Queuing

Recherche opérationnelle
Dimitri Watel - ENSIIE

2024

In this course, we will focus on a simple model of queues. These models allow us to estimate information about the queue, such as its size or waiting time.

Specifically, a queue consists of one or more lists of people and one or more counters to handle the individuals at the front of the line. When a person is served, they leave the queue. We assume that people are well-behaved: those who arrive queue up at the back, and no one cuts in line.

This type of queue can, of course, model a line at a ticket counter, but also different events such as packet processing in a router or the waiting times for patients waiting for an organ donation, stock levels of fruits and vegetables in a supermarket, etc.

One initial challenge in a queue is uncertainty. We do not know exactly when people will arrive in the queue and when they will leave. However, with the help of statistical models, we can estimate a distribution associated with the arrival and departure of individuals. We will see that we can model these queues with an infinite-state Markov chain. Another difficulty is the continuous aspect of the queue, as individuals can arrive and leave at any moment.

1 The M/M/k model

The M/M/k model is a simplified queuing model that assumes that people arrive and leave the queue following a Poisson stochastic process. The mixture of these two processes defines a queue. We define a random variable $X(t)$ that corresponds to the number of people in the queue at time t. The first Poisson process, called the *birth process*, randomly brings a person to the end of the queue and increases X. The second Poisson process, called the *death process*, randomly removes a person from the front of the queue and decreases X.

1.1 Model equations

Since time is continuous, we cannot easily define the probability that $X(t) = n$. Instead, we fix a time step dt, and the following equations indicate the probability of a person appearing or disappearing between the instances t and $t + dt$, so $X(t + dt) - X(t)$. These probabilities depend on the number of people in the queue at time t, allowing the model the flexibility to vary the probabilities based on the size of the queue. This is useful, for example, if we want to model a behavior of abandonment, when people think they will come back later, or attraction, when a crowd makes people curious and they approach the crowd to take a look.

We obtain the following equations that govern X:

$$\begin{aligned}
Pr(X(t+dt) - X(t) = 1 | X(t) = n) &= \lambda_n \cdot dt + o(dt) \\
Pr(X(t+dt) - X(t) = -1 | X(t) = n > 0) &= \mu_n \cdot dt + o(dt) \\
Pr(X(t+dt) - X(t) = 0 | X(t) = n > 0) &= \\
&\quad 1 - (\lambda_n + \mu_n) \cdot dt + o(dt) \\
Pr(X(t+dt) - X(t) = 0 | X(t) = 0) &= 1 - \lambda_0 \cdot dt + o(dt) \\
(|Pr(X(t+dt) - X(t)| > 1 | X(t) = n) &= o(dt))
\end{aligned}$$

where $(\lambda_n)_{n\in\mathbb{N}}$ and $(\mu_n)_{n\in\mathbb{N}^*}$ are strictly positive values.

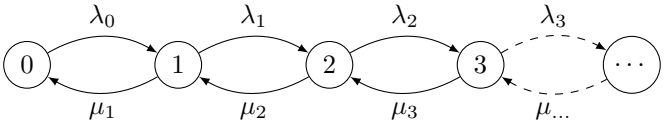
The first equation corresponds to the first process, when a person appears. The second corresponds to the second process, when a person disappears. The third and fourth indicate the probability that no one appears or disappears. It can be seen that there is a special case when $n = 0$ because, in this case, people can only appear. The last is the probability that more than one person appears or disappears. The $o(dt)$ can be understood as *a very small value that decreases very quickly and tends to 0 as dt decreases and tends to 0*. It can be seen that if dt tends to 0, the probability of appearance or disappearance of people is very low; it is more likely that X does not change between t and $t+dt$. It can also be seen that the probability of having more than one event is very low with dt. Indeed, the smaller the time step, the less likely it is that two people arrive or leave the queue between t and $t+dt$. Assuming dt is small enough, we thus obtain a fairly simple model of a queue where sometimes a single person arrives and sometimes a single person leaves.

In concrete terms, what do λ_n and μ_n correspond to? They represent the *arrival rates* and *departure rates*. These rates indicate, on average, the number of people per second who arrive and leave the queue if there are n people in the queue. It can be noted that μ_0 has no meaning since no one can leave the queue if $n = 0$. This is why there is a special case when $n = 0$.

Remark 1. These rates can be expressed in another unit; you just need to ensure that it is the same unit for both.

