Introduction to Stochastic Networks

Draft October 2010

Matthieu Jonckheere
matthieu.jonckheere@gmail.com

Contents

Chapter 1

Introduction to Markov jump processes

This chapter consists only of a short and quite informal summary of the most basic definitions and results for Markov Jump processes that are going to be needed in the following. We often refer to [6] for more (rigorous) details.

We work in this chapter with continuous time stochastic process, i.e. a collection of random variables $(X_t)_{t\in T}$, where T is an interval of $\mathbb{R}^+$ including 0.

To speak about stochastic processes, one needs to appropriately define a probability space and a sigma-algebra. We do not describe this construction in details and refer to [6] for a rigorous introduction. We shall just mention that we generally assume (except mentioned otherwise) the stochastic processes treated in the sequel to be S-valued ($S \subset \mathbb{R}_+^N$, $N \subset \mathbb{N}$) and to have paths in the space $D = D(T, S)$ of right-continuous functions from T to S with finite left limits (càdlàg). Except if explicitly mentioned, N will be finite. A stochastic process with paths in D can then be viewed as a random element on the measurable space $(D, \mathcal{D})$, where $\mathcal{D}$ denotes the Borel σ-algebra generated by the standard Skorokhod topology (again we refer to [6] for more details). The filtration $(\mathcal{F}_t)$ is a $\sigma-$field on the probability space, usually representing the history of a process until time t. The smallest filtration with this property is $\sigma\{X_s, s \leq t\}$ and is called the induced filtration. We say that a process is adapted to a filtration $\mathcal{F} = (\mathcal{F}_t)$ if X_t is $\mathcal{F}_t$-measurable for every t. Finally, we call a random time τ a stopping time if $\{\tau \leq t\} \in \mathcal{F}_t$ for each time t. We now define the notion of Markov process and Markov jump processes.

1.1 Markov processes and Markov jump processes

A process is said to be Markov when the law of its position at time $t + s$, conditionally on its position at time s, does not depend on its history until s, but only on its position at time s. More formally,

Definition 1 *An adapted process X is said to be Markov if for any times $s < t$, $X_t =$*

$f_{s,t}(X_s, \mathcal{U}_{s,t})$ a.s. for some measurable function $f_{s,t}$ and a uniform random variable $\mathcal{U}_{s,t}$ independent of $\mathcal{F}_s$.

In the latter, it will be convenient to introduce a shift operator $\theta_t : D \to D$,

$$\theta_t \omega_s = \omega_{s+t}, \forall s, t.$$

The notation P^x will be used to indicate the law of the process conditionally on its initial position being x. Assuming that the existence of transition kernels $P_{s,t}(X_s, B) = P(X_t \in B | X_s)$ and that they define correctly and completely a Markov process (which is demonstrated by Kolmogorov theorem, see Theorem 8.4 in [6]), the Markov property can also be defined as:

Definition 2 *A process $X(t)$ is Markov if for all $t, s \geq 0$ and x, P^x-almost surely:*

$$E^x(f(X(t+s))|\mathcal{F}_t) = E^{X_t}(f(X(t+s))),$$

for every measurable and bounded function or equivalently:

$$P^x(X(t+s) \in \cdot | \mathcal{F}_t) = P^{X(t)}(X(t+s) \in \cdot).$$

The process is time-homogeneous if $P_{s,t} = P_{t-s}$, in which case P_t is a semi-group satisfying the following Chapman-Kolmogorov equations:

$$P_{t+s} = P_t P_s, \forall s, t.$$

In particular, we have:

$$P^x(X(t+s) \in \cdot | \mathcal{F}_t) = P^{X(t)}(X(s) \in \cdot).$$

We now turn our attention to pure jump processes.

Definition 3 *We shall say that a process is a Markov jump process (MJP) (or a pure-jump Markov process) if it is a (cadlag) Markov process taking values in a subset of $\mathbb{N}^N$*

Remark 1 *This definition implies that the jump times of the process are isolated. However, remark that it does not preclude explosion in finite time.*

We shall moreover always make the assumption that almost surely, there is a finite number of jumps in a finite interval of time. This is verified for instance if the transitions defined below are bounded.

One can further prove that a MJP satisfies the strong Markov property, i.e., replacing the fixed time by a stopping time τ in the Markov property:

$$P(X(\tau + t) \in A | \mathcal{F}_\tau) = P^{X(\tau)}(X_t \in A).$$

Our next aim is to define the rates (infinitesimal generator) of a MJP and as well as its jump transitions. A crucial step consists in describing the distribution of the times between jumps.

Lemma 1 *If a state x is not absorbing, then under P^x, the time τ until the first jump is exponentially distributed and independent of X_τ.*

Proof: Using the Markov property,

$$P^x(\tau > t+s) = P^x(\tau > s, \tau \circ \theta_s > t) = P^x(\tau > s)P(\tau > t).$$

The only non-increasing solutions of these equations are of the form $t \to e^{-\alpha t}$ for a non-negative α. Since x is non-absorbing, $\alpha > 0$ and since we suppose the process to have a finite number of jumps in finite intervals, $\alpha < \infty$.

$$P^x(\tau > t, X_\tau \in B) = P^x(\tau > t, X_\tau \circ \theta_t \in B) = P^x(\tau > t)P^x(X_\tau \ \in B).$$

□

We now define the rate function and the transitions kernel as:

$$r(x) = E^x[\tau]^{-1}.$$

$$p(x, B) = P^x(X(\tau) \in B).$$

The transitions rates are defined as:

$$q(x, y) = r(x)p(x, y).$$

We can explain better the structure of MJP by giving a generic construction in terms of sojourn times and probability of jumping from one state to another which is often easier to recognize. Note that proving the Markov property directly for arbitrary stochastic jump processes is not always an easy task in applications while the following characterization is often simpler to recognize.

Proposition 1 *A process is a Markov jump process (MJP) if and only if it has the following dynamics:*

- *The process stays in a given state x, a random time S_x (called sojourn time), having an exponential distribution with mean $r(x)^{-1}$, with $r(x) = \sum_y q(x, y)$.*
- *Then, it jumps to another state $y \neq x$ with probability*

$$p(x, y) = \frac{q(x, y)}{r(x)}.$$

Proof: We can argue recursively using the previous lemma to prove that the dynamics of a MJP are determined by a sequence of i.i.d exponential random variables and from the transitions of a discrete time Markov chain. For the converse statement, one basically needs to prove the Markov property. See [6] Theorem 12.17. □

1.1.1 Examples and exercises

We illustrate the dynamics explained below on a few examples.

Example 2 (Poisson process) *The simplest example of a MJP is the Poisson process which is a MJP living on $S = \mathbb{Z}^+$ and having non-zeros transitions:*

$$q(x, x+1) = \lambda.$$

$r(x) = \lambda \in \mathbb{R}^+$ is the intensity of the process. Many specific properties of the Poisson process could be stated. In particular, it is the unique point process (i.e., a counting process) being a MJP, with stationary independent increments. It can be defined (independently of the theory of MJP presented so far) at least in three ways:

- *Via sojourn times :*

  define a sequence of i.i.d exponential random variables E_i and $S_n = \sum_{i=1}^{n} E_i$. Then define the Poisson process N_t with $N_0 = 0$ and $N_t = \sup\{n : S_n < t\}$.

- *Via the number of occurrences:*

  For each time t, the number of occurrences N_t in $[0, t]$ has a Poisson distribution. The position of each occurrence is then chosen independently and uniformly distributed in $[0, t]$.

- *Via differential equations:*

  Let $p_i(h)$ the probability to have i occurrences at time h. Assume the p_i differentiable in 0 and such that $p_1'(0) = \lambda$, $p_{\geq 2}'(0) = 0$,and $p_0(t+h) = p_0(t)p_0(h)$.

Example 3 (M/M/1 queue) *Consider a one-server queue with Poisson arrivals (parameter λ) and exponential service times (with parameter μ) and a FIFO (first in first out) scheduling. Let X_t the number of customers in the queue. Given that at time 0, there are x customers in the queue, using the memory-less property of the exponential distribution, the next jump is occurring after an exponential random variable of parameter λ (the next arrival) if $x = 0$ and the minimum of an exponential random variable of parameter λ (the next arrival) and an exponential random variable of parameter μ (the next departure) if $x > 0$. Using now that the minimum of two exponential variables is exponential, we have proved that the times between jumps is exponentially distributed and that the rate function is:*

$$q(0) = \lambda, \tag{1.1}$$
$$q(x) = \lambda + \mu, \ x > 0. \tag{1.2}$$

Now, the probability that the next jump is an arrival is 1 if $x = 0$ and $\frac{\lambda}{\lambda+\mu}$ if $x > 0$ by looking at which of the two competing exponential clocks terminates first. We have hence proved that the process X_t is a MJP using proposition 1.

Example 4 (The Erlang system) *Consider a trunk of N telephone lines. Calls arrive according to a Poisson process and have an exponential duration with parameter μ. The number of calls in progress is a MJP with non-zeros transitions:*

$$q(x, x+1) = \lambda 1_{x<N}, \tag{1.3}$$
$$q(x, x-1) = \mu x, \tag{1.4}$$

for all x.

Exercise 1 *Show that the transitions rates $q(x, y)$ may be defined by:*

$$q(x, y) = \lim_{t \to 0} \frac{1}{t} P(X_t = y | X_0 = x), \ x, y \in \mathcal{X}, \tag{1.5}$$
$$q(x, x) = 0. \tag{1.6}$$

Exercise 2 *Prove that the sequences of jump times diverge to infinity iff $\sum_n r(Y_n)^{-1} = \infty$, where Y_n is a Markov chain on S.*

Exercise 3 *Construct a MJP when the assumption of a finite number of jumps in a finite interval is not verified.*

Exercise 4 *Let X a stochastic process on $\mathbb{N}^d$.The sequence of states it visits is a Markov chain with transitions probabilities $\bar{p}(x, y)$ (possibly with $\bar{p}(x, x) > 0$), The times between two transitions is exponentially distributed with rate $\lambda(x)$. Show that X is a MJP and give its transitions.*

Exercise 5 *Consider a Poisson process N_t such that $N_0 = 0$ and E_i the sequence of times between jumps. Let $S_n = \sum_{i=1}^n E_i$. What is the distribution of $U = S_{N_t+1} - S_{N_t}$.*

1.2 Invariant and stationary measure:

We denote in the sequel $X = \{X_t, t \geq 0\}$ a continuous time Markov jump process (MJP) with state space $\mathcal{X} \subset \mathbb{N}^d$ associated with some transition rates $q(.,.)$. We denote $p_{x,y}(t) = P^x(X_t = y)$. We suppose in the sequel that X is irreducible i.e. $p_{x,y}(t) > 0$ for all times $t > 0$ and all states x, y, which boils down to having a positive probability to reach any state from any state in a finite number of jumps.

