Stochastic Modeling and Analysis of Telecom Networks

Stochastic Modeling and Analysis of Telecom Networks

Laurent Decreusefond
Pascal Moyal

Series Editor
Nikolaos Limnios

First published 2012 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George’s Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Cataloging-in-Publication Data

Decreusefond, Laurent, 1966-
Stochastic modeling and analysis of telecom networks / Laurent Decreusefond, Pascal Moyal.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84821-238-1
1. Mobile communication systems. 2. System analysis--Mathematical models. I. Moyal, Pascal, 1978- II. Title.
TK5103.2.D43 2011
621.3845'6051922--dc23
2011036865

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN: 978-1-84821-238-1

Printed and bound in Great Britain by CPI Group (UK) Ltd., Croydon, Surrey CR0 4YY

Table of Contents

Preface

In mobile telecommunications, ARCEP (the French Regulatory Authority for Electronic Communications and Postal services) publishes an annual analysis of quality of different mobile radio networks. For voice, the two criteria are the ability to start up a communication and to hold it for 2 or 5 minutes as well as the audio quality of the communication. For each data service, the transmission time and integrity of the message (SMS, MMS) are tested in different situations: urban, semi-urban, for pedestrians, cars, high-speed train, etc.

The results of these tests are often used as commercial arguments. On the contrary, bad results may rapidly alter the image of a telecom operator in the public opinion and thus lead to an economic disaster. Hence, these performance tests are a major challenge for the whole telecom industry. The satisfaction of some of these criteria depends directly on the number of resources allocated to the network, including the capacity of the so-called base stations. The operator must have some quantitative means to anticipate demand and its impact on the design of its network. If we want to move beyond the phase of divination, then modelization is needed. This is about putting into equations, although sometimes with a kabbalistic aspect, the phenomenon which we want to study. To each situation may correspond several models depending on whether one is interested in the microscopic or macroscopic scale, the long or short time behavior, and so on. Ideally, the choice should be made only based on purpose but it is also conditioned by the technical and mathematical knowledge of the people who build the model.

Once the problem is raised, it must be solved: in other words, if numbers are given in input, then some numbers should pop up in output. Thanks to advances in computing, the situation has changed dramatically in the last twenty years. It is now possible to calculate quantities that are not only defined by explicit mathematical formulas, but that may result from more or less sophisticated algorithms.

A model is also often a support for simulation, in this way it creates an artificial simplification of reality. If this method gives very often only approximate results and is costly in computation time, it is also often the only possible.

We tried in this book to show for what purpose could stochastic models be used in telecommunications networks, with quantitative as well as qualitive points of view. We wanted to vary the possible approaches (discrete time, continuous time Markov chains or processes, recurrent sequences, spatial modeling) to allow the reader to proceed with his modelization works himself. We have not, far from it, addressed all themes and all the technicalities on which the researchers are currently working. In particular, we did not discuss fluid limits and Palm measures, but we hope that our readers can take the rich literature to extend their thinking. We have tried to be as complete as possible in the mathematical prerequisites. Proofs and results that are missing can be found easily in many books that appear in the references. To emphasize the computational aspects and to help our student readers, we have very often explained the algorithms that to be implemented in order to solve a particular problem. Languages such as Octave, Scilab or Scipy/Numpy (available through the SAGE platform) are particularly well suited to the vector computations that appear here and allow us to instantiate the algorithms described in a few lines only.

This book would not exist without the assistance of a considerable number of people. The first draft of this book is a handout from Telecom ParisTech written by L. Decreusefond, D. Kofman, H. Korezlioglu and S. Tohme. The introduction to the martingale theory owes much to a handout from A.S. Üstünel. We have tried as much as possible to present the underlying network protocols. It must be noted that the decryption of standards of thousands of pages and their translation into human language require much work and fine knowledge in a wide variety of disciplines, and as well as infinite patience. We wish to thank C. Rigault and especially P. Martins, without whom we would not know what POTS was and even less OFDMA.

We heartily thank N.Limnios, who offered us a beginning on this long-term venture, as well as our colleagues C. Graham, Ph. Robert and F. Baccelli, with whom we have had much interaction on these topics for several years. This book would not have been what it is without the inspiration born from reading their books on these topics.

A big thanks to our partners for having supported us at difficult times. Thanks to Adele for her help.

Our students or colleagues, E. Ferraz, I. Flint, P. Martins, A. Vergne, T. T. Vu have reviewed and amended all or part of this opus. We thank them for participating in an often thankless task. The residual errors are ours.

Paris, February 2012.

Chapter 1

Introduction

1.1. Traffic, load, Erlang, etc.

In electricity, we count the amps or volts; in meteorology, we measure the pressure; in telecommunications, we count the Erlangs.

The telephone came into existence in 1870. Most of the concepts and notations were derived during this period. Looking at a telephone connection over a time period of length T, we define its observed traffic flow as the percentage of time during which the connection is busy

$$\rho = \frac{\sum_i t_i}{T}.$$

A priori, traffic is a dimensionless quantity since it is the ratio of the occupation time to the total time. However, it still has a unit, Erlang, in remembrance of Erlang who, along with Palm, was one of the pioneers of the performance assessment of telephone networks. Therefore, a load of 1 Erlang corresponds to an always busy connection.

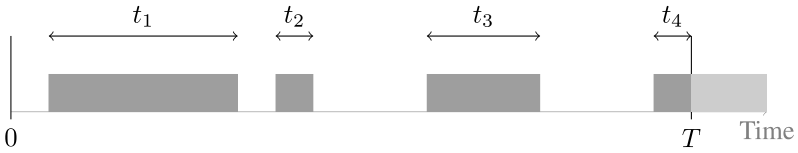

Figure 1.1. *Traffic of a connection: ratio of the occupation time to the total time of observation*

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

Looking at several connections, the traffic carried by this trunk is the sum of the traffic of each connection

$$\rho_{\text{trunk}} = \sum_{\text{connections}} \rho_{\text{connection}}.$$

This is no longer a percentage, but we can give a physical interpretation to this quantity according to the ergodic hypothesis. In fact, assume that the number of junctions is large, then we can calculate the average occupation rate in two different ways: either by calculating the percentage of the occupation time of a particular connection over a large period of time; or by computing the percentage of busy connections at a given time. In statistical physics, the ergodicity of a set of gas molecules implies that the spatial averages (for example, averages calculated on the set of gas molecules) are equal to time averages (i.e. averages calculated over a molecule for a long period of time). By analogy, we now assume that the same holds true for the occupation rate of telephone connections. We, therefore, have

$$p = \lim_{T\to\infty} \frac{1}{T} \sum_j t_j = \lim_{N\to\infty} \frac{1}{N} \sum_n X_n(t), \qquad [1.1]$$

where $X_n = 1$ if the junction n is busy at time t, $X_n = 0$, otherwise. Note that on the right-hand side, the value of t is arbitrary. This implies that we have implicitly assumed that the system is in steady state, that is statistically, its behavior does not change with time. When the number of junctions is large, it is unrealistic to try to define a structure of correlation between them. It is therefore reasonable to assume that a connection is free or busy, irrespective of the situation of other connections. Therefore, at a given time t, the number of busy connections follows a binomial distribution with parameters N (the total number of connections) and p (calculated by equation [1.1]). The average number of busy connections is Np at each moment.

This relation provides a simple and efficient way to estimate p. Telephone switches have among other functions to count the number of ongoing calls at each moment. By averaging this number over 15 seconds, we obtain a fairly accurate estimation of the average number of simultaneous calls, that is an estimation of p.

This raises a question: How to choose T and when to carry out the measurements? It is in fact clear that the traffic fluctuates throughout the day based on the human activities. For we want to reduce and ensure a low failure rates, it is necessary to consider the worst case and conduct measurements during heavy traffic periods. For generations, the observation period has been referred to as one hour and we look at the traffic at the busiest hour of the day.

Let us imagine for a moment that calls occur every $1/\lambda$ seconds and last exactly $1/\mu$ seconds with $\mu > \lambda$.

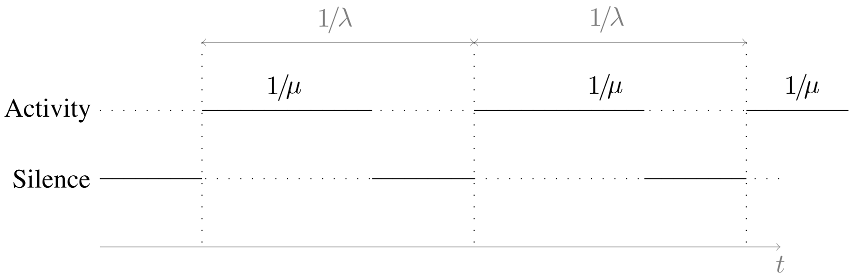

Figure 1.2. *Deterministic calls*

It is obvious that the number of calls between 0 and T is about λT and then the occupation rate of such a line is given by

$$\frac{1}{T}(\lambda T \times 1/\mu) = \lambda/\mu.$$

Of course, in reality, neither the inter-arrivals, nor the holding times are deterministic. Let us imagine a situation in which the holding times and the idle times are independent of each other with a distribution that is common to all busy periods and idle periods, respectively. Mathematically speaking, $(X_n, Y_n, n \geq 1)$ is a sequence of independent random variables. For any n, X_n has distribution $\mathbf{P}_X$ and Y_n has distribution $\mathbf{P}_Y$. Assume that these two distributions have finite moments of order 1 and note

$$1/\mu = \int y \, \mathrm{d}\, \mathbf{P}_Y(y), \; 1/\tau = \int x \, \mathrm{d}\, \mathbf{P}_X(x), \; \lambda = \frac{1}{1/\tau + 1/\mu}.$$

Set

$$T_0 = 0, \; T_n = T_{n-1} + X_n + Y_n, \; T'_n = T_n + X_n$$

and

$$X(t) = \begin{cases} 1 & \text{if } T'_n \leq t < T_{n+1}, \\ 0 & \text{if } T_n \leq t < T'_n. \end{cases}$$

Note that $\mathbf{E}\left[T_{n+1} - T_n\right] = \mathbf{E}\left[X_n + Y_n\right] = 1/\lambda$, so λ represents the average number of arrivals per unit time.

The theory of renewal process, or Little's formula (Chapter 8) show that we have the following limit

$$\frac{1}{T}\int_0^T \mathbf{1}_1(X(s)) \, \mathrm{d}\, s \xrightarrow{T\to\infty} \frac{1}{\mu}\frac{1}{1/\lambda} = \frac{\lambda}{\mu}. \qquad [1.2]$$

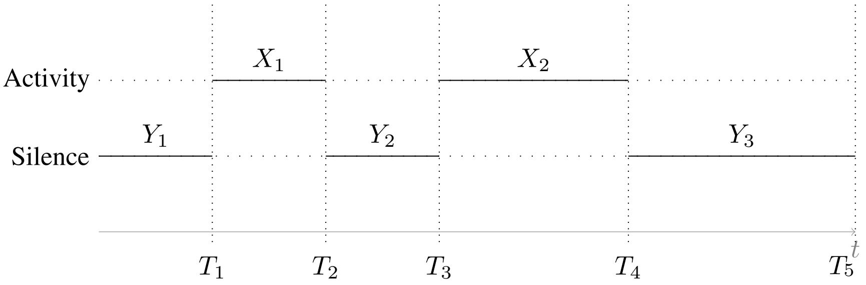

Figure 1.3. *Random calls*

We have shown in a particular but relatively general case that

$$\text{load} = \text{average number of calls per time unit} \times \text{average duration of a call.}$$

This simple formula is important since it allows us to switch to the world of the Internet. The ARPANET, remote ancestor of the Internet, was born in the 1970s following the works done for the U.S. army which required a distributed data transmission network more resistant against a timely attack. Unlike the telephone network where the resource, that is the telephone connection, is reserved for the duration of the communication, data networks are connectionless. The information is sent in packets of a few octets to which we add some identifiers, each following their own path in the intricacies of the network. The packet size, fixed or variable, large or small, is one of the issues to be resolved in such protocols. In this context, there is no notion of connection thus the concept of traffic must be redefined. The last equation [1.2] still has a meaning and it is this meaning that we will retain.

Volts and amps are nothing without Ohm's law, and meteorology is nothing without the equations of fluid mechanics. Erlangs are useless if we do not specify how the arrivals occur or how long the calls last. As demonstrated in Figure 1.4, the load is not sufficient to characterize the number of resources that are necessary to operate the system.

The situation was rather simple until the 1990s. As far as the telephone system is concerned, the process of call arrivals was modeled by a Poisson process (Chapter 6). This was justified by the statistical observations confirming it, and a well-known qualitative reasoning: each telephone subscriber has a low probability to call at a given time, but there are many (mathematically, an infinite number) subscribers. The approximation of a binomial distribution by a Poisson distribution justifies that at least at a given time, the number of simultaneous calls follows a Poisson distribution. Regarding the call duration, measurements on the switches proved that it could be considered to follow an exponential distribution with a mean of 3 minutes. Finally, the traffic generated by a subscriber was considered equal to 0.12 Erlang in the busy hour.

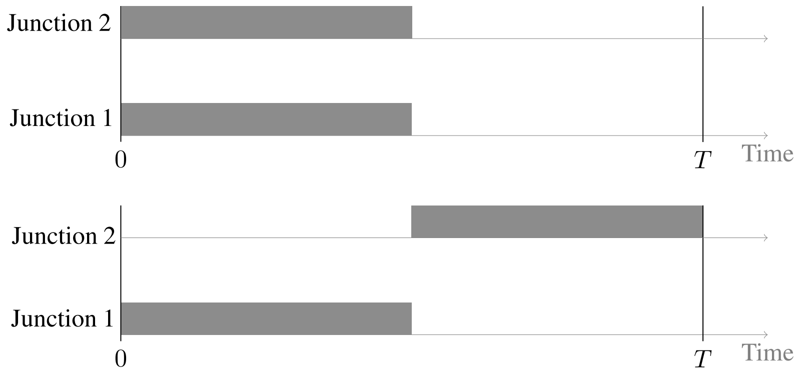

Figure 1.4. *Two systems are required to carry one Erlang. The first can be satisfied with one connection. In the second system, two connections are required*

In the case of data networks, despite serious doubts with regard to its validity, the packets were always supposed to arrive according to a Poisson process and their processing time assumed to follow an exponential distribution whose mean was dependent on the processing speed of the routers and their average length. And then boom! In the early 1990s, Bell Labs researchers showed in an intensive statistical campaign that we cannot possibly compare the traffic in a broadband network to Poisson traffic. In fact, if we consider the number of packets that arrive during 100 seconds, 10 seconds, . . . , 10 milli-seconds, we observe behaviors similar to that of Figure 1.5.

In the case of Poisson traffic, we observe behaviors that are visually similar to that of Figure 1.5 for small time scales, but when we agglomerate the received packets per period of 10 or 100 seconds, we mostly obtain a graph of the type of Figure 1.6.

This invariance of the number of packets at large time scale for the Poisson process is explained by theorem A.36. Indeed, according to this theorem

$$\frac{1}{\lambda t}(N(t) - \lambda t) \xrightarrow{t\to\infty} 0, \text{ that is } \frac{N(t)}{t} \xrightarrow{t\to\infty} \lambda.$$

At the speed at which the packets are sent, λ is the order of a thousands packets per second, hence 100 seconds can be considered as a “infinite” time and we obtain the number of packets sent per period of 100 seconds is almost constant. Consequently, the actual situation seems to be closer to that of a “fractal” system: the system retains the same shape at all time scales. The researchers at Bell Labs even proposed an alternative model in their paper, the fractional Brownian motion.

DEFINITION 1.1.– *A fractional Brownian motion of Hurst index H is a centered Gaussian process with covariance given by*

$$\mathbf{E}\left[B_H(t)B_H(s)\right] = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}).$$

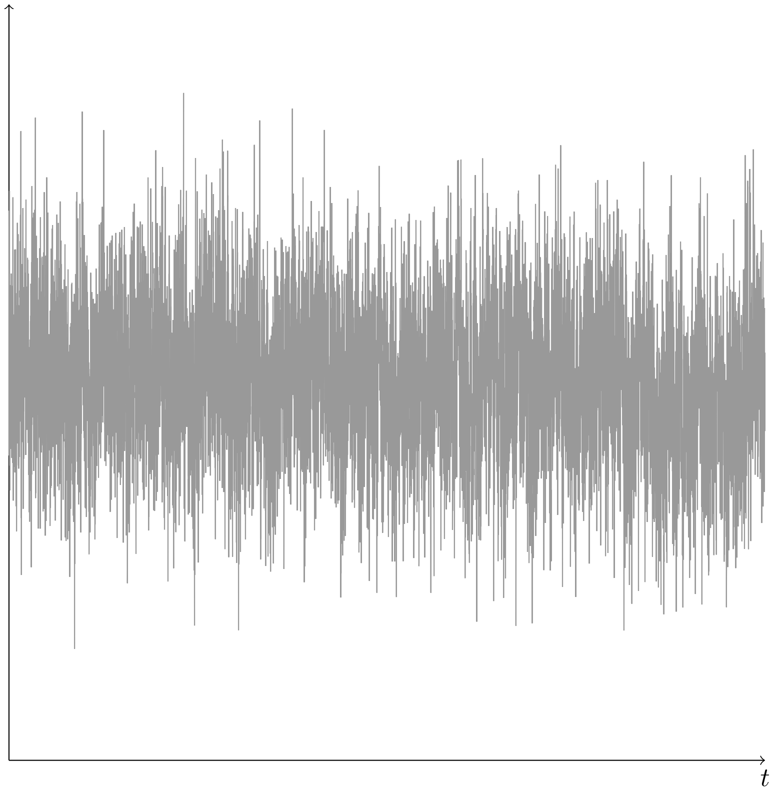

Figure 1.5. *At all observable time scales, after renormalization, the traffic recorded by time intervals resembles this one*

If $H = 1/2$, we get the ordinary Brownian motion, when $H > 1/2$, the increments are with positive covariance, if $H < 1/2$, they are with negative covariance: if the traffic tends to increase, immediately afterwards, it will tend to decrease.

From then on, the entire academic community began to ponder this question: What causes this fractal aspect, is the invariance true at all time scales, should the model be fractal or multi-fractal and most of all what is the impact of this form of traffic on the size of the queues? After 10 years of frantic research, we know how to explain the reasons behind the fractality, but we still do not know how to control it though it may have a major impact on the design.

To explain the fractality, it is enough to consider an emission schema for a source such as that of Figure 1.3: when $X = 1$, it implies that the source is emitting at its maximum speed, when $X = 0$, the source does not emit. Motivated by the statistical

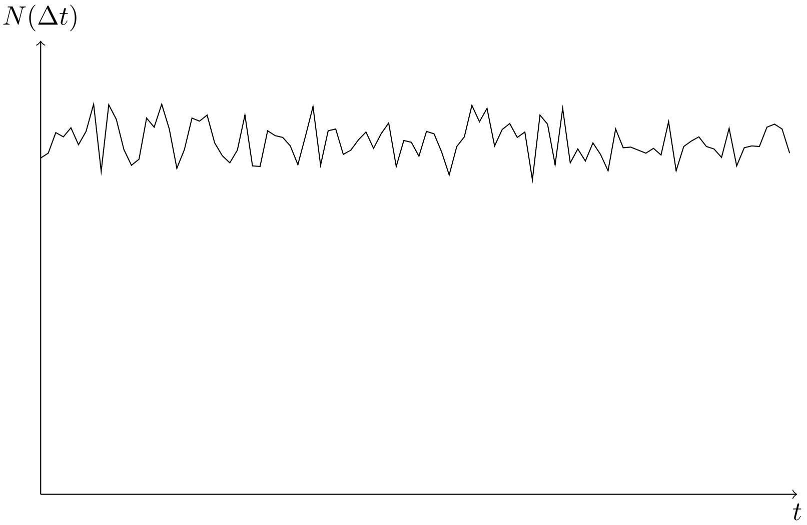

Figure 1.6. *For a Poisson traffic, the number of packets received per time interval becomes almost invariant when the time intervals are sufficiently large*

studies which prove that the length of the files available on the web has a "heavy-tailed" distribution, that is $\mathbf{P}(X > x) \sim x^{-\alpha}$ for $\alpha > 0$, as opposed to the exponential distribution where $\mathbf{P}(X > x) = \exp(-bx)$, the length of the emission period is assumed to follow a Pareto distribution, that is $\mathrm{d}\,\mathbf{P}_X(x) = cx^{-\alpha}\,\mathbf{1}_{[K,\,\infty[}(x)$, and the same is assumed for the idle period. When we superimposed many sources of this type and observe the steady state of this superposition, we find that the resulting process is exactly a fractional Brownian motion whose Hurst index depends on the powers appearing in the Pareto distributions.

In the end, all this matters only if we try to think in terms of packets. However, the current protocols mostly try to agglomerate the packets in flows (to prioritize some traffic for example) and thus virtually recreate the concept of connection specific to our plain old telephone. Under these conditions, only the arrival times and lifetimes of packets matter; however, these are less incorrectly modeled by a Poisson process and independent lifetimes with a heavy-tailed distribution. The Poisson process has a bright future...

1.2. Notations and nomenclature

$T_0 = 0 < T_1 < \dots T_n$ commonly denotes the arrival times of the customers (packets, sessions, calls, etc.) in the queuing system. The quantities $S_n = T_n - T_{n-1}$

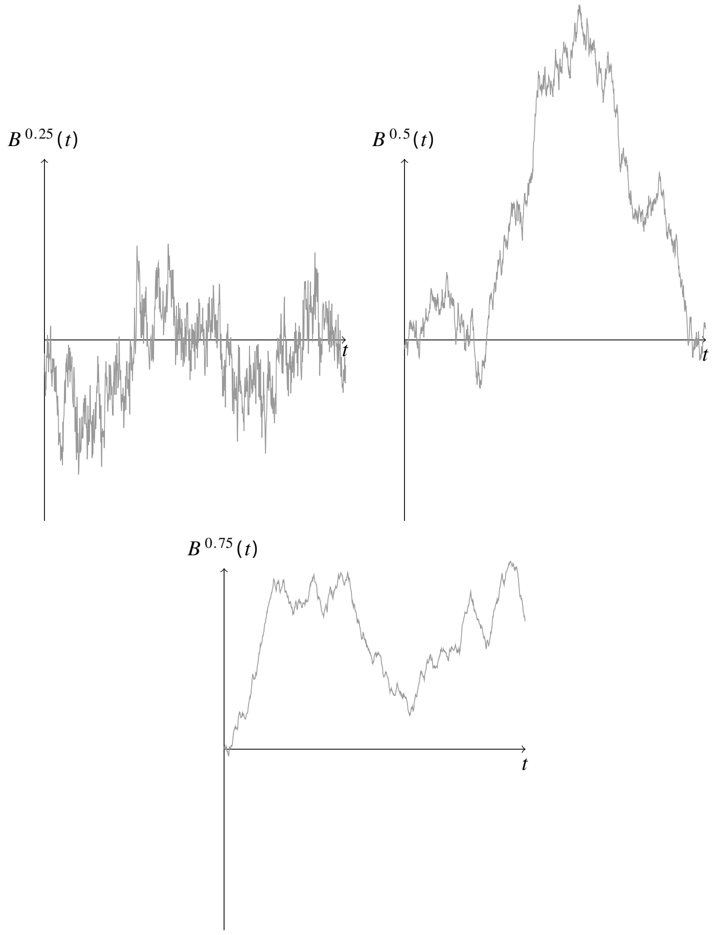

Figure 1.7. *Fractional Brownian motion for different values of H: from left to right $H = 0.2$; $H = 0.5$; $H = 0.8$. Lower the value of H more irregular are the trajectories*

are called inter-arrivals. The service time of the nth customer (processing time, call duration, etc.) is denoted by σ_n.

To distinguish different queues, we use Kendall's notation. A queue is a discrete event dynamic system described by five parameters: the statistical type of inter-arrivals, the statistical type of service time, the number of servers, the total number of resources (servers plus size of the waiting room), and service discipline. Implicitly, the inter-arrivals and service times are independent random variables.

For the first two points, the same abbreviations are used:

M$_\lambda$ to describe independent inter-arrivals (or service times) exponentially distributed of parameter λ.

GI to describe independent inter-arrivals (or service times) of the same distribution.

G to describe random inter-arrivals (or service times).

D$_\lambda$ to describe deterministic inter-arrivals (or service times) equal to λ.

The service discipline describes the order in which the customers are served:

FIFO or FCFS for First In First Out or First Come First Served.

LIFO LCFS or for Last In First Out. The last arrived customer is served first. Such a discipline may be preemptive if the new customer interrupts the current service or non-preemptive if otherwise. If preemptive, we can distinguish the preemptive resume case where the service of an interrupted customer picks up where it stopped, of the preemptive non-resume case where the service restarts at zero.

SRPT or Shortest Remaining Processing Time. The customer who has the lowest residual service time is served first. This discipline may be preemptive or non-preemptive.

EDF or Earliest Deadline First. Each customer has a marker of impatience. The customer with the lowest impatience is served first.

A discipline is said to be conservative when the input traffic is equal to the output traffic. Obviously, if the resources are finite, no discipline can be conservative (except in the deterministic case with traffic strictly less than 1). Even with infinite resources, a discipline is not necessarily conservative: in the EDF discipline, we can consider removing all the customers who are not served before their impatience marker; the non-preemptive resume disciplines are no longer conservative as there is more processed load than input load.

In the absence of information on the number of resources or the service discipline, it is understood that the number of resources is infinite and that the service discipline is the FIFO discipline.

EXAMPLE.– The M/M/1 queue is the queuing system where the inter-arrivals and the service times are independent of exponential distribution and there is one server. The waiting room is of infinite size and the service discipline is FIFO.

The GI/D/S/S+K queue is a queue with S servers, K places in the waiting room, deterministic service times, independent and identically distributed inter-arrivals.

1.3. Lindley and Beneš

1.3.1. *Discrete model*

We often consider the number of customers present in the system but the quantity that contains the most information is the system load, defined at each moment as the time required for the system to empty itself in the absence of new arrivals. The server works at unit speed: it serves a unit of work per unit time. Consequently, the load decreases with speed 1 between two arrivals. Figure 1.8 which represents the load over time depending on the arrivals and required service times is easily constructed.

DEFINITION 1.2.– *A busy period of a queue is a period that begins with the arrival of a customer in an empty system (server plus buffer) and ends with the end of a service after which the system is empty again.*

A cycle is a time period that begins with the arrival of a customer in an empty system and ends on the next arrival of a customer in an empty system. This is the concatenation of a busy period and an idle period, that is the time elapsed between the departure of the last customer of the busy period and the arrival of the next customer.

NOTE.– In Figure 1.8, a busy period begins at T_1 and ends at D_4. The corresponding cycle begins at T_1 and ends at T_5.

Note that as long as a service policy is conservative, the size of a busy period is independent of it: for waiting rooms of infinite size, the busy periods have, for example, the same length for the FIFO policy as that for the non-preemptive or preemptive resume LIFO policy.

Now let us consider the system load just before the arrival of the customer n. If W_{n-1} is the system load at the arrival of the customer $n-1$, it is increased by the load provided by the customer, that is σ_{n-1}, and reduced by the service time elapsed between the arrival times T_{n-1} and T_n, that is S_n exactly. We, therefore, have *a priori*

$$W_n = W_{n-1} + \sigma_{n-1} - S_n.$$

However, if $S_n > W_n + \sigma_{n-1}$, the inter-arrival is so large that the system has emptied, hence $W_n = 0$ and not $W_n < 0$. Consequently, the true formula known as Lindley's formula is given by

$$W_n = \max(W_{n-1} + \sigma_{n-1} - S_n, \, 0). \qquad [1.3]$$

Since the server works at unit speed, the load between T_n and T_{n+1} is given by

$$W(t) = \max(W_n + \sigma_n - (t - T_n), 0).$$

These two equations are used to easily simulate the load in any system, irrespective of the type of arrivals or service times or the service discipline as long as it is conservative.

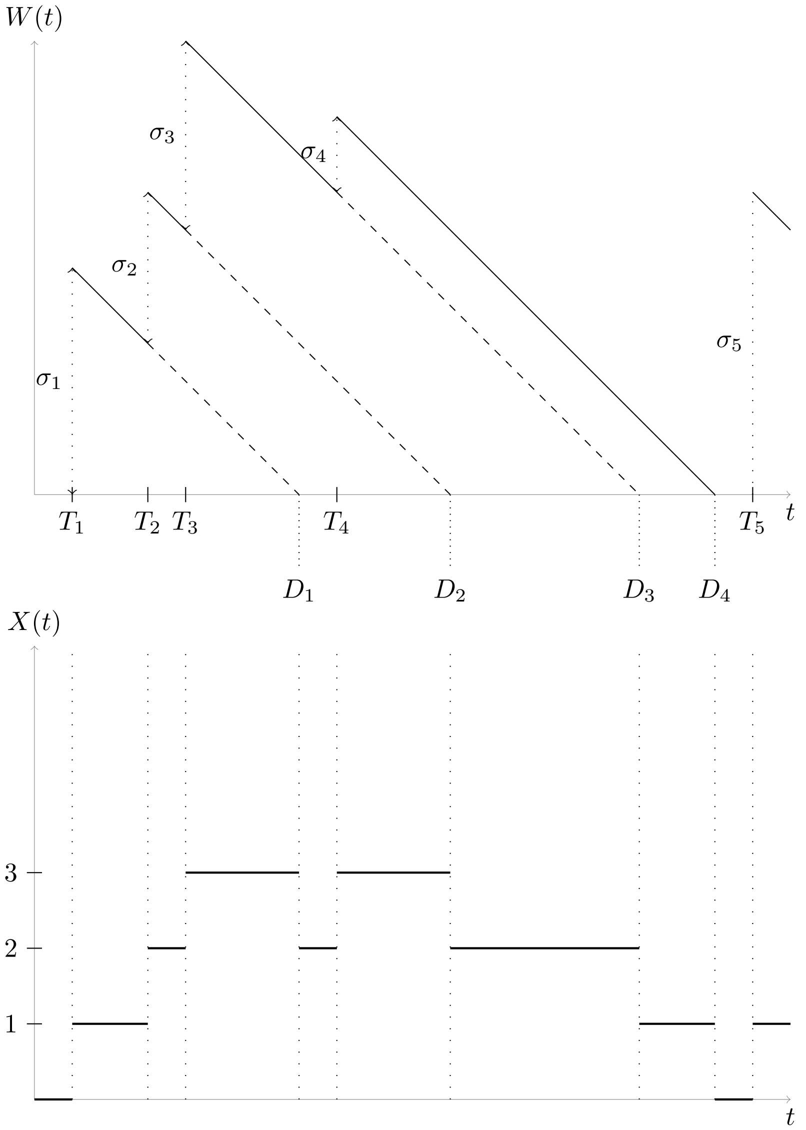

Figure 1.8. *The change in the system load with time. This graph is used to find the departure time in the case of a FIFO discipline, therefore to represent the change in the number of customers in the system*

They are also used to qualitatively analyze the stability of the system in very general cases. This will be dealt with in Chapter 4.

1.3.2. *Fluid model*

A fluid model consists of replacing a queue which is a discrete-time event system by a reservoir of infinite capacity which empties itself at unit speed and is fed by some continuous data flow. We can then obtain qualitative results on models whose study supports no other approaches. On the one hand, the method does not require precise knowledge about the rate of the input process, and on the other hand, it is particularly well adapted to the study of extreme cases: low and high loads, superposition of heterogeneous traffic.

We work in continuous time and we assume that all the processes are right-continuous with left limits. We denote:

1) $S(t)$: the total service time for the requests arrived up to time t;

2) $W(t)$: the virtual waiting time of a customer arriving at time t, that is the time that the customer must wait before starting to be served;

3) $X(t) = S(t) - t$.

As the system has no losses, we have

$$W(t) = X(t) - (t - \int_0^t \mathbf{1}_{\{0\}}(W(s))\,\mathrm{d}\,s). \qquad [1.4]$$

We will focus on showing an equivalent formulation of this equation.

THEOREM 1.1 (Beneš Equation).– *With the previous notations, we obtain the following identity: for $x \geq 0$,*

$$\mathbf{P}\left(W(t) < x\right) = \mathbf{P}\left(X(t) < x\right)$$
$$- \frac{\partial}{\partial x}\int_0^t \mathbf{P}\left(X(t) - X(u) < x \,|\, W(u) = 0\right)\mathbf{P}\left(W(u) = 0\right)\mathrm{d}\,u, \qquad [1.5]$$

and for $-t \leq x \leq 0$,

$$\mathbf{P}\left(X(t) < x\right) = \frac{\partial}{\partial x}\int_0^{t+x} \mathbf{P}\left(X(t) - X(u) < x \,|\, W(u) = 0\right)\mathbf{P}\left(W(u) = 0\right)\mathrm{d}\,u.$$

To go further, we will use the theory of reflection.

1.3.3. *Reflection problem*

DEFINITION 1.3.– *Let $(X(t)),\ t \geq 0)$ be a left-continuous process whose jumps are non-negative, the pair (W, L) solves the reflection equation associated with X if:*

1) $W(t) = X(t) + L(t)$, $\forall t \geq 0$*;*

2) $W(t) \geq 0, \quad \forall t \geq 0$*;*

3) L is a left-continuous, null at zero increasing process such that the measure $dL(s)(\omega)$ *is supported on the set* $\{s : W(s)(\omega) = 0\}$*, that is* L *increases only at moments where* W *is zero.*

THEOREM 1.2.– *The problem of reflection associated with* X *has a unique solution given by*

$$L(t) = \sup_{s \leq t} X(s)^-, \quad W(t) = X(t) + \sup_{s \leq t} X(s)^-$$

where $x^+ = \max(x, 0)$ *and* $x^- = \max(-x, 0)$.

Proof. If (W, L) and $(\tilde{W}, \tilde{L})$ are two solutions

$$\begin{aligned}(W(t) - \tilde{W}(t))^2 = (L(s) - \tilde{L}(s))^2 = 2\int_0^t (L(s) - \tilde{L}(s))d(L - \tilde{L})(s) \\ = 2\int_0^t (W(s) - \tilde{W}(s))d(L - \tilde{L})(s) \\ = -2\int_0^t (W(s)d\tilde{L}(s) + \tilde{W}(s)dL(s)) \leq 0,\end{aligned}$$

where we have successively used:

– the relation between W, X, and L ($\tilde{W}$, $\tilde{L}$, X, respectively);

– at fixed ω, the process $s \mapsto L(\omega, s)$ is an increasing process, therefore differentiable almost everywhere and whose derivative $dL(s)(\omega)$ is a non-negative measure. The process $L - \tilde{L}$ is of finite variation and so we can apply the formula of integration by parts A.13;

– L increases only at moments when W is zero, therefore "$W(s)dL(s)(\omega) = 0$";

– $\tilde{W}$ is a non-negative process and $dL(s)$ is a non-negative measure, therefore $\tilde{W}(s)dL(s) \geq 0$ and the same holds for the other term of the last integral.

Consequently, $W = \tilde{W}$ and uniqueness follows.

It is enough to check whether the process $\sup_{s \leq t} X(s)^-$ is suitable for L. Clearly, the L thus defined is an increasing process which is non-negative and null at 0. On the other hand

$$W(t) = X(t) + L(t) = X(t)^+ - X(t)^- + \sup_{s \leq t} X(s)^- \geq 0$$

We just have to see that L increases only in the set of zeros of W. Let T_0 be a point of increase of L, then for any non-negative h, there exists t_h such that $L_{t_0 - h} \leq X^-_{t_h} \leq L_{t_0}$.

When h tends toward 0, we have by left-continuity, $L_{t_0} = X_{t_0}^- = -X_{t_0}$, therefore $X_{t_0} + L_{t_0} = W_{t_0} = 0$. The proof is thus complete. □

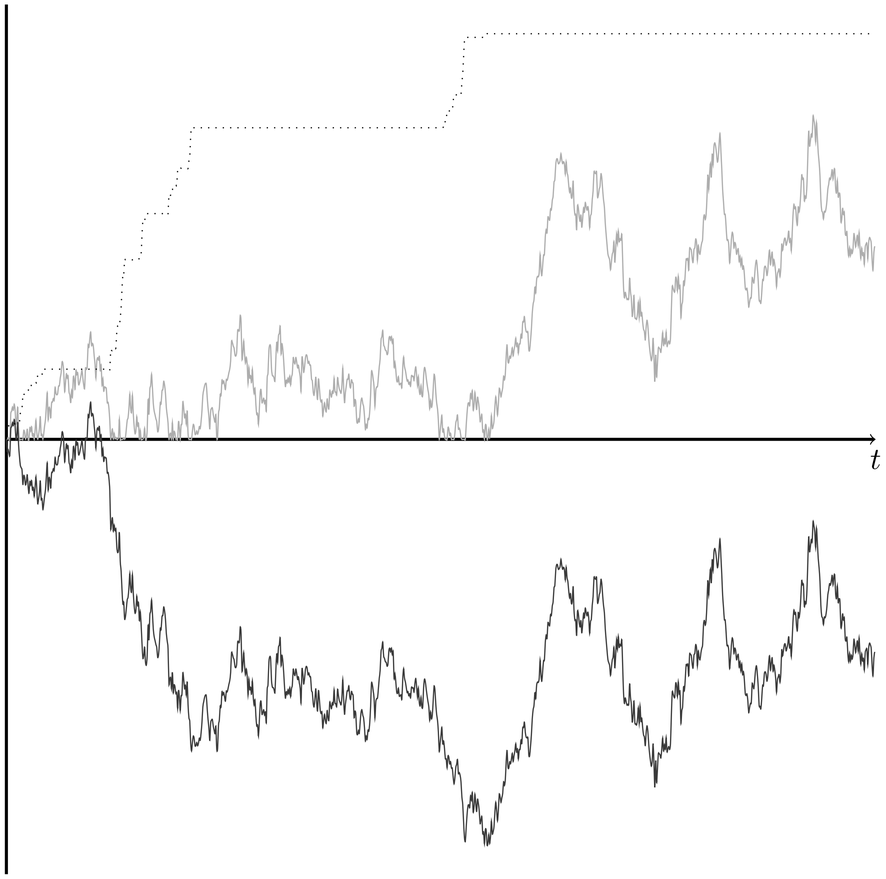

Figure 1.9. *An example of a reflected process. The dark color represents the input process X; dots represent the process L, and light color represents the process W*

COROLLARY 1.3.– *With the previous notations, we have the following identity*

$$\exp\left(-\lambda \int_0^t \mathbf{1}_{\{0\}}(W(s))\,\mathrm{d}\,s\right) = 1 - \lambda \int_0^t e^{\lambda X(s)} \mathbf{1}_{\{0\}}(W(s))\,\mathrm{d}\,s. \qquad [1.6]$$

Proof. From the relation $f(t) - f(0) = \int_0^t f'(u) \,\mathrm{d}\, u$, we deduce

$$\exp\left(-\lambda \int_0^t \mathbf{1}_{\{0\}}(W(s)) \,\mathrm{d}\, s\right) = 1 - \lambda \int_0^t e^{-\lambda \int_0^s \mathbf{1}_{\{0\}}(W(u)) \,\mathrm{d}\, u} \,\mathbf{1}_{\{0\}}(W(s)) \,\mathrm{d}\, s$$
$$= 1 - \lambda \int_0^t e^{-\lambda(W(s) - X(s))} \,\mathbf{1}_{\{0\}}(W(s)) \,\mathrm{d}\, s$$
$$= 1 - \lambda \int_0^t e^{\lambda X(s)} \,\mathbf{1}_{\{0\}}(W(s)) \,\mathrm{d}\, s.$$

Hence the result. □

We can now show an intermediate version of the Beneš equation which is also interesting in itself.

THEOREM 1.4.– *With the previous notations, we have the following identity*:

$$\mathbf{E}\left[f(W(t))\right] = \mathbf{E}\left[f(X(t))\right] + \mathbf{E}\left[\int_0^t f'(X(t) - X(u))\, \mathbf{1}_{\{0\}}(W(u)) \,\mathrm{d}\, u\right]$$
$$= \mathbf{E}\left[f(X(t))\right] + \int_0^t \mathbf{E}\left[f'(X(t) - X(u)) \,|\, W(u) = 0\right] \mathbf{P}\left(W(u) = 0\right) \mathrm{d}\, u. \quad [1.7]$$

Proof. By multiplying the two terms of equation [1.6] by $e^{-\lambda X(t)}$, we get

$$\mathbf{E}\left[e^{-\lambda W(t)}\right] = \mathbf{E}\left[e^{-\lambda X(t)}\right] - \lambda \int_0^t \mathbf{E}\left[e^{-\lambda(X(t) - X(u))}\, \mathbf{1}_{\{0\}}(W(u)) \,\mathrm{d}\, u\right].$$

More generally, for any function which is twice differentiable and bounded, we obtain equation [1.7]. □

Now we can give an idea on the proof of the Beneš equation [1.5].

NOTE.– $\mathbf{P}\left(X(t) - X(u) < x | W(u) = 0\right) = 0$ for $t + x \leq u \leq t$, since on $(W(u) = 0)$,

$$X(t) - X(u) = W(t) - L(t) + L(u) \geq -(L(t) - L(u))$$

however, $L(t) \leq t$, thus $X(t) - X(u) \geq -(t - u)$. Therefore, $X(t) - X(u)$ cannot be smaller than x if x itself is less than $u - t$, that is $u \geq t + x$.

Proof. Proof of the Beneš equation. By multiplying the terms of [1.5] by $e^{-\lambda x}$, then by integrating it from $-t$ to $+\infty$, it appears that equation [1.5] is equivalent to:

$$\begin{aligned}\int_{-t}^{+\infty} e^{-\lambda x}\mathbf{P}\left(W(t) < x\right)\mathrm{d}\,x &= \int_{-t}^{+\infty} e^{-\lambda x}\mathbf{P}\left(X(t) < x\right)\mathrm{d}\,x \\ &- \int_{-t}^{+\infty} e^{-\lambda x}\frac{\partial}{\partial x}\int_{0}^{t+x}\mathbf{P}\left(X(t)-X(u) < x|W(u)=0\right)\mathbf{P}\left(W(u)=0\right)\mathrm{d}\,u\ \mathrm{d}\,x,\end{aligned}$$

according to the note above. By using the formula

$$\int_{0}^{+\infty} e^{-\lambda x}\mathbf{P}\left(X < x\right)\mathrm{d}\,x = \frac{1}{\lambda}\mathbf{E}\left[e^{-\lambda X}\right]$$

and an integration by parts, we obtain

$$\begin{aligned}\mathbf{E}\left[e^{-\lambda W(t)}\right] &= \mathbf{E}\left[e^{-\lambda X(t)}\right] \\ &\quad - \lambda\frac{\partial}{\partial x}\int_{0}^{t}\mathbf{E}\left[e^{-\lambda(X(t)-X(u))}|W(u)=0\right]\mathbf{E}\left[\mathbf{1}_{\{0\}}(W(u))\right]\mathrm{d}\,u \\ &= \mathbf{E}\left[e^{-\lambda X(t)}\right] \\ &\quad - \lambda\left[e^{-\lambda x}\int_{0}^{t+x}\mathbf{P}\left(X(t)-X(u) < x|W(u)=0\right)\mathbf{P}\left(W(u)=0\right)\mathrm{d}\,u\right]_{x=-t}^{x=+\infty} \\ &\quad + \lambda^2\int_{-t}^{+\infty} e^{-\lambda x}\int_{0}^{t+x}\mathbf{P}\left(X(t)-X(u) < x|W(u)=0\right)\mathbf{P}\left(W(u)=0\right)\mathrm{d}\,u\ \mathrm{d}\,x \\ &= \mathbf{E}\left[e^{-\lambda X(t)}\right] \\ &\quad + \lambda^2\int_{0}^{+\infty}\mathbf{P}\left(W(u)=0\right)\int_{u-t}^{+\infty}\mathbf{P}\left(X(t)-X(u) < x|W(u)=0\right)\mathrm{d}\,x\ \mathrm{d}\,u.\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{E}\left[e^{-\lambda W(t)}\right] &= \mathbf{E}\left[e^{-\lambda X(t)}\right] \\ &\quad + \lambda^2\int_{0}^{+\infty}\mathbf{P}\left(W(u)=0\right)\int_{0}^{+\infty}\mathbf{P}\left(X(t)-X(u) < x|W(u)=0\right)\mathrm{d}\,x\ \mathrm{d}\,u \\ &= \mathbf{E}\left[e^{-\lambda X(t)}\right] \\ &\quad + \lambda\int_{0}^{+\infty}\mathbf{P}\left(W(u)=0\right)\int_{0}^{+\infty}\mathbf{E}\left[e^{-}\lambda(X(t)-X(u))|W(u)=0\right]\mathrm{d}\,x\ \mathrm{d}\,u,\end{aligned}$$

always according to the initial note. □

The representation of the workload in the form of a reflected process constitutes the basis of many convergence theorems. This formula also helps us to study the significance of the long memory in the transmission delay. Other Markovian methods could not process this situation.

There is a particular class of processes for which we can deduce complete results:

DEFINITION 1.4.– *A process X is said to be with independent increments if and only if, for any $0 \leq t_1 < \cdots < t_n$, the random variables $X(t_1),\ X(t_2)-X(t_1),\ \ldots,\ X(t_n)-X(t_{n-1})$ are independent. The process is said to have homogeneous increments if for any pair (t, s) of non-negative real numbers, the distribution of $X(t+s)-X(t)$ is that of $X(s)$.*

THEOREM 1.5.– *If $X(t) = S(t) - t$ is a process with homogeneous independent increments and $(W,\ L)$ is the solution of the reflection problem associated with X, we have*

$$\sigma \int_0^{+\infty} e^{-\sigma t} \mathbf{E}\left[e^{-\lambda W(t) - \mu L(t)}\right] \mathrm{d}\,t = \Psi_\sigma^-(-\lambda)\Psi_\sigma^+(-\mu) \qquad [1.8]$$

where:

– Ψ is the Lévy-Khintchine function of X defined by $\mathbf{E}\left[e^{sX(t)}\right] = e^{t\psi(s)}$ which can be written as

$$\Psi(s) = as + \frac{1}{2}\sigma_0^2 s^2 + \int_{|x|<1} (e^{sx} - 1 - sx)\,\mathrm{d}\,\Pi(x) + \int_{|x|\geq 1} (e^{sx} - 1)\,\mathrm{d}\,\Pi(x)$$

for a measure Π integrating $x^2 \wedge 1$.

– η is a Lévy-Khintchine function such that $\psi(\eta(s)) = s$.

$$\begin{cases} \Psi_\sigma^+(s) & = & \dfrac{1}{1 - \dfrac{s}{\eta(s)}} \\ \Psi_\sigma^-(s) & = & \left(1 - \dfrac{s}{\eta(s)}\right)\left(\dfrac{\sigma}{\sigma - \Psi(s)}\right) \end{cases}$$

This result is complementary to the future results on the M/GI/1 queue (Chapter 5) since such a queue is represented by a fluid model by considering

$$S(t) = \sum_{T_n \leq t} Y_n$$

where $(T_n,\ n \geq 1)$ is the sequence of arrival times distributed according to a Poisson process and $(Y_n,\ n \geq 1)$ is a sequence of independent random variables that are identically distributed.

1.4. Notes and comments

For more details on the design of telephone networks, their history and future developments, we may refer to [RIG 98]. The original paper by Bell Labs refers to [LEL 94], the studies which show that files have a length that follows a Pareto distribution are available in [CRO 96]. The mathematical explanation of auto-similarity is available in [SHE 97]. Norros is the first to have studied the impact of auto-similarity in networks which can be referred to in [NOR 94]. Many results on statistical estimation, as well as on the use in networks or more mathematical aspects on long-memory processes may be found in [DOU 02].

Epitome

– The load is also called traffic. It is defined as the average number of calls per unit of time multiplied by the mean processing time of a call. Its unit is Erlang.

– Kendall's notation is used to describe the different queues. The M/M/1 queue is the queue where the inter-arrivals and the service times are independent and follow exponential distributions. There is only one server, the queue is infinite, and the service policy is FIFO.

– We know how to qualitatively examine the M/M /*/*/ FIFO queues, and to a lesser extent, the M/GI/1 and GI/M/1 queues. For the others (other distributions of inter-arrivals or service times, other disciplines), we often have only partial or asymptotic results.

PART 1

Discrete-time Modeling

Chapter 2

Stochastic Recursive Sequences

The modeling of discrete-time deterministic dynamical systems is based on recursive sequences of the form $u_{n+1} = f(u_n)$. One addresses the question of convergence of the sequence as n goes to infinity, and the value of the limit which, assuming that f is continuous, is necessarily the solution of the equation $l = f(l)$.

The purpose of this chapter is to develop the tools that will enable us to answer such questions for stochastic recursive sequences.

For example, let us consider a G/G/1 queue (which will be dealt with in section 4.1). Denoting $(\xi_n,\, n \in \mathbf{N})$ the sequence of inter-arrival times and $(\sigma_n,\, n \in \mathbf{N})$ the sequence of service times, the workload W_{n+1} of the server at the arrival of the $n+1$th customer is deduced from the workload at the arrival of the nth customer by Lindley's equation

$$W_{n+1} = [W_n + \sigma_n - \xi_n]^+ . \qquad [2.1]$$

If the two sequences are independent (GI/GI/1 queue), then the sequence $(W_n,\, n \in \mathbf{N})$ is a Markov chain with values in the uncountable space $\mathbf{R}^+$. Its analysis is impossible with the tools of Chapter 3 since we restrict it to Markov chains having finite or countable state space.

Obviously, there can be no almost sure convergence of the sequence $(W_n,\, n \in \mathbf{N})$ toward the same limit, but we can expect a convergence "in distribution" that is $\mathbf{P}(W_n \in [a,\, b]) \xrightarrow{n\to\infty} \mathbf{P}(W_\infty \in [a,\, b])$ for a random variable W_∞ whose distribution must be determined. We then say that the sequence converges toward its steady state. It is then remarkable that by properly choosing the probability space, we can write a deterministic equation, similar to the equation $l = f(l)$, which is solved by the stationary distribution.

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

More generally, the asymptotic study is essentially based on the properties of the recurrence function (monotonicity, continuity, etc.), on criteria of comparison with other sequences and on the resolution, in a stochastic frame, of a fixed point-type limiting equation (see equation [2.7]).

This chapter is therefore mainly theoretical, but introduces the necessary tools for the study of the stability of queues, under the most general hypothesis.

2.1. Canonical space

The concept of stationarity implies invariance in time, that is : a shift in time does not change the global picture. If the idea is easily understood, its formalization quickly clouds the basic concept.

Let us consider the set $F^{\mathbf{N}}$ of sequences of elements of a set F. The shift operator θ on $F^{\mathbf{N}}$ is then defined by

$$\theta\colon \begin{cases} F^{\mathbf{N}} & \longrightarrow F^{\mathbf{N}} \\ (\omega_n,\, n \geq 0) & \longmapsto (\omega_{n+1},\, n \geq 0) = (\omega_n,\, n \geq 1). \end{cases}$$

Defined in this way, this operator has the drawback of not being bijective: if we consider a sequence $\beta = (\beta_n,\, n \geq 0)$, all the sequences obtained by concatenation of any element of F and β are mapped onto β by θ. To overcome this problem, it is customary to work with sequences indexed by $\mathbf{Z}$ and not by $\mathbf{N}$. This change has no crucial mathematical consequence, as the indexation space remains countable. Philosophically, however, it implies that there is no more origin of time...

The shift operator is thus defined on $F^{\mathbf{Z}}$ by

$$\theta(\omega_n,\, n \in \mathbf{Z}) = (\omega_{n+1},\, n \in \mathbf{Z})$$

and thus becomes bijective!

Now, let us suppose that F is a Polish space, and thereby that $F^{\mathbf{Z}}$ is Polish. It can therefore be equipped with its Borel sigma-field $\mathfrak{B}(F^{\mathbf{Z}})$. Throughout this chapter, the canonical space will be $\Omega = (F^{\mathbf{Z}},\, \mathfrak{B}(F^{\mathbf{Z}}))$. For $n \in \mathbf{Z}$, X_n denotes the "nth coordinate" map

$$X_n\colon \begin{cases} \Omega = F^{\mathbf{Z}} & \longrightarrow F \\ \omega = (\omega_n,\, n \in \mathbf{Z}) & \longmapsto \omega_n. \end{cases}$$

Let us notice the following identity

$$X_k = X_0 \circ \theta^k, \text{ for any } k \in \mathbf{Z}. \qquad [2.2]$$

DEFINITION 2.1.– *Let* $(E, \mathcal{E}, \mathbf{P})$ *be a probability space and* ψ *be a measurable mapping from* $(E, \mathcal{E})$ *to* $(F, \mathcal{F})$*. We denote* $\psi^*\mathbf{P}$ *the image measure of* $\mathbf{P}$ *by* ψ*, that is*

$$\forall A \in \mathcal{F},\ (\psi^*\mathbf{P})(A) = \mathbf{P}(\psi^{-1}(A)),$$

where $\psi^1(A) = \{x \in E,\ \psi(x) \in A\}$.

DEFINITION 2.2.– *A probability* $\mathbf{P}$ *on* Ω *is said stationary if for any* $A \in \mathfrak{B}(F^{\mathbf{Z}})$,

$$\mathbf{P}(A) = \mathbf{P}(\theta^{-1}A).$$

In an equivalent manner, we have $\theta^*\mathbf{P} = \mathbf{P}$.

Particularly, if one considers events of the form

$$A = (X_{k_1} \in A_1, \cdots, X_{k_n} \in A_n),$$

we deduce that

$$\mathbf{P}(X_{k_1} \in A_1, \cdots, X_{k_n} \in A_n) = \mathbf{P}(X_{k_1-1} \in A_1, \cdots, X_{k_n-1} \in A_n). \quad [2.3]$$

A sequence of random variables satisfying [2.3] for any n, all $k_1, \cdots, k_n$ and all $A_1, \cdots, A_n \in \mathfrak{B}(F)$, will be said to be stationary. In particular, if the canonical space $\Omega = F^{\mathbf{Z}}$ is equipped with a stationary probability, the sequence of “coordinate” maps is a stationary sequence of random variables.

NOTE.– Conversely, if the sequence $(\alpha_n,\ n \in \mathbf{Z})$, defined on a probability space $(\tilde{\Omega}, \tilde{\mathbf{P}})$, is stationary, we can consider $\mathbf{P}_\alpha$ its distribution on $F^{\mathbf{Z}}$, that is the image measure of $\tilde{\mathbf{P}}$ by the mapping

$$\begin{cases} \tilde{\Omega} & \longrightarrow F^{\mathbf{Z}} \\ \tilde{\omega} & \longmapsto (\alpha_n(\tilde{\omega}),\ n \in \mathbf{Z}). \end{cases}$$

It is then easily seen that $\mathbf{P}_\alpha$ is a stationary measure. In fact, it suffices to check [2.3] for $\mathbf{P} = \mathbf{P}_\alpha$. But this is true since

$$\mathbf{P}_\alpha(X_{k_1} \in A_1, \cdots, X_{k_n} \in A_n) = \tilde{\mathbf{P}}(\alpha_{k_1} \in A_1, \cdots, \alpha_{k_n} \in A_n).$$

Henceforth, all the stationary sequences will be seen as a stationary probability on $F^{\mathbf{Z}}$.

NOTE.– Constructing a stationary sequence is not easy in general. The simplest example is that of independent and identically distributed random variables, as in this case,

$$\tilde{\mathbf{P}}(\alpha_{k_1} \in A_1, \ldots, \alpha_{k_n} \in A_n) = \prod_{j=1}^{n} \mathbf{P}_{\alpha_0}(A_j),$$

a quantity which does not depend on $(k_1, \ldots, k_n)$.

Another construction of stationary sequences can be obtained from irreducible positive recurrent Markov chains (see Chapter 3) . Let $\tilde{X}$ be a Markov chain on E with transition operator Q and invariant probability π. We denote $\tilde{\mathbf{P}}$ the distribution of $\tilde{X}$ on $E^{\mathbf{N}}$ when π is the distribution of X_0. We know from Kolmogorov's Lemma that to define a probability on $E^{\mathbf{Z}}$, it suffices to define the finite-dimensional distribution. Now, by setting for any n-tuple of relative integers $k_1 < k_2 < \cdots < k_n$ and all $A_1, \ldots, A_n \subset E$,

$$\mathbf{P}(X_{k_1} \in A_1, \ldots, X_{k_n} \in A_n) = \tilde{\mathbf{P}}(\tilde{X}_0 \in A_1, \ldots, \tilde{X}_{k_n - k_1} \in A_n),$$

we define in fact the finite-dimensional marginals of a unique probability measure on $E^{\mathbf{Z}}$. The stationarity of $\mathbf{P}$ thus defined is straightforward.

