)}. \quad (72)
\]

This probability is of wide use in telephony and gives the probability that no trunk (i.e., server) is available for an arriving call (i.e., customer) in a system of \( c \) trunks. It is referred to as *Erlang’s C formula*.
5.4 The M/M/c/c queue

Here we have a situation when there are \( c \geq 1 \) available servers but no waiting room. This is a pure loss queueing system. Each newly arriving customer is given its private server; however, if a customer arrives when all the servers are occupied, that customer is lost. Parameters \( \lambda \) and \( \mu \) are defined as in the previous sections.

The number of busy servers can be modeled as a birth and death process (use the construction rule) with birth rate

\[
\lambda_i = \begin{cases} 
\lambda & \text{if } i = 0, 1, \ldots, c - 1 \\
0 & \text{if } i \geq c
\end{cases}
\]

and death rate \( \mu_i = i\mu \) for \( i = 1, 2, \ldots, c \).

We are interested in determining the limiting distribution function \( \pi(i) \) \( (i = 0, 1, \ldots, c) \) of the number of busy servers.

**Proposition 15** (Stationary server occupation in a M/M/c/c queue).

\[
\pi(i) = \pi(0) \frac{\rho^i}{i!} \quad \text{for } i = 0, 1, \ldots, c, \pi(i) = 0 \text{ for } i > c, \text{ where }
\]

\[
\pi(0) = \left[ \sum_{i=0}^{c} \frac{\rho^i}{i!} \right]^{-1}
\]

This system is also of great interest in telephony. In particular, \( \pi(c) \) gives the probability that all trunks (i.e., servers) are busy, and it is given by

\[
\pi(c) = \frac{\rho^c/c!}{\sum_{i=0}^{c} \rho^i/i!}
\]

This is the celebrated *Erlang’s loss formula* (derived by A. K. Erlang in 1917).

Remarkably enough Proposition 15 is valid for any service time distribution and not only for exponential service times! Such a property is called an *insensitivity property.*

Later on, we will see an extremely useful extension of this model to (in particular) several classes of customers, that has nice applications in the modeling and performance evaluation of multimedia networks.

5.5 The repairman model

It is one of the most useful models. There are \( K \) machines and a single repairman. Each machine breaks down after a time that is exponentially distributed with parameter \( \alpha \). In other words, \( \alpha \) is the rate at which a machine breaks.

When a breakdown occurs, a request is sent to the repairman for fixing it. Requests are buffered. It takes an exponentially distributed amount of time with parameter \( \mu \) for the repairman to repair a machine. In other words, \( \mu \) is the repair rate.

We assume that “lifetimes” and repair times are all mutually independent.

What is the probability \( \pi(i) \) that \( i \) machines are up (i.e., working properly)? What is the overall failure rate?

Let \( X(t) \) be the number of machines up at time \( t \). It is easily seen that \( \{X(t), t \geq 0\} \) is a birth and death process with birth and death rates given by \( \lambda_n = \mu \) for \( n = 0, 1, \ldots, K - 1 \), \( \lambda_n = 0 \) for \( n \geq K \) and \( \mu_n = n\alpha \) for \( n = 1, 2, \ldots, K \), respectively.

We notice that \( \{X(t), t \geq 0\} \) has the same behavior as the queue-length process of an \( M/M/K/K \) queue! Hence, by (73) and (74) we find that

\[
\pi(i) = \frac{(\mu/\alpha)^i/i!}{C(K, \mu/\alpha)}
\]
for \( i = 0, 1, \ldots, K \), where
\[
C(K, a) := \sum_{i=0}^{K} \frac{a^i}{i!}.
\] (76)

The overall failure rate \( \lambda_b \) is given by
\[
\lambda_b = \sum_{i=1}^{K} \alpha_i \pi(i) = \alpha \sum_{i=1}^{K} \frac{\mu/\alpha)^i}{i!} = \mu \frac{C(K-1, \mu/\alpha)}{C(K, \mu/\alpha)}.
\]

Observe that \( \pi(0) = 1/C(K, \mu/\alpha) \). Hence, the mean number \( n_r \) of machines repaired by unit of time is
\[
n_r = \mu (1 - \pi(K)) = \mu \left( 1 - \frac{\mu/\alpha)^K/K!}{C(K, \mu/\alpha)} \right) = \mu \frac{C(K-1, \mu/\alpha)}{C(K, \mu/\alpha)}.
\]

5.6 Little’s formula

So far we have only obtained results for the buffer occupation namely, the limiting distribution of the queue-length, the mean number of customers, etc. These performance measures are of particular interest for a system’s manager. What we would like to do now is to address performance issues from a user’s perspective, such as, for instance, response times and waiting times.

For this, we need to introduce the most used formula in performance evaluation.

**Proposition 16** (Little’s formula).

The long-term average number of customers in a stable system, denoted by \( \bar{N} \), is equal to the long-term average effective arrival rate, denoted by \( \lambda \), multiplied by the average time that a customer spends in the system, denoted by \( \bar{T} \). Or, expressed algebraically:
\[
\bar{N} = \lambda \bar{T}.
\] (77)

This formula is of great interest since very often one knows \( \bar{N} \) and \( \lambda \). This result does not make any specific assumption regarding the arrival distribution or the service time distribution; nor does it depend upon the number of servers in the system or upon the particular queueing discipline within the system.

This result has an intuitive explanation: \( \bar{N} / \bar{T} \) can be interpreted as the departure rate, which has to be equal to the input rate \( \lambda \) since the system is in steady-state.

**Example 3.** Consider an M/M/1 queue with arrival rate \( \lambda \) and service rate \( \mu \). Let \( \bar{T} \) (resp. \( \bar{W} \)) be the mean customer sojourn time, also referred to as the mean customer response time (resp. waiting time).

If \( \rho := \lambda/\mu < 1 \) (i.e., if the queue is stable) then we know that the mean number of customers \( N \) is given by
\[
N = \rho / (1 - \rho) \quad \text{(see 66)}.
\]

Therefore, by Little’s formula,
\[
\bar{T} = \frac{\rho}{\lambda (1 - \rho)} = \frac{1}{\mu (1 - \rho)} \quad \text{(78)}
\]
\[
\bar{W} = \bar{T} - \frac{1}{\mu} = \frac{\rho}{\mu (1 - \rho)}.
\] (79)

Observe that both \( \bar{T} \to \infty \) and \( \bar{W} \to \infty \) when \( \rho \to 1 \).

\[\heartsuit\]
These dates form an increasing sequence of times \((t_n)_{n=1}^{2k}\) such that
\[a_1 = t_1 < t_2 < \cdots < t_{2k-1} < t_{2k} = d_k.\]

The mean sojourn time \(\bar{T}\) of a customer in \((0, C)\) is by definition
\[
\bar{T} = \frac{1}{k} \sum_{i=1}^{k} (d_i - a_i)
\]

since \(d_i - a_i\) is the time spent in the system by the \(i\)-th customer.

Let us now compute \(\bar{N}\), the mean number of customers in the system in \((0, C)\). Denote by \(N(t)\) the number of customers at time \(t\). Then,
\[
\bar{N} = \frac{1}{C} \int_0^C N(t) \, dt
\]
\[
= \frac{1}{C} \sum_{i=1}^{2k-1} N(t_i^+) (t_{i+1} - t_i)
\]

where \(N(t^+)\) is the number of customers in the system \textit{just after} time \(t\).

It is not difficult to see (make a picture) that
\[
\sum_{i=1}^{k} (d_i - a_i) = \sum_{i=1}^{2k-1} N(t_i^+) (t_{i+1} - t_i).
\]

Hence,
\[
\bar{N} = \frac{k}{C} \bar{T}.
\]

The proof is concluded as follows: since the system empties infinitely often we can choose \(C\) large enough so that \(k/C\) is equal to the arrival rate \(\lambda\). Hence, \(\bar{N} = \lambda \bar{T}\).

### 5.7 Comparing different multiprocessor systems

While designing a multiprocessor system we may wish to compare different systems.

The first system is an M/M/2 queue with arrival rate \(2 \lambda\) and service rate \(\mu\).

The second system is an M/M/1 queue with arrival rate \(2 \lambda\) and service rate \(2 \mu\).

Note that the comparison is fair since in both systems the traffic intensity denoted by \(\nu\) is given by \(\nu = \lambda/\mu\).

What system yields the smallest expected customer response time?

Let \(T_1\) and \(T_2\) be the expected customer response time in systems 1 and 2, respectively.

**Computation of \(T_1\):**

Denote \(\bar{N}_1\) by the mean number of customers in the system. From (70) and (71) we get that
\[
\pi(i) = 2 \pi(0) \nu^i \quad \forall i \geq 1,
\]
if \(\nu < 1\), from which we deduce that
\[
\pi(0) = \frac{1 - \nu}{1 + \nu}.
\]

Thus, for \(\nu < 1\),
\[
\bar{N}_1 = \sum_{i=1}^{\infty} i \pi(i)
\]
\[
= 2 \left(\frac{1 - \nu}{1 + \nu}\right) \sum_{i=1}^{\infty} i \nu^i
\]
by using the well-known identity $\sum_{i=1}^{\infty} i z^{i-1} = 1/(1-z)^2$ for all $0 \leq z < 1$.

From Little’s formula we deduce that

$$T_1 = \frac{\nu}{\lambda (1 - \nu)(1 + \nu)}$$

under the stability condition $\nu < 1$.

**Computation of $T_2$:**

For the M/M/1 queue with arrival rate $2\lambda$ and service rate $2\mu$ we have already seen in Example 3 (take $\rho = \nu$) that

$$T_2 = \frac{\nu}{2\lambda(1 - \nu)}$$

under the stability condition $\nu < 1$.

It is easily seen that $T_2 < T_1$ when $\nu < 1$.

### 5.8 The M/G/1 FIFO queue

This is a queue where customers are served according to the First-In-First-Out (FIFO) discipline, the arrivals are Poisson (rate $\lambda > 0$), and the successive customer service times are mutually independent with the *same*, arbitrary, cumulative distribution function $G(x)$. More precisely, if $\sigma_i$ and $\sigma_j$ are the service times of two customers, say customers $i$ and $j$, $i \neq j$, respectively, then

1. $\sigma_i$ and $\sigma_j$ are independent rvs
2. $G(x) = P(\sigma_i \leq x) = P(\sigma_j \leq x)$ for all $x \geq 0$.

Let $1/\mu$ be the mean service time, namely, $1/\mu = E[\sigma_i]$. The service times are further assumed to be independent of the arrival process.

As usual we will set $\rho = \lambda/\mu$.

For this queueing system, the process $(N(t), t \geq 0)$, where $N(t)$ is the number of customers in the queue at time $t$, is *not* a Markov process. This is because the probabilistic future of $N(t)$ for $t > s$ cannot be determined if one only knows $N(s)$, except if $N(s) = 0$ (consider for instance the case when the service times are all equal to the same constant).

#### 5.8.1 Mean queue-length and mean response time

We assume that the queue is empty at time $t = 0$. Customers are served according to the FIFO service discipline.

Let

- $0 < t_i$ be the arrival time of the $i$th customer;
- $W_i$ be the waiting time in queue of the $i$th customer;
- $\overline{W}$ be the expected waiting time in steady-state ($\overline{W} = \lim_{n \to \infty} (1/n) \sum_{i=1}^{n} W_i$ when this limit exists);
- $X(t)$ be the number of customers in the waiting room at time $t > 0$;
- $R(t)$ be the *residual service time* of the customer in the server at time $t$, if any. By convention, $R(t) = 0$ if the system is empty at time $t$;
- $\sigma_i$ the service time of customer $n$. Note that $E[\sigma_n] = 1/\mu$.

We will assume by convention that $X(t_i)$ is the number of customers in the waiting room just *before* the arrival of the $i$-th customer. We have

$$E[W_i] = E[R(t_i)] + E \left[ \sum_{j=i-X(t_i)}^{i-1} \sigma_j \right]$$
To derive (80) we have used the fact that \( \sigma_j \) is independent of \( X(t_i) \) for \( j = i - X(t_i), \ldots, i - 1 \), which implies that \( E[\sigma_j | X(t_i) = k] = 1/\mu \). Indeed, \( X(t_i) \) only depends on the service times \( \sigma_j \) for \( j = 1, \ldots, i - X(t_i) - 1 \) and not on \( \sigma_j \) for \( j \geq i - X(t_i) \) since the service discipline is FIFO.

Letting now \( i \to \infty \) in (80) yields

\[
W = \overline{R}_a + \frac{\overline{X}_a}{\mu} \tag{81}
\]

with

- \( \overline{R}_a := \lim_{i \to \infty} E[R(t_i)] \) is the mean service time at arrival epochs in steady-state, and
- \( \overline{X}_a := \lim_{i \to \infty} E[X(t_i)] \) is the mean number of customers in the waiting room at arrival epochs in steady-state.

Because the arrival process is a Poisson process (PASTA property: Poisson Arrivals See Times Averages), we have that

\[
\overline{R} := \lim_{t \to \infty} \frac{1}{t} \int_0^t R(s) \, ds = \overline{R}_a \tag{82}
\]

\[
\overline{X} := \lim_{t \to \infty} \frac{1}{t} \int_0^t X(s) \, ds = \overline{X}_a. \tag{83}
\]

We shall not prove these results.\(^6\)

In words, (82) says that the mean residual service times at arrival epochs and at arbitrary epochs are the same. Similarly, (83) expresses the fact that the mean number of customers at arrival epochs and at arbitrary epochs are the same.

**Example 4.** If the arrivals are not Poisson then formulae (82) and (83) are in general not true. Here is an example where (82) is not true: assume that the \( n \)th customer arrives at time \( t_n = n \) seconds \((s)\) for all \( n \geq 1 \) and that it requires \( 0.999 \)s of service \((i.e., \sigma_n = 0.999s)\). If the system is empty at time 0, then clearly \( R(t_n) = 0 \) for all \( n \geq 1 \) since an incoming customer will always find the system empty, and therefore the left-hand side of (82) is zero; however, since the server is always working in \((n, n + 0.999)\) for all \( n \geq 1 \) it should be clear that the right-hand side of (82) is \( (0.999)^2/2 \).

Applying Little's formula to the waiting room yields

\[
\overline{X} = \lambda \overline{W}
\]

so that, cf. (81),

\[
\overline{W} (1 - \rho) = \overline{R}. \tag{84}
\]

From now on we will assume that \( \rho < 1 \). Hence, cf. (84),

\[
\overline{W} = \frac{\overline{R}}{1 - \rho}. \tag{85}
\]

The condition \( \rho < 1 \) is the stability condition of the M/G/1 queue. This condition is again very natural. We will compute \( \overline{R} \) under the assumption that the queue empties infinitely often (it can be shown that this occurs with probability 1 if \( \rho < 1 \)). Let \( C \) be a time when the queue is empty and define \( Y(C) \) to be the number of customers served in \((0, C)\).

\(^6\)The proof is not technically difficult but beyond the scope of this course. Please ask me for references if interest.
We have (hint: display the curve \( t \to R(t) \)):

\[
R = \lim_{C \to \infty} \frac{1}{C} \sum_{n=1}^{Y(C)} \frac{\sigma_i^2}{2} = \lim_{C \to \infty} \left( \frac{Y(C)}{C} \right) \lim_{C \to \infty} \left( \frac{1}{Y(C)} \sum_{n=1}^{Y(C)} \frac{\sigma_i^2}{2} \right) = \frac{\lambda E[\sigma^2]}{2}
\]

where \( E[\sigma^2] \) is the second-order moment of the service times (i.e., \( E[\sigma^2] = E[\sigma_i^2] \) for all \( i \geq 1 \)). Hence, for \( \rho < 1 \),

\[
W = \frac{\lambda E[\sigma^2]}{2 (1 - \rho)} \quad (86)
\]

This formula is the Pollaczek-Khinchin (abbreviated as P-K) formula for the mean waiting in an M/G/1 queue.

Thus, the mean system response time \( \bar{T} \) is given by

\[
\bar{T} = \frac{1}{\mu} + \frac{\lambda E[\sigma^2]}{2 (1 - \rho)} \quad (87)
\]

and, by Little’s formula, the mean number of customers \( E[N] \) in the entire system (waiting room + server) is given by

\[
E[N] = \rho + \frac{\lambda^2 E[\sigma^2]}{2 (1 - \rho)} \quad (88)
\]

Consider the particular case when \( P(\sigma_i \leq x) = 1 - \exp(-\mu x) \) for \( x \geq 0 \), that is, the M/M/1 queue. Since \( E[\sigma^2] = 2/\mu^2 \), we see from (86) that

\[
W = \frac{\lambda}{\mu^2 (1 - \rho)} = \frac{\rho}{\mu (1 - \rho)}
\]

which agrees with (79).

It should be emphasised that \( W, T \) and \( E[N] \) now depend upon the first two moments \( (1/\mu \) and \( E[\sigma^2] \)) of the service time distribution function (and of course upon the arrival rate). This is in constrast with the M/M/1 queue where these quantities only depend upon the mean of the service time (and upon the arrival rate).

### 6 Priority Queues

Jobs/programs in a computer systems or packets/sessions in a computer network may not be treated equally: some may receive preferential treatment. Queueing systems in which some customers get preferential treatment are called Priority Queueing Systems.

In the following, we will only focus on priority queueing systems with a single server. In a priority queueing system customers are divided into classes, say class 1, 2, \ldots, K if there are K different classes of customers. We will assume that the lower the priority class the higher the priority number. In other words, customers of priority class \( i \) are given preference over customers in priority class \( j \) if \( i < j \).

We will assume that the customers within a given priority class are served according to the FIFO service discipline, although this assumption may be relaxed in many cases (e.g., it can be relaxed when we study queue-length processes).

There are two basic priority policies:

1. the preemptive-resume priority policy
2. the non-preemptive priority policy.
Under a preemptive-resume priority policy service is interrupted if a newly arrived customer (say with priority $i$) has priority over the customer in the service facility (say with priority $j > i$), and the newly arrived customer begins service at once. The customer whose service was interrupted or preempted (call it customer $C_j$) returns to the head of the $j$ priority class. When there are no longer customers in priority classes $1, 2, \ldots, j-1$, then customer $C_j$ returns to the server and resumes his service at the point of interruption. Under a non-preemptive priority policy service is not interrupted $^7$.

Throughout this section, we consider a M/G/1 priority queue with $K$ classes of customers: customers of class (priority) $k$ arrive at the system arrive to a Poisson process with intensity $\lambda_k$ and have independent service times with common distribution function $G_k(x)$, mean $1/\mu_k$, and second-order moment $\sigma_k^2$ ($k = 1, 2, \ldots, K$). We assume that service times are mutually independent rvs, that the arrival processes are mutually independent, and that the service times and arrival processes are independent.

Our objective is to compute

- $T_k$, the expected sojourn time (or response time) of customers of class $k$
- $W_k$, the expected waiting time of customers of class $k$
- $N_k$, the expected number of customers of class $k$ in the system

for both the non-preemptive and the preemptive policies.

Some further notation: define $X_k$ as the mean number of customers of class $k$ in the waiting room and let $\rho_k := \lambda_k/\mu_k$ be the traffic intensity of customers of class $k$, $k = 1, 2, \ldots, K$.