An invariant measure of the MJP X is a positive measure on $\mathcal{X}$ verifying the global balance equations:

$$\pi(x) \sum_y q(x, y) = \sum_y \pi(y) q(y, x). \tag{1.7}$$

Remark 5 *The property of aperiodicity is irrelevant in continuous time but plays a crucial role in discrete time. Several discrete versions of a continuous time process can be constructed. However, even if the continuous time process has a unique stationary distribution, there is no guarantee in general that a discrete time version of the process is aperiodic with a unique stationary distribution.*

1.2.1 Recurrence:

- If π has a finite l^1 norm, i.e., $\sum_x \pi(x) < \infty$, then the process is said to be **positive recurrent**. The condition of irreducibility ensures the uniqueness, up to a multiplicative constant, of the invariant measure. In particular, π can be normalized such as to become a probability distribution on $\mathcal{X}$. We speak in this case of the **stationary distribution** of X and we say (it will become clear why soon) that X is ergodic. π can also alternatively be called the equilibrium distribution or the limiting distribution since

$$\lim_{t\to\infty} P^\nu(X_t = x) = \pi(x), \forall x, \forall \text{initial distribution } \nu,$$

  for all distribution ν of the initial position X_0.

- If π has an infinite l^1 norm, the process is said to be **transient** if $|X_t| \to \infty$, almost surely and **null recurrent** otherwise.

1.2.2 Ergodicity

Consider a MJP X with stationary distribution π.

Theorem 6 *With probability* 1 *(a.s.),*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X_s)ds \to \sum_x f(x)\pi(x).$$

The proof follows essentially from the law of large numbers for i.i.d sequences. (See [6].)

1.3 Martingales for jump Markov processes and (backward) Kolmogorov equation

We suppose Martingale theory known. An appendix at the end of the chapter recalls the definition and a few fundamental results.

Define a Markov process X_t with bounded transitions q. Define δ the drift operator of the process such that for all bounded function f:

$$\delta f(x) = \sum_y q(x,y)(f(y) - f(x)).$$

Proposition 2 *For every bounded function f, The process M_t defined by:*

$$M_t^f = f(X_t) - f(x) - \int_0^t \delta f(X_s)ds,$$

is a martingale.

Corollary 1 (Kolmogorov equations)

$$E^x[f(X_t)] = f(x) + \int_0^t E^x[\delta f(X_s)]ds. \tag{1.8}$$

This last equation can be obtained directly using the strong Markov property (and without using martingales) as follows. Recall that τ_1 is the time of the first jump. Let $\sigma = \tau_1 \wedge t$. Using the strong Markov property (and property of the conditional expectation):

$$\begin{aligned}
E^x[f(X_t)] &= E^x[f\left((X \circ \theta_\sigma)_{t-\sigma}\right], & (1.9)\\
&= E^x[E^{X_\sigma}[f(X_{t-\sigma})]], & (1.10)\\
&= f(x)P(\tau_1 > t) + E^x\left[E^{X_{\tau_1}}[f(X_{t-\tau_1}); \tau_1 \leq t]\right], & (1.11)\\
&= f(x)e^{-r(x)t} + \int_0^t e^{-sr(x)} \sum_y E^y[f(X(t-s))]q(x,y)ds. & (1.12)
\end{aligned}$$

Hence

$$e^{r(x)t}E^x[f(X_t)] = f(x) + \int_0^t e^{sr(x)} \int q(x,dy)E^y[f(X_s)]ds,$$

which leads to (1.8) by direct computations.

1.4 Discrete time transformation - Uniformization

1.4.1 Embedded chain

Let $(T_n)_{n\in\mathcal{N}}$ the sequence of times at which the process experiences a jump. The embedded chain is the discrete version corresponding to:

$$X_n^e = X_{T_n}.$$

Exercise 6 *Construct an example showing that the stationary measure of X_n^e is generally different from the stationary measure of X_t.*

We nevertheless have the following Proposition.

Proposition 3 *π is invariant for X iff $q(\cdot)\pi(\cdot)$ is invariant for X^e.*

Proof: π is invariant iff

$$\sum_y \pi(y)q(y,x) = \sum_y q(x,y)\pi(x),$$

which gives that:

$$\sum_y \pi(y)q(y)p(y,x) = q(x)\pi(x).$$

□

1.4.2 Uniformization

If the transitions q are bounded, then an easy way to construct a discrete-time analogue of the continuous time MJP is to sample the continuous process along the points of a Poisson process with a sufficiently big intensity γ. Let $\gamma \geq \sup_x \sum_y q(x,y)$.

Define the discrete time Markov chain X^d with the following transitions:

$$q^d(x,y) = \frac{q(x,y)}{\gamma}, \tag{1.13}$$

$$q^d(x,x) = 1 - \sum_x q^d(x,y). \tag{1.14}$$

Then, we can see (the continuous-time) MJP as the composition of a Markov chain and a Poisson process. (Hence, it is sometimes called a pseudo-Poisson process.)

Proposition 4 *A MJP X has bounded rates iff $X = X^d \circ N$ with X^d the Markov chain previously defined and N a Poisson process with intensity γ.*

Proof: If $X = X^d \circ N$, then using that the times between the jumps of the Poisson process are exponentially distributed and the probability of jumping to a given state depends only on the current state as X^d is Markov. Using Proposition 1, we have that X is a MJP. The time between two jumps is exponentially distributed with fixed parameter γ and hence the process has bounded transitions rates.

Conversely, assume X is a MJP with bounded transitions. Define the chain X^d using equations (1.13) and let N a Poisson process with parameters γ. Let $X' = X^d \circ N$. X and X' have clearly the same transitions rates, which implies (see for instance Cor. 6.11 in [6]) that the two processes have the same distribution. □

The following Proposition makes the uniformization procedure helpful (and simpler than consider the embedded chain) for many applications.

Proposition 5 *The discrete time Markov chain X^d with transitions q^d and the continuous time Markov process X_t with transitions q have the same invariant measures.*

Proof: The invariant measure of X^d, π^d satisfies:

$$\pi^d \frac{q}{\gamma} + \pi(1 - \frac{\sum_y q(\cdot, y)}{\gamma}) = \pi^d,$$

hence

$$\pi^d(x) \sum_x q(x,y) = \sum_y q(x,y)\pi^d(y).$$

□

1.5 Coupling and strong ordering

In this Section, we briefly describe the basic tools to compare (in a strong sense) two MJP. Suppose that we give ourself two stochastic processes X_t and Y_t and that we want to compare (in some sense to be made precise) their laws. A very useful technique in the theory of probability consists in constructing a (super) process $Z_t = (\tilde{X}_t, \tilde{Y}_t)$ having marginals $\tilde{X}_t$ and $\tilde{Y}_t$ equal in distribution to (respectively) X_t and Y_t and such that the trajectories (sample paths) of Z_t gives information on the comparisons of the laws of X_t and Y_t.

We begin by defining the notion of stochastic ordering. We first give ourselves a partial order $\prec$, and we say that a function $f : S \to \mathbb{R}$ is increasing if:

$$x \prec y \Rightarrow f(x) \leq f(y).$$

Definition 4 *We say that two random variables are ordered in the strong sense (denoted $X \leq_{st} Y$), if $E[f(X)] \leq E[f(Y)]$ for all functions f increasing.*

This definition extends to stochastic processes:

Definition 5 *We say that two stochastic processes are ordered in the strong sense (denoted $X[x] \leq_{st} Y[y]$ or $X^x \leq_{st} Y^y$), if for all initial conditions $x \prec y$, and for all times times t:*

$$E[f(X_t^x)] \leq E[f(Y_t^y)],$$

for all functions f increasing.

We now introduction the definition of a coupling between random variables and processes.

Definition 6 *We say that $Z = (U, V)$ is a coupling of the random variables X and Y (possibly defined on different probability spaces) if the r.v. X and U as well as Y and V have the same distribution.*

Simirlarly, we say that $Z_t = (U_t, V_t)$ is a coupling of the processes X_t and Y_t (possibly defined on different probability spaces) if the processes (X_t) and (U_t) as well as (Y_t) and (V_t) have the same distribution (process wise).

We now give a very useful Theorem in probability theory linking stochastic ordering and coupling:

Theorem 7 *$X \leq_{st} Y$ if and only if there exists a coupling $(\tilde{X}, \tilde{Y})$ such that $X \leq Y$ almost surely.*

The theorem also extends to stochastic processes.

We illustrate here the power of the coupling technique on a simple example that will be useful for comparing stochastic networks. A point $x \in \mathcal{Z}_+^N$ has coordinates $(x_1, \ldots, x_N)$ and e_i denotes the unit vectors in $\mathcal{Z}_+^N$. We call a MJP a birth and death process if from a point $x \in \mathcal{Z}_+^N$, it can jump only to a neighboring point i.e. a point $y = x + e_i$ or $y = x - e_i$.

Consider two birth and death processes (i.e. processes having transitions only from one state to the neighboring states) defined on $\mathcal{Z}_+^N$.

Lemma 2 *Let $X = (X_1, \ldots, X_N)$ and $Y = (Y_1, \ldots, Y_N)$ be multiclass birth and death processes such that X has birth rates $\lambda_i(x)$ and death rates $\phi_i(x)$, and Y has birth rates $\eta_j(y)$ and death rates $\psi_j(y)$. Assume that for all $i = 1, \ldots, N$, and all $x, y \in \mathcal{Z}_+^N$ such that $x_i = y_i$ and $(x_1, \ldots, x_N) \leq (y_1, \ldots, y_N)$,*

$$\lambda_i(x) \leq \eta_i(y), \tag{1.15}$$

$$\phi_i(x) \geq \psi_i(y). \tag{1.16}$$

Then for all $x, y \in \mathcal{Z}_+^N$ such that $(x_1, \ldots, x_N) \leq (y_1, \ldots, y_N)$,

$$(X_1[x], \ldots, X_N[x]) \leq_{\text{st}} (Y_1[y], \ldots, Y_N[y]),$$

where $X[x]$ and $Y[y]$ are versions of X and Y started in x and y, respectively.

Proof: Let $(\tilde{X}, \tilde{Y})$ be the Markov process with paths in $D(\mathcal{R}_+, U)$, where $U = \{(x, y) \in \mathcal{Z}_+^N \times \mathcal{Z}_+^N : (x_1, \ldots, x_N) \leq (y_1, \ldots, y_N)\}$, having the upward transitions

$$\begin{array}{lll}
(x, y) \mapsto (x + e_i, y) & \text{at rate } \lambda_i(x), & x_i < y_i, \\
(x, y) \mapsto (x, y + e_i) & \text{at rate } \eta_i(y), & x_i < y_i, \\
(x, y) \mapsto (x + e_i, y + e_i) & \text{at rate } \lambda_i(x), x_i = y_i, & \\
(x, y) \mapsto (x, y + e_i) & \text{at rate } \eta_i(y) - \lambda_i(x), & x_i = y_i,
\end{array}$$

and the downward transitions

$$\begin{array}{lll}
(x, y) \mapsto (x - e_i, y) & \text{at rate } \phi_i(x), & 0 < x_i < y_i, \\
(x, y) \mapsto (x, y - e_i) & \text{at rate } \psi_i(y), & 0 < x_i < y_i, \\
(x, y) \mapsto (x - e_i, y - e_i) & \text{at rate } \psi_i(y), & 0 < x_i = y_i, \\
(x, y) \mapsto (x - e_i, y) & \text{at rate } \phi_i(x) - \psi_i(y), & 0 < x_i = y_i.
\end{array}$$

In light of 1.15 and 1.16, we see that all transition rates described above are positive. Moreover, because each of the transitions are mappings from U into U, we can be assured that the process $(\tilde{X}, \tilde{Y})$ exists.

