### Chapter 6

#### Queueing

Systems are often modeled by automata, and discrete events are transitions from one state to another. In this chapter we want to analyze such discrete events systems. We assume that events are stochastic, and we want to know how our system behaves on average.

If the events happen in discrete time (for example, there is an event every hour, on the hour), the tool to model the system is called Discrete Time Markov Chain (DTMC). The system transitions from one state to another, according to the probabilities of the transitions at the current state.

In Continuous Time Markov Chains (CTMCs) events (transitions) happen at arbitrary times, for example, whenever a new customer enters the store. In this lecture we concentrate on CTMCs.

#### 6.1 Continuous Time Markov Chains

**Definition 6.1 (Continuous Time Markov Chain, CTMC).** Let $S$ be a finite or countably infinite set of states. A **Continuous Time Markov Chain (CTMC)** is a continuous time stochastic process \( \{X_t : t \in \mathbb{R}_{\geq 0}\} \) with \( X_t \in S \) for all \( t \) that satisfies the continuous Markov property.

**Definition 6.2 (Continuous Markov Property).** A Markov chain satisfies the **Markov property** if the probability for the next state depends only on the current state, and not the history. Such a system is also called memoryless.

**Remarks:**
- We will only consider time-homogeneous CTMCs for which the transition probability \( p(X_{t_2} = j | X_{t_1} = i) \) from state \( i \) to \( j \) in the time period \( [t_1, t_2) \) depends only on the difference \( \Delta t = t_2 - t_1 \) and not on the times \( t_1, t_2 \) themselves.
- The sojourn times for time-homogeneous CTMCs are exponentially distributed, cf. Definition 6.4.

**Definition 6.3 (Sojourn Time).** The **sojourn time** \( T_i \) of state \( i \) is the time the process stays in state \( i \).

**Definition 6.4 (Exponential Distribution).** A random variable \( Y \) with the cumulative distribution function (CDF)

\[
F_Y(t) = \Pr[Y \leq t] = \begin{cases} 1 - e^{-\lambda t} & \text{for } t \geq 0, \\ 0 & \text{otherwise} \end{cases}
\]

is **exponentially distributed** with parameter \( \lambda \), or \( Y \sim \exp(\lambda) \) for short. The corresponding probability density function (PDF) is

\[
f_Y(t) = \frac{d}{dt}F_Y(t) = \lambda e^{-\lambda t}.
\]

**Remarks:**
- If \( Y \sim \exp(\lambda) \), then \( \mathbb{E}[Y] = 1/\lambda \) and \( \text{Var}[Y] = 1/\lambda^2 \).
- The exponential distribution is the continuous analogue to the discrete-time geometric distribution, i.e., the probability of an event is the same in every discrete time step, where the duration of the discrete steps goes towards 0.
- The exponential distribution is the only memoryless continuous distribution.
- Consider the continuous time stochastic process \( \{X_t : t \in \mathbb{R}_{\geq 0}\} \) counting the number of events up to time \( t \), where the time between two consecutive events is exponentially distributed with parameter \( \lambda \). Then \( X_t \) is a Poisson process with rate \( \lambda \). According to the Poisson distribution we can expect \( \lambda \) events per time unit.
- Let us consider an example of a CTMC.

![Figure 6.5: A CTMC modeling an unreliable system. In state 1 the system is working, in state 0 the system is faulty. The failure rate, i.e., the time until the system fails, is exponentially distributed with parameter \( \lambda \). After a failure, the repair takes some time, exponentially distributed with parameter \( \mu \).](image)

**Lemma 6.6.** Let \( Y_1, \ldots, Y_k \) be independent exponential random variables with corresponding parameters \( \lambda_1, \ldots, \lambda_k \). The random variable \( Y = \min\{Y_1, \ldots, Y_k\} \) is exponentially distributed with parameter \( \lambda_1 + \cdots + \lambda_k \).
Proof. We establish the claim for \( k = 2 \). The general case can be derived by applying the same reasoning. By definition it holds for \( Y \), \( Y_1 \), and \( Y_2 \) that

\[
\Pr(Y > t) = \Pr(\text{min}(Y_1, Y_2) > t) = \Pr(Y_1 > t) \cdot \Pr(Y_2 > t).
\]

Since the random variables \( Y_1 \) and \( Y_2 \) are independent, this is the same as

\[
\Pr(Y > t) = \Pr(Y_1 > t) \cdot \Pr(Y_2 > t) = e^{-\lambda t} e^{-\lambda t} = e^{-(\lambda_1 + \lambda_2) t}.
\]

It follows that the random variable \( Y = \text{min}(Y_1, Y_2) \) is exponentially distributed with parameter \( \lambda_1 + \lambda_2 \).