1.2 Graphical drawing of a queue

One can graphically represent a queue with the following drawing:



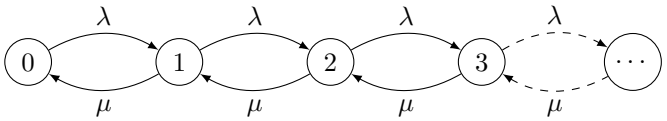
This representation echoes those of Markov chains. **Each state indicates the number of people in the queue.** It does not involve moments or durations. Each arc indicates the arrival and departure rate of a single person (thus the transition from state n to state $n+1$ or $n-1$). There are several differences here compared to the representations made in the Markov chain model. First, the chain is infinite, so it cannot be represented in its entirety, and most results about chains vary when moving from a finite chain to an infinite chain. We will therefore conduct a specific study in this course on this type of chain, which is queues. Secondly, probabilities are not indicated on the arcs. Only the rates are indicated. Finally, to have a complete chain, there would need to be arcs indicating that the number of people in the queue is constant (thus from state n to itself) and those indicating that the number of people varies by more than 1.

Important detail: pay close attention to the placement of the rates as a function of n. Each arrow indicates the transition from state n to state m, which can be $n+1$ or $n-1$. On this arrow, the rate λ_n is indicated if $m = n+1$ and μ_n if $m = n-1$. Thus, the rate arrows λ_n and μ_n are not aligned. Also, it is clear from the drawing that μ_0 is not applicable.

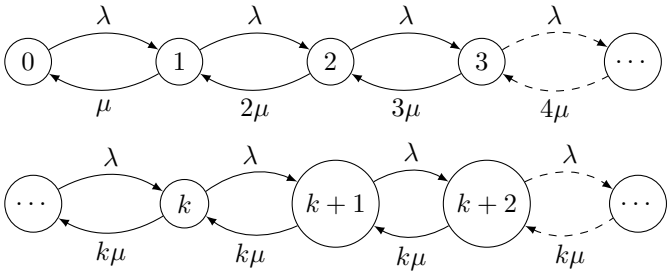
Finally, be careful with another detail. One must take a step back regarding these rates. λ_n and μ_n are arrival rates, in number of people per second. However, just because it is stated at state n that 30 people are arriving every second, it does not mean that we will have $n+30$ people in the queue at the next second. This rate indicates an average and not a deterministic departure. It thus indicates a probability of transitioning from n to $n+1$ between instances t and $t+dt$. Without taking this random aspect into account, one can have a misleading idea of the queue's evolution. For example, just because there are very high rates μ_n and low rates λ_n, it does not mean that the queue will not grow. It is improbable, but it is possible that no one leaves the queue for a certain period of time.

1.3 Particular queues

We can mention two particular queues. The first is the queue where the rates do not depend on n. Thus, we have $\lambda_n = \lambda$ and $\mu_n = \mu$ for all $n \in \mathbb{N}$, where $\lambda, \mu > 0$.



A second recognizable case is that of lines with multiple counters. Each counter means that the number of people leaving the line per second is multiplied by the number of counters open. However, a counter can only process one person in line; a counter only opens if it can process someone. So initially, when there is no one in line, no counter is open. Then one person arrives, we open one counter, then another person arrives, we open another counter, and so on, until all counters are open. The rate increases as counters open until the point where the rate becomes constant. If there are k counters, we have the following representation:



1.4 Name of the model

In the model, the name M/M/k refers respectively to:

- the arrival process, M means Poisson and comes from the fact that we model it as a Markov process.
- the departure process, M also means Poisson
- the number of servers.

Other letters exist such as D for deterministic arrivals and departures, planned in advance, G for general indicating that the distribution is unknown (more precisely it is not specified by the model). These notations are known as Kendall's notations.

2 Stationnary distribution

We want to obtain information about the queue. At the end of this course, we will focus on the distribution law of X as a function of t: we denote $P_n(t)$ the probability

that there are n people in the queue at time t. In other words, $P_n(t) = Pr(X(t) = n)$.

We make two assumptions in this section.

- We assume the queue is empty at time 0.
- We assume that, in all the equations governing $X(t)$, the $o(dt)$ are the same function which we will denote as $g(dt)$ from now on. We thus check that $\frac{g(dt)}{dt} \to 0$ as $dt \to 0$.

Theorem 2.1. *P_n is the solution of the following system of differential equations:*

$$\begin{aligned} P_n'(t) &= \lambda_{n-1} \cdot P_{n-1}(t) + \mu_{n+1} \cdot P_{n+1}(t) \\ &\quad - (\mu_n + \lambda_n) \cdot P_n(t) \ si\ n > 0 \quad (1) \\ P_0'(t) &= \mu_1 \cdot P_1(t) - \lambda_0 \cdot P_0(t) \quad (2) \\ P_0(0) &= 1 \quad (3) \\ P_n(0) &= 0, n > 0 \quad (4) \end{aligned}$$

Proof. The two initial cases correspond to the assumption that the queue is empty at the beginning of the process. To demonstrate the first equation, we will use the definition of the derivative as a rate of change.