By studying the marginals of the transition rates, we see that both $\tilde{X}$ and $\tilde{Y}$ are Markov, and that their intensity matrices coincide with those of X and Y, respectively. Hence, for all x and y such that $(x, y) \in U$, we have constructed versions of $X[x]$ and $Y[y]$ on a common probability space such that $(X_1[x](t), \ldots, X_N[x](t)) \leq (Y_1[y](t), \ldots, Y_N[y](t))$ for all t almost surely. □

1.6 Stability analysis

Often the positive recurrence (also called stability or equivalently here the ergodicity) of MJP that are not defined on a compact space depends in an intricate manner of the transitions parameters and can be difficult to establish when there is no closed-form expression available for the invariant measure. A useful technique to get sufficient conditions of stability is explained below and follows ideas steaming from the analysis of deterministic dynamical systems.

1.6.1 The Foster-Lyapunov criterion

We first need to establish a preliminary result. The idea is to come up with a manipulable representation of the invariant measure.

Proposition 6 *Let F a finite subset of S and T_F the hitting time of F. Let $g(x) = E^x[T_F]$. If g is finite and $E^x[g(X(1))] < \infty$ for any $x \in F$, then X is ergodic.*

Proof: Define the increasing sequence of stopping times t_n by $t_0 = 0$ and $t_n = \inf\{s > t_{n-1}+1, X_s \in F\}$. (We use later on that $t_n - t_{n-1} \geq 1$.) $X(t_n)$ is a Markov chain (because X is a MJP) on F, irreducible since X is irreducible thus having an invariant distribution by the Perron-Frobenius theorem. Let π_F this distribution. Define now $\tilde{\pi}$ by (for any positive function)

$$\tilde{\pi}(f) = E^{\pi_F}\left(\int_0^{t_1} f(X_s)ds\right).$$

We have that $\tilde{\pi}(\mathbb{1}) = E^{\pi_F}(t_1) = 1 + \sum_{x\in F} g(x) < \infty$ by assumption. Hence, we may define π the probability distribution obtained by normalizing $\tilde{\pi}$. Let C the normalization constant. We now observe that:

$$\begin{aligned} E^{\pi}(f(X_t)) &= \sum_{x\in S} \pi(x) E^x[f(X_t)] \\ &= \frac{1}{C}\sum_{x\in S} E^{\pi_F}(\int_0^{t_1} 1_{X_s=x} ds) E^x[f(X_t)] \\ &= \frac{1}{C}\sum_{x\in S} E^{\pi_F}(\int_0^{\infty} 1_{X_s=x, t_1>s} ds) E^x[f(X_t)]. \end{aligned}$$

Using the Markov property

$$E^{\pi_F}[f(X_{t+s})1_{X_s=x,t_1>s}] = E^{\pi_F}[E^{X_s}[f(X_t)]1_{X_s=x,t_1>s}] = E^x[f(X_t)]E^{\pi_F}[1_{X_s=x,t_1>s}],$$

which gives:

$$\begin{aligned} E^{\pi}(f(X_t)) &= \frac{1}{C}\sum_{x\in S} E^{\pi_F}(\int_0^{\infty} f(X_{t+s})1_{X_s=x,t_1>s} ds], \\ &= \frac{1}{C} E^{\pi_F}[\int_0^{t_1} f(X_{t+s}) ds], \\ &= \frac{1}{C} E^{\pi_F}[\int_t^{t+t_1} f(X_s) ds], \\ &= \frac{1}{C}\left(E^{\pi_F}[\int_t^{t_1} f(X_s) ds] + E^{\pi_F}[\int_{t_1}^{t+t_1} f(X_s) ds]\right). \end{aligned}$$

Using the definition of π_F, $E^{\pi_F}[\int_{t_1}^{t+t_1} f(X_s)ds] = E^{\pi_F}[\int_0^t f(X_s)ds]$. Hence

$$E^{\pi}(f(X_t)) = \frac{1}{C} E^{\pi_F}[\int_0^{t_1} f(X_s) ds] = \pi(f).$$

Lyapunov-Foster criterion

We can now state the most used result in the analysis of the stability of Markov processes.

Theorem 8 *If there exists a (so-called Lyapnov) function $L : \mathcal{X} \to \mathbb{R}_+$, some constants $K, > 0$ and an integrable stopping time τ such that:*

1. *$E^x[L(X_\tau)] - L(x) \leq -\gamma E^x[\tau]$, for $L(x) > K$,*
2. *the set $F = x : L(x) \leq K$ is finite,*
3. *$E^x[L(X(1))] < \infty$, for all $x \in S$.*

then the MJP X is ergodic and the hitting time T_F of F satisifes:

$$E^x(T_F) \leq \frac{L(x)}{\gamma}, \ x \notin F$$

.

Proof: Define inductively the sequence of random times (t_n) by $t_0 = 0$ and $t_n = t_{n-1} + \theta_{t_{n-1}} \circ \tau$. Define the filtrations $\mathcal{F}_n^\tau = \mathcal{F}_{t_n}$ and

$$\nu = \inf\{n : t_n \geq T_F\}.$$

ν is a stopping time with respect to $(\mathcal{F}_n^\tau)$. Define further $S_n = L(X_{t_n}) + \gamma t_n$. S_n is $\mathcal{F}_n^\tau$ measurable and

$$\begin{aligned} E^x[S_{n+1}|\mathcal{F}_n^\tau] &= E^x[L(X_{t_n+\tau\circ\theta_{t_n}}) + \gamma(t_n + \tau \circ \theta_{t_n})|\mathcal{F}_n^\tau], & (1.17)\\ &= \gamma t_n + E^{X_{t_n}}[L(X_\tau) + \gamma\tau]. & (1.18) \end{aligned}$$

If $\nu > n$, $t_n < T_F$ and $L(X_{t_n}) > K$, hence, on $\nu > n$, $E^x[S_{n+1}|\mathcal{F}_n^\tau] - S_n = E^{X_{t_n}}[L(X_\tau) - f(X_{t_n}) + \gamma\tau] \leq 0$. This implies that $S_{n\wedge\nu}$ is a non-negative supermartingale. Hence $E^x(S_{n\wedge\nu}) \leq S_0$ and

$$E^x[L(X_{t_{\nu\wedge n}}) + \gamma t_{\nu\wedge n}] \leq L(x),$$

and by monotone convergence this implies (L non negative) $E^x(t_\nu) \leq \frac{L(x)}{\gamma}$, which implies the bound on the hitting time. We then use the previous Proposition.

□

Appendix :Martingales

Any MJP process can be decomposed into two terms, one being the drift part and the other being in some sense the noise part. This appealing decomposition is a powerful tool to study the time evolution of the distribution of a Markov process and is based on the important notion of martingale. We refer to [6] and [10] for an introduction on the subject and for more details. We recall here only the main definitions and very basic results that we might use along the way.

Definition 7 *A continuous time process M_t is called a martingale if*

- $E(M_t|F_s) = M_s, \forall t \geq s.$
- $E\left(\sup_{0\leq s\leq t}|M_s|\right) < \infty, \ \forall t.$

It is a sub-martingale if

$$E(M_t|F_s) \geq M_s, \forall t \geq s.$$

It is a super-martingale if

$$E(M_t|F_s) \leq M_s, \forall t \geq s.$$

We now state three crucial results concerning Martingales:

Proposition 7 *If T is a stopping time (i.e. $\{T \leq t\} \in \mathcal{F}_t, \forall t$), then:*

$$E(M_{T\wedge t}) = E(M_0).$$

Proposition 8 *A non negative super-martingale converges almost surely to a finite limit.*

Proposition 9 (Doob's inequality)

$$P(\sup_{0\leq s\leq t} M_s \geq a) \leq \frac{E(M_t)}{a}.$$

Chapter 2

Markovian queues

This chapter is based on the book [9]. We aim at describing finer properties and approximation schemes for the simplest Markovian queues. This can also be seen as an analysis of very specific birth and death processes.

Another aim is to emphasize the differences of behavior between the $M/M/1$ and the $M/M/\infty$. For more advanced readers, this difference is reflected in the spectral description of the process. See the work of Van Doorn in particular.

2.1 M/M/1

This queue was introduced in the previous chapter. We here give a more precise analysis of some of its main feature. It is essential to notice that whereas the stationary properties of the queue are very simple to obtain (see the next proposition), it is on the other hand very cumbersome to study the transient regime of the queue. We therefore establish useful ways to overcome these difficulties either in terms of scalings and approximations with the stationary regime.

Recall that the number of customers X is a MJP with transitions:

$$q(x, x+1) = \lambda, \tag{2.1}$$
$$q(x, x-1) = \mu 1_{x>0}. \tag{2.2}$$

2.1.1 Stationary distribution:

The following Proposition follows immediately from the results of the next chapter.

Proposition 10 *If $\lambda < \mu$, then X is ergodic and its stationary distribution is geometric with parameter $\rho = \frac{\lambda}{\mu}$. The stationary distribution of the workload (which is Markov but not a MJP) is given by:*

$$P(W > x) = \rho e^{-(\mu-\lambda)x}.$$

2.1.2 Hitting times:

We now turn our attention to the hitting times associated with the MJP X_t. Define $T_z = \inf\{s > 0 : X_s = z\}$ (with the convention $\inf \emptyset = \infty$). Remark that if $X(0) = 1$, the T_0 is the busy period of the queue.

Proposition 11 *If $a \geq b$ we have for $s > 0$:*

$$E^a(e^{-sT_b}) = \left(E^1(e^{-sT_0})\right)^{a-b},$$

and

$$E^1(e^{-sT_0}) = \frac{\lambda + \mu + s - \sqrt{(\lambda + \mu + x)^2 - 4\lambda\mu}}{2\lambda}.$$

Further:

$$E(e^{((\sqrt{\lambda} - \sqrt{\mu})^2)T_0}) \leq \sqrt{\frac{\mu}{\lambda}}.$$

Proof: The proof first consists in remarking that the hitting time of $b < a$ starting from a can be decomposed as the sum of the hitting time of $a - 1$ and then the hitting time of b starting from $a - 1$. Using the strong Markov property for MJP (see the previous chapter), we obtain by induction that:

$$E^a(e^{-sT_b}) = E(e^{-s\sum_{i=1}^{a-b} T_0^i}),$$

where T_0^i is an i.i.d sequence of r.v with the distribution of T_0. Hence:

$$E^a(e^{-sT_b}) = E^1(e^{-sT_0^1})^{a-b}$$

If we define $D(t)$ as a difference of Poisson processes (which can also be seen as the free process for the queue), then we can easily verify that for $u \in \mathbb{R}$, the process:

$$H(t) = u^{D(t)} e^{\lambda t(1-u) + \mu t)(1 - \frac{1}{u})},$$

is a martingale. It follows from the independence of the increments of the Poisson process that:

$$E(H(t)|\mathcal{F}_s) = H(s) e^{(t-s)(\lambda(1-u) + \mu(1-\frac{1}{u})} E(u^{D_t - D_s}) = D_s.$$

Writing $g(u) = \lambda(1-u) + \mu(1-\frac{1}{u})$ and using the optional stopping Theorem for martingales, we obtain that:

$$E(e^{\log(u) D(t \wedge T_0) + g(u)(t \wedge T_0)}) = 1.$$

We now define u_s such that $g(u_s) = -s$ and using Lebesgue convergence Theorem, we obtain:

$$E\left(\frac{e^{-sT_0}}{u_s}\right) = 1,$$

which proves the second claim after simple calculations. Now choose $s_0 = -(\sqrt{\lambda} - \sqrt{\mu})^2$ and use the same technique and the monotone convergence Theorem.