Lemma 6.7. Let \( Y_1, \ldots, Y_k \) be \( k \) independent exponential random variables with corresponding parameters \( \lambda_1, \ldots, \lambda_k \). The probability \( \Pr(Y_1 = \text{min}(Y_1, \ldots, Y_k)) \) is \( \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_k} \).

Proof. Let \( Z \) be the random variable \( Z = \text{min}(Y_1, \ldots, Y_k) \). Lemma 6.6 states that \( Z \) is exponentially distributed with parameter \( \mu = \lambda_2 + \cdots + \lambda_k \). Applying the law of total probability we obtain that the probability for \( Y_1 \) to take on the smallest value is

\[
\Pr(Y_1 < Z) = \int_0^\infty \Pr(Y_1 < Z | Y_1 = t) \cdot f_{Y_1}(t) \, dt
\]

Since \( Z \) is independent of \( Y_1 \), we can simplify to

\[
\Pr(Y_1 < Z) = \int_0^\infty (1 - \Pr[Z \leq t]) \cdot f_{Y_1}(t) \, dt.
\]

Recall that the probability density function of \( Y_1 \) is \( f_{Y_1}(t) = \lambda_1 e^{-\lambda_1 t} \), and that the cumulative distribution function for \( Z \) is \( F_Z(t) = 1 - e^{-\mu t} \). Plugging both in, we obtain

\[
\Pr(Y_1 < Z) = \lambda_1 \int_0^\infty e^{-\lambda_1 t} e^{-\mu t} \, dt = \lambda_1 \int_0^\infty e^{-(\lambda_1 + \mu) t} \, dt
\]

\[
= \lambda_1 \cdot \left( \frac{1}{\lambda_1 + \mu} \right) (\lambda_1 + \mu) = \lambda_1 \cdot \left( 0 - \frac{0}{\lambda_1 + \mu} \right)
\]

\[
= \lambda_1 \cdot \frac{\gamma}{\lambda_1 + \gamma} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_k}.
\]

Theorem 6.9. For all \( t \in S \), the change in the state probability \( q(t) \) is

\[
\frac{dq(t)}{dt} = \sum_{j \in S} q(t) \cdot \lambda_{ij} - q(t) \cdot \lambda_i.
\]

Remarks:

- Theorem 6.9 follows from the memoryless property and relies on the CTMC being time homogeneous.

- Solving such differential equations for exact values of \( t \) can be a laborious task. We can look at the stationary distribution instead. The mathematical notion that captures a Markov chain’s long term behavior is the stationary distribution. Informally, a stationary distribution should satisfy that \( \sum_{t \in S} \hat{q}(t) = 0 \) “after enough time has passed”.

Definition 6.10 (Stationary Distribution). For \( t \to \infty \), \( \pi \) is a stationary distribution if for all \( i \in S \),

\[
0 = \sum_{j \in S} \pi_j \cdot \lambda_{ij} - \pi_i \cdot \lambda_i.
\]
6.2. Kendall’s Notation for Queues

Remarks:
- Thus, one can solve above system of linear equations in order to compute the stationary distribution. Since we are interested in a probability distribution, the solution must additionally satisfy the conditions $\pi_i \geq 0$ and $\sum \pi_i = 1$.

Definition 6.11 (Irreducible). A CTMC is irreducible if for all states $i$ and $j$ it holds that $j$ is reachable from $i$. That is, if there exists some $t \geq 0$ such that $\Pr[X_t = j | X_0 = i] > 0$.

Theorem 6.12. For finite irreducible CTMCs the limits

$$
\lim_{t \to \infty} q(t)
$$

exist for all $i \in S$. Moreover, the entries in $\pi$ are independent of $q(0)$.