Let us define the quadruple $\mathfrak{O} = (\Omega = F^{\mathbf{Z}}, \mathcal{F} = \mathfrak{B}(F^{\mathbf{Z}}), \mathbf{P}, \theta)$.

DEFINITION 2.3.– *A probability measure* $\mathbf{P}$ *on* $F^{\mathbf{Z}}$ *is said to be ergodic if*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Phi \circ \theta^{i} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Phi \circ \theta^{-i} = \mathbf{E}_{\mathbf{P}}\left[\Phi\right], \mathbf{P} - a.s. \qquad [2.4]$$

for any function $\Phi \in L^1(\mathbf{P})$. *The quadruple* $\mathfrak{O}$ *will then be said to be ergodic. A stationary sequence* $(\alpha_n,\, n \in \mathbf{Z})$ *is said to be ergodic if its distribution induces an ergodic measure on* $F^{\mathbf{Z}}$.

EXAMPLE 2.1.– Let a and b be two real numbers. The random sequence $(\alpha_n,\, n \in \mathbf{Z})$ equal to $a, b, a, b, \ldots$ with probability 1/2 and $b, a, b, a, \ldots$ with probability 1/2 is stationary (we have $\mathbf{P}\left(\alpha_n = a\right) = \mathbf{P}\left(\alpha_n = b\right) = 1/2$ for all n) and ergodic ([2.4] is clearly verified).

EXAMPLE 2.2.– The sequence $(\beta_n,\, n \in \mathbf{Z})$ equal to $a, a, ..., a$ with probability 1/2 and $b, b, ..., b$ with probability 1/2 is also stationary. On the other hand, it is easily checked that it is not ergodic since [2.4] does not hold e.g. for $\Phi = \mathbf{1}_{\{a\}}$.

LEMMA 2.1.– *Let* $\mathfrak{O}$ *be an ergodic quadruple. Any event* $A \in \mathfrak{B}(F^{\mathbf{Z}})$ *such that* $A = \theta^{-1}A$ *is trivial:* $\mathbf{P}\left(A\right) = 0$ *or* 1.

Proof. For any integer $n \in \mathbf{N}$, we define $\mathcal{F}_n^0 = \sigma\{X_k,\, k \le n\}$, where X_n is the nth coordinate map, and $W_n = \mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n^0\right]$. The sequence W is clearly a uniformly bounded martingale, which thus converges a.s. and in L^1 to $\mathbf{1}_A$. Temporarily let us assume that $W_n = W_0 \circ \theta^n$. If a sequence $(u_n,\ n \in \mathbf{N})$ converges to a limit, then its Cesaro averages also converge to the same limit. Thus,

$$\frac{1}{n}\sum_{k=1}^{n} W_k \xrightarrow{n\to\infty} \mathbf{1}_A \, .$$

On the other hand,

$$\frac{1}{n}\sum_{k=1}^{n} W_k = \frac{1}{n}\sum_{k=1}^{n} W_0 \circ \theta^k \xrightarrow{n\to\infty} \mathbf{E}\left[W_0\right],$$

according to the hypothesis [2.4]. It follows that the random variable $\mathbf{1}_A$ is constant, hence the result.

It thus remains to prove that $W_n = W_0 \circ \theta^n$ or in an equivalent manner, $\mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n^0\right] = \mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_{n-1}^0\right] \circ \theta$ for all n. First let us observe that if ϕ is $\mathcal{F}_n^0$-measurable, then $\phi \circ \theta^{-1}$ is $\mathcal{F}_{n-1}^0$-measurable. In addition, $\theta^{-1}A = A$ is equivalent to $\mathbf{1}_A \circ\theta = \mathbf{1}_A$. By definition of W_n, for all $\mathcal{F}_n^0$-measurable and bounded ϕ, we have

$$\begin{aligned}\mathbf{E}\left[\phi W_n\right] &= \mathbf{E}\left[\phi\ \mathbf{1}_A\right]\\ &= \mathbf{E}\left[\phi \circ \theta^{-1} \circ \theta\ \mathbf{1}_A \circ\theta\right]\\ &= \mathbf{E}\left[\phi \circ \theta^{-1}\ \mathbf{1}_A\right]\\ &= \mathbf{E}\left[\phi \circ \theta^{-1}\ W_{n-1}\right]\\ &= \mathbf{E}\left[\phi \circ \theta^{-1}\ W_{n-1} \circ \theta \circ \theta^{-1}\right]\\ &= \mathbf{E}\left[\phi\ W_{n-1} \circ \theta\right],\end{aligned}$$

where we have twice used the invariance of $\mathbf{P}$ by θ, that is the stationarity. By identification, we deduce that $W_n = W_0 \circ \theta^n$. □

The following result will be frequently used in Chapter 4.

LEMMA 2.2 (ERGODIC LEMMA).– *Let Y be a random variable defined on the stationary ergodic quadruple $\mathfrak{Q}$, $\mathbf{P}$-a.s. positive and such that $Y \circ \theta - Y$ is integrable. Then,*

$$\mathbf{E}\left[Y \circ \theta - Y\right] = 0.$$

Proof. For all $n \in \mathbf{N}$, the v.a. $Y \wedge n$ is integrable and thus

$$\mathbf{E}\left[(Y \wedge n) \circ \theta - Y \wedge n\right] = \mathbf{E}\left[(Y \wedge n) \circ \theta\right] - \mathbf{E}\left[Y \wedge n\right] = 0. \qquad [2.5]$$

The sequence $((Y \wedge n) \circ \theta - Y \wedge n,\ n \in \mathbf{N})$ converges $\mathbf{P}$-p.s. to $Y \circ \theta - Y$, and it is easy to see that for all n,

$$| (Y \wedge n) \circ \theta - Y \wedge n | = | (Y \circ \theta) \wedge n - Y \wedge n | \leq | Y \circ \theta - Y | .$$

The dominated convergence theorem thus implies that the null expectations of [2.5] tend to $\mathbf{E}\,[Y \circ \theta - Y]$, which concludes the proof. □

It is important to notice (and this is its essential purpose) that the latter result holds true even if Y is not assumed as integrable.

DEFINITION 2.4.– *A stochastic recursive sequence (SRS for short)* $(W_n,\ n \in \mathbf{N})$ *with values in the Polish space E, is defined on a stationary ergodic quadruple by a random variable* Y *valued in* E*, a measurable mapping* φ *from* $E \times F$ *to* E *and the relations*

$$W_0 = Y \text{ and } W_{n+1} = \varphi(W_n,\ X_n) \text{ for } n \geq 1. \qquad [2.6]$$

We then say that the SRS $(W_n,\ n \in \mathbf{N})$ *is driven by* φ *and descends from* Y*. It is often denoted* $\left(W_n^Y,\ n \in \mathbf{N}\right)$ *to emphasize the dependence on the initial condition* Y*.*

EXAMPLE 2.3.– The sequence $(W_n,\ n \in \mathbf{N})$ of the workload of the G/G/1 queue mentioned in the introduction of this chapter has such a form: we set for all n, $X_n = (\sigma_n, \xi_n)$, $E = \mathbf{R}^+$, $F = \mathbf{R}^+ \times \mathbf{R}^+$ and

$$\varphi : \begin{cases} E \times F & \longrightarrow E \\ (x,\ (y,\ z)) & \longmapsto [x + y - z]^+. \end{cases}$$

Hence there are two sources of randomness in the evolution of W: the initial condition Y and the "stimulus" represented by the sequence X. The probability space on which M is defined is therefore

$$\left(E \times F^{\mathbf{Z}},\ \mathfrak{B}(E) \otimes \mathfrak{B}(F^{\mathbf{Z}})\right).$$

As a random variable, W takes values in $E^{\mathbf{N}}$. The law of the sequence W thus defines a probability on $E^{\mathbf{N}}$. This space is also equipped with a shift θ_E defined in a similar way to that of $F^{\mathbf{Z}}$, which we temporarily note θ_F. The shift θ_E is not bijective, but we will not need this property for the time being. The definitions of stationarity and of ergodicity remain valid to the identical.

The stimulus is given by the model, hence we cannot do anything but to act on the initial condition. The question is to know whether one can choose Y as a stimulus function so that the distribution of W is stationary. A sufficient condition is provided by the following theorem. It transforms an identity in distribution in to a trajectorial identity which we hope is easier to prove.

THEOREM 2.3.– *If there is a random variable Y such that*

$$Y \circ \theta_F = \varphi(Y,\, X_0),\ \mathbf{P}_X -\ a.s., \tag{2.7}$$

then the SRS W defined by [2.6] admits a stationary probability.

Proof. After introducing some notations, the result is straightforward. Let us introduce the mappings

$$Y\colon \begin{cases} F^{\mathbf{Z}} & \longrightarrow F^{\mathbf{Z}} \times E \\ \omega & \longmapsto (\omega, Y(\omega)) \end{cases}$$

and

$$W\colon \begin{cases} F^{\mathbf{Z}} \times E & \longrightarrow E^{\mathbf{N}} \\ (\omega, \eta) & \longmapsto (\eta, \varphi(\eta, \omega), \ldots). \end{cases}$$

Hence we have the following diagram

$$\begin{tikzcd} F^{\mathbf{Z}} \arrow[r, "W\circ Y"] \arrow[d, "\theta_F"'] & E^{\mathbf{N}} \arrow[d, "\theta_E"] \\ F^{\mathbf{Z}} \arrow[r, "W\circ Y"'] & E^{\mathbf{N}} \end{tikzcd}$$

The nth component of $W \circ Y \circ \theta_F(\omega)$ is $W_n(Y(\theta_F\omega), \theta_F\omega)$, and that of $\theta_E \circ M \circ Y(\omega)$ is $W_{n+1}(Y(\omega), \omega)$. In particular for $n = 0$, on the one hand we have $Y \circ \theta_F(\omega)$ and on the other $\varphi(Y(\omega), \omega_0)$, then equation [2.7] means that these two quantities are equal. We deduce by induction that it is also the case for all the components, so $W \circ Y \circ \theta_F = \theta_E \circ M \circ Y$. In mathematical terms, we say that the diagram is “commutative”.

From this we can deduce that the image measures of $\mathbf{P}_X$ by these two mappings are identical. Let us note $\mathbf{P}$ as the law of W, that is the image measure of $\mathbf{P}_X$ by $W \circ Y$. On the one hand we have

$$(\theta_E \circ W \circ Y)^*\mathbf{P}_X = \theta_E^*\mathbf{P}$$

and on the other hand, as $\theta_F^*\mathbf{P}_X = \mathbf{P}_X$,

$$(W \circ Y \circ \theta_F)^*\mathbf{P}_X = (W \circ Y)^*\mathbf{P}_X = \mathbf{P}.$$

We have thus proven that $\theta_E^*\mathbf{P} = \mathbf{P}$, which according to Definition 2.2, means that $\mathbf{P}$ is stationary. □

The crucial question of the stationarity of the SRS thus amounts to the resolution of the almost-sure equation [2.7]. We propose later in this chapter, two methods which allow us to conclude in many cases.

2.2. Loynes's scheme

Here we will consider the case where the state space E is equipped with a partial ordering $\preceq$ (see section A.3), and admits a minimal point $\mathbf{0}$ such that $\mathbf{0} \preceq x$ for all $x \in E$. We will assume that on E there exists a metric d_E such that all $\preceq$-increasing sequences converge in $\bar{E}$, the adherence of E.

DEFINITION 2.5.– *A function $\varphi\colon E \times F^{\mathbf{Z}} \to E$ is said $\preceq$-increasing when*

$$\eta \preceq \eta' \Longrightarrow \varphi(\eta, \omega) \preceq \varphi(\eta', \omega), \mathbf{P}_X - a.s..$$

It is said continuous with respect to its first variable when for $\mathbf{P}_X$-almost all ω, the function $(\eta \mapsto \varphi(\eta, \omega))$ is continuous for the metric d_E.

THEOREM 2.4 (LOYNES'S THEOREM).– *If φ is $\preceq$-increasing and continuous, the equation [2.7] admits a solution M_∞ with values in the adherence $\bar{E}$ of E.*

Proof. Let us recall that we have assumed that we know the stimulus through the quadruple $\mathfrak{O}$, whose generic element is denoted ω. We look for a random variable Y valued in E and satisfying [2.7]. We will get Y as the limit of a sequence converging almost surely. To do this, we consider Loynes's sequence $(M_n,\, n \in \mathbf{N})$, defined by

$$M_0(\omega) = \mathbf{0} \text{ and } M_{n+1}(\omega) = \varphi(M_n \circ \theta^{-1}(\omega), \theta^{-1}\omega), \forall n \geq 1. \qquad [2.8]$$

By the definition of $\mathbf{0}$, we have $M_0 = \mathbf{0} \preceq M_1$, and assuming that for some $n > 1$, $M_{n-1} \preceq M_n$ a.s., since φ is increasing we have

$$M_n(\omega) = \varphi\left(M_{n-1}(\theta^{-1}\omega), \theta^{-1}\omega\right) \preceq \varphi\left(M_n(\theta^{-1}\omega), \theta^{-1}\omega\right) = M_{n+1}(\omega)\ \mathbf{P}_X\text{-a.s..}$$

Therefore, the sequence $(M_n,\, n \in \mathbf{N})$ is a.s. increasing. In view of our assumption on the increasing sequences of E, it thus converges a.s. to the random variable $M_\infty = (\preceq -\sup)_{n\in\mathbf{N}} M_n$, which is valued in $\bar{E}$. By continuity of φ, the second relation of [2.8] implies that

$$M_\infty(\omega) = \varphi\left(M_\infty \circ \theta^{-1}(\omega), \theta^{-1}\omega\right)$$

and as θ is bijective, we deduce that M_∞ is a solution of [2.7]. □

NOTE.– In fact, the sequence M has an easy interpretation. Let $\left(W_n^0,\, n \in \mathbf{N}\right)$ be the SRS descending from 0 and driven by φ. It is easy to verify that for all $n \in \mathbf{N}$, a.s.

$$M_n = W_n^0 \circ \theta^{-n}.$$

Indeed, this relation is true for $n = 0$, and if it holds true at rank n, then a.s.

$$\begin{aligned} M_{n+1}(\omega) &= \varphi(M_n(\theta^{-1}\omega),\ \theta^{-1}\omega) \\ &= \varphi(W_n^0 \circ \theta^{-n} \circ \theta^{-1}\omega, \theta^n\theta^{-(n+1)}\omega) \\ &= W_{n+1}^0(\theta^{-(n+1)}\omega). \end{aligned}$$

In a concrete manner, M_n is the value at the instant 0 of the sequence W^0 when descending from 0 at the instant $-n$ and using as stimulus, the values of $X_{-n}, X_{-n+1}, \ldots, X_0$. For this reason, we call the construction of Loynes a *backwards recurrence scheme*. Notice by the way, the ease brought by the construction on $F^{\mathbf{Z}}$ and not $F^{\mathbf{N}}$ of the stimulus. The underlying idea is that by indexing the sequence from $-\infty$, we will have reached the stationary state at time 0.

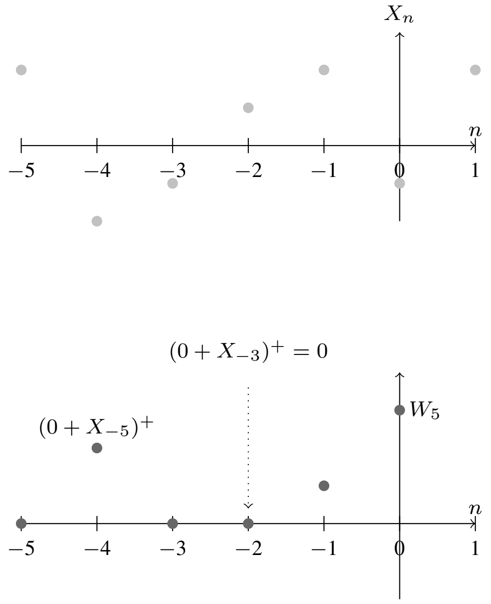

Figure 2.1. *Backwards recurrence scheme* $\varphi(x, z) = (x + z)^+$

EXAMPLE 2.4.– Example 2.3 is a typical example of such a construction, since the function $(x \mapsto (x + z)^+)$ is obviously continuous and increasing for all $z \in \mathbf{R}$. Thus there exists a random variable Y that is solution of [2.7], but we do not know *a priori* if its distribution is "proper", that is if $\mathbf{P}(Y = +\infty) = 0$. This will be one of the subjects of study of section 4.1.

We can now answer the question of weak convergence of an SRS descending from the minimal state.

COROLLARY 2.5.– *Under the assumptions of Theorem 2.4, the SRS descending from* $\mathbf{0}$ *and driven by* φ *converges in distribution to* M_∞.

Proof. Let F be bounded and continuous from E to $\mathbf{R}$. As $\mathbf{P}$ is invariant through θ we have

$$\mathbf{E}\left[F(W_n^0)\right] = \mathbf{E}\left[F(W_n^0 \circ \theta^{-n})\right] = \mathbf{E}\left[F(M_n)\right] \xrightarrow{n\to\infty} \mathbf{E}\left[F(M_\infty)\right],$$

hence the result. □

The following result will be of crucial interest in the applications to queueing.

THEOREM 2.6.– *Under the assumptions of Theorem 2.4, the solution M_∞ constructed by Loynes's scheme is $\preceq$-minimal among the solutions of [2.7].*

Proof. Let Y be a solution of [2.7]. We have $M_0 = \mathbf{0} \preceq Y$ a.s. and $M_n \preceq Y$ a.s. implies that

$$M_{n+1}(\omega) \preceq \varphi(Y \circ \theta^{-1}(\omega), \theta^{-1}\omega) = Y(\omega), \mathbf{P} - \text{ a.s..}$$

This inequality is preserved when taking the almost-sure limit, therefore

$$M_\infty \preceq Y.$$

□

We can apply the previous results to $E = \mathbf{R}^+$ totally ordered by "$\leq$" and the minimal point 0. In this context, Birkhoff's Ergodic Theorem can be seen as a fundamental application of Loynes's Theorem.

THEOREM 2.7 (Birkhoff's Ergodic theorem).– *For any real random variable $Y \in \mathbf{L}^1(\mathbf{P})$,*

$$\mathbf{E}\left[Y\right] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y \circ \theta^{-i}, \mathbf{P} - \textit{a.s..}$$

Proof. Let $\varepsilon > 0$. We define the random variable

$$Y^\varepsilon = Y - \mathbf{E}\left[Y\right] - \varepsilon$$

and the following random application from $\mathbf{R}^+$ into itself:

$$\varphi^\varepsilon \colon x \mapsto [x + Y^\varepsilon]^+ = x - x \wedge (-Y^\varepsilon)\,.$$

The function φ^ε is a.s. increasing and continuous, so Loynes's sequence $(M_n^\varepsilon,\, n \in \mathbf{N})$ driven by φ^ε is a.s. increasing in $\mathbf{R}^+$ and in view of Theorem 2.4, tends a.s. to a random variable M_∞^ε satisfying

$$M_\infty^\varepsilon \circ \theta = \varphi^\varepsilon\left(M_\infty^\varepsilon\right) = \left[M_\infty^\varepsilon + Y^\varepsilon\right]^+. \qquad [2.9]$$

An immediate induction shows that the $(M_n^\varepsilon,\ n \in \mathbf{N})$ are integrable. Therefore, for all $n \in \mathbf{N}$ we have

$$\begin{aligned}0 \geq \mathbf{E}\left[M_n^\varepsilon\right] - \mathbf{E}\left[M_{n+1}^\varepsilon\right] &= \mathbf{E}\left[M_n^\varepsilon - M_{n+1}^\varepsilon \circ \theta\right] \\ &= \mathbf{E}\left[M_n^\varepsilon - \varphi^\varepsilon\left(M_n^\varepsilon\right)\right] = \mathbf{E}\left[M_n^\varepsilon \wedge \left(-Y^\varepsilon\right)\right].\end{aligned}$$

Hence, by dominated convergence,

$$\mathbf{E}\left[M_\infty^\varepsilon \wedge \left(-Y^\varepsilon\right)\right] \leq 0.$$

In view of [2.9], the event $(M_\infty^\varepsilon = +\infty)$ is θ-invariant, and is thus of probability 0 or 1. But $M_\infty^\varepsilon = \infty$ a.s. would imply that $\mathbf{E}\left[-Y^\varepsilon\right] \leq 0$, an absurdity. Therefore, M_∞^ε is a.s. finite. Now, define the random mapping from $\mathbf{R}^+$ into itself $\tilde{\varphi}^\varepsilon\colon x \mapsto x + Y^\varepsilon$, and $(\tilde{M}_n^\varepsilon, n \in \mathbf{N})$ the associated Loynes's sequence. Notice that by construction, $\tilde{M}_n^\varepsilon = \sum_{i=1}^n Y^\varepsilon \circ \theta^{-i}$ (with the convention $\sum_{i=1}^0 = 0$). Moreover, as $\tilde{\varphi}(x) \leq \varphi(x)$ for all x, it is easy to check by induction that $\tilde{M}_n^\varepsilon \leq M_n^\varepsilon$ a.s. for any $n \in \mathbf{N}$. In particular, we have $\tilde{M}_n^\varepsilon \leq M_\infty^\varepsilon$ a.s. for all $n \in \mathbf{N}$, which amounts to saying that

$$\frac{1}{n}\sum_{i=1}^n Y \circ \theta^{-i} \leq \frac{1}{n} M_\infty^\varepsilon + \mathbf{E}\left[Y\right] + \varepsilon.$$

This is true for any $\varepsilon > 0$, we thus have that

$$\limsup_{n \to +\infty} \frac{1}{n}\sum_{i=1}^n Y \circ \theta^{-i} \leq \mathbf{E}\left[Y\right], \mathbf{P} - \text{a.s.}.$$

The latter inequality is also verified by the integrable random variable $-Y$, therefore we also have

$$\liminf_{n \to +\infty} \frac{1}{n}\sum_{i=1}^n Y \circ \theta^{-i} \geq \mathbf{E}\left[Y\right], \mathbf{P} - \text{a.s.},$$

which concludes the proof. □

NOTE.– The quadruple $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$ is stationary ergodic if, and only if, $\left(\Omega, \mathcal{F}, \mathbf{P}, \theta^{-1}\right)$ is so. We can therefore replace the statement of Theorem 2.7 by

$$\mathbf{E}\left[Y\right] = \lim_{n \to \infty} \frac{1}{n}\sum_{i=1}^n Y \circ \theta^i, \mathbf{P} - \text{a.s.},$$

for all real random variables $Y \in \mathbf{L}^1(\mathbf{P})$.

2.3. Coupling

The idea of coupling plays a central role in the asymptotic study of SRS. It is in fact possible to state the conditions under which the trajectories of two SRS (or possibly those of the corresponding backward schemes) coincide at a certain point. These properties imply naturally, in particular, more traditional properties of convergence for random sequences such as convergence in distribution.

Hereafter we only state the results that will be useful to us in the applications to queueing, in their simplest form.

Secondly, we develop the theory of renovating events of Borovkov, which gives sufficient conditions for coupling, and even strong backward coupling. In addition, the results of Borovkov and Foss also allow in many cases to solve the equation [2.7], even when the conditions of continuity and monotonicity of Theorem 2.4 are not satisfied. Particularly, in this framework we can also deal with the intricate question of the transient behavior depending on the initial conditions. In what follows, $\mathfrak{Q} = (\Omega, \mathcal{F}, \mathbf{P}, \theta)$ is a stationary ergodic quadruple.

2.3.1. *Definition*

We begin by defining the different types of coupling.

DEFINITION 2.6.– *Let $(W_n,\, n \in \mathbf{N})$ and $(Y_n,\, n \in \mathbf{N})$ be two random sequences defined on $\mathfrak{Q}$.*

1) We say that $(W_n,\, n \in \mathbf{N})$ and $(Y_n,\, n \in \mathbf{N})$ couple if

$$\mathbf{P}\left(W_n = Y_n; \forall n \geq N\right) \xrightarrow{N\to\infty} 1. \qquad [2.10]$$

2) We say that there is a strong backward coupling between $(W_n,\, n \in \mathbf{N})$ and $(Y_n,\, n \in \mathbf{N})$ if

$$\mathbf{P}\left(W_n \circ \theta^{-n} = Y_n \circ \theta^{-n}; \forall n \geq N\right) \xrightarrow{N\to\infty} 1. \qquad [2.11]$$

In what follows, we denote

$$\tau_F = \inf\left\{N \in \mathbf{N}; W_n = Y_n, \forall n \geq N\right\};$$
$$\tau_B = \inf\left\{N \in \mathbf{N}; W_n \circ \theta^{-n} = Y_n \circ \theta^{-n}, \forall n \geq N\right\},$$

the (random) indexes of coupling of the two sequences, respectively “forward” and “backward”, setting these random variables as infinite whenever the right-hand set is empty. We can then easily see that the forward and the strong backward coupling of $(W_n,\, n \in \mathbf{N})$ with $(Y_n,\, n \in \mathbf{N})$ admit, respectively, the following equivalent definitions

$$\mathbf{P}\left(\tau_F \geq N\right) \xrightarrow{N\to\infty} 0; \qquad [2.12]$$

$$\mathbf{P}\left(\tau_B \geq N\right) \xrightarrow{N\to\infty} 0. \qquad [2.13]$$

We start by noticing an immediate link between coupling and convergence in distribution.

THEOREM 2.8.– *Let* $(W_n,\, n \in \mathbf{N})$ *be a sequence with values in* E*, which couples with the stationary sequence* $(Y \circ \theta^n,\, n \in \mathbf{N})$*. Then,*

$$W_n \xrightarrow[n\to\infty]{\mathcal{L}} Y.$$

Proof. Let G be a bounded continuous function: $E \to \mathbf{R}$, and $\parallel G \parallel_\infty$ its supremum. For all N we have

$$\begin{aligned}\Big|\mathbf{E}\left[G(W_N)\right] - \mathbf{E}\left[G(Y)\right]\Big| &= \Big|\mathbf{E}\left[G(W_N)\right] - \mathbf{E}\left[G(Y)\circ\theta^N\right]\Big| \\ &\leq \mathbf{E}\Big[\Big|G(W_N) - G(Y\circ\theta^N)\Big|\Big] \\ &= \mathbf{E}\Big[\Big|G(W_N) - G(Y\circ\theta^N)\Big|\mathbf{1}_{\tau_F > N}\Big] \\ &\leq 2 \parallel G \parallel_\infty \mathbf{P}\left(\tau_F > N\right),\end{aligned}$$

and the quantity on the right-hand side tends to 0 by hypothesis. □

Let us denote

$$\tau_f = \inf\left\{n \in \mathbf{N}; W_n = Y_n\right\},$$

the first index in which the two sequences $(W_n,\, n \in \mathbf{N})$ and $(Y_n,\, n \in \mathbf{N})$, coincide (setting $\tau_f = \infty$ if the latter set is empty).

NOTE.– We can observe that two SRS driven by the same recurrence function φ (we then denote them $\left(W_n^Y,\, n \in \mathbf{N}\right)$ and $\left(W_n^Z,\, n \in \mathbf{N}\right)$, only their initial random variables possibly differentiate them) couple as soon as τ_f is reached. Indeed, a.s. $W_n^Y(\omega) = W_n^Z(\omega)$ implies that

$$W_{n+1}^Y(\omega) = \varphi(W_n^Y(\omega), \theta^n\omega) = \varphi(W_n^Z(\omega), \theta^n\omega) = W_{n+1}^Z(\omega).$$

Consequently, for all $n \in \mathbf{N}$, $\{\tau_f \leq n\} \subset \{\tau_F \leq n\}$ or in other words

$$\tau_F = \tau_f \text{ a.s.} \qquad [2.14]$$

On the other hand, the two sequences $\left(W^Y \circ \theta^{-n},\, n \in \mathbf{N}\right)$ and $\left(W^Z \circ \theta^{-n},\, n \in \mathbf{N}\right)$ can coincide for a certain index without a strong backward coupling, as for any ω such that $W_n^Y(\theta^{-n}\omega) = W_n^Z(\theta^{-n}\omega)$, we have

$$\begin{aligned} W_{n+1}^Y(\theta^{-(n+1)}\omega) &= \varphi(W_n^Y(\theta^{-(n+1)}\omega), \theta^n(\theta^{-(n+1)}\omega)) \\ &= \varphi(W_n^Y(\theta^{-(n+1)}\omega), \theta^{-1}\omega); \\ W_{n+1}^Z(\theta^{-(n+1)}\omega)] &= \varphi(W_n^Z(\theta^{-(n+1)}\omega, \theta^{-1}\omega), \end{aligned}$$

and these two quantities are not equal in general.

In the case of SRS, the link between the different types of coupling is established in the following theorem.

THEOREM 2.9.– *Let Z and Y be two random variables with values in E and $\left(W_n^Z,\, n \in \mathbf{N}\right)$, an SRS descending from Z and driven by φ. If there is strong backward coupling between $\left(W_n^Z,\, n \in \mathbf{N}\right)$ and the stationary sequence $(Y \circ \theta^n,\, n \in \mathbf{N})$, then these two sequences couple, $W_n^Z \xrightarrow{\mathcal{L}} Y$ and Y is a solution of [2.7].*

Proof. First, for any $\omega \in \theta^{-1}\{\tau_B < \infty\}$ (i.e. such that $\tau_B(\theta\omega) < \infty$), for all $n \geq \tau_B(\theta\omega)$,

$$\begin{aligned} Y \circ \theta(\omega) &= W_{n+1}^Z \circ \theta^{-(n+1)}(\theta\omega) \\ &= \varphi(W_n^Z \circ \theta^{-n}(\omega), \theta^n \circ \theta^{-n}\omega) \\ &= \varphi(Y(\omega), \omega). \end{aligned}$$

The event $\theta^{-1}(\tau_B < \infty)$ is of probability 1 by hypothesis: we therefore have

$$Y \circ \theta(\omega) = \varphi(Y(\omega), \omega), \text{ a.s.}. \qquad [2.15]$$

Moreover, for any $N \in \mathbf{N}$,

$$\begin{aligned} \mathbf{P}(\tau_B \leq N) &= \mathbf{P}\left(W_n^Z \circ \theta^{-n} = Y, \forall n \geq N\right) \\ &= \mathbf{P}\left(\theta^{-N}\left\{W_n^Z \circ \theta^{-n} = Y, \forall n \geq N\right\}\right) \\ &= \mathbf{P}\left(W_n^Z \circ \theta^{N-n} = Y \circ \theta^N, \forall n \geq N\right) \\ &\leq \mathbf{P}\left(W_N^Z = Y \circ \theta^N\right) \\ &\leq \mathbf{P}(\tau_f \leq N). \end{aligned} \qquad [2.16]$$

But according to [2.15], $(Y \circ \theta^n,\, n \in \mathbf{N}) \equiv \left(W_n^Y,\, n \in \mathbf{N}\right)$ the SRS descending from Y and driven by φ and, therefore, we are in the case of the previous Remark: according to [2.14], $\tau_f = \tau_F$ a.s. in this case, and therefore according to [2.16],

$$\mathbf{P}(\tau_B > N) \geq \mathbf{P}(\tau_F > N).$$

The right-hand term thus tends to 0 as N goes to infinity, just as the left-hand term, which shows the coupling property. Finally, the convergence in distribution follows from Theorem 2.8. □

We end this section with the following result.

THEOREM 2.10.– *Let $\left(W_n^0,\ n \in \mathbf{N}\right)$ be an SRS with values in E descending from* $\mathbf{0}$ *and driven by φ, a mapping that is a.s. $\preceq$-increasing and continuous. Let W be a solution of equation [2.7] corresponding to φ, with values in E. Then there is an equivalence between the forward and the strong backward coupling between $\left(W_n^0,\ n \in \mathbf{N}\right)$ and $(W \circ \theta^n,\ n \in \mathbf{N})$.*

Proof. We will in fact show that τ_F and τ_B have the same distribution here. Denote once again $(M_n,\ n \in \mathbf{N}) = \left(W_n^0 \circ \theta^{-n},\ n \in \mathbf{N}\right)$ the corresponding Loynes sequence, and recall that this sequence increases a.s. toward $M_\infty \preceq W$. In particular, for any $N \in \mathbf{N}$ and any $n \geq N$, a.s. $M_N(\omega) = W(\omega)$ implies that $M_n(\omega) = M_\infty(\omega) = W(\omega)$. Thus,

$$\begin{aligned}
\mathbf{P}\left(\tau_F \leq N\right) &= \mathbf{P}\left(W_N^0 = W \circ \theta^N\right) \\
&= \mathbf{P}\left(M_N = W\right) \\
&= \mathbf{P}\left(M_n = W, \forall n \geq N\right) \\
&= \mathbf{P}\left(W_n^0 \circ \theta^{-n} = W, \forall n \geq N\right) \\
&= \mathbf{P}\left(\tau_N \leq N\right),
\end{aligned}$$

which completes the proof. □

2.3.2. *Renovating events*

The theory of renovating events provides, as we are going to see, a simple criteria for the strong backward coupling of an SRS with a solution of [2.7]. Throughout this subsection, $(W_n,\ n \in \mathbf{N})$ denotes an SRS defined on $\mathfrak{O} = (\Omega, \mathcal{F}, \mathbf{P}, \theta)$ with values in E and driven by φ.

DEFINITION 2.7.– *Let N be a strictly positive integer. We say that the sequence of measurable events $(\mathfrak{A}_n,\ n \in \mathbf{N})$ is a sequence of renovating events of length N for $(W_n,\ n \in \mathbf{N})$ if, and only if, there exists a random variable α defined on Ω with values in a measurable space F, and a deterministic mapping $\Phi\colon F^N \to E$ such that for all $n \geq N$, on $\mathfrak{A}_{n-N}$,*

$$W_n = \Phi\left(\alpha \circ \theta^{n-N}, \alpha \circ \theta^{n-2}, ..., \alpha \circ \theta^{n-1}\right).$$

The latter is in a sense a “memoryless property”: on $\mathfrak{A}_{n-N}$, W_n does not depend on anything but a list of N values of the stationary sequence $(\alpha \circ \theta^n,\ n \in \mathbf{N})$. In a concrete manner, if a given event occurs N time slots in the past, W_n depends on its past only up to the moment $n-N$ and no further.

EXAMPLE 2.5.– Let us assume that $(W_n,\, n\in\mathbf{N})$ is an SRS driven by φ. Let $x\in E$ and for all n, $\mathfrak{A}_n=\{W_{n-1}=x\}$. It is then easy to see that $(\mathfrak{A}_n,\, n\in\mathbf{N})$ is a sequence of renovating events of length 1. Indeed, for any $n\geq 1$, on $\mathfrak{A}_{n-1}$, $W_n=\varphi\circ\theta^{n-1}(x)$, and the definition is matched by taking $F=E$, Φ as the identity on E and $\alpha=\varphi(x)$. In the applications (Chapter 4), we will mainly consider this type of renovating events in the particular case $x=0$.

DEFINITION 2.8.– *A sequence of events* $(\mathfrak{A}_n,\, n\in\mathbf{N})$ *is said to be* θ*-compatible if for any* $n\geq 0$, $\mathfrak{A}_n=\theta^{-n}\mathfrak{A}_0$. *We can then define the sequence* $(\mathfrak{A}_n,\, n\in\mathbf{Z})$ *by denoting* $\mathfrak{A}_{-n}=\theta^n\mathfrak{A}_0$ *for any* $n\geq 0$.

EXAMPLE 2.6.– Let β, a random variable defined on Ω and with values in $(E,\mathcal{E})$. Then, for all $\mathcal{B}\in\mathcal{E}$, the sequence of events defined for all n by $\mathfrak{A}_n=\{\beta\circ\theta^n\in\mathcal{B}\}$ is θ-compatible, since for any n, $\omega\in\mathfrak{A}_n$ amounts to $\beta\circ\theta^n(\omega)\in\mathcal{B}$, that is $\beta\circ\theta^{n+1}\left(\theta^{-1}\omega\right)\in\mathcal{B}$ or in other words, $\omega\in\theta\mathfrak{A}_{n+1}$. The sequence of the antecedents of a measurable set by a stationary sequence is thus, as expected, a θ-compatible sequence.

The following theorem is due to Borovkov and Foss.

THEOREM 2.11.– *Let* $\mathcal{Z}$ *be a family of random variables with values in* E. *We assume that all sequences* $\left(W_n^Z,\, n\in\mathbf{N}\right)$, *with* $Z\in\mathcal{Z}$, *admit the same sequence of renovating events* $(\mathfrak{A}_n,\, n\in\mathbf{N})$, *of same length* N, *with the same associated random variable* α *and the same associated application* Φ, *and that the sequence* $(\mathfrak{A}_n,\, n\in\mathbf{N})$ *is* θ*-compatible and such that* $\mathbf{P}\left(\mathfrak{A}_0\right)>0$. *Then, there exists a finite random variable* W *such that for all* $Z\in\mathcal{Z}$, *there is strong backward coupling between* $\left(W_n^Z,\, n\in\mathbf{N}\right)$ *and* $(W\circ\theta^n,\, n\in\mathbf{N})$.

Proof. Let $n\geq N$, $i\in[\![0,\, n-N]\!]$, and $k\geq 0$. The θ-compatibility implies that

$$\mathfrak{A}_{-N-i}=\bigcap_{k\geq 0}\theta^{n+k}\mathfrak{A}_{n+k-N-i}$$

and therefore, for any $Z\in\mathcal{Z}$, for all $\omega\in\mathfrak{A}_{-N-i}$ we have for all $k\geq 0$,

$$\begin{aligned}&\theta^{-(n+k)}\omega\in\mathfrak{A}_{n+k-N-i}\\&\iff W_{n+k-i}^Z(\theta^{-(n+k)}\omega)\\&\qquad\qquad=\Phi(\alpha\circ\theta^{n+k-i-N}(\theta^{-(n+k)}\omega),\ldots,\alpha\circ\theta^{n+k-i-1}(\theta^{-(n+k)}\omega))\\&\qquad\iff W_{n+k-i}^Z\circ\theta^{-(n+k)}(\omega)=\Phi(\alpha\circ\theta^{-i-N}(\omega),\ldots,\alpha\circ\theta^{-i-1}(\omega)),\end{aligned}$$

which is a random variable that does not depend either on Z, or on k. Thus, for any pair of initial conditions Z and $Z'\in\mathcal{Z}$, we thus have on $\mathfrak{A}_{-N-i}$,

$$W_{n+k-i}^{Z'}\circ\theta^{-(n+k)}=W_{n-i}^Z\circ\theta^{-n},\forall k\geq 0.$$

Therefore,

$$\begin{aligned} W^{Z'}_{n+1+k-i} \circ \theta^{-(n+k)} &= \varphi \circ \theta^{n+k-i} \circ \theta^{-(n+k)} \left(W^{Z'}_{n+k-i} \circ \theta^{-(n+k)} \right) \\ &= \varphi \circ \theta^{n-i} \circ \theta^{-n} \left(W^{Z}_{n-i} \circ \theta^{-n} \right) \\ &= W^{Z}_{n+1-i} \circ \theta^{-n}, \end{aligned}$$

which implies by an immediate induction up to the rank i that

$$W^{Z'}_{n+k} \circ \theta^{-(n+k)} = W^{Z'}_{n+i+k-i} \circ \theta^{-(n+k)} = W^{Z}_{n+i-i} \circ \theta^{-n} = W^{Z}_{n} \circ \theta^{-n}.$$

In other words, on $\mathfrak{A}_{-N-i}$ the two sequences $\left(W^{Z}_{n} \circ \theta^{-n},\ n \in \mathbf{N}\right)$ and $\left(W^{Z'}_{n} \circ \theta^{-n},\ n \in \mathbf{N}\right)$ are constant and equal after n. This is true for all $i \in [\![0,\ n-N]\!]$, therefore denoting for all $Z \in \mathcal{Z}$, τ^{Z}_{B} the backward coupling time of the sequence $(W^{Z}_{n}, n \in \mathbf{N})$, we have

$$\begin{aligned} &\bigcup_{i=0}^{n-N} \mathfrak{A}_{-i-N} \subseteq \mathcal{B}_n \\ &\qquad = \left\{ \tau^{Z}_{B} \leq n, \forall Z \in \mathcal{Z} \right\} \bigcap \left\{ W^{Z}_{n} \circ \theta^{-n} = W^{Z'}_{n} \circ \theta^{-n}, \forall Z, Z' \in \mathcal{Z} \right\}, \end{aligned}$$

which implies that

$$\mathfrak{A} = \bigcup_{j=N}^{+\infty} \mathfrak{A}_{-i} = \bigcup_{n=N}^{+\infty} \bigcup_{i=N}^{n} \mathfrak{A}_{-i} \subseteq \bigcup_{n=N}^{+\infty} \mathcal{B}_n.$$

But $\mathfrak{A} \subseteq \theta\mathfrak{A}$, while $\mathfrak{A}_{-N} \subseteq \mathfrak{A}$ has a strictly positive probability by hypothesis. The event $\mathfrak{A}$ is therefore of probability 1. So the event $\bigcup_{n=N}^{+\infty} \mathcal{B}_n$ is almost sure, which means that all the sequences $\left(W^{Z}_{n},\ n \in \mathbf{N}\right)$ couple (in the strong backward sense), with the same random variable since the sequences $\left(W^{Z}_{n} \circ \theta^{-n},\ n \in \mathbf{N}\right)$, $Z \in \mathcal{Z}$ are equal from a certain time. The theorem is proved. □

An immediate, but crucial corollary to the last theorem is the following sufficient condition for the existence of a solution to the stationary equation [2.7].

COROLLARY 2.12.– *If there exists a set of non-empty conditions $\mathcal{Z}$ satisfying the hypothesis of Theorem 2.11, equation [2.7] admits a E-valued solution.*

Proof. According to Theorem 2.11, for all $Z \in \mathcal{Z}$ there is a strong backward coupling for $\left(W^{Z}_{n},\ n \in \mathbf{N}\right)$. This implies the result in view of Theorem 2.9. □

The following second corollary is a criterion for the uniqueness of a solution to [2.7].

COROLLARY 2.13.– *If the set*

$$\mathcal{Z} = \{\textit{solutions of [2.7] with values in } E\}$$

is non-empty and satisfies the assumptions of Theorem 2.11, it is reduced to a singleton.

Proof. For any pair of solutions $Z, Z' \in \mathcal{Z}$, the two stationary sequences $\left(W_n^Z,\, n \in \mathbf{N}\right) = (Z \circ \theta^n,\, n \in \mathbf{N})$ and $\left(W_n^{Z'},\, n \in \mathbf{N}\right) = (Z' \circ \theta^n,\, n \in \mathbf{N})$ couple with strong backward coupling with the same stationary sequence $(W \circ \theta^n,\, n \in \mathbf{N})$. Hence, we naturally have $Z = Z' = W$, $\mathbf{P}$-a.s.. □

2.4. Comparison of stochastic recursive sequences

In this last section, we give two remarkable comparison results for stochastic recursive sequences, which will be applied to queueing systems in Chapter 4. Throughout this section, the Euclidean spaces $\mathbf{R}^K$, $K \geq 1$ are equipped with the partial ordering $\prec$ defined in Appendix A.

DEFINITION 2.9.– *Let W and Y be two random variables with values in $\mathbf{R}^K$, possibly defined on two different probability spaces Ω and $\hat{\Omega}$. We say that Y dominates W stochastically, or for the strong ordering, and we denote $W \leq_{st} Y$, if for any $\prec$-increasing function $F\colon \mathbf{R}^K \to \mathbf{R}$ such as the following integrals exist,*

$$\mathbf{E}\left[F(W)\right] \leq \hat{\mathbf{E}}\left[F(Y)\right].$$

Notice in particular that if W and Y are real random variables,

$$W \leq_{st} Y \Longleftrightarrow \mathbf{P}\left(W \leq x\right) \geq \hat{\mathbf{P}}\left[Y \leq x\right] \text{ for any } x \in \mathbf{R}. \qquad [2.17]$$

The following theorem is the fundamental result of the theory of stochastic comparison.

THEOREM 2.14 (Strassen's Theorem).– *Let W and Y be two random variables with values in $\mathbf{R}^K$. Then $W \leq_{st} Y$ if, and only if, there exists a probability space on which are defined two random variables $\tilde{W}$ and $\tilde{Y}$, of same respective distributions as W and Y on $\mathbf{R}^K$, and such that*

$$\tilde{W} \prec \tilde{Y} \textit{ a.s..}$$

THEOREM 2.15.– *Let α and $\bar{\alpha}$ be two random variables defined on the stationary ergodic quadruple $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$, with values in $\mathbf{R}^m$ and integrable. Let f be a deterministic mapping: $\mathbf{R}^d \times \mathbf{R}^m \to \mathbf{R}^d$. We note $(W_n,\, n \in \mathbf{N})$ and $\left(\bar{W}_n,\, n \in \mathbf{N}\right)$*

the two SRS with values in $\mathbf{R}^d$, descending from 0 *and driven, respectively, by the random mappings*

$$x \mapsto f(x, \alpha) \text{ and } x \mapsto f(x, \bar{\alpha}).$$

If $\bar{\alpha} \leq_{st} \alpha$ *and* f *is increasing in both its arguments, then* $\bar{W}_n \leq_{st} W_n$ *for any* $n \in \mathbf{N}$.

Proof. According to Strassen's theorem, there exists a probability space on which are defined the sequences $(\alpha_n,\ n \in \mathbf{N})$ and $(\bar{\alpha}_n,\ n \in \mathbf{N})$, such that we have the following identities in distribution

$$\alpha_n \overset{\mathcal{L}}{=} \alpha \circ \theta^n \text{ and } \bar{\alpha}_n \overset{\mathcal{L}}{=} \bar{\alpha} \circ \theta^n, \text{ for all } n \in \mathbf{N},$$

and such that

$$\bar{\alpha}_n \prec \alpha_n \text{ a.s. for all } n \in \mathbf{N}.$$

Let $(Y_n,\ n \in \mathbf{N})$ and $\left(\bar{Y}_n,\ n \in \mathbf{N}\right)$, the two SRS defined on this new probability space, descending from 0 and driven, respectively, by the sequences of mappings $(f(.,\alpha_n),\ n \in \mathbf{N})$ and $(f(.,\bar{\alpha}_n),\ n \in \mathbf{N})$. We then have

$$\bar{Y}_n \prec Y_n, \text{ a.s. for any } n \in \mathbf{N},$$

which we show by induction. We have $Y_0 = \bar{Y}_0 = 0$ a.s. and if $\bar{Y}_n \prec Y_n$ for some $n \in \mathbf{N}$, then by the monotonicity of f, we have a.s.

$$\begin{aligned}\bar{Y}_{n+1} &= f\left(\bar{Y}_n, \bar{\alpha}_n\right)\\ &\prec f\left(Y_n, \bar{\alpha}_n\right)\\ &\prec f\left(Y_n, \alpha_n\right)\\ &= Y_{n+1}.\end{aligned}$$

So $\bar{Y}_n \prec Y_n$ a.s. for all $n \in \mathbf{N}$ on the new probability space. As $\bar{Y}_n$ (respectively, Y_n) clearly has the same distribution as $\bar{W}_n$ (respectively, W_n) for all n, the converse of Strassen's theorem allows us to conclude. □

Now let us assume furthermore that the mapping f is continuous in its first variable. According to Loynes's theorem, Loynes's sequences $(W_n \circ \theta^{-n},\ n \in \mathbf{N})$ and $\left(\bar{W}_n \circ \theta^{-n},\ n \in \mathbf{N}\right)$ converge a.s. to the respective minimal stationary versions W_∞ and $\bar{W}_\infty$ of the two SRS.

COROLLARY 2.16.– *Under the assumptions of Theorem 2.15, if Loynes's theorem applies to both SRS and if the minimal solutions* W_∞ *and* $\bar{W}_\infty$ *are a.s. finite, they are such that*

$$\bar{W}_\infty \leq_{st} W_\infty.$$

Proof. According to Theorem 2.15, for any increasing function $F\colon E \to \mathbf{R}$ such that the following expectations exist,

$$\mathbf{E}\left[F(\bar{W}_n)\circ\theta^{-n}\right] = \mathbf{E}\left[F(\bar{W}_n)\right] \leq \mathbf{E}\left[F(W_n)\right] = \mathbf{E}\left[F(W_n)\circ\theta^{-n}\right],$$

by θ-invariance. We conclude easily by monotone convergence. □

The following theorem is the special case (which will be useful to us in this form in Chapter 4) of a more general result, which declines the comparison property of Theorem 2.15 for a stochastic ordering involving the convex test functions, by applying a corollary of Strassen's Theorem for this ordering. In the sequel, if Y is a random variable with values in $\mathbf{R}^K$, $K \geq 1$ and $\mathcal{A}$ is a sigma-field, we classically denote

$$\mathbf{E}\left[Y \mid \mathcal{A}\right] = \left(\mathbf{E}\left[Y(1) \mid \mathcal{A}\right], \mathbf{E}\left[Y(2) \mid \mathcal{A}\right], \ldots\right).$$

THEOREM 2.17.– *We assume that the mapping f is $\prec$-increasing in its first argument, and convex from $\mathbf{R}^d \times \mathbf{R}^m$ into $\mathbf{R}^d$. Furthermore, let us assume that there exists a filtration $(\mathcal{F}_n,\ n \in \mathbf{N})$ such that for all $n \in \mathbf{N}$ and for all $i \in [\![0,\ n]\!]$,*

$$\bar{\alpha}\circ\theta^i = \mathbf{E}\left[\alpha\circ\theta^i \mid \mathcal{F}_n\right].$$

Under these assumptions,

$$\mathbf{E}\left[F\left(\bar{W}_n\right)\right] \leq \mathbf{E}\left[F\left(W_n\right)\right] \text{ for all } n \in \mathbf{N}, \tag{2.18}$$

for all $F\colon \mathbf{R}^d \to \mathbf{R}$, $\prec$-increasing and convex, and such that these integrals are well defined.

Proof. Let us fix $n \in \mathbf{N}$. Let us show by induction on $[\![0,\ n]\!]$ the relation

$$\bar{W}_i \prec \mathbf{E}\left[W_i \mid \mathcal{F}_n\right] \text{ a.s. for all } i \in [\![0, n]\!], \tag{2.19}$$

which is of course checked for $i = 0$. Assuming that it is true for some $i \in [\![0, n]\!]$, Jensen's inequality yields that a.s.

$$\begin{aligned}
\mathbf{E}\left[W_{i+1} \mid \mathcal{F}_n\right] &= \mathbf{E}\left[f\left(W_i, \alpha\circ\theta^i\right) \mid \mathcal{F}_n\right] \\
&\succ f\left(\mathbf{E}\left[W_i \mid \mathcal{F}_n\right], \mathbf{E}\left[\alpha\circ\theta^i \mid \mathcal{F}_n\right]\right) \\
&\succ f\left(\bar{W}_i, \mathbf{E}\left[\alpha\circ\theta^i \mid \mathcal{F}_n\right]\right) \\
&= f\left(\bar{W}_i, \bar{\alpha}\circ\theta^i\right) \\
&= \bar{W}_{i+1},
\end{aligned}$$

and [2.19] is proved. Therefore, we have in particular that

$$\bar{W}_n \prec \mathbf{E}\left[W_n \mid \mathcal{F}_n\right] \text{ a.s.,}$$

which implies with Jensen's inequality that for any increasing and convex function $F\colon \mathbf{R}^d \to \mathbf{R}$, if the integrals are well defined,

$$\begin{aligned} F\left(\bar{W}_n\right) &\leq F\left(\mathbf{E}\left[W_n \,|\, \mathcal{F}_n\right]\right) \\ &\leq \mathbf{E}\left[F(W_n) \,|\, \mathcal{F}_n\right]. \end{aligned}$$

We conclude by taking expectations in the last inequality. □

We can deduce in particular, as for Corollary 2.16,

COROLLARY 2.18.– *Under the hypothesis of Theorem 2.17, if furthermore Loynes's theorem applies to both SRS, and if the minimal solutions $\bar{W}_\infty$ and W_∞ are a.s. finite, they satisfy*

$$\mathbf{E}\left[F\left(\bar{W}_\infty\right)\right] \leq \mathbf{E}\left[F\left(W_\infty\right)\right],$$

for any increasing and convex function $F\colon\ \mathbf{R}^d \to \mathbf{R}$, such that the expectations are well defined.

2.5. Notes and comments

Loynes's Theorem has been introduced in [LOY 62] in the particular case of the G/G/1 queue. It has been generalized in the form presented here, for instance in [NEV 84] and [BAC 02]. The proof of Birkhoff's Ergodic Theorem that is presented here is due to Garsia [GAR 65].

For a more complete picture on the idea of coupling, we refer the reader to [THO 00] and [BRA 90].

The theory of Renovating events is due to Borovkov and Foss. It has been introduced in [BOR 84], and developed in [BOR 92], [BOR 94] and [BOR 98].

For a more complete overview on stochastic comparison of stochastic recursions, see the reference books [BAC 02] and [STO 83]. The construction presented here is due to Baccelli and Makowski [BAC 89]. We only give here a simplifies versions of the results therein.

Epitome

– A stochastic Recursive Sequence (SRS) is of the form $X_{n+1} = f(X_n, \alpha_n)$, where (α_n) is a stationary ergodic sequence.

– The existence of a stationary distribution for (X_n) amounts to that of a random variable X solving the pathwise equation

$$X \circ \theta = f(X, \alpha),$$

on the canonical space of (α_n), where θ is the bijective shift operator.

– Loynes's backward scheme guarantees the existence of a solution X (possibly infinite) if f is a.s. increasing and continuous in its first argument.

– Borovkov and Foss's Theory of renovating events provides conditions of existence and uniqueness of a solution X, and for coupling to occur with the stationary version of the SRS.

– Strassen's Theorem allows us to compare the stationary versions of an SRS based on the ordering of the random sequences that drive it.

Chapter 3

Markov Chains

To describe the evolution of a system, we must prescribe how the future depends on the present or the past. Two major examples of such descriptions are differential equations and recurrent sequences. When a piece of randomness is added, it leads to stochastic differential equations (which are beyond the scope of this book) and stochastic recurrent sequences (SRS), which we already studied in Chapter 2. Among SRS, Markov chains are the most salient category. Behind a seemingly simple description lies a mathematical tool which is quite efficient for applications and rich of many properties.

Remember that it is recommended to read section A.1.1.

3.1. Definition and examples

Consider a sequence of random variables $X = (X_n, n \geq 0)$ with values in E, finite or countable, and the filtration $\mathcal{F}_n = \sigma\{X_j, 0 \leq j \leq n\}$ generated by this sequence.

Trajectories of X are elements of $E^{\mathbf{N}}$, that is to say sequences of elements of E. The shift (see section A) is then defined by

$$\begin{aligned} \theta\colon E^{\mathbf{N}} &\longrightarrow E^{\mathbf{N}} \\ (x_0, x_1, \ldots) &\longmapsto (x_1, x_2, \ldots). \end{aligned}$$

This shift is the non-bijective restriction to $E^{\mathbf{N}}$ of the bijective flow defined on $E^{\mathbf{Z}}$ in section 2.1. As with the flow, we need to define the nth iteration of θ, denoted as θ^n and defined by

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

$$\begin{aligned}\theta^n \colon E^{\mathbf{N}} &\longrightarrow E^{\mathbf{N}}\\ (x_0, x_1, \ldots) &\longmapsto (x_n, x_{n+1}, x_{n+2}, \ldots).\end{aligned}$$

From now on, we identify θ and θ^1.

DEFINITION 3.1.– *The sequence X is a Markov chain when for any $n \le m$, the σ-field $\mathcal{F}_n$ is independent of the σ-field $\sigma(X_m)$, given $\sigma(X_n)$. In other words, for any bounded functions F and G*

$$\begin{aligned}&\mathbf{E}\left[F(X_0, \ldots, X_n)G(X_m) \,|\, X_n\right]\\ &\quad = \mathbf{E}\left[F(X_0, \ldots, X_n) \,|\, X_n\right]\ \mathbf{E}\left[G(X_m) \,|\, X_n\right]. \qquad [3.1]\end{aligned}$$

According to Theorem A.12, we know that this property is equivalent to the independence of the past and the future given the present, and that this can be expressed by

$$\mathbf{E}\left[F(X_0, \ldots, X_n)G \circ \theta^n \,|\, \mathcal{F}_n\right] = F(X_0, \ldots, X_n)\mathbf{E}\left[G \circ \theta^n \,|\, X_n\right]. \qquad [3.2]$$

In particular, for $G = \mathbf{1}_{\{y\}}(X_1)$, for any integer n, we obtain

$$\mathbf{P}(X_{n+1} = y \,|\, \mathcal{F}_n) = \mathbf{P}(X_{n+1} = y \,|\, X_n).$$

DEFINITION 3.2.– *The Markov chain X is said to be homogeneous when $\mathbf{P}(X_{n+1} = y \,|\, X_n = x)$ does not depend on n but only on x and y. We denote this quantity by $p(x, y)$ and $P = (\mathbf{P}(X_1 = y \,|\, X_0 = x), x, y \in E)$ is called the transition operator of X. If E is finite, then P is identified with a matrix that has as many rows and columns as elements in E.*

EXAMPLE 3.1.– A rat moves through the labyrinth of seven squares shown in Figure 3.1. It goes from one box to another by uniformly choosing among the possibilities given to it, that is to say that when there are 2 (respectively 3) outputs in the box where the rat is present, the rat goes into each possible box with a probability of one half (respectively of a third). Its evolution has no memory: each change depends only on the current situation, not the past. X_n is the position of the rat after its nth movement, X_0 is its initial position.