□

Remark 9 *The Laplace transform is continuous in 0 for $\lambda \leq \mu$ and discontinuous for $\mu > \lambda$.*

2.1.3 Convergence to the stationary measure

We can use the previous Proposition combined with a coupling to get interesting results on the speed of convergence of the queue towards its stationary regime.

Define the total variation norm of a signed measure μ on $\mathbb{N}$ as:

$$||\mu|| = \frac{1}{2}\sum_x |\mu(x)|.$$

We shall later use the following property. For two probability measures μ and ν on $\mathbb{N}$, we have:

$$||\mu - \nu|| = \sup_{A\subset\mathbb{N}} |\mu(A) - \nu(A)|.$$

We now state our estimate of the speed of convergence towards equilibrium:

Proposition 12 *For $\rho < 1$, G a geometric distribution with parameter ρ, and G_x the distribution obtained by taking the supremum of x and a r.v. with distribution G, we have:*

$$||P_t^x - G|| \quad \leq \quad P^{\sup(x,G)}(T_0 > t) \leq (\sqrt{\frac{\mu}{\lambda}}^x + 1)e^{-(\sqrt{\mu}-\sqrt{\lambda})^2 t}. \tag{2.3}$$

Proof: We couple a stationary process $\tilde{X}$ and a process X started at x, as follows. Before meeting they evolve independently. After meeting they follow the same trajectory. Let T^c be the meeting time of the two processes. We have that:

$$\begin{aligned} \sup_A |E[1_{X\in A}] - E[1_{\tilde{X}\in A}]| &\leq \sup_A E|1_{X\in A} - 1_{\tilde{X}\in A}|, &(2.4)\\ &\leq P(T^c > t). &(2.5)\end{aligned}$$

We then argue that the hitting time of the two processes is smaller than the hitting time of 0 by the one of the two process started in a higher value. (Obvious when drawing trajectories). The inequality then follows from Chebichev inequality and the results of Proposition 11.

□

2.1.4 Fluid limits

We here look at a scaling corresponding to zooming out the process. This technique is very useful for more complicated multidimensional processes. Define the scaled Markov process $X^K(t) = \frac{X^{Kx}(Kt)}{K}$ started with Kx customers.

Proposition 13 *We have the almost sure convergence:*

$$X^K(t) \to h(t) = (x + (\lambda - \mu)t)^+,$$

where a^+ denotes the maximum of a and 0.

Proof:

Recall that the hitting time of 0 can be decomposed as a sum of independent random variables identically distributed with mean $\frac{1}{\mu-\lambda}$. Let T_K the hitting time of 0 for the scaled process. Using the strong of large numbers, we get that $T_K/K \to t_1 = \frac{x}{\mu-\lambda}$ almost surely.

Before time T_K, $X^K(t)$ can be written as the difference of two indenpendent (and inter-independent) Poisson processes $\mathcal{N}_\lambda$ and $\mathcal{N}_\mu$. Hence, using the strong law of large numbers, there exists K_0 such that for all $k \geq K_0$:

$$\forall t \leq \frac{T^K}{K},\ |X^K(t) - h(t)| \leq \epsilon, a.s.\ .$$

which given the previous result gives that there exists K_0' such that for all $K \geq K_0'$

$$\forall t < t_1,\ |X^K(t) - h(t)| \leq \epsilon, a.s.\ .$$

We conclude the proof by using stochastic comparisons with a scaled process started in $K(x+\epsilon)$.

Exercise 7 *Show that the workload of a $G/G/1$ queue with independent and interdependent services has a also an almost sure fluid limit.*

Exercise 8 *What about the number of customers in a $M/G/1$ queue?*

2.1.5 Fluctuations around the fluid limit

We can use a functional central limit theorem to get a second-order approximation, when zooming out the state of the queue.

Proposition 14 *For $t \leq t_1$:*

$$\sqrt{K}(X^K - h(t)) \to_{distr.} cB_t,$$

with B_t a standart Browninan motion and $c = \sqrt{\lambda+\mu}$.

Proof: Donsker Theorem shows that a renormalized Poisson process converges in distribution to a Brownian. We the use the almost sure convergence of the hitting time.

2.2 M/M/∞

This queue models the case of an infinite supply of servers. Hence, each customer entering the queue stays an exponential time corresponding to its 'service' time. Hence the number of customers X is a MJP with transitions:

$$q(x, x+1) = \lambda, \tag{2.6}$$
$$q(x, x-1) = x\mu. \tag{2.7}$$

2.2.1 Transient distribution:

Proposition 15 *If $X_0 = 0$, the number of customers at time t has a Poisson distribution with parameter*

$$\mu(t) = \lambda \int \min(x,t)\mu \exp(-\mu x)dx.$$

Proof: Marked point processes and Laplace transforms (See [9]). □

2.2.2 Stationary distribution:

The following Proposition follows immediately from the results of the next chapter and is also a direct consequence of the previous Proposition.

Proposition 16 *X is ergodic (for all parameters) and its stationary distribution is Poisson with parameter $\rho = \frac{\lambda}{\mu}$.*

2.2.3 Convergence to the stationary measure

The speed of convergence is relatively easy to obtain as the transient distribution is Poisson.

Proposition 17 *Let P_ρ a Poisson distribution with parameter ρ.*

$$||P_t^0 - P_\rho|| \quad \leq \quad 1 - e^{-\lambda e^{-\mu t}} \tag{2.8}$$

Proof: It follows from the fact that for two Poisson random variables with parameters $h \leq g$,

$$||Pois_h - Pois_g|| \leq 1 - e^{-(g-h)}.$$

□

2.2.4 Fluid limits

In order to look at scaled version of the process, converging to a deterministic limit, we cannot only zoom out the process anymore. The right scaling is clear when writing the martingale decomposition and is the following. The arrival rates are multiplied by K, state is scaled by K while service rates and time are not scaled.

Define the scaled Markov process $X^K(t) = \frac{X^{Kx}(Kt)}{K}$ started with Kx customers and the error estimate:

$$n_K(t) = \sup_{0 \leq s \leq t} |X^K(s) - y(s)|.$$

with $y(t)$ the deterministic process:

$$y(t) = \rho + (x - \rho)e^{\mu t}.$$

Finally let $\bar{n}_K(t) = En_K(t)$.

Proposition 18 *We have the convergence of the scaled process in L_1 uniformly on compact sets, i.e.:*

$$\bar{n}_k(t) \to 0.$$

Proof:

Using the martingale representation:

$$X^K(t) = X^K(0) + M_t^K + \int_0^t \Delta(X^K(s))ds.$$

where M^K is a martingale. Using Cauchy Schwartz and Doob's inequality:

$$E\left(\sup_{0\le s\le t} M_s^K\right)^2 \le 4E\left(\sup_{0\le s\le t} M_s^{K\,2}\right), \tag{2.9}$$

$$\le 4E\left(M_t^{K\,2}\right). \tag{2.10}$$

$$\tag{2.11}$$

$$\langle M_t^K\rangle = \frac{1}{K^2}\int_0^t K\lambda + \mu X_s^K ds.$$

Using that the process is obviously bounded by taking only arrivals into account:

$$E\langle M_t^K\rangle \le \frac{1}{K}(\lambda t + \lambda\mu t^2/2).$$

Hence:

$$E\left(M_t^{K\,2}\right) \le \frac{1}{K}(\lambda t + \lambda\mu t^2/2), \tag{2.12}$$

and:

$$E\left(\sup_{0\le s\le t} M_s^K\right) \le \sqrt{\frac{1}{K}(\lambda t + \lambda\mu t^2/2)}. \tag{2.13}$$

This now allows to obtain that:

$$\bar{n}_K(t) \le \epsilon_K + \int_0^t \bar{n}_K(s)ds.$$

We conclude using Gronwall Lemma. □

2.2.5 Fluctuations

We conclude this chapter by mentioning a functional central limit theorem for fluctuations around the fluid limit (see [**?**], Chapter 6 for a proof). Define the process $Y^K(t) = \sqrt{K}\left(X^K(t) - y(t)\right)$.

Proposition 19 *Suppose* $\lim Y^K(0) = v$. *Then* Y^K *converges in distribution towards a process* Z_t *with independent increments of the form:*

$$Z_t = ve^{-\mu t} + \int_0^t e^{-\mu(t-s)} h(s)^{1/2} dB_s,$$

with B *a standart Brownian motion and* h *a deterministic function given by:*

$$h(t) = 2\lambda + \mu(x - \rho)e^{-\mu t}.$$

Chapter 3

Reversible processes

As transient distributions are usually very complicated objects, queuing theorists are often primarily interested in solving the stationary distributions of the queues and networks they study. A fundamental tool in that direction is a special notion of symmetry linked to the time reversibility of the process.

3.1 Definition

Definition 8 *X is a reversible MJP if there exists a distribution π on S such that :*

$$\pi(x)q(x,y) = \pi(y)q(y,x) \tag{3.1}$$

These equations are called the **detailed balance equations** and π is then the stationary distribution of X. Indeed the summation of (3.1) imply (1.7).

Proposition 20 *The process X is reversible if and only if there exists a distribution on S and a symmetric function γ (i.e., such that $\gamma(x,y) = \gamma(y,x)$) such that:*

$$q(x,y) = \frac{\gamma(x,y)}{\pi(x)}.$$

3.2 Time reversal

Fix a time $\tau > 0$ and consider the reversed time process

$$X_t^\tau = X_{\tau - t}.$$

It is important to remark that it is difficult to interpret X^τ when X is not stationary: in this case, X^τ is not even time-homogeneous. However, if X is stationary, then its time reversal has the simple structure described in the following proposition:

Proposition 21 *X^τ verifies:*

- *X^τ is Markov,*

- *For $s < t$,*

$$P(X_t^\tau = y | X_s^\tau = x) = \frac{P(X_{\tau-t} = y)}{P(X_{\tau-s} = x)} P(X_{\tau-s} = x | X_0 = y). \tag{3.2}$$

- *If X is stationary ergodic with a stationary distribution π, so is X^τ (stationary ergodic with stationary distribution π).*

- *If X is stationary, the transitions of X^τ are given by*

$$q^\tau(x, y) = \frac{\pi(y)}{\pi(x)} q(y, x). \tag{3.3}$$

Proof: Consider a set $\mathcal{A}$ belonging to the σ-algebra $\mathcal{F}_s^\tau$ (generated by $\{X_r^\tau, 0 \leq r \leq s\}$) and take $s < t$.

$$\begin{aligned} P(X_t^\tau = y | X_s^\tau = x, \mathcal{A}) &= \frac{P(X_{\tau-t} = y, X_{\tau-s} = x, \mathcal{A})}{P(X_{\tau-s} = x, \mathcal{A})}, & (3.4) \\ &= \frac{P(X_{\tau-t} = y) P(X_{\tau-s} = x | X_{\tau-t} = y) P(\mathcal{A} | X_{\tau-s} = x)}{P(X_{\tau-s} = x) P(\mathcal{A} | X_{\tau-s} = x)}. & (3.5) \end{aligned}$$

Suppose X stationary with stationary distribution π, then the law of X_t is π for each t, which implies the second statement. In that case, taking $s = 0$ and $t \to 0$ in (3.2) leads to (3.3).