Remarks:
- CTMCs for which the stationary distribution exists are called ergodic. For finite chains this is the same as being irreducible. We will later see examples of irreducible infinite chains that are not ergodic.
- In our examples from Figure 6.5 we obtain the following two equations:

$$
0 = \mu \pi_0 - \lambda \pi_1, \quad \text{and} \quad 0 = \lambda \pi_1 - \mu \pi_0.
$$

Since it must also hold that $\pi_1 + \pi_0 = 1$, we conclude that in the long run, the probability of being in the working respectively faulty state are

$$
\pi_1 = \frac{\mu}{\lambda + \mu} \quad \text{and} \quad \pi_0 = \frac{\lambda}{\lambda + \mu}.
$$

6.2 Kendall’s Notation for Queues

Queueing theory can be a diversion to think about while queueing at the cash register, but it is also used in modeling telecommunication networks, traffic, factories, or internet servers, as illustrated in Figure 6.14.

Definition 6.13 (Jobs, Servers). A queueing system consists of a queue with one or more servers which process jobs. The queue acts as a buffer for jobs that arrived but cannot be processed yet, because the server is busy processing another job.

Remarks:
- A job may be a shopper, a phone call, a web request, etc. A server may model a checkout clerk, a factory, or a telephone network.

Figure 6.14. A queueing system with one server. Jobs arrive at the queue from a Poisson process with rate $\lambda$, i.e., the inter-arrival time between two jobs is exponentially distributed with parameter $\lambda$. If the system is empty, the job is processed immediately, otherwise the job waits in the queue. The time it takes to process a single job is exponentially distributed with parameter $\mu$, and after one job has been served, if there is a job waiting, the server starts to process the next job.

Definition 6.15 (Kendall’s Notation). Let $a$ and $s$ be symbols describing the arrival and service rates, and let $m, n, j \in \mathbb{N}$. The Kendall notation for a queueing system $Q$ is $a/(m/n)/j$. The symbols $a$ and $s$ can be $D$, $M$, or $G$, where

- $D$ means that the rate distribution is degenerate, i.e., of fixed length,
- $M$ means that the arrival/service process is memoryless, and
- $G$ means that the corresponding rate stems from a generic distribution.

The parameter $m$ is the number of servers, $n$ is the number of places in the system (in the queue and at servers), and $j$ determines the external population of jobs that may enter the system. The latter two parameters are omitted if the respective number is unbounded.

Remarks:
- Extensions to Kendall’s notation include other kinds of distributions for the arrival and service times. We will only consider memoryless processes, i.e., the arrival and service times are exponentially distributed.
- One reason is of course that the memoryless property allows for simpler math. But more importantly, memoryless processes turn out to be a good approximation for many real world systems, and thus memoryless queueing theory is a good tool to model such cases.
- When using this tool, one should be aware that for instance bursty behavior, where batches of jobs sometimes arrive in quick succession (think of a new trend appearing on Twitter) is not captured well by memoryless distributions.
- The parameter $m$ in Kendall’s notation limits how many jobs may be present in the system, and how many jobs are rejected by the queueing system. The parameter $j$ affects the arrival rate—if a large fraction of the population is already in the queue, then jobs are less likely to arrive, and vice versa.
- Another parameter may be added to indicate the queueing discipline, i.e., in which order jobs are served. For our discussion this distinction
is not necessary, and you may assume a First In First Out (FIFO) order. Other queueing disciplines are, e.g., Last In First Out (LIFO), random order, or queues where jobs have different priorities.

6.3 The $M/M/1$ Queue

Figure 6.16: A CTMC modeling an $M/M/1$ system. In state 0 the system is empty. When the chain is in state $i \geq 1$, then there are $i-1$ jobs in the queue, and one job is being served with rate $\mu$. New jobs arrive with rate $\lambda$. Since the exponential distribution is memoryless, switching from state $i$ to $i+1$ does not change the probability distribution for the service time of the currently processed job.

Theorem 6.17. An $M/M/1$ queueing system has a stationary distribution if and only if $\rho = \lambda/\mu < 1$. In that case the stationary distribution is $\pi_k = \rho^k/(1 - \rho)$.