Here, $E = \{1, 2, 3, 4, 5, 6, 7\}$ and the transition matrix is easily deduced from Figure 3.1.

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0\\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0\\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0\\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3}\\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0\\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

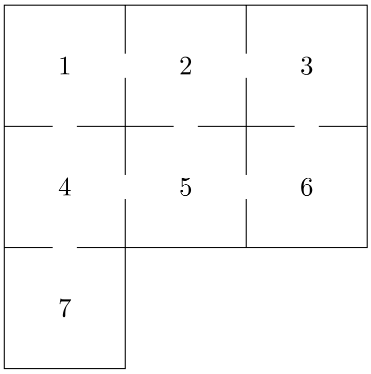

Figure 3.1. *The maze*

EXAMPLE 3.2.– Starting with a null score, let us throw two unbiased dice. If the sum of their figures is different from 7, we add this amount to the current score and play again. Otherwise the score is zeroed and the game stops. X_n represents the score after the nth throw. We must distinguish two "states" 0 if we want X to be an homogeneous chain. In fact, the score can quit the 0 at the beginning but cannot leave the 0 subsequent to a 7. Hence, we take as state space $E = \mathbf{N} \cup \{\delta\}$ where δ is called a cemetery point. The transitions are thus given for all $i \neq \delta$ by

$$\begin{aligned}
&\mathbf{P}(X_{n+1} = i + 2 \,|\, X_n = i) = \mathbf{P}(X_{n+1} = i + 12 \,|\, X_n = i) = 1/36\\
&\mathbf{P}(X_{n+1} = i + 3 \,|\, X_n = i) = \mathbf{P}(X_{n+1} = i + 11 \,|\, X_n = i) = 2/36\\
&\mathbf{P}(X_{n+1} = i + 4 \,|\, X_n = i) = \mathbf{P}(X_{n+1} = i + 10 \,|\, X_n = i) = 3/36\\
&\mathbf{P}(X_{n+1} = i + 5 \,|\, X_n = i) = \mathbf{P}(X_{n+1} = i + 9 \,|\, X_n = i) = 4/36\\
&\mathbf{P}(X_{n+1} = i + 6 \,|\, X_n = i) = \mathbf{P}(X_{n+1} = i + 8 \,|\, X_n = i) = 5/36\\
&\mathbf{P}(X_{n+1} = \delta \,|\, X_n = i) = 1/6\\
&\mathbf{P}(X_{n+1} = \delta \,|\, X_n = \delta) = 1.
\end{aligned}$$

The very definition of a Markov chain implies that its evolution is determined by the distribution of the initial position, denoted henceforth by ν, and the transition operator P. This is mathematically expressed in the following theorem.

THEOREM 3.1.– *For any n, the joint distribution of $(X_0, \cdots, X_n)$ is determined by the distribution of X_0 and P by the following formula*

$$\mathbf{P}(X_0 = x_0, \dots, X_m = x_m) = \nu(\{x_0\}) \prod_{l=0}^{m-1} p(x_l, x_{l+1}),$$

for all n and all $x_0, \dots, x_n$ in E.

NOTE.– In the following, $\mathbf{P}_\nu$ denotes the distribution of a Markov chain with initial distribution ν. By abuse of notation, $\mathbf{P}_x$ represents the distribution of the Markov chain starting from $x \in E$. Since E is at most countable, we can always number the states, using an injection between E and N. Thus we can assume that $E \subset \mathbf{N}$. Therefore, we can use the formalism of vectors and matrices, even if it may be necessary to handle such objects with an infinite number of components... We often consider the "vector" π_n defined by $\pi_n(i) = \mathbf{P}(X_n = i)$ for $i \in E \subset \mathbf{N}$. It is common to consider it as a row vector. For all n, for all $j \in E$, the relation

$$\begin{aligned}\mathbf{P}(X_{n+1} = j) &= \sum_{i \in E} \mathbf{P}(X_{n+1} = j \,|\, X_n = i)\mathbf{P}(X_n = i) \\ &= \sum_{i \in E} \mathbf{P}(X_n = i)p(i,j),\end{aligned}$$

reads in matrix notation

$$\pi_{n+1} = \pi_n.P \text{ thus } \pi_n = \pi_0.P^n, \qquad [3.3]$$

where P^n is the nth power of P. For instance, if π_0 is composed only of 0 but a 1 in ith position (that is to say $\nu = \mathbf{P}_i$) then for any $j \in E$,

$$\mathbf{P}_i(X_n = j) = p^{(n)}(i,j),$$

where $p^{(n)}(i,j)$ is the term in ith row and jth column of P^n.

As $P^{n+m} = P^n P^m$, we deduce from [3.3] the so-called Chapman-Kolmogorov equation

$$p^{(n+m)}(x,y) = \sum_{z \in E} p^{(n)}(x,z)p^{(m)}(z,y), \qquad [3.4]$$

valid for any n, m, any initial condition and any final state. Note this equation is written "intrinsically", that is to say regardless of the injection mentioned above.

3.1.1. *Simulation*

Let us first recall how to simulate ν a distribution on a denumerable set E. The states are supposed to be numbered with an bijection ϕ between E and a subset of $\mathbf{N}$. Then we put

$$r_0 = \nu(\{\phi^{-1}(0)\}) \text{ and } r_n = \sum_{j=0}^{n} \nu(\{\phi^{-1}(j)\}) = \nu(\phi^{-1}(\{0,\dots,n\})).$$

Algorithm 3.1. Realization of a random variable of distribution ν

```
Data: $r_0, r_1, \ldots$
Result: An element of $E$ chosen according to the distribution $\nu$
x← sample of a uniform distribution on $[0,1]$;
n← 0;
while $x > r_n$ do
    $n \leftarrow n+1$
end
return $\phi^{-1}(n)$
```

When a Markov chain X is in state x, it moves to state y with probability $p(x,y)$. To move from one stage to another, we simply have to apply the previous algorithm to the distribution $\mu_x = (p(x,y), y \in E)$.

Algorithm 3.2. Simulation of a trajectory of a Markov chain (ν, P)

```
Data: $\nu$, $P$, $N$
Result: A path of length $N$ of the Markov chain $(\nu, P)$
Choose $x_0$ initial condition according to $\nu$;
for counter ← 1 to $N$ do
    Choose $x_{\text{counter}}$ according to the distribution $(p(x_{\text{counter}-1}, y), y \in E)$;
end
return $x_0, x_1, \ldots, x_N$
```

3.2. Strong Markov property

For a stopping time T, on $(T < \infty)$, we define θ^T by

$$\theta^T(\omega) = (\omega_{T(\omega)}, \omega_{T(\omega)+1}, \ldots).$$

For $x \in E$, the visiting times to x are defined by

$$\tau_x^1 = \begin{cases} \infty & \text{if } X_n \neq x \text{ for all } n > 0, \\ \inf\{n > 0, X_n = x\} & \text{otherwise ;} \end{cases}$$

$$\tau_x^k = \begin{cases} \infty & \text{if } \tau_x^{k-1} = \infty \text{ or } X_n \neq x \text{ for all } i > \tau_x^{k-1}, \\ \inf\{n > \tau_x^{k-1}, X_n = x\} & \text{otherwise.} \end{cases}$$

LEMMA 3.2.– *For x fixed in E, on $(\tau_x^1 < \infty)$, we have*

$$\tau_x^k = \tau_x^{k-1} + \tau_x^1 \circ \theta^{\tau_x^{k-1}}. \qquad [3.5]$$

Proof. If $\tau_x^{k-1} = \infty$, then we have ∞ on both sides of the equality. If $\tau_x^{k-1} < \infty$, the result is immediate once we are convinced that $\theta^{\tau_x^{k-1}}(\omega)$ represents the part of the trajectory posterior to the $(k-1)$th visit to the state x. Therefore, the first visit (if any) after the $(k-1)$th is the kth visit since the beginning. □

THEOREM 3.3.– *Let T be an a.s. finite stopping time and $F\colon \Omega \to \mathbf{R}^+$ an integrable random variable. Then, we have the following identity*

$$\mathbf{E}\left[F \circ \theta^T \mid \mathcal{F}_T\right] = \mathbf{E}\left[F \mid X_0 = X_T\right]. \qquad [3.6]$$

To calculate the right hand side, we first calculate $\mathbf{E}\left[F \mid X_0 = x\right] = \phi(x)$ and we set

$$\mathbf{E}\left[F \mid X_0 = X_T\right] = \phi(X_T).$$

Proof. Since $A \in \mathcal{F}_T$, $A \cap \{T = n\} \in \mathcal{F}_n$. Moreover, using [3.2] and the properties of conditional expectation, we have

$$\begin{aligned}
\mathbf{E}\left[F \circ \theta^T . \mathbf{1}_A\right] &= \sum_{n=0}^{\infty} \mathbf{E}\left[F \circ \theta^n . \mathbf{1}_{A \cap \{T=n\}}\right] \\
&= \sum_{n=0}^{\infty} \mathbf{E}\left[\mathbf{E}\left[F \circ \theta^n \mid \mathcal{F}_n\right] \mathbf{1}_{A \cap \{T=n\}}\right] \\
&= \sum_{n=0}^{\infty} \mathbf{E}\left[\mathbf{E}\left[F \mid X_0 = X_n\right] \mathbf{1}_{A \cap \{T=n\}}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[F \mid X_0 = X_T\right] \mathbf{1}_A\right].
\end{aligned}$$

The equality is true by linearity for all step functions, and thus for all positive functions. □

NOTE.– Let X be the Markov chain with two states 0 and 1, and transition matrix $p_{0,0} = 0.9, p_{1.1} = 1$. Let $T = \sup\{n \geq 1, X_n = 0\}$. Under $\mathbf{P}_0$, $T = Y - 1$ where Y has a geometric distribution of parameter 0.1, and therefore T is almost surely finite. However, $\mathbf{P}_0(X_{T+1} = 1 \mid X_T = 0) = 1$, which is different from $\mathbf{P}_0(X_{n+1} = 1 | X_n = 0) = 0.1$.

This example illustrates that one cannot avoid the "T stopping time" hypotheses in the strong Markov property. Here, it is clear that T is not a stopping time since to know whether T is less than n requires knowing the trajectory after time n to be sure that it does not return to 0 after time n.

EXAMPLE (Example 3.1 (continued)).– Suppose that there is a piece of cheese in box 3 and a battery in box 7. We want to calculate the probability that the rat could eat before being electrocuted. Consider the two random variables

$$\tau_3 = \inf\{n \geq 0, X_n = 3\} \text{ and } \tau_7 = \inf\{n \geq 0, X_n = 7\}.$$

For any $i \in \{1, \cdots, 7\}$, set $u_i = \mathbf{P}_i(\tau_3 < \tau_7)$. It is clear that $u_3 = 1$ and that $u_7 = 0$. For $i \notin \{3; 7\}$

$$u_i = \sum_{j=1}^{7} \mathbf{P}_i(\tau_3 < \tau_7 \,|\, X_1 = j)\mathbf{P}_i(X_1 = j).$$

Since i is different from 3 and 7, the event $\{\tau_3 < \tau_7\}$ is $\mathbf{P}_i$ a.s. equal to A_1 where

$$\begin{aligned} A_l &= \{\omega,\ \exists i \geq l \text{ such that } \omega_i = 3 \text{ and } \omega_j \in \{1, 2, 4, 5, 6\} \text{ for any } l \leq j < i\} \\ &= \{ \text{ after time } l\text{, we get 3 before 7 } \}. \end{aligned}$$

As $\mathbf{1}_{A_1} = \mathbf{1}_{A_0} \circ\theta$, we have

$$\mathbf{P}_i(\tau_3 < \tau_7 \,|\, X_1 = j) = \mathbf{P}_j(\tau_3 < \tau_7).$$

Since $\mathbf{P}_i(X_1 = j) = p(i, j)$, we see that $(u_i,\, i = 1, \cdots, 7)$ is the solution of the linear system

$$u_3 = 1,\ u_7 = 0,\ u_i = \sum_{j=1}^{6} p(i, j)u_j \text{ for } i \notin \{3; 7\}.$$

Solving this system gives $u_1 = 7/12, u_2 = 3/4, u_4 = 5/12, u_5 = 2/3, u_6 = 5/6.$

Without cheese and battery, let us now calculate the mean hitting time of box 3. For any $i \in \{1, \cdots, 7\}$, set $v_i = \mathbf{E}_i[\tau_3]$. It is clear $v_3 = 0$. Moreover, for $i \neq 3$, we have

$$\mathbf{E}_i[\tau_3] = \sum_{j=1}^{7} \mathbf{E}_i[\tau_3 \,|\, X_1 = j]\, p(i, j).$$

For the trajectory $\omega = (1, 2, 5, 2, 5, 6, 3, \ldots)$, $\tau_3(\omega) = 6$ but $\tau_3(\theta\omega) = 5$. More generally, given $X_0 \neq 3$, we have $\tau_3 = \tau_3 \circ \theta + 1$. Therefore,

$$v_i = \sum_{j=1}^{7} \big(\mathbf{E}_i[\tau_3 \circ \theta \,|\, X_1 = j] + 1\big)p(i, j) = \sum_{j=1}^{7} p(i, j)v_j + 1,$$

according to equation [3.2]. Hence, $(v_i,\, i = 1, \cdots, 7)$ is the solution of a linear system with six equations and six variables.

3.3. Classification of states

Let N_x denote the number of visits to state x, not including the initial state

$$N_x = \sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = x\}} .$$

LEMMA 3.4.– *For any k, the two events $\{N_x \geq k\}$ and $\left\{\tau_x^k < \infty\right\}$ coincide.*

Proof. $N_x \geq k$ means that there were more than k visits to the state x which is exactly equivalent to say that $\tau_x^k < \infty$. □

DEFINITION 3.3.– *A state x is called recurrent when $\mathbf{P}_x\left(\tau_x^1 < \infty\right) = 1$. Otherwise, x is said to be transient. The X chain is called recurrent (respectively transient) if all its states are recurrent (respectively transient).*

LEMMA 3.5.– *For any $(x, y) \in E \times E$, the following equality holds true,*

$$\mathbf{P}_y(\tau_x^k < \infty) = \mathbf{P}_x(\tau_x^1 < \infty)^{k-1}\mathbf{P}_y(\tau_x^1 < \infty). \tag{[3.7]}$$

In particular, if $x = y, \mathbf{P}_x\left(\tau_x^k < \infty\right) = \mathbf{P}_x(\tau_x^1 < \infty)^k$. Moreover,

$$\mathbf{E}_y\left[N_x\right] = \frac{\mathbf{P}_y(\tau_x^1 < \infty)}{1 - \mathbf{P}_x(\tau_x^1 < \infty)} = \sum_{n \geq 1} p^{(n)}(y, x). \tag{[3.8]}$$

By taking the current time as that of the kth visit to the state x, according to the strong Markov property, the past and the future given this visit are independent. Therefore, knowing that we have already visited k times the state x, the probability that one comes back to x a $(k+1)$th time is the same as during the first visit to x we return at least once. In addition, these two events are independent.

Proof. For $k > 2$, according to [3.2] and [3.6], we have

$$\begin{aligned}
\mathbf{P}_y(\tau_x^k < \infty) &= \mathbf{P}_y(\tau_x^{k-1} < \infty, \tau_x^1 \circ \theta^{\tau_x^{k-1}} < \infty) \\
&= \mathbf{E}_y\left[\mathbf{1}_{\{\tau_x^{k-1} < \infty\}}\, \mathbf{P}_y(\tau_x^1 \circ \theta^{\tau_x^{k-1} < \infty} \,|\, \mathcal{F}_{\tau_x^{k-1}})\right] \\
&= \mathbf{E}_y\left[\mathbf{1}_{\{\tau_x^{k-1} < \infty\}}\, \mathbf{P}_y(\tau_x^1 < \infty \,|\, X_0 = X_{\tau_x^{k-1}})\right] \\
&= \mathbf{E}_y\left[\mathbf{1}_{\{\tau_x^{k-1} < \infty\}}\, \mathbf{P}_y(\tau_x^1 < \infty \,|\, X_0 = x)\right] \\
&= \mathbf{E}_y\left[\mathbf{1}_{\{\tau_x^{k-1} < \infty\}}\, \mathbf{P}_y(\tau_x^1 < \infty)\right] \\
&= \mathbf{P}_y(\tau_x^{k-1} < \infty)\mathbf{P}_y(\tau_x^1 < \infty),
\end{aligned}$$

and we find [3.7] by induction.

Now, according to Fubini's theorem

$$\mathbf{E}_y[N_x] = \sum_{k\geq 1} \mathbf{P}_y(N_x \geq k) = \sum_{k\geq 1} \mathbf{P}_y(\tau_x^k < \infty),$$

and the first equality of [3.8] follows. Also according to Fubini's theorem and [3.4]

$$\mathbf{E}_y[N_x] = \mathbf{E}_y\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=x\}}\right] = \sum_{n=1}^{\infty} \mathbf{E}_y\left[\mathbf{1}_{\{X_n=x\}}\right] = \sum_{n=1}^{\infty} p^{(n)}(y,x).$$

Hence the result. □

The following theorem gives different characterizations of recurrence and transience.

THEOREM 3.6.– *Let x be a fixed state. Then:*

1) The following assertions are equivalent:

a) x is recurrent;

b) $\mathbf{P}_x(N_x = \infty) = 1$;

c) $\mathbf{E}_x[N_x] = \infty$.

2) The following assertions are equivalent:

a) x is transient;

b) $\mathbf{P}_x(N_x < \infty) = 1$;

c) $\mathbf{E}_x[N_x] < \infty$.

Proof. First let us show that $a \Rightarrow b$. According to [3.4] and Lemma 3.5

$$\mathbf{P}_x(N_x > k) = \mathbf{P}_x(\tau_x^k < \infty) = \mathbf{P}_x(\tau_x < \infty)^k, \qquad [3.9]$$

and according to the monotone convergence theorem, we have

$$\mathbf{P}_x(N_x = \infty) = \lim_{k\to\infty} \mathbf{P}_x(N_x > k). \qquad [3.10]$$

The recurrence of x means $\mathbf{P}_x(\tau_x < \infty)$ and hence implies $N_x = \infty$, $\mathbf{P}_x$ almost surely. Therefore, x recurrent implies that $\mathbf{P}_x(\tau_x^1 < \infty) = 1$. By the same argument, x transient implies $\mathbf{P}_x(\tau_x^1 < \infty) < 1$.

$b \Rightarrow c$. Easy when x is recurrent. For the other case, use the relation

$$\mathbf{E}_x[N_x] = \sum_{k\geq 0} \mathbf{P}_x(\tau_x < \infty)^k. \qquad [3.11]$$

As N_x is finite almost surely, according to [3.10], $\mathbf{P}_x(N_x > k)$ tends to 0 when k tends to infinity. According to [3.9] this implies that $\mathbf{P}_x(\tau_x < \infty) < 1$ so that the series converges.

$c \Rightarrow a$. In both cases, relation [3.11] leads to the conclusion. □

DEFINITION 3.4.– *We say that a state x communicates with a state y, what is denoted by $x \longrightarrow y$, if there is a strictly positive m integer such that $p^{(m)}(x, y) > 0$. This means that $\mathbf{P}_x(\tau_y^1 < \infty) = 1$.*

THEOREM 3.7.– *If x is a recurrent state and $x \longrightarrow y$, then $y \longrightarrow x$ and y is recurrent.*

Starting from x, we eventually reach y. If from y there is a risk of not coming back to x we eventually do not return to it, we therefore make only a finite number of visits to x, and there is a contradiction with the hypothesis of recurrence about x. Moreover, if from y we are almost certain to come back to x and that we pass an infinite number of times by x, we are likely to pass an infinite number of times to y as well.

Proof. Let us show, by contradiction, that y communicates with x by writing the probability of starting from x and never returning to x is greater than the probability of the same thing but visiting y at least once

$$\begin{aligned}\mathbf{P}_x(\tau_x = \infty) &\geq \mathbf{P}_x(\tau_x \circ \theta^{\tau_y} = \infty, \tau_y < \infty)\\ &= \mathbf{P}_x(\tau_y < \infty)\mathbf{P}_y(\tau_x = \infty),\end{aligned}$$

according to the strong Markov property. If y does not communicate with x, this quantity is positive which contradicts the recurrence of x. Similarly

$$\begin{aligned}\mathbf{P}_y(\tau_y < \infty) &\geq \mathbf{P}_y(\tau_y \circ \theta^{\tau_x} < \infty, \tau_x < \infty)\\ &= \mathbf{P}_y(\tau_x < \infty)\mathbf{P}_x(\tau_y < \infty) = 1,\end{aligned}$$

hence y is recurrent. □

THEOREM 3.8.– *The relation $\longrightarrow$ restricted to recurrent states is an equivalence relation.*

Proof. Reflexivity, that is, $x \longrightarrow x$, is induced by the very definition of a recurrent state. Symmetry, that is, $x \longrightarrow y \Longrightarrow y \longrightarrow x$, follows from Theorem 3.7. Let x, y and z three states of E such that $x \longrightarrow y$ and $y \longrightarrow z$. By definition, there are two positive integers which we call r and s such that $p^{(r)}(x, y) > 0$ and $p^{(s)}(y, z) > 0$. The Chapman-Kolmogorov equation tells us that

$$p^{(r+s)}(x, z) = \sum_{\ell \in E} p^{(r)}(x, \ell)p^{(s)}(\ell, z).$$

All the terms of this sum are non-negative and there is at least a positive term: $p^{(r)}(x, y)p^{(s)}(y, z)$. Thus, we have found a positive integer, $r + s$, such that $p^{(r+s)}(x, z) > 0$, hence the result. □

The set of recurrent points can be partitioned into equivalence classes. By definition, a state belonging to a class communicates with the other states of this class and does not communicate with any recurrent state belonging to another class neither with any transient state . In contrast, a transient state can communicate with both transient and recurrent states.

DEFINITION 3.5.– *A subset F of E is said to be closed, if for all x and y*

$$(x \in F \text{ and } x \longrightarrow y) \Longrightarrow y \in F.$$

In other words, $\sum_{y\in F} p(x,y) = 1$ for all $x \in F$.

THEOREM 3.9.– *A closed set of finite cardinal contains at least one recurrent point.*

Proof. Let F be a closed set. If all states are transient, we have

$$\mathbf{E}_y\left[N_x\right] = \mathbf{P}_y(\tau_x^1 < \infty)\mathbf{E}_x\left[N_x\right] < \infty,$$

for every pair (x, y) of F. Since F is finite, $\sum_{x\in F} \mathbf{E}_y\left[N_x\right] < \infty$. On the other hand

$$\sum_{x\in F} \mathbf{E}_y\left[N_x\right] = \sum_{x\in F} \mathbf{E}_y\left[\sum_{n\geq 0} \mathbf{1}_{\{X_n=x\}}\right] = \sum_{n\geq 0} \mathbf{E}_y\left[\sum_{x\in F} \mathbf{1}_{\{X_n=x\}}\right] = \sum_{n\geq 0} 1 = \infty,$$

since F is closed, thus a contradiction. Hence, there exists at least one recurrent point. □

EXAMPLE 3.3.– It is often easier to have a graphic representation of the transition matrix of a Markov chain. To do this, we construct a directed graph whose vertices correspond to the states. The x, y edge (oriented) has the weight of the transition probability from x to y. If this probability is zero, the edge does not exist. Let us consider the Markov chain transition matrix

$$P = \begin{pmatrix} 0.8 & 0.1 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{1}{1000} & \frac{999}{1000} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

Its graphic representation is that of Figure 3.2. It is clear that the sets $\{2, 3\}$ and $\{4, 5, 6\}$ are closed sets. Since they are both finite, both of them contain at least one recurrent point. Within each of them, all the points communicate with one another, so that they are all recurrent. If the chain leaves the point 1 to point 2 or 4, it is sure not to come back to 1 so that the probability of not returning to 1 is $2.1/10 > 0$, hence 1 is transient.

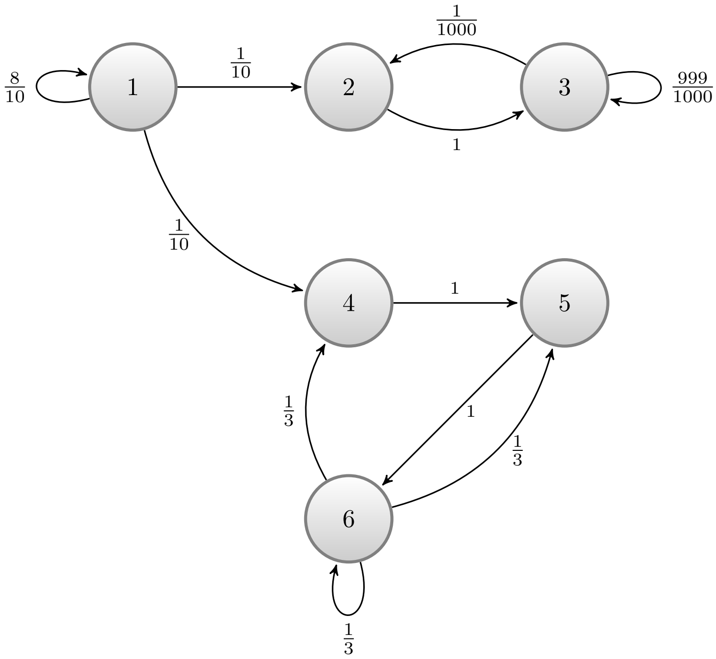

Figure 3.2. *Graphic representation of a Markov chain*

Let us observe that even if the set E is closed, Theorem 3.9 does not lead to a conflict over the status of state 1. In fact, we know that there exists at least one recurrent point in E, but we do not know which one and we cannot, *a priori*, say anything more since all the states do not communicate with each other.

EXAMPLE (Example 3.1 (continued)).– All the states communicate with one another so that the only closed subset is E itself. As it is finite, there is at least one recurrent state so that they are all recurrent.

EXAMPLE (Example 3.2 (continued)).– As long as it does not reach δ, the score can only grow, so that all the states of N are transient. δ is recurrent.

EXAMPLE.– A set of N cards is mixed by cutting it into two parts which are then interchanged. Each mixture of the set is represented by a permutation of $\{1, \dots, N\}$. If $N = 3$ and the mixture is represented by (3, 2, 1), meaning that the card 3 is in position 1, the card 2 in position 2 and the card 1 in position 3. X_n denotes the state of the deck of cards after the nth mixing. The state space is hence the group of permutations of $\{1, \dots, N\}$, denoted by $\mathfrak{S}_N$. If $X_0 = (3, 2, 1)$ and the cut is between the first and second card, we have $X_1 = (2, 1, 3)$. In other words, we have just made a circular permutation on the cards but did not change their relative order. In fact, to define the transition probabilities, let us consider the set of N elements of $\mathfrak{S}_N$

$$E_1 = \big\{\sigma \in \mathfrak{S}_N, \exists k \in \{1,\ldots,N\}, \sigma = (k+1,k+2,\ldots,N,1,\ldots,k)\big\}.$$

When we cut the pack at the level of the kth card we apply the cycle $(k+1,k+2,\ldots,N,1,\ldots,k)$ to the current situation. As the choice of the location of the cut is assumed to be uniform on $\{1,\ldots,N\}$ we have

$$\mathbf{P}(X_1 = \tau \,|\, X_0 = \sigma) = \frac{1}{N} \text{ if } \tau\sigma^{-1} \in E_1.$$

Equivalence classes of relation $\longrightarrow$ are those of the relation $\sigma \mathfrak{R} \tau \equiv \tau\sigma^{-1} \in E_1$. In other words, σ communicates with τ, if and only if there exists $\rho \in E_1$ such that $\tau = \rho\sigma$. Hence there are $(n-1)!$ equivalence classes of cardinal n each. All these classes form some closed sets of finite cardinal which contain all or at least one recurrent point. As all the states communicate with each other within these classes, they are all recurrent. Hence, the chain is recurrent.

When the state space is infinite, we cannot apply Theorem 3.9.

DEFINITION 3.6.– *A Markov chain is called irreducible when all the states communicate. In particular, the smallest closed subspace is E itself, and all the states have the same nature.*

NOTE.– If the number of transient states is finite, as we pass only a finite number of times in each of them, the Markov chain eventually enters a recurrence class and thus remains there. Let us observe that, according to Lemma 3.7, a recurrence class is always an irreducible subset. If the number of transient states is infinite, the above reasoning no longer applies automatically, but the cases in which a Markov chain never enters an irreducible closed subset, are out of our discussion. As far as we are interested in the asymptotic behavior of Markov chains, there is no loss of generality to assume that the Markov chains studied are irreducible.

When x is recurrent, we know that starting from x we eventually come back to x in a finite time, but what about the average time of return to x?

DEFINITION 3.7.– *A recurrent state x is said to be*

– *positive recurrent if* $\mathbf{E}_x\left[\tau_x^1\right] < \infty$;

– *null recurrent if* $\mathbf{E}_x\left[\tau_x^1\right] = \infty$.

The chain X is called positive recurrent (null recurrent respectively) if all its states are positive recurrent (null recurrent respectively).

The following construction is used several times thereafter.

DEFINITION 3.8.– *Let X be an irreducible Markov chain and recurring on E and F, a subset of E. Set*

$$\tau_F^1 = \inf\{n \geq 1, X_n \in F\} \text{ and } \tau_F^{k+1} = \tau_F^k + \tau_F^1 \circ \theta^{\tau_F^k},$$

the times of the successive visits to the set F. We consider the random sequence X^F, defined by $X_n^F = X_{\tau_F^n}, n \in \mathbf{N}$. We easily check that X^F is a Markov chain on F, called the Markov chain restricted to F.

THEOREM 3.10.– *Let X be an irreducible Markov chain and F a finite subset of E. If for any $x \in F, \mathbf{E}_x\left[\tau_F^1\right] < \infty$ then X is positive recurrent.*

Proof. For any $x \in F$, define $\sigma_x = \inf\{n > 0, X_n^F = x\}$ and for any $k \in \mathbf{N}^*$, $Y_k = \tau_F^k - \tau_F^{k-1}$. Since F is finite, then X^F is positive recurrent $\mathbf{E}_x\left[\sigma_x\right] < \infty$ for any $x \in F$. We must prove that $\mathbf{E}_x\left[\tau_x\right] < \infty$. By the very definition of Y_k

$$\begin{aligned}
\mathbf{E}_x\left[\tau_x\right] &= \mathbf{E}_x\left[\sum_{k=1}^{\sigma_x} Y_k\right] \\
&= \sum_{n\geq 1} \mathbf{E}_x\left[\sum_{k=1}^{\sigma_x} Y_k \, \mathbf{1}_{\{\sigma_x = n\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbf{E}_x\left[Y_k \sum_{n \geq k} \mathbf{1}_{\{\sigma_x = n\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbf{E}_x\left[Y_k \, \mathbf{1}_{\{\sigma_x \geq k\}}\right].
\end{aligned}$$

Using the strong Markov property, we obtain

$$\begin{aligned}
\mathbf{E}_x\left[Y_k \, \mathbf{1}_{\{\sigma_x \geq k\}}\right] &= \mathbf{E}_x\left[\mathbf{E}_x\left[Y_k \,|\, \mathcal{F}_{\tau_{k-1}^F}\right] \mathbf{1}_{\{\sigma_x \geq k\}}\right] \\
&= \mathbf{E}_x\left[\mathbf{E}_{X_{\tau_F^{k-1}}}\left[Y_1\right] \mathbf{1}_{\{\sigma_x \geq k\}}\right] \\
&\leq \sup_{y \in F} \mathbf{E}_y\left[Y_1\right] \mathbf{P}_x(\sigma_x \geq k).
\end{aligned}$$

Since F is finite, the supremum is finite. We therefore obtain

$$\mathbf{E}_x\left[\tau_x\right] \leq c \sum_{k=1}^{\infty} \mathbf{P}_x(\sigma_x \geq k) = c\mathbf{E}_x\left[\sigma_x\right].$$

According to the initial observation, this proves the positive recurrence of X. □

LEMMA 3.11.– *Let X be a Markov chain and $h: E \times E \longrightarrow \mathbf{R}$, bounded. For any integer n*

$$\mathbf{E}\left[h(X_n, X_{n+1}) \,|\, \mathcal{F}_n\right] = P(h(X_n, .))(X_n) = \sum_{y \in E} p(X_n, y) h(X_n, y). \quad [3.12]$$

Proof. Since h is bounded, we only have to compute the conditional expectation. According to the Markov property

$$\mathbf{E}\left[h(X_n, X_{n+1}) \,|\, \mathcal{F}_n\right] = \mathbf{E}\left[h(X_n, X_{n+1}) \,|\, X_n\right].$$

Now, let $\phi: E \to \mathbf{R}$ be bounded,

$$\begin{aligned}\mathbf{E}\left[h(X_n, X_{n+1})\phi(X_n)\right] &= \int \phi(x) \int h(x, y) \,\mathrm{d}\,\mathbf{P}_{X_{n+1} \,|\, X_n = x}(y) \,\mathrm{d}\,\mathbf{P}_{X_n}(x) \\ &= \int \phi(x) \sum_{y \in E} h(x, y) p(x, y) \,\mathrm{d}\,\mathbf{P}_{X_n}(x) \\ &= \mathbf{E}\left[\phi(X_n) P(h(X_n, .))(X_n)\right].\end{aligned}$$

The previous equation is true for any function, thus [3.12] holds true. □

THEOREM 3.12 (Foster criterion of recurrence).– *Let E_0 be a finite subset of E. Assume there exists a function $h: E \to \mathbf{R}$ such that $\{x \in E, h(x) < K\}$ is finite for any K finite and that*

$$h(y) \geq \mathbf{E}_y\left[h(X_1)\right] \text{ for all } y \in E_0^c.$$

Then, X is recurrent.

Proof. According to the hypothesis, h is lower bounded hence up to an additive constant, we can assume $h \geq 0$. Consider the stopping time $\tau = \inf\{n, X_n \in E_0\}$ and the sequence Y defined by $Y_n = h(X_n)\,\mathbf{1}_{\{\tau > n\}}$. Let us show that Y is a positive supermartingale as soon as $X_0 \in E_0^c$. Let $x \in E_0^c$

$$\begin{aligned}\mathbf{E}_x\left[h(X_{n+1})\,\mathbf{1}_{\{\tau > n+1\}} \,|\, \mathcal{F}_n\right] &= \mathbf{1}_{\{\tau > n+1\}}\,\mathbf{E}_{X_n}\left[h(X_{n+1})\right] \\ &\leq \mathbf{1}_{\{\tau > n\}}\, h(X_n) = Y_n,\end{aligned}$$

since on $\{\tau > n+1\}$, X_n does not belong to E_0. Thus, Y converges almost surely to a random variable Y_∞.

Suppose that X is transient. Let $x \notin E_0$, for any K, the set $\{x, h(x) < K\}$ is finite and is thus visited by X only a finite number of times. Hence X is not bounded.

As Y_∞ is finite, τ is necessarily finite almost surely. This means that for $x \notin E_0$, $\mathbf{P}_x(\tau < \infty) = 1$. Starting from E_0^c, the Markov chain eventually reaches E_0. Either we stay in E_0 forever and as E_0 is finite, E_0 is recurrent and the chain is irreducible; or the chain leaves E_0 and in accordance with what we have just proved, it eventually returns to it. The number of visits to E_0 is hence infinite, which implies again that E_0 is recurrent. This the Markov chain is recurrent. □

3.4. Invariant measures and invariant probability

DEFINITION 3.9.– *Let E be a countable set and P a transition operator on $E \times E$. A positive finite measure ν on E is said to be invariant with respect to P if and only if*

$$\nu = \nu P \text{ that is to say } \nu(y) = \sum_{x \in E} \nu(x)p(x,y) \text{ for all } y \in E. \qquad [3.13]$$

If moreover $\sum \nu(x) = 1$, ν is an invariant probability.

NOTE.– If $\pi_0 = \nu$ then $\pi_n = \pi_0 P^n = \pi_0$.

THEOREM 3.13.– *Let x be a recurrent state, then the measure ν defined by*

$$\nu(y) = \mathbf{E}_x \left[\sum_{n=0}^{\tau_x^1 - 1} \boldsymbol{1}_{\{X_n = y\}} \right] = \sum_{n=0}^{\infty} \mathbf{P}_x(X_n = y, \tau_x^1 > n)$$

is an invariant measure.

Proof. Let us first show the equality of the two expressions of ν. Since x is recurrent, τ_x^1 is almost surely finite then $\cup_{n \geq 1} \{\tau_x^1 = n\}$ is a partition of Ω. According to Fubini Theorem

$$\begin{aligned}
\mathbf{E}_x \left[\sum_{n=0}^{\tau_x^1 - 1} \mathbf{1}_{\{X_n = y\}} \right] &= \sum_{\ell=1}^{\infty} \mathbf{E}_x \left[\sum_{n=0}^{\ell - 1} \mathbf{1}_{\{X_n = y\}} \mathbf{1}_{\{\tau_x^1 = \ell\}} \right] \\
&= \sum_{n=0}^{\infty} \mathbf{E}_x \left[\sum_{\ell > n} \mathbf{1}_{\{\tau_x^1 = \ell\}} \mathbf{1}_{\{X_n = y\}} \right] \\
&= \sum_{n=0}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 > n\}} \mathbf{1}_{\{X_n = y\}} \right].
\end{aligned}$$

Under $\mathbf{P}_x$, $X_0 = X_{\tau_x^1} = x$, thus we can write $\nu(y) = \mathbf{E}_x \left[\sum_{n=1}^{\tau_x^1} \mathbf{1}_{\{X_n = y\}} \right]$, which gives the same calculations with a different index range

$$\nu(y) = \sum_{n=1}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{1}_{\{X_n = y\}} \right]. \qquad [3.14]$$

We have already observed that the event $\{\tau_x^1 \geq n\}$ belongs to $\mathcal{F}_{n-1}$ since it is the complementary of the event $\{\tau_x^1 \leq n-1\}$. For $y \neq x$, by using the properties of conditional expectation and the strong Markov property

$$\begin{aligned}
&\sum_{n=0}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{1}_{\{X_n = y\}}\right] \\
&\qquad = \sum_{n=1}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{E}_x \left[\mathbf{1}_{\{X_n=y\}} \,|\, \mathcal{F}_{n-1}\right]\right] \\
&\qquad = \sum_{n=1}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{E}_x \left[\mathbf{1}_{\{X_n=y\}} \,|\, X_{n-1}\right]\right] \\
&\qquad = \sum_{n=1}^{\infty} \sum_{z \in E} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{1}_{\{X_{n-1}=z\}} \mathbf{E}_x \left[\mathbf{1}_{\{X_n=y\}} \,|\, X_{n-1}=z\right]\right] \\
&\qquad = \sum_{z \in E} p(z,y) \sum_{n=1}^{\infty} \mathbf{E}_x \left[\mathbf{1}_{\{\tau_x^1 \geq n\}} \mathbf{1}_{\{X_{n-1}=z\}}\right] \\
&\qquad = \sum_{z \in E} \nu(z) p(z,y).
\end{aligned}$$

For $y = x$, it is clear that $\nu(x) = 1$ and on the other hand

$$\begin{aligned}
\sum_{z \in E} \nu(z) p(z,x) &= \sum_{n=0}^{\infty} \sum_{z \in E} p(z,x) \mathbf{P}_x(X_n = z, \tau_x^1 > n) \\
&= \sum_{n=0}^{\infty} \sum_{z \in E} \mathbf{P}_x(X_n = z, X_{n+1} = x, \tau_x^1 > n) \\
&= \sum_{n=0}^{\infty} \sum_{z \in E} \mathbf{P}_x(X_n = z, \tau_x^1 = n+1) \\
&= \sum_{n=0}^{\infty} \mathbf{P}_x(\tau_x^1 = n+1) \\
&= \mathbf{P}_x(\tau_x^1 < \infty) = 1.
\end{aligned}$$

Hence we have $\nu = \nu P$, and it only remains to verify that $\nu(y) < \infty$ for any y. This is true for $x = y$. For $y \neq x$, either x does not communicate with y and hence $\nu(y) = 0$, or x communicates with y and as x is recurrent, according to Theorem 3.7, y communicates x, that is to say that there exists m such that $p^{(m)}(y,x) > 0$. As ν is invariant, $\nu.P^m = \nu$, which implies that

$$1 = \nu(x) = \sum_{z \in E} \nu(z) p^{(m)}(z, x) \geq \nu(y) p^{(m)}(y, x),$$

and therefore $\nu(y) < \infty$. □

COROLLARY 3.14.– *Let X be an irreducible recurrent Markov chain of invariant measure ν. Let F be a subset of E and X^F the chain restricted to F. Then, X^F is irreducible and recurrent and admits as an invariant measure the measure given Theorem 3.13.*

Proof. The first two points are obvious. For $y \in F$, the number of visits to y of X^F and of X are the same, hence X and X^F have the same invariant measure given by Theorem 3.13. □

THEOREM 3.15.– *If the Markov chain X is irreducible and recurrent, then there exists a unique (up to a multiplicative constant) invariant measure ν such that for any y, $0 < \nu(y) < \infty$. The uniqueness "up a to a multiplicative constant" means that if ν and ν' are two such measures then there exists $c > 0$ such that $\nu(x) = c\nu'(x)$ for any $x \in E$.*

Proof. Let μ an invariant measure and let $a \in E$. Let ν be the invariant measure constructed in Theorem 3.13 with a as starting point. By construction, $\nu(a) = 1$ then for any invariant measure μ, $\mu(a) = \nu(a)\mu(a)$. By definition, for $z \in E \backslash \{a\}$,

$$\mu(z) = \sum_{y \in E} \mu(y) p(y, z) = \mu(a) p(a, z) + \sum_{y \neq a} \mu(y) p(y, z).$$

By iterating this relation,

$$\mu(z) = \mu(a) p(a, z) + \mu(a) \sum_{y \neq a} p(a, y) p(y, z) + \sum_{i \neq a} \sum_{y \neq a} \mu(x) p(x, y) p(y, z).$$

This can be rewritten

$$\mu(z) = \mu(a) \mathbf{P}_a(X_1 = z)$$
$$+ \mu(a) \mathbf{P}_a(X_1 \neq a, X_2 = z) + \mathbf{P}_\mu(X_0 \neq a, X_1 \neq a, X_2 = z).$$

By induction on n, for any n,

$$\mu(z) = \mu(a) \sum_{m=1}^{n} \mathbf{P}_a(\tau_a^1 > m, X_m = z) + \mathbf{P}_\mu \left(\bigcap_{y=0}^{n} (X_y \neq a) \cap X_n = z \right).$$

The last probability is a positive term and when n tends to infinity, we recognize in the first sum the definition of ν. Hence:

$$\mu(z) \geq \mu(a) \nu(z) \text{for any } z \in E.$$

On the other hand, since for any n, $\mu.P^n = \mu$, we also have

$$\mu(a) = \sum_x \mu(x)p^{(n)}(x,a) \geq \mu(a)\sum_x \nu(x)p^{(n)}(x,a) = \mu(a)\nu(a) = \mu(a).$$

Therefore, the intermediate inequality is an equality and as $\mu(x) \geq \mu(a)\nu(x)$, we must have $\mu(x) = \mu(a)\nu(x)$ when n is such that $p^{(n)}(x,a) > 0$. Given that X is irreducible, such an integer n always exists and thus $\mu(x) = \mu(a)\nu(x)$ for any $x \in E$. □

THEOREM 3.16.– *If there is an invariant probability then all the states satisfying $\nu(y) > 0$ are recurrent.*

Proof. As $\nu = \nu P^n$, Fubini's theorem implies that

$$\sum_{x\in E}\nu(x)\sum_{n\geq 1}p^{(n)}(x,y) = \sum_{n\geq 1}\nu(y) = \infty \text{ if } \nu(y) > 0.$$

On the other hand, according to Lemma 3.5

$$\sum_{n\geq 1}p^{(n)}(x,y) = \frac{\mathbf{P}_x(\tau_y^1 < \infty)}{1-\mathbf{P}_y(\tau_y^1 < \infty)}.$$

As $\mathbf{P}_x(\tau_y^1 < \infty) \leq 1$,

$$\infty \leq \sum_{x\in E}\nu(x).\frac{1}{1-\mathbf{P}_y(\tau_y^1 < \infty)}.$$

Therefore $\mathbf{P}_y(\tau_y^1 < \infty) = 1$ since ν is finite, which means that y is recurrent. □

THEOREM 3.17.– *If X is irreducible and admits ν as invariant probability, then*

$$\nu(x) = \frac{1}{\mathbf{E}_x\left[\tau_x^1\right]}.$$

Proof. If there exists x such that $\nu(x) = 0$ then as for any n

$$\nu(x) = \sum_{y\in E}p^{(n)}(y,x)\nu(y),$$

for any n and any y, the product of $\nu(y)$ and of $p^{(n)}(y,x)$ is zero. Now, the chain is irreducible, hence for any y, there exists n_y such that $p^{(n_y)} > 0$ so that $\nu(y) = 0$. Therefore ν is not a probability, thus for any $x \in E, \nu(x) > 0$. According to the previous theorem, all states are recurrent. Hence we know that ν is, up to a multiplicative constant, given by Theorem 3.13. This multiplicative constant c must satisfy $c\sum_{y\in E}\nu(y) = 1$,. Since we know that for any $x \in E$,

$$\sum_{y\in E}\nu(y) = \sum_{y\in E}\mathbf{E}_x\left[\sum_{n=0}^{\tau_x^1-1}\mathbf{1}_{\{X_n=y\}}\right] = \mathbf{E}_x\left[\tau_x^1\right],$$

the result follows. □

The following theorem summarizes the above mentioned principal results.

THEOREM 3.18.– *If X is irreducible, the three following assertions are equivalent:*

1) One of the states is positive recurrent;

2) There is an invariant probability;

3) All the states are positive recurrent.

Moreover, the invariant probability is given by

$$\nu(y) = \frac{1}{\mathbf{E}_x\left[\tau_x^1\right]}\mathbf{E}_x\left[\sum_{n=0}^{\tau_x^1-1}\boldsymbol{1}_{\{X_n=y\}}\right].$$

Proof. 1) ⇒ 2). By combining Theorem 3.13 and Theorem 3.17, we see that

$$\nu(y) = \frac{1}{\mathbf{E}_x\left[\tau_x^1\right]}\mathbf{E}_x\left[\sum_{n=0}^{\tau_x^1-1}\mathbf{1}_{\{X_n=y\}}\right],$$

defines an invariant probability. As the term on the left does not depend on x, we can choose $x = y$ and we find $\nu(y) = \mathbf{E}_y\left[\tau_y^1\right]^{-1}$.

2)⇒ 3). Since X is irreducible, we know that the invariant probability is a multiple of that built in Theorem 3.13 and therefore $\pi(y) > 0$ for any $y \in E$. According to Theorem 3.17, this means that all the states are positive recurrent.

3) ⇒ 1) is trivial. □

COROLLARY 3.19.– *Any irreducible Markov chain on a finite cardinal E is positive recurrent.*

Proof. There exists an invariant measure μ. As the state space is finite, we can always normalize it by requiring

$$\nu(x) = \frac{1}{\sum_{y\in E}\mu(y)}\mu(y),$$

and we obtain an invariant probability. According to point 2) of the previous theorem, we deduce that the Markov chain is positive recurrent. □

THEOREM 3.20 (Foster criterion of positive recurrence).– *Assume that there exists* $h: E \to \mathbf{R}$ *and* $\epsilon > 0$ *such that:*

– $\liminf_y h(y) > -\infty$*;*

– $h(X_1)$ *is integrable;*

– *for any* $y \in E_0^c$,

$$h(y) - \epsilon \geq \mathbf{E}\left[h(X_1) \,|\, X_0 = y\right].$$

Under these conditions, X *is positive recurrent.*

Let X be an irreducible recurrent Markov chain on a Polish space E. For any $x \in E$ we put, $(Y_k^n, k \geq 0) = \left(X_{k\wedge\tau_x^1}\right)$ and for any $n \geq 1$, we define the nth excursion Y^n of X from x by $(Y_k^n, k \geq 0)$, where

$$Y_k^n = X_{k\wedge\tau_x^1} \circ \theta^{\tau_x^n}.$$

The 0th excursion coincides with X until the first visit to x, after Y^0 remains in x. The nth excursion is a Markov chain which starts from x and follows the pattern of the initial chain until it hits x. Then, it remains equal to x. The evolution of the chain is then captured by Y^{n+1}.

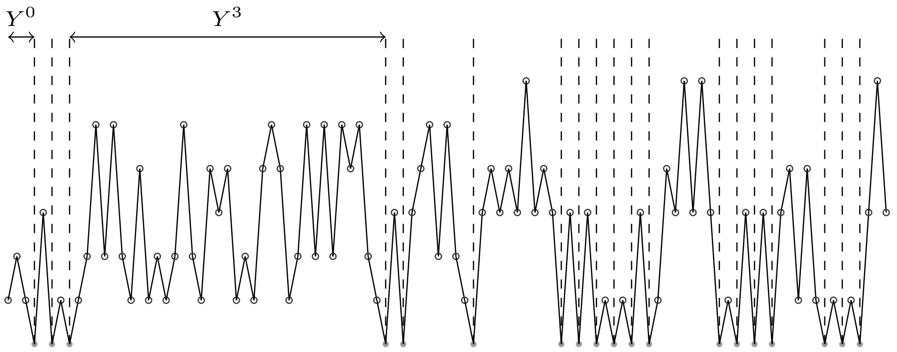

Figure 3.3. *Excursions of the Markov chain of example 3.1*

According to the strong Markov property, the processes $(Y^n, n \geq 0)$ are independent and for $n \geq 1$, they all have the same law: for any function $\psi: E^{\mathbf{N}} \longrightarrow \mathbf{R}$,

$$\mathbf{E}\left[\psi(Y^n)\right] = \mathbf{E}\left[\psi(Y^1 \circ \theta^{\tau_x^n})\right] = \mathbf{E}\left[\psi(Y^1)\right].$$

THEOREM 3.21.– *Let X be recurrent, irreducible of invariant distribution ν. For any initial condition $x \in E$, for any function f in $L^1(\nu)$, for any function $g \geq 0$ such that $\sum_y g(y)\nu(y) > 0$, we have*

$$\frac{\sum_{j=0}^n f(X_j)}{\sum_{j=0}^n g(X_j)} \xrightarrow{n\to\infty} \frac{\sum_{y\in E} f(y)\nu(y)}{\sum_{y\in E} g(y)\nu(y)},\ \mathbf{P}_x \text{ a.s.}$$

As a consequence, for $f \in L^1(\nu)$,

$$\frac{1}{n}\sum_{j=0}^n f(X_j) \xrightarrow{n\to\infty} \sum_{y\in E} f(y)\nu(y),\ \mathbf{P}_x \text{ a.s.}$$

We can cut any additive functional into pieces depending on each excursion. According to the independence and equidistribution of these pieces, we can apply the strong law of large numbers. It remains to prove that the side-effects are negligible, that is to say that the term which depends on Y^0 and the term which depends on the incomplete excursion disappear when divided by n.

Proof. The invariant probability is proportional to the invariant measure given in Theorem 3.13. Thus, there exists $c > 0$ such that for any function $g \geq 0$,

$$c\sum_{y\in E} g(y)\nu(y) = \mathbf{E}_x\left[\sum_{n=0}^{\tau_x^1-1} g(X_n)\right].$$

By homogeneity we can assume that $c = 1$. Let $Z = (Z_k, k \geq 1)$ is defined by

$$Z_k = \sum_{n=\tau_x^k}^{\tau_x^k-1} f(X_n) = \sum_{n=0}^{\tau_x^1-1} f(Y_n^k) = Z_1 \circ \theta^{\tau_x^k}.$$

According to the strong Markov property, the random variables $(Z_k, k \geq 1)$ are independent and identically distributed. Moreover,

$$\mathbf{E}_x\left[|Z_1|\right] \leq \mathbf{E}_x\left[\sum_{n=0}^{\tau_x^1-1} |f(X_n)|\right] = \sum_{y\in E} \mid f(y) \mid \nu(y) < \infty,$$

since $f \in L^1(\nu)$. We can therefore apply the strong law of large numbers, which states that

$$\frac{1}{n}\sum_{k=0}^{\tau_x^n-1} f(X_k) = \frac{1}{n}\sum_{k=1}^{n} Z_k \xrightarrow{n\to\infty} \mathbf{E}_x\left[Z_1\right] = \sum_{y\in E} f(y)\nu(y),\ \mathbf{P}_x \text{ a.s.} \quad [3.15]$$

If we apply this result to $f \equiv 1$, we obtain

$$\frac{\tau_x^n}{n} \xrightarrow{n\to\infty} 1, \ \mathbf{P}_y \text{ a.s.} \qquad [3.16]$$

Let $e(n)$ be the number of visits to x between 0 and n times. Let us observe that $e(n)$ is also the number of excursions starting from x completed before time n. By definition, $\tau_x^{e(n)} \le n < \tau_x^{e(n)+1}$, therefore

$$\frac{\tau_x^{e(n)}}{e(n)} \le \frac{n}{e(n)} < \frac{\tau_x^{e(n)+1}}{e(n)+1}\,\frac{e(n)+1}{e(n)}.$$

According to [3.16], the lower and upper bounds of the previous line converge a.s. to 1, therefore so does $n^{-1}e(n)$. Hence,

$$\begin{aligned}&\frac{1}{n}\sum_{k=0}^{n} f(X_k)\\ &= \frac{1}{n}\sum_{k=0}^{\tau_x^1-1} f(X_k)\frac{e(n)}{n}\frac{1}{e(n)}(Z_1+\cdots+Z_{e(n)}) + \frac{1}{n}\sum_{k=\tau_x^{e(n)}+1}^{n} f(X_k).\end{aligned} \qquad [3.17]$$

Assume that $f \ge 0$. The first term tends to 0 $\mathbf{P}_y$-almost surely for any $y \in E$. According to the definition of almost-sure convergence, this is equivalent to

$$\mathbf{P}_y\left(\limsup_n \left(\frac{1}{n}\sum_{k=0}^{\tau_x^1-1} f(X_k) > \epsilon\right)\right) = 0, \qquad [3.18]$$

for any $\epsilon > 0$. Therefore, by the strong Markov property, for any $\epsilon > 0$,

$$\begin{aligned}&\mathbf{P}_y\left(\limsup_n \left(\frac{1}{n}\sum_{k=\tau_x^{e(n)}}^{\tau_x^{e(n+1)}-1} f(X_k) > \epsilon\right)\right)\\ &\qquad = \mathbf{P}_x\left(\limsup_n \left(\frac{1}{n}\sum_{k=0}^{\tau_x^1-1} f(Y_k^n) > \epsilon\right)\right)\\ &\qquad = \mathbf{P}_x\left(\limsup_n \left(\frac{1}{n}\sum_{k=0}^{\tau_x^1-1} f(X_k) > \epsilon\right)\right) = 0,\end{aligned}$$

according to the [3.18]. As

$$\sum_{k=\tau_x^{e(n)}+1}^{n} f(X_k) \le \sum_{k=\tau_x^{e(n)}}^{\tau_x^{e(n+1)}-1} f(X_k),$$

we have

$$\frac{1}{n} \sum_{k=\tau_x^{e(n)}}^{n} f(X_k) \xrightarrow{n\to\infty} 0,\ \mathbf{P}_y \text{ a.s.}$$

For general f, by applying the above reasoning to $|f|$, we show that the first and third terms of [3.17] tend a.s. to 0. According to the first part of the proof (see equation [3.15]), the middle term of [3.17] tends almost surely toward $\sum_{y\in E} f(y)\nu(y)$. The special case is obtained by taking $g = 1$. □

DEFINITION 3.10.– *A state x is periodic if there exists an integer $\delta \ge 2$ such that*

$$\sum_{k=1}^{\infty} \mathbf{P}_x(\tau_x^1 = \delta k) = 1. \qquad [3.19]$$

The smallest δ satisfying [3.19] is called the period of x and we denote it as $d(x)$. The states that are not periodic are called aperiodic.