### 6.1 Non-preemptive priority

The response time of a customer of class $k$ is simply the sum of its waiting time in the queue and of its service time. Therefore,

$$T_k = W_k + \frac{1}{\mu_k} \quad (89)$$

for $k = 1, 2, \ldots, K$.

Let us first compute $W_1$.

By the same argument used to compute the mean waiting time in an M/G/1 queue with one class of customers (see (85)) we have

$$W_1 = \frac{R}{1-\rho_1} \quad (90)$$

provided that $\rho_1 < 1$, where $R$ is the expected residual time.

[Hint: write $W_1$ as $W_1 = R + (1/\mu_1) E[X_1]$ and note that $X_1 = \lambda_1 W_1$ by Little’s formula, from which (90) follows].

We will compute $R$ later on. Let us now compute $W_2$.

$W_2$ is the sum of four terms: the mean residual service time, the expected time to serve all customers of class 1 (resp. class 2) in the system when a class 2 customer arrives, and the expected time to serve all customers of class 1 that will arrive during the total waiting time in the waiting room of a class 2 customer. Thus,

$$W_2 = R + \frac{1}{\mu_1} X_1 + \frac{1}{\mu_2} X_2 + \frac{1}{\mu_1} E[Z_1] \quad (91)$$

where $Z_1$ is the number of customers of class 1 that arrive during the wait in queue of a customer of class 2. By Little’s formula

$$X_1 = \lambda_1 W_1$$

$$X_2 = \lambda_2 W_2.$$
Denote by $W_2$ be the time spent in the waiting by a customer of class 2. We have
\[
E[Z_1] = \int_0^\infty E[Z_1 \mid W_2 = x] P(W_2 \in dx)
= \int_0^\infty \lambda_1 x P(W_2 \in dx)
= \lambda_1 W_2
\]
where (92) follows from the assumption that the arrival process of class 1 customers is Poisson with rate $\lambda_1$ (see Appendix, Proposition C.2).

Hence,
\[
W_2 = \frac{\overline{R} + \rho_1 W_1}{1 - \rho_1 - \rho_2}
\]
provided that $\rho_1 + \rho_2 < 1$.

Using the value of $W_1$ obtained in (90) gives
\[
W_2 = \frac{\overline{R}}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}
\]
provided that $\rho_1 + \rho_2 < 1$.

By repeating the same argument, it is easily seen that
\[
W_k = \frac{\overline{R}}{(1 - \sum_{j=1}^{k-1} \rho_j)(1 - \sum_{j=1}^{k} \rho_j)}
\]
provided that $\sum_{j=1}^{k} \rho_j < 1$.

Hence, it is seen that the condition $\rho := \sum_{j=1}^{K} \rho_j < 1$ is the stability condition of this model. From now on we will assume that $\rho < 1$.

Let us now compute $\overline{R}$, the expected residual service time. This quantity is nothing but the expected residual service time in an ordinary M/G/1 queue with arrival intensity $\sum_{j=1}^{K} \lambda_j$ (since the superposition of $K$ independent processes with intensities $\lambda_1, \ldots, \lambda_K$ is a Poisson process with intensity $\sum_{j=1}^{K} \lambda_j$) and service time distribution given by $\sum_{j=1}^{K} \lambda_j G_j(x)/\sum_{j=1}^{K} \lambda_j$ (since with probability $\lambda_k/\sum_{j=1}^{K} \lambda_j$ an arrival customer is of class $k$). In particular, the second-order moment of the service times is $\sum_{j=1}^{K} \lambda_j \sigma_j^2/\sum_{j=1}^{K} \lambda_j$. Note that the traffic intensity in this ordinary M/G/1 is the same as in the original priority queue, namely, $\sum_{j=1}^{K} \rho_j$.

Substituting $\lambda$ by $\sum_{j=1}^{K} \lambda_j$, $\rho$ by $\sum_{j=1}^{K} \rho_j$ and $E[\sigma^2]$ by $\sum_{j=1}^{K} \lambda_j \sigma_j^2/\sum_{j=1}^{K} \lambda_j$ in (86), we find that
\[
\overline{R} = \sum_{j=1}^{K} \lambda_j \sigma_j^2/2
\]
(94)

Finally, cf. (93), (94),
\[
W_k = \frac{\sum_{j=1}^{K} \lambda_j \sigma_j^2}{2 \left(1 - \sum_{j=1}^{k-1} \rho_j\right) \left(1 - \sum_{j=1}^{k} \rho_j\right)}
\]
(95)
for all $k = 1, 2, \ldots, K$, and, by (98)
\[
T_k = \frac{1}{\mu_k} + \frac{\sum_{j=1}^{K} \lambda_j \sigma_j^2}{2 \left(1 - \sum_{j=1}^{k-1} \rho_j\right) \left(1 - \sum_{j=1}^{k} \rho_j\right)}
\]
(96)

Applying again Little’s formula to (96) gives
\[
\overline{N}_k = \rho_k + \frac{\lambda_k \sum_{j=1}^{K} \lambda_j \sigma_j^2}{2 \left(1 - \sum_{j=1}^{k-1} \rho_j\right) \left(1 - \sum_{j=1}^{k} \rho_j\right)}
\]
(97)
for all $k = 1, 2, \ldots, K$. 

32
6.2 Preemptive-resume priority

We consider the same model as in the previous section but we now assume that the policy is preemptive-resume. We keep the same notation and we assume that $\rho < 1$.

Our objective is to compute $T_k$ and $\overline{N}_k$.

Let us first consider $T_1$. Actually, there is nothing to compute since $T_1$ is simply the sojourn time in an ordinary M/G/1 queue from the very definition of a preemptive-resume priority policy, that is,

$$T_1 = \frac{1}{\mu_1} + \frac{\lambda_1 \sigma_1^2}{2(1 - \rho_1)}.$$  \hfill (98)

The sojourn of a customer a class $k$, $k \geq 2$, is the sum of three terms:

$$T_k = \frac{1}{\mu_k} + T_{k,1} + T_{k,2}$$

where

1. $T_{k,1}$ is the expected time required to service all customers of class 1 to $k$ which are in the system at the arrival of a class $k$ customer;

2. $T_{k,2}$ is the expected cumulated sojourn times of customers of class 1 to $k - 1$ who arrive while a customer of class $k$ is in the system.

**Computation of $T_{k,1}$:**

When a customer of class $k$ arrives his waiting time before entering the server for the first time is the same as his waiting time in an ordinary M/G/1 queue (without priority) where customers of class $k + 1, \ldots, K$ are neglected (i.e., $\lambda_i = 0$ for $i = k + 1, \ldots, K$). The reason is that the sum of remaining service times of all customers in the system is independent of the service discipline of the system. This is true for any system where the server is always busy when the system is nonempty.

Hence, after substitution of $\lambda$ by $\sum_{i=1}^{k} \lambda_i$, $\rho$ by $\sum_{i=1}^{k} \rho_i$ and $E[\sigma^2]$ by $\sum_{i=1}^{k} \lambda_i \sigma_i^2 / \sum_{i=1}^{k} \lambda_i$ in (86), we find that

$$T_{k,1} = \frac{\sum_{i=1}^{k} \lambda_i \sigma_i^2}{2 \left(1 - \sum_{i=1}^{k} \rho_i\right)}.$$  \hfill (99)

**Computation of $T_{k,2}$:**

Let $\tau_k$ be the sojourn time of a class-$k$ customers. Note that $E[\tau_k] = T_k$. We have

$$T_{k,2} = \sum_{i=1}^{k-1} E[\text{nb. class-}i \text{ to arrive during } \tau_k] \frac{1}{\mu_i}$$

with

$$E[\text{nb. class-}i \text{ to arrive during } \tau_k] = \int_0^\infty E[\text{nb. class-}i \text{ to arrive during } \tau_k \mid \tau_k = x]P(\tau_k \in dx)$$

$$= \int_0^\infty \lambda_i x P(\tau_k \in dx)$$

$$= \lambda_i T_k$$

since class-$i$ arrivals are Poisson with rate $\lambda_i$ for $i = 1, \ldots, K$. Therefore,

$$T_{k,2} = T_k \sum_{i=1}^{k-1} \rho_i$$

Finally,

$$T_k = \frac{1}{1 - \sum_{i=1}^{k-1} \rho_i} \left( \frac{1}{\mu_k} + \frac{\sum_{i=1}^{k} \lambda_i \sigma_i^2}{2 \left(1 - \sum_{i=1}^{k} \rho_i\right)} \right)$$
as long as $\sum_{i=1}^{k} \rho_i < 1$ and, by Little’s formula,
\[
N_k = \frac{1}{1 - \sum_{i=1}^{k-1} \rho_i} \left( \rho_k + \frac{\lambda_k \sum_{i=1}^{k} \lambda_i \sigma_i^2}{2 \left( 1 - \sum_{i=1}^{k} \rho_i \right)} \right).
\]

7 Single-Class Queueing Networks

So far, the queueing systems we studied were only single resource systems: that is, there was one service facility, possibly with multiple servers. Actual computer systems and communication networks are multiple resource systems. Thus, we may have online terminals or workstations, communication lines, etc., as well as the computer itself. The computer, even the simplest personal computer, has multiple resources, too, including main memory, virtual memory, coprocessors, I/O devices, etc. There may be a queue associated with each of these resources. Thus, a computer system or a communication network is a network of queues.

A queueing network is open if customers enter from outside the network, circulate among the service centers (or queues or nodes) for service, and depart from the network. A queueing network is closed if a fixed number of customers circulate indefinitely among the queues. A queueing network is mixed if some customers enter from outside the network and eventually leave, and if some customers always remain in the network.

7.1 Networks of Markovian queues: open Jackson network

Consider an open network consisting of $K$ M/M/1 queues. Jobs arrive from outside the system joining queue $i$ according to a Poisson process with rate $\lambda_i^0$. After service at queue $i$, which is exponentially distributed with parameter $\mu_i$, the job either leaves the system with probability $p_{i0}$, or goes to queue $j$, with probability $p_{ij}$. Clearly, $\sum_{j=0}^{K} p_{ij} = 1$, since each job must go somewhere.

As usual, the arrival times and the service times are assumed to be mutually independent rvs.

Let $Q_i(t)$ be the number of customers in queue (or node) $i$ at time $t$ and define
\[
Q(t) = (Q_1(t), \ldots, Q_K(t)) \quad \forall t \geq 0.
\]

As usual we will be interested in the computation of
\[
\pi(n) = \lim_{t \to \infty} P(Q(t) = n)
\]

with $n := (n_1, \ldots, n_K) \in \mathbb{N}^K$.

Because of the Poisson and exponential assumptions the continuous-time, discrete-space stochastic process $(Q(t), t \geq 0)$ is seen to be a continuous-time Markov chain on the state-space $I = \mathbb{N}^K$.

The balance equations for this C-MC are (cf. (37))
\[
\pi(n) \left( \sum_{i=1}^{K} \lambda_i^0 + \sum_{i=1}^{K} 1(n_i > 0)(1-p_{i0})\mu_i \right) = \sum_{i=1}^{K} 1(n_i > 0) \lambda_i^0 \pi(n - e_i)
\]

\[
+ \sum_{i=1}^{K} p_{i0} \mu_i \pi(n + e_i)
\]

\[
+ \sum_{i=1}^{K} \sum_{j=1}^{K} 1(n_j > 0)p_{ij}\mu_i \pi(n + e_i - e_j)
\]

where $1(k > 0) = 1$ if $k > 0$ and 0 otherwise, and where $e_i$ is the vector with all components zero, except the $i$-th one which is one.

Proposition 17 (Open Jackson network).
Define \( \rho_i = \lambda_i / \mu_i \) for all \( i = 1, 2, \ldots, K \). If \( \lambda_i < \mu_i \) for all \( i = 1, 2, \ldots, K \) then

\[
\pi(n) = \prod_{i=1}^{K} (1 - \rho_i)^{n_i}, \quad \forall n = (n_1, \ldots, n_K) \in \mathbb{N}^K
\]  

(102)

where \( (\lambda_1, \lambda_2, \ldots, \lambda_K) \) is the unique nonnegative solution of the system of linear equations

\[
\lambda_i = \lambda_i^0 + \sum_{j=1}^{K} p_{ji} \lambda_j \quad i = 1, 2, \ldots, K.
\]  

(103)

In matrix form (103) writes

\[
(\lambda_1, \ldots, \lambda_K) = (\lambda_1^0, \ldots, \lambda_K^0) + (\lambda_1, \ldots, \lambda_K)P
\]  

(104)

with \( P = [p_{ij}]_{1 \leq i, j \leq K} \) the routing matrix. If the matrix \( (I - P)^{-1} \) exists, the unique solution of (104) is given by

\[
(\lambda_1, \ldots, \lambda_K) = (\lambda_1^0, \ldots, \lambda_K^0)(I - P)^{-1}.
\]  

(105)

The existence of matrix \( (I - P)^{-1} \) is equivalent to the irreducibility of the Markov process \( \{Q(t), t \geq 0\} \) on \( \mathbb{N}^K \). Therefore, the statement of Proposition 17 can be reformulated as

**Proposition 18** (Open Jackson network - alternative statement).

If the matrix \( I - P \) is invertible and if \( \lambda_i < \mu_i \) for \( i = 1, \ldots, K \) then (102) holds, where \( (\lambda_1, \ldots, \lambda_K) \) is given in (105).

Let us comment this fundamental result of queueing network theory obtained by J. R. Jackson in 1957. Equations (103) are referred to as the traffic equations. Let us show that \( \lambda_i \) is the total arrival rate at node \( i \) when the system is in steady-state. To do so, let us first determine the total throughput of a node. The total throughput of node \( i \) consists of the customers who arrive from outside the network with rate \( \lambda_i^0 \), plus all the customers who are transferred to node \( i \) after completing service at node \( j \) for all nodes in the network. If \( \lambda_i \) is the total throughput of node \( i \), then the rate at which customers arrive at node \( i \) from node \( j \) is \( p_{ji} \lambda_j \). Hence, the throughput of node \( i \), \( \lambda_i \) must satisfy (103).

Since, in steady-state, the throughput of every node is equal to the arrival rate at this node, we see that \( \lambda_i \) is also the total arrival rate in node \( i \).

Hence, the conditions \( \lambda_i < \mu_i \) for \( i = 1, 2, \ldots, K \), are the stability conditions of an open Jackson network.

Let us now discuss the form of the limiting distribution (102). We see that (102) is a product of terms, where the \( i \)-th term \( (1 - \rho_i)^{n_i} \) is actually the limiting distribution function of the number of customers in an isolated M/M/1 queue with arrival rate \( \lambda_i \) and service rate \( \mu_i \). This property is usually referred to as the product-form property.

Therefore, the network state probability (i.e., \( \pi(n) \)) is the product of the state probabilities of the individual queues.

It is important to note that the steady-state probabilities behave as if the total arrival process at every node (usually referred to as the flow) were Poisson (with rate \( \lambda_i \) for node \( i \)), but the flows are not Poisson in general! The flows are Poisson if and only if \( p_{i,i+1} = 1 \) for \( i = 1, 2, \ldots, K \), and \( p_{K0} = 1 \), that is, for a network of queues in series.

**Proof of Proposition 17:** Since the network is open we know that there exists \( i_0 \) such that \( p_{i0} > 0 \) since otherwise customers would stay forever in the network. This ensures that the matrix \( I - P \) is invertible, with \( P = [p_{ij}]_{1 \leq i, j \leq K} \), from which we may deduce that the balance equations (101) have a unique nonnegative solution \( (\lambda_1, \ldots, \lambda_K) \) (this proof is omitted).
On the other hand, because every node may receive and serve infinitely many customers, and because of the markovian assumptions it is seen that the C-M.C \((Q(t), t \geq 0)\) is irreducible (hint: one must show that the probability of going from state \((n_1, \ldots, n_K)\) to state \((m_1, \ldots, m_K)\) in exactly \(t\) units of time is strictly positive for all \(t > 0, n_i \geq 0, m_i \geq 0, i = 1, 2, \ldots, K\). Let \(s < t\). Since the probability of having 0 external arrival in \([0, s)\) is strictly positive we see that the probability that the system is empty at time \(s\) is also strictly positive. On the other hand, starting from an empty system, it should be clear that we can reach any state in exactly \(t - s\) units of time).

Hence, Proposition 5 applies to the irreducible C.-M.C. \((Q(t), t \geq 0)\). Thus, it suffices to check that \((102)\) satisfies the balance equations together with the condition \(\sum_{n \in \mathbb{N}^K} \pi(n) = 1\).

Observe from \((102)\) that the latter condition is trivially satisfied. It remains to check that \((102)\) satisfies the traffic equations \((101)\).

Observe that from \((102)\) that

\[
\begin{aligned}
\pi(n - e_i) &= \frac{\pi(n)}{\rho_i} \\
\pi(n + e_j) &= \pi(n) \rho_i \\
\pi(n + e_i - e_j) &= \frac{\pi(n) \rho_i}{\rho_j}.
\end{aligned}
\]

By plugging these values in the r.h.s. of \((101)\), then by dividing both sides of the resulting equation by \(\pi(n)\) yields

\[
\begin{aligned}
\sum_{i=1}^K \lambda_i^0 + \sum_{i=1}^K 1(n_i > 0) (1 - p_{ii}) \mu_i &= \sum_{i=1}^K 1(n_i > 0) \frac{\lambda_i^0 \mu_i}{\lambda_i} + \sum_{i=1}^K p_{i0} \lambda_i \\
&+ \sum_{i=1}^K \sum_{j \neq i}^K 1(n_j > 0) p_{ij} \frac{\lambda_i \mu_j}{\lambda_j} 
\end{aligned}
\]

that we can rewrite as

\[
\begin{aligned}
\sum_{i=1}^K \lambda_i^0 - \sum_{i=1}^K p_{i0} \lambda_i &= -\sum_{i=1}^K 1(n_i > 0) (1 - p_{ii}) \mu_i + \sum_{i=1}^K 1(n_i > 0) \frac{\lambda_i^0 \mu_i}{\lambda_i} \\
&+ \sum_{i=1}^K \sum_{j \neq i}^K 1(n_j > 0) p_{ij} \frac{\lambda_i \mu_j}{\lambda_j}. \tag{106}
\end{aligned}
\]

The l.h.s of \((106)\) is equal to 0. Indeed, by summing up the traffic equations \((103)\) over all values of \(i = 1, 2, \ldots, K\), we find that

\[
\begin{aligned}
\sum_{i=1}^K \lambda_i &= \sum_{i=1}^K \lambda_i^0 + \sum_{i=1}^K \sum_{j=1}^K p_{ji} \lambda_j \\
&= \sum_{i=1}^K \lambda_i^0 + \sum_{j=1}^K \lambda_j \sum_{i=1}^K p_{ji} \\
&= \sum_{i=1}^K \lambda_i^0 + \sum_{j=1}^K \lambda_j (1 - p_{j0})
\end{aligned}
\]

so that \(\sum_{i=1}^K \lambda_i^0 - \sum_{i=1}^K p_{i0} \lambda_i = 0\) as announced earlier.

