Dr. Nawel ARRAR REMITA

November 2, 2023

Continuous-Time Markov Chains

DTMCs are totaly synchronized, in that the state only changes at discrete time steps, whereas in CTMCs the state can change at any time. This makes CTMCs more realistic for modeling computer systems, where events can occur at any time. In preparation for CTMCs, we need to dicuss the exponential distribution and Poisson arrival process.

Let's consider the residence time T_i at state i in the context of a continuous-time process $\{X_t\}_{t \geq 0}$.

By hypothesis, we have the following property: The residence time T_i is a continuous random variable with truly positive values, thus prohibiting instantaneous visits. In fact, it follows an exponential law with parameter $0 \leq \lambda_i < \infty$, with $\lambda_i = 0$ when the i is an absorbing state.

Exponential Distribution

Definition

We say that a random variable X is distributed exponentially with rate $\lambda, X \rightsquigarrow \mathcal{E}\left(\lambda\right)$, if X has the probability density function:

$$f\left(x\right) = \begin{cases} \lambda e^{-\lambda t}, t \geq 0 \\ 0, t < 0 \end{cases}$$

and the cummulative distribution function,
$F\left(x\right) = \left(1 - e^{-\lambda t}\right) \mathbf{1}_{t \geq 0}$. So $\overline{F}\left(x\right) = e^{-\lambda t}, t \geq 0$.

In addition,

$$\begin{aligned} \mathbb{E}\left[X\right] &= \frac{1}{\lambda}, \mathbb{E}\left[X^2\right] = \frac{2}{\lambda^2}, Var\left(X\right) = \frac{1}{\lambda^2} \\ \mathbb{E}\left[e^{-tX}\right] &= \frac{\lambda}{\lambda - t} \text{ for } t < \lambda. \end{aligned}$$

Exponential Distribution

Remark

And the square coefficient of variation of random varaible X is defined as

$$CV_X^2 = \frac{Var\left(X\right)}{\mathbb{E}\left[X\right]^2}$$

This can be thought of as the scaled or normalized variance. When $X \rightsquigarrow \mathcal{E}\left(\lambda\right)$, $CV_X^2 = 1$. A random variable X is said to be memoryless if

$$\mathbb{P}\left(X > s + t/X > s\right) = \mathbb{P}\left(X > t\right), \forall s, t \geq 0.$$

Exponential Distribution

Proposition

Given $X_1, X_2, ..., X_n$ *independant r.v of exponential law with respective parameters* $\lambda_1, \lambda_2, ..., \lambda_n$. *We define* $X = \min\left(X_1, X_2, ..., X_n\right)$.

The variables $X_1, X_2, ..., X_n$ represent times before different events occur, and then X is the time until the first of these events occurs. We have the following properties:

Exponential Distribution

Theorem

Given $X_1, X_2, ..., X_n$ *i.i.d of* $\mathcal{E}\left(\lambda_i\right), 1 \leq i \leq n$. *Let* $X = \min\left(X_1, X_2, ..., X_n\right)$. *Then*

a)

$$X \rightsquigarrow \mathcal{E}\left(\lambda_1 + \lambda_2 + ... + \lambda_n\right);$$

b)

$$\mathbb{P}\left(X \leq X_i\right) = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n}, \forall i = 1, 2, ..., n;$$

c) *The event* $X \leq X_i$ *is independent of the r.v* X.

Processus de Poisson

A Poisson process is a continuous-time Markov chain $\{X_t, t \geq 0\}$, which counts the number of arrivals of an event that occur randomly over time at a constant rate, for example customers or accidents. The *intensity* of a Poisson process is the instantaneous rate $\lambda > 0$ with which arrivals occur.
A flow of random events can be described mathematically in two different ways:

1. The number of events X_t occurring in $[0, t]$ and seek to determine the probability law of this discrete random variable. The process $\{X_t, t \geq 0\}$, is called the "*counting process*".
2. The times intervals that separate the instants of occurrence of two consecutive events is called "*inter-arrival time*". These are independent and identically distributed r.v with law $\mathcal{E}(\lambda)$. This process is called the "*birth process*".