EXAMPLE (Example 3.1 (continued)).– In this case, a moment of thought shows that we reach only an odd-numbered box every two steps, and the same holds for even-numbered boxes: if the rat starts in box 1, it can be in boxes 3, 5 or 7 after its second move. Hence, the period is 2. We see that the states can be classified into two packets, boxes $\{1, 3, 5, 7\}$ on the one hand, boxes $\{2, 4, 6\}$ on the other.

More generally, we have the following theorem.

THEOREM 3.22.– *Let X be an irreducible recurrent Markov chain of period d. Let x be fixed in E, there exists a partition of E in d sets $C_0, C_1, \ldots, C_{d-1}$ such that:*

1) x belongs to C_0;

2) Let $y \in C_r$ and $z \in C_s$, if $p^{(n)}(y, z) > 0$ then $n = (s - r) \mod d$;

3) $C_0, \ldots, C_{d-1}$ are irreducible recurrent aperiodic classes for the Markov chain of transition matrix P^d.

The factorization is unique up to a renumbering. The Markov chain transition matrix P^d is irreducible, recurrent, aperiodic. If its initial condition is in C_r for $r \in [0, d-1]$, then all its subsequent states are in C_r.

The proof of this theorem requires two technical lemmas.

LEMMA 3.23.– *Let $a_1, \dots, a_n$ be relative prime integers, any integer $m \geq \prod_x (1+ax)$ can be written as*

$$m = \sum_k x_k a_k \text{ with } x_k \geq 0 \text{ for any } k. \qquad [3.20]$$

Proof. Let us show by induction on n that if $a_1, \dots, a_n$ are n integers (not necessarily relatively prime) and that $m \in \mathbf{N}$ is written as $m = \sum_k x_k a_k$ then we can always find another expression satisfying the conditions of [3.20]. Specifically, there exists a permutation σ of $\{1, \dots, N\}$ into itself such that

$$x_{\sigma(i)} \leq \min_{l \neq \sigma(1), \dots, \sigma(i-1)} (a_l) \text{ for all } l \leq n-1.$$

First assume that $n = 2$. As $m \geq 0$, one of the two coefficients x_1 or x_2 is positive. Without loss of generality, we can always assume that it concerns x_1. Let us show that we can always assume that $x_1 < a_2$. If this is not the case, then we can write it as $x_1 = ka_2 + r$ with $0 \leq r < a_2$

$$\begin{aligned} m &= x_1 a_1 + x_2 a_2 + k a_1 a_2 - k a_1 a_2 \\ &= (x_1 - k a_2) a_1 + (x_2 + k a_1) a_2 = r a_1 + (x_2 + k a_1) a_2. \end{aligned}$$

In conclusion, any integer m can be written as $m = x_1 a_1 + x_2 a_2$ with $0 \leq x < b$. As a consequence, if $m \geq a_1 a_2$, x_2 must be positive.

Assume the result is shown for $(n-1)$. Up to a renumbering, we can always assume that x_1 is positive, and apply the recurrence hypothesis to $m - x_1 a_1$ and to the $(n-1)$ remaining numbers. The renumbering that has been applied during this step defines the permutation σ.

Now, if $a_1, \dots, a_n$ are relatively prime, Bezout's lemma guarantees the existence of the representation $m = \sum_k x_k a_k$ for any integer. According to the first part of the demonstration, we can always assume that $\sum_{k \leq n-1} x_k a_k$ is positive and less than

$$\sup_x \left(a_1 \dots a_n + \prod_{y \neq x} a_y + \dots \right) \leq a_1 \dots a_n + \sum_x \prod_{y \neq x} a_y + \dots = \prod (1 + a_x) - 1.$$

Therefore, for m greater than or equal to $\prod(1 + a_x)$, there is still an expression of the form [3.20]. □

LEMMA 3.24.– *If x is aperiodic then there exists n_0 such that if $n \geq n_0$ then $p^{(n)}(x, x) > 0$.*

Proof. Define the set

$$I_x = \{n \in \mathbf{N}, p^{(n)}(x,x) > 0\}.$$

According to the Markov property, I_x is a semi-group: if m and n belong to I_x then $m+n$ also belong to it. Indeed

$$p^{(m+n)}(x,x) \geq p^{(m)}(x,x)p^{(n)}(x,x).$$

I_x is equipped with the usual order. Let u_n be the number of common divisors of the first n elements of I_x. $(u_n,\ n \in \mathbf{N})$ is a non-negative decreasing sequence, therefore it is convergent, and since x is aperiodic, its limit is 1. As $(u_n,\ n \in \mathbf{N})$ is integer valued, there must be a rank from which it is constant, let n_0 be this rank and let $a_1, \ldots, a_{n_0}$ the n_0 first elements of I_x. According to the previous lemma, for n sufficiently large, $n \in I_x$. □

Proof of Theorem 3.22.
Let $K_y = \{n, p^{(n)}(x,y) > 0\}$. For two integers k and l, according to the Markov property

$$\mathbf{P}_x(X_{k+l} = x) \geq \mathbf{P}_x(X_k = y)\mathbf{P}_y(X_l = x).$$

Therefore, n can belong to K_y only if d divides $n+l$, that is to say, if n is written as αdr where $r \in \{0, \ldots, d-1\}$ is the remainder of the division of l by d. We define C_r as the set of points of E which have the same r. These sets are clearly forming a partition and $x \in C_0$.

Let m and n such that $p^{(m)}(y,z) > 0$ and $p^{(n)}(x,y) > 0$. As $p^{(n+m)}(x,z) > 0$, it follows from i) that $n+m \equiv s \mod d$ and as $n \equiv r \mod d$, the result follows.

The irreducibility follows immediately from the previous point, the aperiodicity of the definition of the period. □

We can now state the result.

THEOREM 3.25.– *Let X be an irreducible, positive recurrent, aperiodic Markov chain of transition matrix P and invariant probability ν. Then,*

$$\lim_{n\to\infty} p^{(n)}(x,y) = \nu(y), \textit{ for all } x \textit{ and all } y.$$

This can be proved by coupling: two independent Markov chains of the same transition matrix but of different initial conditions always end up meeting each other. Let us observe that from this moment of meeting, they coincide in distribution.

Proof. On $E \times E$, we define the Markov chain $Z_n = (W_n, y_n)$ of transition matrix

$$\hat{p}\big((x_1, x_2), (y_1, y_2)\big) = p(x_1, y_1)p(x_2, y_2).$$

In other words, both W and Y coordinates evolve independently from one another according to the distribution of the original Markov chain.

We first show that Z is an irreducible Markov chain. As all states are aperiodic, according to Lemma 3.24, from a certain rank M,

$$p^{(l)}(y_1,\, y_1) > 0 \text{ and } p^{(l)}(x_2, x_2) > 0.$$

As X is irreducible and recurrent, there exists $K \geq M$ and $L \geq M$ such that

$$p^{(K)}(x_1, x_2) > 0 \text{ and } p^{(L)}(y_1, y_2) > 0.$$

Therefore, the path

$$(x_1, y_1) \to (x_2, y_1) \to (x_2, y_2)$$

has a positive probability for the index $K + L + M$. Indeed, according to the Markov property

$$p^{(K+L)}\big((x_1, y_1), (x_2, y_2)\big) \geq p^{(K)}(x_1, x_2)p^{(K)}(y_1, y_1).p^{(L)}(x_2, x_2)p^{(L)}(y_1, y_2) > 0.$$

It is clear that $\hat{\nu}(x, y) = \nu(x)\nu(y)$ defines an invariant probability for the Markov chain Z. Therefore according to Theorem 3.18, all states are positive recurrent. Let T be the hitting time of the diagonal of $E \times E$ by Z

$$\Delta = \{(x, y) \in E \times E, x = y\}$$
$$T = \inf \{n > 0, Z_n \in \Delta\}.$$

Since Z is irreducible, recurrent, the hitting time of a state (x, x) of the diagonal is almost surely finite. Since T is the minimum of all these hitting times, it is almost certainly finite. Let us show that on $\{T \leq n\}$, W_n and Y_n have the same distribution

$$\begin{aligned}
\mathbf{P}(W_n = y, T \leq n) &= \sum_x \mathbf{E}\left[\mathbf{1}_{\{W_n=y\}}\,\mathbf{1}_{\{W_T=x\}}\,\mathbf{1}_{\{T\leq n\}}\right] \\
&= \sum_x \mathbf{E}\left[\mathbf{1}_{\{W_T=x\}}\,\mathbf{1}_{\{T\leq n\}}\,\mathbf{E}\left[\mathbf{1}_{\{W_n=y\}}\mid \mathcal{F}_T\right]\right] \\
&= \sum_x \mathbf{E}\left[\mathbf{1}_{\{W_T=x\}}\,\mathbf{1}_{\{T\leq n\}}\,\mathbf{E}_x\left[\mathbf{1}_{\{W_{n-T}=y\}}\right]\right] \\
&= \sum_x \mathbf{E}\left[\mathbf{1}_{\{Y_T=x\}}\,\mathbf{1}_{\{T\leq n\}}\,\mathbf{E}_x\left[\mathbf{1}_{\{Y_{n-T}=y\}}\right]\right] \\
&= \mathbf{P}(Y_n = y, T \leq n).
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbf{P}(W_n = y) &= \mathbf{P}(W_n = y, T \leq n) + \mathbf{P}(W_n = y, T > n) \\
&= \mathbf{P}(Y_n = y, T \leq n) + \mathbf{P}(W_n = y, T > n) \\
&\leq \mathbf{P}(Y_n = y) + \mathbf{P}(W_n = y, T > n).
\end{aligned}$$

Symmetrically, we have

$$\mathbf{P}(Y_n = y) \leq \mathbf{P}(W_n = y)\mathbf{P}(Y_n = y, T > n),$$

from which we deduce that

$$|\mathbf{P}(W_n = y) - \mathbf{P}(W_n = y)| \leq \mathbf{P}(Y_n = y, T > n) + \mathbf{P}(W_n = y, T > n).$$

Summing over all the possible values of y, we get

$$\sum_y |\mathbf{P}(Y_n = y) - \mathbf{P}(W_n = y)| \leq 2\mathbf{P}(T > n).$$

Since T is almost surely finite, the right-side tends to 0 when n to infinity. If we take $W_0 = x$ and Y having the distribution ν, we deduce

$$\sum_y |p^{(n)}(x, y) - \nu(y)| \xrightarrow{n\to\infty} 0.$$

Hence the result. □

NOTE.– We observe that the aperiodicity hypothesis is only used to prove the irreducibility of the Markov chain Z. To be convinced that this is essential, consider again the example of the rat in its maze. Let us form the Markov chain $Z_n = (X_n, Y_n)$ which represents the positions of two rats released in the same maze, which evolve independently of each other according to the same rules as before. Let C_1 be the cyclic class of 1 and C_2 that of 2 for the Markov chain X. If Z starts from a state of $C_1 \times C_2$

then Z evolves between the states of this set and the states of $C_2 \times C_1$, but never reaches the states of $C_1 \times C_1$, therefore Z is not irreducible.

In the periodic case, however, we have the following result:

THEOREM 3.26.– *Let X be an irreducible, positive recurrent Markov chain periodic of period d and of invariant probability ν. Let $x \in E$ and $C_0, \ldots, C_{d-1}$ be the cyclic classes associated with x. If $y \in C_r$*

$$\lim_{n \to \infty} p^{(nd+r)}(x, y) = d\nu(y).$$

The idea is to apply the previous theorem to the Markov chain of transition matrix P^d. It is necessary to determine the invariant probability of this Markov chain. Observe that according to Theorem 3.13, up to a multiplicative constant c, the invariant probability of a state y is equal to c times the proportion of the number of visits to this state between two visits to a fixed state x. As in the Markov chain of matrix P^d we divide the number of steps by d, this proportion is multiplied by d.

Proof. According to the definitions of the period and of C_k, C_k is a closed subset for the chain X^k defined by $X^k_n = X_{nd+k}$ for $k = 0, \ldots, d-1$. These chains are irreducible and positive recurrent. According to Corollary 3.14, the invariant probability ν^k of X^k is proportional to ν, that is, there exist α_k such that $\nu^k(y) = \alpha_k \nu(y)$ for any $y \in C_k$. In addition, since ν is the invariant probability of X, for any k and any l belonging to $0, \ldots, d-1$,

$$\alpha_k = \mathbf{P}_\nu(X_{nd+k} \in C_k) = \mathbf{P}_\nu(X_{nd+k} \in C_k \cup C_l) = \mathbf{P}_\nu(X_{nd+l} \in C_l) = \alpha_l.$$

It follows that $\alpha_k = d^{-1}$. The final point results from Theorem 3.25. □

The final result useful for simulations is the central theorem limit which states that:

THEOREM 3.27.– *Let X be a positive recurrent Markov chain of invariant probability ν. For $f\colon E \times E \to \mathbf{R}$,*

$$Pf\colon \begin{cases} E & \longrightarrow \mathbf{R} \\ x & \longmapsto Pf(x) = \sum_y f(x,y)p(x,y) = \mathbf{E}_x\left[f(X_0, x_1)\right] \end{cases}.$$

For any function f such that$\mathbf{E}_\nu[P(f^2)] < \infty$*,*

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\left(f(X_{k-1},X_k)-Pf(X_{k-1})\right)\xrightarrow[n\to\infty]{Law}\mathcal{N}(0,\sigma^2(f)),$$

where $\mathcal{N}(0,\sigma^2(f))$ *is a Gaussian distribution of variance given by*

$$\sigma^2(f)=\mathbf{E}_\nu[P(f^2)]-\mathbf{E}_\nu[(Pf)^2].$$

Proof. Lemma 3.12 implies that for f bounded, the process

$$M_n^f=\sum_{j=0}^{n}f(X_n,X_{n+1})-\sum_{j=0}^{n}Pf(X_j,.)(X_j)$$

is a martingale. Furthermore, its increasing process is

$$\begin{aligned}\Delta\langle M^f\rangle_n&=\mathbf{E}\left[(\Delta M_n^f)^2\,|\,\mathcal{F}_n\right]\\&=\mathbf{E}\left[\Big(f(X_n,X_{n+1})-Pf(X_n)\Big)^2\,|\,\mathcal{F}_n\right]\\&=Pf^2(X_n)+Pf(X_n)^2-2Pf(X_n)^2\\&=\Gamma f(X_n),\end{aligned}$$

where $\Gamma f = P(f^2)-(Pf)^2$ is *the carré du champ* operator associated with P. Therefore

$$\langle M^f\rangle_n=\sum_{j=0}^{n}\Gamma f(X_j).$$

By hypothesis, Γf is integrable with respect to ν, Theorem 3.21 implies that

$$\frac{\langle M^f\rangle_n}{n}\xrightarrow{n\to\infty}\sigma^2(f),\mathbf{P}_x\text{a.e.}$$

The result follows from the central limit theorem for martingales increments. □

If we take as a particular case $f(X_{k-1},X_k)=\mathbf{1}_{\{X_k=x\}}$, we get

$$\mathbf{P}\left(\sqrt{n}(N_x^n-\pi(x))\in[a,b]\right)\xrightarrow{n\to\infty}\int_a^b\exp(-x^2/2\sigma^2)\frac{\mathrm{d}\,x}{\sigma\sqrt{2\pi}},$$

with $\sigma^2=\nu(y)-\sum_x p(x,y)^2\nu(x)$.

EXAMPLE (Example 3.1 (continued)).– This is the simplest case in which we have to solve the system $\nu = \nu P$ and $\sum \pi(x) = 1$. After making all the calculations,

$$\nu = \left(\frac{1}{8}, \frac{3}{16}, \frac{1}{8}, \frac{3}{16}, \frac{3}{16}, \frac{1}{8}, \frac{1}{16}\right).$$

EXAMPLE (Example 3.3 (continued)).– It is necessary to restrict the Markov chain to any equivalence class of the "communicate" relation. In this case, it is clear that the invariant probability is the uniform measure on these states.

3.5. Effective calculation of the invariant probability

The principle is simple: the invariant probability is the only vector with non-negative components, of total weight 1 which satisfies the equation $\nu(P - \mathrm{Id}) = 0$. To solve such a system by computer, it is must be taken into account that this system has co-rank 1, that is, it is necessary to remove a column of P (e.g. the last one) and replace it with a column of 1. Let $\hat{P}$ be the matrix thus obtained. We must then solve the system

$$\pi(\hat{P} - \hat{I}) = b, \text{ with } b = (0, \ldots, 0, 1) \text{ and } \hat{I} = \begin{pmatrix} 1 & & & 0 \\ & 1 & (0) & 0 \\ & (0) & \ddots & \\ & & & 0 \end{pmatrix}.$$

In practice, the chains that are used have a finite state space but very large cardinal (several thousands of states). This requires the use of numerical analysis methods.

3.5.1. *Iterative method*

We have to solve the equation $\pi = \pi P$ where P is the transition matrix. According to Theorem 3.25, if the chain is aperiodic then $\pi_{n+1} = \pi_n P$ tends to the invariant probability. In practice, we take any π_0 and we iterate the relation $\pi_{n+1} = \pi_n P$. This process can be expensive if the calculation of the coefficients of P is long. However, convergence is exponentially fast with scale factor given by the modulus of the second largest eigenvalue of P.

When the chain is periodic (see the example of the rat) of period d, we must be more cautious. Theorem 3.25 shows the sequence $(\pi_n,\, n \in \mathbf{N})$ has d cluster points. Specifically, by definition of a cyclic class, if π_0 is a Dirac mass at x, the terms π_{kn} have positive components only for the states of the cyclic class x, the terms π_{kn+j} have positive components only for the states of the cyclic class C_j, for all $j \in \{1, \ldots, d-1\}$.

EXAMPLE (Example 3.1 (continued)).– Consider the initial condition $\pi_0 = (1, 0, \ldots)$, we then have

$$v(2) = \left(0, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0\right)$$
$$v(3) = \left(\frac{1}{3}, 0, \frac{1}{6}, 0, \frac{1}{3}, 0, \frac{1}{6}\right)$$
$$v(4) = \left(0, \frac{13}{36}, 0, \frac{4}{9}, 0, \frac{7}{36}, 0\right)$$
$$v(5) = \left(\frac{29}{108}, 0, \frac{47}{216}, 0, \frac{79}{216}, 0, \frac{4}{27}\right)$$
$$v(6) = \left(0, \frac{473}{1296}, 0, \frac{131}{324}, 0, \frac{299}{1296}, 0\right)$$
$$v(7) = \left(\frac{997}{3888}, 0, \frac{1843}{7776}, 0, \frac{2891}{7776}, 0, \frac{131}{972}\right).$$

One way to avoid this is to consider the sum of Cesaro $\hat{\pi}_n = d^{-1}\sum_{i=n-d}^{n} \pi_i$. This requires knowledge of the period, if it is not possible then we can rely on the ergodic method.

3.5.2. *Ergodic method*

Theorem 3.21 states that for an irreducible and positive recurrent Markov chain of invariant probability ν, we have for any initial condition and all $x \in E$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k=x\}} = \nu(x).$$

We simulate a trajectory of the Markov chain as long as possible and compute the percentage of time that we move in the state x. Theorem 3.27 indicates that the speed of convergence is $1/\sqrt{n}$ which compares very unfavorably with the previous two methods. However, we do not store all the $\nu(x)$ but only the values of interest. It is actually very common that only a few components of ν are interesting.

EXAMPLE (File M/GI/1/K).– In this queuing system, there is a buffer of size K, and X_n denotes the number of customers in the system just after the departure of the customer n. Then $(X_n,\ n \in \mathbf{N})$ satisfies the recurrence $X_{n+1} = \min((X_n - 1)^+ + A_{n+1}, K + 1)$. Hence, we have an irreducible Markov chain which is necessarily recurrent since the state space is finite. We cannot calculate the invariant probabilities by generating function because of the side effects. However, for dimensioning the buffer, we are only interested in the probability of loss, that is to say $\nu(K + 1)$. It is obtained by the following formula

$$\nu(K+1) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k = K+1\}}$$

and

$$\hat{N} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k$$

is the mean number of customers in the queue at the equilibrium state.

3.6. Problems

EXERCISE 1.– On a chess board, we place a knight in the corner $A1$. The knight moves randomly (it chooses a move at random among those possible) and without memory. We recall that a knight moves two squares in one direction (either horizontal or vertical) and a square in the other direction. Using the reversibility and symmetry considerations, calculate the average time back to square $A1$.

Same question, if we identify the opposite edges of the board, then the knight moves on a torus!

EXERCISE 2.– Build (whenever it is possible) a Markov chain with two states such that:

- the two states are recurrent;
- the two states are transient;
- one state is transient, and the other recurrent;
- both are transient;
- both are zero recurrent.

EXERCISE 3.– Consider the Markov chain with values in {1,2,3} whose transition matrix is given by

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ f(p) & 0 & 1-f(p) \\ 1-f(p) & 0 & f(p) \end{pmatrix}$$

where $p \in [0,1]$ and $f(p)$ is defined by:

$$f(p) = \begin{cases} 0 & \text{if } p \leq 1/4 \\ 2p - 1/2 & \text{if } 1/4 < p \leq 3/4 \\ 1 & \text{if } p \geq 3/4 \end{cases}$$

1) Give the classification of states according to the values of p;

2) For what values of p is there a probability invariant? Compute it when it exists;

3) Starting from 2, what is the mean return time to 2?

4) Let h be the function defined by

$$h(1) = -1,\ h(2) = 1,\ h(3) = 1.$$

What is the limit

$$\frac{1}{n}\sum_{j=1}^{n} h(X_j)$$

when n tends to $+\infty$ for $p < 3/4$?

5) If one has an arbitrarily large number of sample paths, how do we know if p is greater than $3/4$? How do we know if $p < 1/4$? How can we estimate p if it is between $1/4$ and $3/4$?

EXERCISE 4.– Let X be an irreducible recurrent Markov chain on E, and F a subset of E. Show that the chain $\left(X_n^F,\ n \in \mathbf{N}\right)$ (see definition 3.8) is a Markov chain on E.

EXERCISE 5.– Consider the homogeneous Markov chain X with two states A and B and transition matrix

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

We seek the time of first occurrence of the sequence ABA. To do this, we construct the process $Y_n = (X_n, X_{n+1}, X_{n+2})$:

1) Show that Y is an homogeneous Markov chain and give its transition matrix (as a matrix or graph);

2) Is this chain irreducible? aperiodic? positive recurrent?

3) Calculate the invariant probability of Y. We can number the states in lexicographic order: $AAA = 1, AAB = 2, \ldots$;

4) Derive the average time between two occurrences of ABA;

5) We assume that $X_0 = A, X_1 = B$. Give the equations to calculate $\mathbf{E}\left[\tau_{ABA}^1\right]$.

EXERCISE 6.– A device emits on a transmission line some packets of constant size. T denotes the transmission duration of a packet. We consider a discrete-time model of the system, that is, a model in which the time is divided into intervals of constant length, which we assume equal to T. Each interval is called a slot. The transmission line can introduce errors and we define a sequence (Y_n) such that $Y_n = 1$ if, at time $n+1$, the line is in a state in which it introduces errors and $Y_n = 0$ if at time $n+1$, it is in a state where it does not introduce errors. Assume that (Y_n) is a Markov chain and invariant $P(Y_1 = 1 \,|\, Y_0 = 1) = 0.9$ and $P(Y_1 = 0 \,|\, Y_0 = 0) = 0.1$.

The emission is made with the protocol “stop and wait”. According to this protocol, each packet must be acquitted. If there is no error, the packet is positively acknowledged and the next packet can be transmitted. Otherwise, the packet must be retransmitted. To simplify the problem, we consider that the acknowledgment arrives instantly.

1) Calculate the invariant probability distribution of Y_n;

2) Assume that the packets arrive according to a geometric process. That is, at the nth slot there is an arrival with probability q and no arrivals with probability $1 - q$. A packet can be transmitted in its arrival slot. Let X_n be the number of packets in the system at slot n. The pair (X_n, y_n) is a Markov chain. We order the states in lexicographic order, that is to say:

0	1	2	3	4	5	6	7	8	9	...
00	01	10	11	20	21	30	31	40	41	...

3) Find Q the transition matrix of (X_n, y_n);

4) Show that

$$\nu_0 = 1$$

$$\nu_{2n} = 9\left(3\sqrt{q/1-q}\right)^{2n}$$

$$\nu_{2n+1} = 9\nu_{2n}$$

is an invariant measure for the Markov chain (X_n, y_n).

5) Find all values of q for which all states are positive recurrent. Compare the result with the result of 1. Conclusion.

EXERCISE 7.– Consider a packet of N cards. To mix, we proceed as follows: one chooses a card at random and we put that card on top of the deck.

1) How to represent the state of the deck, denoted by X_n, after the nth operation?

2) By introducing the special permutations

$$\tau_k = \begin{pmatrix} 1 & 2 & \dots & k-1 & k & k+1 & \dots & N \\ 2 & 3 & \dots & k & 1 & k+1 & \dots & N \end{pmatrix}$$

for $k \in \{1, \dots, N\}$; write the transition probabilities of X.

3) Show that this chain is irreducible (first analyze small values of N as $N = 4$, for example).

4) Show that after a sufficiently large number of operations we obtain a “good” mixture, characterized by the equal probability of all possible states of the deck.

EXERCISE 8.– Set $E = \{1, \dots, 10\}$. We define on E, the addition modulo 10, that is to say $10 + 1 = 1$. We consider X, the Markov chain of transition matrix $P = (p_{i,j})$ given by

$$p_{i,i+1} = p, \; p_{i,i-1} = 1 - p.$$

We assume that p is not equal to 0 or 1:

1) Is this chain irreducible ? recurrent? aperiodic?

2) What is its invariant probability?

We now consider, X_1 and X_2, two independent copies of this chain. We put $Y = (X_1, X_2)$.

3) Is this chain irreducible?

4) What is its invariant probability?

5) We set $Z_n = Y_{2n}$. Is this chain irreducible? What are its closed subsets? Is it recurrent? aperiodic?

EXERCISE 9.– Let $A = \{A_n: n \geq 1\}$ be a sequence of random variables independent and identically distributed with values in $\mathbf{R}^k$, let h be an application of $E \times \mathbf{R}^k$ in E, let X_0 be a random variable independent of the sequence A. We define the sequence $X = \{X_n: n \in \mathbf{N}\}$ by X_0 for $n = 0$ and by $X_n = h(X_{n-1}, A_n)$, for $n \geq 1$. Show that X is a Markov chain.

3.7. Notes and comments

The number of books on Markov chains is incalculable, we cannot list them all. Among the most recent, close or complementary to our approach, we can refer to [BAL 01, GRA 08]. Markov chains are still a very active field of investigation because of their universality. Current problems focus on the calculation of the speed of convergence to the stationary probability and its relationship to the "spectral gap", on the reduction of the state space to calculate easily approximates of the invariant probability, on the applications to simulation and to statistical methods through the MCMC.

Epitome

– A Markov chain is defined by its initial distribution ν and its transition operator P.

– A recurrent state is a state visited infinitely often. A state is a transient when it is visited a finite number of times.

– A chain is irreducible if all states communicate.

– A stationary measure identified with a row vector is solution of the equation $\pi P = \pi$.

– If we can find π such that $\sum_{x \in E} \pi(x) = 1$ then π is an invariant probability, and the chain is recurrent.

– In this case, whatever the initial condition, $\mathbf{P}(X_n = x) \xrightarrow{n\to\infty} \pi(x)$.

– To calculate π, we can either solve the system $\pi P = \pi, \sum_{x \in E} \pi(x) = 1$ or consider the limit of the sequence $\pi_{n+1} = \pi_n P$, for any π_0.

Chapter 4

Stationary Queues

In Chapters 8 and 9, it is assumed that the generic distributions of the sequences $(\xi_n, n \in \mathbf{N})$ and $(\sigma_n, n \in \mathbf{N})$ of service and inter-arrival times are exponential. This allows us to represent many of the models by Markov processes in continuous time. Many quantitative results can then be obtained regarding the performances of these systems.

Unfortunately, this hypothesis is unrealistic in many cases, and we are led to consider sequences of random variables which are independent and identically distributed with general distributions (GI/GI/...queues). The architectures of the systems under consideration often lead to further weaken these hypotheses. In fact, a queue often models the traffic in a node that is integrated within a network, and it is desirable that the probabilistic characteristics of a queue are the same as that of the following one, in other words that the input traffic of a queue be the same type as the output traffic. However, aside from the particular case where the input is Poissonian (then, the output is also Poissonian - see Theorem 8.8), it is easy to see that the inter-arrivals time in the second queue (which are the intervals between the departure times from the first queue) are not independent in general, even if the inter-arrivals in the first queue are independent, since their order, for example, depends on the order of service in the first queue.

It is therefore of crucial interest to consider queuing models where stationarity, but not necessarily independence, is assumed. In this context, we can easily understand that accurate quantitative results may be more difficult to obtain. However, in the framework of Chapter 2 we can, in many cases, handle the study of essential questions: existence and uniqueness of an equilibrium and qualitative study of the stationary state (comparison of models, dependence on the distribution of the random variables involved, etc.).

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

We first address the classical G/G/1 queue, and then the multiple server queue. The results are primarily based on Loynes's Theorem 2.4, and hence on the monotonicity of the SRS involved. Then we consider several queuing models admitting a representation in more complex state spaces (such as processor sharing queues or infinite servers queues), or whose representation by an SRS is not monotonic, such as loss queues and queues with impatient customers.

4.1. Single server queues

4.1.1. *Stability*

We consider a queue with a single server working without vacations, processing jobs according to a conservative service discipline. These requests enter the system according to a stationary point process. Specifically, we take two random sequences $(\xi_n, n \in \mathbf{Z})$ and $(\sigma_n, n \in \mathbf{Z})$, taking values respectively in $\mathbf{R}_*^+$ and $\mathbf{R}^+$, which represent the intervals of time between the arrivals of the customers, and their service time, respectively, counted in units of time.

ASSUMPTION– $((\xi_n, \sigma_n), n \in \mathbf{Z})$ is stationary and ergodic. Moreover, $\mathbf{E}[\xi_0] + \mathbf{E}[\sigma_0] < \infty$.

According to Kendall's notation, we thus consider a G/G/1 queue. As in Chapter 2, we can assume that the canonical probability space is $\Omega = F^{\mathbf{Z}}$, where $F = \mathbf{R}_*^+ \times \mathbf{R}^+$ is equipped with the product sigma-field. The probability measure $\mathbf{P}$ is the image measure of the sequence of couples, and the shift towards right θ operates on the two components simultaneously. The quadruple $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$ hence defined is called the *Palm space of arrivals and services*.

The random variables σ and ξ are defined on Ω by

$$\sigma\left(((\xi_n, \sigma_n), n \in \mathbf{Z})\right) = \sigma_0 \text{ and } \xi\left(((\xi_n, \sigma_n), n \in \mathbf{Z})\right) = \xi_0,$$

so that for all $n \in \mathbf{Z}$,

$$\xi_n = \xi \circ \theta^n \text{ and } \sigma_n = \sigma \circ \theta^n.$$

For any $n \in \mathbf{N}, \xi \circ \theta^n$ is interpreted as the time between the arrivals of the nth and the $n + 1$th customers (respectively denoted C_n and C_{n+1}) and $\sigma \circ \theta^n$ as the service time requested by C_n. We fix the time origin $T_0 = 0$ at the arrival time of C_0, and we write for any $n \geqslant 1, T_n = \sum_{i=0}^{n-1} \xi \circ \theta^i$, which is interpreted as in the rest of the book, as the arrival time of C_n.

For any $n \in \mathbf{N}$, let W_n denote the workload to be completed by the server, measured in time units just before the arrival of customer C_n (i.e. at T_n-), starting from an arbitrary workload W_0 at T_0-. The quantity W_n corresponds to the sum of the service times of the customers possibly in line, plus the service time remaining to be processed for the customer possibly in service at this time. In the special case where the service discipline is FCFS (First Come, First Served), W_n then represents the amount of work having priority over that brought by C_n, or in other words, the waiting time of W_n before reaching the server.

LEMMA 4.1.– *For a G/G/1 queue, the workload sequence* $(W_n, n \in \mathbf{N})$ *satisfies the following recurrence relation, commonly referred to as Lindley's equation*

$$W_{n+1} = [W_n + \sigma \circ \theta^n - \xi \circ \theta^n]^+. \qquad [4.1]$$

Proof. For any $n \geqslant 0$, the amount of work W_{n+1} equals the sum of the remaining amount of work W_n just before the arrival of C_n, plus the service time $\sigma \circ \theta^n$ brought by C_n, minus the amount of work J_n processed by the server between the arrivals of C_n and C_{n+1}. Therefore,

$$W_{n+1} = W_n + \sigma \circ \theta^n - J_n.$$

Let us observe that just after the arrival of C_n, the server has a workload of $W_n + \sigma \circ \theta^n$. Thus, there are two cases:

– if $W_n + \sigma \circ \theta^n \geqslant \xi \circ \theta^n$, the server is busy without interruption between the arrival of C_n and that of C_{n+1}, and therefore completes an amount of work $\xi \circ \theta^n$ between these two moments. In this case, $J_n = \xi \circ \theta^n$;

– if $W_n + \sigma \circ \theta^n < \xi \circ \theta^n$, the server becomes available before the arrival of C_{n+1}, who finds an empty system, and hence is served upon arrival. In particular, $J_n = W_n + \sigma \circ \theta^n$.

In both cases, we obtain [4.1] □

We know from Example 2.4 that there exists a unique random variable W, which could be the limit in distribution of $(W_n,\ n \in \mathbf{N})$, and that W satisfies the equation

$$W \circ \theta = [W + \sigma - \xi]^+ = \varphi(W), \text{a.s.}. \qquad [4.2]$$

Let us examine the conditions under which W takes values in $\mathbf{R}+$. Under these conditions, a finite workload exists, describing the steady state of the system.

THEOREM 4.2.– *If*

$$\mathbf{E}[\sigma] < \mathbf{E}[\xi], \qquad [4.3]$$

the random variable W defined by [2.8] is the only a.s. finite solution of [4.2]. If

$$\mathbf{E}[\sigma] > \mathbf{E}[\xi],$$

there is no a.s. finite solution to [4.2].

Proof. W can be explicitly constructed. Let us recall (see the proof of Theorem 2.4) that $M_0 = 0$ a.s., and then

$$\begin{aligned} M_1(\omega) &= \varphi(M_0,\, \theta^{-1}\omega) \\ &= [\sigma\left(\theta^{-1}\omega\right) - \xi\left(\theta^{-1}\omega\right)]^+, \\ M_2(\omega) &= \varphi(M_1\left(\theta^{-1}\omega\right),\, \theta^{-1}\omega) \\ &= \left[[\sigma \circ \theta^{-1}\left(\theta^{-1}\omega\right) - \xi \circ \theta^{-1}\left(\theta^{-1}\omega\right)]^+ + \sigma\left(\theta^{-1}\omega\right) - \xi\left(\theta^{-1}\omega\right)\right]^+ \\ &= \left[\max_{1\leq k\leq 2}\sum_{i=1}^{k}(\sigma\left(\theta^{-i}\omega\right) - \xi\left(\theta^{-i}\omega\right))\right]^+, \end{aligned}$$

and a simple induction shows that for any $n \in \mathbf{N}^*$,

$$M_n = \left[\max_{1\leq k\leq n}\sum_{i=1}^{k}(\sigma \circ \theta^{-i} - \xi \circ \theta^{-i})\right]^+.$$

Thus, the minimal solution of [4.2] is a.s. given by

$$W = M_\infty = \left[\sup_{k\geq 1}\sum_{i=1}^{k}(\sigma \circ \theta^{-i} - \xi \circ \theta^{-i})\right]^+. \qquad [4.4]$$

The random variable $\sigma - \xi$ being integrable, from the Ergodic Theorem 2.7 we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sigma \circ \theta^{-i} - \xi \circ \theta^{-i}\right) \xrightarrow{n\to\infty} \mathbf{E}\left[\sigma - \xi\right] \ a.s.. \qquad [4.5]$$

Denote for any n,

$$S_n = \sum_{i=1}^{n}(\sigma \circ \theta^{-i} - \xi \circ \theta^{-i}).$$

Let us assume that $\mathbf{E}[\sigma - \xi] > 0$. In this case, according to [4.5], the sequence S_n tends to $+\infty$ almost surely. But in view of [4.4], $W = \limsup_n S_n^+$, therefore $\mathbf{P}(W = +\infty) = 1$. Hence there is no finite solution to [4.2] in view of the minimality of W.

Let us now assume that $\mathbf{E}[\sigma - \xi] < 0$. According to [4.5], the sequence $(S_n, n \in \mathbf{N})$ tends to $-\infty$, therefore $\mathbf{P}(\limsup_n S_n^+ < +\infty) = 1$, that is $\mathbf{P}(W < +\infty) = 1$. It remains to check that W is the only a.s. finite solution to [4.2]. At first, let us observe that

$$\mathbf{P}\left(Y = 0\right) > 0 \text{ for any finite solution } Y \text{ of [4.2].} \qquad [4.6]$$

Indeed, if $Y > 0$ a.s. then $Y \circ \theta = Y + \sigma - \xi$ a.s., which implies that

$$\mathbf{E}[Y \circ \theta - Y] = \mathbf{E}[\sigma - \xi] < 0,$$

contradicting Lemma 2.2. On another hand,

$$\{Y = 0\} \subset \{Y \leq W\},$$

which implies according to [4.6] that

$$\mathbf{P}\left(Y \leq W\right) > 0. \tag{4.7}$$

But on $\{Y \leqslant W\}$, by monotonicity we have that

$$Y \circ \theta = \varphi\left(Y\right) \leq \varphi\left(W\right) = W \circ \theta.$$

The event $\{Y \leqslant W\}$ is hence $\theta-$ contracting, it is therefore almost sure according to [4.7], that is $Y \leqslant W$ a.s.. By the minimality of W, it follows that $Y = W$ a.s.. □

The case where $\mathbf{E}[\sigma] = \mathbf{E}[\xi]$ is a limit case where a finite stationary workload can exist or not, depending on the distributions of σ and ξ. To illustrate this fact, let us consider two simple examples:

EXAMPLE 4.1.– Let us assume that $\sigma_n = \xi_n = 1$ for any n, which amounts to $\sigma = \xi = 1$ a.s.. Then, it is straightforward that for any $x \geqslant 0$, the random variable $W \equiv x$ is a solution of [4.2].

EXAMPLE 4.2.– Let us assume that $(\sigma_n, n \in \mathbf{N})$ and $(\xi_n, n \in \mathbf{N})$ are two independent sequences of random variables independent and identically distributed, with the same mean expectation and respective variances σ_1^2 and σ_2^2. Let $0 < \epsilon < \frac{1}{2}$ and $a \in \mathbf{R}_+^*$ be such that $1 - F(a) = \epsilon$, where F denotes the distribution function of $\mathcal{N}(0, 1)$. Then, for any $x > 0$, there is a sufficiently large index n such that $\sqrt{n}\sqrt{\sigma_1^2 + \sigma_2^2}a \geqslant x$, and therefore

$$\mathbf{P}\left(W > x\right) \geq \mathbf{P}\left(S_n > x\right) \geq \mathbf{P}\left(\frac{S_n}{\sqrt{n}\sqrt{\sigma_1^2 + \sigma_2^2}} > a\right) \xrightarrow{n \to \infty} \epsilon,$$

according to the Central Limit Theorem. This shows that $\mathbf{P}(W = \infty) \geqslant \epsilon$, and therefore $W = +\infty$ a.s., since the event $\{W = \infty\}$ is θ-contracting.

For the remainder of this section, let us assume that the stability condition [4.3] holds.

THEOREM 4.3.– *For any random variable Y, $\mathbf{P}$-a.s. finite and positive, the workload sequence $(W_n^Y, n \in \mathbf{N})$ with initial value Y couples with $(W \circ \theta^n, n \in \mathbf{N})$. Particularly, $W_n^Y \xrightarrow[n \to \infty]{\mathcal{L}} W$. On the other hand, there is strong backward coupling if $Y \leqslant W$, a.s.*

Proof. Let ψ be the random mapping defined by

$$\psi : \begin{cases} \mathbf{R}^+ \times \Omega & \longrightarrow \mathbf{R} \\ (x,\, \omega) & \longmapsto x + (\sigma(\omega) - \xi(\omega)) \end{cases}$$

and $(Z_n, n \in \mathbf{N})$, the SRS driven by ψ, with initial value Y. Theorem 2.7 shows that on an event $\mathcal{A}$ of probability 1,

$$\frac{1}{n} Z_n = \frac{Y}{n} + \frac{1}{n} \sum_{i=1}^{n} \left(\sigma \circ \theta^i - \xi \circ \theta^i \right) \xrightarrow{n \to \infty} \mathbf{E}\left[\sigma - \xi \right] < 0.$$

On $\mathcal{A}$, Z_n tends to $-\infty$, thus

$$N^Y = \inf\{n \geq 0,\, Z_n < 0\}$$

is a.s. finite. Let us observe that $W_n^Y = Z_n > 0$ for any $n \leqslant N^Y$, and that $W_N^Y = 0$. In addition, as $Y \geqslant 0$, an immediate induction shows that $W_n^Y \geqslant W_n^0$ a.s. for all n. This implies that $W_{N^Y}^0 = 0 = W_{N^Y}^Y$, from where it follows by the definition of the SRS that $W_n^0 = W_n^Y$ for any $n \geqslant N^Y$. For another initial condition, say $Y = W$, there exists $N^W < \infty$ such that $W_n^W = W_n^0$ for any $n \geqslant N^W$. For $n \geqslant N^Y \vee N^W$, $W_n^Y = W_n^0 = W_n^W$, which shows the coupling property between $(W_n^Y, n \in \mathbf{N})$ and $(W \circ \theta^n, n \in \mathbf{N})$.

Now, let Y be a random variable a.s. upper-bounded by W. By monotonicity, we clearly have $W_n^Y \leqslant W_n^W = W \circ \theta^n$ a.s. for any $n \in \mathbf{N}$. On $\theta^{-n}\{W = 0\}$, we therefore have $W_n^Y = 0$, which shows that $(\theta^{-n}\{W = 0\}, n \in \mathbf{N})$ is a sequence of renovating events of length 1 for $(W_n^Y, n \in \mathbf{N})$. Theorem 2.11 thus guarantees that strong backwards coupling holds. □

We now turn to the property of the queue to get empty infinitely often almost surely. It is clear, that if the queue is empty between the arrival of C_{n-1} and that of C_n, we have $W_n = 0$. If $(W_n,\, n \in \mathbf{N})$ was a Markov chain, the property to get empty a.s. infinitely often would correspond to the recurrence of 0 for the chain. We show that this property holds under condition [4.3].

COROLLARY 4.4.– *Under the stability condition [4.3], the G/G/1 queue starting from the initial finite workload Y upon the arrival of C_0, empties $\mathbf{P}$-a.s. infinitely often.*

Proof. From Theorem 2.7 and [4.6], we have $\mathbf{P}$-a.s.

$$\begin{aligned} 0 < \mathbf{P}\left(W = 0\right) &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_0 \left(W \circ \theta^i \right) \\ &= \lim_{n \to \infty} \frac{1}{n} \left\{ \sum_{i=1}^{N-1} \mathbf{1}_0 \left(W \circ \theta^i \right) + \sum_{i=N}^{n} \mathbf{1}_0 \left(W_i^Y \right) \right\}. \end{aligned}$$

In other words,

$$\sum_{i=N}^{+\infty} \mathbf{1}_0(W_i^Y) = +\infty, \ \mathbf{P} - \text{a.s.},$$

which exactly means that the queue becomes empty an infinite number of times. □

4.1.2. *Comparisons of G/G/1 queues*

LEMMA 4.5.– *Consider two G/G/1 queues, carried respectively by the random variables (σ, ξ) and $(\bar{\sigma}, \bar{\xi})$, defined on their respective Palm spaces $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$ and $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbf{P}}, \bar{\theta})$. If we have*

$$\bar{\sigma} - \bar{\xi} \leq_{st} \sigma - \xi, \qquad [4.8]$$

then the respective solutions W and $\bar{W}$ of the equations

$$W \circ \theta = [W + \sigma - \xi]^+, \ \mathbf{P} - a.s.$$

$$\bar{W} \circ \bar{\theta} = \left[\bar{W} + \bar{\sigma} - \bar{\xi}\right]^+, \ \bar{\mathbf{P}} - a.s.$$

verify

$$\bar{W} \leq_{st} W. \qquad [4.9]$$

Proof. Apply Theorem 2.15 to $\bar{\alpha} = \bar{\sigma} - \bar{\xi}$, $\alpha = \sigma - \xi$ and $\varphi(x, y) = (x + y)^+$. □

NOTE.– In particular, if we assume that $(\sigma \circ \theta^n, n \in \mathbf{N})$ (respectively $(\bar{\sigma} \circ \bar{\theta}^n, n \in \mathbf{N})$) is independent of $(\xi \circ \theta^n, n \in \mathbf{N})$ (resp. $(\bar{\xi} \circ \bar{\theta^n}, n \in \mathbf{N})$), it is easy to see that [4.8], and therefore [4.9] holds under either one of the two conditions

$$\xi \overset{\mathcal{L}}{=} \bar{\xi} \text{ and } \sigma \leq_{st} \bar{\sigma};$$

$$\sigma \overset{\mathcal{L}}{=} \bar{\sigma} \text{ and } \bar{\xi} \leq_{st} \xi.$$

THEOREM 4.6.– *Consider now a G/G/1 queue driven by the random variables σ and ξ, where it is assumed that the sequences $(\sigma \circ \theta^n, \ n \in \mathbf{N})$ and $(\xi \circ \theta^n, \ n \in \mathbf{N})$ are independent and satisfy the stability condition $\mathbf{E}[\sigma - \xi] < 0$. Let us define on the same probability space, the following two alternative queues:*

– The first one is a stable G/D/1 queue having the same load, where the generic service time is given by $\hat{\sigma} = \mathbf{E}[\sigma]$ and the generic inter-arrival time by ξ. We denote $\hat{W}$, its stationary workload.

– The second one is a D/G/1 queue carried by $\check{\xi} = \mathbf{E}[\xi]$ and σ. We denote by $\check{W}$ its stationary workload.

Then, for any increasing convex function F, we have

$$\mathbf{E}\left[F\left(\hat{W}\right)\right] \leq \mathbf{E}\left[F\left(W\right)\right]; \qquad [4.10]$$

$$\mathbf{E}\left[F\left(\check{W}\right)\right] \leq \mathbf{E}\left[F\left(W\right)\right]. \qquad [4.11]$$

Proof. Theorem 2.17 is verified by setting, for $\hat{W}$,

$$\alpha = \sigma - \xi,\ \bar{\alpha} = \mathbf{E}\left[\sigma\right] - \xi \text{ and } \mathcal{F}_n = \sigma\left(\xi \circ \theta^j;\ j = 0, \ldots, n\right)$$

and for $\check{W}$,

$$\alpha = \sigma - \xi,\ \bar{\alpha} = \sigma - \mathbf{E}\left[\xi\right] \text{ and } \mathcal{F}_n = \sigma\left(\sigma \circ \theta^i;\ i = 0, \ldots, n\right).$$

Indeed, for example in the first such case, by independence, for any $n \in N$ and any $i \leqslant n$, we have a.s.

$$\begin{aligned}\mathbf{E}\left[\alpha \circ \theta^i \mid \mathcal{F}_n\right] &= \mathbf{E}\left[\sigma \circ \theta^i - \xi \circ \theta^i \mid \xi \circ \theta^j;\ j = 0, \ldots, n\right] \\ &= \mathbf{E}\left[\sigma\right] - \xi \circ \theta^i \\ &= \bar{\alpha} \circ \theta^i.\end{aligned}$$

The other case is treated similarly. We conclude with Corollary 2.18. □

In this sense, determinism minimizes the average workload at equilibrium and therefore the average waiting time if the discipline is FIFO. If we assume that in a router, the packet processing time is proportional to its length, this means that the average delay is minimized when taking packets of constant length.

To the limit, let us observe that the deterministic system, of inter-arrival and service times equal to their respective means admits, clearly, the only stationary solution $W = 0$ a.s.

4.1.3. *Representation of service disciplines*

In a queueing system, the service discipline characterizes the policy applied by the server(s) to select a customer when he (they) is (are) available, and there are several customers in line. Note that all the results obtained so far in section 4.1 are independent of the discipline we are dealing with. Thus, to represent the service discipline in the state of the system, we have to enrich the model, since the workload alone is not sufficient to recover this information.

To account for the service discipline, we represent the system in a larger state space. Specifically, we describe the system just before the arrival of the customer $C_n, n \geqslant 0$, by an ordered sequence S_n, representing the residual service times of the customers in the system at this time. In other words, if X_n is the number of customers in the system at T_n^-, for any $i < n$ such that C_i is in the system at T_n^-, we denote $\varphi_n(i) \in \{1, \ldots, X_n\}$, the place of C_i in the queue in the order of priorities, the first one being occupied by the customer in service at T_n^-. For every such customer C_i, the remaining service time at T_n^- is denoted by $R_n(\varphi_n(i))$, that is

$$R_n(\varphi_n(i))$$
$$= \begin{cases} \sigma \circ \theta^i & \text{if } C_i \text{ has not received service before } T_n^-, \\ \sigma \circ \theta^i - \gamma_i & \text{if } C_i \text{ has already received the amount of service } \gamma_i < \sigma \circ \theta^i \text{ at } T_n^-. \end{cases}$$

For any $n \in \mathbf{Z}$, we define $S_n \in \mathcal{S}$ (see Appendix A.3), the sequence representing the residual service times of the customers in the system at that same time, sorted in the reverse order of priorities, and setting to 0 the other components of S_n, that is

$$S_n(i) = R_n\left(X_n + 1 - i\right), \, i \leq X_n \text{ and } S_n(i) = 0, \, i > X_n,$$

or in other words,

$$S_n = \big\{R_n(X_n), \, R_n(X_n - 1), \, \ldots, \, R_n(2), \, R_n(1), \, 0, \, 0, \, \ldots\big\}.$$

The sequence S_n will be termed *service profile* of the queue at T_n^-.

Now, we make precise the dynamics of the sequence of sequences $(S_n, n \in \mathbf{N})$, in function of the service discipline. We start with an arbitrary profile $S_0 \in S$ at the arrival of C_0. Let S_n be the value of the profile just before the arrival of C_n. At T_n, the service time $\sigma \circ \theta^n$ of the incoming customer is inserted in the service profile arbitrarily in the first place, and it shifts the other terms of the sequence of one slot to the right. By denoting S_{n+} as the resulting sequence, we therefore have

$$S_{n+} = \big\{\sigma \circ \theta^n, \, S_n(1), \, S_n(2), \, \ldots\big\} = F^1\left(S_n, \, \sigma \circ \theta^n\right). \qquad [4.12]$$

Then, the service discipline Φ is represented by a mapping $F^\Phi : \mathcal{S} \to \mathcal{S}$ as follows:

1) FCFS: F^{FCFS} is the identity since the incoming customer has the lowest priority;

2) Non-preemptive LCFS

$$F^{\text{LCFS}}(u) = \big\{u(2), \, u(3), \, \ldots, \, u\big(N(u) - 1\big), \, u(1), \, u\big(N(u)\big), \, 0, \, \ldots\big\},$$

since the incoming customer is inserted just after the customer in service in the order of priorities;

3) Preemptive LCFS: in this case,

$$F^{\text{LCFS}}(u) = \big\{u(2),\, u(3),\, \dots,\, u\big(N_u - 1\big),\, u\big(N_u\big),\, u(1),\, 0,\, \dots\big\},$$

since the entering customer shall immediately substitutes the customer in service (if any);

4) SRPT (Shortest Remaining Processing Time): we give a preemptive priority to the customer who has the smallest residual service time. Therefore,

$$\begin{aligned}
&F^{\text{SRPT}}(u) = \\
&\big\{u(2),\, u(3),\, \dots,\, u(i-1),\, u(1),\, u(i),\, \dots,\, u\big(N(u)\big),\, 0, \dots\big\} \\
&\hspace{8cm}\text{if } u(i-1) \geq u(1) \geq u(i); \\
&\big\{u(1),\, u(2),\, u(3),\, \dots,\, u\big(N(u)\big),\, 0, \dots\big\} \text{ if } u(1) \geq u(2); \\
&\big\{u(2),\, u(3),\, \dots,\, u\big(N(u)\big),\, u(1),\, 0,\, \dots\big\} \text{ if } u(1) < u\big(N(u)\big).
\end{aligned}$$

It follows that $F^{\text{SRPT}}(u)$ is ordered in decreasing order whenever u is so.

5) SPT (Shortest Processing Time): it gives non-preemptive priority to the customer who has the smallest residual service time. Therefore F^{SPT} equals F^{SRPT} except that $u(1)$ is inserted just before $u(N(u))$ even if $u(1) < u(N(u))$.

We can thus represent by such a permutation of S, any service discipline depending only on the arrival dates and service requests from the customers since the last arrival in an empty system, or at the most since the moment T_{0-}. Such a discipline is said to be *admissible*.

We denote S_{n++} as the profile of the queue just after the scheduling of the customers, so that

$$S_{n++} = F^{\Phi}\,(S_{n+}). \qquad [4.13]$$

Then, the customer in service at T_n (having a residual service $S_{n+}(X_n)$ at this time) has, just before the arrival of C_{n+1}, a residual service time equal to

$$S_{n+1}(X_n) = [S_{n+1}(X_n) - \xi \circ \theta^n]^+ .$$

Any customer following him (hence having a residual service time given at T_n by $S_{n+}(j)$ for some $j \in [\![0, X_n - 1]\!]$) receives some service before T_{n+1} if and only if

$$\xi \circ \theta^n > \sum_{i=j+1}^{+\infty} S_n(i),$$

in quantity given by

$$\left(\xi \circ \theta^n - \sum_{i=j+1}^{+\infty} S_n(i)\right) \wedge S_{n+}(j).$$

In other words, for any $j \in N$,

$$\begin{aligned} S_{n+1}(j) &= S_{n+}(j) - \left(\xi \circ \theta^n - \sum_{i=j+1}^{+\infty} S_n(i)\right)^+ \wedge S_{n+}(j) \\ &= \left(S_{n+}(j) - \left(\xi \circ \theta^n - \sum_{i=j+1}^{+\infty} S_n(i)\right)^+\right)^+ . \end{aligned}$$

We denote $F^3(., \xi \circ \theta^n)$ the corresponding mapping, so that

$$S_{n+1} = F^3\left(S_{n++}, \, \xi \circ \theta^n\right). \qquad [4.14]$$

Equations [4.12]–[4.14] indicate that for a fixed Φ, the sequence $(S_n, n \in N)$ is an SRS, since for any $n \in N$,

$$S_{n+1} = F^3(., \, \xi \circ \theta^n) \circ F^{\Phi} \circ F^1(., \, \sigma \circ \theta^n)(S_n). \qquad [4.15]$$

A stationary sequence of service profiles hence corresponds uniquely to a random variable S^{Φ} with values in $\mathcal{S}$, solving the equation

$$S^{\Phi} \circ \theta = G^{\Phi}\left(S^{\Phi}\right), \qquad [4.16]$$

where the mapping G^{Φ} is defined by

$$G^{\Phi}\colon \begin{cases} \mathcal{S} & \longrightarrow \mathcal{S} \\ u & \longmapsto F^3(u, \, \xi) \circ F^{\Phi} \circ F^1(u, \, \sigma). \end{cases}$$

THEOREM 4.7.– *Let Φ be an admissible service discipline. Under the stability condition [4.3], there exists a unique solution S^{Φ} to [4.16] such that $S^{\Phi} \in \mathcal{S}$ a.s.. In addition, there is a strong backward coupling between the sequences $\left(S_n^{\Phi,\mu}, \, n \in \mathbf{N}\right)$ and $\left(S^{\Phi} \circ \theta^n, \, n \in \mathbf{N}\right)$ for any $\mu \in \mathcal{S}$ such that*

$$Z(\mu) := \sum_{i \in \mathbf{N}^*} \mu(i) \leq W \text{ a.s.,} \qquad [4.17]$$

where W is the only a.s. finite solution of [4.2].

Proof. Let us observe that for any $n \in \mathbf{N}$, the workload at T_n^- (starting from a given initial condition) is deduced from S_n by

$$W_n = \sum_{i \in \mathbf{N}^*} S_n(i).$$

Let $\mu \in \mathcal{S}$ satisfying [4.17] and $S_n^{\Phi,\mu}$ be the service profile of the queue at T_n^- under the discipline Φ, when starting from the profile μ at T_0^-. We clearly have

$$\sum_{i \in \mathbf{N}^*} S_n^{\Phi,\mu}(i) = W_n^{Z(\mu)} \text{ a.s.,}$$

where $W_n^{Z(\mu)}$ is the nth value of the SRS driven by φ defined by [4.2] and starting from $Z(\mu)$.