Proof. In the stationary distribution, the change in probability mass at every node must be zero. We obtain the equations

$$0 = \mu \cdot \pi_1 - \lambda \pi_0$$

for state 0, and

$$0 = \lambda \cdot \pi_{k+1} + \mu \cdot \pi_k + (\lambda + \mu)\pi_k$$

for all $k \geq 1$. Rearranging yields

$$\mu \cdot \pi_{k+1} - \lambda \cdot \pi_{k+1} + \mu \cdot \pi_k - \lambda \cdot \pi_k - \cdots - \mu \cdot \pi_1 - \lambda \cdot \pi_0 = 0$$

$$\Rightarrow \mu \cdot \pi_k - \lambda \cdot \pi_{k+1} = 0 \Rightarrow \pi_k = \rho \cdot \pi_{k+1} \Rightarrow \pi_0 = \rho^0 \cdot \pi_0$$

In the case where $\rho \geq 1$ the only solution is $\pi = (0, 0, \ldots)$. This means that the queuing system does not converge, and that the length of the queue grows indefinitely. If on the other hand $\rho < 1$, then:

$$1 = \sum_{k=0}^{\infty} \pi_k = \pi_0 \sum_{k=0}^{\infty} \rho^k = \pi_0 \frac{1}{1 - \rho} \Rightarrow \pi_0 = \frac{1}{1 - \rho}.$$
6.4 LITTLE’S LAW

In many cases, for \( t \to \infty \), the expected values are equal to the limit of the random variables with probability 1.

So far we suggested a FIFO (first in first out) queueing discipline. To prove Little’s Law this assumption was not required, i.e., Theorem 6.21 also holds for systems other than \( M/M/1 \) queues.

Applying Little’s Law we conclude that in the steady state the average response time is

\[
T = \frac{\lambda}{\mu} = \frac{\lambda}{\rho},
\]

since for \( M/M/1 \) queueing systems we know that \( \rho = \frac{\lambda}{\mu} \).

Definition 6.23 (Waiting Time, Jobs in the Queue). We denote by \( N \) the average number of jobs waiting in the queue.

Remarks:

- Similar to the time in the system, for \( M/M/1 \) queueing systems the average waiting time of a job is
  \[
  W = \frac{\rho}{\mu} = \frac{\rho}{\rho - 1}.
  \]
- The average number of jobs in the queue is
  \[
  Q = \frac{\rho}{1 - \rho}.
  \]

6.5 Birth-Death Processes

Our CTMC for the \( M/M/1 \) queueing system is a special case of a so-called Birth-Death Process.

Remarks:

- As before we can compute the stationary distribution. We obtain
  \[
  \pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \frac{\lambda_k}{\mu_{k+1}}},
  \text{ and }
  \pi_k = \pi_0 \frac{\lambda_k}{\mu_{k+1}} \text{ for } k \geq 1.
  \]

\( M/M/m \) Queues

What if there is a single queue for multiple servers, e.g., in a service hotline? In Kendall’s notation such systems are written as \( M/M/m \) systems, where \( m \) denotes the number of servers.
6.5. BIRTH-DEATH PROCESSES

Figure 6.25: Birth-Death process modeling an $M/M/3$ queueing system. If there are less than 3 jobs, then the number of active servers is the number of jobs in the system. When 3 or more jobs are in the system all servers are active.

Remarks:
- In $M/M/m$ queueing systems, the utilization $\rho$ is the average fraction of active servers.
- If $\rho = \frac{\lambda}{\mu} < 1$, then the stationary distribution is
  $$\pi_k = \left\{ \begin{array}{ll}
  \pi_0 \frac{\mu^m}{m!} & \text{for } 1 \leq k \leq m \\
  \pi_0 \rho^m & \text{for } k \geq m,
  \end{array} \right.$$ 
  and
  $$\pi_0 = \frac{1}{\sum_{k=0}^{m} \frac{\mu^m}{m!} + \pi_0 \rho^m}.$$ 
- The probability that in the stationary distribution an arriving job has to wait in the queue is
  $$P_Q = \sum_{k=m}^{\infty} \pi_k = \sum_{k=m}^{\infty} \rho^m (1 - \rho)^{k-m} = \pi_0 (1 - \rho)^m.$$ 
- The average number of jobs in the queue $N_Q$ can be calculated in a similar fashion. With $P_Q$ the number can be expressed as
  $$N_Q = P_Q \frac{\rho}{1 - \rho}.$$ 

The $M/M/m/n$ Queue

Often, the space in the queue is bounded, i.e., the system is $M/M/m/n$. Recall that $n$ is the number of places in the system, so the maximum length of the queue is $n - m$.