Let us now show that the r.h.s. of \((106)\) is also equal to 0. We have

\[
-\sum_{i=1}^K 1(n_i > 0) (1 - p_{ii}) \mu_i + \sum_{i=1}^K 1(n_i > 0) \frac{\lambda_i^0 \mu_i}{\lambda_i} + \sum_{i=1}^K \sum_{j \neq i}^K 1(n_j > 0) p_{ij} \frac{\lambda_i \mu_j}{\lambda_j}
\]
\[
\begin{align*}
&= - \sum_{i=1}^{K} 1(n_i > 0) \mu_i + \sum_{i=1}^{K} 1(n_i > 0) \frac{\lambda_i^0 \mu_i}{\lambda_i} + \sum_{i=1}^{K} \sum_{j=1}^{K} 1(n_j > 0) p_{ij} \frac{\lambda_i \mu_j}{\lambda_j} \\
&= - \sum_{i=1}^{K} 1(n_i > 0) \mu_i + \sum_{i=1}^{K} 1(n_i > 0) \frac{\lambda_i^0 \mu_i}{\lambda_i} + \sum_{i=1}^{K} \sum_{j=1}^{K} 1(n_i > 0) p_{ij} \frac{\lambda_j \mu_i}{\lambda_i} \\
&= - \sum_{i=1}^{K} 1(n_i > 0) \frac{\mu_i}{\lambda_i} \left( \lambda_i - \lambda_i^0 - \sum_{j=1}^{K} p_{ji} \lambda_j \right) \\
&= 0
\end{align*}
\]
from (103). This concludes the proof.

This result actually extends to the case when the network consists of \( K \) M/M/c queues. Assume that node \( i \) is an M/M/c queue. The following result holds:

**Proposition 19 (Open Jackson network of M/M/c queues).**

Define \( \mu_i(r) = \mu_i \min(r, c_i) \) for \( r \geq 0 \), \( i = 1, 2, \ldots, K \) and let \( \rho_i = \lambda_i / \mu_i \) for \( i = 1, 2, \ldots, K \).

If \( \lambda_i < c_i \mu_i \) for all \( i = 1, 2, \ldots, K \), then

\[
\pi(n) = \prod_{i=1}^{K} C_i \left( \frac{\lambda_i^n}{\prod_{r=1}^{n} \mu_i(r)} \right) \quad \forall n = (n_1, \ldots, n_K) \in \mathbb{N}^K,
\]

where \( (\lambda_1, \lambda_2, \ldots, \lambda_K) \) is the unique nonnegative solution of the system of linear equations (103), and where \( C_i \) is given by

\[
C_i = \left[ \sum_{r=0}^{c_i-1} \rho_i^r \frac{\rho_i^{c_i}}{c_i!} \left( \frac{1}{1 - \rho_i/c_i} \right) \right]^{-1}.
\]

**Example 5.** Consider a switching facility that transmits messages to a required destination. A NACK (Negative ACKnowledgment) is sent by the destination when a packet has not been properly transmitted. If so, the packet in error is retransmitted as soon as the NACK is received.

We assume that the time to send a message and the time to receive a NACK are both exponentially distributed with parameter \( \mu \). We also assume that packets arrive at the switch according to a Poisson process with rate \( \lambda^0 \). Let \( p, 0 < p \leq 1 \), be the probability that a message is received correctly.

Thus, we can model this switching facility as a Jackson network with one node, where \( c_1 = 1 \) (one server), \( p_{10} = p \) and \( p_{11} = 1 - p \). By Jackson’s theorem we have that, \( \pi(n) \), the number of packets in the service facility in steady-state, is given by

\[
\pi(n) = \left( 1 - \frac{\lambda}{\mu} \right) \left( \frac{\lambda}{\mu} \right)^n, \quad n \geq 0
\]

provided that \( \lambda < \mu \), where \( \lambda \) is the solution of the traffic equation

\[
\lambda = \lambda^0 + (1 - p) \lambda.
\]

Therefore, \( \lambda = \lambda^0 / p \), and

\[
\pi(n) = \left( 1 - \frac{\lambda^0}{p \mu} \right) \left( \frac{\lambda^0}{p \mu} \right)^n, \quad n \geq 0
\]

provided that \( \lambda^0 < p \mu \).

The mean number of packets (denoted as \( \overline{X} \)) in the switching facility is then given by (see (66))

\[
\overline{X} = \frac{\lambda^0}{p \mu - \lambda^0}
\]
and by, Little’s formula, the mean response time (denoted as $T$) is

$$T = \frac{1}{p\mu - \lambda^0}.$$ 

\[\heartsuit\]

**Example 6.** We consider the model in Example 5 but we now assume that the switching facility is composed of $K$ nodes in series, each modeled as an $M/M/1$ queue with common service rate $\mu$. In other words, we now have a Jackson network with $K$ $M/M/1$ queues where $\lambda_i^0 = 0$ for $i = 2, 3, \ldots, K$ (no external arrivals at nodes 2, $\ldots$, $K$), $\mu_i = \mu$ for $i = 1, 2, \ldots, K$, $p_{ii+1} = 1$ for $i = 1, 2, \ldots, K - 1$, $p_{K,0} = p$ and $p_{K,1} = 1 - p$.

For this model, the traffic equations read

$$\lambda_i = \lambda_{i-1}$$

for $i = 2, 3, \ldots, K$, and

$$\lambda_1 = \lambda^0 + (1 - p) \lambda_K.$$

It is easy to see that the solution to this system of equations is

$$\lambda_i = \frac{\lambda^0}{p} \quad \forall i = 1, 2, \ldots, K.$$

Hence, by Jackson’s theorem, the joint distribution function $\pi(n)$ for the number of packets in the system is given by

$$\pi(n) = \left(\frac{p\mu - \lambda^0}{p\mu}\right)^K \left(\frac{\lambda^0}{p\mu}\right)^{n_1 + \cdots + n_K} \forall n = (n_1, n_2, \ldots, n_K) \in \mathbb{N}^K$$

provided that $\lambda^0 < p\mu$. In particular, the probability $q_{ij}(r, s)$ of having $r$ packets in node $i$ and $s$ packets in node $j > i$ is given by

$$q_{ij}(r, s) = \sum_{n_i \geq 0, n_j \geq 0, i \neq j} \pi(n_1, \ldots, n_{i-1}, r, n_{i+1}, \ldots, n_{j-1}, s, n_{j+1}, \ldots, n_K)$$

$$= \left(\frac{p\mu - \lambda^0}{p\mu}\right)^2 \left(\frac{\lambda^0}{p\mu}\right)^{r+s}.$$

Let us now determine for this model the expected sojourn time of a packet. Since queue $i$ has the same characteristics as an $M/M/1$ queue with arrival rate $\lambda^0/p$ and mean service time $1/\mu$, the mean number of packets (denoted as $X_i$) is given by

$$X_i = \frac{\lambda^0}{p\mu - \lambda^0}$$

for every $i = 1, 2, \ldots, K$. Therefore, the total expected number of packets in the network is

$$\sum_{i=1}^K E[X_i] = \frac{K \lambda^0}{p\mu - \lambda^0}.$$ 

Hence, by Little’s formula, the expected sojourn time is given by

$$T = \frac{1}{\lambda^0} \sum_{i=1}^K E[X_i] = \frac{K}{p\mu - \lambda^0}.$$ 

\[\heartsuit\]

**Example 7** (The open central server network). Consider a computer system with one CPU and several I/O devices. A job enters the system from the outside and then waits until its execution begins. During its execution by the CPU, I/O requests may be needed. When an I/O request has been fulfilled the job then returns to the CPU for additional treatment. If the latter is available then the service begins at once;
otherwise the job must wait. Eventually, the job is completed (no more I/O requests are requested) and it leaves the system.

We are going to model this system as an open Jackson network with 3 nodes: one node (node 1) for the CPU and two nodes (nodes 2 and 3) for the I/O devices. In other words, we assume that $K = 3$, $\lambda_i^0 = 0$ for $i = 2, 3$ (jobs cannot access the I/O devices directly from the outside) and $p_{21} = p_{31} = 1$, $p_{10} > 0$.

For this system the traffic equations are:

$$
\begin{align*}
\lambda_1 &= \lambda_1^0 + \lambda_2 + \lambda_3 \\
\lambda_2 &= \lambda_1 p_{12} \\
\lambda_3 &= \lambda_1 p_{13}.
\end{align*}
$$

The solution of the traffic equations is

$$
\pi(n) = \left(1 - \frac{\lambda_1^0}{\mu_1 p_{10}}\right) \left(\frac{\lambda_1^0}{\mu_1 p_{10}}\right)^{n_1} \prod_{i=2}^{3} \left(1 - \frac{\lambda_i^0}{\mu_i p_{10}}\right) \left(\frac{\lambda_i^0}{\mu_i p_{10}}\right)^{n_i} \quad \forall n = (n_1, n_2, n_3) \in \mathbb{N}^3
$$

and

$$
T = \frac{1}{\mu_1 p_{10} - \lambda_1^0} + \sum_{i=2}^{3} \frac{p_{1i}}{\mu_i p_{10} - \lambda_i^0}. 
$$

7.2 Networks of Markovian queues: closed Jackson networks

We now discuss closed Markovian queueing networks. In such networks the number of customers in the network is always constant: no customer may enter from the outside and no customer may leave the network. More precisely, a closed Jackson is an open Jackson network where $\lambda_i^0 = 0$ for $i = 1, 2, \ldots, K$. However, because the number of customers is constant in a closed Jackson network— and assumed equal to $N$— a particular treatment is needed. Indeed, letting $\lambda_i^0 = 0$ for $i = 1, 2, \ldots, K$ in Proposition 19 does not yield the correct result (see Proposition 20 below).

Without loss of generality we shall assume that each node in the network is visited infinitely often by the customers (simply remove the nodes that are only visited a finite number of times).

For this model, the balance equations are

$$
\pi(n) \sum_{i=1}^{K} 1(n_i > 0) \mu_i = \sum_{i=1}^{K} \sum_{j=1}^{K} 1(n_i \leq N - 1, n_j > 0) p_{ij} \mu_i \pi(n + e_i - e_j) 
$$

for all $n = (n_1, \ldots, n_K) \in \{0, 1, \ldots, N\}^K$ such that $\sum_{i=1}^{K} n_i = N$.

**Proposition 20** (Closed Jackson network).

*Let $(\lambda_1, \ldots, \lambda_K)$ be any non-zero solution of the equations

$$
\lambda_i = \sum_{j=1}^{K} p_{ji} \lambda_j \quad i = 1, 2, \ldots, K. 
$$

Then,

$$
\pi(n) = \frac{1}{G(N, K)} \prod_{i=1}^{K} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} \quad \forall n = (n_1, \ldots, n_K) \in S(N, K),
$$

for all $n = (n_1, \ldots, n_K) \in S(N, K)$, with

$$
S(N, K) := \left\{ n = (n_1, \ldots, n_K) \in \{0, 1, \ldots, N\}^K \text{ such that } \sum_{i=1}^{K} n_i = N \right\}
$$

39
where
\[
G(N, K) := \sum_{\mathbf{n} \in S(N, K)} \prod_{i=1}^{K} \left( \frac{\lambda_i}{\mu_i} \right)^{n_i}.
\]

The constant \(G(N, K)\) is called the normalization constant. It has been named like this since it ensures that
\[
\sum_{\mathbf{n} \in S(N, K)} \pi(\mathbf{n}) = 1.
\]

Unlike the corresponding result for the open Jackson result, (111) shows that the number of customers in two different nodes in steady-state are not independent rvs. This follows from the fact that the right-hand side of (111) does not write as a product of terms of the form \(f_1(n_1) \times \cdots \times f_K(n_K)\).

This result is obvious: assume that there are \(n_i\) customers in node \(i\). Then, the number of customers in node \(j\) is necessarily less than or equal to \(M - n_i\). Thus, the rvs \(X_i\) and \(X_j\) representing the number of customers in nodes \(i\) and \(j\), respectively, cannot be independent (take \(K = 2\); then \(P(X_j = M) = 0\) if \(X_i = 1\) whereas \(P(X_j = M) = 1\) if \(X_i = 0\)).

Nevertheless, and with a slight abuse of notation, (111) is usually referred to as a product-form theorem for closed Jackson networks.

**Proof of Proposition 20:** Check that (111) satisfies (109) and apply Proposition 5.

Like for open Jackson networks, Proposition 20 extends to closed Jackson network of \(M/M/c\) queues (we keep the notation introduced earlier) We have:

**Proposition 21** (Closed Jackson network of \(M/M/c\) queues).

Let \((\lambda_1, \ldots, \lambda_K)\) be any non-zero solution to the traffic equations (110). Then,
\[
\pi(\mathbf{n}) = \frac{1}{G(N, K)} \prod_{i=1}^{K} \left( \frac{\lambda_i^{n_i}}{\prod_{r=1}^{n_i} \mu_i(r)} \right),
\]
for all \(\mathbf{n} = (n_1, \ldots, n_K) \in S(N, K)\), where
\[
G(N, K) := \sum_{\mathbf{n} \in S(N, K)} \prod_{i=1}^{K} \left( \frac{\lambda_i^{n_i}}{\prod_{r=1}^{n_i} \mu_i(r)} \right).
\]

The computation of the normalizing constant \(G(N, K)\) seems to be an easy task: it suffices to add a bunch of terms and to do a couple of multiplications. Yes? Well...

Assume that \(K = 5\) (five nodes) and \(N = 10\) (10 customers). Then, the set \(S(5, 10)\) already contains 1001 elements. If \(K = 10\) (ten queues) and \(N = 35\) customers, then \(S(10, 35)\) contains 52,451,256 terms! In addition, each term requires the computation of ten constants so that the total number of multiplications is over half a billion! (A communication network may have several hundredths of nodes...) More generally, it is not difficult to see that \(S(N, K)\) has \(\binom{N+K-1}{K} \) elements.

Therefore, a brute force approach (direct summation) may be both very expensive and numerically unstable. There exist stable and efficient algorithms to compute \(G(N, K)\). The first algorithm was obtained by J. Buzen in 1973 and almost every year one (or sometimes several) new algorithms appear! (for, however, more general “product-form” queueing networks to be described later on in this course).

### 7.2.1 The convolution algorithm

In this section we present the so-called convolution algorithm for the computation of \(G(N, K)\). It is due to J. Buzen.
By definition

\[ G(n, m) = \sum_{n \in S(n, m)} \prod_{i=1}^{m} \left( \frac{\lambda_i}{\mu_i} \right)^{n_i} \]

\[ = \sum_{k=0}^{n} \sum_{n_{m} = k}^{n} \prod_{i=1}^{m} \left( \frac{\lambda_i}{\mu_i} \right)^{n_i} \]

\[ = \sum_{k=0}^{n} \left( \frac{\lambda_m}{\mu_m} \right)^{k} \sum_{n \in S(n-k, m-1)}^{n} \prod_{i=1}^{m-1} \left( \frac{\lambda_i}{\mu_i} \right)^{n_i} \]

\[ = \sum_{k=0}^{n} \left( \frac{\lambda_m}{\mu_m} \right)^{k} G(n-k, m-1). \quad (115) \]

From the definition of \( G(n, m) \) the initial conditions for the algorithm are

\[ G(n, 1) = \left( \frac{\lambda_1}{\mu_1} \right)^{n} \quad \text{for } n = 0, 1, \ldots, N \]

\[ G(0, m) = 1 \quad \text{for } m = 1, 2, \ldots, K. \]

This convolution-like expression accounts for the name “convolution algorithm”. A similar algorithm exists for a closed Jackson network with M/M/c queues.

### 7.2.2 Performance measures from normalization constants

Performance evaluation can be expressed as functions of the equilibrium state probabilities. Unfortunately, this approach can lead to the same problems of excessive and inaccurate computations that were encountered in the calculation of the normalization constant. Fortunately, a number of important performance measures can be computed as functions of the various normalization constants which are a product of the convolution algorithm. In this section, it will be shown how this can be done.

#### Marginal Distribution of Queue-Length

Denote by \( X_i \) the number of customers in node \( i \) in steady-state. Define \( \pi_i(k) = P(X_i = k) \) be the steady-state p.f. of \( X_i \).

We have

\[ \pi_i(k) = \sum_{n \in S(N,K), n_i = k} \pi(n). \]

To arrive at the marginal distribution \( \pi_i \) it will be easier to first calculate

\[ P(X_i \geq k) = \sum_{n \in S(N,K), n_i \geq k} \pi(n) \]

\[ = \sum_{n \in S(N,K), n_i \geq k} \frac{1}{G(N,K)} \prod_{j=1}^{K} \left( \frac{\lambda_j}{\mu_j} \right)^{n_j}. \]

\[ = \left( \frac{\lambda_i}{\mu_i} \right)^{k} \frac{1}{G(N,K)} \sum_{n \in S(N-k,K)} \prod_{j=1}^{K} \left( \frac{\lambda_j}{\mu_j} \right)^{n_j} \]

\[ = \left( \frac{\lambda_i}{\mu_i} \right)^{k} \frac{G(N-k,K)}{G(N,K)}. \quad (116) \]

Now the key to calculating the marginal distribution is to recognize that

\[ P(X_i = k) = P(X_i \geq k) - P(X_i \geq k + 1) \]
so that
\[ \pi_i(k) = \left( \frac{\lambda_i}{\mu_i} \right)^k \frac{1}{G(N, K)} \left[ G(N - k, K) - \left( \frac{\lambda_i}{\mu_i} \right) G(N - k - 1, K) \right]. \]

### Expected Queue-Length

Perhaps the most useful statistic that can be derived from the marginal queue-length distribution is its mean. Recall the well-known formula for the mean of a rv X with values in N:

\[ E[X] = \sum_{k \geq 1} P(X \geq k). \]

Therefore, cf. (116),

\[ E[X_i] = \sum_{k=1}^{N} \left( \frac{\lambda_i}{\mu_i} \right)^k \frac{G(N - k, K)}{G(N, K)}. \]

### Utilization

The utilization of node \( i \) — denoted as \( U_i \) — is defined to be the probability that node \( i \) is non-empty in steady-state, namely, \( U_i = 1 - \pi_i(0) \).

From (117), we have that

\[ U_i = \left( \frac{\lambda_i}{\mu_i} \right) \frac{G(N - 1, K)}{G(N, K)}. \]

### Throughput

The throughput \( T_i \) of node \( i \) is defined as

\[ T_i = \sum_{k=1}^{N} \pi_i(k) \mu_i \]

or, equivalently, \( T_i = \mu_i (1 - \pi_i(0)) = \mu_i U_i \).

Therefore, cf. (119),

\[ T_i = \lambda_i \frac{G(N - 1, K)}{G(N, K)}. \]

### 7.3 The central server model

This is one a very useful model. There is one CPU (node M+1), K I/O devices (nodes M+2,\ldots,M+K+1), and M terminals (nodes 1,2,\ldots,K) that send jobs to the CPU.

When a user (i.e., terminal) has sent a job to the CPU he waits the end of the execution of his job before sending a new job. (we could also consider the case when a user is working on several jobs at a time). Hence, there are exactly \( K \) jobs in the network. This is a closed network.

For the time being, we are going to study a variant of this system. The central server system will be studied later on in this course.