Note that the proof here is made for $n > 0$. It must be adjusted for the case $n = 0$.

$$P_n'(t) = \lim_{dt \to 0} \frac{P_n(t + dt) - P_n(t)}{dt}$$

We need to describe $P_n(t+dt)$. The probability of having n people at the time $t + dt$ depends on the number of people at the time t. By the law of total probability:

$$P_n(t + dt) = \sum_{i=0}^{+\infty} Pr(X(t + dt) = n | X(t) = i) P_i(t)$$

Using the equations that govern X

$$\begin{aligned} P_n(t + dt) &= (\lambda_{n-1} \cdot dt + g(dt)) \cdot P_{n-1}(t) \\ &\quad + (\mu_{n+1} \cdot dt + g(dt)) \cdot P_{n+1}(t) \\ &\quad + (1 - (\lambda_n - \mu_n) \cdot dt + g(dt)) \cdot P_n(t) \\ &\quad + \sum_{\substack{i \neq n-1 \\ i \neq n \\ i \neq n+1}} g(dt) P_i(t) \\ P_n(t + dt) &= \lambda_{n-1} \cdot dt \cdot P_{n-1}(t) \\ &\quad + \mu_{n+1} \cdot dt \cdot P_{n+1}(t) \\ &\quad + (1 - (\lambda_n - \mu_n) \cdot dt) \cdot P_n(t) \\ &\quad + g(dt) \sum_{i=0}^{+\infty} P_i(t) \end{aligned}$$

On rappelle que $\sum_{i=0}^{+\infty} P_i(t) = 1$.

$$\begin{aligned} \frac{P_n(t + dt) - P_n(t)}{dt} &= \lambda_{n-1} \cdot P_{n-1}(t) + \mu_{n+1} \cdot P_{n+1}(t) \\ &\quad + (-(\lambda_n - \mu_n)) \cdot P_n(t) \\ &\quad + \frac{g(dt)}{dt} \\ P_n'(t) &= \lambda_{n-1} \cdot P_{n-1}(t) + \mu_{n+1} \cdot P_{n+1}(t) \\ &\quad + (-(\lambda_n - \mu_n)) \cdot P_n(t) \end{aligned}$$

□

We will solve this system in a simple case, called stationnary distribution. The stationnary distribution is a condition of the queue where the probabilities $P_n(t)$ are constant. It may or may not exist and can be reached more or less quickly. Here we ask the question of the existence of this state.

Definition 1. A stationary distribution exists if and only if there exists a set of probability functions $P_n(t)$ that are solutions to equations 1 and 2 of the system in theorem 2.1 such that $P_n'(t) = 0$. We then denote $P_n(t) = P_n$.

Remark 2. It is not because $P_n(t)$ becomes constant after a certain rank t that the number of people in the queue becomes constant. Only the probabilities of having n people in the queue become constant. Observing the queue will show that its size varies randomly over time; however, the distribution managing the number of people in the queue will not depend on time anymore.

Theorem 2.2. *The serie $\sum_{n=1}^{+\infty} \frac{\lambda_0 \cdot \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdots \mu_n}$ is convergent if and only if the stationary distribution exists. In that case, we have*

$$\begin{aligned} P_n &= \frac{\lambda_0 \cdot \lambda_1 \cdot \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdot \mu_3 \cdots \mu_n} \cdot P_0 \\ P_0 &= \frac{1}{1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \cdot \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdots \mu_n}} \end{aligned}$$

Proof. $(P_n)_{n \in \mathbb{N}}$ is a stationnary distribution if and only if

$$\begin{aligned} 0 &= \lambda_{n-1} \cdot P_{n-1} + \mu_{n+1} \cdot P_{n+1} \\ &\quad - (\mu_n + \lambda_n) \cdot P_n \text{ si } n > 0 \\ 0 &= \mu_1 \cdot P_1 - \lambda_0 \cdot P_0 \\ \sum_{i=1}^{+\infty} P_i &= 1 \end{aligned}$$

We can then show that $(P_n)_{n \in \mathbb{N}}$ is a stationnary distribution if and only if

$$P_n = \frac{\lambda_0 \cdot \lambda_1 \cdot \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdot \mu_3 \cdots \mu_n} \cdot P_0$$
$$\sum_{i=1}^{+\infty} P_i = 1$$