Figure 6.26: Birth-Death process modeling an $M/M/5/5$ queueing system.

Remarks:
- The case $m = n$ is often used to model communication networks. Such a system can accommodate $m$ simultaneous calls, and the duration of a call is distributed with $\text{exp}(\mu)$. One can calculate that in this case
  $$\pi_k = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} \text{ for } 1 \leq k \leq m.$$ 
  Using that $\sum_{k=0}^{m} \pi_k = 1$ yields that the probability to be in state 0 is
  $$\pi_0 = \frac{1}{\sum_{k=0}^{m} \left( \frac{\lambda}{\mu} \right)^k}.$$ 
- The blocking probability, i.e., the probability that an arriving job is rejected, is thus
  $$\pi_n = \frac{1}{\sum_{k=0}^{m} \left( \frac{\lambda}{\mu} \right)^k}.$$ 

This so-called Erlang-B formula also holds for $M/G/m/m$ systems where the service times are $1/\mu$ in expectation, regardless of their distribution.

The $M/M/m/\infty$ Queue

In telephone networks the population is assumed to be much larger than the number of places in the system. Thus, it is justified to assume that the arrival rate is independent of the number of jobs in the system. Cases where this assumption cannot be made can be modeled as $M/M/n/m/j$ systems.

Figure 6.27: Birth-Death process modeling an $M/M/1/m/m$ queueing system.
Remarks:

- For $M/M/1/m$ systems, one can calculate that
  \[ \tau_k = \frac{\lambda_k}{\mu} \] for $1 \leq k \leq m$
  \[ \tau_0 = \sum_{k=1}^{m} \left( \frac{\lambda}{\mu} \right)^k - 1 \]
  where $\lambda_k = m(m-1)(m-2)\ldots(m-k+1)$.

- Many other queueing systems are possible. For example, jobs might have either high or low priority, and low priority jobs will only be served when no high priority job is waiting, with or without preemption when a new high priority job is arriving. Also such a priority-based queueing system can be modeled by CTMCs, in the case of two priorities the CTMC becomes two-dimensional.

### 6.6 Queueing Networks

Sometimes, systems consist of more than a single queueing system. Consider, for instance, a support call center where calls are initially handled by first-line support. Customers with problems that cannot be solved by the first-line support are handed over to technicians with a separate queue. See Figure 6.28 for an illustration.

Figure 6.28: A queueing network modeling a two-tier support hotline. Jobs arrive from the outside with rate $\lambda_1$ and enter the queue for first-line support. After the first-line support served the job, with rate $\mu_1$, a fraction $(1 - p_{1.2})$ of the jobs are satisfied and leave the system. The remaining $p_{1.2}$ fraction of jobs need in-depth technical assistance, which is provided by a technician in the technical support queue. Technical support takes time exponentially distributed with parameter $\mu_2$, and afterwards the job leaves the system.

Remarks:

- Before looking at the whole network, let us look at a single queueing system. If the queueing system is stable, i.e., $\rho < 1$, what is the inter-departure time between consecutive departing jobs?

**Theorem 6.29 (Burke’s Theorem).** Consider a $M/M/1$ queue with arrival rate $\lambda$ and service rate $\mu$. If the system is stable, then in the steady state the time between two departures is exponentially distributed with parameter $\lambda$.

**Proof.** Consider any point in time, and let $T$ be the random variable for the time until the next job leaves the queueing system. Denoting by $\rho$ the probability that the system is not empty, we can write

\[ \Pr[T \leq t] = \rho \cdot \Pr[T \leq t \mid \text{system not empty}] + (1 - \rho) \cdot \Pr[T \leq t \mid \text{system empty}] \]