We assume that

\[
\begin{align*}
p_{i,i+1} &= 1 & \text{for } i = M + 2, \ldots, M + K + 1 \\
p_{i,M+1} &= 1 & \text{for } i = 1, 2, \ldots, M \\
p_{M+1,i} &> 0 & \text{for } i = 1, 2, \ldots, M + K + 1 \\
p_{ij} &= 0 & \text{otherwise.}
\end{align*}
\]

Note here that several jobs may be waiting at a terminal.

Let \( \mu_i \) be the service rate at node \( i \) for \( i = 1, 2, \ldots, M + K + 1 \). For \( i = 1, 2, \ldots, M, 1/\mu_i \) can be thought of as the mean thinking time of the user before sending a new job to the CPU.

For this model, the traffic equations are:

\[
\lambda_i = \begin{cases} 
\lambda_{M+1} p_{M+1,i} & \text{for } i \neq M + 1 \\
\sum_{j=1}^{M+K+1} \lambda_j p_{j,M+1} & \text{for } i = M + 1
\end{cases}
\]

72x673}
Setting (for instance) \( \lambda_{M+1} = 1 \) yields \( \lambda_i = p_{M+1} \) for all \( i = 1, 2, \ldots, M + K + 1, i \neq M + 1 \).

The mean performance measures \( E[X_i], U_i \) and \( T_i \) for \( i = 1, 2, \ldots, M + K + 1 \), then follow from the previous section.

8 Multiclass Queueing Networks

We are now going to consider more general queueing networks that still enjoy the product-form property.

8.1 Multiclass open/closed/mixed Jackson queueing networks

The class of systems under consideration contains an arbitrary but finite number of nodes \( K \). There is an arbitrary but finite number \( R \) of different classes of customers. Customers travel through the network and change class according to transition probabilities. Thus a customer of class \( r \) who completes service at node \( i \) will next require service at node \( j \) in class \( s \) with a certain probability denoted \( p_{i,r;j,s} \).

We shall say that the pairs \((i, r)\) and \((j, s)\) belong to the same subchain if the same customer can visit node \( i \) in class \( r \) and node \( j \) in class \( s \). Let \( m \) be the number of subchains, and let \( E_1, \ldots, E_m \) be the sets of states in each of these subchains.

Let \( n_{ir} \) be the number of customers of class \( r \) at node \( i \). A closed system is characterized by

\[
\sum_{(i,r)\in E_i} n_{ir} = \text{constant}
\]

for all \( j = 1, 2, \ldots, m \). In other words, if the system is closed, then there is a constant number of customer circulating in all the subchains.

In an open system, customers may arrive to the network from the outside according to independent Poisson process. Let \( \lambda_{ir}^0 \) be the external arrival rate of customers of class \( r \) at node \( i \). In an open network a customer of class \( r \) who completes service at node \( i \) may leave the system. This occurs with the probability \( p_{i,r;0} \), so that \( \sum_s p_{i,r;j,s} + p_{i,r;0} = 1 \).

A subchain is said to be open if it contains at least one pair \((i, r)\) such that \( \lambda_{ir}^0 > 0 \); otherwise the subchain is closed. A network that contains at least one open subchain and one closed subchain is called a mixed network.

At node \( i, i = 1, 2, \ldots, K \), the service times are still assumed to be independent exponential rvs, all with the same parameter \( \mu_i \), for all \( i = 1, 2, \ldots, K \). We shall further assume that the service discipline at each node is FIFO (more general nodes will be considered later on).

Define \( Q(t) = (Q_1(t), \ldots, Q_K(t)) \) with \( Q_i(t) = (Q_{i1}(t), \ldots, Q_{iR}(t)) \), where \( Q_{ir}(t) \) denotes the number of customers of class \( r \) in node \( i \) at time \( t \).

The process \((Q(t), t \geq 0)\) is not a continuous-time Markov chain because the class of a customer leaving a node is not known.

Define \( X_i(t) = (I_{i1}(t), \ldots, I_{iQ_i(t)}(t)) \), where \( I_{ij}(t) \in \{1, 2, \ldots, R\} \) is the class of the customer in position \( j \) in node \( i \) at time \( t \). Then, the process \((X_1(t), \ldots, X_K(t)), t \geq 0)\) is a C.-M.C.

We can write the balance equations (or, equivalently, the infinitesimal generator) corresponding to this C.-M.C. (this is tedious but not difficult) and obtain a product-form solution. By aggregating states we may obtain from this result the limiting joint distribution function for \((Q_1(t), \ldots, Q_K(t))\), denoted as usual by \( \pi(\cdot) \).

The result is the following:

**Proposition 22** (Multiclass Open/Closed/Mixed Jackson Network). For \( k \in \{1, \ldots, m\} \) such that \( E_k \) is an open subchain, let \((\lambda_{ir})_{(i,r)\in E_k}\) be the unique strictly positive solution of the traffic equations

\[
\lambda_{ir} = \lambda_{ir}^0 + \sum_{(j,s)\in E_k} \lambda_{js} p_{j,s;i,r} \quad \forall (i,r) \in E_k.
\]

For every \( k \) in \( \{1, 2, \ldots, m\} \) such that \( E_k \) is a closed subchain, let \((\lambda_{ir})_{(i,r)\in E_k}\) be any non-zero solution of the equations

\[
\lambda_{ir} = \sum_{(j,s)\in E_k} \lambda_{js} p_{j,s;i,r} \quad \forall (i,r) \in E_k.
\]
If $\sum_{(r:i,r)} \lambda_{ir} < \mu_i$ for all $i = 1, 2, \ldots, K$ (stability condition), then

$$\pi(n) = \frac{1}{G} \prod_{i=1}^{K} n_i! \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left( \frac{\lambda_{ir}}{\mu_i} \right)^{n_{ir}}$$

(122)

for all $n = (n_1, \ldots, n_K)$ in the state-space, where $n_i = (n_{i1}, \ldots, n_{ir}) \in \mathbb{N}^R$ and $n_i = \sum_{r=1}^{R} n_{ir}$.

Let us illustrate this result through a simple example.

**Example 8.** There are two nodes (node 1 and node 2) and two classes of customers (class 1 and class 2). There are no external arrivals at node 2. External customers enter node 1 in class 1 with the rate $\lambda$. Upon service completion at node 1 a customer of class 1 is routed to node 2 with the probability 1. Upon service completion at node 2 a customer of class 1 leaves the system with probability 1. There are always exactly $K$ customers of class 2 in the system. Upon service completion at node 1 (resp. node 2) of customer of class 2 is routed back to node 2 (resp. node 1) in class 2 with the probability 1.

Let $\mu_i$ be the service rate at node $i$, $i = 1, 2$. The state-space $S$ for this system is

$$S = \{(n_{11}, n_{12}, n_{21}, n_{22}) \in \mathbb{N}^4 : n_{11} \geq 0, n_{21} \geq 0, n_{12} + n_{22} = K\}.$$

There are two subchains, $E_1$ and $E_2$, one open (say $E_1$) and one closed. Clearly, $E_1 = \{(1,1),(2,1)\}$ and $E_2 = \{(1,2),(2,2)\}$.

We find: $\lambda_{11} = \lambda_{21} = \lambda$ and $\lambda_{12} = \lambda_{22}$. Take $\lambda_{12} = \lambda_{22} = 1$, for instance.

The product-form result is:

$$\pi(n) = \frac{1}{G} \left( \frac{\lambda}{\mu_1} \right)^{n_{11}} \left( \frac{\lambda}{\mu_2} \right)^{n_{21}} \left( \frac{1}{\mu_1} \right)^{n_{12}} \left( \frac{1}{\mu_2} \right)^{n_{22}}$$

with $n \in S$, provided that $\lambda < \mu_i$ for $i = 1, 2$, i.e., $\lambda < \min(\mu_1, \mu_2)$ (stability condition).

Let us compute the normalizing constant $G$. By definition,

$$G = \sum_{n_{11} \geq 0, n_{21} \geq 0, n_{12} + n_{22} = K} \left( \frac{\lambda}{\mu_1} \right)^{n_{11}} \left( \frac{\lambda}{\mu_2} \right)^{n_{21}} \left( \frac{1}{\mu_1} \right)^{n_{12}} \left( \frac{1}{\mu_2} \right)^{n_{22}}$$

$$= \left( \sum_{n_{11} \geq 0} \frac{\lambda}{\mu_1} \right)^{n_{11}} \left( \sum_{n_{21} \geq 0} \frac{\lambda}{\mu_2} \right)^{n_{21}} \sum_{n_{12} + n_{22} = K} \left( \frac{1}{\mu_1} \right)^{n_{12}} \left( \frac{1}{\mu_2} \right)^{n_{22}}$$

$$= \left( \prod_{i=1}^{2} \frac{\mu_i}{\mu_i - \lambda} \right)^{n_{11}} \left( \frac{1}{\mu_1} \right)^{n_{12}} \sum_{i=0}^{K} \left( \frac{\mu_1}{\mu_2} \right)^{n_{22}} \sum_{i=0}^{K} \left( \frac{\mu_2}{\mu_1} \right)^{n_{12}} \sum_{i=0}^{K} \left( \frac{\mu_1}{\mu_2} \right)^{n_{22}}$$

Thus,

$$G = \frac{K+1}{\mu^K} \left( \frac{\mu}{\mu - \lambda} \right)^2$$

if $\mu_1 = \mu_2 := \mu$, and

$$G = \left( \prod_{i=1}^{2} \frac{\mu_i}{\mu_i - \lambda} \right)^{n_{11}} \left( \frac{1}{\mu_1} \right)^{n_{12}} \frac{1 - (\mu_1/\mu_2)^K + 1}{1 - (\mu_1/\mu_2)}$$

if $\mu_1 \neq \mu_2$.

If $K = 0$, then $G = \prod_{i=1}^{2} (\mu_i/(\mu_i - \lambda))$ as expected (open Jackson network with two M/M/1 nodes in series).

---

8Observe that this condition is automatically satisfied for node $i$ if this node is only visited by customers belonging to closed subchains.
The product-form result in Proposition 22 can be dramatically extended. We now give two simple extensions.

The first, not really surprising extension owing to what we have seen before, is the extension of the product-form result to multiclass open/closed/mixed Jackson networks with M/M/c queues. Let $c_i \geq 1$ be the number of servers at node $i$, and define $\alpha_i(j) = \min(c_i, j)$ for all $i = 1, 2, \ldots, K$. Hence, $\mu_i \alpha_i(j)$ is the service rate in node $i$ when there are $j$ customers.

Then, the product-form result (122) becomes (the stability condition is unchanged except that $\mu_i$ must be replaced by $\mu_i c_i$):

$$\pi(n) = \frac{1}{G} \prod_{i=1}^{K} \left[ \left( \prod_{j=1}^{n_i} \frac{1}{\alpha_i(j)} \right) n_i! \left( \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left( \frac{\lambda_{ir}}{\mu_i} \right)^{n_{ir}} \right) \right].$$  \hspace{1cm} (123)

The second extension is maybe more interesting. We may allow for state depending external arrival rates. Let us first introduce some notation: when the network is in state $n$ let $M(n)$ be the total number of customers in the network, that is, $M(n) = \sum_{i=1}^{K} n_i$.

We shall assume that the external arrival rate of customers of class $r$ at node $i$ maybe a function of the total number of customers in the network. More precisely, we shall assume that the external arrival rate of customers of class $r$ at node $i$ when the system is in state $n$ is of the form $\lambda_{ir}^0 \gamma(M(n))$, where $\gamma$ is an arbitrary function from $\mathbb{N}$ into $[0, \infty)$.

We have the following result:

The product-form property is preserved in a multiclass open/closed/mixed Jackson network with M/M/c nodes and state-dependent external arrivals as defined above, and is given by

$$\pi(n) = \frac{d(n)}{G} \prod_{i=1}^{K} \left[ \left( \prod_{j=1}^{n_i} \frac{1}{\alpha_i(j)} \right) n_i! \left( \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left( \frac{\lambda_{ir}^0}{\mu_i} \right)^{n_{ir}} \right) \right].$$  \hspace{1cm} (124)

for all $n$ in the state-space, where

$$d(n) = \prod_{j=0}^{M(n)-1} \gamma(j),$$

where $d(n) = 1$ if the network is closed.

This extension is particularly appealing in the context of communication networks since it may be used to model flow control mechanisms, namely, mechanisms that prevent congestion by denying access of packets to the network when it is too loaded.

This result may be further extended to the case when the external arrival rate of customers of class $r$ at node $i$ depends on the current total number of customers in the subchain that contains $(i, r)$.

So far, only networks composed of M/M/c queues have been considered. Such nodes have obvious limitations in terms of the systems they can model:

- the service times must be exponential;
- the service times must all have the same distribution function, namely, there are all exponential with the same parameter;
- the service discipline must be FIFO.

We now introduce three new types of nodes that will preserve the product-form result: Processor Sharing (PS), Last-In-First-Out (LIFO) and the Infinite Servers (IS) nodes.

### 8.2 The multiclass PS queue

There are $R$ classes of customers. Customers of class $r$ enter the network according to a Poisson process with rate $\lambda_r$ and require an exponential amount of service time with parameter $\mu_r$. The $K$ Poisson processes and service time processes are all assumed to be mutually independent processes.

\footnote{Actually, (123) and (124) hold for any mapping $\alpha_i : \mathbb{N} \to [0, \infty)$ such that $\alpha_i(0) = 0.$}
There is a single server equipped with an infinite buffer. If there are \( n_1 \) customers of class 1, \( \ldots, n_R \) customers of class \( R \) in the system, then any customer of class \( r \) is served at the rate \( \mu_r/|n| \). Therefore, the total service rate allocated to customers of class \( r \) is \( \mu_r n_r/|n| \).

Let \( X_r(t) \) be the number of customers of class \( r \) in the system at time \( t \geq 0 \). We define by \( \pi(n) = \lim_{t \to \infty} P(X_1(t) = n_1, \ldots, X_R(t) = n_R) \) the joint stationary probability distribution for the number customers of class \( 1, \ldots, R \) in the system.

Define \( \rho_r := \lambda_r/\mu_r \) for \( r = 1, 2, \ldots, R \).

We recall the following identity for the multinomial distribution, which we will use several times:

\[
\sum_{n_1 + \ldots + n_R = n} n! \prod_{r=1}^{R} \frac{x_r^{n_r}}{n_r!} = (x_1 + \ldots + x_R)^n.
\] (125)

We have the following result:

**Proposition 23** (Joint probability distribution for the queue-lengths in a multiclass PS queue).

If \( \sum_{r=1}^{R} \rho_r < 1 \) (stability condition) then

\[
\pi(n) = \left(1 - \sum_{r=1}^{R} \rho_r\right) |n|! \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!}
\] (126)

for all \( n = (n_1, \ldots, n_R) \in \mathbb{N}^R \).

**Proof.** Since the arrival processes are Poisson, the service times are exponentially distributed and that all of these rvs are mutually independent, it is easily seen by using the rule of construction for C.-M.C. that \(((X_1(t), \ldots, X_R(t)), t \geq 0)\) is an irreducible C.-M.C. with state-space \( \mathbb{N}^R \). Therefore, it suffices to check that (126) satisfies the balance equations and that it is a probability distribution.

The balance equations are:

\[
\pi(n) \sum_{r=1}^{R} \left( \mu_r \frac{n_r}{|n|} + \lambda_r \right) = \sum_{r=1}^{R} \pi(n - \varepsilon_r) \lambda_r \mathbf{1}(n_r > 0) + \sum_{r=1}^{R} \pi(n + \varepsilon_r) \mu_r \frac{n_r + 1}{|n + \varepsilon_r|}
\] (127)

for all \( n \in \mathbb{N}^R \).

Introducing (126) (after substituting \( n \) for \( n + \varepsilon_r \)) in the second term in the right-hand side of (127) yields

\[
\sum_{r=1}^{R} \pi(n + \varepsilon_r) \mu_r \frac{n_r + 1}{|n + \varepsilon_r|} = \frac{1}{G} \sum_{r=1}^{R} |n + \varepsilon_r|! \left( \prod_{s=1, s \neq r}^{R} \frac{\rho_s^{n_s}}{n_s!} \right) \frac{\rho_r^{n_r} + 1}{(n_r + 1)!} \mu_r \frac{n_r + 1}{|n + \varepsilon_r|}
\]

\[
= \pi(n) \sum_{r=1}^{R} \frac{|n + \varepsilon_r|!}{|n|!} \frac{\lambda_r}{n_r + 1} \frac{n_r + 1}{|n + \varepsilon_r|}
\]

\[
= \pi(n) \sum_{r=1}^{R} \lambda_r.
\] (128)

Similarly, we get that the first term in the right-hand side of (127) satisfies

\[
\sum_{r=1}^{R} \pi(n - \varepsilon_r) \lambda_r \mathbf{1}(n_r > 0) = \pi(n) \sum_{r=1}^{R} \mu_r \frac{n_r}{|n|!}.
\] (129)
Adding now the terms in the right-hand sides of (128) and (129) gives us the left-hand side of (127).

Let us now check that (126) is a probability distribution. We have

\[
\sum_{n \in \mathbb{N}} \left( 1 - \sum_{r=1}^{R} \rho_r \right) \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!} = \left( 1 - \sum_{r=1}^{R} \rho_r \right) \sum_{n \geq 0} \sum_{n_1 + \cdots + n_R = n} n! \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!} \\
= \left( 1 - \sum_{r=1}^{R} \rho_r \right) \left( \rho_1 + \cdots + \rho_R \right)^n \text{ from (125)} \\
= 1
\]

which concludes the proof.

The remarkable result about the PS queue is the following: the joint distribution function \( \pi(n) \) for the number of customers of class 1, 2, \ldots, R is given by (126) for any service time distribution of the customers of class 1, 2, \ldots, R. In other words, (126) is insensitive to the service time distributions. The proof is omitted. In particular, if the service time of customers of class \( r \) is constant and equals to \( S_r \) for \( r = 1, 2, \ldots, R \), then

\[
\pi(n) = \left| n \right|! \frac{\rho_1 S_1}{1 - \sum_{r=1}^{R} \rho_r} \\
\prod_{r=1}^{R} \frac{\left( \lambda_r S_r \right)^{n_r}}{n_r!}
\]

for all \( n \in \mathbb{N}^R \), provided that \( \lambda_r < 1/S_r \) for \( r = 1, 2, \ldots, R \).

**Corollary 1** (Probability distribution of the stationary number of customers of class \( i \)).

Assume that that queue is stable, i.e., \( \sum_{r=1}^{R} \rho_r < 1 \) and let \( X_i \) be the stationary number of customers of class-\( i \) in the system.

For \( i = 1, \ldots, R \), we have

\[
P(X_i = n) = \frac{1 - \sum_{r=1}^{R} \rho_r}{1 - \sum_{r=1}^{R} \rho_r} \left( \frac{\rho_i}{1 - \sum_{r=1}^{R} \rho_r} \right)^n
\]

for \( n \geq 0 \), which yields

\[
E[X_i] = \frac{\rho_i}{1 - \sum_{r=1}^{R} \rho_r}.
\]

The expected sojourn time and the expected waiting time of customers of class-\( i \) are given by

\[
T_i = \frac{1}{\mu_i \left( 1 - \sum_{r=1}^{R} \rho_r \right)} \text{ and } W_i = \frac{\sum_{r=1}^{R} \rho_r}{\mu_i \left( 1 - \sum_{r=1}^{R} \rho_r \right)}
\]

respectively, for \( i = 1, \ldots, R \).