Processus de Poisson

Let's recall the important properties of the Poisson process:
- it is homogeneous, which means that the rate of appearance (birth) λ is independent of time ;
- it is a law with "*independent increases*": the numbers of events occurring X_1 and X_2 over two disjoint time intervals T_1 and T_2, are independent random variables.
- the probability of "simultaneous" occurrence (i.e. between t and $t + \Delta t$) of several events (birth) is negligible: $o\Delta t$.

Processus de Poisson

Definition

A Poisson process having rate λ is a sequence of events such that

$$\begin{aligned} \mathbb{P}\left(X\left(t\right)=k\right) &= p_k\left(t\right)=\frac{\left(\lambda t\right)^k}{k!}e^{-\lambda t}, \lambda>0, k\geq 0; \\ E\left[X\left(t\right)\right] &= \lambda t \text{ et } Var\left[X\left(t\right)\right]=\lambda t. \end{aligned}$$

Definition

These relations define the transitional state of the Poisson process. No stationary state exists, since $p_k=\lim_{t\to\infty}p_k\left(t\right)=0, \forall k\geq 0.$

Processus de Poisson

Definition

The interval time time $T_{i,}(i \geq 1)$ that separates any instant from the next event is a random variable distributed according to an $\mathcal{E}(\lambda)$ law.

Particular Markov Processes

Birth and Death Processes

The processes in question can generally be used to write the temporal evolution of the size of a population of a given type. They are widely used to model waiting phenomena or systems subject to repairable failures. They are obtained by superimposing a birth process and a death process. This Markov process X_t represents the size of a population at time t. These are stochastic processes with continuous time and discrete states space ($S = \{0, 1, 2, ...\}$). They are characterized by two important conditions:

- Memoryless;
- transitions are only possible to one or other of the neighboring states,from state n the possible transitions are $n-1$ and $n+1$ with $n \geq 1$.

Birth and Death Process

Let $X_t = N(t)$ be the size of a population at time t (number of individuals present). Define

$$p_n\left(t\right) = \mathbb{P}\left(N\left(t\right) = n\right)\left(\text{during } \left[0, t\right]\right);$$

and $p_{ij}(t)$ be the probability that at the moment t the number of individuals is j such that there were already i individuals in the population. We have that

$p_{i,j}(t) = \mathbb{P}(N(t+s) = j / N(s) = i)$ does not depend on s (the process is homogeneous), so

$$\begin{aligned} p_{i,i+1}\left(\Delta t\right) &= \lambda_i \Delta t + o\Delta t, \\ p_{i,i-1}\left(\Delta t\right) &= \mu_i \Delta t + o\Delta t, \\ p_{i,i}\left(\Delta t\right) &= 1 - (\lambda_i + \mu_i)\Delta t + o\Delta t, \\ p_{i,j}(\Delta t) &= o\Delta t \text{ if } \left| i - j \right| \geq 2 \text{ and } p_{i,j}(0) = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \end{aligned}$$

with $\lambda_i > 0, \mu_i > 0$ and $\mu_0 = 0$.

Birth and Death Process

Transitory State

Let the transitions probabilities (or states probabilities) $p_n(t) = \mathbb{P}(N(t) = n), \forall n \geq 0$ during the interval time $[(0, t]$ The transitory state is described by

$$\mathbb{P}(t + \ \Delta t) = \mathbb{P}(t) \times M$$

where M is transition matrix. Letting $\Delta t \longrightarrow 0$, we obtain the following **Kolmogorov equations** system:

$$\left\{ \begin{array}{c} p_0'(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t) \\ p_1'(t) = \lambda_0 p_0(t) - (\lambda_1 + \mu_1) p_1(t) + \mu_2 p_2(t) \\ - - - - - - - - - - - - \\ p_n'(t) = \lambda_{n-1} p_{n-1}(t) - (\lambda_n + \mu_n) p_n(t) + \mu_{n+1} p_{n+1}(t), n \geq 1. \end{array} \right. \tag{1}$$

Birth and Death Process

Stationory state

In state of equilibrium, also know as steady state, the behavior of the process is independent of the time and of the initial state; i,e, $p_n = \lim_{t \to \infty} p_n(t)$, $n = 0, 1, 2 ...$which is the stationary distribution of the process under study. Therefore, $\lim_{t \to \infty} p'_n(t) = 0$.