The meaning and the term of time-reversibility is better understood through the following theorem:

Theorem 10 *X is reversible if and only if X^τ is ergodic and*

$$(X_{\tau-t_1}, \dots X_{\tau-t_n}) =_{distr} (X_{t_1}, \dots X_{t_n}), \tag{3.6}$$

for all sequences of time $t_1, \dots t_n$.

Proof: X is reversible if and only if $q^\tau = q$. Two MJP have the same transitions if and only if they have the same distribution (i.e. same distribution for all finite marginals). □

3.2.1 Kolmogorov criterion

In this section, we give a practical criterion to recognize reversibility without calculating the stationary measure.

Theorem 11 (Kolmogorov) *The following statements are equivalent:*

- *there exists a positive measure π satisfying the detailed balance equations for q ,*
- *$\forall n \in \mathbb{N}$, $\forall x^0, x^1, \dots, x^n \in \mathcal{X}$ with $x^n = x^0$,*

$$\prod_{i=1}^{n} q(x^{i-1}, x^i) = \prod_{i=1}^{n} q(x^i, x^{i-1}). \tag{3.7}$$

- *For each path* $\forall x^0, x^1, \ldots, x^n \in \mathcal{X}$*, such that each* x^i *communicates with* x^{i+1} *for all* i*, the quantity (defined for non-zero transitions)* $\frac{\prod_{i=1}^n q(x^{i-1},x^i)}{\prod_{i=1}^n q(x^i,x^{i-1})}$ *depends only on* x^0 *and* x^n*.*

Proof: We prove that q reversible implies (3.7). There exists π such that $\prod_{i=1}^n q(x^{i-1}, x^i)\pi(x^{i-1}) = \prod_{i=1}^n \pi(x^i)q(x^i, x^{i-1})$. Because $x^0 = x^n$, we can cancel the π. □

If reversibility is satisfied, the stationary measure can easily be expressed in terms of the transitions.

Theorem 12 *If there exists a positive measure* π *satisfying the detailed balance equations for* q*, an invariant measure is given by:*

$$\pi(x) = \prod_{i=1}^{n} \rho(x^{i-1}, x^i), x \in \mathcal{X} \setminus x^0, \tag{3.8}$$

$$\pi(x^0) = 1, \tag{3.9}$$

$$\rho(x^{i-1}, x^i) = \frac{q(x^{i-1}, x^i)}{q(x^i, x^{i-1})}. \tag{3.10}$$

The following proposition is essential to study processes that are restriction of reversible processes.

Proposition 22 *Consider a MJP* X *reversible with stationary distribution* π *on* S*. Then, its restriction* $\tilde{S}$ *to* $\tilde{S} \subset S$ *is reversible with a stationary measure* $\tilde{\pi}$*. If* X *is ergodic, then* $\tilde{X}$ *is ergodic with stationary distribution:*

$$\tilde{\pi}(x) = \frac{\pi(x)}{\sum_{y \in \tilde{S}} \pi(y)}. \tag{3.11}$$

Proof: $x, y \in \tilde{S}, \pi(x)q(x, y) = \pi(y)q(y, x)$. $\tilde{X}$ is reversible and π is invariant for $\tilde{X}$ on $\tilde{\mathcal{X}}$. □

3.2.2 Reversible state dependent birth and death processes

As previously defined, a birth and death process is a process with transitions from $x \in \mathcal{X}$ to $y = x + e_i$ or $y = x - e_i$. The analysis of many processes representing the state of some queuing network boils down to find the stationary measure of birth and death processes.

One-dimensional birth and death processes A one dimensional birth and death death process with strictly positive transitions is always reversible. It follows from the Kolmogorov criterion (or alternatively from a simple manipulation of the global balance equations) that:

Proposition 23 *The transitions of a birth and death process on $\mathbb{N}$ with strictly positive transitions $\lambda(n)$ and $\mu(n)$ satisfy the detailed balance equations and has the following invariant measure:*

$$\pi(n) = \prod_{i=1}^{n} \frac{\lambda(i-1)}{\mu(i)}.$$

Exercise 9 *Prove the previous Proposition.*

Fixed birth rates, state dependent death rates: We consider a birth and death process X on $\mathbb{N}^d$ with the following strictly positive transitions:

$$\begin{aligned} q(x, x+e_i) &= \lambda_i, \ i = 1, \dots d \\ q(x, x+e_i) &= \phi_i(x), \ i = 1, \dots d. \end{aligned}$$

Proposition 24 *The following statements are equivalent:*

- *There exists π solution of the detailed balance equations for q.*
- *The transitions $\phi_i(x)$ verify:*

$$\phi_i(x)\phi_j(x-e_i) = \phi_j(x)\phi_i(x-e_j), \forall i, j \in \{1 \dots, d\}, \forall x \in \mathbb{N}^d. \tag{3.12}$$

- *There exists a strictly positive function Φ such that the transitions $\phi_i(x)$ write:*

$$\phi_i(x) = \frac{\Phi(x-e_i)}{\Phi(x)}, \forall i \in \{1 \dots, d\}, \forall x \in \mathbb{N}^d. \tag{3.13}$$

Exercise 10 *Prove the previous Proposition.*

Fixed death rates, state-dependent birth rates: The same analysis can be carried out for the symmetric situation:

$$\begin{aligned} q(x, x+e_i) &= \lambda_i(x), \ i = 1, \dots d \\ q(x, x+e_i) &= \phi_i, \ i = 1, \dots d. \end{aligned}$$

Using Kolmogorov criterion, we get the following Proposition:

Proposition 25 *The following statements are equivalent:*

- *There exists π solution of the detailed balance equations for q.*
- *The transitions $\lambda_i(x)$ verify:*

$$\lambda_i(x)\lambda_j(x+e_i) = \lambda_j(x)\lambda_i(x+e_j), \forall i, j \in \{1 \dots, d\}, \forall x \in \mathbb{N}^d. \tag{3.14}$$

- *There exists a strictly positive function Λ such that the transitions $\lambda_i(x)$ write:*

$$\lambda_i(x) = \frac{\Lambda(x+e_i)}{\Lambda(x)}, \forall i \in \{1 \dots, d\}, \forall x \in \mathbb{N}^d. \tag{3.15}$$

Chapter 4

Jackson and Whittle networks

4.1 The tandem queue

The simplest example of two interacting queues consists of a two queues in tandem: customers arrive at queue 1 according to a Poisson process (of rate λ) and demand an exponentially distributed service time of mean $1/\mu_1$. After being served at the first queue, they joined the second queue, when they once again demand an exponentially distributed service time of mean $1/\mu_2$. The assumption of exponential inter-arrival and service times has two important consequences:

- Any non-anticipating and work conserving scheduling (not taking into account the residual service times of the customers in the system) is equivalent in the sense that the number of customers has the same probabilistic description.
- The process describing the number of customers at each queue is a Markov process.

Though the dynamics of the second queue are clearly impacted by the dynamics of the first one, we will see that in the stationary regime, the 2 queues behave as if they were independent. This property is closely related to some reversibility properties of the system. Verify however that the MJP process describing the number of customers in the 2 queues is not strictly reversible. (Consider for instance the states $(0,0)$ and $(1,0)$.) The process is going to be called quasi-reversible or partially reversible and the theory of partial reversibility is going to be extensively studied in the next Chapter.

4.1.1 Burke-Reich theorem

Reversibility of the first queue The number of customers in the first queue is itself a Markov process. It is thus a one-dimensional birth and death process (corresponding to a $M/M/1$ queue) and it is reversible. It is stationary if $\lambda < \mu_1$. (More about ergodicity conditions is established in Chapter **??**.)

Theorem 13 (The output Theorem) *If $\lambda < \mu_1$, the departure process from the first queue in the stationary regime is a Poisson process.*

Proof: Suppose $\lambda < \mu_1$, then X_1 is stationary. Using the reversibility and stationarity of X_1, the time reversal process of X_1 has the same transitions as X. Hence, its arrival process is a Poisson process. The departure process of the queue can be interpreted as the arrival process of the time reversed process of X. □

Independence of the two queues in the stationary regime: Denote $D(t)$ the process couting the departures from the first queue. Using the reversibility of the first queue, the following result can be established:

Proposition 26 *Assume X stationary. For each time t_0, $X_1(t_0)$ is independent of $D(t), t < t_0$.*

Proof: In the stationary regime, the reversed time process X_1^τ has the same dynamics as X_1. The arrival process of the reversed-time is the departure process of X_1 . The arrival process after t_0 of X_1^τ is independent of $X_1^\tau(s), s \leq t_0$. Hence, the departure process and the arrival process of X_1 before t_0 are independent of $X_1(t_0)$. □

We deduce from the previous Proposition that:

- In the stationary regime, the arrival process of the second queue is a Poisson process and is independent of the state of the first queue.
- In the stationary regime, the two queues are independent. Thus, the stationary measure has a product form $\pi(x) = \pi_1(x_1)\pi_2(x_2)$.
- Solving the balance equations, (given we have a product form) leads to:

$$\pi(x) = \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^{x_1}\left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^{x_2} \tag{4.1}$$

4.1.2 Partial reversibility

In this section, we show the same result using a different approach, that will turn out to be more powerful in the sequel. Since the routing between queue 1 and 2 breaks the strict reversibility property of the 2-dimensional process X, it is natural to consider an auxiliary process $\bar{X}$ (that we shall call the adjoint process), that has the same input and service rates that X, but different transitions, given by:

$$\bar{q}(x, x + e_i) = \lambda, \ i = 1, 2, \tag{4.2}$$
$$\bar{q}(x, x - e_i) = \mu_i, \ i = 1, 2. \tag{4.3}$$

This process can directly be interpreted as two independent one-dimensional process and it is obvious that is stationary measure has a product form, given by (4.1). The partial balance equations for $\bar{X}$ write:

$$\pi(x)\bar{q}(x, x + e_i) = \pi(x + e_i)\bar{q}(x, x - e_i), \ i = 1, 2, . \tag{4.4}$$

which can be rewritten as:

$$\begin{aligned} \pi(x)\lambda &= \mu_2\pi(x+e_2), && (4.5)\\ \pi(x)\mu_2 &= \lambda\pi(x-e_2) = \mu_1\pi(x+e_1-e_2), && (4.6)\\ \pi(x)\mu_1 &= \lambda\pi(x-e_1). && (4.7) \end{aligned}$$

Summing these equalities,

$$\pi(x)(\lambda+\mu_1+\mu_2) = \mu_1\pi(x+e_1-e_2) + \mu_2\pi(x+e_2) + \lambda\pi(x-e_1),$$

This approach does actually not need any assumption on the routing matrix and shall considerably be generalized in the next chapter.