**Proof.** An easy way to calculate \( P(X_i = n) \) for all \( n \geq 0 \) is to find its z-transform \( F_{X_i}(z) := \sum_{n \geq 0} P(X_i = n) z^n \) for \( 0 < z < 1 \), and then to invert it w.r.t. the variable \( z \).

We have

\[
F_{X_i}(z) = \sum_{n \in \mathbb{N}^R} \pi(n) z^n \\
= \left( 1 - \sum_{r=1}^{R} \rho_r \right) \sum_{n \in \mathbb{N}^R} \left| n \right|! \prod_{r=1}^{R} \frac{(\rho_r)^{n_r}}{n_r!} z^{n_i} \text{ by using (126)}
\]
\[
\frac{1}{(1 - \sum_{r=1}^{R} \rho_r)} \left( \rho_1 + \cdots + \rho_{i-1} + \rho_i z + \rho_{i+1} + \cdots + \rho_R \right)^n \quad \text{by using (125)}
\]

By writing \( F_{X_i}(z) \) in the form

\[
F_{X_i}(z) = \frac{b}{1 - az}
\]

with \( a := \rho_i / \left( 1 - \sum_{r \neq i} \rho_r \right) \) and \( b := \left( 1 - \sum_{r=1}^{R} \rho_r \right) / \left( 1 - \sum_{r \neq i} \rho_r \right) \), we immediately deduce that

\[
P(X_i = n) = ba^n
\]

for all \( n \geq 0 \), which is the result announced in (130).

From \( E[X_i] = (d/dz)F_{X_i}(z)|_{z=1} \) we get (131), and from the latter and Little’s result we find \( T_i \). Finally, \( W_i = T_i - 1/\mu_i \).

8.3 The multiclass LIFO queue

There are \( R \) classes of customers. Customers of class \( r \) enter the network according to a Poisson process with rate \( \lambda_r \) and require an exponential amount of service time with parameter \( \mu_r \). The \( K \) Poisson processes and service time processes are all assumed to be mutually independent processes.

There is a single server equipped with an infinite buffer. The service discipline is LIFO.

Define \( \pi(n) := \lim_{t \to \infty} P(X_1(t) = n_1, \ldots, X_R(t) = n_R) \) to be the joint distribution function for the number of customers of class 1, \ldots, \( R \) in the system in steady-state, where \( X_r(t) \) is the number of customers of class \( r \) in the system at time \( t \).

The following result holds:

**Proposition 24** (Joint probability distribution for the queue-lengths in a multiclass LIFO queue). If \( \sum_{r=1}^{R} \rho_r < 1 \) (stability condition) then

\[
\pi(n) = \left( 1 - \sum_{r=1}^{R} \rho_r \right) |n|! \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!} \quad (133)
\]

for all \( n = (n_1, \ldots, n_R) \in \mathbb{N}^R \).

Observe that the joint distribution function \( \pi(n) \) is the same as in the PS queue (see (126) in Proposition 23)! Therefore, all the results in Corollary 1 apply to the multiclass LIFO queue.

**Proof.** Here, \((X_1(t), \ldots, X_R(t)), t \geq 0\) is not a C.-M.C. (can you see why?). A C.-M.C. for this queue is given \((I_1(t), \ldots, I_{N(t)}), t \geq 0\) where \( I_j(t) \in \{1, 2, \ldots, R\} \) is the class of the customer in the \( j \)-th position in the waiting room at time \( t \) and \( N(t) \) is the total number of customers in the queue at time \( t \).
The balance equations for this C.-M.C. are:

$$\pi'(r_1, \ldots, r_{n-1}, r_n) \left( \mu_{r_n} + \sum_{r=1}^{R} \lambda_r \right) = \lambda_r \pi'(r_1, \ldots, r_{n-1}) + \sum_{r=1}^{R} \pi'(r_1, \ldots, r_n, r) \mu_r$$ (134)

for all $(r_1, \ldots, r_{n-1}, r_n) \in \{1, 2, \ldots, R\}^n$, $n = 1, 2, \ldots$, and

$$\pi'(0) = \frac{1}{G}$$ (135)

where $\pi'(0)$ is the probability that the system is empty.

It is straightforward to check that

$$\pi'(r_1, \ldots, r_n) = \frac{1}{G} \prod_{i=1}^{n} \rho_{r_i} \quad \text{for } n = 1, 2, \ldots$$ (136)

$$\pi'(0) = \frac{1}{G}$$

satisfies the balance equations, where $G$ is the normalizing constant. Let us find $G$.

We want

$$1 = \frac{1}{G} + \sum_{n \geq 1} \sum_{r_1=1}^{R} \cdots \sum_{r_R=1}^{R} \pi'(r_1, \ldots, r_n) = \frac{1}{G} \sum_{n \geq 0} \left( \sum_{r=1}^{R} \rho_r \right)^n = \frac{1}{G} \left( \frac{1}{1 - \sum_{r=1}^{R} \rho_r} \right)$$

which implies that $G = 1/ \left( 1 - \sum_{r=1}^{R} \rho_r \right)$.

Recall that $|\mathbf{n}| = \sum_{i=1}^{n} n_i$ for all $\mathbf{n} = (n_1, \ldots, n_R)$. For every fixed vector $\mathbf{n} = (n_1, \ldots, n_R) \in \mathbb{N}^R$, let $S(n_1, \ldots, n_R)$ denote the set of all vectors in $\{1, 2, \ldots, R\}^{|\mathbf{n}|}$ that have exactly $n_1$ components equal to 1, $n_2$ components equal to 2, $\ldots$, and $n_R$ components equal to $R$.

Clearly, we have

$$\pi(\mathbf{n}) = \sum_{(i_1, \ldots, i_{|\mathbf{n}|}) \in S(n_1, \ldots, n_R)} \pi'(i_1, \ldots, i_{|\mathbf{n}|})$$ (137)

for all $\mathbf{n} \in \mathbb{N}^R$, $\mathbf{n} \neq 0$, and $\pi(0, \ldots, 0) = \pi'(0) = 1/G$. From (136) and (137) we readily get (133).

Again, we have the remarkable result that the product-form result (133) for the LIFO queue is insensitive to the service time distribution. The proof is omitted.

8.4 The multiclass IS server queue

There is an infinite number of servers so that a new customer enters directly a server. This queue is used to model delay phenomena in communication networks, for instance.

We keep the notation introduced in the previous sections. In particular, $\pi(\mathbf{n})$ is the joint probability distribution for the number of customers of class $1, \ldots, R$ in steady-state (or, equivalently, the number of busy servers) is $n_1, \ldots, n_R$.

We have the following result:
Proposition 25 (Joint probability distribution for the queue-lengths in a multiclass IS queue). For all \( \underline{n} = (n_1, \ldots, n_R) \in \mathbb{N}^R \),

\[
\pi(\underline{n}) = e^{-\sum_{r=1}^{R} \rho_r n_r} \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!}.
\tag{138}
\]

The expected number of class-\( i \) \( (i = 1, \ldots, R) \) customers is

\[ E[X_i] = \rho_i. \]

Observe that \( \pi(\underline{n}) \) is not equal to the corresponding quantity for PS (resp. LIFO) queues.

Proof. The process \( ((X_1(t), \ldots, X_R(t)) | t \geq 0) \) is a C.-M.C. The balance equations are:

\[
\pi(\underline{n}) \sum_{r=1}^{R} (\lambda_r + \mu_r n_r) = \sum_{r=1}^{K} \pi(\underline{n} - e_r) \lambda_r + \sum_{r=1}^{K} \pi(\underline{n} + e_r) \mu_r (n_r + 1)
\tag{139}
\]

for all \( \underline{n} = (n_1, \ldots, n_R) \in \mathbb{N}^R \).

It is easily checked that (138) satisfies the balance equations (139) and that it is a proper probability distribution.

Let us now compute \( E[X_i] \). We have

\[
E[X_i] = e^{-\sum_{r=1}^{R} \rho_r n_r} \sum_{n_i \geq 1, n_r \geq 0, r \neq i} n_i^{n_i} \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!} = \rho_i e^{-\sum_{r=1}^{R} \rho_r^{n_r}} \sum_{n_r \geq 0}^{n_r \geq 0} \prod_{r=1}^{R} \frac{\rho_r^{n_r}}{n_r!} = \rho_i
\]

which completes the proof.

Note that (138) holds for any values of the parameters \( (\rho_r)_r \), or, equivalently, that an IS queue is always stable.

Again the product-form result (138) is insensitive to the service time distributions (proof omitted).

8.5 BCMP networks

We now come to one of the main results of queueing network theory. Because of this result modeling and performance evaluation became popular in the late seventies and many queueing softwares based on the BCMP theorem became available at this time \(^\text{10}\) (QNAP2 (INRIA-BULL) nowadays commercialized by SIMULOG as MODLINE, PAW (AT&T), PANACEA (IBM), etc.). Since then, queueing softwares have been continuously improved in the sense that they contain more and more analytical (e.g., larger class of product-form queueing models) and simulation tools (e.g., animated, monitored, and controlled simulations), and are more user-friendly (e.g., graphical interfaces). Most queueing softwares are essentially simulation oriented (PAW, for instance); a few are hybrid (QNAP, PANACEA), meaning that they can handle both analytical models — queueing networks for which the “solution” is known explicitly, like open product-form queueing networks, or can be obtained through numerical procedures, like closed product-form queueing networks and closed markovian queueing networks — and simulation models. Simulation is more time and memory consuming but this is sometimes the only way to get (reasonably) accurate results. However, for large (several hundreds of nodes or more) non-product queueing networks both analytical and simulation models are not feasible in general, and one must then resort to approximation techniques. Approximation and simulation techniques will be discussed later on in this course.

We now introduce the BCMP network. A BCMP network is similar to a multiclass open/mixed/closed Jackson network (see Section 8.1) except for the fact that a node may be of the following type:

\(^\text{10}\)This list is not at all exhaustive!
(1) a **FIFO** node with $c_i$ servers. All customers are served according to an exponential distribution with mean $1/\mu_i$;

(2) a **Processor Sharing (PS)** node. If there are $n_i$ customers in such a node, then any customer of class $r = 1, 2, \ldots, R$ is served at rate $\mu_{ir}/n_i$. Hence, the total service rate allocated to customers of class $r$ is $\mu_{ir}n_{ir}/n_i$, with $n_{ir}$ the number of customers of class $r$. Within a class customers have the same service time distribution (with mean $1/\mu_{ir}$ for class $r$) but it is arbitrary. Service time distributions of customers belonging to different classes may be different;

(3) an **Infinite Server (IS)** node. Once a customer enters an IS node he always finds an available server that serves him according to the distribution of its service time. Alike for PS nodes, within a class customers have the same service time distribution (with mean $1/\mu_{ir}$ for class $r$) but it is arbitrary. Service time distributions of customers belonging to different classes may be different;

(4) a **LIFO** node. Among all customers present the server always serves the one who last joined the node. Alike for PS and IS nodes, within a class customers have the same service time distribution (with mean $1/\mu_{ir}$ for class $r$) but it is arbitrary. Service time distributions of customers belonging to different classes may be different.

Alike in a multiclass open Jackson network exogeneous arrivals are Poisson and we denote by $\lambda_{ir}$ the external arrival rate of customers of class $r = 1, 2, \ldots, R$ at node $i = 1, 2, \ldots, K$.

Let $\lambda_{ir}$ be the solution to the traffic equations, that is

- If $(i, r)$ belongs to an open chain $E_k$ then $(\lambda_{ir})_{(i, r)\in E_k}$ is the unique strictly positive solution of the traffic equations

$$
\lambda_{ir} = \lambda_{ir}^0 + \sum_{(j, s)\in E_k} \lambda_{js} p_{j, s; i, r} \quad \forall (i, r) \in E_k.
$$

If $0$ is the only solution to these equations then the BCMP result below does not apply (underlying Markov process not irreducible).

- If $(i, r)$ belongs to a closed chain $E_k$ then $(\lambda_{ir})_{(i, r)\in E_k}$ is any non-zero solution of the equations

$$
\lambda_{ir} = \sum_{(j, s)\in E_k} \lambda_{js} p_{j, s; i, r} \quad \forall (i, r) \in E_k.
$$

We define $\rho_{ir} := \lambda_{ir}/\mu_{ir}$ when $i$ is a PS, IS or LIFO node, and $\rho_{ir} := \lambda_{ir}/\mu_i$ when $i$ is a FIFO node. Here is the BCMP\textsuperscript{11} result:

**Theorem 1** (BCMP). *For a BCMP network with $K$ nodes and $R$ classes of customers, which is open, closed or mixed in which each node is of type FIFO, PS, IS or LIFO, the equilibrium state probabilities are given by

$$
\pi(n) = \frac{1}{G} \prod_{i=1}^{K} f_i(n_i),
$$

with $G$ a normalizing constant given by $G = \sum_{n\in S} \prod_{i=1}^{K} f_i(n_i)$. Formula (140) holds*

(1) for any state $n = (n_1, \ldots, n_K)$ in the state-space $S$ (that depends on the network under consideration) with $n_i = (n_{i1}, \ldots, n_{iR})$, where $n_{ir}$ is the number of customers of class $r$ in node $i$;

(2) if $\sum_{(i, r)\in E_k \text{ belongs to an open subchain}} n_{ir} < 1$ for all $i = 1, 2, \ldots, K$ (stability condition).

Moreover (with $|n_i| = \sum_{r=1}^{R} n_{ir}$ for $i = 1, 2, \ldots, K$),

• if node $i$ is of type FIFO, then

$$f_i(n) = \frac{n!}{\prod_{j=1}^{n_i} \alpha_i(j)} \prod_{r=1}^{R} \rho_{ir}^{n_{ir}} n_{ir}!$$

with $\alpha_i(j) = \min(c_i, j)$;

• if node $i$ is of type PS or LIFO, then

$$f_i(n) = \frac{n! \prod_{r=1}^{R} \rho_{ir}^{n_{ir}}}{n_{ir}!};$$

• if node $i$ is of type IS, then

$$f_i(n) = \prod_{r=1}^{R} \rho_{ir}^{n_{ir}}.$$  

(141)

(142)

(143)

The proof of Proposition 140 consists in writing down the balance equations for an appropriate C.-M.C. (can you see which one?), guessing a product-form solution, checking that this solution satisfies the balance equations, and finally aggregating the states in this solution (has done in Section 8.3) to derive (140).

Remarks 1.

- Observe that (140) is a product-form result: node $i$ behaves as if it was an isolated FIFO (resp. PS, LIFO, IS) node with input rates $(\lambda_{ir})_{r=1}^{R}$ if $i \in \{FIFO\}$ (resp. $i \in \{PS\}, i \in \{LIFO\}, i \in \{IS\}$);

- The BCMP theorem still holds if exogeneous arrivals are state-dependent (up to some modifications in (140) that can be found in the BCMP paper). This is the case if (1) the external arrival rate at any node depends on the total number of customers in the system and (2) if the total external arrival rate in a subchain depends on the total number of customers in that subchain. This allows one, for instance, to reject new customers if the network (or a part of the network) is congested;

- It is worth observing that the convergence of the series $\sum_{n \in S} \pi(n)$ imposes conditions on the parameters of the model, referred to as the stability conditions.

Further generalizations of this result are possible: state-dependent routing probabilities, arrivals that depend on the number of customers in the subchain they belong to, more detailed state-space that gives the class of the customer in any position in any queue, etc. The last two extensions can be found in the BCMP paper whose reference has been given earlier.

Obtaining this result was not a trivial task; the consolation is that it is easy to use (which, overall, is still better than the contrary!). Indeed, the only thing that has to be done is to solve the traffic equations to determine the arrival rates $(\lambda_{ir})_{r=1}^{R}$ and to compute the normalizing constant $G$. For the latter computation, an extended convolution algorithm exists as well as many others. When you use a (hybrid) queueing software, these calculations are of course performed by the software. What you only need to do is to enter the topology of the network, that is, $K$ the number of nodes and their type, $R$ the number of classes, $[p_{i,r,j,s}]$ the matrix of routing probabilities, and to enter the values of the external arrival and service rates, namely, $(\lambda_{ir}^j, \gamma(j), j = 0, 1, 2 \ldots)_{r=1}^R$ the external arrival rates, $(\mu_{i,\alpha_i(j)}, j = 0, 1, 2 \ldots)_{j=0}^{\infty}$ the service rates for nodes of type FIFO, and $(\mu_{ir})_{i,r}$ the service rates of customers of class $r = 1, 2, \ldots, R$ visiting nodes of type $\{PS, LIFO, IS\}$.

A further simplification in the BCMP theorem is possible if the network is open and if the arrivals do not depend upon the state of the system.

Introduce $R_i = \{\text{class } r \text{ customers that may require service at node } i\}$. Define $\rho_i = \sum_{r \in R_i} \rho_{ir}$ ($\rho_{ir}$ has been defined earlier in this section). Let $\pi_i(n)$ be the probability that node $i$ contains $n$ customers.

We have the following result:
Proposition 26 (Isolated node). For all \( n \geq 0 \),

\[
\pi_i(n) = (1 - \rho_i) \rho_i^n
\]

if \( i \in \{ \text{FIFO, PS, LIFO} \} \), and

\[
\pi_i(n) = e^{-\rho_i} \frac{\rho_i^n}{n!}
\]

if \( i \in \{ \text{IS} \} \).

\( \square \)

8.6 Kelly Networks

We define \( M_{KR} \) as the set of all \( K \)-by-\( R \) matrices with entries in \( \{0, 1, 2, \ldots \} \). Any matrix \( M \) in \( M_{KR} \) with \((i, k)\)-entry \( m_{ik} \) will be denoted by the shorthand \( M = [m_{ik}] \).

In Jackson networks, and more generally in BCMP networks, customers follow random routes. There are however many systems in which paths or routes are deterministic. Typical examples are flexible manufacturing systems and connection-oriented communication networks (e.g. ATM networks). For such systems one can use Kelly networks. In a Kelly network customers belong to different classes where each class corresponds to a deterministic route. More specifically, to each customer of class \( k \) is associated a deterministic route \( r_k = (r^1_k, \ldots, r^n_k) \) where \( r^t_k \) is the identity of the \( t \)th node to be visited and \( n_k \) the total number of visits. A customer may visit several times the same node.

There are \( K \) FIFO nodes and \( R \) classes of customers. Each node is equipped with a single server and an infinite capacity waiting room. Customers of class \( k \) arrive to the system according to a Poisson process (rate \( \lambda_k \)) and require exponential service times with node dependent rates. Let \( \mu_i \) be the service rate of customers at node \( i \).