Moreover, as φ is a.s. non-decreasing, it is easy to show by induction from [4.17] that a.s.,

$$W_n^{Z(\mu)} \leq W \circ \theta^n, \; n \in \mathbf{N}.$$

Therefore, on the event $\mathcal{A}_n = \{W \circ \theta^n = 0\}$, we have $W_n^{Z(\mu)} = 0$ and therefore $S_n^{\Phi,\mu} = \mathbf{0}$, the null sequence of $\mathcal{S}$. Thus, $(\mathcal{A}_n, \; n \in \mathbf{N}) = (\{W = 0\} \circ \theta^{-n}, \; n \in \mathbf{N})$ is a stationary sequence of renovating events of length 1 for the sequence $\left(S_n^{\Phi,\mu}, \; n \in \mathbf{N}\right)$. As $\mathbf{P}(W = 0) > 0$ according to [4.6], Theorem 2.11 applied to the class of initial conditions

$$\mathcal{Z} = \left\{\mu \in \mathcal{S}; \mu \text{ satisfy [4.17]}\right\} \qquad [4.18]$$

implies the existence of a solution S^Φ to equation [4.16] for the discipline Φ. This solution (which reads as the limit of a sequence that is a.s. constant from a certain rank – see the proof of Theorem 2.11) is a.s. finite, which means that its components are a.s. finite and that it admits an a.s. finite number of components. In other words, $S^\Phi \in \mathcal{S}$ a.s..

Now, let S and S' be two solutions of [4.16] with values in $\mathcal{S}$. By denoting

$$Z = \sum_{i \in \mathbf{N}^*} S(i) \text{ and } Z' = \sum_{i \in \mathbf{N}^*} S'(i)$$

as respective workloads corresponding to these two profiles, it is easy to check with [4.12] and [4.14] that Z and Z' are two a.s. finite solutions of [4.2]. According to Theorem 4.2, we then have $Z = Z' = W$. Therefore,

$$\{W = 0\} \subset \{Z = Z' = 0\} \subset \{S = S'\}.$$

Since the event on the left-hand side is of non-zero probability and the one on the right-hand side is θ-invariant, we have $S = S'$ a.s..

Theorem 2.11 applied to the class $\mathcal{Z}$ particularly implies the property of strong backward coupling for $(S_n^{\Phi,\mu}, n \in \mathbf{N})$ with $(S^\Phi \circ \theta^n, n \in \mathbf{N})$. □

4.1.4. *Other features at equilibrium*

For a given admissible discipline Φ, the sequence S^Φ therefore provides more information on the steady-state of the system than the workload W. Let us show how this information can be used to deduce from this, other characteristics of the system at equilibrium, such as congestion and waiting time.

Let us denote $(X_n^\Phi, n \in \mathbf{N})$ the sequence counting for any $n \in \mathbf{N}$ the number of customers found in the system by C_n, starting from the initial profile S^Φ. Under these conditions, the sequence of the profiles found by the successive customers upon arrival is stationary and equals $(S^\Phi \circ \theta^n, n \in \mathbf{N})$. In particular, the customer C_n finds a service profile $S^\Phi \circ \theta^n$ upon arrival, and therefore

$$X_n^\Phi = N\left(S^\Phi \circ \theta^n\right) = N\left(S^\Phi\right) \circ \theta^n \text{ a.s. for any } n \in \mathbf{N},$$

where $N(.)$ is the number of non-zero coordinates of the sequence (see A.3). This means that $(X_n^\Phi, n \in \mathbf{N})$ is stationary and that a stationary congestion exists, given by $X^\Phi = N\left(S^\Phi\right)$.

We can apply the same arguments to show the existence of a waiting time at equilibrium, using the service profile. Let us denote TA_n^Φ as the waiting time of the customer C_n before entering service under the admissible discipline Φ (let us recall that for $\Phi = \mathrm{FIFO}, \mathrm{TA}_n^\Phi = W_n$).

Once again, start from the service profile S^Φ. The profile of the system at the arrival of customer C_n equals $S^\Phi \circ \theta^n$ and becomes $(S^\Phi \circ \theta^n)_{++}$ after inserting $\sigma \circ \theta^n$ (see [4.13]). If i is the rank of $\sigma \circ \theta^n$ in the sequence $(S^\Phi \circ \theta^n)_{++}$, then the waiting time of C_n equals the sum $\sum_{j>i}(S^\Phi \circ \theta^n)_{++}(j)$ of the service times of the customers having priority upon C_n, plus those of the customers who have arrived after C_n, and left the system before his departure (or its entry into service if the discipline Φ is non-preemptive).

The form of this quantity may be very intricate, depending on the discipline Φ. The crucial point is to get convinced that it depends only on S_n^Φ, on $\sigma \circ \theta^n$ and on the service times $\{\sigma \circ \theta^j;\ j > n\}$ and inter-arrival times $\{\xi \circ \theta^j;\ j > n\}$ of the customers who have arrived after C_n. In other words, there is a deterministic function

$J^\Phi : (\mathbf{R}^+)^{\mathrm{N}} \times (\mathbf{R}^+)^{\mathrm{N}} \rightarrow (\mathbf{R}^+)$ depending only on Φ such that for any $n \in \mathbf{N}$, almost surely

$$\begin{aligned}
\mathrm{TA}_n^\Phi &= \sum_{j>i} \left(S^\Phi \circ \theta^n\right)_{++}(j) \\
&\quad + J^\Phi\left(\left\{\sigma \circ \theta^{n+1}, \sigma \circ \theta^{n+2}, \dots\right\}, \left\{\xi \circ \theta^{n+1}, \xi \circ \theta^{n+2}, \dots\right\}\right) \\
&= \Bigg(\sum_{j>i} \left(F^\Phi \circ F^1(., \sigma)\left(S^\Phi\right)\right)(j) \\
&\qquad + J^\Phi\Big(\left\{\sigma \circ \theta, \sigma \circ \theta^2, \dots\right\}, \left\{\xi \circ \theta, \xi \circ \theta^2, \dots\right\}\Big)\Bigg) \circ \theta^n,
\end{aligned}$$

where we used [4.15]. Therefore, this again shows the existence of a stationary waiting time

$$\begin{aligned}
\mathrm{TA}^\Phi &= \sum_{j>i} \left(F^\Phi \circ F^1(., \sigma)\left(S^\Phi\right)\right)(j) \\
&\quad + J^\Phi\left(\left\{\sigma \circ \theta, \sigma \circ \theta^2, \dots\right\}, \left\{\xi \circ \theta, \xi \circ \theta^2, \dots\right\}\right).
\end{aligned} \qquad [4.19]$$

NOTE.– Under the stability condition [4.3], for any admissible discipline Φ there also exists a *stationary sojourn time* TS^Φ in the system, given by

$$\mathrm{TS}^\Phi = \mathrm{TA}^\Phi + \sigma. \qquad [4.20]$$

EXAMPLE 4.2.1.– Let us write explicitly TA^Φ for Φ = non-preemptive LIFO. Initially, as the service time of the incoming customer C_0 is placed directly on priority just behind the customer already in service, the sum on the left-hand side of [4.19] equals the remaining service time of the customer in service at the arrival of C_0, given by the last non null term of $\tilde{S}^{\mathrm{LIFO}}$, that is $\tilde{S}^{\mathrm{LIFO}}(N(\tilde{S}^{\mathrm{LIFO}}))$.

Then the term on the right-hand side equals the sum of the service times of the customers already entered before C_0 could reach the server. In other words,

$$J^{\mathrm{LIFO}}(\dots) = \sum_{i=1}^{i_0-1} \sigma \circ \theta^i,$$

where, setting $\sum_{j=1}^0 = 0$,

$$i_0 = \inf\left\{ j \in \mathbf{N}^*;\ \tilde{S}^{\mathrm{LIFO}}(N(\tilde{S}^{\mathrm{LIFO}})) + \sum_{j=1}^{i-1} \sigma \circ \theta^j - \sum_{k=1}^{i} \xi \circ \theta^k \le 0 \right\}$$

is the first index of a customer who entered after C_0 and completed his service before the next customer could enter.

4.1.5. *Optimality of SRPT*

With the exhaustive representation introduced in the previous section in hand, we now show how to compare at equilibrium, the different service disciplines. Hereafter, "$\prec_c$" denotes the Schur-convex ordering, presented in Appendix A.3.

THEOREM 4.8.– *Consider a G/G/1 queue satisfying the stability condition [4.3]. Let Φ be an admissible discipline. Then,*

$$S^{\Phi} \prec_c S^{\mathrm{SRPT}} \text{ a.s.,}$$

where S^{Φ} and S^{SRPT} are the only solutions of [4.16] for Φ and SRPT, respectively.

Proof. With the previous notations, we have a.s.

$$S^{\Phi} \xrightarrow{F^1(.,\sigma)} S^{\Phi}_{+} \xrightarrow{F^{\Phi}} S^{\Phi}_{++} \xrightarrow{F^2(.,\xi)} S^{\Phi} \circ \theta.$$

Fix $j \in \mathbf{N}$. If $\sigma < \underline{S}^{\Phi}(j-1)$, then $\sigma \in \left\{\underline{S}^{\Phi}_{+}(i);\, i \geqslant j\right\}$, whereas if $\sigma \geqslant \underline{S}^{\Phi}(j-1)$, we have $\underline{S}^{\Phi}_{+}(i) = \underline{S}^{\Phi}(i-1)$ for any $i \geqslant j-1$. We therefore have in all the cases,

$$\sum_{i=j}^{+\infty} \underline{S}^{\Phi}_{+}(i) = \left(\sigma + \sum_{i=j}^{+\infty} \underline{S}^{\Phi}(i)\right) \wedge \left(\sum_{i=j-1}^{+\infty} \underline{S}^{\Phi}(i)\right). \qquad [4.21]$$

Moreover, customers of service time initially equal to $\underline{S}^{\Phi}_{++}(i), i \geqslant j$ receive during a time interval of length ξ, a total service time at the most equal to

$$\left(\sum_{i=j}^{+\infty} \underline{S}^{\Phi}_{++}(i)\right) \wedge \xi.$$

Therefore,

$$\begin{aligned} \sum_{i=j}^{+\infty} \left(\underline{S}^{\Phi} \circ \theta\right)(i) &\geq \left[\sum_{i=j}^{+\infty} \underline{S}^{\Phi}_{++}(i) - \xi\right]^{+} \\ &= \left[\sum_{i=j}^{+\infty} \underline{S}^{\Phi}_{+}(i) - \xi\right]^{+} \qquad [4.22] \\ &= \left[\min\left\{\sigma + \sum_{i=j}^{+\infty} \underline{S}^{\Phi}(i)\,;\, \sum_{i=j-1}^{+\infty} \underline{S}^{\Phi}(i)\right\} - \xi\right]^{+}, \end{aligned}$$

with [4.21].

Now, observe that we have $S^{\mathrm{SRPT}}_{++} = \underline{S}^{\mathrm{SRPT}}_{+}$ by the very definition of SRPT, which with [4.21] and [4.22] implies that $S^{\mathrm{SRPT}} \circ \theta$ is ordered and that we have

$$\sum_{i=j}^{+\infty} \left(\underline{S}^{\mathrm{SRPT}} \circ \theta\right)(i) = \sum_{i=j}^{+\infty} \left(S^{\mathrm{SRPT}} \circ \theta\right)(i)$$
$$= \left[\sum_{i=j}^{+\infty} \min\left\{\sigma + \sum_{i=j}^{+\infty} \underline{S}^{\mathrm{SRPT}}(i)\,;\ \sum_{i=j-1}^{+\infty} \underline{S}^{\mathrm{SRPT}}(i)\right\} - \xi\right]^{+}. \qquad [4.23]$$

Therefore, for any admissible Φ, on $\{S^{\Phi} \prec S^{\mathrm{SRPT}}\}$ we have for any $j \in \mathbf{N}$,

$$\sum_{i=j}^{+\infty} \underline{S}^{\Phi}(i) \geq \sum_{i=j}^{+\infty} \underline{S}^{\mathrm{SRPT}}(i).$$

This with [4.22] and [4.23] yields

$$\sum_{i=j}^{+\infty} \left(\underline{S}^{\Phi} \circ \theta\right)(i) \geq \sum_{i=j}^{+\infty} \left(\underline{S}^{\mathrm{SRPT}} \circ \theta\right)(i).$$

Since this last equality holds for all $j \in \mathbf{N}$, and since for $j = 1$,

$$\sum_{i=1}^{+\infty} \left(\underline{S}^{\Phi} \circ \theta\right)(i) = \sum_{i=j}^{+\infty} \left(\underline{S}^{\mathrm{SRPT}} \circ \theta\right)(i) = W \circ \theta,$$

we have $\underline{S}^{\Phi} \circ \theta \prec \underline{S}^{\mathrm{SRPT}} \circ \theta$. The event $\{S^{\Phi} \prec S^{SRPT}\}$ is hence θ-contracting. In addition, it is of positive probability, since it includes the event $\{W = 0\}$. The theorem is proved. □

The proof of the following corollary is left to the reader.

COROLLARY 4.9.– *Let Φ be admissible. Then,*

(a) $X^{\mathrm{SRPT}} \leqslant X^{\Phi}$ a.s., where X^{SRPT} and X^{Φ} denote the number of customers at equilibrium in the system under SRPT and Φ, respectively;

(b) $\mathbf{E}\left[\mathrm{Ts}^{\mathrm{SRPT}}\right] \leqslant \mathbf{E}\left[\mathrm{Ts}^{\Phi}\right]$*, where $\mathrm{Ts}^{\mathrm{SRPT}}$ and Ts^{Φ} denote the stationary sojourn time, under SRPT and Φ, respectively (see [4.20]).*

4.1.6. *GI/GI/1 queue: optimality of FIFO*

Let us consider a GI/GI/1 queue where λ, μ, and ρ denote the usual parameters. In addition to the common hypotheses, we assume that the sequences of inter-arrivals and service times are identically distributed and are independent of each other. We assume again that the stability condition [4.3] holds. We denote again for any n, TA_n, the waiting time of C_n before reaching the server, $\text{Ts}_n = \text{TA}_n + \sigma \circ \theta^n$ the sojourn time of C_n and $T'_n = T_n + \text{Ts}_n$, the departure time of C_n. Subsequently, we emphasize the dependence on the service discipline whenever necessary by adding exponents $^{\text{FIFO}}$ and $^{\Psi}$ to the various parameters. In particular, we know that a stationary waiting time TA^{FIFO} (respectively TA^{Ψ}) and a stationary sojourn time Ts^{FIFO} (respectively Ts^{Ψ}) exist under FIFO (respectively, Ψ).

THEOREM 4.10.– *For any convex function g: $\mathbf{R} \to \mathbf{R}$ and any admissible discipline Ψ non-preemptive and independent on the service times,*

$$\mathbf{E}\left[g\left(\text{Ts}^{\text{FIFO}}\right)\right] \leq \mathbf{E}\left[g\left(\text{Ts}^{\Psi}\right)\right]. \qquad [4.24]$$

NOTE.– The FIFO discipline is thus optimal for the sojourn time among all the acceptable disciplines non-preemptive and independent of service times.

Proof. We couple two systems having the same input, the first one processed with FIFO and the other by Ψ. We assume that C_0 finds an empty system upon arrival. As the system is stable, there exists $\mathbf{P}$-a.s. a finite integer τ (common to both systems) such that C_τ enters an empty system. Let us denote for any $k \geqslant 0$, $\psi(k)$ as the index of the kth customer served by Ψ, by considering that C_0 is served in the "0th" position (since it is the only one in the system upon arrival), that is $\psi(0) = 0$. Consider the two following vectors of size τ

$$N = ((\xi_0, \sigma_0), \ldots, (\xi_{\tau-1}, \sigma_{\tau-1}))\,; N^{\psi} = \left(\left(\xi_0, \sigma_{\psi(0)}\right), \ldots, \left(\xi_{\tau-1}, \sigma_{\psi(\tau-1)}\right)\right), \qquad [4.25]$$

which represent, respectively, during the first busy period, the inter-arrival and service times of the customers, and the inter-arrival and service times when re-arranging the service times following ψ. The interchange argument for i.i.d. sequences (see the references at the end of the chapter) is the following intuitive result

$$N \text{ and } N^{\psi} \text{ have the same distribution.} \qquad [4.26]$$

The underlying idea is, that as the service times are identically distributed and are independent of everything else, we do not change the distribution of the various parameters of the system by exchanging the service times of the customers. Everything happens as if the server was deciding the service times of the customers by making an independent draw of service time at each arrival in service. We stress the fact that [4.26] holds true provided that the service discipline is independent of the service times, as one can easily understand.

We will add subsequently, when necessary, an argument (N) (respectively $\left(N^{\psi}\right)$) when the input during the first busy period is given by N (resp., N^{ψ}). For any $n \in [\![0,\, \tau-1]\!]$, the moment where the customer C_n ends his service in FIFO, if the service times follow N^{ψ}, is given by

$$\begin{aligned} T_n^{\prime\text{FIFO}}\left(N^{\psi}\right) &= \sum_{i=0}^{n} \sigma_{\psi(i)} \\ &= \sum_{i=0}^{\psi^{-1}(\psi(n))} \sigma_{\psi(i)} \\ &= T_{\psi(n)}^{\prime\psi}(N), \end{aligned} \qquad [4.27]$$

that is to say, the moment where the customer $C_{\psi(n)}$ ends his service in ψ if the input is N. Therefore,

$$\text{Ts}_n^{\text{FIFO}}\left(N^{\psi}\right) = T_n^{\prime\text{FIFO}}\left(N^{\psi}\right) - T_n = T_{\psi(n)}^{\prime\psi}(N) - T_n. \qquad [4.28]$$

Until the end of the proof, denote in bold letters, the vectors of τ components representing the different quantities for each customer of the first busy period, for instance

$$\mathbf{T}^{\prime\Psi}(N) = \left(T_0^{\prime\Psi}(N),\, \ldots,\, T_{\tau-1}^{\prime\Psi}(N)\right).$$

Notice, that $\mathbf{T}_{\psi}^{\prime\Psi}$ is, by definition, the fully ordered version of $\mathbf{T}^{\prime\Psi}$. Hence, according to the assertion (ii) of Lemma A.14,

$$\mathbf{T}_{\psi}^{\prime\Psi}(N) - \mathbf{T} \prec_c \mathbf{T}^{\prime\Psi}(N) - \mathbf{T} = \mathbf{A}^{\Psi}(N),$$

that is with [4.28],

$$\mathbf{Ts}^{\text{FIFO}}\left(N^{\psi}\right) \prec_c \mathbf{Ts}^{\Psi}(N).$$

Hence, according to (i) of Lemma A.14, for any convex symmetric function $F\colon \mathbf{R}^{\tau} \to \mathbf{R}$,

$$F\left(\mathbf{Ts}^{\text{FIFO}}\left(N^{\psi}\right)\right) \leq F\left(\mathbf{Ts}^{\Psi}(N)\right),$$

and in particular for any convex function $g\colon \mathbf{R} \to \mathbf{R}$,

$$\sum_{n=1}^{\tau-1} g\left(\text{Ts}_n^{\text{FIFO}}\left(N^{\psi}\right)\right) \leq \sum_{n=1}^{\tau-1} g\left(\text{Ts}_n^{\Psi}(N)\right).$$

Finally, as the busy periods are independent and indistinguishable in distribution according to the independence assumptions,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} g\left(\mathrm{Ts}_n^{\mathrm{FIFO}}\left(N^{\psi}\right)\right) \leq \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{\mathbf{N}} g\left(\mathrm{Ts}_n^{\Psi}(N)\right),$$

and [4.24] follows from [4.26] and Birkhoff's Theorem. □

The proofs of the following two corollaries are left to the reader.

COROLLARY 4.11.– *For any convex function $g\colon \mathbf{R}\to\mathbf{R}$, the stationary waiting times under FIFO and Ψ satisfy*

$$\mathbf{E}\left[g\left(\mathrm{TA}^{\mathrm{FIFO}}\right)\right] \leq \mathbf{E}\left[g\left(\mathrm{TA}^{\Psi}\right)\right].$$

COROLLARY 4.12.– *For any convex function g,*

$$\mathbf{E}\left[g\left(\mathrm{Ts}^{\Psi}\right)\right] \leq \mathbf{E}\left[g\left(\mathrm{Ts}^{\mathrm{LIFO}}\right)\right] \textit{ and } \mathbf{E}\left[\mathrm{TA}^{\Psi}\right] \leq \mathbf{E}\left[g\left(\mathrm{TA}^{\mathrm{LIFO}}\right)\right].$$

4.1.7. *Queues with deadlines: optimality of EDF*

We now assume that the customers have deadlines to enter in service. We denote E_n the deadline of customer C_n and $D_n = E_n - T_n$, the initial remaining time before the deadline (termed *lead time*) of C_n. We assume that the sequence $(D_n,\, n\in\mathbf{Z})$ is stationary and we work on the canonical space $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$ of arrivals, services and lead times. We denote then D the projection of $(D_n,\, n\in\mathbf{Z})$ on its first coordinate, interpreted as the lead time of customer C_0.

We assume that $(\sigma_n,\, n\in\mathbf{Z})$ is an i.i.d. sequence, independent of the arrival process (and therefore of $(\xi_n,\, n\in\mathbf{Z})$ and of $(D_n,\, n\in\mathbf{Z})$), and that the random variables ξ, σ, and D are integrable. The deadlines of the customers are smooth, as opposed to the case of hard deadlines (or impatience times) discussed in section 4.6. Indeed, a customer who did not enter service before his deadline does not leave the system, but continues to wait for his turn. The deadlines must then be seen here as indicators of the timing requirement of the customers.

We study hereafter the capacity of the system to minimize the lateness of the customers with respect to this requirement, by comparing the different service disciplines. Let us assume that the stability condition [4.3] holds. We denote again TA_n the waiting time of C_n before reaching the server, and $B_n = T_n + \mathrm{TA}_n$, the moment where C_n enters service. At any time $t \geq T_n$, the residual lead time (i.e. the remaining time before the deadline) of C_n at t is given by $R_n(t) = E_n - t$, and the lead time of C_n at the beginning of his service is hence given by

$$R_n = R_n\,(B_n),$$

whereas the lateness of C_n with respect to its deadline can be written as

$$L_n = (R_n)^- = -R_n \wedge 0.$$

We define two disciplines sensitive to the deadlines:

– the Earliest Deadline First (EDF) discipline always gives priority to the customer with the earliest deadline;

– the Latest Deadline First (LDF) discipline gives priority to the customer with the latest deadline.

In addition, as the system is stable, there exists for any Φ a residual lead time at the entry in service and a stationary lateness. These are given, respectively, by

$$R^\Phi = D - \mathrm{TA}^\Phi \ \text{ and } \ L^\Phi = \left(R^\Phi\right)^- .$$

We establish an analog of Theorem [4.10] in the case of a queue with deadlines.

THEOREM 4.13.– *For any convex function $g\colon \mathbf{R} \to \mathbf{R}$ and any admissible and non-preemptive discipline Ψ, independent of the service times,*

$$\mathbf{E}\left[g\left(R^{\text{EDF}}\right)\right] \leq \mathbf{E}\left[g\left(R^{\Psi}\right)\right]. \qquad [4.29]$$

Proof. The notations already introduced in the proof of theorem [4.10] are not repeated here. We note for every $j \geq 0$, $C_{\alpha(j)}$ the jth customer in the order of increasing deadlines (i.e. $D_{\alpha(i)} \leq D_{\alpha(j)}$ for $i < j$), and define the mapping

$$\zeta = \alpha \circ \psi \circ \phi^{-1} \circ \alpha^{-1},$$

where for any k, $\phi(k)$ is the index of the kth customer served by EDF. The stopping time τ, independent of the discipline under consideration, is defined as above, and we define the following random vectors

$$N = ((\xi_0,\, \sigma_0,\, D_0), \ldots, (\xi_{\tau-1},\, \sigma_{\tau-1},\, D_{\tau-1}))\,,$$
$$N^\zeta = \left(\left(\xi_0,\, \sigma_{\psi(0)},\, D_0\right), \ldots, \left(\xi_{\tau-1},\, \sigma_{\psi(\tau-1)},\, D_{\tau-1}\right)\right),$$

in other words N^ζ rearranges the service times according to ζ during the first busy period. Then, N and N^ζ have the same distribution as in [4.26]. For any $n \in [\![0,\, \tau-1]\!]$,

$$\begin{aligned}
B^{\text{EDF}}_{\alpha(n)}\left(N^{\zeta}\right) &= \sum_{i=0}^{F^{-1}(n)-1} \sigma_{\zeta\circ\alpha\circ\phi(i)} \\
&= \sum_{i=0}^{\phi^{-1}(n)-1} \sigma_{\alpha\circ\psi\circ\phi^{-1}\circ\alpha^{-1}\circ\alpha\circ\phi(i)} \\
&= \sum_{i=0}^{\phi^{-1}(n)-1} \sigma_{\alpha\circ\psi(i)} \\
&= B^{\psi}_{\zeta\circ\alpha(n)}\left(N\right).
\end{aligned}$$

We have therefore

$$\begin{aligned}
R^{\text{EDF}}_{\alpha(n)}\left(N^{\zeta}\right) &= E_{\alpha(n)}\left(N^{\zeta}\right) - B^{\text{EDF}}_{\alpha(n)}\left(N^{\zeta}\right) \\
&= E_{\alpha(n)}\left(N^{\zeta}\right) - B^{\Psi}_{\zeta\circ\alpha(n)}(N) \qquad [4.30] \\
&= E_{\alpha(n)}(N) - B^{\Psi}_{\zeta\circ\alpha(n)}(N).
\end{aligned}$$

But on the other hand, reminding the reader of the notions introduced in Appendix A.3,

LEMMA 4.14.– *ζ is the composition of ordering permutations of* $\mathbf{B}^{\Psi}_{\alpha}(N)$.

Proof of Lemma 4.14. The first integer n, if any, satisfying $\zeta(\alpha(n)) \neq \alpha(n)$ is such that at the $\alpha(n)$th end of service under Ψ (which is also the $\alpha(n)$th end of service under EDF since $\zeta(k) = k$ for $k \in [\![0,\, \alpha(n)-1]\!]$), there are in the system two customers C_{i_1} and C_{i_2} such that $D_{i_1} < D_{i_2}$ and EDF chooses C_{i_1} whereas Ψ chooses C_{i_2}. In other words, by denoting for $\ell = 1,\, 2,\, j_\ell = \alpha^{-1}(i_\ell)$, we have $B^{\Psi}_{\alpha(j_2)}(N) < B^{\Psi}_{\alpha(j_1)}(N)$ while $i_2 = \alpha(j_2) > \alpha(j_1) = i_1$. But as EDF gives priority to C_{i_1} over C_{i_2}, we have $\phi^{-1}(j_1) < \phi^{-1}(j_2)$. So

$$B^{\Psi}_{\zeta\circ\alpha(j_1)}(N) = B^{\Psi}_{\alpha\circ\psi\circ\phi^{-1}(j_1)}(N) < B^{\Psi}_{\alpha\circ\psi\circ\phi^{-1}(j_2)}(N) = B^{\Psi}_{\zeta\circ\alpha(j_2)}(N).$$

Thus, the permutation ζ_1 exchanging i and j fully orders $B^{\Psi}_{\alpha}(N)$. We conclude by noticing that ζ reads $\zeta = \zeta_p \circ \ldots \circ \zeta_1$, where ζ_i are such permutations. □

According to Lemma A.14 and Lemma 4.14, we thus have

$$\mathbf{E}_{\alpha}(N) - \mathbf{B}^{\text{EDF}}_{\zeta\circ\alpha}(N) \prec_c \mathbf{E}_{\alpha}(N) - \mathbf{B}^{\Psi}_{\alpha}(N) = \mathbf{R}^{\Psi}_{\alpha}(N),$$

and therefore, according to [4.30],

$$\mathbf{R}^{\text{EDF}}_{\alpha}\left(N^{\zeta}\right) \prec_c \mathbf{R}^{\Psi}_{\alpha}(N).$$

We conclude as in the proof of Theorem 4.10. □

Since $g: x \mapsto x^-$ is a convex function, and by definition of EDF and LDF we have particularly:

COROLLARY 4.15.– *The average tardiness at equilibrium is minimized by EDF and maximized by LDF, that is for any admissible discipline independent of the service times,*

$$\mathbf{E}\left[L^{\text{EDF}}\right] \leq \mathbf{E}\left[L^{\Psi}\right] \leq \mathbf{E}\left[L^{\text{LDF}}\right].$$

4.2. Processor sharing queue

We now introduce a system of a particular type, which has the capacity to serve all the customers simultaneously (thus there is no waiting room). The price for such a mechanism (which models many physical systems) is that the instantaneous processing speed for each customer is divided by the number of customers in the system. That is, if there are p customers in the system at a given time, their respective residual service time decrease by $1/p$ per unit of time.

We make the same probabilistic hypotheses, and keep the same notation as before. Since the server is working, whatever happens, at speed unit when the system is not empty, it is easy to be convinced that the workload sequence $(W_n, n \in \mathbf{N})$ satisfies Lindley's equation [4.1]. So there exists a stationary workload to the condition [4.3].

To characterize more accurately the equilibrium state, we aim to construct the stationary versions of remarkable characteristics, such as congestion of the system, waiting time or sojourn time. However, the service profile at equilibrium, from which we will deduce these quantities, has a different form for this system as for a classical G/G/1 queue. We show here how to construct the latter, using the renovating events.

Once again, we recall the notation and definitions introduced in Appendix A.3. We define for every n, S_n^{PS} the service profile at T_n^-, starting from an arbitrary profile $S_0^{\text{PS}} \in \mathcal{S}$, by ordering by convention, the non-zero terms of S_n^{PS} in decreasing order. Clearly, $S_n^{\text{PS}} \in \mathcal{S}$ for any $n \in \mathbf{N}$. We have the following result.

LEMMA 4.16.– *The sequence $\left(S_n^{\text{PS}}, n \in \mathbf{N}\right)$ is recursive in $\mathcal{S}$: denote for every $u \in \mathcal{S}$ and $x \in \mathbf{R}^+$,*

$$\begin{cases} \gamma_i(u, x) = \frac{1}{i}\left(x - \displaystyle\sum_{j=i+1}^{+\infty} u(j)\right) & \textit{for any } i \in \mathbf{N}^*; \\ i_0(u, x) = \min\{i \in \mathbf{N}^*;\, u(i) \leq \gamma_i(u, x)\}; & \\ \gamma(u, x) = \gamma_{(i_0(u,x)-1)\vee 1}(u, x); & \\ \left(F^{\text{PS}}(u, x)\right)(k) = [u(k) - \gamma(u, x)]^+ & \textit{for any } k \in \mathbf{N}^*. \end{cases}$$

For every $n \in \mathbf{N}$, we have

$$S^{\text{PS}}_{n+1} = F^{\text{PS}}\big(\underline{F^1\left(S^{\text{PS}}_n, \sigma \circ \theta^n\right)}, \xi \circ \theta^n\big), \qquad [4.31]$$

where $\underline{u}$ denotes the reordered version of u in decreasing order and $F^1(., \sigma)$ is defined by [4.12].

Proof. We denote as above

$$S^{\text{PS}}_{n++} = \underline{F^1\left(S^{\text{PS}}_n, \sigma \circ \theta^n\right)},$$

as the profile just after T_n, and the reordering of residual service times in decreasing order.

Denote for any $i \in \mathbf{N}^*$ such that $S^{\text{PS}}_{n++}(i) \neq 0$, $\tilde{C}_i$ the customer with residual service time $S^{\text{PS}}_{n++}(i)$ at T_n and $\tilde{T}'_i$, the virtual exit time of $\tilde{C}_i$ if no customer had arrived after T_n. Of course, $\tilde{T}'_i$ is not equal to $T_n + S^{\text{PS}}_{n++}(i)$ if C_n did not enter an empty system.

Let us recall that $N(S^{\text{PS}}_{n++})$ denotes the number of non-zero terms of S^{PS}_{n++}. For any $i \in [\![1, N(S^{\text{PS}}_{n++})]\!]$, $\tilde{C}_i$ and $\tilde{C}_{i-1}$ both receive the amount of service $S^{\text{PS}}_{n++}(i)$ on the interval of time $[T_n, \tilde{T}'_i]$. Thus, the remaining service time of $\tilde{C}_{i-1}$ at $\tilde{T}'_i$ equals $S^{\text{PS}}_{n++}(i-1) - S^{\text{PS}}_{n++}(i)$, and this customer as well as those who follow him are served at rate $\frac{1}{i-1}$ on the interval $\left[\tilde{T}'_i, \tilde{T}'_{i-1}\right]$. We therefore have the recurrence formula

$$\tilde{T}'_{i-1} = \tilde{T}'_i + \left(S^{\text{PS}}_{n++}(i-1) - S^{\text{PS}}_{n++}(i)\right)(i-1),\ i \in [\![2, N\left(S^{\text{PS}}_{n++}\right)]\!],$$

from which we deduce that for any $i \in [\![1, N(S^{\text{PS}}_{n++})]\!]$,

$$\tilde{T}'_i = T_n + iS^{\text{PS}}_{n++}(i) + \sum_{j=i+1}^{N\left(S^{\text{PS}}_{n++}\right)} S^{\text{PS}}_{n++}(j).$$

For any i, $\tilde{C}_i$ is served before T_{n+1} if $\tilde{T}'_i - T_n \leq \xi \circ \theta^n$, or in other words, if

$$S^{\text{PS}}_{n++}(i) < \frac{1}{i}\left(\xi \circ \theta^n - \sum_{j=i+1}^{N\left(S^{\text{PS}}_{n++}\right)} S^{\text{PS}}_{n++}(j)\right).$$

In particular, $i_0 = i_0\left(S^{\text{PS}}_{n++}, \xi \circ \theta^n\right)$ is the index of the last customer to leave the system before T_{n+1} (or 0 if there is no departure between T_n and T_{n+1}).

Hence, the system is not empty just before T_{n+1} if $i_0 < N$, and in this case $\left\{\tilde{C}_i,\, i \in [\![1,\, i_0 - 1]\!]\right\}$ is the family of customers present in the system at this time. For such a customer $\tilde{C}_i$, the remaining service time at T_{n+1} equals his remaining service time at $\tilde{T}'_{i_0}$ minus the amount of work received between $\tilde{T}'_{i_0}$ and T_{n+1}, i.e.

$$S^{\mathrm{PS}}_{n++}(i) - S^{\mathrm{PS}}_{n++}(i_0) - \frac{T_{n+1} - T'_{i_0}}{i_0 - 1} = S^{\mathrm{PS}}_{n++}(i) - \gamma\left(S^{\mathrm{PS}}_{n++},\, \xi \circ \theta^n\right).$$

Hence the result. □

A stationary profile corresponds to the only solution S^{PS} of the equation

$$S^{\mathrm{PS}} \circ \theta = G^{\mathrm{PS}}\left(S^{\mathrm{PS}}\right),\ \mathbf{P} - \text{ a.s.,} \tag{4.32}$$

where G^{PS} is the random mapping : $\mathcal{S} \to \mathcal{S}$ defined by

$$G^{\mathrm{PS}}(u) = F^{\mathrm{PS}}\left(\underline{F^1(u,\, \sigma)},\, \xi\right).$$

THEOREM 4.17.– *Provided [4.3] holds, [4.32] admits a unique solution with values in $\mathcal{S}$.*

Proof. The workload at a given time equals the sum of the terms of the service profile at this time. Hence, as in the proof of Theorem 4.7, the same sequence $(\mathcal{A}_n,\, n \in \mathbf{N})$ is a stationary sequence of renovating events of length 1 for every sequence $\left(S^{\mathrm{PS},\,\mu}_n,\, n \in \mathbf{N}\right)$ starting from $\mu \in \mathcal{S}$ such that $\sum_{i \in \mathbf{N}^*} \mu(i) \leq W$, where W is the only solution of [4.2]. Here again, Theorem 2.11 implies the existence of a solution to [4.32]. The uniqueness follows once again from the fact than any two solutions coincide on the non-negligible event $\{W = 0\}$. □

We can then argue as in section 4.1.4.

COROLLARY 4.18.– *There exists a unique stationary congestion X^{PS} under condition [4.3]. In addition, for any initial condition of the family $\mathcal{Z}$ (defined by [4.18]), the sequence converges with strong backward coupling toward X^{PS}.*

NOTE.– As in section 4.1.4, we can also build on S^{PS} to construct a stationary sojourn time (or in other words, a service time) in the system. This is left to the reader.

4.3. Parallel queues

Let us now consider a system fed by a G/G/ input, but where S servers (where $S \geq 1$) process the customer requests without loss or vacations. There is a waiting line of unlimited capacity for each server. We allocate customers upon arrival to one of the free servers, or if there is none, to the one having the smallest workload. Such an assignment policy will be termed *Join the Shortest Workload*, or JSW for short. Once assigned to a server, any customer remains as such until he leaves the system – hence there is no exchange.

4.3.1. *Preliminary result*

We begin by introducing a technical result, useful in the following. We work on a stationary ergodic quadruple $(\Omega,\, \mathcal{F},\, \mathbf{P},\, \theta)$, on which two random variables α and β are defined, integrable and with values in $\mathbf{R}_+$. We assume further that $\mathbf{P}\,(\beta > 0) > 0$. Let $F^{\alpha,\,\beta}$ be the random map: $\mathbf{R}^+ \to \mathbf{R}^+$ defined for any $x \in \mathbf{R}$ by

$$F^{\alpha,\,\beta}(x) = [x \vee \alpha - \beta]^+ . \qquad [4.33]$$

The following result stems from Loynes theory, as for the G/G/1 queue. Its proof is left as an exercise.

THEOREM 4.19.– *There exists a unique* $\mathbf{P}$*-a.s. finite solution to the equation*

$$Z \circ \theta = F^{\alpha,\,\beta}\,(Z)\,, \qquad [4.34]$$

given by

$$Y^{\alpha,\,\beta} = \left[\sup_{j\in\mathbf{N}^*}\left(\alpha\circ\theta^{-j} - \sum_{i=1}^{j}\beta\circ\theta^{-i}\right)\right]^+ . \qquad [4.35]$$

In addition, for any random variable Z *a.s. finite and positive, the SRS* $\left(Y_n^Z,\; n \in \mathbf{N}\right)$ *couples with* $\left(Y^{\alpha,\,\beta}\circ\theta^n,\; n\in\mathbf{N}\right)$*, and there exists a.s. an infinity of indexes* n *such that* $Y_n^Z = 0$ *if, and only if*

$$\mathbf{P}\left(Y^{\alpha,\,\beta} = 0\right) > 0. \qquad [4.36]$$

4.3.2. *The service profile*

We represent the system with S parallel queues upon the arrival of each customer by a random variable with values in the space $\overline{(\mathbf{R}^+)^S}$ (see A.3), representing the workload of each server at that moment, arranged in the increasing order.

We start at time 0 with an initial state $V_0 = (V_0(1),\, \ldots,\, V_0(s)) \in \overline{(\mathbf{R}^+)^S}$, where for every i, $V_0(i)$ represents the workload of the server having the ith smallest workload. Then, we represent the system at the arrival of customer C_n, $n \geq 0$ by the vector $V_n \in \overline{(\mathbf{R}^+)^S}$, where $V_n(i)$ is the ith smallest workload of a server just before the arrival of C_n. We call once again V_n, the *service profile* at this time.

The recurrence relation known as Kiefer and Wolfowitz's equation is then easy to check: for every $n \in \mathbf{N}$,

$$V_{n+1} = \left[\overline{V_n + \sigma \circ \theta^n.\mathbf{e}_1 - \xi \circ \theta^n.\mathbf{1}}\right]^+ . \qquad [4.37]$$

A service profile thus corresponds uniquely to a solution Y with values in $\overline{(\mathbf{R}^+)^S}$ for the equation

$$Y \circ \theta = \overline{[Y + \sigma.\mathbf{e}_1 - \xi.\mathbf{1}]^+} = G(Y). \qquad [4.38]$$

The mapping G is clearly a.s. continuous, and it is easy to observe that it is a.s. $\prec$-increasing: if u and v are such that $u \prec v$ in $\overline{(\mathbf{R}^+)^S}$, then for every $i \in [\![1, S]\!]$, a.s.

$$\begin{aligned} G(u)(i) &= [u(i) \vee ((u(1) + \sigma) \wedge u(i+1)) - \xi]^+ \\ &\leq [v(i) \vee ((v(1) + \sigma) \wedge v(i+1)) - \xi]^+ \qquad [4.39] \\ &= G(v)(i), \end{aligned}$$

setting $u(S + 1) = v(S + 1) = \infty$. Therefore, we can apply Theorem 2.4: there exists a $\prec$-minimal solution Y_∞, given by the almost sure limit of Loynes's sequence corresponding to $(V_n,\, n \in \mathbf{N})$, denoted in this case $(Y_n,\, n \in \mathbf{N})$.

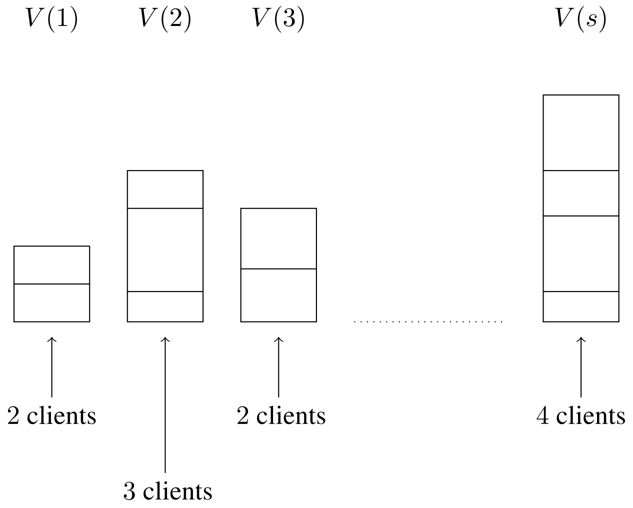

Figure 4.1. *The workload vector. The portions of the column $V(i)$ represent the service times of the customers who will be served by the server having a workload $V(i)$*

4.3.3. *Stability*

As for the G/G/1 queue, the a.s. finiteness of the solution Y_∞ (in the sense that all coordinates are finite a.s.) is not granted in general. We provide hereafter a stability condition for this system, that is a sufficient condition such that Y_∞ takes a.s. values in $\overline{(\mathbf{R}^+)^S}$.

THEOREM 4.20.– *If*

$$\mathbf{E}\left[\sigma\right] < S\mathbf{E}\left[\xi\right], \tag{4.40}$$

$Y_\infty(i) < \infty$ *a.s. for every* $i \in [\![1,\, S]\!]$ *and if*

$$\mathbf{E}\left[\sigma\right] > S\mathbf{E}\left[\xi\right],$$

$Y_\infty(i) = \infty$ *a.s. for every* $i \in [\![1,\, S]\!]$.

Proof. Loynes's sequence reads for all $n \in \mathbf{N}$,

$$Y_{n+1} = \overline{[Y_n \circ \theta^{-1} + \sigma \circ \theta^{-1}.\mathbf{e}_1 - \xi \circ \theta^{-1}.\mathbf{1}]^+},$$

which implies in particular according to [4.32] that

$$Y_{n+1}(S) = \left[\left((Y_n \circ \theta^{-1}(1) + \sigma \circ \theta^{-1}) \vee Y_n \circ \theta^{-1}(S)\right) - \xi \circ \theta^{-1}\right]^+. \tag{4.41}$$

As for any i, $(Y_n(i),\, n \in \mathbf{N})$ tends increasingly a.s. to $Y_\infty(i)$, taking the almost sure limit in [4.41] yields

$$\begin{aligned}(Y_\infty(S)) \circ \theta &= \Big[((Y_\infty(1) + \sigma) \vee Y_\infty(S)) - \xi\Big]^+ \\ &= F^{Y_\infty(1)+\sigma\,,\xi}\,(Y_\infty(S)),\end{aligned} \tag{4.42}$$

recalling the notation [4.33]. According to Theorem 4.19, we thus have a.s.

$$Y_\infty(S) = \left[\sup_{j \in \mathbf{N}^*} \left((Y_\infty(1) + \sigma) \circ \theta^{-j} - \sum_{i=1}^{j} \xi \circ \theta^{-i}\right)\right]^+.$$

Therefore, as $\mathbf{E}\left[\xi\right] \geq 0$, we have

$$\{Y_\infty(S) = \infty\} \subset \{Y_\infty(1) = \infty\}$$

up to a negligible event. Hence, as Y_∞ is arranged in increasing order,

$$\{Y_\infty(S) = \infty\} \subset \{Y_\infty(i) = \infty,\, i = 1\ldots,\, S\}.$$

Equality [4.42] implies that the event on the left-hand side is θ-contracting. So we are in the following alternative

$$Y_\infty(i) < \infty \text{ a.s. for any } i \in [\![1,\, S]\!], \tag{4.43}$$

or

$$Y_\infty(i) = \infty \text{ a.s. for any } i \in [\![1,\, S]\!]. \tag{4.44}$$

As each server works at unit speed if it has someone to serve, we have a.s. for every $n \in \mathbf{N}$ that

$$\left(\sum_{i=1}^{S} Y_{n+1}(i)\right) \circ \theta = \sum_{i=2}^{S} [Y_n(i) - \xi]^+ + [Y_n(1) + \sigma - \xi]^+. \tag{4.45}$$

Therefore, by denoting

$$S_n = \sum_{i=1}^{S} Y_n(i),$$

the sum of the workloads of the servers at the arrival time of C_n, we have

$$\begin{aligned} S_{n+1} \circ \theta - S_n &= \sum_{i=2}^{S} \Big([Y_n(i) - \xi]^+ - Y_n(i)\Big) + [Y_n(1) + \sigma - \xi]^+ - Y_n(1) \\ &= -\sum_{i=1}^{S} (\xi \wedge Y_n(i)) - ((\xi - \sigma) \wedge Y_n(1)). \end{aligned} \tag{4.46}$$

As $(Y_n,\, n \in \mathbf{N})$ is a.s. $\prec$-increasing, $(S_n,\, n \in \mathbf{N})$ is a.s. increasing. Particularly, by θ-invariance $\mathbf{E}\,[S_{n+1} \circ \theta] - \mathbf{E}\,[S_n] \geq 0$, which gives with [4.46] that

$$\sum_{i=2}^{S} \mathbf{E}\,[\xi \wedge Y_n(i)] + \mathbf{E}\,[(\xi - \sigma) \wedge Y_n(1)] \leq 0.$$

Taking the limit, by monotone convergence,

$$\sum_{i=2}^{S} \mathbf{E}\,[\xi \wedge Y_\infty(i)] + \mathbf{E}\,[(\xi - \sigma) \wedge Y_\infty(1)] \leq 0. \tag{4.47}$$

So [4.44] implies that

$$\mathbf{E}\,[\sigma] \geq S\mathbf{E}\,[\xi],$$

which shows the sufficiency of the condition.

On the other hand, as $x^+ + y^+ \geq (x+y)^+$ for all real numbers x and y, we have according to [4.45] that for every $n \in \mathbf{N}$, a.s.

$$S_{n+1} \circ \theta \geq [S_n + \sigma - s\xi]^+ .$$

Consider Loynes's sequence $\left(\tilde{M}_n,\, n \in \mathbf{N}\right)$ corresponding to the workload of a G/G/1 queue of generic inter-arrival $\tilde{\xi} = S\xi$: we have $\tilde{M}_0 = 0$ a.s. and for any $n \in \mathbf{N}$,

$$\tilde{M}_{n+1} \circ \theta = \left[\tilde{M}_n + \sigma - S\xi\right]^+ .$$

The mapping $x \mapsto [x + \sigma - S\xi]^+$ being a.s. increasing, as $S_0 = 0 = \tilde{M}_0$ a.s., an immediate induction shows that $S_n \geq \tilde{M}_n$ a.s. for every n. Let us denote $\tilde{M}_\infty$ as the minimal solution of [4.4] for the r.v. $\tilde{\xi}$. According to Theorem 4.2, to the limit we obtain that provided $\mathbf{E}[\sigma] > \mathbf{E}\left[\tilde{\xi}\right] = S\mathbf{E}[\xi]$,

$$S_\infty \geq \tilde{M}_\infty = \infty,$$

which concludes the proof. □

NOTE.– It is easy to construct examples where Y_∞ is not the only solution to [4.38] with finite coordinates. It is possible to construct a maximal finite solution to this equation by focusing on translated versions of Loynes's sequence by a constant - see the references at the end of the Chapter.

We now show that, similarly to the single server queue, the stable queue returns almost surely infinitely often into a state of small congestion (see [4.4]).

THEOREM 4.21.– *Provided [4.40] holds, the minimal solution Y_∞ of [4.38] satisfies*

$$\mathbf{P}(Y_\infty(1) = 0) > 0. \qquad [4.48]$$

In particular, there are $\mathbf{P}$*-a.s. an infinite number of times where the system has at less than S customers.*

Proof. According to [4.47], if we had $Y_\infty(1) \geq \xi - \sigma$ and $Y_\infty(2) \geq \xi$ a.s., we would then have $\mathbf{E}[\sigma] \geq s\mathbf{E}[\xi]$, an absurdity. Therefore, on a non-negligible event, we have $Y_\infty(1) < \xi - \sigma$ or $Y_\infty(2) < \xi$, implying that $Y_\infty(1) \circ \theta = 0$.

We show in Section 4.4 that a server cannot be idle if some customer is in line in another queue: that customer would have chosen the empty queue, which had upon his arrival, the least workload. The event $\{Y_\infty(1) = 0\}$ thus corresponds to the equilibrium states at which at most $S-1$ servers are busy, that is, there are at most $S-1$ customers in the system. We can then deduce from [4.48] that the system visits a similar state almost surely infinitely often, as in Corollary [4.4]. □

NOTE.– It can be verified through examples that the condition [4.40] does not imply that

$$\mathbf{P}\left(X_\infty(s) = 0\right) > 0,$$

and therefore that the system empties almost surely infinitely often: just set the pair of real numbers $(x,\, y)$, such that $x > 0$ and $x < y \leq 2x$, and consider the deterministic system where $\xi = x$, $\sigma = y$ a.s. and $S = 2$. Then the minimal solution is the smallest ordered pair v such that

$$v = \overline{\Big((v(1) + y - x)^+,\, (v(2) - x)^+\Big)},$$

clearly given by

$$v = (0,\, y - x).$$

4.3.4. *Comparison of systems*

As in section 4.1.2, the $\prec$-monotonicity of the SRS of service profiles allows us to compare the equilibrium states of various systems under the JSW policy, according to the stochastic orderings of the random variables under consideration.

THEOREM 4.22.– *Let two systems of S parallel queues driven respectively by the random variables $(\sigma,\, \xi)$ and $(\bar\sigma,\, \bar\xi)$, possibly defined on two different quadruples. If it holds that*

$$\left(\bar\sigma,\, -\bar\xi\right) \leq_{st} (\sigma,\, -\xi),$$

then the respective minimal solutions W and $\bar W$ of [4.38] for $(\sigma,\, \xi)$ and $(\bar\sigma,\, \bar\xi)$ satisfy

$$\bar W \leq_{st} W.$$

Proof. Apply Theorem 2.15 to $\bar\alpha = (\bar\sigma,\, \bar\xi)$, $\alpha = (\sigma,\, \xi)$ and

$$f\ :\left\{\begin{array}{ll} \mathbf{R}^S \times \mathbf{R}^2 & \to \mathbf{R}^S \\ \big(x,\, (y(1), y(2)\big) & \mapsto \left[\overline{x + y(1).\mathbf{e}_1 + y(2).\mathbf{1}}\right]^+ . \end{array}\right.$$

We deduce easily from [4.39] that f is $\prec$-non-decreasing in its two arguments. □

NOTE.– As above, if one assumes that $(\sigma \circ \theta^n,\, n \in \mathbf{N})$ (respectively $\left(\bar\sigma \circ \bar\theta,\, n \in \mathbf{N}\right)$) is independent of $(\xi \circ \theta^n,\, n \in \mathbf{N})$ (respectively $\left(\bar\xi \circ \bar\theta^n,\, n \in \mathbf{N}\right)$), then it is easily checked that the above theorem applies under either one of the following two conditions:

$$\xi \overset{\mathcal{L}}{=} \bar\xi \text{ and } \bar\sigma \leq_{st} \sigma,$$

$$\sigma \overset{\mathcal{L}}{=} \bar\sigma \text{ and } \xi \leq_{st} \bar\xi.$$

4.3.5. *The optimal allocation*

Here, we show in what sense the JSW policy is optimal. Let us consider a system where, starting from an arbitrary service profile U at the arrival of C_0, we assign every incoming customer to the Ith server (where I is a fixed integer in $[\![1,\, S]\!]$) in the order of increasing workloads, rather than the first one. For this model, the sequence $\left(\tilde{V}_n^U,\, n \in \mathbf{N}\right)$ of service profiles satisfies the recurrence relation

$$\tilde{V}_{n+1}^U = \left[\overline{\tilde{V}_n^U + \sigma \circ \theta^n.\mathbf{e}_I - \xi \circ \theta^n.\mathbf{1}}\right]^+, \text{ a.s..} \qquad [4.49]$$

Let us denote as above, $\left(V_n^U,\, n \in \mathbf{N}\right)$ the SRS of service profiles initially equal to U, when applying the JSW policy. In the following theorem, "$\prec_*$" denotes the partial order on $\overline{(\mathbf{R}^+)^S}$ introduced in Definition A.21.