4.2 Open feed-forward Jackson networks

The direct reversibility arguments involved for the proof of the Burke-Reich theorem can easily be extended to any kind of feed-forward networks, i.e., such that once a node is visited, it is not visited twice. Hence, defining the traffic equations

$$\omega_i = \lambda_i + \sum_j p_{ji}\omega_j, \tag{4.8}$$

the product form of a N-dimensional feed forward networks with constant service rates at each queue directly follows:

$$\pi(x_1,\ldots,x_N) = \prod_{i=1}^{N}\left(1-\frac{\omega_i}{\mu_i}\right)\left(\frac{\omega_i}{\mu_1}\right)^{x_i}. \tag{4.9}$$

4.3 Whittle networks

We will consider here an open queuing network of N nodes. At any node i, exogenous arrivals form a **Poisson process of intensity** ν_i.

In all the following, we will focus on the process $X(t) = (x_1(t), ..., x_N(t))$ representing the number of clients at each node of a Whittle network. We may consider either a finite or infinite state space $\mathcal{X}$. e_i shall denote i-th the unit vector in $\mathbb{N}^N$.

4.3.1 Definitions

Definition 9 (Processor-sharing queues and networks) • *In a processor-sharing queue of capacity $\phi_1(t)$, the service capacity of the queue at time t is equally shared between all the clients present in the queue at every instant t. (Hence, any client is served at time t with service rate $\frac{\phi_1(t)}{x_1(t)}$)*

- *We shall call a processor-sharing network a queueing network of processor-sharing nodes with state-dependent capacities. The dynamics of a Processor sharing network with exponential service requirements can be characterized by the following principle: whenever the network is in state x, the time to the next movement of a single unit from node i to node j is exponentially distributed with rate $p_{ij}\phi_i(x)$. (p_{ij} represents the routing probability from node i to node j while $\phi_i(x)$ represents the allocation of badwidth at node i). At node i, customers require i.i.d.* `exponential services` *services of mean $1/\mu_i$. After service completion at node i, a customer is routed to node j with probability p_{ij} and leaves the network with probability $p_i \equiv 1 - \sum_j p_{ij}$.*

  The transition rates of the number of customers or units at each node are hence given by:

$$q(x,y) = \begin{array}{c} p_{jk}\phi_j(x) \quad if \quad y = x - e_j + e_k \\ p_j\phi_j(x) \quad if \quad y = x - e_j \\ \nu_j \quad if \quad y = x + e_j \\ 0 \quad otherwise \end{array} \tag{4.10}$$

Definition 10 (Whittle network) *A Whittle network is a processor sharing network such that the service rates have the property of balance :*

$$\phi_i(x)\phi_j(x - e_i) = \phi_j(x)\phi_i(x - e_j), \quad i,j = 1,\ldots,N,\ x_i > 0,\ x_j > 0. \tag{4.11}$$

We assume moreover that $\phi_i(x) > 0$ if and only if $x_i > 0$.

Remark 14 *These networks are somehow the most general instances of tractable multi-class networks as the capacity (or speed) ϕ_i of node i may depend on the* **whole state** *of the system $x = (x_1,\ldots,x_N)$, where x_i is the number of customers at node i.*

Remark 15 *We shall relax the assumption of exponential services later on in the sense that a network with non-exponential services at each node but all other characteristics preserved will have the same stationary measure.*

The balance capacity is directly linked to the reversibility of the process representing the number of customers in the network (see the previous Chapter). Let $\langle x, x - e_{i_n}, \ldots, x - e_{i_{n-1}} \ldots x - e_{i_1} x, 0\rangle$ be a direct path from state x to state 0, i.e., a path of length n where n is the number of customers in state x. The balance property implies that the expression

$$\Phi(x) = 1/\phi_{i_1}(x)\phi_{i_2}(x - e_{i_1}) \ldots \phi_{i_n}(x - e_{i_{n-1}} \ldots - e_{i_1}), \tag{4.12}$$

is independent of the considered direct path. In particular, the capacities are uniquely characterized by the function Φ, referred to as the balance function:

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}, \quad i = 1,\ldots,N,\ x_i > 0. \tag{4.13}$$

Conversely, if there exists a function Φ such that the capacities satisfy (4.15), it can be easily verified that these capacities are balanced. We say that the capacities are balanced by Φ.

Remark 16 (Jackson networks) *A Jackson network is a Whittle network in which the service rate at each node depends only of the number of clients at this node.*

4.3.2 Invariant measures

Open whittle networks

For the sake of simplicity, we focus thereafter on the case of open or closed networks. However, this assumption is not necessary and the same analysis could be done for mixed networks. The arguments to obtain the stationary measure of these networks are a systematic extension of those given in the previous Chapter for the tandem queue. Let $\rho_i = \frac{\lambda_i}{\mu_i}$.

Theorem 17 (Invariant measure) *The invariant measure of an open Whittle network is given by:*

$$\pi(x) = \Phi(x) \prod_{i=1}^{N} \rho_i^{x_i}, \tag{4.14}$$

where Φ is the balance function of the network:

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}, \quad i = 1, \ldots, N, \ x_i > 0. \tag{4.15}$$

The effective arrival rate λ_i at node i is uniquely defined by the equations:

$$\lambda_i = \nu_i + \sum_j \lambda_j p_{ji}, \quad i = 1, \ldots, N.$$

and $\rho_i = \lambda_i / \mu_i$ the traffic intensity at node i.

Theorem 18 *Define another Markov Jump process $\bar{X}$, called the adjoint process, by the following transitions:*

$$\begin{aligned} x &\mapsto x + e_i \quad \text{at rate } \lambda_i, \\ x &\mapsto x - e_i \quad \text{at rate } \phi_i(x) 1(x_i > 0), \end{aligned}$$

The process $\bar{X}$ is reversible and has the same stationary measure as X.

Proof: We prove both theorems simultaneously. It follows from the results on Chapter 3, that $\bar{X}$ is reversible. Hence its stationary measure is given by 4.14. We then verify the global balance equations for X using π. We have:

$$\lambda_i \pi(x) = \pi(x + e_i) \phi_i(x + e_i),$$

which leads to:

$$\begin{aligned} \sum_i \phi_i(x) \pi(x) &= \sum_i \pi(x - e_i) \lambda_i, & (4.16) \\ &= \sum_i \pi(x - e_i)(\nu_i + \sum_j p_{ji} \lambda_j), & (4.17) \\ &= \sum_i \pi(x - e_i) \nu_i + \sum_j p_{ji} \phi(x - e_i + e_j) \pi(x - e_i + e_j), & (4.18) \end{aligned}$$

Summing the traffic equations gives $\sum \nu_i = \sum_i \lambda_i p_i$. Hence:

$$\sum_i \nu_i \pi(x) = \sum_i \pi(x)\lambda_i p_i, \tag{4.19}$$

$$= \sum_i p_i \phi_i(x + e_i)\pi(x + e_i). \tag{4.20}$$

□

A key question is then to characterize the allocations $(\phi_j)_{j=1..N}$ for which a PS network is a Whittle network for any distribution of services and hence to find a necessary and sufficient condition of "insensitivity". We prove in the next Section that the property of balance is a necessary and sufficient condition of insensitivity for processor-sharing networks with fixed arrival rates.

4.3.3 Examples of Whittle networks

We describe here various models useful to develop performance evaluation and dimensioning tools for various real-life networks functioning according it specific topologies.

Example 19 (Loss network)

Our first example deals with the modeling of the so called streaming traffic corresponding mainly to voice or video traffic and using a circuit-switched like technology. Equivalently one may think of a telephone network as follows. The network is constituted of a set of links each one having several lines or circuits. Te number of circuits in each link is called the capacity of the link. In the network, there are several routes corresponding to different source destinations pairs that can communicate with each other. Calls arrive to the network and if accepted, occupy a circuit in each of the links of the route that follow, during the whole duration of the call. Suppose now that calls arrive according to a Poisson process and having an exponentially distributed duration with parameter 1*. Let X_r denote the number of r. Let $X = (X_1, \ldots, X_N) \in \mathbb{N}^N$ be the N-dimensional stochastic process on describing the numbers of calls in progress on each route.*

Assume first that all the links have an infinite number of circuits. Then the coordinate of X are independent and each one behave as an $M/M/\infty$ queue. I.e., the transitions of X are given by:

$$q(x, x - e_i) = x_i,$$
$$q(x, x + e_i) = \lambda_i,$$

and the stationary distribution of the system is given by:

$$\pi(x) = \prod_i \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!}.$$

Now assume that a given link l can support at most C_l calls. Then the state space is reduced to $\mathcal{S} = \{x \in\in \mathbb{N}^N, Ax \leq C$, where the matrix A is the link-route incidence

matrix and C is the vector (C_l) of capacities of the links. Hence the process $\tilde{X}$ has the same dynamics as the process X previously described if we truncate the whole orthant to $\mathcal{S}$. Using the reversibility of X, and Proposition 22 of the previous Chapter, we obtain that $\tilde{X}$ is reversible and has the stationary distribution:

$$\tilde{\pi}(x) = \frac{1}{Z}\pi(x)1_{x\in\mathcal{S}},$$

where Z is the normalizing constant associated to the state space $\mathcal{S}$. This normalizing constant can be difficult to calculate for high dimensional networks. However there are several techniques to (try to) overcome this difficulty:

- *Using recursive calculations to obtain Z. The efficiency of this program depends crucially on the shape of the state space (i.e. on the specific constraints).*
- *Using bounds (involving for instance only the usual Erlang formula).*

Example 20 (A processor sharing queue with several classes of customers)

We now turn our attention to a different kind of traffic, the so-called date traffic, corresponding mainly to file and webpages downloads and emails.

Consider a single link sharing its capacity between several competing flow transfers. Here the capacity has to be understood as the number of bits per second that can be transmitted through the link. Suppose that flows of class-i arrive according to a Poisson process of parameter λ_i and have an exponentially distributed size with parameter μ_i.

Let $X = (X_1, \ldots, X_N) \in \mathbb{Z}^N$ be the N-dimensional stochastic process describing the numbers of flows in progress. Thus, X_i represents the number of concurrent flows of class i. In the absence of additional priority mechanisms, and supposing a fair-sharing of the capacity of the link, X is a multi-dimensional birth-and-death process with transition rates:

$$\begin{aligned} q(x, x - e_i) &= \mu_i \frac{x_i}{|x|}, \\ q(x, x + e_i) &= \lambda_i, \end{aligned}$$

where $|x| = \sum_{j=1}^N x_j$.