Once routes and external arrival rates on each route have been specified one can calculate the rate of customers of class \( k \) entering node \( i \), denoted by \( \hat{\lambda}_{ik} \). It is given by

\[
\hat{\lambda}_{ik} = \lambda_k \sum_{j=1}^{n_k} 1(r^t_j = i).
\]

In particular, if class \( k \) customers enter node \( i \) exactly once then \( \hat{\lambda}_{ik} = \lambda_k \). The total arrival rate in node \( i \) - denoted by \( \hat{\lambda}_i \) - is given by

\[
\hat{\lambda}_i = \sum_k \hat{\lambda}_{ik}.
\]

Let \( M_{ik}(t) \) be the number of customers of class \( k \) in node \( i \) at time \( t \). It is important to note that the stochastic process defined by \( \{ M(t) := [M_{ik}(t)] \in M_{KR}, t \geq 0 \} \) is not a continuous-time Markov chain for two reasons:

1. the available information does not allow us to calculate transition rates of the type \( M \rightarrow M - E(i, k) \) when \( m_{ik} > 0 \) where \( M = [m_{ik}] \in M_{KR} \) and \( E(i, k) \in M_{KR} \) has all entries equal to zero except entry \((i, k)\) that is one (this transition corresponds to a departure from node \( i \) of a customer of class \( k \)). This is so because we do not know the class of the customer in the server.

2. If a customer visits the same node more than once, then upon its departure we do not know where to route it.

The way to tackle this is to introduce a more detailed description of the state of the network – in the same spirit as done for the multiclass LIFO M/M/1 queue in Section 8.3. We will obtain a continuous-time Markov chain if the state of the network give the class of the customer and the position in its route at every position of every queue. By position in its route, we mean that if a customer visits the \( j \)th queue then its position in its route is \( j \) at node \( j \). Balance equations can be written for this Markov process and solved - see [4]. By aggregating the states, the following result is obtained:
Theorem 2 (Kelly network).

If \( \hat{\lambda}_i < \mu_i \) for all \( i = 1, \ldots, K \) the stationary probability \( \pi(M) \) that the network is in state \( M = [m_{ik}] \in \mathcal{M}_{KR} \) is

\[
\pi(M) = \prod_{i=1}^{K} \left( 1 - \frac{\hat{\lambda}_i}{\mu_i} \right) \left( \sum_{k=1}^{R} m_{ik} \right)! \prod_{k=1}^{R} \frac{1}{m_{ik}! \mu_i}.
\] (144)

This is a product-form result where the \( i \)th term of the product is the steady-state of a node \( i \) taken in isolation.

From this result one can calculate various metrics of interest such as the expected number of customers of class \( k \) in node \( i \)

\[
\overline{N}_{ik} = \frac{\hat{\lambda}_{ik}}{\mu_i - \hat{\lambda}_i},
\] (145)

the expected number of customers in node \( i \)

\[
\overline{N}_i = \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i},
\] (146)

the expected sojourn time of customers of class \( k \) in the network (Hint: use Little’s formula)

\[
\overline{T}_k = \frac{1}{\lambda_k} \sum_{j=1}^{n_k} \frac{\hat{\lambda}_{r_{j,k}}}{\mu_{r_{j,k}} - \hat{\lambda}_{r_{j,k}}},
\] (147)

and the expected sojourn time in the network of an arbitrary customer

\[
\overline{T} = \frac{1}{\sum_{k=1}^{R} \lambda_k} \sum_{i=1}^{K} \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i}.
\] (148)

This version of a Kelly network is a very basic version. Alike for Jackson networks vs. BCMP networks, many extensions are possible – see [4].

9 Queueing Theory (continued)

9.1 The FIFO GI/GI/1 queue

A GI/GI/1 queue is a single server queue with arbitrary, but independent, service times and interarrival times. More precisely, let \( s_n \) be the service time of the \( n \)-th customer, and let \( \tau_n \) be the time between arrivals of customers \( n \) and \( n + 1 \).

In a GI/GI/1 queue we assume that:

A1 \( (s_n)_n \) is a sequence of independent random variables with the common cumulative distribution function \( G(x) \), namely, \( P(s_n \leq x) = G(x) \) for all \( n \geq 1, x \geq 0 \);

A2 \( (s_n)_n \) is a sequence of independent random variables with the common c.distribution function \( F(x) \), namely, \( P(\tau_n \leq x) = F(x) \) for all \( n \geq 1, x \geq 0 \);

A3 \( s_n \) and \( \tau_m \) are independent rvs for all \( m \geq 1, n \geq 1 \).

We shall assume that \( F(x) \) and \( G(x) \) are both differentiable in \( [0, \infty) \) (i.e., for every \( n \geq 1, s_n \) and \( \tau_n \) each have a density function).

In the following, we shall assume that the service discipline is FIFO. Let \( \lambda = 1/E[\tau_n] \) and \( \mu = 1/E[s_n] \) be the arrival rate and the service rate, respectively.

Let \( W_n \) be the waiting time in queue of the \( n \)-th customer.
The following so-called Lindley’s equation holds:
\[ W_{n+1} = \max(0, W_n + s_n - \tau_n) \quad \forall n \geq 1. \tag{149} \]

From now on, we shall assume without loss of generality that \( W_1 = 0 \), namely, the first customer enters an empty system.

It can be shown that the system is stable, namely, \( \lim_{n \to \infty} P(W_n \leq x) = P(W \leq x) \) for all \( x \geq 0 \), where \( W \) is an almost surely finite rv, if
\[ \lambda < \mu \tag{150} \]

which should not be a surprising result. The proof of (150) is omitted (if \( \lambda > \mu \) then the system is always unstable, that is, the queue is unbounded with probability one).

Let \( W := \lim_{n \to \infty} W_n \) under the stability condition (150).

Define \( \phi(\theta) = E[\exp(\theta(s_n - \tau_n))] \) the Laplace transform of the rv \( s_n - \tau_n \). We shall assume that there exists \( c > 0 \) such that \( \phi(c) < \infty \). In practice, namely for all interesting c.distribution function’s for the interarrival times and service times, this assumption is always satisfied.

Therefore, since \( \phi(0) = 1 \) and since
\[ \phi'(0) = E[s_n - \tau_n] = (\lambda - \mu)/(\lambda \mu) < 0 \]

from (150), we know that there exists \( \theta > 0 \) such that \( \phi(\theta) < 1 \).

Our goal is to show the following result:

**Proposition 27** (Exponential bound for the GI/GI/1 queue). Assume that \( \lambda < \mu \). Let \( \theta > 0 \) be such that \( \phi(\theta) \leq 1 \).

Then,
\[ P(W_n \geq x) \leq e^{-\theta x} \quad \forall x > 0, n \geq 1 \tag{151} \]

and
\[ P(W \geq x) \leq e^{-\theta x} \quad \forall x > 0. \tag{152} \]


**Proof of Proposition 27:** Let \( f \) be a non-increasing function on \((0, \infty)\) with \( 0 \leq f(x) \leq 1 \), such that, for all \( x > 0 \),
\[ \int_{-\infty}^{x} f(x - y) dH(y) + 1 - H(x) \leq f(x) \tag{153} \]

where \( H(x) \) is the c.distribution function of the rv \( s_n - \tau_n \), namely, \( H(x) = P(s_n - \tau_n \leq x) \) for all \( x \in (-\infty, \infty) \), and where \( dH(x) := H'(x) \, dx \).

Let us show that
\[ P(W_n \geq x) \leq f(x) \quad \forall x > 0, n \geq 1. \tag{154} \]

We use an induction argument.

The result is true for \( n = 1 \) since \( W_1 = 0 \). Therefore, \( P(W_1 \geq x) = 0 \leq f(x) \) for all \( x > 0 \). Assume that \( P(W_m \geq x) \leq f(x) \) for all \( x > 0, m = 1, 2, \ldots, n \) and let us show that this is still true for \( m = n + 1 \).

We have, for all \( x > 0 \) (cf. (149)),
\[ P(W_{n+1} \geq x) = P(\max(0, W_n + s_n - \tau_n) \geq x) = P(W_n + s_n - \tau_n \geq x) \]
since for any rv \( X \), \( P(\max(0, X) \geq x) = 1 - P(\max(0, X) < x) = 1 - P(0 < x, X < x) = 1 - P(X < x) = P(X \geq x) \) for all \( x > 0 \).

Thus, for all \( x > 0 \),
\[ P(W_{n+1} \geq x) = P(W_n + s_n - \tau_n \geq x) \]

\[ P(W_{n+1} \geq x) \leq f(x) \quad \forall x > 0, n \geq 1. \]
\[ P(W_n \geq x - y | s_n - \tau_n = y) dH(y) \]
\[ = \int_{-\infty}^{\infty} P(W_n \geq x - y) dH(y) \]
\[ = \int_{-\infty}^{\infty} P(W_n \geq x - y) dH(y) \] since \( W_n \) is independent of \( s_n \) and \( \tau_n \)
\[ = \int_{-\infty}^{x} P(W_n \geq x - y) dH(y) + \int_{x}^{\infty} dH(y) \] since \( P(W_n \geq u) = 1 \) for \( u \leq 0 \)
\[ \leq \int_{-\infty}^{x} f(x - y) dH(y) + 1 - H(x) \] from the induction hypothesis
\[ \leq f(x) \]

from (153).
Letting now \( n \to \infty \) in (154) gives
\[ P(W \geq x) \leq f(x) \quad \forall x > 0. \] (155)

Let us now show that the function \( f(x) = \exp(-\theta x) \) satisfies (153) which will conclude the proof. We have
\[
\int_{-\infty}^{x} e^{\theta(x-y)} dH(y) + 1 - H(x)
\]
\[ = e^{-\theta x} \int_{-\infty}^{\infty} e^{\theta y} dH(y) - \int_{x}^{\infty} e^{\theta(y-x)} dH(y) + 1 - H(x)
\]
\[ = e^{-\theta x} \phi(\theta) - \int_{x}^{\infty} (e^{\theta(y-x)} - 1) dH(y)
\]
\[ \leq e^{-\theta x} \]

since \( \phi(\theta) \leq 1 \) by assumption and \( \exp(\theta(y-x)) - 1 \geq 0 \) for all \( y \geq x \).

**Remark 1** (Independence of \( W_n \) and \((s_n, \tau_n)\)). In the proof of Proposition 27, we have used the fact that \( W_n \) is independent of the rvs \( s_n \) and \( \tau_n \). This result comes from the fact that, for all \( n \geq 2 \), \( W_n \) is only a function of \( s_1, \ldots, s_{n-1}, \tau_1, \ldots, \tau_{n-1} \). More precisely, it is straightforward to see from (149) and the initial condition \( W_1 = 0 \) that

\[ W_n = \max \left( 0, \max_{i=1,2,\ldots,n-1} \sum_{j=i}^{n-2} (s_j - \tau_i) \right) \quad \forall n \geq 2. \] (156)

Therefore, \( W_n \) is independent of \( s_n \) and \( \tau_n \) since \( s_n \) and \( \tau_n \) are independent of \( (s_i, \tau_i)_{i=1}^{n-1} \) from assumptions A1, A2 and A3.

**Remark 2** ((\(W_n)\) is a Markov chain). Let us show that \((W_n)\) is a discrete-time continuous-time Markov chain. For this, we must check that \( P(W_{n+1} \leq y | W_1 = x_1, \ldots, W_{n-1} = x_{n-1}, W_n = x) \) is only a function of \( x \) and \( y \), for all \( n \geq 1, x_1, \ldots, x_{n-1}, x, y \) in \([0, \infty)\).

We have from (149)
\[ P(W_{n+1} \leq y | W_1 = x_1, \ldots, W_{n-1} = x_{n-1}, W_n = x)
\[ = P(\max(0, x + s_n - \tau_n) \leq y | W_1 = x_1, \ldots, W_{n-1} = x_{n-1}, W_n = x)
\]
\[ = P(\max(0, x + s_n - \tau_n) \leq y)
\]
since we have shown in Remark 1 that \( s_n \) and \( \tau_n \) are independent of \( W_j \) for \( j = 1, 2, \ldots, n \).

The best exponential decay in (151) and (152) is the largest \( \theta > 0 \), denoted as \( \theta^* \), such that \( \phi(\theta) \leq 1 \). One can show (difficult) that \( \theta^* \) is actually the largest exponential decay, which means there does not exist \( \theta \) such that \( \phi(\theta) > 1 \) and \( \theta > \theta^* \).

From Proposition 27 we can easily derive an upper bound for \( E[W_n] \) and \( E[W] \).
Proposition 28: (Upper bound for the transient and stationary mean waiting time). Assume that \( \lambda < \mu \). Then,
\[
E[W_n] \leq \frac{1}{\theta^*} \quad \forall n \geq 2, \tag{157}
\]
and
\[
E[W] \leq \frac{1}{\theta^*}. \tag{158}
\]

Proof. It is known that \( E[X] = \int_0^{\infty} P(X > x) \, dx \) for any nonnegative rv \( X \). From Proposition 27 and the above identity we readily get (157) and (158).

The bound on the stationary mean waiting time obtained in (158) can be improved using more sophisticated techniques.

9.2 Application: effective bandwidth in multimedia networks

In high-speed multimedia networks, admission control plays a major role. Because of the extreme burstiness of some real-time traffic (e.g., video), accepting a new session in a network close to congestion may be dramatic. On the other hand, rejecting to many users maybe very costly. Also, because of the high speeds involved, admission control mechanisms must be very fast in making the decision to accept/reject a new session.

On the other hand, an interesting feature about real-time traffic applications is that they are able to tolerate a small fraction of packets missing their deadline (e.g., approx. 1% for voice). Therefore, bounds on the tail distribution of quantities such as buffer occupancy and response times can be used by designers to size the network as well as to develop efficient admission control mechanisms.

Assume that the system may support \( K \) different types of sessions (e.g., voice, data, images). Assume that there are \( n_1 \) active sessions of type 1, \( \ldots \), \( n_K \) active sessions of class \( K \), upon arrival of a new session of type \( i \). We would like to answer the following questions: should we accept or reject this new session so that

Problem 1:
\[
P(W \geq b) \leq q \tag{159}
\]

Problem 2:
\[
E[W] < \alpha \tag{160}
\]

where \( \alpha, b > 0 \) and \( q \in (0, 1) \) are prespecified numbers. Here, \( W \) is the stationary delay (i.e., waiting time in queue).

For each problem, the decision criterion has to be simple enough so that decisions to admit/reject new sessions can be made very rapidly, and easily implementable.

We will first consider the case when the inputs are independent Poisson processes.

9.2.1 Effective bandwidth for Poisson inputs

Consider a M/G/1 queue with \( N \) (non-necessarily distinct) classes of customers. Customers of class \( k \) are generated according to a Poisson process with rate \( \lambda_k \); let \( G_k(x) \) be the c.distribution function of their service time and let \( 1/\mu_k \) be their mean service time. We assume that the arrival time and service time processes are all mutually independent processes.

We first solve the optimization problem (159).

The form of the criterion (159) strongly suggests the use of Proposition 27. For this, we first need to place this multiclass M/G/1 queue in the setting of Section 1, and then to compute \( \phi(\theta) \).

First, let us determine the c.distribution function of the time between two consecutive arrivals. Because the superposition of \( N \) independent Poisson processes with rates \( \lambda_1, \ldots, \lambda_N \) is a Poisson process with rate
\[
\lambda := \sum_{i=1}^{N} \lambda_i \tag{161}
\]
we have that \( P(\tau_n \leq x) = 1 - \exp(-\lambda x) \) for all \( x \geq 0 \).
Let us now focus on the service times. With the probability $\frac{\lambda_k}{\lambda}$ the $n$-th customer will be a customer of class $k$. Let us prove this statement. Let $X_k$ be the time that elapses between the arrival of the $n$-th customer and the first arrival of a customer of class $k$. Since the arrival process of each class is Poisson, and therefore memoryless, we know that $X_k$ is distributed according to an exponential rv with rate $\lambda_k$, and that $X_1, \ldots, X_N$ are independent rvs.

Therefore,

$$P((n+1)\text{-st customer is of class } k) = P(X_k < \min_{i \neq k} X_i)$$

$$= \int_0^\infty P(x < \min_{i \neq k} X_i) \lambda_k e^{-\lambda_k x} \, dx$$

$$= \int_0^\infty \prod_{i \neq k} P(x < X_i) \lambda_k e^{-\lambda_k x} \, dx$$

$$= \lambda_k \int_0^\infty \prod_{i=1}^N e^{-\lambda_i x} \, dx$$

$$= \lambda_k \int_0^\infty e^{-\lambda x} \, dx$$

$$= \frac{\lambda_k}{\lambda}.$$

Let us now determine the c.distribution function $G(x)$ of the service time of an arbitrary customer. We have

$$G(x) = P(\text{service time new customer } \leq x)$$

$$= \sum_{k=1}^N P(\text{service time new customer } \leq x \text{ and new customer of type } k)$$

$$= \sum_{k=1}^N P(\text{service time new customer } \leq x \mid \text{new customer of type } k) \frac{\lambda_k}{\lambda}$$

(164)

Here, (162) and (163) come from the law of total probability and from Bayes’ formula, respectively.

Thus,

$$G(x) = \sum_{k=1}^N \frac{\lambda_k}{\lambda} G_k(x).$$

(164)

In particular, the mean service time $\frac{1}{\mu}$ of this multiclass M/G/1 queue is given by

$$\frac{1}{\mu} = \int_0^\infty x \, dG(x)$$

$$= \sum_{k=1}^N \frac{\lambda_k}{\lambda} \int_0^\infty x \, dG_k(x)$$

$$= \sum_{k=1}^N \rho_k$$

(165)

with $\rho_k := \frac{\lambda_k}{\mu_k}$.

In other words, we have “reduced” this multiclass M/G/1 queue to a G/G/1 queue where $F(x) = 1 - \exp(-\lambda x)$ and $G(x)$ is given by (164).

The stability condition is $\lambda < \mu$, that is from (161) and (165),

$$\sum_{k=1}^N \rho_k < 1.$$  

(166)

From now on we shall assume that (166) holds.
We are now in position to determine $\phi(\theta)$. We have

\[
\phi(\theta) = E[e^{\theta(s_n - \tau_n)}] = E[e^{\theta s_n}] E[e^{-\theta \tau_n}] \quad \text{since } s_n \text{ and } \tau_n \text{ are independent rvs}
\]

\[
= \left( \frac{\lambda}{\lambda + \theta} \right) E[e^{\theta s_n}] \quad \text{since } P(\tau_n \leq x) = 1 - e^{-\lambda x}.
\]

It remains to evaluate $E[e^{\theta s_n}]$. We have from (164)

\[
E[e^{\theta s_n}] = \int_0^\infty e^{\theta y} dG(y) = \sum_{k=1}^N \frac{\lambda_k}{\lambda} \int_0^\infty e^{\theta y} dG_k(y).
\]

Therefore,

\[
\phi(\theta) = \sum_{k=1}^N \frac{\lambda_k}{\lambda + \theta} \int_0^\infty e^{\theta y} dG_k(y).
\] (167)

From (152) we deduce that $P(W \geq b) \leq q$ if $\phi(-(\log q)/b) \leq 1$, that is from (167) if

\[
\sum_{k=1}^N \frac{\lambda_k}{\lambda - (\log q)/b} \int_0^\infty e^{-(\log q)/b} dG_k(y) \leq 1.
\] (168)

Let us now get back to the original Problem 1. Let $n_i, i = 0, 1, \ldots, K$, be integer numbers such that $\sum_{i=1}^K n_i = N$ and $n_0 = 0$, and assume that customers with class in $\{1 + \sum_{j=1}^i n_j, \ldots, \sum_{j=1}^{i+1} n_j\}$ are all identical ($i = 0, 1, \ldots, K - 1$).