THEOREM 4.23.– *For any random variable U with values in $\overline{(\mathbf{R}+)^S}$, for any $n \in \mathbf{N}$, a.s.*

$$V_n^U \prec_* \tilde{V}_n^U; \qquad [4.50]$$

$$\forall\, \ell \geq I,\, V_n^U(\ell) \leq \tilde{V}_n^U(\ell). \qquad [4.51]$$

Proof. We proceed by induction. Relations [4.50] and [4.51] are clearly satisfied for $n = 0$. Suppose that they hold at rank n. Setting again $u(S+1) = \infty$ for any $u \in \overline{(\mathbf{R}^+)^S}$, we then have a.s. for any $\ell \geq I$,

$$\begin{aligned} V_{n+1}^U(\ell) &= \left[V_n^U(\ell) \vee \left(\left(V_n^U(1) + \sigma \circ \theta^n\right) \wedge V_n^U(\ell+1)\right) - \xi \circ \theta^n\right]^+ \\ &\leq \left[\tilde{V}_n^U(\ell) \vee \left(\left(\tilde{V}_n^U(I) + \sigma \circ \theta^n\right) \wedge \tilde{V}_n^U(\ell+1)\right) - \xi \circ \theta^n\right]^+ \\ &= \tilde{V}_{n+1}^U(\ell), \end{aligned}$$

and [4.51] holds at rank $n+1$. Particularly, this implies that

$$\sum_{i=k}^{S} V_{n+1}^U(i) \leq \sum_{i=k}^{n} \tilde{V}_{n+1}^U(i) \text{ for all } k \geq I.$$

Therefore, it is sufficient to show that

$$\sum_{i=k}^{S} V_{n+1}^U(i) \leq \sum_{i=k}^{n} \tilde{V}_{n+1}^U(i) \text{ for all } k \leq I - 1 \qquad [4.52]$$

to check [4.50] at rank $n+1$. Let us therefore fix $k \leq I-1$ and form the following sums

$$\begin{aligned}\sum_{i=k}^{S} V_{n+1}^{U}(i) = & \sum_{i=k+1}^{S} \left[V_n^U(i) - \xi\circ\theta^n\right]^+ \\ & + \left[\left(V_n^U(1) + \sigma\circ\theta^n\right) \vee V_n^U(k) - \xi\circ\theta^n\right]^+ ;\end{aligned} \qquad [4.53]$$

$$\begin{aligned}\sum_{i=k}^{S} \tilde{V}_{n+1}^{U}(i) = & \sum_{i=k;k\neq I}^{S} \left[\tilde{V}_n^U(i) - \xi\circ\theta^n\right]^+ \\ & + \left[\tilde{V}_n^U(I) + \sigma\circ\theta^n - \xi\circ\theta^n\right]^+ .\end{aligned} \qquad [4.54]$$

If $V_n^U(k) \geq V_n^U(1) + \sigma\circ\theta^n$, then [4.53] equals

$$\sum_{i=k}^{S} \left[V_n^U(i) - \xi\circ\theta^n\right]^+ \leq \sum_{i=k}^{S} \left[\tilde{V}_n^U(i) - \xi\circ\theta^n\right]^+ \leq \sum_{i=k}^{S} \tilde{V}_{n+1}^U(i) \text{ a.s.,}$$

where we used (i) of Lemma A.15 in the first inequality. It remains only to consider the case where $V_n^U(k) \leq V_n^U(1) + \sigma\circ\theta^n$. Then [4.53] equals

$$\left[V_n^U(1) + \sigma\circ\theta^n - \xi\circ\theta^n\right]^+ + \sum_{i=k+1}^{S} \left[V_n^U(i) - \xi\circ\theta^n\right]^+ . \qquad [4.55]$$

The vector $\left(V_n^U(1),\, V_n^U(k+1),\, \ldots,\, V_n^U(S)\right)$ is fully ordered and a.s.

$$(\xi-\sigma,\, \xi,\, \ldots,\, \xi)\circ\theta^n = \overline{\left(\xi,\, \ldots,\, \underbrace{\xi}_{I-1},\, \underbrace{\xi-\sigma}_{I},\, \underbrace{\xi}_{I+1},\, \ldots,\, \xi\right)\circ\theta^n}.$$

Hence, according to [A.9] and [A.6],

$$\begin{aligned}&\left(V_n^U(1),\, V_n^U(k+1),\, \ldots,\, V_n^U(S)\right) - (\xi-\sigma,\, \xi,\, \ldots,\, \xi)\circ\theta^n \\ &\prec_c \left(V_n^U(1),\, V_n^U(k+1),\, \ldots,\, V_n^U(S)\right) \\ &\qquad\qquad - \left(\xi,\, \ldots,\, \underbrace{\xi}_{I-1},\, \underbrace{\xi-\sigma}_{I},\, \underbrace{\xi}_{I+1},\, \ldots,\, \xi\right)\circ\theta^n ,\end{aligned}$$

where $\prec_c$ denotes the Schur-convex ordering. As, for any p, the function $u \rightarrow \sum_{i=1}^{p} u^+$ is symmetric and convex from $\overline{(\mathbf{R}^+)^p}$ in $\mathbf{R}$, according to [A.7] the sum [4.55] satisfies a.s.

$$
\begin{aligned}
&\left[V_n^U(1)+\sigma\circ\theta^n-\xi\circ\theta^n\right]^+ + \sum_{i=k+1}^{S}\left[V_n^U(i)-\xi\circ\theta^n\right]^+ \\
&\quad\le \left[V_n^U(1)-\xi\circ\theta^n\right]^+ + \sum_{i=k+1;i\neq I}^{S}\left[V_n^U(i)-\xi\circ\theta^n\right]^+ \\
&\qquad + \left[V_n^U(I)+\sigma\circ\theta^n-\xi\circ\theta^n\right]^+ \\
&\quad\le \sum_{i=k;i\neq I}^{S}\left[V_n^U(i)-\xi\circ\theta^n\right]^+ + \left[V_n^U(I)+\sigma\circ\theta^n-\xi\circ\theta^n\right]^+ .
\end{aligned}
\qquad [4.56]
$$

Moreover, as $V_n^U(I)\le\tilde{V}_n^U(I)$ from [4.50], the assertions (ii) and (i) of Lemma A.15 show that [4.51] implies that a.s.

$$\left[\overline{V_n^U+\sigma\circ\theta^n.\mathbf{e}_I}-\xi\circ\theta^n.\mathbf{1}\right]^+ \prec_* \left[\overline{\tilde{V}_n^U+\sigma\circ\theta^n.\mathbf{e}_I}-\xi\circ\theta^n.\mathbf{1}\right]^+ .$$

Particularly,

$$
\begin{aligned}
&\sum_{i=k;i\neq I}^{S}\left[V_n^U(i)-\xi\circ\theta^n\right]^+ + \left[V_n^U(I)+\sigma\circ\theta^n-\xi\circ\theta^n\right]^+ \\
&\quad\le \sum_{i=k;i\neq I}^{S}\left[\tilde{V}_n^U(i)-\xi\circ\theta^n\right]^+ + \left[\tilde{V}_n^U(I)+\sigma\circ\theta^n-\xi\circ\theta^n\right]^+
\end{aligned}
\qquad [4.57]
$$

and we deduce [4.52] from [4.53, 4.55, 4.56], and [4.57]. Relation [4.50] is thus verified at rank $n+1$. □

In particular, the above result shows that, starting from the same initial service profile and subject to the same traffic, the JSW policy optimizes the total workload with respect to any other fixed allocation to another server, since for every $I\in[1,\,S]$, a.s.

$$\sum_{i=1}^{S}V_n^U(i)\le\sum_{i=1}^{s}\tilde{V}_n^U(i).$$

On the other hand, as $\tilde{V}_n^U$ is fully ordered, [4.51] implies that $V_n^U(1)\le\tilde{V}_n^U(I)$ a.s., that is, the proposed waiting time to the nth customer is as well minimized.

We can extend these results to the steady state. According to [4.49], a stationary service profile for the allocating to the Ith server is a $\overline{(\mathbf{R}^+)^S}$-valued solution to the equation

$$\tilde{V}\circ\theta=\left[\overline{\tilde{V}+\sigma.\mathbf{e}_I}-\xi.\mathbf{1}\right]^+=\tilde{G}(\tilde{V})\text{ a.s.} \qquad [4.58]$$

For any u and v such that $u \prec v$, for every $i \geq I$, a.s.

$$\begin{aligned}\tilde{G}(u)(i) &= [u(i) \vee ((u(I) + \sigma) \wedge u(i+1)) - \xi]^+\\ &\leq [v(i) \vee ((v(I) + \sigma) \wedge v(i+1)) - \xi]^+\\ &= \tilde{G}(v)(i),\end{aligned}$$

whereas for all $i < I$, a.s.

$$\tilde{G}(u)(i) = [u(i) - \xi]^+ \leq [v(i) - \xi]^+ = \tilde{G}(v)(i).$$

The mapping $\tilde{G}$ is therefore a.s. $\prec -$ increasing, and clearly continuous. Loynes's Theorem then yields the minimal solution of [4.58], given by $\tilde{Y}_\infty$, that is the almost sure coordinatewise limit of the corresponding Loynes sequence $(\tilde{Y}_n, n \in \mathbf{N}) = \left(\tilde{V}_n^0 \circ \theta^{-n},\ n \in \mathbf{N}\right)$. According to Theorem 4.23, to the limit, the minimal solutions satisfy a.s.

$$Y_\infty \prec_* \tilde{Y}_\infty \text{ and } Y_\infty(1) \leq \tilde{Y}_\infty(I).$$

The JSW policy hence minimizes the total workload at equilibrium and the proposed waiting time.

It is also immediate to observe that for every $n \in \mathbf{N}$, a.s. $V_n^0(\ell) = 0$ for every $\ell < I$, as no service is ever provided by the first $I-1$ servers, always inactive. Therefore, the restriction of $\tilde{Y}_\infty$ to its $S-(I-1)$ last coordinates clearly reads as the minimal solution of [4.38], that is the stationary profile for a JSW system of $S \equiv S-(I-1)$ queues. We summarize these results in the following two corollaries.

COROLLARY 4.24.– *For every $I \in [\![1,\ S]\!]$, the $\prec$-minimal solution $\tilde{Y}_\infty$ of [4.58] satisfies $\tilde{Y}_\infty(s) < \infty$ a.s. provided that*

$$\mathbf{E}[\sigma] < (S - I + 1)\mathbf{E}[\xi].$$

In addition, if Y_∞ denotes the minimal solution of [4.38] we have

$$\begin{gathered}Y_\infty \prec_* \tilde{Y}^I_\infty :\\ \forall\, \ell \geq I,\ Y_\infty(\ell) \leq Y^I_\infty(\ell),\end{gathered}$$

and in particular $Y_\infty(1) \leq \tilde{Y}_\infty(I)$.

COROLLARY 4.25.– *Let $1 \leq S' \leq S$. Denote Y^S_∞ and $Y^{S'}_\infty$ as the minimal solutions of equation [4.38], respectively for S and S' servers. Under the condition $\mathbf{E}[\sigma] < S'\mathbf{E}[\xi]$, where both are finite a.s., they satisfy*

$$Y_{\infty}^{S}(S-i) \leq Y_{\infty}^{S'}(S'-i) \textit{ a.s. for all } i \in [\![0,\, S'-1]\!].$$

The latter result states precisely the (intuitively clear) property, that in a JSW system, an increase in the number of servers reduces the workload at equilibrium: if both systems are stable, the workload of each server of the small system is larger than that of the corresponding server (in the decreasing order of workloads) in the big system. The last inequality means that that the waiting time is minimized by the bug system.

4.4. The queue with S servers

We now consider a system closely related to the previous one. There are S servers processing the requests without loss nor interruptions, but the architecture of the queueing system is different: if all S servers are busy, the customers are queued in a *single* queue of infinite size, and are assigned to the first server available, on a First come, First served basis. Notations and probabilistic hypotheses are the same as above – we thus consider a stationary G/G /S/∞/FIFO queue.

In this section, we show that this system amounts to S parallel queues, under the JSW policy. Particularly, the stability condition remains [4.40].

We again represent the queue by the sequence of service profiles, keeping track of the service times of *all* the customers in the system at current time. Specifically, we fix $\hat{V}_0 \in \mathcal{S}$ and we denote for every $n \in \mathbf{N}$, $\hat{V}_n$ the element of $\mathcal{S}$ which represents the residual service time of all customers in the system at the arrival of nth customer:

(i) If the S servers are busy:

- the first S coordinates of $\hat{V}_n$ are the residual service times of the S customers in service, ranked in decreasing order;

- the following coordinates represent the service times requested by customers in queue, arranged in the order of priorities. In other words, for every $i \in [\![S+1,\, N(\hat{V}_n)]\!]$, $\hat{V}_n(i)$ represents the service time of the ith customer in queue, according to the order of arrivals. Particularly, the customer of service time $\hat{V}_n\,(N(S+1))$ will be the next to enter service, and so on.

(ii) If $j \leq S$ servers are busy, $N(\hat{V}_n) = j$ and the coordinates $\hat{V}_n(i)$, $i \in [\![1,\, j]\!]$ represent the residual service time of the customers in service, arranged in decreasing order.

It is then easy to see that the sequence $(\hat{V}_n, n \in \mathbf{N})$ so defined is recursive on the canonical space of arrivals and services, and to make the recursive function explicit. In that purpose, we construct for every $u \in \mathcal{S}$, the family of sets of indexes $\mathcal{A}_1(u)$, $\mathcal{A}_2(u)$, $\ldots$, $\mathcal{A}_S(u)$ by induction, in the following manner:

– we start by setting $\mathcal{A}_1^0(u) = \mathcal{A}_2^0(u) = \cdots = \mathcal{A}_S^0(u) = \emptyset$;

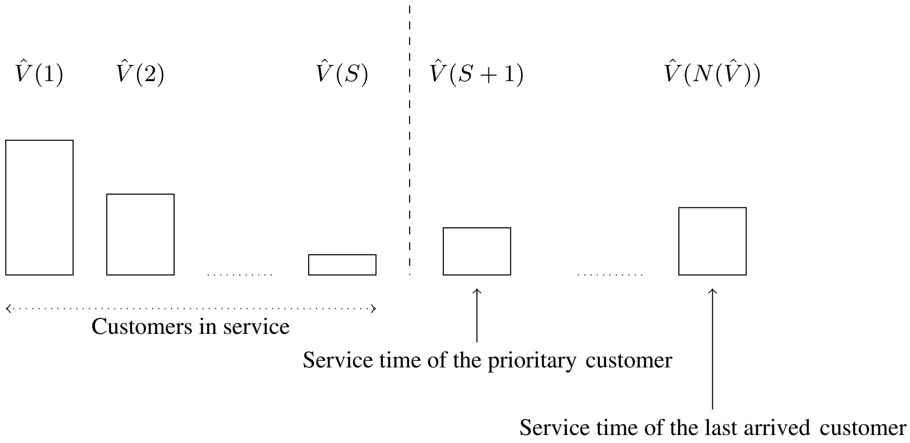

Figure 4.2. *The service profile of the G/G/S queue*

– then, for every $j \in [\![1,\ N(u) - S]\!]$, we denote

$$\varphi_u(j) = \underset{i \in [1,\ S]}{\text{Argmin}} \left\{ u(i) + \sum_{k \in \mathcal{A}_i^{j-1}(u)} u\left(S + k\right) \right\} \qquad [4.59]$$

and we set

$$\begin{cases} \mathcal{A}^j_{\varphi_u(j)}(u) = \mathcal{A}^{j-1}_{\varphi_u(j)}(u) \cup \{j\} \\ \mathcal{A}^j_i(u) = \mathcal{A}^{j-1}_i(u),\ \text{for every } i \neq \varphi_u(j); \end{cases}$$

– we finally set

$$\mathcal{A}_i(u) = \mathcal{A}_i^{N(u)-S}(u) \text{ for every } i \in [\![1,\ S]\!].$$

As usual, it is understood that $\sum_{k \in \emptyset} \ldots = 0$ and we fix $\mathcal{A}_i(u) = \emptyset$ if $N(u) \leq S$. We have the following result.

THEOREM 4.26.– *Starting from $\hat{V}_0 \in \mathcal{S}$, we have for every $n \in \mathbf{N}$,*

$$\hat{V}_{n+1} = \hat{G}^3 \circ \hat{G}^2\left(., \xi \circ \theta^n\right) \circ \hat{G}^1\left(., \sigma \circ \theta^n\right)(\hat{V}_n),$$

where the mappings $\hat{G}^1, \hat{G}^2$ and $\hat{G}^3$ are, respectively, defined by [4.60], [4.63] and [4.64].

Proof. Assume that C_n finds a system having a service profile $\hat{V}_n$ upon arrival. First, the service time $\sigma \circ \theta^n$ brought by C_n is placed at the place of lowest priority, in other words the service profile becomes at first

$$\hat{V}_{n+} = \hat{V}_n + (\sigma \circ \theta^n) \, .\mathbf{e}_{N\left(\hat{V}_n\right)+1} =: \hat{G}^1(\hat{V}_n, \sigma \circ \theta^n). \qquad [4.60]$$

Then, the customers possibly in line are assigned to the various servers.

(i) If $N(\hat{V}_n) < S$, there are available servers at the arrival of C_n, and therefore the service time $\sigma \circ \theta^n$ is assigned to the first that becomes available. $\hat{V}_{n+}$ remains unchanged since the service time of the arriving customer is by construction given by $\hat{V}_{n+}(N(\hat{V}_{n+}) + 1)$.

(ii) If $N\left(\hat{V}_{n+}\right) \geq S$, there is no available server upon the arrival of C_n. It suffices to understand the construction of the sets $\mathcal{A}_i(\hat{V}_n)$ to write easily the recurrence function. Let us call "server i", $i \in [\![1,\, S]\!]$, the server whose customer in service has a residual service equal to $\hat{V}_n(i)$ upon the arrival of C_n (particularly, the server 1 has the largest remaining workload and the server S the smallest one at this instant). Let us also denote $\tilde{C}_j$, $j \in [\![1,\, N(\hat{V}_{n+}) - S]\!]$ the customer (if any) in line at the arrival of C_n, whose service time is given by $\hat{V}_{n+}(i)$ (particularly, $\tilde{C}_1$ is the customer on priority at this time and $\tilde{C}_{N\left(\hat{V}_{n+}\right)-S} = \tilde{C}_{N(\hat{V}_n)+1-S}$ is the customer C_n just arrived). Notice that both these indexations are related to the situation at the arrival time of C_n, in other words they depend on n.

First, notice that $1 \in \mathcal{A}_S(\hat{V}_{n+})$ by definition. The first customer to possibly enter service after the arrival of C_n is $\tilde{C}_1$. This customer will join the first server that becomes available, that is the server S. The second customer to enter service $\tilde{C}_2$ will then join the server S if

$$\hat{V}_{n+}(S) + \hat{V}_{n+}(S+1) \leq \hat{V}_{n+}(S-1),$$

or the server $S - 1$ if

$$\hat{V}_{n+}(S) + \hat{V}_{n+}(S+1) > \hat{V}_{n+}(S-1).$$

Notice, that in the first case $2 \in \mathcal{A}_S(\hat{V}_{n+})$ and in the second case, $2 \in \mathcal{A}_{S-1}(\hat{V}_{n+})$, by the very definition of the sets $\mathcal{A}_j(\hat{V}_{n+})$. And so on, we observe that for every $j \in [\![1,\, N(\hat{V}_{n+}) - S]\!]$, $\varphi_{\hat{V}_{n+}}(j)$ defined by [4.59] represents the index (upon the arrival of C_n) of the server that actually serves the customer $\tilde{C}_j$, as it is the first one for which the remaining workload vanishes after the beginning of service of the customer $\tilde{C}_{j-1}$ (or after the arrival of C_n if $j = 1$). In other words, for any $i \in [\![1,\, S]\!]$ and any $j \in [\![1,\, N(\hat{V}_{n+}) - S]\!]$, $\mathcal{A}_i^j(\hat{V}_{n+})$ represents the set of indexes (in the indexation of the $\tilde{C}_k$'s) of those customers in the system after the arrival of C_n, arrived strictly before $\tilde{C}_{j+1}$, and who will enter service with server i. Therefore, $\mathcal{A}_i(\hat{V}_{n+})$ denotes the set of indexes of all customers present just after the arrival of C_n and who will be assigned to server i.

It is easy to see, as the discipline is FCFS, that the sets $\mathcal{A}_i^j(\hat{V}_{n+})$ differ only from $\mathcal{A}_i^j(\hat{V}_n)$ for the index $j = N(\hat{V}_{n+}) - S$ of the customer just entered, which is added to the set $\mathcal{A}_{\varphi_{\hat{V}_{n+}}(N(\hat{V}_{n+})-S)}^{N(\hat{V}_n)-S}$.

Thus, between the arrivals of C_n and C_{n+1}, the server of index i at the arrival of C_n provides a quantity of service equal to

$$\xi \circ \theta^n \wedge \left(\hat{V}_{n+}(i) + \sum_{k \in \mathcal{A}_i(\hat{V}_{n+})} \hat{V}_{n+}\left(S + k\right) \right).$$

If $\xi \circ \theta^n$ is less than the latter sum, the server i is still busy when C_{n+1} enters the system. The last customer to have come into service at server i before the arrival of C_{n+1} is then:

- the customer who was already in service at the arrival of C_n if $\hat{V}_{n+}(i) > \xi \circ \theta^n$;
- otherwise, the customer $\tilde{C}_{\psi(i)}$, where

$$\psi(i) = \begin{cases} \min \left\{ j \in \mathcal{A}_i(\hat{V}_{n+}) \mid \hat{V}_{n+}(i) + \sum_{k \in \mathcal{A}_i^j(\hat{V}_{n+})} \hat{V}_{n+}\left(S + k\right) > \xi \circ \theta^n \right\}, \\ \text{or} \\ \max \mathcal{A}_i(\hat{V}_{n+}) \text{ if the previous set is empty,} \end{cases}$$

since, as easily checked, $\psi(i)$ denotes the index of the last customer who had the time to reach the server of index i (at the arrival of C_n) between the arrival times of C_n and C_{n+1}.

In other words, for every $j \in [\![1,\, N(\hat{V}_{n+}) - S]\!]$, the customer $\tilde{C}_j$ comes into service (with the server $\varphi_{\hat{V}_{n+}}(j)$) before the arrival of C_{n+1} if and only if $j \leq \psi(\varphi_{\hat{V}_{n+}}(j))$;.

In both cases (i) and (ii), the sequence $\hat{V}_{n++}$ representing the service profile just before the arrival of C_{n+1} and before reordering, reads

$$\hat{V}_{n++}(i) = \left[\hat{V}_{n+}(i) + \sum_{k \in \mathcal{A}_i^{\psi(i)}(\hat{V}_{n+})} \hat{V}_{n+}\left(k\right) - \xi \circ \theta^n \right]^+ ; i \in [\![1,\, S]\!], \quad [4.61]$$

and for every $j \in [\![1,\, N(\hat{V}_{n+}) - S]\!]$,

$$\hat{V}_{n++}\left(S + j\right) = \begin{cases} 0 & \text{if } j \leq \psi(\varphi_{\hat{V}_{n+}}(j)); \\ \hat{V}_{n+}(S + j) & \text{otherwise,} \end{cases} \quad [4.62]$$

with the convention $\sum_{k \in \emptyset} \ldots = 0$.

As above, denote $\tilde{G}^2(., \xi \circ \theta^n)$ the application: $(\mathbf{R}^+)^{\mathbf{N}} \times \mathbf{R} \to (\mathbf{R}^+)^{\mathbf{N}}$ defined by [4.61] and [4.62], and such that

$$\hat{V}_{n++} = \hat{G}^2(\hat{V}_{n+}, \xi \circ \theta^n). \qquad [4.63]$$

Finally, we rearrange in decreasing order, the remaining service times of the customer in service (which are the first S coordinates of $\hat{V}_{n++}$). The possible following non-zero coordinates represent the customers in line in the order of priorities, until the index $N\left(\hat{V}_{n++}\right)$. By denoting $\hat{G}^3$ as the mapping : $(\mathbf{R}^+)^{\mathbf{N}} \to \mathcal{S}$ which arranges the first S components of a sequence in decreasing order, and which deletes the following null components, retaining their order (keeping only the residual services time of the customers still in line at the arrival of C_{n+1}), we therefore have

$$\hat{V}_{n+1} = \hat{G}^3\left(\hat{V}_{n++}\right). \qquad [4.64]$$

Hence the result. □

Define the mapping

$$\Psi\colon \begin{cases} \mathcal{S} & \to (\mathbf{R}^+)^S \\ u & \mapsto \Psi(u) \text{ such that} \\ & \quad \Psi(u)(i) = u(i) + \displaystyle\sum_{j \in \mathcal{A}_i(u)} u(S+j) \text{ for all } i \in [\![1,\, S]\!]. \end{cases}$$

In this case, for every $n \in \mathbf{N}$ and $i \in [\![1,\, S]\!]$, the amount $\Psi(\hat{V}_n)(i)$ represents the "virtual" workload of the server of index i just before the arrival of C_n, consisting in the service times of all the customers to be served by this server. We have the following result.

LEMMA 4.27.– *For any $n \in \mathbf{N}$,*

$$\overline{\Psi(\hat{V}_{n+1})} = \left[\overline{\overline{\Psi(\hat{V}_n)} + \sigma \circ \theta^n . \boldsymbol{e}_1 - (\xi \circ \theta^n)\, .\mathbf{1}}\right]^+. \qquad [4.65]$$

Proof. Upon the arrival of C_n,

(i) if some servers are free (i.e. $N(\hat{V}_n) < S$), $\overline{\Psi(\hat{V}_n)}$ is nothing but the arranged version in increasing order of the restriction of $\hat{V}_n$ to its first S components. Then,

$$\hat{V}_{n+} = \hat{V}_n + \sigma \circ \theta^n . \mathbf{e}_{N(\hat{V}_n)+1};$$

$$\mathcal{A}_i(\hat{V}_{n+1}) = \emptyset,\ i \in [\![1,\, S]\!],$$

which implies that

$$\hat{V}_{n+1} = \left[\hat{V}_n + \sigma \circ \theta^n . \mathbf{e}_{N(\hat{V}_n)+1} - (\xi \circ \theta^n) . \mathbf{1}\right]^+ .$$

Thus [4.65] is satisfied in this case.

(ii) If all servers are busy, C_n will be assigned to the server of index $\varphi_{\hat{V}_{n+}}(N(\hat{V}_{n+}) - S)$. In FCFS, the assignments of the customers to the various servers do not depend on future arrivals, hence any customer in line at the arrival of C_n remains at the same server after the arrival of C_n, after the following arrivals, and so on until his service (even if the server index may change through the successive arrivals). Hence, for any server i to which C_n will not be assigned, the set of indexes of those customers to be served by i just before the arrival of C_n, is the same as just after the arrival of C_n. In other words,

$$\Psi(\hat{V}_{n+})(i) = \Psi(\hat{V}_n)(i); \; i \neq \varphi_{\hat{V}_{n+}}(N(\hat{V}_{n+}) - S).$$

Then, for any i the set of indexes of customers to be served by a given server just before the arrival of C_{n+1} equals that just after the arrival of C_n, from which we remove the indexes of the customers entered in service between the two arrivals, during a time interval of duration $\xi \circ \theta^n$. On the other hand, the index of the server may possibly vary from i to ℓ thereby following the order of the residual service times of the customers in service upon the arrival of C_{n+1}. Therefore, we have

$$\forall i \neq \varphi_{\hat{V}_{n+}}(N(\hat{V}_{n+}) - S), \; \exists \ell \in [1, \, S]; \; \Psi(\hat{V}_{n+1})(\ell) = [\Psi(\hat{V}_n)(i) - \xi \circ \theta^n]^+ . \tag{4.66}$$

In addition, C_n is actually assigned to the server having, upon his arrival, the least virtual workload (made by the customers in line at the arrival of C_n), that is

$$\Psi(\hat{V}_{n+})(\varphi_{\hat{V}_{n+}}(N(\hat{V}_{n+}) - s)) = \overline{\Psi(\hat{V}_n)}(1) + \sigma \circ \theta^n,$$

and therefore

$$\exists k \in [\![1, \, S]\!] \text{ s. t. } \Psi(\hat{V}_{n+1})(k) = \left[\overline{\Psi(\hat{V}_n)}(1) + \sigma \circ \theta^n - \xi \circ \theta^n\right]^+ . \tag{4.67}$$

Clearly, [4.66] and [4.67] also imply [4.65] in this case.

□

The SRS of the virtual workloads of the servers thus satisfies Kiefer and Wolfowitz's equation. This amounts to saying that this system is equivalent to that of S parallel queues under the JSW policy : each queue corresponds at a given time to a given server, and to the customers he is about to serve.

We can now address the stability of the system. Denote $\hat{G}$, the random map: $\mathcal{S} \to \mathcal{S}$ defined for every $u \in \mathcal{S}$ by

$$\hat{G}(u) = \hat{G}^3 \circ \hat{G}^2\left(.,\xi\right) \circ \hat{G}^1\left(.,\sigma\right)\big(u\big) \text{ a.s..} \qquad [4.68]$$

A stationary service profile thus corresponds to a solution $\hat{V}$ to the equation

$$\hat{V} \circ \theta = \hat{G}(\hat{V}) \text{ a.s..} \qquad [4.69]$$

THEOREM 4.28.– *Equation [4.69] admits a $\mathcal{S}$-valued solution provided that [4.40] holds. Otherwise, there is no $\mathcal{S}$-valued solution.*

Proof. Let $(V_n,\, n \in \mathbf{N})$ be the sequence of the service profiles of the system of S parallel queues and Y_∞, be the minimal solution of [4.38]. Let $U \in \mathcal{S}$ be such that

$$\overline{\Psi(U)} = Y_\infty \text{ a.s..} \qquad [4.70]$$

We can then deduce from [4.37] and [4.65] that the SRS $\Big(\overline{\Psi(\hat{V}_n^U)},\, n \in \mathbf{N}\Big)$ and $\big(V_n^{Y_\infty},\, n \in \mathbf{N}\big) = (Y_\infty \circ \theta^n,\, n \in \mathbf{N})$ coincide a.s..

Let the event

$$\mathcal{E} = \{Y_\infty(1) = 0\}.$$

According to the previous remark, for any $n \in \mathbf{N}$, on the event $\theta^{-n}\mathcal{E}$ we have

$$\overline{\Psi(\hat{V}_n^U)}(1) = 0,$$

and therefore $\mathcal{A}_s(\hat{V}_n^U) = \emptyset$, which implies in turn that

$$\hat{V}_n^U(S+1) = \hat{V}_n^U(S+2) = \ldots = 0,$$

since for any $u \in \mathcal{S}$, $S \in \mathcal{A}_S(u)$ when $u(S+1) > 0$. Hence, on the event $\theta^{-n}\mathcal{E}$,

$$\begin{aligned}\hat{V}_{n+1}^U &= \left[\hat{V}_n^U + \sigma \circ \theta^n.\mathbf{e}_{N(\hat{V}_n^U)+1} - \xi \circ \theta^n\right]^+ \\ &= \Bigg(\underbrace{\left[\underline{\Psi(\hat{V}_n^U) + \sigma \circ \theta^n.\mathbf{e}_S - \xi \circ \theta^n}\right]^+}_{\text{first } S \text{ components}}, 0, 0, \ldots\Bigg) \\ &= \Bigg(\left[\underline{Y_\infty \circ \theta^n + \sigma \circ \theta^n.\mathbf{e}_1 - \xi \circ \theta^n}\right]^+, 0, 0, \ldots\Bigg).\end{aligned}$$

The sequence $(\theta^{-n}\mathcal{E},\, n \in \mathbf{N})$ is hence a sequence of renovating events of length 1 for $\big(S_n^U,\, n \in \mathbf{N}\big)$, for any initial condition U satisfying [4.70]. As a conclusion,

(i) If [4.40] is satisfied, in view of the fact that $\mathcal{E}$ is non-negligible according to Theorem 4.21, the Corollary 2.12 implies the existence of a solution $\hat{V}$ taking values in $\mathcal{S}$ for the equation [4.69]. In addition, it is easy to show that $\overline{\Psi(\hat{V})} = Y_\infty$ a.s..

(ii) If [4.40] is not verified, if [4.69] would admit a solution $\hat{V}$ with values in $\mathcal{S}$, a $(\mathbf{R}^+)^S$-valued solution V to the equation [4.38], would clearly be given by $V = \overline{\Psi(\hat{V})}$. This is an absurdity according to Theorem 4.20.

This completes the proof. □

4.5. Infinite servers queue

We now consider an ideal system where all the customers are served simultaneously at full velocity. In other words, there are an infinite number of servers, so that every customer is accepted for service upon arrival. We assume again that the input is of the G/G/ type, and keep the same notation as before. We hence note G/G/∞, such a system. It is easily seen that in this case, the workload sequence is not recursive, as the amount of work processed by the server between two successive arrivals depends on the number of customers in the system at any time between these two dates. We give hereafter the stability condition of this system, and a representation at equilibrium using the service profiles.

4.5.1. *The service profile*

As above, we work on the space $\mathcal{S}$ (see A.3). We denote S_n^∞, the service profile at T_n^- starting from a profile S_0^∞ originally. We arrange the profiles S_n^∞, $n \in \mathbf{N}$ arbitrarily in decreasing order. It is immediate to observe that the service profile sequence is recurrent on $\mathcal{S}$: for any $n \in \mathbf{N}$,

$$S_{n+1}^\infty = \left[\underline{F^1\left(S_n^\infty,\, \sigma \circ \theta^n\right)} - \xi \circ \theta^n.\mathbf{1}\right]^+,$$

where $F^1(., \sigma \circ \theta^n)$ is defined by [4.12].

By working on the Palm space of arrivals and services, the existence of a stationary service profile thus amounts to that of a $\mathcal{S}$-valued solution S^∞ to the equation

$$S^\infty \circ \theta = G^\infty\left(S^\infty\right),\ \mathbf{P}-\ \text{a.s.}, \qquad [4.71]$$

where $G^\infty(.) = \left[\underline{F^1(., \sigma)} - \xi.\mathbf{1}\right]^+$, $\mathbf{P}$-a.s.. The stability condition of the system is given below.

THEOREM 4.29.– *We assume that σ and ξ are integrable and that $\mathbf{P}(\xi > 0) > 0$. Provided that*

$$\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left(\sigma\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\le 0\right)>0, \qquad [4.72]$$

equation [4.71] admits a unique $\mathcal{S}$-valued solution S^∞. In addition, by denoting Z, the only solution of the equation

$$Z\circ\theta=[Z\vee\sigma-\xi]^+, \qquad [4.73]$$

then for every $\mu\in\mathcal{S}$ such that

$$\mu(1)\le Z, \qquad [4.74]$$

there is a strong backward coupling between the sequences $(S_n^{\infty,\mu},\, n\in\mathbf{N})$ and $(S^\infty\circ\theta^n,\, n\in\mathbf{N})$.

Proof. Starting with any service profile at the origin, for every $n\in\mathbf{N}$ the largest residual service time at T_{n+1}^- is the maximum between the largest term at T_n^- and the initial service time of C_n, minus the quantity of service provided to all customers between T_n and T_{n+1}. In other words,

$$S_{n+1}^\infty(1)=[S_n^\infty(1)\vee(\sigma\circ\theta^n)-\xi\circ\theta^n]^+\,.$$

Therefore, the existence of a largest residual service time amounts to that of a proper solution to equation [4.73]. According to Theorem 4.19, this equation admits a unique finite solution, given by

$$Z=\left[\sup_{j\in\mathbf{N}^+}\left(\sigma\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\right]^+\,. \qquad [4.75]$$

We then apply the arguments of the proof of Theorem 4.7. As $u(1)=0$ implies $u=\mathbf{0}$ for any $u\in\mathcal{S}$, denoting for all n, $\mathcal{B}_n=\{Z\circ\theta^n=0\}$, $(\mathcal{B}_n,\, n\in\mathbf{N})$ is a stationary sequence of renovating events of length 1 for any sequence $(S_n^{\infty,\mu},\, n\in\mathbf{N})$ initially equal to μ, where μ is a $\mathcal{S}$-valued random variable such that

$$\mu(1)\le Z, \text{a.s.}.$$

As [4.72] amounts to $\mathbf{P}(\mathcal{B}_0)>0$, Theorem 2.11 implies again the existence of a solution to [4.71]. The uniqueness of the $\mathcal{S}$-valued solution follows from the fact that $S(1)=Z$ for any solution S and therefore, that two solutions coincide and are $\mathcal{S}$-valued on the non-negligible event $\{Z=0\}$. Finally, the property of strong backward coupling follows from Borovkov's Theorem, as in Theorem 4.7. □

As before, it follows naturally:

COROLLARY 4.30.– *Provided [4.72] holds, there exists a unique stationary congestion X^∞, and any sequence of congestions starting from any μ satisfying [4.74], converges with strong back coupling towards X^∞.*

COROLLARY 4.31.– *Provided [4.72] holds, the G/G/∞ queue starting from any finite profile empties a.s. infinitely often.*

Proof. As the largest component $S^\infty(1)$ of the stationary profile satisfies [4.73], it is given explicitly by [4.75] and the assumption [4.72] implies that

$$\mathbf{P}\left(S^\infty = \mathbf{0}\right) = \mathbf{P}\left(S^\infty(1) = 0\right) > 0.$$

We conclude with an argument similar to that of Corollary 4.4. □

4.5.2. *The GI/GI/∞ queue*

Let us assume that the service and inter-arrival times form two independent and identically distributed sequences, that are independent of one another (this is thus a GI/GI/∞ system). In this particular case, the stability condition of the system can be rewritten more explicitly.

COROLLARY 4.32.– *In the case of a GI/GI/∞ queue, the conclusions of Theorem 4.29 and of Corollaries 4.30 and 4.31 remain valid, under the stability condition*

$$\mathbf{P}\left(\sigma \leq \xi\right) > 0. \qquad [4.76]$$

Proof. It suffices to check that [4.72] is equivalent to [4.76] in this case. Of course, [4.72] always implies [4.76] since

$$\mathbf{P}\left(\sigma \leq \xi\right) = \mathbf{P}\left(\sigma \circ \theta^{-1} - \xi \circ \theta^{-1} \leq 0\right)$$

and, a.s.,

$$\sup_{j \in \mathbf{N}^*} \left(\sigma \circ \theta^{-j} - \sum_{i=1}^{j} \xi \circ \theta^{-i}\right) \geq \sigma \circ \theta^{-1} - \xi \circ \theta^{-1}.$$

For the converse, we have a.s. for any $j \in \mathbf{N}^*$,

$$\sigma \circ \theta^{-j} - \sum_{i=1}^{j} \xi \circ \theta^{-i}$$
$$= \sigma \circ \theta^{-1} - \xi \circ \theta^{-1} + \sum_{i=1}^{j-1} \left(\left(\sigma \circ \theta^{-1} - \sigma\right) \circ \theta^{-i} - \left(\xi \circ \theta^{-1}\right) \circ \theta^{-i}\right),$$

we the usual convention $\sum_{i=1}^{0} \dots = 0$. Therefore, in virtue of the independence assumptions,

$$\begin{aligned}
&\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left(\sigma\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\leq 0\right)\\
&\geq \mathbf{P}\Bigg[\{\sigma\circ\theta^{-1}-\xi\circ\theta^{-1}\leq 0\}\\
&\qquad\bigcap\left\{\sup_{j\in\mathbf{N}^*}\sum_{i=1}^{j-1}((\sigma\circ\theta^{-1}-\sigma)\circ\theta^{-i}-\left(\xi\circ\theta^{-1}\right)\circ\theta^{-i})\leq 0\right\}\Bigg]\\
&=\mathbf{P}\left(\sigma-\xi\leq 0\right)\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\sum_{i=1}^{j}((\sigma\circ\theta^{-1}-\sigma)\circ\theta^{-i}-\left(\xi\circ\theta^{-1}\right)\circ\theta^{-i})\leq 0\right).
\end{aligned}$$

The last probability on the right-hand side is strictly positive by applying [4.6] to the random variables $\sigma\circ\theta^{-1}-\sigma$ and $\xi\circ\theta^{-1}$ and by observing that

$$\mathbf{E}\left[\sigma\circ\theta^{-1}-\sigma-\xi\circ\theta^{-1}\right]=\mathbf{E}\left[-\xi\right]<0$$

according to Lemma 2.2. So [4.76] implies [4.72] in this case. □

4.6. Queues with impatient customers

Let us consider a system with one server and of infinite capacity, in which the customers enter according to a G/G input (we keep the same notation as in section 4.1). It is further assumed that these customers are impatient: the customer C_n requires to be served before his patience D_n elapses, where we assume that the sequence $((\sigma_n,\xi_n,D_n);n\in\mathbf{Z})$ is stationary. This patience time therefore sets a "deadline" at time T_n+D_n, at which the customer leaves the system forever, if he has not been satisfied. We note such a system as G/G/1/1+G-X where the third G characterizes the sequence of the patience times, and where X denotes the service discipline.

We are primarily interested in two types of systems:

(i) The customer C_n requires us to *enter* the service booth before T_n+D_n. In such a case, we will consider that the customer will remain in the system until the end of his service, even if the latter occurs after T_n+D_n. Otherwise, the customer leaves the system forever at T_n+D_n, without reaching the server. The time is then said to be *up to the beginning of service*. We will then say that the queue is of type (b), as "Beginning", and we will note it as G/G/1/1+G(b)-X

(ii) The customer C_n requires us to have been fully *served* before T_n+D_n. Otherwise, he leaves the system forever at T_n+D_n, even if his service was initiated before that date. We then say that the patience time runs *until the end of service*, and that the queue is of type (e) as “end”. We will denote it as G/G/1/1+G(e)-X.

Throughout this section, we work on the canonical space $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$ of the sequence $((\sigma_n, \xi_n, D_n);\ n \in \mathbf{Z})$ or in other words, on the Palm space of arrivals, services, and patiences. The random variables σ, ξ, and D are then defined as in the previous sections, and we assume that they are all integrable, and that $\mathbf{P}\,(\xi > 0) > 0$.

4.6.1. *The profile of service and patience times*

In this section, we give an exhaustive representation of the system by an SRS keeping track of all residual service times and all residual patience times of the customers in the system. To simplify the presentation, we consider here only a G/G/1/1 + G(b) system with non-preemptive service discipline. It will appear, however, that similar representations may be proposed for queues (e) and/or multiple server queues. This is left to the reader.

As above, we denote X_n the number of customers in the system just before the arrival of customer C_n (at T_n^-) and for any $i < n$ such that C_i is in the system at T_n^-, we note $\varphi_n(i) \in [\![1,\ X_n]\!]$, the place of C_i in the queue in the order of priority, the first being occupied by the customer in service at T_n^-. For each such customer C_i, we denote :

– $R_n(\varphi_n(i))$ as the residual service time of C_i at T_n^- (already defined in section 4.1.3);

– $\tilde{R}_n(\varphi_n(i))$ as the *residual patience time* of C_i at T_n^-, i.e. the residual time at T_n before the end of the patience of C_i. In other words,

$$\tilde{R}_n\left(\varphi_n(i)\right) = T_i + D_i - T_n. \qquad [4.77]$$

For any $n \in \mathbf{N}$, we define $\nu_n \in \mathcal{S}^2$ (see the formal definition of $\mathcal{S}^2$, and additional notations in appendix A.3), the sequence representing the residual service and patience times of the customers in the system at this time, arranged in the reverse order of priorities. By convention, we set as 0, the residual patience time of the customer in service (this customer will not be removed anyway), and as $(0, 0)$, the other components of ν_n. In other words,

$$\nu_n(i) = \begin{cases} \Big(R_n\left(X_n + 1 - i\right), \tilde{R}_n\left(X_n + 1 - i\right)\Big) & \text{for } i < X_n; \\ \Big(R_n\left(X_n + 1 - i\right), 0\Big) & \text{for } i = X_n; \\ (0,\ 0) & \text{for } i > X_n. \end{cases}$$

We call ν_n, the *service and patience profile* at T_n^-.

Let us make precise the dynamics of the profile process within the Palm space of arrivals, services and patiences. Assume that the customer C_n finds upon arrival a profile ν_n. First, the service time and the patience of C_n are inserted in the profile arbitrarily in the first position, i.e.

$$\nu_{n+} = \Big\{(\sigma \circ \theta^n, D \circ \theta^n), \nu_n(1), \nu_n(2), \dots \Big\} =: H^1\left(\nu_n, \sigma \circ \theta^n\right). \qquad [4.78]$$

Then, as in the case of the service profiles of a G/G/1 queue, we apply to S_{n+} the map $H^{2,\Phi} : \mathcal{S}^2 \to \mathcal{S}^2$, to rearrange the components ν_{n+} following the order of priorities for the discipline Φ:

1) If Φ depends only on the arrivals and service times of the customers (i.e. $\Phi =$ LIFO SRPT, etc. . . .), $H^{2,\Phi}$ is nothing but the "extension" of the application F^Φ defined in [4.13] to $\mathcal{S}^2$ in the sense that for any $u \in \mathcal{S}^2$,

$$H^\Phi(u)(i) = (F^\phi(u^1)(i), u^2(j)), \text{where } j \text{ is such that } F^\phi(u^1)(i) = u^1(j),$$

that is the second coordinate "follows" the first one, re-arranged following F^Φ.

2) The discipline Φ may also depend on the patience times of the customers:

- The *Earliest Deadline First* discipline (EDF) gives a non-preemptive priority, at the end of each service, to the customer in line whose residual patience is the shortest. Therefore,

$$H^{\text{EDF}}(u) = \Big\{u(2), u(3), \dots, u(i), u(1), u(i+1), \dots \Big\} \text{ if } u^2(i+1) \leq u^2(1) < u^2(i),$$

in other words the sequence of second coordinates of $H^{2,\text{EDF}}(u)$ is ordered in decreasing order.

- The *Latest Deadline First* discipline (LDF) gives non-preemptive priority, at the end of each service, to the customer in line having the largest residual patience. Therefore,

$$H^{\text{LDF}}(u) = \Big\{u(2), u(3), \dots, u(i), u(1), u(i+1), \dots \Big\} \text{ if } u^2(i+1) \geq u^2(1) > u^2(i),$$

i.e. the second coordinates of $H^{2,\text{LDF}}(u)$ are arranged in increasing order.

We denote as above, for any discipline Φ,

$$\nu_{n++} = H^{2,\Phi}\left(\nu_{n+}\right). \qquad [4.79]$$

Then, the customers are served successively in the order of priorities for a duration $\xi \circ \theta^n$. To describe this, we define for any $x \in \mathbf{R}^+$ and $u \in \mathcal{S}^2$, the sets of indexes $\mathcal{B}_x^j(u)$, $j \in [\![1, N(u) - 1]\!]$ in a similar way as the sets $\mathcal{A}_i^j(u)$ of section 4.4, as follows

$$\mathcal{B}_x^0(u) = \{0\},$$

and for all $j \in \mathbf{N}^*$,

$$\mathcal{B}_x^j(u) = \begin{cases} \mathcal{B}_x^{j-1}(u) \cup \{j\} & \text{if } \sum_{k \in \mathcal{B}_x^{j-1}(u)} u^1\left(N(u)-k\right) < x \wedge u^2\left(N(u)-j\right); \\ \mathcal{B}_x^{j-1}(u) & \text{otherwise.} \end{cases}$$

Finally, we denote

$$\mathcal{B}_x(u) = \mathcal{B}_x^{N(u)-1}(u) \text{ and } \psi_x(u) = \max \mathcal{B}_x(u).$$

Let us call $\tilde{C}_j, j \in [\![0,\ N(\nu_{n++})-1]\!]$ the customer in the system just after the arrival of C_n, and whose service time and residual patience are given by $\nu_{n++}(N(\nu_{n++})-j)$, so that $\tilde{C}_0$ is the customer in service, $\tilde{C}_1$ the following customer in the order of priority, etc. and $\tilde{C}_{N(\nu_{n++})-1}$ is the last customer in the order of priorities.

The set $\mathcal{B}^j_{\xi\circ\theta^n}(\nu_{n++})$ then contains all the indexes (up to j included) of those customers who make it to enter service before the arrival of C_{n+1}. Indeed, customer $\tilde{C}_j$ can enter service before the arrival of C_{n+1} if and only if the following two conditions are met:

(i) the time needed to serve the customers on priority with respect to this customer is less than $\xi \circ \theta^n$, that is

$$\sum_{k \in \mathcal{B}^{j-1}_{\xi\circ\theta^n}(\nu_{n++})} \nu^1_{n++}\left(N\left(\nu_{n++}\right)-k\right) < \xi \circ \theta^n;$$

(ii) his patience has not ended before the end of the services of the customers on priority with respect to him, that is

$$\sum_{k \in \mathcal{B}^{j-1}_{\xi\circ\theta^n}(\nu_{n++})} \nu^1_{n++}\left(N\left(\nu_{n++}\right)-k\right) < \nu^2_{n++}\left(N\left(\nu_{n++}\right)-j\right).$$

The integer $\psi_{\xi\circ\theta^n}(\nu_{n++})$ thus represents the largest index of a customer who came into service before the arrival of C_{n+1} (or the customer in service at the arrival of C_n if this service has not been completed before the arrival of C_{n+1}).

Just before the arrival of C_{n+1}, the components of ν_{n++} hence read as follows.

– First,

$$\nu_{n+++}(N\left(\nu_{n++}\right)-j) = (0,0),\ \ j \in [\![0, \psi_{\xi\circ\theta^n}\left(\nu_{n++}\right)-1]\!], \qquad [4.80]$$

as all the customers of corresponding indexes have either finished their service, or their patience has ended before the arrival of C_{n+1}.

– The components corresponding to the customer in service at the arrival of C_{n+1} are given by

$$\nu^1_{n+++}(N(\nu_{n++}) - \psi_{\xi\circ\theta^n}(\nu_{n++})) = \left[\sum_{k=0}^{\psi_{\xi\circ\theta^n}(\nu_{n++})} \nu^1_{n++}(N\left(\nu_{n++}\right) - k) - \xi \circ \theta^n\right]^+ ; \quad [4.81]$$

$$\nu^2_{n+++}(N\left(\nu_{n++}\right) - \psi_{\xi\circ\theta^n}\left(\nu_{n++}\right)) = \left[\nu^2_{n++}(N\left(\nu_{n++}\right) - \psi_{\xi\circ\theta^n}\left(\nu_{n++}\right)) - \xi \circ \theta^n\right]^+, \quad [4.82]$$

the last quantity being zero whenever the customer's patience ran out after the beginning of his service.

– Finally, for any $j \in [\![\psi_{\xi\circ\theta^n}(\nu_{n++}), N((\nu_{n++}) - 1]\!]$,

$$\nu^2_{n+++}(N\left(\nu_{n++}\right) - j) = \left[\nu^2_{n++}(N\left(\nu_{n++}\right) - j) - \xi \circ \theta^n\right]^+; \quad [4.83]$$

$$\nu^1_{n+++}(N\left(\nu_{n++}\right) - j) = \nu^1_{n++}(N\left(\nu_{n++}\right) - j)\, \mathbf{1}_{\left\{\nu^2_{n+++}(N(\nu_{n++})-j)>0\right\}}, \quad [4.84]$$

the possible customers having less priority than the one currently in service hence have their service time remained unchanged, and their residual patience reduced by the time elapsed between the two arrivals. They are eliminated (the corresponding coordinate is set at $(0, 0)$) if the latter quantity is negative.

Equations [4.80–4.84] thus define a mapping $H^1 \colon \mathcal{S}^2 \times \mathbf{R}^+ \to \mathcal{S}^2$ such that

$$\nu_{n+++} = H^3\left(\nu_{n++}, \xi \circ \theta^n\right). \quad [4.85]$$

Finally, ν_{n+1} is obtained by removing the components equal to $(0, 0)$ intercalated in between non-zero components (which correspond to the customers served or eliminated between the arrivals of C_n and C_{n+1}), keeping unchanged the order of the remaining components. We then write

$$\nu_{n+1} = H^4\left(\nu_{n+++}\right). \quad [4.86]$$

We have thus proven the following result.

THEOREM 4.33.– *The sequence* $(\nu_n,\ n \in \mathbf{N})$ *is recurrent for any admissible discipline* Φ*: for any initial value* $\nu_0 \in \mathcal{S}^2$*, for any* $n \in \mathbf{N}$*,*

$$\nu_{n+1} = H^\Phi \circ \theta^n\left(\nu_n\right),$$

where

$$H^\Phi = H^4 \circ H^3(.,\xi) \circ H^{2,\Phi} \circ H^1(.,\sigma),$$

defined by [4.78, 4.79, 4.85] and [4.86].

As usual, for a given admissible service discipline Φ, a stationary profile is a $\mathcal{S}^2$-valued random variable ν^Φ satisfying the equation

$$\nu^\Phi \circ \theta = H^\Phi\left(\nu^\Phi\right) \text{ a.s..} \qquad [4.87]$$

We have the following.

THEOREM 4.34.– *For any admissible discipline Φ, if*

$$\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left((\sigma+D)\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\leq 0\right)>0, \qquad [4.88]$$

the G/G/1/1+G(b)-Φ queue admit a unique stationary profile ν^Φ.

Proof. Existence. According to Theorem 4.19 applied to $\alpha \equiv \sigma + D$ and $\beta \equiv \xi$, provided that [4.88] holds there exists a unique positive random variable, solution of the equation

$$Y\circ\theta = \left[Y \vee (\sigma+D)-\xi\right]^+,$$

given by

$$Y^{\sigma+D,\xi} = \left[\sup_{j\in\mathbf{N}^*}\left((\sigma+D)\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\right]^+.$$

Let us define for any $u \in \mathcal{S}^2$,

$$Z(u) = \sup_{i\in\mathbf{N}^*}\left(u^1(i)+u^2(i)\right), \qquad [4.89]$$

as the largest sum of the two coordinates of a component of u. Let χ be a $\mathcal{S}^2$-valued random variable such that

$$Z(\chi) \in \mathcal{Z} := \left\{\text{positive random variables } Z \text{ s.t. } Z \leq Y^{\sigma+D,\xi} \text{ a.s.}\right\}. \qquad [4.90]$$

Let $(\nu_n^\chi,\, n \in \mathbf{N})$ be the sequence of profiles initially equal to χ for a fixed service discipline Φ. As we noticed during the computation of the recurrence function of $(\nu_n^\chi,\, n \in \mathbf{N})$, for any $n \in \mathbf{N}$ and $i \in \mathbf{N}^*$, there exists a $j \in \mathbf{N}^*$ such that

$$\nu_{n+1}^{\chi,1}(i)+\nu_{n+1}^{\chi,2}(i) \leq \left[\nu_{n+}^{\chi,1}(j)+\nu_{n+}^{\chi,2}(j)-\xi\circ\theta^n\right]^+, \qquad [4.91]$$

since between the arrivals of C_n and C_{n+1}, the residual patience of any customer decreases from $\xi \circ \theta^n$ and the residual service time is at the most constant (the index of the customer in the queue may change from i to j between ν_{n+} and ν_{n+1}, based on priorities). But according to [4.78],

$$\left\{\nu_{n+}^{\chi,1}(j) + \nu_{n+}^{\chi,2}(j);\ j \in \mathbf{N}^*\right\} = \left\{\nu_{n}^{\chi,1}(j) + \nu_{n}^{\chi,2}(j);\ j \in \mathbf{N}^*\right\} \cup \{\sigma \circ \theta^n + D \circ \theta^n\},$$

which, taking the supremum, implies with [4.91] that

$$Z\left(\nu_{n+1}^{\chi}\right) \leq \left[Z\left(\nu_n^{\chi,\cdot}\right) \vee (\sigma \circ \theta^n + D \circ \theta^n) - \xi \circ \theta^n\right]^+ \text{ a.s..} \qquad [4.92]$$

Therefore, for any $n \in \mathbf{N}$ such that $Z(\nu_n^{\chi}) \leq Y^{\sigma+D,\xi} \circ \theta^n$, we obtain by monotonicity with [4.92] that

$$\begin{aligned} Z\left(\nu_{n+1}^{\chi}\right) &\leq \left[\left(Y^{\sigma+D,\xi} \circ \theta^n\right) \vee (\sigma \circ \theta^n + D \circ \theta^n) - \xi \circ \theta^n\right]^+ \\ &= Y^{\sigma+D,\xi} \circ \theta^{n+1}, \end{aligned}$$

by definition of $Y^{\sigma+D,\xi}$. We have thus shown by induction with [4.90] that

$$Z\left(\nu_n^{\chi}\right) \leq Y^{\sigma+D,\xi} \circ \theta^n \text{ a.s. for any } n \in \mathbf{N}. \qquad [4.93]$$

Now, set the event

$$\mathcal{A} = \left\{Y^{\sigma+D,\xi} = 0\right\}. \qquad [4.94]$$

According to [4.93], on $\theta^{-n}\mathcal{A}$ we have $Z(\nu_n^{\chi}) = 0$ and therefore $\nu_n^{\chi} = \mathbf{0}$, the null sequence of $\mathcal{S}^2$. Therefore, $(\theta^{-n}\mathcal{A},\ n \in \mathbf{N})$ is a sequence of renovating events of length 1 for any initial condition $\chi \in \mathcal{Z}$ as defined by [4.90]. In view of the assumption [4.88], Corollary 2.12 implies the existence of a $\mathcal{S}^2$-valued solution for [4.87].

Uniqueness. Let ς be a solution of [4.87] for a fixed discipline Φ. According to [4.92], we have a.s.

$$Z\left(\varsigma \circ \theta\right) \leq \left[Z(\varsigma) \vee (\sigma + D) - \xi\right]^+. \qquad [4.95]$$

If we had $Z(\varsigma) > \sigma + D$ a.s. (which implies in particular that $Z(\varsigma) \circ \theta > 0$ a.s.), according to [4.95] we would have a.s. that

$$Z(\varsigma) \circ \theta \leq \left[Z(\varsigma) \vee (\sigma + D) - \xi\right]^+ = Z(\varsigma) \vee (\sigma + D) - \xi = Z(\varsigma) - \xi,$$

and therefore $\mathbf{E}\left[Z(\varsigma) \circ \theta - Z(\varsigma)\right] \leq -\mathbf{E}\left[\xi\right] < 0$, a contradiction to Lemma 2.2. We have therefore

$$\mathbf{P}\left(Z\left(\varsigma\right) \leq \sigma + D\right) > 0. \qquad [4.96]$$

According to [4.95], on the event $\{Z(\varsigma) \leq Y^{\sigma+D,\xi}\}$,

$$Z(\varsigma) \circ \theta \leq \left[Y^{\sigma+D,\xi} \vee (\sigma + D) - \xi\right]^+ = Y^{\sigma+D,\xi} \circ \theta.$$

Hence, $\{Z(\varsigma) \leq Y^{\sigma+D,\xi}\}$ is θ-contracting. On the other hand, on $\{Z(\varsigma) \leq \sigma + D\}$,

$$\begin{aligned}Z(\varsigma) \circ \theta \leq [Z(\varsigma) \vee (\sigma + D) - \xi]^+ &= [\sigma + D - \xi]^+ \\ &\leq \left[Y^{\sigma+D,\xi} \vee (D + \sigma) - \xi\right]^+ = Y^{\sigma+D,\xi} \circ \theta,\end{aligned}$$

and [4.96] implies that $\mathbf{P}\left(Z(\varsigma) \leq Y^{\sigma+D,\xi}\right) > 0$, so

$$Z(\varsigma) \leq Y^{\sigma+D,\xi} \text{ a.s..}$$

In other words, for any solution ς, $Z(\varsigma)$ belongs to the set $\mathcal{Z}$, and therefore $(\nu_n^\varsigma, n \in \mathbf{N}) = (\varsigma \circ \theta^n, n \in \mathbf{N})$ admits $(\theta^{-n}\mathcal{A}, n \in \mathbf{N})$ as a sequence of renovating events of length 1. According to Corollary 2.13, $\mathbf{P}(\mathcal{A}) > 0$ implies the uniqueness of the solution ν^Φ. □

As the queue is empty at the arrival of C_n if and only if $\nu_n = \mathbf{0}$, and as this is true if and only if $Z(\nu_n) = 0$, we obtain as usual the following result by domination.