Then, these probabilities satisfy the following **balance equations** *or* **statistical equilibrium system** (obtained from (1) by taking):

$$\left\{ \begin{array}{c} \lambda_0 p_0 = \mu_1 p_1 \\ (\lambda_1 + \mu_1) p_1 = \lambda_0 p_0 + \mu_2 p_2 \\ - - - - - - - - - - - - \\ (\lambda_n + \mu_n) p_n = \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}, n \geq 1. \end{array} \right. \tag{2}$$

To solve this system, we add $\sum_{n=0}^{\infty} p_n = 1$.

Simple Markovian Queueing Systems

The formation of queues is a frequent phenomenon that occurs whenever service demands exceed the permitted capacity of service devices.

The general model of a queueing phenomenon (waiting system) can be summarized as follows: requests (customers) arrive at a certain location and demand a certain service. If a service device is free, the incoming request immediately goes to this device, where it is served. Otherwise, there are two possibilities: either the request leaves the system (denied request systems), or it goes into a queue (queuing system). At some point, the request is selected for service according to a discipline. On completion of the service, the request leaves the system.

A queuing system therefore comprises a service area with one or more service devices (servers), and a waiting area in which a possible queue is formed.

Simple Markovian Queueing Systems

Classification of queueing systems

To identify a queuing system, we need to specify the entry rate, the service mechanism and the waiting discipline.

- The size of a potential customer population (or the number of sources) can be finite or infinite.
- The process (flow) of arrivals can be regular or random. The time between two successive arrivals follows a probability law.
- The service mechanism includes the number of servers and the distribution service times.
- Queuing discipline or service discipline, customers can be selected and served in order of arrival **FCFS**, or **LCFS**, or selected at random **RANDOM**.
- Waiting space capacity can be unlimited or not.

Simple Markovian Queueing Systems

Classification of queueing systems

We use a symbolic notation "**Kendall notation**" comprising 4 symbols arranged in the order $A/B/s/m$, where A and B describe respectively the distribution of times between two successive arrivals and the distribution of service times, s is the number of servers (connected in parallel), m is the system capacity (the number of servers plus the number of waiting positions). The last symbol can be omitted if $m = \infty$.

To specify distributions A and B, we introduce the following symbols:

M Exponential distribution ;

E_k Erlang distribution of degree k ;

H_k Hyperexponential distribution of degree k ;

D Deterministic ;

G General distribution.

Simple Markovian Queueing Systems

Mathematical analysis

The mathematical study of a queuing system involves the introduction of an appropriately defined stochastic process. The stationary distribution of the introduced stochastic process allows us to obtain the system's performance characteristics, such as such as:
- the mean number of customers in system $\overline{n}$;
- the mean number of customers in the queue $\overline{n_q}$;
- the mean waiting time for a customer $\overline{W}$;
- the mean amount of time a customer spends in the system $\overline{Ws}$.

Simple Markovian Queueing Systems

Mathematical analysis

One of a more important and useful relationship in queueing theory is what is communly known as **Little's Law:**

$$\overline{n} = \lambda \overline{Ws}, \overline{n_q} = \lambda \overline{W},$$

and

$$\overline{n} = \overline{n_q} + \frac{\lambda}{\mu}, \overline{Ws} = \overline{W} + \frac{1}{\mu},$$

where λ is the rate at which customers enter the system, $\frac{1}{\mu}$ is the mean service time $(\mu > 0)$. Another important measure for a queueing system, the one that measures the degree of saturation of the system, is the traffic intensity ρ. It is defined by

$$\rho = \frac{\text{mean service time}}{\text{mean time between two successive arrivals}}.$$