Using (161) it is easily seen that (168) can be rewritten as

\[
\sum_{i=1}^K n_i \alpha_i \leq 1
\] (169)

with

\[
\alpha_i := \frac{\lambda_i b (1 - \phi_i(-(\log q)/b))}{\log q}
\] (170)

where $\phi_i(\theta) := \int_0^\infty \exp(\theta y) dG_i(y)$.

Thus, a new session, that is a new class of customers, say class $i$, can be admitted in the system when there are already $n_1$ active sessions of class 1, ..., $n_K$ active sessions of class $K$, if

\[
\sum_{i=1}^K n_i \alpha_i + \alpha_i \leq 1.
\] (171)

This result is called an effective bandwidth-type result since $\alpha_i$ may be interpreted as the effective bandwidth required by a session of type $i$. So, the decision criterion for admitting/rejecting a new session in Problem 1 consists in adding the effective bandwidth requirements of all the active sessions in the system to the effective bandwidth of the new session and to accept this new session if and only if the sum does not exceed 1.

Let us now focus on Problem 2. We could follow the same approach and use the bounds in Proposition 28. We shall instead use the exact formula for $E[W]$ (see the Polloczec-Khinchin formula (86)).

We have

\[
E[W] = \frac{\lambda \sigma^2}{(1 - \rho)}
\]

where $\rho := \sum_{k=1}^N \rho_k = \sum_{i=1}^K n_i \rho_i$ and where $\sigma^2$ is the second-order moment of the service time of an arbitrary customer.
Let $\sigma_i^2$ be the second-order moment of customers of class $i$, for $i = 1, 2, \ldots, K$. Using (164) we see that

$$\sigma^2 = \sum_{i=1}^{K} n_i \frac{\lambda_i}{\lambda} \sigma_i^2.$$ 

Hence,

$$E[W] = \frac{\sum_{i=1}^{K} n_i \lambda_i \sigma_i^2}{2 \left(1 - \sum_{i=1}^{K} n_i \rho_i\right)}.$$ 

Thus, the condition $E[W] < \alpha$ in Problem 2 becomes

$$\sum_{i=1}^{K} n_i \lambda_i \sigma_i^2 < 2\alpha \left(1 - \sum_{i=1}^{K} n_i \rho_i\right).$$

Rearranging terms, this is equivalent to

$$\sum_{i=1}^{K} n_i \alpha_i < 1 \quad (172)$$

where

$$\alpha_i = \rho_i + \lambda_i \frac{\sigma_i^2}{2\alpha}$$

for $i = 1, 2, \ldots, K$.

The analytical expression for the effective bandwidth $\alpha_i$ is illuminating. Observe that $\rho_i$ is the mean workload brought to the system per unit of time by a customer of type $i$. Therefore, the effective bandwidth requirement for session $i$ is seen to be larger than $\rho_i$. Let $\alpha \to \infty$. Then, the constraint $E[W] < \alpha$ becomes $\sum_{i=1}^{K} n_i \rho_i < 1$, which is nothing but that the stability condition.

Effective bandwidth-type results have lately been obtained for much more general input processes than Poisson processes (that do not match well bursty traffic).

### Appendices

#### A Probability Refresher

**A.1 Sample-space, events and probability measure**

A *Probability space* is a triplet $(\Omega, \mathcal{F}, P)$ where

- $\Omega$ is the set of all *outcomes* associated with an experiment. $\Omega$ will be called the sample-space

- $\mathcal{F}$ is a set of subsets of $\Omega$, called *events*, such that

  (i) $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$

  (ii) if $A \in \mathcal{F}$ then the complementary set $A^c$ is in $\mathcal{F}$

  (iii) if $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

$\mathcal{F}$ is called a $\sigma$-algebra.

- $P$ is a *probability measure* on $(\Omega, \mathcal{F})$, that is, $P$ is a mapping from $\mathcal{F}$ into $[0, 1]$ such that

  (a) $P(\emptyset) = 0$ and $P(\Omega) = 1$

  (b) $P(\bigcup_{n \in I} A_n) = \sum_{n \in I} P(A_n)$ for any countable (finite or infinite) family $\{A_n, n \in I\}$ of mutually exclusive events (i.e., $A_i \cap A_j = \emptyset$ for $i \in I, j \in I$ such that $i \neq j$).
Example 9. The experiment consists in rolling a die. Then

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$  

$A = \{1, 3, 5\}$ is the event of rolling an odd number. Instances of $\sigma$-algebras on $\Omega$ are $\mathcal{F}_1 = \{\emptyset, \Omega\}$, $\mathcal{F}_2 = \{\emptyset, \Omega, A, A^c\}$, $\mathcal{F}_3 = \{\emptyset, \Omega, \{1, 2, 3, 5\}, \{4, 6\}\}$, $\mathcal{F}_4 = \mathcal{P}(\Omega)$ (the set of all subsets of $\Omega$). Is $\{\emptyset, \Omega, \{1, 2, 3\}, \{3, 4, 6\}, \{5, 6\}\}$ a $\sigma$-algebra? $\mathcal{F}_1$ and $\mathcal{F}_4$ are the smallest and the largest $\sigma$-algebras on $\Omega$, respectively. If the die is not biased, the probability measure on, say, $(\Omega, \mathcal{F}_3)$, is defined by $P(\emptyset) = 0$, $P(\Omega) = 1$, $P(\{1, 2, 3, 5\}) = 4/6$ and $P(\{4, 6\}) = 2/6$.


\[
\begin{align*}
\text{Example 10.} & \quad \text{The experiment consists in rolling two dice. Then} \\
\Omega & = \{(1,1), (1,2), \ldots, (1,6), (2,1), (2,2), \ldots, (6,6)\}, \\
A & = \{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)\} \text{ is the event of rolling a seven.}
\end{align*}
\]


\[
\begin{align*}
\text{Example 11.} & \quad \text{The experiment consists in tossing a fair coin until head appears. Then,} \\
\Omega & = \{\text{H, TH, TTH, TTTH, \ldots}\}, \\
A & = \{\text{TTH, TTTH}\} \text{ is the event that 3 or 4 tosses are required.}
\end{align*}
\]


\[
\begin{align*}
\text{Example 12.} & \quad \text{The experiment consists in measuring the time that elapses from the instant the last character of a request is typed on an inter-active terminal until the last character of the response from the computer has been received and displayed (referred to as response time). We assume that the response time is at least of 1 second. Then,} \\
\Omega & = \{\text{real } t : t \geq 1\}, \\
A & = \{10 \leq t \leq 20\} \text{ is the event that the response time is between 10 and 20 seconds.}
\end{align*}
\]

A.2 Combinatorial analysis

A permutation of order $k$ of $n$ elements is an ordered selection of $k$ elements taken from the $n$ elements. A combination of order $k$ of $n$ elements is an unordered selection of $k$ elements taken from the $n$ elements. Recall that $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ for any nonnegative integer $n$ with $0! = 1$ by convention.

Proposition 29. The number of permutations of order $k$ of $n$ elements is

$$A(n, k) = \frac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-k+1).$$

Proposition 30. The number of combination of order $k$ of $n$ elements is

$$C(n, k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$
Example 13. Suppose that 5 terminals are connected to an on-line computer system via a single communication channel, so that only one terminal at a time may use the channel to send a message to the computer. At every instant, there may be 0, 1, 2, 3, 4, or 5 terminals ready for transmission. One possible sample-space is

\[ \Omega = \{ (x_1, x_2, x_3, x_4, x_5) : \text{each } x_i \text{ is either 0 or 1} \}. \]

\( x_i = 1 \) means that terminal \( i \) is ready to transmit a message, \( x_i = 0 \) that it is not ready. The number of points in the sample-space is 2^5 since each \( x_i \) of \( (x_1, x_2, x_3, x_4, x_5) \) can be selected in two ways.

Assume that there are always 3 terminals in the ready state. Then, \( \Omega = \{ (x_1, x_2, x_3, x_4, x_5) : \text{exactly 3 of the } x_i \text{'s are 1 and 2 are 0} \}. \)

In that case, the number \( n \) of points in the sample-space is the number of ways that 3 terminals that are ready can be chosen from the 5 terminals, that is from Proposition 30,

\[ \binom{5}{3} = \frac{5!}{3!(5-3)!} = 10. \]

Assume that each terminal is equally likely to be in the ready condition.

If the terminals are polled sequentially (i.e., terminal 1 is polled first, then terminal 2 is polled, etc.) until a ready terminal is found, the number of polls required can be 1, 2 or 3. Let \( A_1, A_2, \text{ and } A_3 \) be the events that the required number of polls is 1, 2, 3, respectively.

\( A_1 \) can only occur if \( x_1 = 1 \), and the other two 1’s occur in the remaining four positions. The number \( n_1 \) of points favorable to \( A_1 \) is calculated as \( n_1 = \binom{4}{2} = 6 \) and therefore \( P(A_1) = n_1/n = 6/10 \).

\( A_2 \) can only occur if \( x_1 = 0, x_2 = 0 \), and the remaining two 1’s occur in the remaining three positions. The number \( n_2 \) of points favorable to \( A_1 \) is calculated as \( n_2 = \binom{3}{2} = 3 \) and therefore \( P(A_1) = 3/10 \).

Similarly, \( P(A_3) = 1/10 \).

\[ \heartsuit \]

A.3 Conditional probability

The probability that the event \( A \) occurs given the event \( B \) has occured is denoted by \( P(A|B) \).

Proposition 31 (Bayes’ formula).

\[ P(A|B) = \frac{P(A \cap B)}{P(B)}. \]

The conditional probability is not defined if \( P(B) = 0 \). It is easily checked that \( P(\cdot|B) \) is a probability measure.

Interchanging the role of \( A \) and \( B \) in the above formula yields

\[ P(B|A) = \frac{P(A \cap B)}{P(A)}. \]

provided that \( P(A) > 0 \).

Let \( A_i, i = 1, 2, \ldots, n \) be \( n \) events. Assume that the events \( A_1, \ldots, A_{n-1} \) are such that \( P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0 \). Then,

Proposition 32 (Generalized Bayes’ formula).

\[ P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) P(A_2|A_1) \cdots \times P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \]
The proof if by induction on \( n \). The result is true for \( n = 2 \). Assume that it is true for \( n = 2, 3, \ldots, k \), and let us show that it is still true for \( n = k + 1 \).

Define \( A = A_1 \cap A_2 \cap \cdots \cap A_k \). We have

\[
P(A_1 \cap A_2 \cap \cdots \cap A_{k+1}) = P(A \cap A_{k+1})
\]

\[
= P(A) P(A_{k+1} | A)
\]

\[
= P(A_1) P(A_2 | A_1) \cdots P(A_k | A_1 \cap A_2 \cap \cdots \cap A_{k-1})
\]

\[
\times P(A_{k+1} | A)
\]

from the induction assumption, which completes the proof.

Example 14. A survey of 100 computer installations in a city shows that 75 of them have at least one brand \( X \) computer. If 3 of these installations are chosen at random, what is the probability that each of them has at least one brand \( X \) machine?

**Answer:** let \( A_1, A_2, A_3 \) be the event that the first, second and third selection, respectively, has a brand \( X \) computer.

The required probability is

\[
P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2 | A_1) P(A_3 | A_1, A_2)
\]

\[
= \frac{75}{100} \times \frac{74}{99} \times \frac{73}{98}
\]

\[
= 0.418.
\]

The following result will be extensively used throughout the course.

**Proposition 33** (Law of total probability). Let \( A_1, A_2, \ldots, A_n \) be events such that

(a) \( A_i \cap A_j = \emptyset \) if \( i \neq j \) (mutually exclusive events)

(b) \( P(A_i > 0) \) for \( i = 1, 2, \ldots, n \)

(c) \( A_1 \cup A_2 \cup \cdots \cup A_n = \Omega \).

Then, for any event \( A \),

\[
P(A) = \sum_{i=1}^{n} P(A | A_i) P(A_i).
\]

To prove this result, let \( B_i = A \cap A_i \) for \( i = 1, 2, \ldots, n \). Then, \( B_i \cap B_j = \emptyset \) for \( i \neq j \) (since \( A_i \cap A_j = \emptyset \) for \( i \neq j \)) and \( A = B_1 \cup B_2 \cup \cdots \cup B_n \). Hence,

\[
P(A) = P(B_1) + P(B_2) + \cdots + P(B_n)
\]

from axiom (b) of a probability measure. But \( P(B_i) = P(A \cap A_i) = P(A | A_i) P(A_i) \) for \( i = 1, 2, \ldots, n \) from Bayes’ formula, and therefore \( P(A) = \sum_{i=1}^{n} P(A | A_i) P(A_i) \), which concludes the proof.

Example 15. Requests to an on-line computer system arrive on 5 communication channels. The percentage of messages received from lines 1, 2, 3, 4, 5, are 20, 30, 10, 15, and 25, respectively. The corresponding probabilities that the length of a request will exceed 100 bits are 0.4, 0.6, 0.2, 0.8, and 0.9. What is the probability that a randomly selected request will be longer than 100 bits?

**Answer:** let \( A \) be the event that the selected message has more than 100 bits, and let \( A_i \) be the event that it was received on line \( i \), \( i = 1, 2, 3, 4, 5 \). Then, by the law of total probability,

\[
P(A) = \sum_{i=1}^{5} P(A | A_i) P(A_i)
\]
Two events $A$ and $B$ are said to be *independent* if

$$ P(A \cap B) = P(A) P(B). $$

This implies the usual meaning of independence; namely, that neither influences the occurrence of the other. Indeed, if $A$ and $B$ are independent, then

$$ P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) P(B)}{P(B)} = P(A) $$

and

$$ P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) P(B)}{P(A)} = P(B). $$

The concept of independent should not be confused with the concept of their being mutually exclusive (i.e., $A \cap B = \emptyset$). In fact, if $A$ and $B$ are mutually exclusive then

$$ 0 = P(\emptyset) = P(A \cap B) $$

and thus $P(A \cap B)$ cannot be equal to $P(A) P(B)$ only if at least one event has the probability 0. Hence, mutually exclusive events are not independent except in the trivial case when at least one of them has zero probability.

### A.4 Random variables

In many random experiments we are interested in some number associated with the experiment rather than the actual outcome (i.e., $\omega \in \Omega$). For instance, in Example 10 one may be interested in the sum of the numbers shown on the dice. We are thus interested in a function that associates a number with an experiment. Such function is called a random variable (rv).

More precisely, a *real-valued* rv $X$ is a mapping from $\Omega$ into $\mathbb{R}$ such that

$$ \{ \omega \in \Omega : X(\omega) \leq x \} \in \mathcal{F} $$

for all $x \in \mathbb{R}$.

As usual, we shall denote $X = x$ for the event $\{ \omega \in \Omega : X(\omega) = x \}$, $X \leq x$ for the event $\{ \omega \in \Omega : X(\omega) \leq x \}$, and $y \leq X \leq x$ for the event $\{ \omega \in \Omega : y \leq X(\omega) \leq x \}$.

The requirement that $X \leq x$ must be an event for $X$ to be a rv is necessary so that probability calculations can be made.

For each rv $X$ we define its cumulative distribution function (c.distribution function) $F$ (also called the probability distribution of $X$ or the law of $X$) as

$$ F(x) = P(X \leq x) $$

for each $x \in \mathbb{R}$.

$F$ satisfies the following properties: $\lim_{x \to +\infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$, and $F(x) \leq F(y)$ if $x \leq y$ (i.e., $F$ is nondecreasing).

A rv is *discrete* if it takes only discrete values. The distribution function $F$ of a discrete rv $X$ with values in a countable (finite or infinite) set $I$ (e.g. $I = \mathbb{N}$) is simply given by

$$ F(x) = P(X = x) $$

for each $x \in I$. We have $\sum_{x \in I} F(x) = 1$. 

\[
= 0.2 \times 0.4 + 0.3 \times 0.6 + 0.1 \times 0.2 + 0.15 \times 0.8 + 0.25 \times 0.9 \\
= 0.625.
\]
Example 16 (The Bernoulli distribution). Let \( p \in (0, 1) \). A rv variable \( X \) taking values in the set \( I = \{0, 1\} \) is said to be a Bernoulli rv with parameter \( p \), or to have a Bernoulli distribution with parameter \( p \) if \( \Pr(X = 1) = p \) and \( \Pr(X = 0) = 1 - p \).

A rv is continuous if \( \Pr(X = x) = 0 \) for all \( x \). The density function of a continuous rv is a function \( f \) such that

(a) \( f(x) \geq 0 \) for all real \( x \)

(b) \( f \) is integrable and \( \Pr(a \leq X \leq b) = \int_a^b f(x) \, dx \) if \( a < b \)

(c) \( \int_{-\infty}^{+\infty} f(x) \, dx = 1 \)

(d) \( F(x) = \int_{-\infty}^{x} f(t) \, dt \) for all real \( x \).

The formula \( F(x) = \frac{\partial f(x)}{\partial x} \) that holds at each point \( x \) where \( f \) is continuous, provides a mean of computing the density function from the distribution function, and conversely.

Example 17 (The exponential distribution). Let \( \alpha > 0 \). A rv \( X \) is said to be an exponential rv with parameter \( \alpha \) or to have an exponential distribution with parameter \( \alpha \) if

\[
F(x) = \begin{cases} 
1 - \exp(-\alpha x) & \text{if } x > 0 \\
0 & \text{if } x \leq 0.
\end{cases}
\]

The density function \( f \) is given by

\[
f(x) = \begin{cases} 
\alpha \exp(-\alpha x) & \text{if } x > 0 \\
0 & \text{if } x \leq 0.
\end{cases}
\]

Suppose that \( \alpha = 2 \) and we wish to calculate the probability that \( X \) lies in the interval \((1, 2] \). We have

\[
P(1 < X \leq 2) = P(X \leq 2) - P(X \leq 1) = F(2) - F(1) = (1 - \exp(-4)) - (1 - \exp(-2)) = 0.117019644.
\]

Example 18 (The exponential distribution is memoryless). Let us now derive a key feature of the exponential distribution: the fact that it is memoryless. Let \( X \) be an exponential rv with parameter \( \alpha \). We have

\[
P(X > x + y | X > x) = \frac{P(X > x + y, X > x)}{P(X > x)} \quad \text{from Bayes' formula}
\]

\[
= \frac{P(X > x + y)}{P(X > x)}
\]

\[
= e^{-\alpha y}
\]

\[
= P(X > y)
\]

which does not depend on \( x \)!
A.5 Parameters of a random variable

Let $X$ be a discrete rv taking values in the set $I$. The mean or the expectation of $X$, denoted as $E[X]$, is the number

$$E[X] = \sum_{x \in I} x P(X = x)$$

provided that $\sum_{x \in I} |x| P(X = x) < \infty$.