Using the results of this Chapter, it is not difficult to realize that X is stable if $\sum_{i=1}^N \frac{\lambda_i}{\mu_i} < 1$ in which case it is reversible and has the stationary distribution:

$$\pi(x) = \binom{|x|}{x_1, \ldots, x_N} \prod_{i=1}^N \frac{(\lambda_i}{\mu_i)}^{x_i} (1 - \sum_{i=1}^N \frac{\lambda_i}{\mu_i}).$$

Example 21 (Bandwidth sharing network with balanced fairness allocation)

Bandwidth-sharing network models have become quite a standard modeling tool over the past decade for communication networks. In particular, they have been used extensively to represent the flow level dynamics of data traffic in wireline and wireless networks as well as for the integration of voice and data traffic Bandwidth-sharing networks generalize more traditional voice traffic models, e.g. [7]. We give in the following a succint description of these models.

Consider a communication with N traffic classes. Each traffic class uses a given route through a wired network or consumes a certain amount of a wireless link. To keep the description simpler, we shall only consider the case of wired networks.

Consider for example the tree network represented in Figure 4.1, with two traffic routes, each passing through a dedicated link, followed by a common link. If each dedicated link has a capacity $c_i \leq 1$, $i = 1, 2$, and the common link has capacity 1, the flow on each route gets a capacity $\phi_i(x)$ that lies in the polyhedron $\mathcal{C}$:

$$\begin{aligned} \phi_i(x) &\leq c_i, \quad i = 1, 2, && (4.21) \\ \sum_{i=1}^{2} \phi_i(x) &\leq 1. && (4.22) \end{aligned}$$

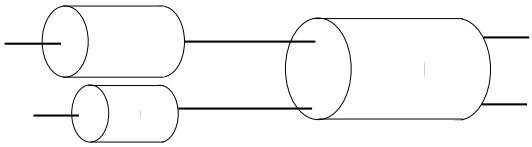

Figure 4.1: Tree network

Another example is the linear network represented in Figure 4.2 with 3 routes sharing two links. While the first route passes through both links, routes 2 and 3 only use one of the links (one each). This gives the following capacity constraints:

$$\begin{aligned} \phi_1(x) + \phi_2(x) &\leq c_1, && (4.23) \\ \phi_1(x) + \phi_3(x) &\leq c_2. && (4.24) \end{aligned}$$

*In general, like for the specific foregoing examples, the capacity constraints determine the space over which a network controller can choose a desired allocation function. It has been argued in [**?**] that a good approximation of current congestion control algorithms such as TCP (the Internet's predominant Transfer Control Protocol) can be obtained by using the weighted proportional fair allocation, which solves an optimization problem for each vector x of instantaneous numbers of flows. Specifically, the weighted proportional fair allocation $\eta(x)$ for state vector x maximizes*

$$\sum_{i=1}^{N} w_i x_i \log(\eta_i), \eta \in \mathcal{C},$$

where the weights w_i are class-dependent control parameters.

*This framework has been generalized to so-called weighted α-fair allocations, which provide flexibility to model different levels of fairness in the network. Another important alternative is the balanced fair allocation [**?**], which allows a closed form expression for the stationary distribution of the numbers of flows in progress. In addition, the balanced fair allocation gives a good approximation of the proportional fair allocation while being easily evaluated, which is attractive for performance evaluation.*

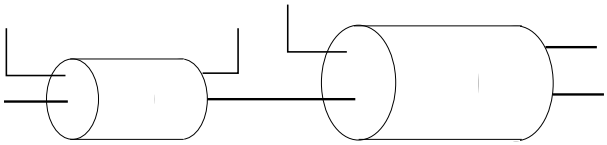

Figure 4.2: Linear network

From a "queuing network perspective, even if there is no queue here, resources are shared according to a processor-sharing service discipline between routes and we can define a processor sharing network to describe the stochastic dynamics of the communication network. The service rates are state-dependent: they may depend on the number of flows within the same class, as well as on the numbers of flows in all the other classes. . Note that the service rate function ϕ captures the allocation of bandwidth which is determined by the specific network topology, the link rates, and the congestion control mechanisms.

Let $X = (X_1, \ldots, X_N) \in \mathbb{Z}^N$ be the N-dimensional stochastic process describing the numbers of flows in progress. Thus, X_i represents the number of concurrent flows of class i. In the absence of additional priority mechanisms, X is a multi-dimensional birth-and-death process with transition rates:

$$\begin{aligned} q(x, x - e_i) &= \mu_i \phi_i(x), \\ q(x, x + e_i) &= \lambda_i, \end{aligned}$$

where the i-th coordinate of the vector $e_i \in \mathbb{Z}^N$ is 1 and all its others coordinates equal 0.

4.3.4 Closed Whittle networks, representation of the stationary measure via an open network

Consider now a closed network $\mathcal{N}$ with K nodes, N customers and fixed Markovian routing p_{ij}. In the sequel, we suppose that at least one of the following assumption is verified:

- the service times are exponentially distributed,
- the scheduling disciplines are PS at each node.

Suppose moreover that the traffic equations have a unique solution $\omega_i, i = 1 \ldots K$, up to multiplicative constant and that the service rates of the networks are balanced, i.e., there exists a so-called balance function Φ such that,

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)},$$

with $x = (x_1, \ldots, x_K)$, and e_i the unit vectors in $\mathbb{N}^K$.

Proposition 27 *The stationary measure of the network writes (for all λ):*

$$\pi(x) = \prod_{i=1}^{K} (\lambda \omega_i)^{x_i} \Phi(x).$$

The stationary distribution of the network can be interpreted as the conditional measure of an equivalent open network as follows:

$$\pi^N(x) = P(X^N = x) = \frac{P(Y = x)}{P(|Y| = N)},$$

where Y is an open network with no internal routing, external arrival ω_i, and the same service rates ϕ_i.

Proof: We prove that the assumptions of the Proposition implies the partial reversibility of the network. Define the adjoint process $\bar{X}$ with transitions $\bar{q}$ given by:

$$\forall i, j, x, \bar{q}(x, x - e_i + e_j) = \phi_i(x)\omega_j.$$

Using the balance conditions of the service rates, it is easily verified that:

$$\pi(x)\bar{q}(x, x + e_i - e_j) = \pi(x + e_i - e_j)\bar{q}(x + e_i - e_j, x), \forall x, i, j.$$

This shows that the network $\mathcal{N}$ is partially reversible and has the stationary measure π Define Y as the process of the number of customers of the open network $\bar{\mathcal{N}}$ with external intensities $\lambda\omega_i$, $i = 1 \dots K$. The balance of the service rates automatically implies the reversibility of Y and that Y has the stationary measure π. The previous Proposition allows to conclude that $P(X^N = x) = \frac{P(Y=x)}{P(|Y|=N)}$. □

Large networks

The last Proposition tells us that the asymptotic behavior of the closed network when $N \to \infty$ is defined by the properties of the open network $\mathcal{N}$ for a large total amount of customers. A natural example is a Jackson network with the same traffic characteristics at each node (see the following section). It is clear from the previous that Y being in this case distributed as a set of independent variables, the central limit theorem allows to approximate the behaviour of the closed network.

Another interesting example is a closed network with a fair sharing of a unique capacity between the K classes of the network, we have, for x such that $|x| = N$,

$$\Phi(x) = \binom{N}{x_1, \dots x_K}.$$

Choosing λ such that $\lambda \sum \omega_i = 1$, and writing $(p_i)_{i=1,\dots,K} = (\lambda\omega_i)_{i=1,\dots K}$, we get that π^N is a multinomial distribution and using the central limit theorem, the marginal distributions $\pi_i^N(x_i)$ verifies :

$$\frac{(\pi_i^N - Np_i)}{\sqrt{N}} \to \mathcal{N}(0, 1).$$

Chapter 5

Insensitivity

A stochastic network is generally a "stochastic dynamical" system which is not Markovian, e.g. when the service requirements distributions are not exponentially distributed. We show in the sequel how we can nevertheless use the Markovian analysis developed so far for Whittle networks to deal with much more general cases.

Definition 11 *A network with Poisson arrivals is said to be insensitive when its stationary distribution depends on the service time distribution only via its mean.*

5.1 Renewal processes and residual life paradox

To proceed with the analysis of the insensitivity of Whittle networks, we first need to recall a basic fact of renewal theory. Define $N_t = (T_i)$ a point process such that $T_0 = 0$, the sequence $(T_{i+1} - T_i)$ is i.i.d with absolutely continuous distribution function θ and a mean m. We aim at calculating the distribution of the remaining time to the next point of the point process starting from an arbitrary point t, i.e. the distribution of $R_t = t_{N_t+1} - t$. This can also be thought as the residual time until the next arrival seen by an observer arriving uniformly between two points of N. Think for example that the point process corresponds to arrivals of buses, and that you arrive at the bus stop.

We first develop an intuitive argument. Let A the distribution of the total inter-arrival time seen by the observer. A basic fact is that A is not distributed as θ. An observer has more chances to get into a longer inter-arrival time. Pushing a little bit the intuition, one might convince herself that the (density of) probability of finding an inter-arrival time of length s should be proportional to s as well as the frequency of original inter-arrival times of this length, i.e.

$$d\theta^A(s) = Csd\theta(s).$$

Normalizing gives:

$$d\theta^A(s) = \frac{1}{m}sd\theta(s).$$

Now it is easy to accept that an observe arrives uniformly in A, which gives that:

$$d\bar{\theta}(t) = \frac{1}{m}\int_t^\infty \frac{1}{s} d\theta^A(s) = \frac{1-\theta(t)}{m}dt.$$

We hence have the following Proposition:

Proposition 28 *Given a stationary renewal point process $N_t = (T_i)$, with $T_1 - T_0 \equiv \theta$, the distribution of the residual lifetime is*

$$\bar{\theta}(t) = \frac{1 - \int_t^\infty (1-\theta(s))ds}{m}.$$

Proof: A formal proof can be given using renewal theory (see [1]). We give here a sketch of the proof based on the renewal equation. Given the stationarity of N_t, it is clear that the distribution of R_t does not depend on t. Further using the stationarity we have that $E(N_{t+s}) = E(N_t) + E(N_s)$ which implies that there exists a constant c such that:

$$E(N_t) = ct.$$

Using the renewal equation $N = \bar{\theta} + F * U$ (where $*$ denotes convolution) and taking expectations:

$$ct = \bar{\theta}(t) + \int_0^t c(t-y)d\theta_y.$$

Integrating by parts and normalizing give the desired result.

□

This Proposition can also be seen as a simple consequence of the definition of the Palm probability using the framework of Palm calculus (see [**?**]): For any stationary ergodic process,

$$E(g(S_0)) = E^0(\lambda S_0 g(S_0)),$$

where E^0 is the mean according to the Palm probability, obtained by conditioning on the (0-probability) event $T_0 = 0$.

5.2 The coupling construction for a PS queue with state dependent rates

Consider a processor sharing queue with Poissonian arrivals with state dependent rate $\lambda(n)$ and service rates $\phi(n)$. Customers have a generally distributed (absolutely continuous) service requirement with mean 1 and distribution Θ. Processor sharing means that the residual service time of each customers decrease with speed $\frac{\beta(n)}{n}$, when there are n customers in the system.