As $\sum_{i=1}^{+\infty} P_i = 1$ we can deduce P_0 from P_n for all $n > 1$.

$$P_n = \frac{\lambda_0 \cdot \lambda_1 \cdot \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdot \mu_3 \cdots \mu_n} \cdot P_0$$
$$P_0 = \frac{1}{1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \cdot \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdots \mu_n}}$$

The series $\sum_{n=1}^{+\infty} \frac{\lambda_0 \cdot \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdots \mu_n}$ converges if and only if $P_0 \neq 0$. If the series diverges, then $P_0 = 0$. From this, we deduce that $P_n = 0$ for all n, leading to a contradiction, and thus there is no stationnary distribution. If the series converges, $P_0 \neq 0$ and the values are well-defined, and the stationnary distribution exists. □

Now that we have a characterization of the existence of the stationnary distribution in the general case, we can apply it to simpler cases. For example, if the arrival and departure rates are constant and equal to λ and μ respectively, then a stationnary distribution exists if and only if the geometric series with coefficient λ/μ converges, which is equivalent to $\lambda < \mu$. This conclusion is quite logical since a queue with $\lambda > \mu$ means there is a higher chance that a person arrives than that a person leaves. Therefore, for all i, the probability $P_i(t)$ gradually increases until there are approximately i people in the queue, then it decreases and converges to 0 as more people arrive in the queue over time. Thus, in the case of a stationnary distribution, the only possible solution would be $P_i = 0$ for all i, which is excluded. Conversely, if $\lambda < \mu$, then as soon as one person arrives in the queue, there is a chance that they will leave before another person arrives. The more people enter the queue, the more likely we are to bring the number of people back to 0. We eventually reach a sort of equilibrium and the probabilities converge.

3 Expected size of the queue

Definition 2. The expected number of people in the queue is defined by

$$L(t) = E(X(t)) = \sum_{n=0}^{+\infty} nP_n(t)$$

In a queue with k counters, the expected number of people waiting is defined by

$$L'(t) = E(\max(0, X(t) - k)) = \sum_{n=k}^{+\infty} (n-k)P_n(t)$$

Below is an example of calculating L and L' in a particular case.

Theorem 3.1. *In an $M/M/1$ queue where the arrival and departure rates are constant and equal to λ and μ respectively, with $\lambda < \mu$, then in stationnary distribution, $L = \frac{\lambda}{\mu - \lambda}$ and $L' = \frac{\lambda^2}{\mu(\mu-\lambda)}$.*

Proof. As $\lambda < \mu$ then

$$P_0 = \frac{1}{1 + \sum_{n=1}^{+\infty} \frac{\lambda^n}{\mu^n}}$$

As

$$\sum_{n=0}^{+\infty} \frac{\lambda^n}{\mu^n} = \frac{1}{1 - \frac{\lambda}{\mu}}$$
$$P_0 = 1 - \frac{\lambda}{\mu}$$
$$P_n = \frac{\lambda^n}{\mu^n}\left(1 - \frac{\lambda}{\mu}\right)$$

$$L = \sum_{n=0}^{+\infty} nP_n = \sum_{n=1}^{+\infty} nP_n = \sum_{n=1}^{+\infty} n\frac{\lambda}{\mu}^n \left(1 - \frac{\lambda}{\mu}\right)$$

We perform a change of variable, and then we recognize a standard series.

$$L = \left(\sum_{n=0}^{+\infty} (n+1)\frac{\lambda}{\mu}^n\right) \frac{\lambda}{\mu}\left(1 - \frac{\lambda}{\mu}\right)$$
$$L = \frac{1}{\left(1 - \frac{\lambda}{\mu}\right)^2} \frac{\lambda}{\mu}\left(1 - \frac{\lambda}{\mu}\right)$$
$$L = \frac{\lambda}{\mu - \lambda}$$

For L', the idea is similar.

$$L' = \sum_{n=1}^{+\infty} (n-1)P_n = \sum_{n=2}^{+\infty} (n-1)P_n$$
$$L' = \sum_{n=2}^{+\infty} (n-1)\frac{\lambda}{\mu}^n \left(1 - \frac{\lambda}{\mu}\right)$$
$$L' = \sum_{n=0}^{+\infty} (n+1)\frac{\lambda}{\mu}^n \frac{\lambda}{\mu}^2 \left(1 - \frac{\lambda}{\mu}\right)$$
$$L' = \frac{\lambda}{\mu} L$$
$$L' = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

□