Queueing Model M/M/1

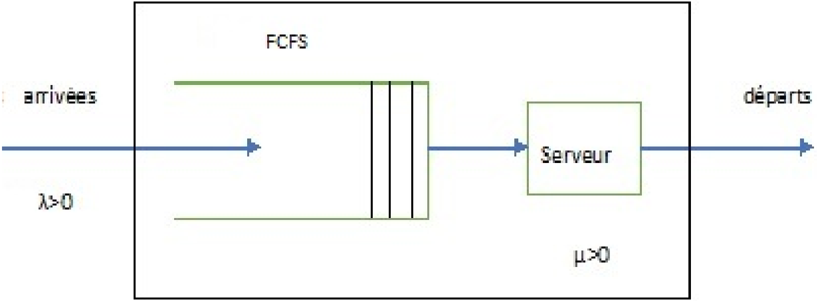
Model Description

Customers arrive at the system according to a Poisson process of rate $\lambda > 0$, (mean number of customers arriving during a unit of time) ; i.e., the time interval between two successive arrivals follows an exponential law with parameter $\lambda > 0$. The service is provided by a single server. When a customer arrives, if the server is free, it is immediately taken over. Otherwise, the customer is put on a queue. Queuing capacity is unlimited (the number of positions is infinite and no other restrictions are imposed). Waiting discipline is FCFS. Service times follow an exponential distribution with parameter $\mu > 0$. Consequently, the service rate is μ (mean number of customers served during a unit of time), and the mean time of service for a customer is $\frac{1}{\mu}$.

Queueing Model M/M/1

Model Description

Finaly, we assume that the r.v representing the times between two consecutive arrivals and the service times are mutually independent.



Queueing Model M/M/1

Model analysis

The state of the system at date t can be described by the stochastic process

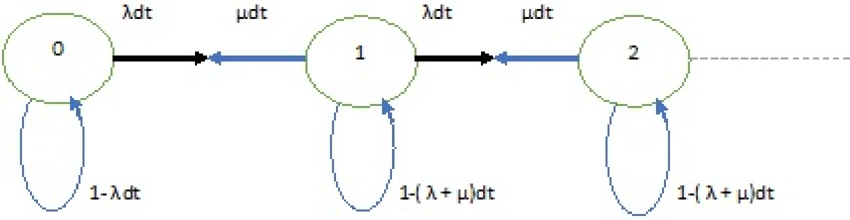
$$\left\{ N\left(t\right) ,t\geq 0\right\} , \tag{3}$$

which is a birth and death process and states space $S = \{0, 1, 2, ...\}$. The rates of transitions are $\lambda_n = \lambda, \forall n \geq 0$ and $\mu_n = \mu$, for $n \geq 1, (\mu_0 = 0)$.

Queueing Model M/M/1

Transitory State

Let $p_n(t) = \mathbb{P}(N(t) = n), \forall n \geq 0$ be the probability of observing n customers in the system at time t. The representative transition graph of an M/M/1 queue is as follows



Queueing Model M/M/1

For the limiting probabilities $\lim_{t\to\infty} P_n(t) = p_n$, the state balance equations are obtained. Hence, we get

$$p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\,\rho^n, n \geq 0,$$

where $\rho = \frac{\lambda}{\mu} < 1$.

Queueing Model M/M/1

Mesures de performance

- Mean number of customers in the system $\overline{n} = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{(1-\rho)}$.
- Mean number of customers in the queue $\overline{n_q} = \frac{\lambda^2}{\mu(\mu-\lambda)}$.
- Mean time a customer spends in the system $\overline{W_s} = \frac{1}{\mu-\lambda}$.
- Mean waiting time for a customer $\overline{W} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$.

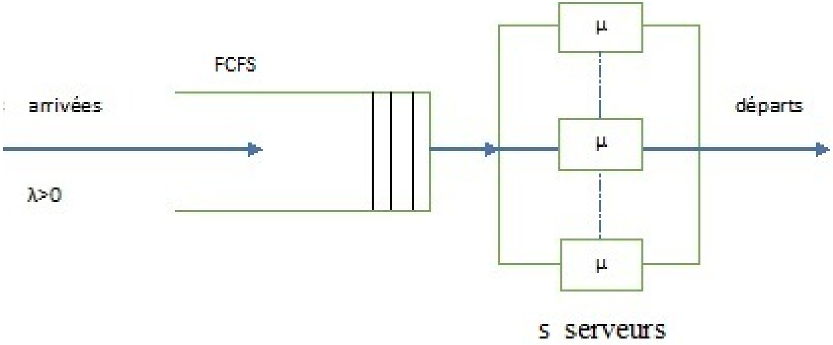
Queueing Model M/M/s

Model description

Customers arrive at the system according to a Poisson process of rate $\lambda > 0$. Service is provided by $s \geq 1$ servers connected in parallel. When a customer arrives, if one of the servers is free, the customer immediately begins service. In the opposite case (all servers are occupied by the service), the customer takes his place in the queue, common to all servers. Waiting capacity is unlimited. When a server is free, the customer at the head of the queue occupies the freed-up server. Consequently, the waiting discipline is FCFS. Service times are exponentially distributed with finite mean $1/\mu$. Times between two consecutive arrivals and service times are mutually independent.