To keep a Markovian description of the system, it is now necessary to record the residual service times of each customer. The state space S is the union of $\mathbb{N}$ and of the union of $\mathbb{R}^n$, for each $n \in \mathbb{N}$ (here the ordering of customers is irrelevant) and we define the appropriate σ-algebra. $X_t = (X_t^n, R_t)$ is hence the number of customers and their

residual service times at time t. Note that X is not a MJP but belongs to wider class of Markov process, sometimes designated as piece-wise deterministic Markov processes.

Define $\bar{\theta}^n$ the product distribution on $\mathbb{R}^n$ with $\bar{\theta}$ the distribution of the residual life time for a renewal point process with inter-arrival distribution θ (see the previous Proposition).

Now for a distribution π (on S), define $\bar{\theta}_\pi = \sum_n \pi(n)\bar{\theta}^n$.

Theorem 22 *If the distribution π is solution of the detailed balanced equations with respect to the transitions $\lambda(\cdot)$ and $\phi(\cdot)$, then $\bar{\mu}_\pi$ is invariant for X_t.*

Proof:

This proof is based on the ideas gathered in [12].

Suppose without lost of generality that θ has mean $m = 1$. Define $\hat{\theta}_n = \bar{\theta}^{n-1} \times \theta$. We are going to couple X_t with a process $\hat{X}$ which has the following dynamics:

- on completion of the workload of a customer, it is immediately replaced by a new customer, having an independent workload with distribution θ.
- there are no arrivals.

For any distribution p, $\bar{\mu}_p$ is stationary for $\hat{X}$. Let P and $\hat{P}$ the transition kernels of both process and $\mathcal{D}$ the set of functions such that both kernels are continuous in $t = 0$. Let $f \in \mathcal{D}$.

$$\begin{aligned}
\bar{\theta}_\pi P_h f - \bar{\theta}_\pi \hat{P}_h f &= E^{\bar{\theta}_\pi} f(X_h) - E^{\bar{\theta}_\pi} f(\hat{X}_h) && (5.1)\\
&= h \sum_n \pi(n)[\lambda(n)(\hat{\theta}_{n+1} f - \bar{\theta}_n f) + \phi(n)(\bar{\theta}_{n-1} f - \hat{\theta}_n f)] + o(h), && (5.2)\\
&= h \sum_n [\pi(n)\lambda(n) - \phi(n+1)\pi(n+1)]](\hat{\theta}_{n+1} f - \bar{\theta}_n f) + o(h), && (5.3)\\
&= o(h). && (5.4)
\end{aligned}$$

In equation (5.2), we have used that:

- The service discipline being processor sharing, all customers are symmetric in the sense that if they were numbered in the queue, then every customer receives the same amount of the service.
- Given the symmetry of service, a completion of service for the original process occurs before time h with probability

$$\begin{aligned}
P(\min_{i=1,\ldots,n} \bar{\theta}_i \leq \frac{\beta(n)}{n} h) &= 1 - P(\bar{\theta}_i \geq h\frac{\beta(n)}{n})^n && (5.5)\\
&= 1 - (1 - \int_0^{h\frac{\beta(n)}{n}} (1 - \theta(s))ds)^n, && (5.6)\\
&= h\beta(n) + o(h). && (5.7)
\end{aligned}$$

- The arrival being Poisson, there is no need to keep track of the time since the last arrival. An arrival occurs before h with probability $\alpha(n)h + o(h)$. When an arrival occurs, the new law of the residual times is changed into $\hat{\theta}_{n+1}$ (keeping in mind that the order of the customers does not matter).
- The probability that more than one events (arrival or service completion) is of order $o(h)$.

Using that $\bar{\theta}_\pi$ is stationarity for $\hat{X}$, this gives that

$$|\bar{\theta}_\pi P_h f - \bar{\theta}_\pi f| = o(h),$$

uniformly on $\mathcal{D}$. This implies that $\bar{\theta}_\pi$ is stationary for X.

5.3 Insensitivity of processor sharing networks

We give a necessary and sufficient for a processor sharing network to be insensitive:

Theorem 23 *A processor sharing network with fixed arrival rates is insensitive if and only if it is a Whittle network.*

In other words, partial reversibility or reversibility of the adjoint process is a sufficient and necessary condition of insensitivity.

Proof: Proving that the conditions of the theorem are sufficient for insensitivity can be proved following the same approach as for a single PS queue, combined with the detailed balance equations for the adjoint process.

The second part of the proof (necessary conditions) uses ideas from Bonald and Proutiere [4]. We use an induction of the number of classes. If $N = 1$, note that the process and the adjoint process coincide (once one has get rid of the loops, see Chapter 1) and both are reversible. Suppose the result true for $N-1$ classes. (Assume for simplicity that there are no transitions for one node to itself.) Consider a network of N classes and take the service requirements of node N to be:

- 0, with probability $1 - \alpha$,
- $1/\alpha$, with probability α.

Remark that such a distribution has mean 1. The network with these service requirements at node N is equivalent to a new network with:

$$\tilde{\nu}_i = \nu_i + (1-\alpha)p_{Ni}\nu_N,\ i = 1 \dots N-1,$$

$$\tilde{p}_{ij} = p_{ij} + (1-\alpha)p_{iN}p_{Ni},\ i,j = 1 \dots N-1,$$

$$\tilde{\nu}_N = \alpha\nu_N,$$

$$\tilde{p}_{iN} = \alpha p_{iN},\ i = 1 \dots N-1,$$

$$\tilde{\phi}_i(x) = \phi(x),\ i = 1 \dots N-1,$$
$$\tilde{\phi}_N(x) = \alpha\phi(x).$$

Insensitivity implies that the network has the same invariant measure for all $\alpha > 0$. Letting $\alpha \to 0$, we obtain:

$$\tilde{\nu}_i = \nu_i + p_{Ni}\nu_N,\ i = 1 \dots N-1,$$
$$\tilde{p}_{ij} = p_{ij} + p_{iN}p_{Ni},\ i,j = 1 \dots N-1,$$
$$\tilde{\nu}_N = 0,$$
$$\tilde{p}_{iN} = 0,\ i = 1 \dots N-1,$$
$$\tilde{\phi}_i(x) = \phi(x),\ i = 1 \dots N-1,$$
$$\tilde{\phi}_N(x) = 0.$$

Look at he limiting balance equations. They are the balance equations of a $N-1$ dimensional process. Using the induction assumption, the associated adjoint process is reversible. Now note that the transitions $\bar{q}$ of the adjoint process associated with the $N-1$ first queues are equal to the transitions of the adjoint process associated with the original N queues. Hence for all states such that $x_N = y_N$, we have that:

$$\pi(x)\bar{q}(x,y) = \pi(y)\bar{q}(y,x).$$

Since node N was chosen arbitrarily, the equality is verified for all states.

□

Chapter 6

Appendix: queuing relations

6.1 Little's law

Proposition 29 *Let λ the arrival rate of customers to an ergodic queuing system in steady state and X the process describing the number of customers or jobs in the system (X does not need to be a MJP). Finally, let S the sojourn time of a customer. We have:*

$$E[X] = \lambda E[S].$$

Remark 24 *Note that $\frac{E[X]}{E[S]}$ should intuitively be the departure rate of the system, which in steady state should be equal to the arrival rate.*

Proof: Assume that the system empties infinitely often (otherwise both quantities converge to infinity). Let T a time when the system gets empty (after having being busy) and $k(T_0)$ the number of customers served between 0 and T. Let A_n and D_n the sequences of arrival and departures times of customers (say of the n-th customers). We denote C_n the reunion of both sequences. The empirical mean sojourn time of customers at time T is:

$$\bar{S}_k = \frac{1}{k}\sum_{i=1}^{k}(D_i - A_i) \to E[S],\ k \to \infty.$$

The empirical mean number of customers of a customer at time T is:

$$\bar{X}_T = \frac{1}{T}\int_0^T X(s)ds = \frac{1}{T}\sum_{i=1}^{2k-1} X(C_i^+)(C_{i+1} - C_i).$$

The relation (easily proved with a picture)

$$\sum_{i=1}^{k}(D_i - A_i) = \sum_{i=1}^{2k-1} X(C_i^+)(C_{i+1} - C_i),$$

and the ergodicity of the process $\bar{S}_k =\to E[S],\ k \to \infty$, $\bar{X}_T =\to E[X],\ T \to \infty$, allow to conclude.

6.2 PASTA

Consider a stationary ergodic queuing system where the inter-arrival times are independent from the service times. Define p_x the stationary probability of having x jobs in the system. Define now p_x^0 the number of customers seen at the moment of arrivals.

Proposition 30 *If the arrival processes are Poisson, then* $p_x = p_x^0$.

Proof: We prove the Proposition for a single queue. The general case being more cumbersome but conecptually equivalent. Let $A(t, t+\delta)$ the event of having an arrival in the interval $[t, t+\delta]$

$$\begin{aligned} p_x^0 &= \lim_{t\to\infty}\lim_{\delta\to 0} P(X_t = x | A(t, t+\delta)), && (6.1)\\ &= \lim_{t\to\infty}\lim_{\delta\to 0} \frac{P(X_t = x, A(t, t+\delta))}{P(A(t, t+\delta))}, && (6.2)\\ &= \lim_{t\to\infty}\lim_{\delta\to 0} \frac{P(A(t, t+\delta)|X_t = x)P(X_t = x)}{P(A(t, t+\delta))} && (6.3) \end{aligned}$$

The memory less of the exponential distribution implies that that $A(t, t+\delta)$ is independent of X_t, which allows to conclude.

□

Bibliography

[1] S. Asmussen, Applied probabilities and queues, Springer, 2003

[2] F. Baccelli, P. Bremaud, Elements of queuing theory, Springer, 1994.

[3] T. Bonald, L. Massoulié, A. Proutière and J. Virtamo, A queueing analysis of max-min fairness, proportional fairness and balanced fairness, Queueing Syst. Theory Appl., 53,1-2,65–84, 2006.

[4] T. Bonald and A. Proutière Insensitivity in processor-sharing networks , Proc. of Performance 2002

[5] W. Feller, *An introduction to probability theory and its application, Volume II*, Wiley, 1991.

[6] Olav Kallenberg, *Foundations of Modern Probability*, Second, Springer, 2002

[7] F.P. Kelly, *Reversibility and stochastic networks*, Wiley, 1979.

[8] W. Massey, Stocastic orderings for Markov processes on partially odered spaces. *Mathematics of Oper. Res.* vol 12, No. 2, 1987 350–367.

[9] P. Robert, *Stochastic Networks and Queues*, Stochastic Modelling and Applied Probability Series, Vol. 52, (Springer Verlag, New York), 2003.

[10] L. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, Vol I, Wiley, 1994

[11] R. Serfozo, *Introduction to stochastic networks*, Springer, 1999.

[12] S. Zachary, A note on insensitivity in stochastic networks. Journal of Applied Probability, 44 (1), 238-248 (2007)