COROLLARY 4.35.– *For any admissible discipline Φ, the G/G/1/1+ G (b)-Φ queue a.s. empties infinitely often if [4.88] holds.*

As above, we can deduce the following result from Theorem 4.33.

COROLLARY 4.36.– *There exists a unique stationary congestion X^Φ and a unique stationary workload W^Φ provided [4.88] holds.*

4.6.2. *GI/GI/1/1+GI queue*

We can apply the same arguments as in section 4.5.2 to give the stability condition of the system GI/GI/1/1 + GI(b):

COROLLARY 4.37.– *For any admissible discipline Φ, the queue GI/GI/1/1 + GI (b)-Φ is stable, and the conclusions of Theorem 4.33 and Corollaries 4.34 and 4.35 remain valid, under the condition*

$$\mathbf{P}(\sigma + D \leq \xi) > 0. \qquad [4.97]$$

4.6.3. *Optimality of EDF*

Consider the special case of a G/M/1/1+G(b) queue: the service times of the customers $(\sigma_n,\ n \in \mathbf{Z})$ are assumed to be independent and identically distributed with an exponential distribution, and are independent of the inter-arrival times $(\xi_n,\ n \in \mathbf{Z})$ and of the patience times $(D_n, n \in \mathbf{Z})$. We show in this case that the EDF discipline is optimal, as it is the one that loses the least amount of customers, in the sense that we are going to specify.

Let $x \in \mathbf{R}^+$ and $u,\ v \in \mathcal{S}^2$. We shall modify the two sequences u and v to obtain two other ones $\hat{u}$ and $\hat{v}$, as follows:

(i) We set their first non-zero first component arbitrarily equal to the first non-zero component of v (we could have taken u either way), i.e.

$$\hat{u}^1\left(N(u)\right) = \hat{v}^1\left(N(v)\right) = v^1\left(N(v)\right).$$

(ii) the second components of $\hat{u}$ are those of u, and second components of $\hat{v}$ are those of v.

(iii) Let

$$\ell_1 = \min\left\{j \geq 0;\ j \in \mathcal{B}_x^j(u) \text{ or } j \in \mathcal{B}_x^j(v)\right\},$$

be the first index belonging to $\mathcal{B}_x(u)$ or to $\mathcal{B}_x(v)$. We then fix for any $j < \ell_1$, $\hat{u}^1(N(u)-j) = u^1(N(u)-j)$ and $\hat{v}^1(N(v)-j) = v^1(N(v)-j)$. Then, if $\ell_1 \in \mathcal{B}_x^j(u) = \mathcal{B}_x^j(\hat{u})$, denoting

$$k_1 = \min\left\{j \geq 0;\ j \in \mathcal{B}_x^j(v) = \mathcal{B}_x^j(\hat{v})\right\} \geq \ell_1$$

the first index of $\mathcal{B}_x(v)$, we set

$$\hat{v}^1\left(N(v) - k_1\right) = u^1\left(N(u) - \ell_1\right),$$

i.e. we make the first coordinates corresponding to the first indexes of $\mathcal{B}_x(u)$ and $\mathcal{B}_x(v)$ artificially equal.

(iv) Then, we denote

$$\ell_2 = \min\left\{j > \ell_1;\ j \in \mathcal{B}_x^j(\hat{u})\right\} \wedge \min\left\{j > k_1;\ j \in \mathcal{B}_x^j(\hat{v})\right\},$$

the second index belonging to $\mathcal{B}_x(u)$ or $\mathcal{B}_x(v)$, and we start again the same construction. If again the minimum is given by the left one, we leave the first intermediate components unchanged, and we set

$$\hat{v}^1\left(N(v) - k_2\right) = u^1\left(N(u) - \ell_2\right) = \hat{u}^1\left(N(u) - \ell_2\right),$$

where k_2 is the second index of $\mathcal{B}_x(\hat{v})$, and so on until $\psi_x(\hat{u}) \wedge \psi_x(\hat{v})$.

These notations are complicated, but they describe the simple idea that we can transform step by step, starting from the last positive components, the two sequences u and v to obtain two other sequences $\hat{u}$ and $\hat{v}$, whose second components are unchanged and whose first components corresponding to the indexes of $\mathcal{B}_x(\hat{u})$ and $\mathcal{B}_x(\hat{v})$, are equal up to $\psi_x(\hat{u}) \wedge \psi_x(\hat{v})$.

We refer the reader to the definition of "$\prec$" in the set $\mathcal{S}$ (Definition A.22). We have the following result.

LEMMA 4.38.– *Let u and v be two non-null elements of $\mathcal{S}^2$ such that*

$$u^2 \prec v^2 \text{ in } \mathcal{S}. \tag{4.98}$$

Then, for any $x \in \mathbf{R}^+$,

$$\left(H^3\left(\hat{u},\, x\right)\right)^2 \prec \left(H^3\left(\hat{v},\, x\right)\right)^2 \text{ in } \mathcal{S},$$

where H^3 is defined by [4.85].

Proof. Let us observe that by the very definition of the space $\mathcal{S}^2$, in view of [4.98] we necessarily have $N(\hat{u}) \leq N(\hat{v})$.

Then, we show by induction that there exists, for any $j \in [\![0,\, \psi_x(\hat{u})]\!]$, a bijection

$$F_j \colon \begin{cases} \mathcal{B}_x^j(\hat{u}) & \to \mathcal{B}_x^j(\hat{v}) \\ i & \mapsto F_j(i) \leq i. \end{cases}$$

This property is obvious for $j = 0$, and if it holds for the integer $j-1$ we have that Card $\mathcal{B}_x^{j-1}(\hat{u}) =$ Card $\mathcal{B}_x^{j-1}(\hat{v})$, thus according to the construction of $\hat{u}$ and $\hat{v}$,

$$\sum_{k \in \mathcal{B}_x^{j-1}(\hat{u})} \hat{u}^1\left(N(\hat{u}) - k\right) = \sum_{k \in \mathcal{B}_x^{j-1}(\hat{v})} \hat{v}^1\left(N(\hat{v}) - k\right). \tag{4.99}$$

In addition, $j \in \mathcal{B}_x^j(\hat{u})$ means that the term on the left-hand side is less than

$$x \wedge \hat{u}^2\left(N(\hat{u}) - j\right) \leq x \wedge \hat{v}^2\left(N(\hat{u}) - j\right),$$

in view of [4.98]. So is the case with the term on the right-hand side of [4.99], hence there exists an integer $\ell \leq j + N(\hat{u}) - N(\hat{v}) \leq j$ such that $\ell \in \mathcal{B}_x^\ell(\hat{v}) \subset \mathcal{B}_x^j(\hat{v})$. The induction is completed, by taking $F_j(i) = F_{j-1}(i)$ for all $i \in \mathcal{B}_x^{j-1}(\hat{u})$, and $F_j(j) = \ell$.

Therefore, $\psi_x(\hat{u}) \leq \psi_x(\hat{v})$ and there exists an injection

$$F \colon \begin{cases} \mathcal{B}_x(\hat{u}) & \to \mathcal{B}_x(\hat{v}) \\ i & \mapsto F(i) \leq i. \end{cases}$$

The result is finally a consequence of the fact that $\psi_x(\hat{u}) \leq \psi_x(\hat{v})$ and that for all $j > \psi_x(\hat{u})$, $j \in \mathcal{B}_x(\hat{v})$ implies that

$$\begin{aligned}\hat{v}^2\left(N(\hat{v})-j\right)\wedge x &> \sum_{k\in\mathcal{B}_x^{j-1}(\hat{v})} \hat{v}^1\left(N(\hat{v})-k\right)\\ &\geq \sum_{k\in\mathcal{B}_x(\hat{u})} \hat{u}^1\left(N(\hat{u})-k\right)\\ &\geq \hat{u}^2\left(N(\hat{u})-j\right)\wedge x,\end{aligned}$$

which implies in turn that

$$\hat{v}^2\left(N(\hat{v})-j\right) > \hat{u}^2\left(N(\hat{u})-j\right).$$

Therefore, any term of $H^3(\hat{u},\, x)$ that is different from $(0,0)$ has a second component smaller than the corresponding term of $H^3(\hat{v},\, x)$: these two terms are, respectively, the term of same index of $\hat{u}$ and of $\hat{v}$, left the same. □

We can now state the main result of this section. Let Φ be an admissible discipline that is independent of the service times of the customers. In the following, we add exponents $^{\Phi}$ and $^{\text{EDF}}$ to emphasize the dependence of the various parameters in the service discipline. Assume that the stability condition [4.88] holds. Then there exist, respectively under Φ and EDF, two stationary profiles ν^{Φ} and ν^{EDF}, and two stationary congestions $X^{\Phi} = N\left(\nu^{\Phi}\right)$ and $X^{\text{EDF}} = N\left(\nu^{\text{EDF}}\right)$. The optimality of EDF is stated in the following terms.

THEOREM 4.39.– *EDF maximizes stochastically the stationary congestion for G/M/1/1/G (b) queues: for any* $x \in \mathbf{R}$,

$$\mathbf{P}\left(X^{\text{EDF}} \geq x\right) \geq \mathbf{P}\left(X^{\Phi} \geq x\right).$$

Proof. Let us place ourselves on the event

$$\left\{\left(\nu^{\Phi}\right)^2 \prec \left(\nu^{\text{EDF}}\right)^2\right\} = \left\{\left(\hat{\nu}^{\Phi}\right)^2 \prec \left(\hat{\nu}^{\text{EDF}}\right)^2\right\}.$$

Then, readily

$$\left(H^1\left(\hat{\nu}^{\Phi},\sigma\right)\right)^2 \prec \left(H^1\left(\hat{\nu}^{\text{EDF}},\sigma\right)\right)^2.$$

It is then easy to check by the definition of $H^{2,\cdot}$, and the very definition of the EDF discipline, that

$$\left(H^{2,\Phi}\left(H^1\left(\hat{\nu}^{\Phi},\sigma\right)\right)\right)^2 \prec \left(H^{2,\text{EDF}}\left(H^1\left(\hat{\nu}^{\text{EDF}},\sigma\right)\right)\right)^2,$$

as EDF arranges the terms of the sequences in the decreasing order of their second coordinates. Therefore, Lemma 4.38 and the definition of the mapping H^4 allow us to conclude that

$$\left(H^{\Phi}\left(\hat{\nu}^{\Phi}\right)\right)^2 \prec \left(H^{\text{EDF}}\left(\hat{\nu}^{\text{EDF}}\right)\right)^2.$$

The event $\left\{\left(\hat{\nu}^{\Phi}\right)^2 \prec \left(\hat{\nu}^{\text{EDF}}\right)^2\right\}$ is thus θ-contracting. As it includes the non-negligible event $\mathcal{A}$ defined in [4.94], it is almost sure. In particular, the respective stationary congestion $N\left(\hat{\nu}^{\Phi}\right)$ and $N\left(\hat{\nu}^{\text{EDF}}\right)$ satisfy

$$N\left(\hat{\nu}^{\Phi}\right) \leq N\left(\hat{\nu}^{\text{EDF}}\right) \text{ a.s.}. \qquad [4.100]$$

For both Φ and EDF, the term $\nu^2\left(N(\nu)\right)$ reads as the residual service time of a customer in service. Thus, in view of Theorem 7.3, $\nu^2\left(N(\nu)\right)$ is exponentially distributed with the same parameter as that of the initial service times, say μ, and independent of the service time already completed for the customer in service. Consequently, in view of the independence assumption, the non-zero terms of ν^2 form nothing but a family $\mathcal{F}$ of $N(\nu)$ random variables that are i.i.d. with distribution $\varepsilon(\mu)$.

Furthermore, for any n the number $N(\nu_n)$ of customers in the system upon the arrival of C_n only depends on the arrival times and patience times of the customers, on the service times of the already departed customers and on the service time already completed for the customer in service at this time. Again, in view of the independence assumption, this implies that all r.v.'s of the family $\mathcal{F}$ are independent of $N(\nu)$.

For a moment, write $\nu(\mathcal{F})$ to emphasize the dependence of ν on the family $\mathcal{F}$. Then, we have that

$$\hat{\nu}(\mathcal{F}) = \nu(\hat{\mathcal{F}}),$$

where the family $\hat{\mathcal{F}}$ is obtained from $\mathcal{F}$ in the following way : if we write

$$\mathcal{F} = \left(U(1), U(2), ..., U\left(N(\nu)\right)\right),$$

where the U_i's are i.i.d. of distribution $\varepsilon(\mu)$, then

$$\hat{\mathcal{F}} = \Psi\left(U(\gamma(1)), ..., U\left(\gamma\left(N(\nu)\right)\right)\right),$$

where :

– γ is a random permutation of $\left\{1, ..., N(\nu^{\Phi})\right\}$ independent of $\mathcal{F}$, corresponding to an exchange of the service times of the customers waiting in line (in order to match the upcoming service times of the customers in both disciplines);

– for all family $\mathcal{G}$ of $N(\nu^{\Phi})$ random variables, $\Psi(\mathcal{G})$ is the family obtained by substituting a random variable Y of distribution $\varepsilon(\mu)$ to the I-th component $\mathcal{G}(I)$ of $\mathcal{G}$, where Y and I are independent of the $\mathcal{G}(i)$'s. Here, this component corresponds to the residual service time of the customer in service, set to Y under both disciplines.

It is then a simple consequence of the interchange argument [4.26] that under both Φ and EDF,

$$\nu(\hat{\mathcal{F}}) = \nu(\mathcal{F}) \text{ in distribution},$$

and thus that $\hat{\nu}(\mathcal{F})$ has the same distribution as $\nu(\mathcal{F})$.

Finally, this with [4.100] implies that for all $x \in \mathbf{R}$,

$$\mathbf{P}\left(N\left(\nu^{\Phi}\right) \geq x\right) \leq \mathbf{P}\left(N\left(\nu^{\mathrm{EDF}}\right) \geq x\right).$$

□

This optimality property is crucial, in that it implies that the loss under EDF is less than that of any other discipline independent of the service times.

4.6.4. *FIFO queues*

Let us consider the particular case where the server processes the requests in First In, First Out. As we shall see, in this case the system can be described by the workload sequence, as in a classical $G/G/1$ queue.

4.6.4.1. *Single server* (b) *queue*

Let us first assume that the system is G/G/1/1 + G (b)-FIFO: the patience times run only until the beginning of the service, thus the customer C_n is served until the end of his service, without interruption, once he reached the server before his deadline $T_n + D_n$. As in section 4.1, let us denote for any n, W_n as the amount of work (measured in unit time) submitted to the server just before the arrival of C_n. In FIFO, this workload thus represents the proposed waiting time to C_n before reaching the server.

We aim to establish the dynamics of the sequence $(W_n,\, n \in \mathbf{N})$ starting from a given initial state upon the arrival of C_0. At the arrival of C_n, we are in the following alternative:

(i) If the patience of C_n is greater than its proposed waiting time (i.e. $D \circ \theta^n \geq W_n$), the customer C_n will reach the server, and thus brings a contribution of $\sigma \circ \theta^n$ to the workload.

(ii) Otherwise, $D \circ \theta^n < W_n$ and C_n does not contribute to the workload as he will never reach the server, even if he stays in line for a duration $D \circ \theta^n$.

Consequently, starting from an arbitrary workload W_0 just before the arrival of C_0, we have for any $n \in \mathbf{N}$,

$$W_{n+1} = \left[W_n + (\sigma \circ \theta^n)\, \mathbf{1}_{\{[0,\, D\circ\theta^n]\}}(W_n) - \xi \circ \theta^n\right]^+. \qquad [4.101]$$

The sequence $(W_n,\, n \in \mathbf{N})$ is hence an SRS driven by the non-monotonic random map

$$\varphi\colon x \mapsto \left[x + \sigma\, \mathbf{1}_{\{[0,\, D]\}}(x) - \xi\right]^+ \ \text{ a.s.},$$

and a stationary workload W^{FIFO} solves the equation

$$W^{\text{FIFO}} \circ \theta = \varphi\left(W^{\text{FIFO}}\right) \ \text{ a.s..} \qquad [4.102]$$

THEOREM 4.40.– *Under condition [4.88], [4.102] admits a unique finite solution* W^{FIFO}*, satisfying*

$$Y^{\sigma \wedge D,\, \xi} \leq W^{\text{FIFO}} \leq Y^{\sigma + D,\, \xi}, \ \textit{a.s.}, \qquad [4.103]$$

where $Y^{\sigma+D,\,\xi}$ *and* $Y^{\sigma\wedge D,\,\xi}$ *are defined as in [4.35]. Moreover, for any random variable* Z *such that* $Z \leq Y^{\sigma+D,\,\xi}$ *a.s., there is strong backwards coupling for* $\left(W_n^Z,\, n \in \mathbf{N}\right)$ *and* $\left(W^{\text{FIFO}} \circ \theta^n,\, n \in \mathbf{N}\right)$.

Proof. It suffices to notice that, under FIFO, the system workload can be obtained readily from the service and patience profile. For doing so, define for all $u \in \mathcal{S}^2$,

$$\mathcal{B}^0(u) = \{0\},$$

for all $j \in \mathbf{N}^*$,

$$\mathcal{B}^j(u) = \begin{cases} \mathcal{B}^{j-1}(u) \cup \{j\} & \text{if } \displaystyle\sum_{k \in \mathcal{B}_\infty^{j-1}(u)} u^1\left(N(u) - k\right) < u^2\left(N(u) - j\right); \\ \mathcal{B}^{j-1}(u) & \text{otherwise,} \end{cases}$$

and finally

$$\mathcal{B}(u) = \mathcal{B}^{N(u)-1}(u).$$

In other words, the set $\mathcal{B}(u)$ can be seen as the limit of $\mathcal{B}_x(u)$ as x goes to infinity.

In FIFO, the set of indexes of the customers who will be served among those already in the system is explicitly known at any time, and will never change except for adding the indexes of new accepted customers. Indeed, no future customer will ever have a higher priority than those already in the system. As a consequence of the construction of the sets $\mathcal{B}_x(u)$, at equilibrium the latter set of indexes is precisely given by $\mathcal{B}\left(\nu^{\text{FIFO}}\right)$. Therefore, the workload of the system is nothing but the sum of the service times requested by the latter customers, or in other words the sum along $\mathcal{B}\left(\nu^{\text{FIFO}}\right)$ of the first coordinates of the terms of ν^{FIFO}, i.e.

$$W^{\text{FIFO}} = \sum_{i \in \mathcal{B}(\nu^{\text{FIFO}})} \left(\nu^{\text{FIFO}}\right)^1(i).$$

Consequently, the existence and uniqueness of W^{FIFO} are straightforward consequences of that of ν^{FIFO}, i.e. they follow from Theorem 4.34.

Now, first notice that if we had $W^{\text{FIFO}} > D$ a.s. (which implies in particular that $W^{\text{FIFO}} \circ \theta > 0$ a.s.), we would have

$$W^{\text{FIFO}} \circ \theta = W^{\text{FIFO}} - \xi \text{ a.s.,}$$

a contradiction to Lemma 2.2. Hence, we have

$$\mathbf{P}\left(W^{\text{FIFO}} \leq D\right) > 0. \qquad [4.104]$$

On another hand, recall the definition of the map φ of [4.102], and remark that for any $x \in \mathbf{R}^+$, a.s.

$$\begin{aligned} \varphi(x) &= \left[x + \sigma\, \mathbf{1}_{\{(-\infty,\, D]\}}(x) + x\, \mathbf{1}_{\{(D,\, D+\sigma]\}}(x) + x\, \mathbf{1}_{\{(D+\sigma,\, \infty)\}}(x) - \xi\right]^+ \\ &\leq \left[(D+\sigma)\, \mathbf{1}_{\{(-\infty,D]\}}(x) + (D+\sigma)\, \mathbf{1}_{\{(D,D+\sigma]\}}(x) + x\, \mathbf{1}_{\{(D+\sigma,\infty)\}}(x) - \xi\right]^+ \\ &= [x \vee (\sigma + D) - \xi]^+ \\ &= F^{\sigma+D,\,\xi}(x), \end{aligned} \qquad [4.105]$$

where $F^{\sigma+D,\xi}$ is defined as in [4.33]. As the latter mapping is a.s. increasing, for all $x \leq y$,

$$\varphi(x) \leq F^{D+\sigma,\,\xi}(y) \text{ a.s..} \qquad [4.106]$$

Consequently, on the event $\{W^{\text{FIFO}} \leq Y^{\sigma+D,\,\xi}\}$,

$$W^{\text{FIFO}} \circ \theta = \varphi(W) \leq F^{\sigma+D,\,\xi}\left(Y^{\sigma+D,\,\xi}\right) = Y^{\sigma+D,\,\xi} \circ \theta.$$

Hence, $\{W^{\text{FIFO}} \leq Y^{\sigma+D,\,\xi}\}$ is θ-contracting. But on $\{W^{\text{FIFO}} \leq D\}$,

$$\begin{aligned} W^{\text{FIFO}} \circ \theta = \left[W^{\text{FIFO}} + \sigma - \xi\right]^+ &\leq [D + \sigma - \xi]^+ \\ &\leq \left[Y^{\sigma+D,\,\xi} \vee (D+\sigma) - \xi\right]^+ = Y^{\sigma+D,\,\xi} \circ \theta. \end{aligned}$$

So [4.104] implies that $\mathbf{P}\left(W^{\text{FIFO}} \leq Y^{\sigma+D,\,\xi}\right) > 0$ and in turn, the upper bound of [4.103]. As a consequence, the announced strong backwards coupling property follow as usual, using Renovating events.

Regarding the lower bound, apply the same argument after remarking that for all $x \in \mathbf{R}$, a.s.

$$\begin{aligned} F^{\sigma\wedge D,\xi}(x) &= \left[(\sigma \wedge D)\, \mathbf{1}_{\{(-\infty,D\wedge\sigma]\}}(x) + x\, \mathbf{1}_{\{(D\wedge\sigma,D]\}}(x) + x\, \mathbf{1}_{\{(D,\infty)\}}(x) - \xi\right]^+ \\ &\leq \left[(x+\sigma)\left(\mathbf{1}_{\{(-\infty,D\wedge\sigma]\}}(x) + \mathbf{1}_{\{(D\wedge\sigma,D]\}}(x)\right) + x\, \mathbf{1}_{\{(D,\infty)\}}(x) - \xi\right]^+ \\ &= \varphi(x). \end{aligned} \qquad [4.107]$$

□

NOTE.– Equation [4.102] can be solved explicitly using Renovating events theory and the upper-bound of [4.103], without using Theorem 4.34 (see the references at the end of the chapter). This alternative proof is left as an exercise.

The loss queue

Let us now turn to a G/G/1/1 queue: the system has a single server, and no waiting line. Consequently, the customers (entering according to a G/G input) are either immediately served if the system is empty upon arrival, or immediately lost if the server is busy when they arrive.

This classical system has been widely studied in the literature (see the references at the end of the chapter), and is often called *Loss queue*. It is easy to see that it is in fact a special case of G/G/1/1+G(b)-FIFO queue, where it is assumed that the patience times are identically zero. Then, by keeping the same notation as in the previous Section, the sequence of the workloads at the arrivals of the customers satisfies

$$W_{n+1} = \left[W_n + (\sigma \circ \theta^n)\, \mathbf{1}_{\{0\}}(W_n) - \xi \circ \theta^n\right]^+, \; n \in \mathbf{N}, \;\text{ a.s..}$$

On the Palm space of arrivals and services, a stationary workload W^0 thus satisfies the equation

$$W^0 \circ \theta = \left[W^0 + \sigma\, \mathbf{1}_{\{0\}}(W^0) - \xi\right]^+ \text{ a.s..} \qquad [4.108]$$

The following result follows directly from Theorem 4.40:

COROLLARY 4.41.– *If*

$$\mathbf{P}\left(\sup_{j \in \mathbf{N}^*}\left(\sigma \circ \theta^{-j} - \sum_{i=1}^{j} \xi \circ \theta^{-i}\right) \leq 0\right) > 0, \qquad [4.109]$$

[4.108] admits a unique finite solution W^0 such that $W^0 \leq Y^{\sigma,\xi}$, where $Y^{\sigma,\xi}$ is defined as in [4.35].

The queue with load limitation

Another classical example of loss system can be addressed within this framework: let us assume that the single server of the system accepts the arrival of the customers only if the workload at this time does not exceed a given threshold, denoted as $d > 0$. The server then serves the accepted customers, following any service discipline X (preemptive or not) until the end of their service. Under the current hypotheses, and by using the same notation, the workload sequence hence satisfies the dynamics

$$W_{n+1} = \left[W_n + (\sigma \circ \theta^n)\, \mathbf{1}_{\{[0,\, d]\}}(W_n) - \xi \circ \theta^n\right]^+, \; n \in \mathbf{N}, \;\text{ a.s..}$$

The sequence $(W_n,\, n \in \mathbf{N})$ for this system equals that of the system G/G/1/1+G(b)-FIFO with the same input and for $D = d$ a.s., even if the two systems are different in general as they do not serve the same customers if $X \neq$ FIFO. A stationary workload for this model is a finite solution of

$$W \circ \theta = \left[W + \sigma\, \mathbf{1}_{\{[0,\,d]\}}(W) - \xi\right]^+ \text{ a.s.,} \qquad [4.110]$$

and we thus have the following corollary.

COROLLARY 4.42.– *If*

$$\mathbf{P}\left(\sup_{j \in \mathbf{N}^*}\left(\sigma \circ \theta^{-j} - \sum_{i=1}^{j} \xi \circ \theta^{-i}\right) \leq -d\right) > 0, \qquad [4.111]$$

[4.110] accepts a unique finite solution W^d such that

$$Y^{\sigma \wedge d,\, \xi} \leq W^d \leq Y^{\sigma + d,\, \xi} \textit{ a.s..}$$

Impatience until the end of service

We now consider a G/G/1/1+G(e)-FIFO queue. In this case, a customer can reach the end of his patience, and get lost, while he is in service. Such a customer then contributes to the workload of the server, by providing an amount of work corresponding only to the time he has spent in service before the end of his patience. More specifically, the workload I_n brought by the customer C_n upon arrival is given by

$$I_n = \begin{cases} \sigma \circ \theta^n, & \text{if } W_n \leq (D \circ \theta^n - \sigma \circ \theta^n)^+; \\ \sigma \circ \theta^n - (W_n + \sigma \circ \theta^n - D \circ \theta^n) = D \circ \theta^n - W_n, & \\ & \text{if } (D \circ \theta^n - \sigma \circ \theta^n)^+ < W_n \leq D \circ \theta^n; \\ 0, & \text{if } W_n > D \circ \theta^n. \end{cases}$$

In the first case, the patience of C_n lasts beyond the end of his service. In the second one, it ends while C_n is in service, and this customer remains in service for a period of time of duration $D \circ \theta^n - W_n$. Finally, in the third case C_n does not have time to reach the server before the end of his patience time. Hence, starting from a workload W_0 at the arrival of C_0, for any $n \in \mathbf{N}$,

$$W_{n+1} = \left[W_n + I_n - \xi_n\right]^+,$$

which can be rewritten in the following more compact form

$$W_{n+1} = \left[W_n + (\sigma \circ \theta^n - (W_n + \sigma \circ \theta^n - D \circ \theta^n)^+)^+ - \xi \circ \theta^n\right]^+. \qquad [4.112]$$

Consequently, a stationary workload on the Palm space of arrivals, services and patience times is a $\mathbf{R}_+$-valued random variable S that solves the equation

$$S \circ \theta = \psi(S) := \left[S + (\sigma - (S + \sigma - D)^+)^+ - \xi\right]^+. \qquad [4.113]$$

THEOREM 4.43.– *There exists an a.s. finite solution S to [4.113], such that*

$$S \leq Y^{D,\,\sigma} \text{ a.s.,} \qquad [4.114]$$

where $Y^{D,\,\sigma}$ is defined as in [4.35]. Moreover, this solution is unique if

$$\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left(D\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\leq 0\right)>0. \qquad [4.115]$$

Proof. First, it is easy to verify that the application ψ defined in [4.113] is a.s. increasing and continuous. Hence we can apply Loynes's Theorem, and obtain the minimal solution S to [4.113], which is the almost sure limit of the corresponding Loynes's sequence.

According to Theorem 4.19, the unique finite solution $Y^{D,\,\xi}$ of the equation

$$Y\circ\theta = F^{D,\,\xi}(Y)$$

is given by

$$Y^{D,\,\xi}=\left[\sup_{j\in\mathbf{N}^*}\left(D\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)\right]^+.$$

By noticing that for any $x\in\mathbf{R}^+$, a.s.

$$\begin{aligned}\psi(x)&=\big[((x+\sigma)\wedge D)\,\mathbf{1}_{\{[0,\,D]\}}(x)+x\,\mathbf{1}_{\{(D,\,\infty)\}}(x)-\xi\big]^+\\&\leq\big[(x\vee D)\wedge\big(x+\sigma\,\mathbf{1}_{\{[0,\,D]\}}(x)\big)-\xi\big]^+\\&=\varphi(x)\wedge F^{D,\,\xi}(x),\end{aligned} \qquad [4.116]$$

we clearly check that the event $\{S\leq Y_{D,\,\xi}\}$ is θ-contracting. On the other hand, we have $\mathbf{P}\,(S\leq D)>0$, since the contrary would imply that $S\circ\theta=S-\xi$ a.s., a contradiction to Lemma 2.2. But on the event $\{S\leq D\}$,

$$S\circ\theta=[((S+\sigma)\wedge D)-\xi]^+\leq[D\vee Y_{D,\,\xi}-\xi]^+=Y_{D,\,\xi}\circ\theta.$$

This implies that a.s.,

$$S\leq Y_{D,\,\xi}<\infty.$$

Finally, for any solution S' of [4.113], $\{S=S'\}$ is θ-invariant. In view of [4.114] and of the minimality of S, this event includes $\{Y_{D,\,\xi}=0\}$. It is therefore almost sure when [4.115] holds true. □

The proof of the following Lemma follows the same arguments as above, and is left to the reader.

LEMMA 4.44.– *Provided [4.115] holds,*

1) for any random variable Z such that $Z \leq Y^{D,\,\xi}$ a.s., there is a strong backwards coupling between the sequences $\left(W_n^Z,\, n \in \mathbf{N}\right)$ and $(S \circ \theta^n,\, n \in \mathbf{N})$;

2) moreover, if the condition [4.88] is satisfied, we have a.s.

$$S \leq W, \tag{4.117}$$

where W denotes the only solution of [4.102] and

$$S \geq Y^{\sigma\wedge D,\xi}, \tag{4.118}$$

where $Y^{\sigma\wedge D,\xi}$ is defined as in [4.35].

The loss probability at equilibrium is a crucial feature of the system's performance. It can be intuitively defined on the Palm space of arrivals, services, and patience times as the asymptotic proportion of lost customers in a stable system.

First, let us consider the G/G/1/G(b)-FIFO queue, and assume that the stability condition [4.88] is verified. For any $n \geq 0$, the customer C_n has a patience of $D \circ \theta^n$ and is offered a waiting time $W \circ \theta^n$, if C_0 had found a workload W upon arrival . Hence C_n is lost if and only if $W \circ \theta^n > D \circ \theta^n$. Consequently, if we denote $\pi(b)$ as the loss probability for this system and

$$B = \left\{(x,\, y) \in (\mathbf{R}^+)^2;\, x > y\right\},$$

we have a.s.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \mathbf{1}_{\{B\}} \left((W,\, D) \circ \theta^i\right) = \mathbf{E}\left[\mathbf{1}_{\{B\}}(W,\, D)\right] = \mathbf{P}\left(W > D\right). \tag{4.119}$$

According to [4.35] and [4.103], we therefore have that

$$\begin{aligned} &\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left((\sigma \wedge D)\circ\theta^{-j} - \sum_{i=1}^{j}\xi\circ\theta^{-i}\right) > D\right) \\ &\qquad \leq \pi(b) \leq \mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left((\sigma + D)\circ\theta^{-j} - \sum_{i=1}^{j}\xi\circ\theta^{-i}\right) > D\right). \end{aligned} \tag{4.120}$$

Let us now get back to the system G/G/1/G(e)-FIFO. The customer C_n is lost (i.e. his service cannot be completed) if the sojourn time $W_n + \sigma\circ\theta^n$ offered to him exceeds his patience $D \circ \theta^n$. The stationary loss probability $\pi(e)$ is given, for the same reason as before, by

$$\pi(e) = \mathbf{P}\left(S > D - \sigma\right).$$

According to [4.114] and [4.118], we therefore have that

$$\begin{aligned}&\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left((\sigma\wedge D)\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)>D-\sigma\right)\\&\leq\pi(e)\leq\mathbf{P}\left(\sup_{j\in\mathbf{N}^*}\left(D\circ\theta^{-j}-\sum_{i=1}^{j}\xi\circ\theta^{-i}\right)>D-\sigma\right).\end{aligned}\qquad[4.121]$$

Finally, the stationary probability $\hat{\pi}(e)$ that a customer cannot even reach the server is given by

$$\hat{\pi}(e)=\mathbf{P}\left(S>D\right).$$

According to [4.117], we therefore have that

$$\hat{\pi}(e)\leq\pi(b).$$

4.7. Notes and comments

Loynes's Theorem for the single server queue was first introduced in [LOY 62]. The approach we propose for the stability study, based on the study of stochastic recursive sequences, concurs in many ways with that introduced and developed in [BAC 02].

The proof of the optimality of SRPT, as presented here, can be found in [FLI 81]. The optimality of FIFO has been treated under different aspects by several authors. We present here a proof similar to that of [FOS 81]. The optimality of EDF for smooth deadlines has been shown in a similar framework in [MOY 08a]. The exchange argument is proven on p.267 of [BAC 02].

The representation of Processor Sharing queues and infinite servers queues by profiles sequences are due to [MOY 08b].

The construction of the stationary state of the system with S parallel queues is due to [NEV 84]. The optimal allocation results, as well as the comparison with the queue with S servers, are original.

The construction of the stationary workload of the FIFO queue with impatient customers follows the representation of [BAC 84]. It is explicitly obtained in [MOY 10], for impatient times until the beginning and until the end of service. The optimality of EDF for hard deadlines has been formulated in similar terms in [PAN 88.] The description by service and patience profiles, and the proof of the optimality of EDF in these settings, are original.

Epitome

– In a single server queue, the stability condition is $\rho < 1$, where ρ is the traffic load.

– In a queue with S servers, this condition becomes $\rho < S$.

– A system with S parallel queues, where each customer joins the shortest queue in terms of workload, is equivalent to a FIFO queue with S servers. It has hence the same stability condition.

– Packets of deterministic size minimize the average waiting time.

– The SRPT service discipline minimizes the waiting time among all admissible disciplines.

– In the independent case, the FIFO discipline minimizes the waiting time and sojourn time among all disciplines independent of the service times.

– The GI/GI/∞ system is stable provided $\mathbf{P}\left(\sigma \leq \xi\right)$.

– The GI/GI/1 queue with impatient customers is stable if $\mathbf{P}\left(\sigma + D \leq \xi\right)$, where D is the patience time.

– The EDF discipline minimizes the lateness in the case of smooth deadlines, and the system loss for hard deadlines.

Chapter 5

The M/GI/1 Queue

In the range of single server queues that can be studied with the stochastic tools introduced in this book, after the M/M/1 queue which will be discussed in Chapter 8, the following ones in term of generality are the GI/M/1 and M/GI/1 queues. In the latter, the inter-arrival times (respectively, service times) are independent and identically distributed, but not necessarily of exponential distribution.

Unfortunately, in both cases the process counting the number of customers in the system is no longer Markov. In fact, at a given time we cannot repeat the argument of example 7.1, as the exponential distribution is the only one that satisfies Theorems 6.6 and 7.3.

In order to circumvent this difficulty, the system is observed in discrete time, at instants suitably chosen. In the case of the M/GI/1 queue, these are the departure times of the customers. To keep this chapter as simple as possible, we only address the embedded Markov chain, which gives the most important results.

5.1. The number of customers in the queue

Let us recall the main notation concerning this queue. The arrivals form a Poisson process of intensity λ and the service times are independent and of the same distribution $\mathbf{P}_\sigma$, of mean $1/\mu$. We denote $\rho = \lambda/\mu$ the traffic load. For all $n \geq 1$, let σ_n be the service time of customer n and X_n be the number of customers in the system just after the departure of customer n. We have seen in the introduction that X_n satisfies the recurrence equation

$$\begin{cases} X_1 & = A_1; \\ X_{n+1} & = (X_n - 1)^+ + A_{n+1},\ n \geq 1, \end{cases}$$

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

where A_n is the number of customers arriving during the service of customer n.

THEOREM 5.1.– *$(A_n,\ n \geq 1)$ is a sequence of independent and identically distributed random variables, of distribution given by*

$$\mathbf{P}(A_n = k) = \int_0^\infty \exp(-\lambda t) \frac{(\lambda t)^k}{k!} \, \mathrm{d}\, \mathbf{P}_\sigma(t).$$

In particular,

$$\mathbf{E}\left[A_n\right] = \rho = \lambda/\mu.$$

Proof. Denote, as usual, T'_n as the departure time of customer n and X as the stochastic process in continuous time counting the number of customers in the system. The process X is adapted to the filtration

$$\mathcal{G}_t = \sigma(N(u),\ u \leq t) \vee \sigma(\sigma_n,\ n \geq 1).$$

As the Poisson process N is independent of the service times, it remains a Poisson process of intensity λ with respect to $\mathcal{G}$ and thus the process $(t \mapsto N(t) - \lambda t)$ is a $\mathcal{G}$-martingale. As we can write

$$T'_1 = \inf\{t > 0,\ \Delta X(t) = -1\} \text{ and } T'_{n+1} = \inf\{t > T'_n,\ \Delta X(t) = -1\},$$

the random variables $(T'_n,\ n \geq 1)$ are $\mathcal{G}$-stopping times that are all constructed in the same way

$$T'_{n+1} = T'_n + T'_1 \circ \theta^{T'_n},\ n \geq 1.$$

According to the strong Markov property,

$$(T'_1,\ T'_2 - T'_1,\ \cdots,\ T'_n - T'_{n-1})$$

is a sequence of independent and identically distributed random variables. Consequently, as

$$\begin{cases} A_1 & = N\left(T'_1\right); \\ A_{n+1} & = N\left(T'_{n+1}\right) - N\left(T'_n\right),\ n \geq 1, \end{cases}$$

the strong Markov property together with Theorem 6.7 entail that the A_n are independent and identically distributed, we thus focus on A_1.

By construction, A_1 is the number of arrivals during the first service time, which lasts by definition σ_1 units of time. To calculate A_1, we condition by the value of σ_1;

as N independent of σ_1, we have

$$\begin{aligned}\mathbf{P}(A_1 = k) &= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_k(A_n) \,|\, \sigma_1\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_k(N(\sigma_1)) \,|\, \sigma_1\right]\right] \\ &= \mathbf{E}\left[\mathbf{P}(N(\sigma_1) = t)\right] \\ &= \mathbf{E}\left[\exp(-\lambda\sigma_1)\frac{(\lambda\sigma_1)^k}{k!}\right] \\ &= \int_0^\infty \exp(-\lambda t)\frac{(\lambda t)^k}{k!}\,\mathrm{d}\,\mathbf{P}_\sigma(t).\end{aligned}$$

We then derive the mean expectation of A_n by its very definition. We have

$$\begin{aligned}\mathbf{E}\left[A_n\right] &= \sum_{k=0}^\infty k\mathbf{P}(A_n = k) \\ &= \int_0^\infty \exp(-\lambda t)\sum_{k=0}^\infty k\frac{(\lambda t)^k}{k!}\,\mathrm{d}\,\mathbf{P}_\sigma(t) \\ &= \lambda\int_0^\infty t\,\mathrm{d}\,\mathbf{P}_\sigma(t) \\ &= \rho,\end{aligned}$$

as the last integral equals the mean expectation of a service time, i.e. $1/\mu$. □

We already know from Theorem 4.2 that the system is stable in the sense that there exists a stationary workload and a stationary congestion if, and only if, the traffic load $\rho = \lambda/\mu$ is strictly less than 1. As X is a Markov chain, we can specify the mode of recurrence.

THEOREM 5.2.– *Let $a_k = \mathbf{P}(A_n = k)$ and $\rho = E[A_n]$. Assume that*

$$0 < a_0 \leq a_0 + a_1 < 1.$$

Then,

1) the chain X is transient if and only if $\rho > 1$;

2) X is recurrent if and only if $\rho = 1$;

3) X is positive recurrent if and only if $\rho < 1$.

Proof. It is obvious that X is irreducible, as it ranges in $\mathbf{N}$ by increments of 1 and of -1, as the condition on a_0 and a_1 ensures that there can be no arrival, or more than one arrival during a service time.

Let us assume that $\rho > 1$. Let $X_0 = l > 0$. As long as X has not decreased from l units, i.e. as long as n is less than τ_0^1 (the first hitting time of 0, see the notation of Chapter 3), the dynamics of X is given by

$$X_n = l + (A_1 - 1) + \ldots + (A_n - 1).$$

Let $\hat{X}_n = \sum_{j=1}^{n}(A_j - 1)$ and $\hat{\tau}_0 = \inf\{n > 0,\ \hat{X}_n = 0\} = \tau_l^1$. As $\rho > 1$, from the strong Law of Large Numbers,

$$\frac{\hat{X}_n}{n} \longrightarrow \rho - 1, \quad \mathbf{P}_i \text{ a.s. for any } i.$$

Therefore, for almost all sample ω, there exists $N(\omega)$ such that

$$n \geq N(\omega) \Longrightarrow 0 < (\rho - 1)/2 \leq \frac{\hat{X}_n}{n} \leq 3(\rho - 1)/2. \qquad [5.1]$$

Consequently, the hitting time of 0 occurs before $N(\omega)$.

Let us assume that there exists a value l_0 of l for which

$$0 < \alpha = \mathbf{P}(\omega : N(\omega) < l_0) < 1. \qquad [5.2]$$

As X can decrease at most by one per step, if $X_0 = l_0$ and $N(\omega) < l_0$ then

$$X_1(\omega) > 0, \ldots, X_{N(\omega)}(\omega) > 0$$

and from [5.1], $X_n(\omega) > 0$ for any $n > N(\omega)$. Consequently, starting from 0, as the chain is irreducible there is a non-zero probability to reach l_0. From this point, with probability α the chain will never return to 0, hence 0 is transient, which by irreducibility implies that the chain is transient.

It remains to show [5.2]. If for all $i \geq 0$, $N(\omega)$ is a.s. less than i, this means that $N(\omega)$ is a.s. null and therefore that $X_n \geq X_0$ for any value of n, which is contrary to the irreducibility assumption. If for all $i \geq 0$, $N(\omega)$ is a.s. greater than i, this means that $N(\omega)$ is a.s. infinite, which contradicts the convergence of $\hat{X}_n/n$. Hence [5.2].

Let us finally assume that $\rho \leq 1$. Take h as the identity function and $F = \{0\}$ in the Foster's criteria. As $X_n = f(X_0, A_1, \ldots, A_n)$ is independent of A_{n+1}, for $i > 0$ we have

$$\begin{aligned}\mathbf{E}\left[X_{n+1} \mid X_n = i\right] &= \mathbf{E}\left[i + A_{n+1} - 1 \mid X_n = i\right] \\ &= i + \mathbf{E}\left[A_{n+1}\right] - 1 \\ &= i + \rho - 1,\end{aligned}$$

so from Foster's criteria, the chain is recurrent null if $\rho = 1$ and positive recurrent if and only if $\rho < 1$. □

5.2. Pollacek-Khinchin formulas

The computation of the stationary probability in the case $\rho < 1$ is done by using a subtle but widely used tool, the Laplace transform or in probabilistic terms, the generating function. We already know that the stationary probability exists since from the Foster's criteria, the chain is positive recurrent. Let π be the stationary probability and X_∞ be a random variable of distribution π.

THEOREM 5.3.– *Let π be the invariant probability of a M/GI/1 queue of traffic load $\rho < 1$. We have the following identity, known as Pollacek-Khinchin formula*

$$\sum_k z^k \pi(k) = \frac{(z-1)\mathcal{Q}_A(z)}{z - \mathcal{Q}_A(z)}(1-\rho), \qquad [5.3]$$

where $\mathcal{Q}_A$ is the generating function of A_1, that is $\mathcal{Q}_A(z) = \mathbf{E}\left[z^{A_1}\right]$ for any z.

Proof. Set

$$\mathcal{Q}_{X_\infty}(z) = \mathbf{E}\left[z^{X_\infty}\right] = \sum_k z^k \pi(k), \text{ for } \mid z \mid \leq 1.$$

If $\rho < 1$, we know that X_n converges in distribution to X_∞, therefore

$$\lim_{n\to\infty} \mathbf{E}\left[z^{X_n}\right] = \mathcal{Q}_{X_\infty}(z).$$

But

$$\begin{aligned}\mathcal{Q}_{X_n}(z) &= \mathbf{E}\left[z^{X_{n+1}}\right] \\ &= \mathbf{E}\left[z^{A_{n+1}} z^{(X_n-1)^+}\right] \\ &= \mathbf{E}\left[z^{A_{n+1}}\right]\Big(\mathbf{E}\left[z^{X_n-1}\mathbf{1}_{\{X_n>0\}}\right] + \mathbf{P}(X_n=0)\Big)\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}\left[z^{X_n-1}\mathbf{1}_{\{X_n>0\}}\right] &= \frac{1}{z}\Big(\mathbf{E}\left[z^{X_n}\right] - \mathbf{E}\left[z^{X_n}\mathbf{1}_{\{X_n=0\}}\right]\Big) \\ &= \frac{1}{z}\mathbf{E}\left[z^{X_n}\right] - \frac{1}{z}\mathbf{P}(X_n=0).\end{aligned}$$

To the limit,

$$\mathcal{Q}_{X_\infty}(z) = \frac{(z-1)\mathcal{Q}_A(z)}{z - \mathcal{Q}_A(z)}\pi(0). \qquad [5.4]$$

In a neighborhood of 1,

$$\begin{aligned}\mathcal{Q}_A(z) &= 1 + (z-1)\mathbf{E}\left[A_1\right] + o(z-1) \\ &= 1 + \rho(z-1) + o(z-1)\end{aligned}$$

and

$$\mathcal{Q}_{X_\infty}(z) = 1 + O(z-1),$$

hence the term on the right-hand side of [5.4] must tend toward 1 and on the other hand, be equivalent to $(1-\rho)\pi(0)$. Therefore,

$$\pi(0) = 1 - \rho.$$

Finally, we conclude that

$$\mathcal{Q}_{X_\infty}(z) = \frac{(z-1)\mathcal{Q}_A(z)}{z - \mathcal{Q}_A(z)}(1-\rho).$$

As the function $\mathcal{Q}_{X_\infty}(z)$ is expandable in series, we have

$$\mathcal{Q}_{X_\infty}(z) = 1 + \sum_{k=1}^{\infty} \frac{\mathcal{Q}_{X_\infty}^{(k)}(0)}{k!} z^k, \qquad [5.5]$$

so it is sufficient to differentiate k times Q_{X_∞} at point 0 to obtain $\pi(k)$. □

We aim to recover $\mathcal{Q}_A$ from the service time distribution and the intensity of the arrival process. Let $\mathcal{L}_\sigma$ be the Laplace-Stieltjes transform of the service time distribution, i.e. for $z \geq 0$,

$$\mathcal{L}_\sigma(z) = \int_0^\infty e^{-zt} \, \mathrm{d}\,\mathbf{P}_\sigma(t).$$

LEMMA 5.4.– *The generating function of the number of arrivals during a service time in the M/GI/1 queue at equilibrium is given by*

$$\mathcal{Q}_A(z) = \mathcal{L}_\sigma(\lambda - \lambda z) \textit{ for all } z \in [0,1].$$

Proof. By definition, $\mathcal{Q}_A(z) = \mathbf{E}\left[z^{N_\sigma}\right]$, from which we deduce that

$$\begin{aligned}\mathcal{Q}_A(z) &= \mathbf{E}\left[\mathbf{E}\left[z^{N_\sigma} \mid \sigma\right]\right]\\ &= \int_0^\infty \mathbf{E}\left[z^{N_u} \mid \sigma = u\right] \mathrm{d}\,\mathbf{P}_\sigma(u)\\ &= \int_0^\infty \mathbf{E}\left[z^{N_u}\right] \mathrm{d}\,\mathbf{P}_\sigma(u)\\ &= \int_0^\infty \exp(-(1-z)\lambda u)\, \mathrm{d}\,\mathbf{P}_\sigma(u)\\ &= \mathcal{L}_\sigma(\lambda - \lambda z).\end{aligned}$$

□

LEMMA 5.5 (Second Pollacek-Khinchin Formula).– *If $\rho < 1$, the average number of customers in steady state is given by*

$$\mathbf{E}\left[X_\infty\right] = \rho + \frac{\rho^2(1+C_b^2)}{2(1-\rho)}, \quad \textit{where} \;\; C_b^2 = \frac{\mathit{Var}[\sigma]}{\mathbf{E}\left[\sigma\right]^2}, \qquad \text{[P-K 2]}$$

with σ a r.v. of distribution $\mathbf{P}_\sigma$.

Proof. Recall that $\mathcal{Q}'_A(1) = \mathbf{E}\left[X_\infty\right]$. In addition, by derivation in the integral we get

$$\frac{\mathrm{d}\,\mathcal{L}_\sigma}{\mathrm{d}\,s}(s) = -\int_0^\infty xe^{-sx}\, \mathrm{d}\,\mathbf{P}_\sigma(x),$$

from where it follows that

$$\frac{d\mathcal{Q}_A}{dz}(1) = \rho.$$

Differentiating the latter leads to [P-K 2]. □

NOTE.– In the M/M/1 case, we have $C_b = 1$ and thus obtain the result known from Chapter 8:

$$\mathbf{E}\left[X_\infty\right] = \rho(1-\rho)^{-1}.$$

NOTE (Optimality of determinism).– With [P-K 2], we verify the result known from Chapter 4: ρ being fixed, the average number of customers in the system at equilibrium is minimized for $\mathrm{Var}[\sigma] = 0$, that is for deterministic service times.

According to Little's Formula, we also minimize the average sojourn time at equilibrium, and this is the main reason why we will prefer, in packet networks, packets of fixed size.

EXAMPLE (M/ Γ/1 queue).– We denote so the queue where the service times follow a Γ distribution, that is to say for all x,

$$d\mathbf{P}_\sigma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, \mathbf{1}_{\mathbf{R}^+}(x)\,\mathrm{d}\,x,$$

where α and β are strictly positive. A quick computation shows that

$$\mathbf{E}\,[\sigma] = \alpha/\beta \text{ and } \mathrm{Var}[\sigma] = \alpha/\beta^2, \text{ thus } C_b^2 = 1/\alpha.$$

We see that when α becomes small and β is chosen in a way that the ratio α/β is constant, the traffic load is constant but the average number of customers goes to infinity. But as $\mathrm{Var}[\sigma]$ can also be written as $\mathbf{E}\,[\sigma]\,/\beta$, a small α induces a small β, hence a very high variance of service times. Given this result and the previous one on M/D/1 queue, we see that is advantageous to limit the fluctuations of the service times.

The distributions for which C_b^2 is greater than 1 are called *super-variants*. In the case of the Gamma distribution, this super-variance is due to the fact that the density tends to 0 at infinity slower than the exponential one, and the probability of having a long service time is much higher than for the M/M/1 queue. This phenomenon is quite common and crucial in the applications to networking, for it has been observed statistically that on the web, the files length follows a Pareto distribution, i.e. of the form

$$\mathrm{d}\,\mathbf{P}_\sigma(x) = x^{-\alpha}\,\mathbf{1}_{[K,+\infty)}(x)\,\mathrm{d}\,x,$$

where $K > 0$ represents the minimum length of the files and $\alpha > 0$ represents the decay rate at infinity. The situation is even more dramatic in this case than for the Gamma distribution, since if $\alpha < 2$, the variance of σ is infinite.

NOTE.– Using Little's Formula, we easily show that

$$\frac{\mathbf{E}[\mathrm{TA}]}{\mathbf{E}[X]} = \frac{\rho(1 + C_b^2)}{2(1-\rho)},$$

where a linear dependence appears between the average waiting time and the coefficient of variation for a fixed traffic load.

5.3. Sojourn time

To compute the stationary distribution of the sojourn time in the system, let us build on the following observation: given that the service policy is FIFO, the customer leaving the system leaves behind him all the customers who arrived during his stay in the system. Therefore, for all n the number of customers at the departure time coincides with the number of arrivals during the sojourn time, i.e.

$$N(\mathrm{Ts}_n) = X_n, \qquad [5.6]$$

where $\textsc{Ts}_n$ is the sojourn time of customer n. Similarly, as the sojourn time depends only on the customers who arrived before customer n, $\textsc{Ts}_n$ is independent of the number of arrivals that occur during the stay of customer n. Knowing the distribution of X_∞ at departure times by using the Pollaczek-Khinchin formula, we can deduce that of the sojourn time. Notice, that we can decompose $\textsc{Ts}_n$ as the sum of the waiting time $\textsc{Ta}_n$ in the queue plus the service time σ_n.

THEOREM 5.6.– *If $\rho < 1$, then the random variables* $\textsc{Ts}_n$ *and* $\textsc{Ta}_n$ *converge in distribution to the random variables* $\textsc{Ts}$ *and* $\textsc{Ta}$*, respectively. In addition, for any* $s \geq 0$,

$$\mathcal{L}_{\textsc{Ts}}(s) = \mathbf{E}\left[e^{-s\textsc{Ts}}\right] = \mathcal{L}_\sigma(s)\frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{L}_\sigma(s)};$$

$$\mathcal{L}_{\textsc{Ta}}(s) = \mathbf{E}\left[e^{-s\textsc{Ta}}\right] = \frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{L}_\sigma(s)}.$$

Proof. Using relation [5.6], we have for all n and z,

$$\begin{aligned}
\mathcal{Q}_{X_n}(z) &= \mathbf{E}\left[\mathbf{E}\left[z^{N_{[T_n, T_n+\textsc{Ts}_n[}} \,|\, \textsc{Ts}_n\right]\right]\\
&= \int_0^\infty \mathbf{E}\left[z^{N_{[T_n, T_n+u[}} \,|\, \textsc{Ts}_n = u\right] \mathrm{d}\,\mathbf{P}_{\textsc{Ts}_n}(u)\\
&= \int_0^\infty \mathbf{E}\left[z^{N_{[T_n, T_n+u[}}\right] \mathrm{d}\,\mathbf{P}_{\textsc{Ts}_n}(u)\\
&= \int_0^\infty \mathbf{E}\left[z^{N_u}\right] \mathrm{d}\,\mathbf{P}_{\textsc{Ts}_n}(u)\\
&= \int_0^\infty \exp(-(1-z)\lambda u)\,\mathrm{d}\,\mathbf{P}_{\textsc{Ts}_n}(u)\\
&= \mathcal{L}_{\textsc{Ts}_n}(\lambda - \lambda z).
\end{aligned}$$

As X_n converges in distribution to X_∞, $\mathcal{Q}_{X_n}$ converges pointwise to $\mathcal{Q}_{X_\infty}$, therefore $\mathcal{L}_{\textsc{Ts}_n}$ also admits a pointwise limit, which is equivalent to saying that $\textsc{Ts}_n$ converges in distribution to $\textsc{Ts}$. As $\textsc{Ta}_n = \textsc{Ts}_n - \sigma_n$, and as the distribution of σ_n is constant, we deduce from this that $\textsc{Ta}_n$ converges in distribution, to the r.v. $\textsc{Ta}$. In view of [5.3], we deduce from this that

$$\mathcal{L}_{\textsc{Ts}}(s) = \mathcal{L}_\sigma(s)\frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{L}_\sigma(s)}.$$

As $\textsc{Ta}_n$ depends only on what happened before the arrival of customer n, $\textsc{Ta}_n$ and σ_n are independent, hence

$$\mathbf{E}\left[e^{-s\textsc{Ts}_n}\right] = \mathbf{E}\left[e^{-s\textsc{Ta}_n}\right]\mathbf{E}\left[e^{-s\sigma_n}\right],$$

from which we deduce the desired formula for $\mathcal{L}_{\textsc{Ta}}$. □

5.4. Tail distribution of the waiting time

If we aim to size a buffer, we need to know *a priori* the loss probability for each possible value of the size. In other cases than that of the M/M/1 queue, these quantities are not analytically tractable. However, in the general M/GI/1 case, we can access the tail distribution of the virtual waiting time, that is to say $\mathbf{P}(\text{T}_{\text{A}} > x)$ for large x. From there, we deduce a dimensioning of the system by choosing x such that the probability of exceeding the threshold is less than the tolerated loss rate. By doing so, it is clearly plausible that we oversize (sometimes significantly) the buffer.