Queueing Model M/M/s

Model description



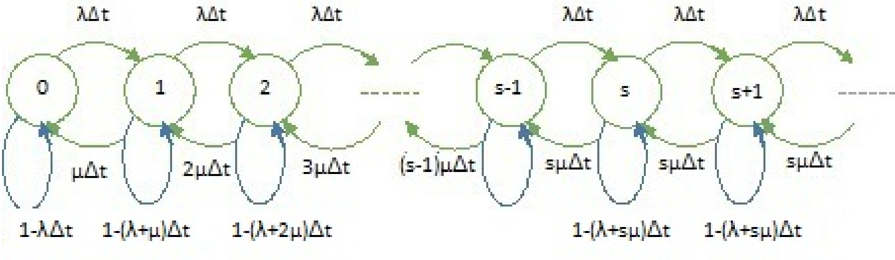
Queueing Model M/M/s

Analyse du modèle

The state of the system at date t can be described using the process (3), whose state space $S = \{0, 1, 2, ...\}$. The latter is a birth and death process whose transition rates are:

$$\lambda_n = \lambda, \forall n \geq 0 \text{ and } \mu_n = \min\{n, s\} \times \mu, n \geq 1 (\mu_0 = 0).$$

Let the state probabilities $p_n(t) = \mathbb{P}(N(t) = n), \forall n \geq 0$. The representative graphe of transitions



Queueing Model M/M/s

Stationary state

As before, for the limit probabilities, from the transition graph and simplification, we obtain :

$$\begin{cases} p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, 1 \leq n < s; \\ p_n = \frac{1}{s!}\frac{1}{s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n p_0, n \geq s. \end{cases}$$

where,

$$p_0 = \left[\sum_{n=0}^{s-1}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!\left(1-\frac{\lambda}{s\mu}\right)}\right]^{-1}.$$

Queueing Model M/M/s

Stationary state

We get

$$
\begin{aligned}
p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, 1 \leq n < s \\
p_n &= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \rho^{n-s} p_0, n \geq s
\end{aligned}
$$

provided $\frac{\lambda}{s\mu} = \rho < 1$.
ρ is the total traffic intensity .

Queueing Model M/M/s

Stationary state

We can write

$$p_n = \rho^{n-s} p_s, n \geq s$$

The probabilité p_s that the customer will have to wait for service if only the number of customers $n \geq s$, called **Erlang formula**, it is given by

$$p_s = \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!\left(1 - \frac{\lambda}{s\mu}\right)} p_0 = \frac{(s\rho)^s}{s!\,(1-\rho)} p_0.$$

Queueing Model M/M/s

Mesures de performance

- Mean number of customers in the system

$$\overline{n} = \frac{\lambda}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^{s+1}}{ss!\left(1-\frac{\lambda}{s\mu}\right)^2} p_0 = s\rho + \frac{\rho p_s}{(1-\rho)}.$$

- Mean number of customers in the queue

$$\overline{n_q} = \frac{\left(\frac{\lambda}{\mu}\right)^{s+1}}{ss!\left(1-\frac{\lambda}{s\mu}\right)^2} p_0 = \frac{\rho p_s}{(1-\rho)}.$$

- Mean time a customer spends in the system

$$\overline{W_s} = \frac{1}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^{s}}{s\mu s!\left(1-\frac{\lambda}{s\mu}\right)^2} p_0 = \frac{1}{\mu} + \frac{p_s}{s\mu(1-\rho)}.$$

- Mean waiting time for a customer

$$\overline{W} = \frac{\left(\frac{\lambda}{\mu}\right)^{s}}{s!s\mu(1-\rho)^2} p_0 = \frac{p_s}{s\mu(1-\rho)}.$$