When the queueing system is not empty, we know that the arrival and service rates are exponentially distributed, the term can be rewritten as

\[ \Pr[T \leq t] = \rho \cdot \Pr[S \leq t] + (1 - \rho) \cdot \Pr[A + S \leq t] \]

where $A \sim \exp(\lambda)$ and $S \sim \exp(\mu)$ are random variables describing the arrival and service time of the next arriving job, respectively. By conditioning on $S$ we obtain

\[ \Pr[T \leq t] = \rho \cdot \Pr[S \leq t] + (1 - \rho) \cdot \int_0^t \Pr[A + S \leq t \mid S = s] \cdot f_S(s) \, ds \]

\[ = \rho \cdot \Pr[S \leq t] + (1 - \rho) \cdot \int_0^t \Pr[A \leq t - s] \cdot f_S(s) \, ds \]

Plugging in the probability density and distribution function and solving the integral yields

\[ \Pr[T \leq t] = \rho \cdot (1 - e^{-\lambda t}) + (1 - \rho) \cdot (1 - e^{-\mu t}) - (1 - \rho) \cdot \mu \cdot \left( \frac{e^{-\lambda t} - e^{-\mu t}}{\lambda - \mu} \right) \]

By rearranging we get that $\Pr[T \leq t] = 1 - e^{-\lambda t}$, which means that $T$ is exponentially distributed with parameter $\lambda$, as desired.

Remarks:

- Burke’s theorem also holds for the more general $M/M/m$ queues.
- It simplifies the analysis of $M/M/m$ queueing systems in the stationary case. Perhaps surprisingly, the departure process does not depend on the time it takes to serve a job, but just on the rate of arrivals.
- The stochastic process counting the number of arrivals or departures from a memoryless queueing system up to time $t$ is a Poisson process.
- What about networks of queues?
\[ \pi(k_1, \ldots, k_n) = \prod_{i=1}^{n} \pi_i(k_i) \]

Here \( \pi(k_1, \ldots, k_n) \) denotes the stationary distribution for the network, i.e., the probability that \( k_i \) jobs are in queueing system \( i \), and \( \pi_i(k_i) \) is the probability that \( k_i \) jobs are in \( v_i \) when considering \( v_i \) as a single \( M/M/m_i \) queue with arrival rate \( \lambda_i \), i.e., the corresponding entry in \( v_i \)'s stationary distribution.

Remarks:
- Jackson's Theorem allows us to compute the stationary distribution of an open queueing network containing memoryless queues. The distribution is obtained by computing the product of each queue's stationary distribution when considered in isolation (with arrival rate \( \lambda_i \) as above).
- Before applying the theorem, one needs to check that each queue is stable. This is done by computing the values \( \rho_i = \lambda_i / (m_i \cdot \mu_i) \) and checking that each \( \rho_i < 1 \). 
- Little's Law also applies to networks of queueing systems as a whole.
- For closed networks the stationary distribution can be computed as follows.

\[ \pi(k_1, \ldots, k_n) = \frac{1}{G(K)} \prod_{i=1}^{n} \rho_i^k_i, \]

where \( G(K) \) is the normalizing constant

\[ G(K) = \sum_{k_1, \ldots, k_n} \prod_{i=1}^{n} \rho_i^{k_i}, \]

and the values \( \rho_i \) are obtained from the \( \lambda_i \) satisfying the equations

\[ \lambda_i = \sum_{j=1}^{m_i} \lambda_j \cdot p_{j,i}. \]

Chapter Notes
The founder of queueing theory is Agner Karpov Erlang (1878–1929), who wanted to understand how the telephone network needs to be dimensioned. He already described the stationary solutions to \( M/M/m \) and \( M/M/m/n \) queues, also referred to as Erlang C and Erlang B models, respectively, and was particularly interested in the probability that the system loses a call [1]. Since then many other kinds of queues were studied, and in 1953 Kendall introduced the notation described in Definition 6.15 to better categorize previous results [4].

For a long time Little's Law (Theorem 6.21) was believed to be true without a formal proof. In a book from 1958 Morse challenged his readers to find a
A counterexample [7], but Little found a proof for the statement instead [6]. A series of papers studied variants and extensions, thus widening the applicability of the law. Fifty years later Little summarized the progress in [5].

Jackson’s Theorem for open networks (Theorem 6.32) was a first step in understanding networks of queues [3]. The stationary distribution for the closed network case (Theorem 6.33) was described by Gordon and Newell [2].

This chapter was written in collaboration with Jochen Seidel.

Bibliography