Example 19 (Expectation of a Bernoulli rv). Let $X$ be a Bernoulli rv with parameter $p$. Then,

$$E[X] = 0 \times P(X = 0) + 1 \times P(X = 1) = p.$$  

If $X$ is a continuous rv with density function function $f$, we define the expectation or the mean of $X$ as the number

$$E[X] = \int_{-\infty}^{+\infty} x f(x) \, dx$$

provided that $\int_{-\infty}^{+\infty} |x| f(x) \, dx.$

Example 20 (Expectation of an exponential rv). Let $X$ be an exponential rv with parameter $\alpha > 0$. Then,

$$E[X] = \int_{0}^{+\infty} x \alpha \exp(-\alpha x) \, dx = \frac{1}{\alpha}.$$  

by using an integration by parts (use the formula $\int u \, dv = uv - \int v \, du$ with $u = x$ and $dv = \alpha \exp(-\alpha x) \, dx$, together with the formula $\lim_{x \to +\infty} x \exp(-\alpha x) = 0$).

Let us give some properties of the expectation operator $E[\cdot]$.

Proposition 34. Suppose that $X$ and $Y$ are rv, such that $E[X]$ and $E[Y]$ exist, and let $c$ a real number. Then, $E[c] = c$, $E[X + Y] = E[X] + E[Y]$, and $E[cX] = c E[X]$.

The $k$-th moment or the moment of order $k$ ($k \geq 1$) of a discrete rv $X$ taking values in the set $I$ is given by

$$E[X^k] = \sum_{x \in I} x^k P(X = x)$$

provided that $\sum_{x \in I} |x^k| P(X = x) < \infty$.

The $k$-th moment or the moment of order $k$ ($k \geq 1$) of a continuous rv $X$ is given by

$$E[X^k] = \int_{-\infty}^{+\infty} x^k f(x) \, dx$$

provided that $\int_{-\infty}^{+\infty} |x^k| f(x) \, dx < \infty$.

The variance of a discrete or continuous rv $X$ is defined to be

$$\text{var} (X) = E (X - E[X])^2 = E[X^2] - (E[X])^2.$$
Example 21 (Variance of the exponential distribution). Let $X$ be an exponential rv with parameter $\alpha > 0$. Then,

$$
\text{var}(X) = \int_0^{+\infty} x^2 \alpha \exp(-\alpha x) \, dx - \frac{1}{\alpha^2}
$$

$$
= \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}.
$$

Hence, the variance of an exponential rv is the square of its mean.

Example 22. Consider the situation described in Example 13 when the terminals are polled sequentially until one terminal is found ready to transmit. We assume that each terminal is ready to transmit with the probability $p$, $0 < p \leq 1$, when it is polled.

Let $X$ be the number of polls required before finding a terminal ready to transmit. Since $P(X = 1) = p$, $P(X = 2) = (1 - p)p$, and more generally since $P(X = n) = (1 - p)^{n-1} p$ for each $n$, we have

$$
E[X] = \sum_{n=1}^{\infty} n (1 - p)^{n-1} p = 1/p.
$$

Observe that $E[X] = 1$ if $p = 1$ and $E[X] \to \infty$ when $p \to 0$ which agrees with the intuition.

A.6 Jointly distributed random variables

Sometimes it is of interest to investigate two or more rvs. If $X$ and $Y$ are defined on the same probability space, we define the joint cumulative distribution function (j.c.d.f.) for $X$ and $Y$ for all real $x$ and $y$ by

$$
F(x, y) = P(X \leq x, Y \leq y) = P(\{X \leq x\} \cap \{Y \leq y\}).
$$

Define $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$ for all real $x$ and $y$. $F_X$ and $F_Y$ are called the marginal cumulative distribution functions of $X$ and $Y$, respectively, corresponding to the joint distribution function $F$.

Note that $F_X(x) = \lim_{y \to +\infty} F(x, \infty)$ and $F_Y(y) = \lim_{x \to +\infty} F(\infty, y)$.

If there exists a nonnegative function $f$ of two variables such that

$$
P(X \leq x, Y \leq Y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) \, du \, dv
$$

then $f$ is called the joint density function for the rvs $X$ and $Y$.

Suppose that $g$ is a function of two variables and let $f$ be the joint density function of $X$ and $Y$. The expectation $E[g(X, Y)]$ is defined as

$$
E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \, f(x, y) \, dx \, dy
$$

provided that the integral exists.

Consider now the case when $X$ and $Y$ are discrete rvs taking values in some countable sets $I$ and $J$, respectively. Then the joint distribution function for $X$ and $Y$ for all $x \in I$, $y \in J$, is given by

$$
F(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\}).
$$

Define $F_X(x) = P(X = x)$ and $F_Y(y) = P(Y = y)$ for all $x \in I$ and $y \in J$ to be the marginal distribution functions of $X$ and $Y$, respectively, corresponding to the joint distribution function $F$.

From the law of total probability, we see that $F_X(x) := \sum_{y \in J} F(x, y) = P(X = x)$ for all $x \in I$ and $F_Y(y) := \sum_{x \in I} F(x, y) = P(Y = y)$ for all $y \in J$.

Suppose that $g$ is a nonnegative function of two variables. The expectation $E[g(X, Y)]$ is defined as

$$
E[g(X, Y)] = \sum_{x \in I, y \in J} g(x, y) P(X = x, Y = y)
$$

provided that the summation exists.
A.7 Independent random variables

Two rvs $X$ and $Y$ are said to be independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

for all real $x$ and $y$ if $X$ and $Y$ are continuous rvs, and if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all $x \in I$ and $y \in J$ if $X$ and $Y$ are discrete and take their values in $I$ and $J$, respectively.

**Proposition 35.** If $X$ and $Y$ are independent rvs such that $E[X]$ and $E[Y]$ exist, then

$$E[X Y] = E[X]E[Y].$$

Let us prove this result when $X$ and $Y$ are discrete rvs taking values in the sets $I$ and $J$, respectively. Let $g(x, y) = xy$ in the definition of $E[g(X, Y)]$ given in the previous section. Then,

$$E[X Y] = \sum_{x \in I, y \in J} xyP(X = x, Y = y)$$

$$= \sum_{x \in I, y \in J} xyP(X = x)P(Y = y) \quad \text{since } X \text{ and } Y \text{ are independent rvs}$$

$$= \sum_{x \in I} xP(X = x)\left(\sum_{y \in J} yP(Y = y)\right)$$

$$= E[X]E[Y].$$

The proof when $X$ and $Y$ are both continuous rvs is analogous and is therefore omitted.

A.8 Conditional expectation

Consider the situation in Example 13. Let $X$ be the number of polls required to find a ready terminal and let $Y$ be the number of ready terminals. The mean number of polls given that $Y = 1, 2, 3, 4, 5$ is the conditional expectation of $X$ given $Y$ (see the computation in Example 23).

Let $X$ and $Y$ be discrete rvs with values in the sets $I$ and $J$, respectively.

Let $P_{X|Y}(x, y) := P(X = x | Y = y)$ be the conditional probability of the event $(X = x)$ given the event $(Y = y)$. From Proposition 31 we have

$$P_{X|Y}(x, y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

for each $x \in I$, $y \in J$, provided that $P(Y = y) > 0$.

$P_{X|Y}(\cdot | y)$ is called the conditional distribution function of $X$ given $Y = y$.

The conditional expectation of $X$ given $Y = y$, denoted as $E[X | Y = y]$, is defined for all $y \in J$ such that $P(Y = y) > 0$, by

$$E[X | Y = y] = \sum_{x \in I} x P_{X|Y}(x, y)$$

**Example 23.** Consider Example 13. Let $X \in \{1, 2, 3, 4, 5\}$ be the number of polls required to find a terminal in the ready state and let $Y \in \{1, 2, 3, 4, 5\}$ be the number of ready terminals. We want to compute $E[X | Y = 3]$, the mean number of polls required given that $Y = 3$.

We have

$$E[X | Y = 3] = 1 \times P_{X|Y}(1, 3) + 2 \times P_{X|Y}(2, 3) + 3 \times P_{X|Y}(3, 3)$$
\[
\begin{align*}
1 \times P(A_1) &+ 2 \times P(A_2) + 3 \times P(A_3) \\
= \frac{6}{10} + 2 \times \frac{3}{10} + 3 \times \frac{1}{10} &= \frac{15}{10} = 1.5.
\end{align*}
\]

The result should not be surprising since each terminal is equally likely to be in the ready state.

Consider now the case when \(X\) and \(Y\) are both continuous rvs with density functions \(f_X\) and \(f_Y\), respectively, and with joint density function \(f\). The conditional probability density function of \(X\) given \(Y = y\), denoted as \(f_{X|Y}(x, y)\), is defined for all real \(y\) such that \(f_Y(y) > 0\), by

\[
f_{X|Y}(x, y) = \frac{f(x, y)}{f_Y(y)}.
\]

The conditional expectation of \(X\) given \(Y = y\) is defined for all real \(y\) such that \(f_Y(y) > 0\), by

\[
E[X | Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x, y) \, dx.
\]

Below is a very useful result on conditional expectation. This is the version of the law of total probability for the expectation.

**Proposition 36 (Law of conditional expectation).** For any rvs \(X\) and \(Y\),

\[
E[X] = \sum_{y \in J} E[X | Y = y] P(Y = y)
\]

if \(X\) is a discrete rv, and

\[
E[X] = \int_{-\infty}^{+\infty} E[X | Y = y] f_Y(y) \, dy
\]

if \(X\) is a continuous rv.

We prove the result in the case when \(X\) and \(Y\) are both discrete rvs. Since \(E[X | Y = y] = \sum_{x \in I} x P(X = x | Y = y)\) we have from the definition of the expectation that

\[
\sum_{y \in J} E[X | Y = y] P(Y = y) = \sum_{y \in J} \left( \sum_{x \in I} x P(X = x | Y = y) \right) P(Y = y)
\]

\[
= \sum_{x \in I} \left( \sum_{y \in J} P(X = x | Y = y) P(Y = y) \right)
\]

\[
= \sum_{x \in I} x P(X = x) \quad \text{by using the law of total probability}
\]

\[
= E[X].
\]

The proof in the case when \(X\) and \(Y\) are continuous rvs is analogous and is therefore omitted.

\[\Box\]

## B Stochastic Process

### B.1 Definitions

All the rvs considered from now on are supposed to be constructed on a common probability space \((\Omega, \mathcal{F}, P)\).

**Notation:** We shall denote by \(\mathbb{N}\) the set of all nonnegative integers and by \(\mathbb{R}\) the set of all real numbers.
A collection of rvs $X = (X(t), t \in T)$ is called a stochastic process. In other word, for each $t \in T$, $X(t)$ is a mapping from $\Omega$ into some set $E$ where $E = \mathbb{R}$ or $E \subset \mathbb{R}$ (e.g., $E = [0, \infty)$, $E = \mathbb{N}$) with the interpretation that $X(t)(\omega)$ (also written as $X(t, \omega)$) is the value of the stochastic process $X$ at time $t$ on the outcome (or path) $\omega$.

The set $T$ is the index set of the stochastic process.

If $T$ is countable (e.g., $T = \mathbb{N}$, $T = \{\ldots, -2, -1, 0, 1, 2 \ldots\}$) then $X$ is called a discrete-time stochastic process; if $T$ is continuous (e.g., $T = \mathbb{R}$, $T = [0, \infty)$) then $X$ is called a continuous-time stochastic process.

When $T$ is countable one will in general substitute the notation $X(t)$ for $X(n)$ (or $X(n)$, $t_n$, etc.).

The space $E$ is called the state-space of the stochastic process $X$. If the set $E$ is countable then $X$ is called a discrete-space stochastic process; if the set $E$ is continuous then $X$ is called a continuous-space stochastic process.

When speaking of “the process $X(t)$” one should understand the process $X$. This is a common abuse of language.

**Example 24** (Discrete-time discrete-space stochastic process). $X(n) =$ number of jobs processed during the $n$-th hour of the day. The stochastic process $(X(n), n = 1, 2, \ldots, 24)$ is a discrete-time discrete-space stochastic process.

**Example 25** (Discrete-time continuous-space stochastic process). $X(n) =$ response time of the $n$-th inquiry to a central processing system of an interactive computer system. The stochastic process $(X(n), n = 1, 2, \ldots)$ is a discrete-time continuous-space stochastic process.

**Example 26** (Continuous-time discrete-space stochastic process). $X(t) =$ number of messages that have arrived at a given node of a communication network in the time period $(0, t)$. The stochastic process \{$(X(t), t \geq 0)$\} is a continuous-time discrete-space stochastic process.

**Example 27** (Continuous-time continuous-space stochastic process). $X(t) =$ waiting time of an inquiry received at time $t$. The stochastic process \{$(X(t), t \geq 0)$\} is a continuous-time continuous-space stochastic process.

Introduce the following notation: a function $f$ is $o(h)$ if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$ 

For instance, $f(h) = h^2$ is $o(h)$, $f(h) = h$ is not, $f(h) = h^r$, $r > 1$, is $o(h)$, $f(h) = \sin(h)$ is not. Any linear combinaison of $o(h)$ functions is also $o(h)$.

**Example 28.** Let $X$ be an exponential rv with parameter $\lambda$. In other words, $P(X \leq x) = 1 - \exp(-\lambda x)$ for $x \geq 0$ and $P(X \leq x) = 0$ for $x < 0$. Then, $P(X \leq h) = \lambda h + o(h)$.

Similarly, $P(X \leq t + h \mid X > t) = \lambda h + o(h)$ since $P(X \leq t + h \mid X > t) = P(X \leq h)$ from the memoryless property of the exponential distribution.

**C Poisson Process**

A Poisson process is one of the simplest interesting stochastic processes.

Consider a point process $(t_n)_n$, that is a collection of random points in time such that $0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1} < \cdots$, where $t_n$ records the occurrence time of the $n$-th event in some experiment. For instance, $t_n$ will be the arrival time of the $n$-th request to a database.
For any interval \([s, t]\) define the integer rv \(N([s, t]) = \sum_{n \geq 1} 1(t_n \in [s, t])\). In words, \(N([s, t])\) is the number of occurrences (or events) of the point process \((t_n)\) in \([s, t]\).

We say that \((N([s, t]), 0 \leq s < t)\) is a Poisson process if

1. \(\{(t_{n+1} - t_n), n = 0, 1, \ldots\}\) is a collection of independent and identically distributed rvs (with \(t_0 = 0\) by convention);
2. \(t_{n+1} - t_n\) is exponentially distributed with rate \(\lambda > 0\), namely,

\[
P(t_{n+1} - t_n < x) = 1 - \exp(-\lambda x), \quad x > 0.
\]

It is a common abuse of terminology to say that \((t_n)\) is a Poisson process if (1)-(2) hold. We will also use this terminology.

One of the original applications of the Poisson process in communications was to model the arrivals of calls to a telephone exchange (the work of A. K. Erlang in 1919). The use of each telephone, at least in a first analysis, can be modeled as a Poisson process.

Below are important consequences of the definition of a Poisson process.

**Proposition 37** (Stationary and independent increments).

(i) Stationary increments: The number of occurrences of a Poisson process in a given time interval only depends on the length of this interval.

(ii) Independent increments: The number of occurrences of a Poisson process in two disjoint time intervals are independent rvs.

Both results immediately follow from the memoryless property of the exponential distribution and from the assumption that the inter-event times \((t_{n+1} - t_n)\) are iid rvs.

**Proposition 38** (Poisson distribution).

Let \(P_n(t)\) be the probability that exactly \(n\) events occur in an interval of length \(t\). We have, for each \(n \in \mathbb{N}, t \geq 0\),

\[
P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.
\] (173)

The probability distribution in (173) is called a Poisson distribution.

**Proof.** The proof uses the fact that the sum of \(n\) iid exponential defines an Erlang-n rv whose probability density function is known in explicit form. More specifically, if \(Y_1, \ldots, Y_n\) are iid rvs with common exponential probability distribution \(P(Y_1 < x) = 1 - \exp(-\lambda x)\) then the probability density function of \(Y_1 + \cdots + Y_n\) is given by

\[
\frac{d}{dx} P(Y_1 + \cdots + Y_n < x) = \frac{\lambda^n x^{n-1} \exp(-\lambda x)}{(n-1)!}.
\] (174)

The proof goes as follows. First observe that

\[
t_k = \tau_1 + \cdots + \tau_k
\] (175)

with \(\tau_k := t_k - t_{k-1}\) (recall that \(t_0 = 0\) by convention).

We know from Proposition 37 that \(P_n(t) = P(N([0, t]) = n)\) for any \(t\). For this reason we will only focus on the probability distribution of \(N([0, t])\). We have

\[
P(N([0, t]) = n) = P(t_n > t, t_{n+1} < t)
\]

\[
= P(\tau_1 + \cdots + \tau_n > t, \tau_1 + \cdots + \tau_{n+1} < t) \text{ from (175)}
\]

\[
= \int_t^\infty P(\tau_1 + \cdots + \tau_{n+1} < t \mid \tau_1 + \cdots + \tau_n = y) \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!} dy \text{ from (174)}
\]

\[
= \int_t^\infty P(\tau_{n+1} < t - y \mid \tau_1 + \cdots + \tau_n = y) \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!} dy
\]
\[
\begin{align*}
\int_t^\infty P(\tau_{n+1} < t - y) \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!} dy 
&= \int_t^\infty \left(1 - e^{-\lambda(t-y)}\right) \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!} dy 
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\end{align*}
\]

which is obtained by performing an integration by part.

A function \( f \) is \( o(h) \) if \( f(h)/h \to 0 \) as \( h \to 0 \). From Proposition 38 we conclude that

**Corollary 2.**

\[
\text{Prob(a single event in an interval of duration } h) = \lambda h + o(h)
\]

and

\[
\text{Prob(more than one event in an interval of duration } h) = o(h).
\]

Let us now compute \( E[N(t)] \), the expected number of events of a Poisson process with rate \( \lambda \) in an interval of length \( t \).

**Proposition 39** (Expected number of events in an interval of length \( t \)).

For each \( t \geq 0 \)

\[
E[N(t)] = \lambda t.
\]

Proposition 39 says that the expected number of events per unit of time, or equivalently, the rate at which the events occur, is given by \( E[N(t)]/t = \lambda \). This is why \( \lambda \) is called the rate of the Poisson process.

**Proof of Proposition 39.** We have by using Proposition 38

\[
E[N(t)] = \sum_{k=0}^{\infty} k P_k(t) = \sum_{k=1}^{\infty} k P_k(t)
\]

\[
= \left( \sum_{k=1}^{\infty} k \frac{(\lambda t)^k}{k!} \right) e^{-\lambda t}
\]

\[
= \lambda t \left( \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} \right) e^{-\lambda t}
\]

\[
= \lambda t.
\]

**Proposition 40.** The superposition of two independent Poisson processes with rates \( \lambda_1 \) and \( \lambda_2 \) is a Poisson process with rate \( \lambda_1 + \lambda_2 \).

The proof is omitted.

**Example 29.** Consider the failures of a link in a communication network. Failures occur according to a Poisson process with rate 2.4 per day. We have:

(i) \( P(\text{time between failures } \leq T \text{ days}) = 1 - e^{-2.4T} \)

(ii) \( P(k \text{ failures in } T \text{ days}) = \frac{(2.4T)^k}{k!} e^{-2.4T} \)

(iii) Expected time between two consecutive failures = 10 hours

(iv) \( P(0 \text{ failure in next day}) = e^{-2.4} \)
(v) Suppose 10 hours have elapsed since the last failure. Then,

\[ \text{Expected time to next failure} = 10 \text{ hours (memoryless property)}. \]