LEMMA 5.7.– *For any $s > 0$,*

$$\mathbf{E}\left[e^{-s\text{T}_{\text{A}}}\right] = 1 - \int_0^\infty \mathbf{P}(\text{T}_{\text{A}} > x) s e^{-sx} \, \mathrm{d}\, x.$$

Proof. From Fubini's Theorem, we have

$$\begin{aligned}
\int_0^\infty \mathbf{P}(\text{T}_{\text{A}} > x) s e^{-sx} \, \mathrm{d}\, x &= \int_0^\infty \left(\int_x^\infty \mathrm{d}\, \mathbf{P}_{\text{T}_{\text{A}}}(u) \right) s e^{-sx} \, \mathrm{d}\, x \\
&= \int_0^\infty \left(\int_0^u s e^{-sx} \, \mathrm{d}\, x \right) \mathrm{d}\, \mathbf{P}_{\text{T}_{\text{A}}}(u) \\
&= \int_0^\infty (1 - e^{-su}) \, \mathrm{d}\, \mathbf{P}_{\text{T}_{\text{A}}}(u) \\
&= 1 - \mathbf{E}\left[e^{-s\text{T}_{\text{A}}}\right].
\end{aligned}$$

□

In other words, denoting $\text{T}_{\text{A}}^c(x) = \mathbf{P}(\text{T}_{\text{A}} > x)$, we have

$$\mathcal{L}_{\text{T}_{\text{A}}^c}(s) = \frac{1}{s}(1 - \mathcal{L}_{\text{T}_{\text{A}}}(s)).$$

The asymptotic study of $\mathcal{L}_{\text{T}_{\text{A}}^c}$ thus allows us to study the behavior of the tail distribution of T_{A}.

EXAMPLE (M/PH/1 queue).– Let us consider that a proportion p of the requests can be processed locally in a time of exponential distribution of parameter μ_1, and that the other requests may need a remote processing, in an exponential time of parameter μ_2. We then say that the service times have a "phase"-type distribution, hence the name of the M/PH/1 queue. We then have

$$1/\mu = \mathbf{E}\left[\sigma_1\right] = p/\mu_1 + (1-p)/\mu_2 \text{ and } \text{Var}[G_1] = 2p/\mu_1^2 + 2(1-p)/\mu_2^2.$$

Similarly,

$$\mathcal{L}_{\sigma}(s) = \mathbf{E}\left[e^{-s\sigma_1}\right] = p\frac{\mu_1}{\mu_1 + s} + (1-p)\frac{\mu_2}{\mu_2 + s}.$$

After some algebra, we obtain

$$\mathcal{L}_{\text{TA}^c_\infty}(s) = \frac{(\rho\,\mu_2 + s\rho - 1 + \rho\,\mu_1)\,\lambda}{\mu_1\,\mu_2 + \mu_2\,s - \mu_2\,\lambda\,\mu_1\,\rho + s^2 - \lambda\,s + \mu_1\,s}.$$

The denominator vanishes in two real points

$$\alpha_+ = -\frac{1}{2}(\mu_1 + \mu_2 - \lambda + \sqrt{\Delta});$$
$$\alpha_- = -\frac{1}{2}(\mu_1 + \mu_2 - \lambda - \sqrt{\Delta}),$$

where

$$\Delta = (\mu_1 + \mu_2 - \lambda)^2 - 4\mu_1\mu_2(1-\rho).$$

We therefore have the simple elements factorization

$$\mathcal{L}_{\text{TA}^c_\infty}(s) = \frac{\rho}{2}\left(\frac{\mu_1 + \mu_2 - \lambda + \sqrt{\Delta}}{s - \alpha_-} + \frac{\mu_1 + \mu_2 - \lambda - \sqrt{\Delta}}{s - \alpha_+}\right).$$

NOTE.– It is not absolutely clear that if $\mu_1 + \mu_2 - \lambda$ is negative then α_+ is negative. In fact, by using the condition

$$\rho = \lambda(p/\mu_1 + (1-p)/\mu_2) < 1,$$

we show that $\mu_1 + \mu_2 - \lambda$ is necessarily positive.

Using the Laplace inversion formulas we obtain that

$$\mathbf{P}(\text{TA} > x) = \frac{\rho}{2}\left((\mu_1 + \mu_2 - \lambda + \sqrt{\Delta})e^{\alpha_- x} + (\mu_1 + \mu_2 - \lambda - \sqrt{\Delta})e^{\alpha_+ x}\right).$$

In fact, as x becomes large, only matters the term which has the slowest decrease, that is the one with the exponential containing α_-. Hence we have

$$\mathbf{P}(\text{TA} > x) \simeq \frac{\rho}{2}(\mu_1 + \mu_2 - \lambda + \sqrt{\Delta})e^{\alpha_\infty x}.$$

This result has to be compared to that for the M/M/1 queue,

$$\mathbf{P}(\text{TA} > x) \sim \rho e^{-\mu(1-\rho)x}.$$

Let us fix the unit of time so that $\lambda = 1$ and let us fix ρ, which amounts to fix the average service time $1/\mu$. In this case, to ensure a probability of exceeding the threshold x below ϵ, we have to take

$$x \geq -\frac{1}{\mu - 1} \ln(\mu\epsilon);$$

whereas in the case of the M/PH/1 queue,

$$x \geq \frac{1}{\alpha_-} \ln\left(\mu\epsilon . \frac{2}{\mu_1 + \mu_2 - 1 + \sqrt{\Delta}}\right).$$

Fixing ρ amounts to linking the three parameters μ_1, μ_2 and p but there are still two degrees of freedom which are chosen arbitrarily as μ_1 and μ_2. We see that the dimensioning of the M/PH/1 queue is determined not only by ρ, but also by the product and the sum of μ_1 and μ_2. In other words, the knowledge of the traffic load alone is not sufficient to size the buffer and to guarantee a given loss. The situation is extremely different from that of the M/M/1 queue, for which the knowledge of ρ alone is enough to calculate the threshold.

5.5. Busy periods

Always assuming that the queue is stable ($\rho < 1$), a *busy cycle* of the queue consists of an *idle period* I which ends with the arrival of a customer, followed by a *busy period* U that ends when the last customer departs, leaving behind an empty system.

Given the memoryless property of the exponential distribution, I follows the same distribution as that of an inter-arrival time, that is

$$\mathbf{P}\left(I \leq t\right) = 1 - e^{-\lambda t}.$$

To analyze the busy period U, we will take time $t = 0$ as time origin, which is the moment of arrival of the first customer in an empty system, or equivalently, the departure of a customer leaving behind a single customer in the system. We denote in the sequel, $(X(t),\ t \geq 0)$ the process in continuous time counting the number of customers in the system at any time.

DEFINITION 5.1.– *An* elementary busy period *is the random variable U defined as follows.*

$$U = \begin{cases} \inf\left\{t > 0,\ X_t = 0 \mid X_0 = 1\right\} & \textit{if } \rho < 1; \\ \infty & \textit{if } \rho \geq 1. \end{cases}$$

It is possible to generalize the definition of the busy period, if the initial time coincides with the departure of a customer leaving behind n other customers.

DEFINITION 5.2.– *A busy period of initial condition n is the random variable U_n defined as follows.*

$$U_n = \begin{cases} \inf\{t > 0,\, Y_t = 0 \mid Y_0 = n\} & \textit{if } \rho < 1 \\ \infty & \textit{if } \rho \geq 1. \end{cases}$$

Notice that $U_1 \equiv U$.

THEOREM 5.8.– *The probability distribution of U_n is the convolution product of order n of the distribution of U.*

Proof. The duration of the busy period is insensitive to the service policy, provided that it is conservative. The server serves all the customers entering the system, so it reads as the sum of the service times of all the customers arrived in the system during this duration.

To compute U_n, we thus factorize it as follows. The n customers present at the beginning of time will be called the *fathers*:

1) serve the lead customer (the first father);

2) serve all the customers who arrived during the service of the father (the sons);

3) serve all the customers who arrived during the service of the son (the grand sons);

4) repeat the previous steps until there are no more descendants to serve;

5) repeat steps 1,2,3 and 4 above for the second father, then the third father, . . . , until the nth father.

On the basis of the i.i.d. nature of the service times, and the memoryless property of the Poisson process, the duration of the time periods subsuming steps 1 to 4 for all fathers form a n sample of the distribution of U. So the distribution of the sum is the convolution of order n of that of U. □

NOTE.– We have chosen a suitable way of ordering the service times in order to easily expedite the proof of the last result. Compare this ordering to the concrete FIFO case: the server takes care in the order of arrivals, of the n present customers at time 0, then the 'sons' of the first customer, i.e. the customers entered during the service of the first father, in the order of arrivals, then the sons of the second father, and so on... so the genealogical tree is read from left to right, then from top to bottom.

THEOREM 5.9.– *For any $s \geq 0$,*

$$\mathcal{L}_U(s) = \mathbf{E}\left[e^{-sU}\right] = \mathcal{L}_\sigma\left[s + \lambda - \lambda\mathcal{L}_U(s)\right]. \qquad [5.7]$$

Proof. We factorize U in a similar manner that U_n, that is:

1) serve the father, during this service, V sons have arrived;

2) serve the first son followed by all his descendants;

3) repeat the previous step for the $V-1$ sons remaining.

Therefore,

$$U = \sigma_1 + \sum_{i=1}^{V} \Phi_i,$$

where each Φ_i represents the service duration of the son i and its descendants, and has the same distribution as U from the memoryless property of the exponential distribution. To compute the distribution of U, we first calculate the conditional distribution on V and σ_1, then we successively un-condition on V, then on σ_1. For any x and k,

$$\begin{aligned}\mathbf{E}[e^{-sU} \,|\, \sigma_1 = x, V = k] &= \mathbf{E}[e^{-s(x+\sum_{i=1}^{k} \Phi_i)}] \\ &= e^{-sx} \prod_{i=1}^{k} \mathbf{E}[e^{-s\Phi_i}] \\ &= e^{-sx}[\mathcal{L}_U(s)]^k.\end{aligned}$$

By un-conditioning on V,

$$\begin{aligned}\mathbf{E}[e^{-sU} \,|\, \sigma_1 = t] &= \sum_{k=0}^{\infty} \mathbf{E}[e^{-sU} \,|\, \sigma_1 = t, V = k] \mathbf{P}\left(V = k \,|\, \sigma_1 = x\right) \\ &= e^{-sx} \sum_{k=0}^{\infty} [\mathcal{L}_U(s)]^k \frac{(\lambda x)^k}{k!} e^{-\lambda x} \\ &= e^{-x[s+\lambda-\lambda\mathcal{L}_U(s)]}.\end{aligned}$$

Finally, by un-conditioning on σ_1,

$$\begin{aligned}\mathcal{L}_U(s) &= \int_0^{\infty} \mathbf{E}[e^{-sU} \,|\, \sigma_1 = x] \, \mathrm{d}\,\mathbf{P}_\sigma(x) \\ &= \int_0^{\infty} e^{-x[s+\lambda-\lambda\mathcal{L}_U(s)]} \, \mathrm{d}\,\mathbf{P}_\sigma(x).\end{aligned}$$

□

EXAMPLE (M/M/1 queue).– In this case, everything is easily calculable and we obtain the following.

THEOREM 5.10.– *For the M/M/1 queue,*

$$\mathcal{L}_U(s) = \frac{1}{2\lambda} \left[(\lambda + \mu + s) - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu} \right].$$

By inverting $\mathcal{L}_U(s)$ and by differentiating we obtain the density of the elementary busy period:

$$g(t) = \frac{d}{\mathrm{d}\,t}G(t) = \frac{1}{\sqrt{\rho}t}e^{-(\lambda+\mu)t}I_1\left[2t\sqrt{\lambda\mu}\right],$$

where $I_k(t)$ is the modified Bessel function of the first kind of order k, defined by

$$I_k(t) = \sum_{m=0}^{\infty} \frac{(\frac{t}{2})^{(k+2m)}}{(k+m)!m!}.$$

Proof. As

$$\mathcal{L}_\sigma(s) = \frac{\mu}{s+\mu},$$

it follows that

$$\mathcal{L}_U(s) = \frac{\mu}{s+\lambda-\lambda\mathcal{L}_U(s)+\mu}.$$

Hence,

$$\lambda[\mathcal{L}_U(s)]^2 - (\lambda+\mu+s)\mathcal{L}_U(s) + \mu = 0.$$

The latter equation has two solutions:

$$\alpha(s) = \frac{1}{2\lambda}\left[(\lambda+\mu+s) - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}\right];$$
$$\beta(s) = \frac{1}{2\lambda}\left[(\lambda+\mu+s) + \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}\right],$$

where $s \in \mathcal{C}$, $\Re e(s) > 0$ and $\lambda\mu > 0$.

On the other hand, $\mathcal{L}_U(s) \in \mathbf{C}$ and as the system is assumed stable, the distribution of U is not degenerated, hence $\mid \mathcal{L}_U(s) \mid \; \leq 1$. Therefore, only the roots of modulus less than or equal to 1 are suitable. We shall now check whether the two roots meet this condition.

Let us first notice that $\mid \alpha(s) \mid \; < \; \mid \beta(s) \mid$ since $\lambda\mu\Re e(s) > 0$, for any s in the right half-plane of the complex plane. When $\mid \mathcal{L}_U(s) \mid \; = 1$ (on the border of the unit disc), we have

$$\mid (\lambda+\mu+s)\mathcal{L}_U(s) \mid \; = \; \mid (\lambda+\mu+s) \mid \; > \lambda+\mu \geq \; \mid \mu + \lambda\mathcal{L}_U(s)^2 \mid.$$

Hence we have two complex functions

$$f(z) = -(\lambda + \mu + s)z;$$
$$h(z) = \mu + \lambda z^2,$$

which are analytic inside and on the border of a closed domain of the complex plane defined by its contour

$$\mathcal{C} = \{z \in \mathcal{C};\ |\,z\,| \ = 1\} \bigcap \{s; \Re e(s) > 0\}$$

and such that $|\,f(z)\,| \ > \ |\,h(z)\,|$ on the contour. Rouché's Theorem allows us to say that the largest (in modulus) of the two functions (that is to say $f(z)$) and the sum of the two functions have the same number of zeros in the area defined by the contour $\mathcal{C}$. However, in this area $f(z)$ has a single zero. Hence only one of the roots $\alpha(s)$, $\beta(s)$ is in the area bounded by $\mathcal{C}$. It must be the smallest of the two, i.e. $\alpha(s)$, which is appropriate. The result is shown. □

Theoretically, it is possible to obtain numerically from the analytic expression of $g(.)$, the distribution of $G(.)$. However, the operation is not simple in view of the complexity of the form of the above series. It is also theoretically possible to compute the distribution of U_n by inverting the function $(\mathcal{L}_U(s))^n$. In practice, it is very easy to calculate the first moment of U, as

$$\mathbf{E}\,[U] = \int_0^\infty t \,\mathrm{d}\,\mathbf{P}_\sigma(t) = -\frac{d}{\mathrm{d}\,s}\mathcal{L}_U(s)\,|_{\{s=0\}},$$

or in other words

$$\mathbf{E}\,[U] = \begin{cases} \dfrac{1}{\mu - \lambda} & \text{if } \rho < 1; \\ \infty & \text{if } \rho \geq 1. \end{cases}$$

THEOREM 5.11.– *If* $\rho > 1$,

$$\mathbf{P}\,(U < \infty) = \frac{1}{\rho} < 1 \textit{ and } \mathbf{P}\,(U = \infty) = 1 - \frac{1}{\rho} > 0.$$

Proof. Observe that

$$\lim_{s \to 0} \mathcal{L}_U(s) = \lim_{s \to 0} \int_0^\infty e^{-sx}\,\mathrm{d}\,\mathbf{P}_\sigma(x) = \lim_{x \to \infty} G(x).$$

But the latter tends to $\mathbf{P}\,(U < \infty)$. If the queue is stable ($\rho < 1$), the state 0 is positive recurrent: starting from 0, the chain comes back to it almost surely after a finite time, with a finite average excursion time. Therefore,

$$\rho < 1 \Longrightarrow G(\infty) = \mathbf{P}\,(U < \infty) = 1.$$

If $\rho = 1$, then 0 is null recurrent. The chain returns to zero almost surely, but the average time between two visits is infinite. We still have $G(\infty) = 1$. Finally if $\rho > 1$, then 0 is transient, and there is a non-zero probability of never returning to 0 starting from this state. Hence,

$$\rho > 1 \Longrightarrow G(\infty) = \mathcal{L}_U(0) = \frac{1}{2\lambda}[\lambda + \mu - \sqrt{(\lambda + \mu)^2 - 4\lambda\mu}] = \frac{\mu}{\lambda}.$$

Hence the result. □

Epitome

– In a M/GI/1 queue, the process counting the number of customers in the system is not Markov. We restrict ourselves to the embedded chain taken at the departure times of the customers.

– We can calculate all the characteristics of this queue (waiting time, sojourn time, length of the busy period, etc.) using their Laplace transforms.

– The waiting time depends not only on the traffic load. It also depends on the variability of the service times. Unfortunately, the variance analysis of the service times is not enough to characterize this variability.

PART 2

Continuous-time Modeling

Chapter 6

Poisson Process

The modeling of a physical system must comply with two constraints. On the one hand, it must reflect the reality as accurately as possible, and on the other hand, it must have a predictive role, in other words it must provide computational tools for the analysis. Beyond the difficulty to qualitatively and quantitatively determine the pertinent parameters of a physical system, the experience shows that the more one wants an accurate model, the less it will be tractable in practice.

Within the framework of queuing systems, we must, in the first place, model the process of arrivals of the requests. The Poisson process which we study in this Chapter, is the most frequently used model, primarily because it is one of the rare models with which we can make computations. This modeling is found to be highly pertinent for the telephone calls to a commutator. Unfortunately, this is not the same for other types of network, where the traffic is much more versatile. However, as we will see at the end of this chapter, the Poisson process can be modified, so as to reflect this versatility to a certain extent.

The definition of a point process and the associated notations are given in A.5.2. Let us recall that an integrable point process is a strictly increasing sequence of positive random variables $(T_1,\, T_2,\, \ldots)$ such that $T_n \to \infty$ a.s.. By convention, we adjoin the random variable $T_0 = 0$ a.s. to this sequence. These random variables will represent the arrival times of requests to the system. We can also describe the sequence by the differences in time which elapses between the successive arrivals: $\xi_n = T_{n+1} - T_n$ is the nth *inter-arrival* time. The sequence $(\xi_n,\, n \in \mathbf{N})$ also characterizes the point process by the relation $T_n = \sum_{i \leq n-1} \xi_i$. We will finally denote $N(t)$, the number of points (that is, of arrivals), up to time t.

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

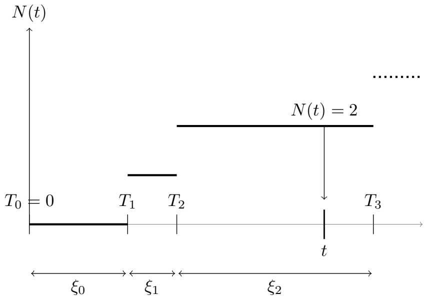

Figure 6.1. *Notations related to point processes*

6.1. Definitions

The Poisson process admits multiple characterizations. As each one of them can be considered as a definition, and the others as properties, we give to all the status of definition and then show that they are equivalent.

DEFINITION 6.1.– *The point process N is a Poisson process of intensity λ if, and only if, the random variables $(\xi_n,\ n \in \mathbf{N})$ are independent and of the same exponential distribution with parameter λ.*

DEFINITION 6.2.– *The point process N is a Poisson process of intensity λ if, and only if, the following two conditions are satisfied:*

1) $N(t)$ follows a Poisson distribution with parameter λt;

2) conditionally to $\{N(t) = n\}$, the family $(T_1,\ \dots,\ T_n)$ is uniformly distributed over $[0, t]$.

DEFINITION 6.3.– *The point process N is a Poisson process of intensity λ if, and only if the following two conditions are satisfied:*

1) for any $0 = t_0 < t_1 < \cdots < t_n$, the random variables $(N(t_i + 1) - N(t_i),\ 1 \leq i \leq n - 1)$ are independent;

2) for any $t,\ s$, the random variables $N(t + s) - N(t)$ follow a Poisson distribution of parameter λs, i.e.

$$\mathbf{P}(N(t + s) - N(t) = k) = \exp(-\lambda s)\frac{(\lambda s)^k}{k!}.$$

DEFINITION 6.4.– *The point process N is a Poisson process of intensity λ if, and only if, for any function $f : \mathbf{R}^+ \to \mathbf{R}^+$ (or for any function f with compact support on $\mathbf{R}^+$), the following identity holds.*

$$\mathbf{E}\left[\exp\left(-\sum_{n\geq 1} f(T_n)\right)\right] = \exp\left(-\int_0^\infty (1 - e^{-f(s)})\,\lambda\,\mathrm{d}\,s\right).$$

DEFINITION 6.5.– *The point process N is a Poisson process of intensity λ if, and only if, the process $(N(t) - \lambda t,\, t \geq 0)$ is a martingale with respect to the filtration $\mathcal{F}$ generated by N, i.e. $\mathcal{F}_t = \sigma\{N(s),\, s \leq t\}$.*

In order to show the equivalence between these definitions, we must introduce three technical results.

LEMMA 6.1.– *The density of the distribution of $(T_1, \ldots, T_n)$ is given by*

$$\mathrm{d}\,\mathbf{P}_{(T_1,\ldots,T_n)}(x_1, \ldots, x_n) = \lambda^n \exp(-\lambda x_n) \mathbf{1}_{\mathcal{C}}(x_1, \ldots, x_n)\ \mathrm{d}\,x_1 \ldots \mathrm{d}\,x_n, \qquad [6.1]$$

where

$$\mathcal{C} = \big\{(y_1, \ldots, y_n) \in (\mathbf{R}^+)^n, 0 \leq y_1 \leq \cdots \leq y_n\big\}.$$

In particular, T_n follows a gamma distribution with parameters n and λ, defined by

$$\mathrm{d}\,\mathbf{P}_{T_n}(x) = \lambda^n \exp(-\lambda x) \frac{x^{n-1}}{(n-1)!} \mathbf{1}_{\mathbf{R}^+}(x)\ \mathrm{d}\,x. \qquad [6.2]$$

Proof. We proceed by identification. For all bounded measurable f,

$$\mathbf{E}\left[f(T_1, \ldots, T_n)\right]$$
$$= \int \cdots \int_{(\mathbf{R}^+)^n} f\left(x_0,\, x_0 + x_1,\, \ldots,\, x_0 + \ldots + x_{n-1}\right) \mathrm{d}\,\mathbf{P}_{\xi_0}(x_0) \ldots \mathrm{d}\,\mathbf{P}_{\xi_{n-1}}(x_{n-1}).$$

Perform the change of variable

$$u_1 = x_0,\ u_2 = x_0 + x_1,\ \ldots,\ u_n = x_0 + \ldots + x_{n-1},$$

whose Jacobian equals 1. The conditions $x_0 \geq 0, \ldots, x_{n-1} \geq 0$ amounts to $0 \leq u_1 \leq u_2 \ldots \leq u_n$. We therefore have

$$\mathbf{E}\left[f(T_1, \ldots, T_n)\right] = \int \cdots \int_{(\mathbf{R}^+)^n} f(u_n) \lambda^n e^{-\lambda u_n} \mathbf{1}_{\mathcal{C}}(u_1, \ldots, u_n)\ \mathrm{d}\,u_1 \ldots \mathrm{d}\,u_n.$$

The density of the joint distribution follows from it. If f depends only on T_n, we obtain

$$\begin{aligned}\mathbf{E}\left[f(T_n)\right] &= \int \ldots \int_{0 \leq u_1 \ldots \leq u_n} f(u_n) \lambda^n \exp(-\lambda u_n)\ \mathrm{d}\,u_1 \ldots \mathrm{d}\,u_n \\ &= \int_0^\infty f(u_n) \lambda^n \exp(-\lambda u_n) \left(\int_0^{u_n} \mathrm{d}\,u_{n-1} \int \ldots \int_0^{u_2} \mathrm{d}\,u_1 \right) \mathrm{d}\,u_n \\ &= \int_0^\infty f(u_n) \lambda^n \exp(-\lambda u_n) \frac{u_n^{n-1}}{(n-1)!}\ \mathrm{d}\,u_n.\end{aligned}$$

The result follows. □

LEMMA 6.2.– *Let X be a random variable of Poisson distribution with parameter λ. We have*

$$\mathbf{E}\left[e^{-sX}\right] = \exp(-\lambda(1-e^{-s})).$$

Proof. By definition of the Poisson distribution, we have

$$\mathbf{E}\left[e^{-sX}\right] = \sum_{k=0}^{\infty} e^{-sk} e^{-\lambda} \frac{\lambda^k}{k!} = \exp(-\lambda + \lambda e^{-s}),$$

hence the result. □

LEMMA 6.3.– *Let $(U_1, \ldots, U_n)$ be n independent random variables of uniform distribution on $[0, t]$. Let $\bar{U}$ represent the reordering of the n-tuple in increasing order, that is*

$$\bar{U}_1(\omega) \leq \bar{U}_2(\omega) \leq \ldots \leq \bar{U}_n(\omega), \text{ a.s.}.$$

The distribution of $\bar{U}$ is given by

$$\mathrm{d}\,\mathbf{P}_{(\bar{U}_1, \ldots, \bar{U}_n)}(x_1, \ldots, x_n) = \frac{n!}{t^n} \mathbf{1}_{\mathcal{C}}(x_1, \ldots, x_n)\, \mathrm{d}\,x_1 \ldots \mathrm{d}\,x_n.$$

Proof. Denote σ, the random variable with values in the group of permutations $\mathfrak{S}_n$ of $[\![1, n]\!]$, representing the permutation of indexes necessary to arrange the values of $U_i(\omega)$ in increasing order, e.g. if we have

$$U_2(\omega) \leq U_3(\omega) \leq U_1(\omega),$$

$\sigma(\omega)$ is defined by

$$\sigma(\omega) = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

The image of i by $\sigma(\omega)$ is the index of the random variable which is in the ith position for the sample ω. Therefore, by definition we have $\bar{U}_i(\omega) = U_{\sigma(\omega)(i)}(\omega)$. As the random variables U_i, $i \in [\![1, n]\!]$ are independent and of the same distribution, for any $\tau \in \mathfrak{S}_n$,

$$\mathrm{d}\,\mathbf{P}_{(U_{\tau(1)}, \ldots, U_{\tau(n)})}(u_1, \ldots, u_n) = \otimes_{i=1}^{n} \frac{1}{t} \mathbf{1}_{[0,t]}(u_i)\, \mathrm{d}\,u_i.$$

Notice, in particular, that this distribution does not depend on τ. Therefore,

$$\begin{aligned}\mathbf{P}(\sigma = \tau) = \mathbf{P}\left(U_{\tau(1)} \leq \ldots \leq U_{\tau(n)}\right) \\ &= \int \cdots \int \mathbf{1}_{\mathcal{C}}(u_1, \ldots, u_n)\, \mathrm{d}\,\mathbf{P}_{(U_{\tau(1)}, \ldots, U_{\tau(n)})}(u_1, \ldots, u_n) \\ &= \mathbf{P}(\sigma = \mathrm{Id}).\end{aligned}$$

Thus, σ follows a uniform distribution on $\mathfrak{S}_n$, that is to say

$$\mathbf{P}(\sigma = \tau) = \frac{1}{n!}.$$

To compute the distribution of the n-tuple $\bar{U}$, we partition the probability space in $\cup_{\tau \in \mathfrak{S}_n} (\sigma = \tau)$. For any bounded continuous function f, we have

$$\begin{aligned}
\mathbf{E}\left[f(\bar{U}_1, \ldots, \bar{U}_n)\right] &= \sum_{\tau \in \mathfrak{S}_n} \mathbf{E}\left[f(\bar{U}_1, \ldots, \bar{U}_n);\, \sigma = \tau\right] \\
&= \sum_{\tau \in \mathfrak{S}_n} \mathbf{E}\left[f(U_{\tau(1)}, \ldots, U_{\tau(n)}) \mathbf{1}_{\mathcal{C}}(U_{\tau(1)}, \ldots, U_{\tau(n)})\right] \\
&= \sum_{\tau \in \mathfrak{S}_n} \int \cdots \int f(u_1, \ldots, u_n) \mathbf{1}_{\mathcal{C}}(u_1, \ldots, u_n)\, \mathrm{d}\, \mathbf{P}_{(U_{\tau(1)}, \ldots, U_{\tau(n)})}(u_1, \ldots, u_n) \\
&= \sum_{\tau \in \mathfrak{S}_n} \int \cdots \int f(u_1, \ldots, u_n) \mathbf{1}_{\mathcal{C}}(u_1, \ldots, u_n)\ \otimes_{i=1}^n \frac{1}{t} \mathbf{1}_{[0,t]}(u_i)\, \mathrm{d}\, u_i \\
&= \frac{n!}{t^n} \int \cdots \int f(u_1, \ldots, u_n) \mathbf{1}_{\mathcal{C}}(u_1, \ldots, u_n)\ \mathrm{d}\, u_1 \ldots \mathrm{d}\, u_n.
\end{aligned}$$

Hence the result. □

Proof of equivalence between the definitions. We are going to show the implication chain: 6.1 ⟹ 6.2 ⟹ 6.3 ⟹ 6.4 ⟹ 6.5 ⟹ 6.1.

6.1 ⟹ 6.2 Let us first show that $N(t)$ follows a Poisson distribution. Since it is clear that the events $\{N(t) = k\}$ and $\{T_k \le t < T_{k+1}\}$ coincide, we have

$$\begin{aligned}
\mathbf{P}(N(t) = k) &= \mathbf{P}(T_k \le t < T_k + \xi_{k+1}) \\
&= \iint \mathbf{1}_{\{x \le t\}}\, \mathbf{1}_{\{x+y > t\}}\, \mathrm{d}\, \mathbf{P}_{T_k}(x)\, \mathrm{d}\, \mathbf{P}_{\xi_k}(y) \\
&= \int_0^t \left(\int_{t-x}^{\infty} \lambda e^{-\lambda y}\, dy \right)\ \lambda^k \frac{x^{k-1}}{(k-1)!} \exp(-\lambda x)\, \mathrm{d}\, x \\
&= e^{-\lambda t} \int_0^t \lambda^k \frac{x^{k-1}}{(k-1)!}\ \mathrm{d}\, x \\
&= e^{-\lambda t} \frac{(\lambda x)^k}{k!}.
\end{aligned}$$

For the conditional distribution, we proceed similarly:

$$\begin{aligned}
&\mathbf{P}\left((T_1,\,\ldots,\,T_n) \in A \,|\, N(t)=n\right) \mathbf{P}(N(t)=n)\\
&\quad = \mathbf{P}\left((T_1,\,\ldots,\,T_n) \in A, T_n \le t < T_{n+1}\right)\\
&\quad = \int\!\!\cdots\!\!\int_{0\le u_1\le\ldots\le u_{n+1}} \mathbf{1}_A(u_1,\,\ldots,\,u_n)\mathbf{1}_{[u_n,\,u_{n+1})}(t)\lambda^{n+1}e^{-\lambda u_{n+1}}\,\mathrm{d}\,u_1\ldots\mathrm{d}\,u_{n+1}\\
&\quad = \lambda^n \int\!\!\cdots\!\!\int_{0\le u_1\le\ldots\le u_{n+1}} \mathbf{1}_A(u_1,\,\ldots,\,u_n)\mathbf{1}_{[u_n,\,\infty)}(t)\left(\int_t^\infty \lambda e^{-\lambda u_{n+1}}\right)\mathrm{d}\,u_1\ldots\mathrm{d}\,u_n\\
&\quad = \lambda^n e^{-\lambda t} \int\!\!\cdots\!\!\int_{0\le u_1\le\ldots\le u_{n+1}} \mathbf{1}_A(u_1,\,\ldots,\,u_n)\mathbf{1}_{[u_n,\,\infty)}(t)\,\mathrm{d}\,u_1\ldots\mathrm{d}\,u_n.
\end{aligned}$$

By dividing the term on the right-hand side by $e^{-\lambda t}(\lambda t)^n/n!$, we obtain

$$\begin{aligned}
&\mathbf{P}\left((T_1,\,\ldots,\,T_n) \in A \,|\, N(t)=n\right)\\
&\quad = \frac{n!}{t^n} \int\!\!\cdots\!\!\int_{0\le u_1\le\ldots\le u_{n+1}} \mathbf{1}_A(u_1,\,\ldots,\,u_n)\mathbf{1}_{[u_n,\,\infty)}(t)\,\mathrm{d}\,u_1\cdots\mathrm{d}\,u_n,
\end{aligned}$$

which, in view of Lemma 6.3, means that $(T_1,\,\ldots,\,T_n)$ has conditionally to $\{N(t)=n\}$, the same distribution as the vector $(\bar{U}_1,\,\ldots,\,\bar{U}_n)$ defined therein. In other words, conditionally to $\{N(t)=n\}$, the n-tuple $(T_1,\,\ldots,\,T_n)$ is uniformly distributed over $[0,t]$.

6.2 $\Longrightarrow$ **6.3** Let $t_0 = 0 < t_1 < \cdots < t_n$ be a family of $n+1$ real numbers and $i_0,\,\ldots,\,i_{n-1}$, a family of n integers. We aim to prove that

$$\mathbf{P}\left(\bigcap_{l=0}^{n-1}\{N(t_{l+1})-N(t_l)=i_l\}\right) = \prod_{l=1}^{n}\mathbf{P}(N(t_{l+1})-N(t_l)=i_l).$$

We can always write that

$$\begin{aligned}
&\mathbf{P}\left(\bigcap_{l=0}^{n-1}\{N(t_{l+1})-N(t_l)=i_l\}\right)\\
&\quad = \sum_{k\in\mathbf{N}} \mathbf{P}\left(\bigcap_{l=0}^{n-1}\{N(t_{l+1})-N(t_l)=i_l\}\,|\,N(t_n)=k\right)\mathbf{P}(N(t_n)=k).
\end{aligned}$$

The unique value of k for which the conditional probabilities of the latter quantity are non-zero is $k_0 = \sum_l i_l$. In order to derive the corresponding conditional probability, we know that the points between 0 and t_n are uniformly distributed. This quantity thus equals the probability that k points that are uniformly distributed over an interval, divides into i_1 points in the interval $[0, t_1]$, i_2 points in the interval $]t_1, t_2]$, and so on. Each point belongs to an interval of length x length with probability x/t_n,. We thus have

$$\mathbf{P}\left(\bigcap_{l=0}^{n-1}\{N(t_{l+1}) - N(t_l) = i_l\} \,|\, N(t_n) = k_0\right) = \frac{k_0!}{i_1! \dots i_n!} \prod_{l=0}^{n-1} \left(\frac{t_{l+1} - t_l}{t_n}\right)^{i_l}.$$

As $N(t_n)$ follows a Poisson distribution with parameter λt_n and $k_0 = \sum_{l=0}^{n-1} i_1$, we deduce that

$$\begin{aligned}\mathbf{P}&\left(\bigcap_{l=0}^{n-1}\{N(t_{l+1}) - N(t_l) = i_l\}\right)\\ &= e^{-\lambda t_n}\frac{(\lambda t_n)^{k_0}}{k_0!}\frac{k_0!}{i_1! \dots i_n!} \prod_{l=0}^{n-1} \left(\frac{t_{l+1} - t_l}{t_n}\right)^{i_l}\\ &= \prod_{l=0}^{n-1} e^{-\lambda(t_{l+1}-t_l)}\frac{(\lambda(t_{l+1} - t_l))^{i_l}}{i_l!}.\end{aligned}$$

The probability of the intersection of events thus reads as a product of probabilities, therefore the random variables are independent. By taking $n = 2$, $t_1 = t$, $t_2 = t + s$, $i_0 = i$ and $i_1 = j$, we obtain

$$\mathbf{P}\left(N(t) = i,\, N(t+s) - N(t) = j\right) = e^{-\lambda t}\frac{(\lambda t)^i}{i!}e^{-\lambda s}\frac{(\lambda s)^j}{j!}.$$

Finally, summing over all the values of i yields the desired result.

6.3 $\Longrightarrow$ **6.4** Notice, that taking $f(s) = \mathbf{1}_{[a,b]}(s)$ leads to

$$\sum_n f(T_n) = N(b) - N(a).$$

From Lemma 6.2, we thus deduce that the result is true for the indicator functions and by linearity, for the piece-wise constant functions (that is to say, the linear combinations of indicator functions). By monotone convergence, we deduce that the result holds true for any positive measurable function.

6.4 $\Longrightarrow$ **6.5** It suffices to write $N(t+s) = (N(t+s) - N(t)) + N(t)$ and to use the independence of these random variables to prove that

$$\mathbf{E}\left[N(t+s) \,|\, \mathcal{F}_t\right] = \mathbf{E}\left[N(t+s) - N(t)\right] + N(t) = \lambda s + N(t),$$

hence the result.

6.5 $\Longrightarrow$ 6.1 As N remains constant between two jumps, for any bounded f we have

$$\begin{aligned}
&f(N(t)) - f(N(s))\\
&\quad = \sum_{s<r\leq t,\, \Delta N(r)=1} f(N(r)) - f(N(r_-))\\
&\quad = \int_s^t (f(N(r_-)+1) - f(N(r_-)))\,\mathrm{d}\,N(r)\\
&\quad = \int_s^t (f(N(r_-)+1) - f(N(r_-)))(\mathrm{d}\,N(r) - \lambda\,\mathrm{d}\,r)\\
&\qquad + \int_s^t (f(N(r_-)+1) - f(N(r_-)))\lambda\,\mathrm{d}\,r.
\end{aligned}$$

As the process $(r \mapsto f(N(r_-))$ is predictable, Theorem A.34 implies that the stochastic integral is in fact a martingale. So the process defined by

$$t \longmapsto f(N(t)) - \int_0^t (f(N(r_-)+1) - f(N(r_-)))\lambda\,\mathrm{d}\,r$$

is a martingale. According to Theorem 7.15, N is a Markov process of infinitesimal generator

$$Af(x) = \lambda(f(x+1) - f(x)) \text{ for any bounded } f\colon \mathbf{N} \to \mathbf{R}.$$

From Theorem 7.9, Definition 6.1 is verified. □

6.2. Properties

Definition 6.2 might lead to a misinterpretation, and should be clearly understood. The latter stipulates that, conditionally to the number of points on an interval, the points are uniformly distributed over this interval. When we observe a sample path of the process, knowing t and the number of impacts in this interval, we should observe a cloud of point that is uniformly distributed. However, we observe distributions that are similar to that of Figure 6.2. This is the "clusterization" phenomenon: the arrivals give the impression of being grouped. The same observation can be made in actual stores, where after an idle period, many customers may arrive almost at the same time.

In fact, the very concept of uniform distribution is vague, and should not be confused with the constancy of the difference between the arrivals. As shown in Figure 6.2, the uniformity in the distribution is "seen" on several sample paths: here there are almost no parts of $[0, 1]$ that does not have any impact in one or the other of the sample paths. There is a primary difference between the apparent clusterization phenomenon of the

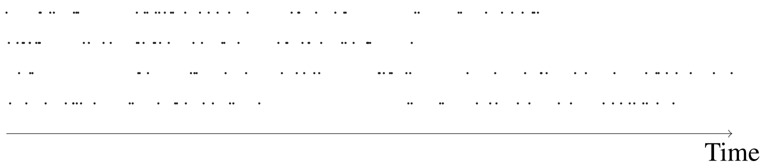

Figure 6.2. *Four trajectories of a Poisson process. There are no areas of the segment that is not covered by any one of the trajectories, but each of the trajectories present "bursts" of arrivals*

Poisson process and the "bursts" phenomenon observed in the LAN-WAN Broadband. In the first case, the instantly average throughput (calculated averaging over a large number of sample paths) does not depend on time (as it is equal to λ) whereas in the other case, it will largely vary over time (think of the variable throughput video traffic flow). Thus, we cannot represent such a traffic by a Poisson process.

THEOREM 6.4.– *Let N be a Poisson process of intensity λ. For any function f with compact support, we have*

$$\mathbf{E}\left[\exp\left(-\sum_{n\geq 1} f(T_n)\right)\right] = \exp\left(-\int_{\mathbf{R}^+}(1-e^{-f(s)})\lambda\,\mathrm{d}\,s\right). \qquad [6.3]$$

Proof. By taking $f(s) = \alpha\mathbf{1}_{[a,b]}(s)$, we have

$$\sum_n f(T_n) = \alpha(N(b) - N(a)).$$

We know that $N(b) - N(a)$ follows a Poisson distribution with parameter $\lambda(b-a)$. Therefore,

$$\begin{aligned}\mathbf{E}\left[\exp\left(-\sum_{n\geq 1} f(T_n)\right)\right] &= \sum_{n=0}^{\infty} e^{-\alpha n} e^{-\lambda(b-a)}\frac{(\lambda(b-a))^n}{n!}\\ &= \exp(-\lambda(b-a) + \lambda(b-a)e^{-\alpha})\\ &= \exp\left(-\int\left(1-e^{-f(s)}\right)\lambda\,\mathrm{d}\,s\right).\end{aligned}$$

By independence of the increments, equation [6.3] thus holds true for the step functions. By dominated convergence, this is also the case for the functions with compact support. □

6.2.1. *Superposition, thinning*

When we set the two point processes N^1 and N^2, the superposition of these processes is the point process, denoted as $N = N^1 + N^2$, whose points are those of N^1 and N^2. With Theorem 6.4 in hand, the following result is straightforward.

THEOREM 6.5.– *The superposition of two independent Poisson processes of respective intensities λ_1 and λ_2 is a Poisson process of intensity $\lambda_1 + \lambda_2$.*

NOTE.– In particular, during the superposition of two independent Poisson processes, two clients cannot arrive simultaneously. This result holds true for any two independent point processes.

Let us now assume that a Poisson process N of intensity λ is split into two processes N^1 and N^2 according to a random draw of Bernoulli of probability p, that is to say that at for any point of N, we decide that it belongs to N^1 with probability p and to N^2 with probability $1 - p$. This draw is assumed as independent of everything else in the model, and in particular of the previous draws (Figure 6.3). It is said that the Poisson process N is *thinned*.

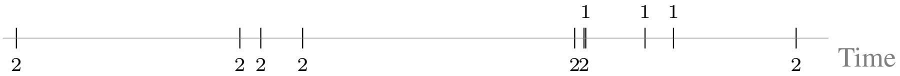

Figure 6.3. *Thinning of a Poisson process. The figure above or below each point represents the number of flow to which this point is attributed*

Theorem 6.4 ensures that a Poisson process on $\mathbf{R}^+$ is a special case of a spatial Poisson process (see section 10.3). Therefore, Theorem 10.6 implies the following result.

THEOREM 6.5.1.– *The processes N^1 and N^2 processes resulting from the thinning of N, are two independent Poisson processes of respective intensities λp and $\lambda(1 - p)$.*

6.2.2. *Bus paradox*

This is a specific result of the one dimensional case, known as the bus paradox (or inspection paradox). Let us interpret the points of a Poisson process as the arrival times of buses at a given bus stop. Let

$$\begin{cases} W(t) = T_{N(t)+1} - t, \\ Z(t) = t - T_{N(t)}, \\ \xi(t) = W(t) + Z(t) = T_{N(t)+1} - T_{N(t)}, \end{cases}$$

be the waiting time of the bus when arriving at the stop at time t, the time elapsed since the last bus went by, and the length of the time interval between the bus that we take and that we missed, respectively.

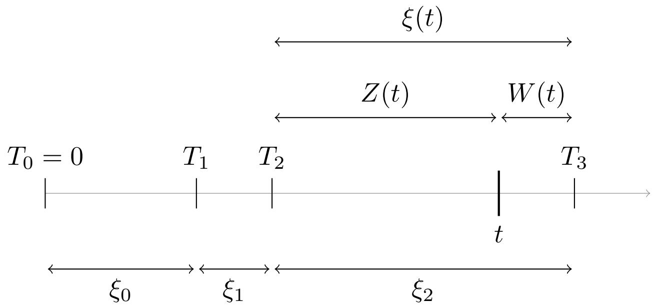

Figure 6.4. *Notations*

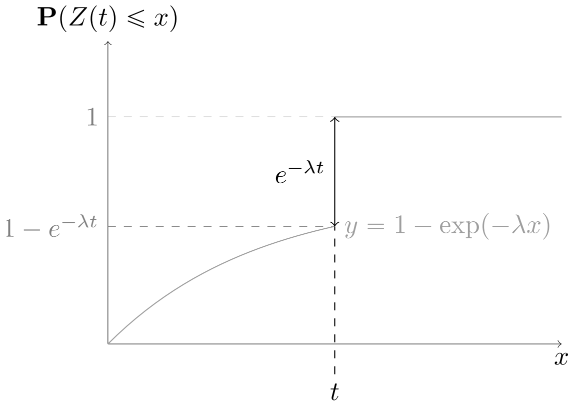

Figure 6.5. *Distribution function of $Z(t)$*

THEOREM 6.6 (BUS PARADOX).– *$W(t)$ follows an exponential distribution of parameter λ and is independent of $Z(t)$, whose distribution is given by*

$$\mathbf{P}(Z(t) \leq x) = \begin{cases} 1 - e^{-\lambda x} & \textit{if } 0 \leq x < t; \\ 1 & \textit{if } x \geq t. \end{cases}$$

This may seem paradoxical since the average value of W_t, that is to say, the average waiting time, is therefore $1/\lambda$, whereas the average time between two passages of the bus also equals $1/\lambda$. This property is another manifestation of what is commonly called the memoryless property of the exponential distribution, which will be discussed in the next chapter. In fact, everything happens as if, at the moment when we arrive at the bus stop, the counter of time which elapses between two bus arrivals was reset to zero, and if we recounted a time of exponential distribution until the next arrival.

Clearly speaking, this approach is mathematically wrong because if ξ_n follows an exponential distribution of parameter λ for any fixed n, $\xi(t)$ does not have the distribution of a ξ_n. Indeed, the number of the buses that we have missed changes along the samples paths. Conditionally to $\{N(t) = n\}$, $\xi(t)$ indeed has an exponential distribution of parameter λ, but contrary to what we might believe, when we uncondition, the distribution of $\xi(t)$ is no longer exponential.

Proof. Let us observe that $Z(t)$, by its very definition, cannot be larger than t. For $0 \le x < t$ and $y \ge 0$, by using the independence of T_n and ξ_n for any $n \in \mathbf{N}$, we have

$$\begin{aligned}\mathbf{P}(Z(t) \le x,\, W(t) > y) &= \sum_{n\in\mathbf{N}} \mathbf{P}(Z(t) \le x,\, W(t) > y,\, N(t) = n)\\ &= \sum_{n\in\mathbf{N}} \mathbf{P}(t - T_n \le x,\, T_n + \xi_n - t > y,\, T_n \le t < T_n + \xi_n).\end{aligned}$$

For $n = 0$ the latter probability equals

$$\mathbf{P}(t \le x,\, \xi_0 > t + y) = 0,$$

since we have assumed that $x < t$. The previous sum therefore reads

$$\begin{aligned}&\sum_{n\in\mathbf{N}^*} \iint \mathbf{1}_{\{t-u\le x\}}\,\mathbf{1}_{\{u+v-t>y\}}\,\mathbf{1}_{\{u\le t\}}\,\mathrm{d}\,\mathbf{P}_{T_n}(u)\,\mathrm{d}\,\mathbf{P}_{\xi_{n+1}}(v)\\ &\quad= \sum_{n\in\mathbf{N}^*} \int_{t-x}^{t} \lambda^n e^{-\lambda u} \frac{u^{n-1}}{(n-1)!} \left(\int_{t+y-u}^{\infty} \lambda e^{\lambda v}\,\mathrm{d}\,v\right) \mathrm{d}\,u\\ &\quad= \sum_{n\in\mathbf{N}^*} \lambda^n e^{-\lambda(t+y)} \int_{t-x}^{t} \frac{u^{n-1}}{(n-1)!}\,\mathrm{d}\,u\\ &\quad= \sum_{n\in\mathbf{N}^*} \lambda^n e^{-\lambda(t+y)} \left[\frac{t^n}{n!} - \frac{(t-x)^n}{n!}\right]\\ &\quad= e^{-\lambda(t+y)} \left(\sum_{n\in\mathbf{N}^*} \frac{(\lambda t)^n}{n!} - \sum_{n\in\mathbf{N}^*} \frac{(\lambda(t-x))^n}{n!}\right)\\ &\quad= e^{-\lambda(t+y)}(e^{\lambda t} - 1 - (e^{\lambda(t-x)} - 1))\\ &\quad= e^{-\lambda y}(1 - e^{-\lambda x}).\end{aligned}$$

As $\lim_{x\nearrow t}(1 - e^{-\lambda x}) = 1 - e^{-\lambda t} < 1$, we deduce from it that there is a jump in the distribution function of $Z(t)$ and therefore $\mathbf{P}(Z(t) = t) = e^{-\lambda t} > 0$. □

NOTE.– The mean expectation of $\xi(t)$ is derived in the following manner.

$$\begin{aligned}\mathbf{E}\left[\xi(t)\right] &= \mathbf{E}\left[W(t)\right] + \mathbf{E}\left[Z(t)\right] \\ &= \frac{1}{\lambda} + t.e^{-\lambda t} + \int_0^t \lambda s e^{-\lambda s}\,\mathrm{d}\,s \\ &= \frac{1}{\lambda}(2 - e^{-\lambda t}).\end{aligned}$$

Therefore, as t goes large the expectation of $\xi(t)$ tends to $2/\lambda$ and the average waiting time, which equals $1/\lambda$, represents half of it. This is in accordance with the intuition.

6.3. Discrete analog: the Bernoulli process

The analog of the Poisson process in discrete time is defined as follows.

DEFINITION 6.6.– *Let* $p \in]0,1[$ *and* $\left(\tilde{\xi}_n,\ n \in \mathbf{N}\right)$ *be a sequence of random variables, independent and identically distributed, having a geometric distribution of parameter* p*, that is*

$$\mathbf{P}(\tilde{\xi}_0 = k) = p(1-p)^{k-1};\ k \in \mathbf{N}^*.$$

We then set $\tilde{T}_0 = 0$ *and* $\tilde{T}_{n+1} - \tilde{T}_n = \tilde{\xi}_n$ *for any* $n \in \mathbf{N}$*. The point process* $(T_0, T_1, \ldots, T_n, \ldots)$ *hence defined is called a* Bernoulli process *of parameter* p*.*

NOTE.– As for the Poisson process, we can define the random process $\tilde{N}$ with rcll paths, defined for all $t \geq 0$ by

$$\tilde{N}(t) = \sum_{n\in\mathbf{N}} \mathbf{1}_{\{\tilde{T}_n \leq t\}},$$

which counts the number of points of the process until t. It is clear that $\tilde{N}(t)$ follows for any $t \geq 0$, the binomial distribution $\mathcal{B}(\lfloor t \rfloor, p)$, where $\lfloor . \rfloor$ denotes the entire part.

We verify hereafter that the Bernoulli process satisfies the bus paradox, similarly to the Poisson process. By analogy to section 6.2.2, let us denote, for any $k \in \mathbf{N}$,

$$\begin{cases}\tilde{W}_k = k - \tilde{T}_{\tilde{N}(k)}, & \text{the time elapsed since the last point before } k; \\ \tilde{Z}_k = \tilde{T}_{\tilde{N}(k)+1}, & \text{the time to wait before the first point after } k.\end{cases}$$

By definition, $\tilde{W}_k$ is zero whenever there is precisely a point at k, while $\tilde{Z}_k$ is necessarily strictly positive. We have the following result.

THEOREM 6.6.1.– *For any $k \geq 0$, $\tilde{W}_k$ follows the geometric distribution with parameter p and is independent of $\tilde{Z}_k$, whose distribution is given by*

$$\mathbf{P}(\tilde{Z}_k = i) = \begin{cases} p(1-p)^i & \textit{if } i < k; \\ (1-p)^i & \textit{if } i = k; \\ 0 & \textit{if } i > k. \end{cases}$$

Proof. The scheme of proof is the same as that of Theorem 6.6. Let us fix $j \in \mathbf{N}^*$. It is clear by definition that $\tilde{Z}_k$ cannot be greater than k. Moreover, for any $i < k$, $\tilde{Z}_k = i$ if $\tilde{N}(k)$ is strictly positive (else $k = i$) and strictly less than $k - i$, because there is no point between times $k - i$ and k. Therefore,

$$\begin{aligned}\mathbf{P}(\tilde{Z}_k = i,\, \tilde{W}_k = j) &= \sum_{1 \leq n \leq k-i} \mathbf{P}(\tilde{Z}_k = i,\, \tilde{W}_k = j,\, \tilde{N}(k) = n) \\ &= \sum_{1 \leq n \leq k-i} \mathbf{P}(\tilde{T}_n = k - i,\, \tilde{\xi}_n = i + j) \\ &= \sum_{1 \leq n \leq k-i} \mathbf{P}(\tilde{T}_n = k - i) p (1-p)^{i+j-1},\end{aligned}$$

by independence. But $T_n = k - i$ amounts to saying that there is a point at $k - i$ (and therefore a success to a Bernoulli draw of parameter p) and if $k - i - 1 > 0$, that there has been $n - 1$ successes during the previous $k - i - 1$ independent Bernoulli draws. In other words,

$$\mathbf{P}(\tilde{T}_n = k - i) = p\mathbf{P}(B = n - 1) = pC_{k-i-1}^{n-1} p^{n-1} (1-p)^{k-i-1-(n-1)},$$

where B is a random variable of binomial distribution $\mathcal{B}(k - i - 1, p)$ and by setting possibly $C_0^0 = 1$. Therefore, for any $i < k$, we have

$$\begin{aligned}\mathbf{P}(\tilde{Z}_k = i,\, \tilde{W}_k = j) &= \sum_{1 \leq n \leq k-i} pC_{k-i-1}^{n-1} p^{n-1} (1-p)^{k-i-1-(n-1)} p (1-p)^{i+j-1} \\ &= p^2 (1-p)^{i+j-1} \sum_{0 \leq n \leq k-i-1} C_{k-i-1}^{n} p^n (1-p)^{k-i-1-n} \\ &= p(1-p)^{j-1} p (1-p)^i, \end{aligned} \qquad [6.4]$$

according to Newton's binomial formula. Moreover, $\tilde{Z}_k = k$ means that $\tilde{N}(k) = 0$, and therefore

$$\begin{aligned}\mathbf{P}(\tilde{Z}_k = k,\, \tilde{W}_k = j) &= \mathbf{P}(\tilde{\xi}_0 = i + j) \\ &= p(1-p)^{i+j-1} = p(1-p)^{j-1}(1-p)^i. \end{aligned} \qquad [6.5]$$

From [6.4] and [6.5], we deduce that $\tilde{Z}_k$ and $\tilde{W}_k$ are independent and follow the announced distributions. □

This memoryless property can be easily understood: a random variable $\tilde{\xi}_i$ counts the number of independent Bernoulli trials needed to achieve a success. The r.v. $\widetilde{W}_k$ counts the number of trials that are still required from k to obtain the first success after k. The trials being independent, we clearly see that, here again, the waiting time "capitalized" since the last success up to k does not increase the probability of success at each attempt after k, and therefore $\widetilde{W}_k$ has the same distribution as anyone of the $\tilde{\xi}_i$'s.

The latter result is thus intuitively clear, and gives an insight on the bus paradox in continuous time. In fact, the Poisson process is nothing but a somewhat "macroscopic" version of a Bernoulli process. More precisely, we set $\lambda > 0$ and for any $n \in \mathbf{N}^*$ such that $\lambda/n < 1$, we denote $\tilde{N}^n$, a Bernoulli process of parameter λ/n and of associated variables $\tilde{\xi}_0^n, \tilde{\xi}_1^n, \ldots$. Finally, for any n we define the point process $\bar{N}^n$ as follows.

$$\begin{cases} \bar{\xi}_i^n = \tilde{\xi}_i^n/n,\ i \in \mathbf{N}; \\ \bar{T}_0^n = 0,\ \bar{T}_{i+1}^n - \bar{T}_i^n = \bar{\xi}_i^n,\ i \in \mathbf{N}; \\ \bar{N}^n(t) = \sum_{i \in \mathbf{N}} \mathbf{1}_{\{\bar{T}_i^n \leq t\}}\,. \end{cases}$$

Starting from a Bernoulli process of parameter λ, in order to obtain $\bar{N}^n$, we divide the probability of occurrence of a point at any time by n and we compensate by an "acceleration of the time", by a factor n, by dividing the inter-points times by n. We therefore have the following result.

THEOREM 6.6.2.– *The sequence of processes* $\left\{\bar{N}^n, n > \lfloor\lambda\rfloor\right\}$ *converges in distribution to the Poisson process of intensity* λ.

Proof. The concept of convergence in distribution for processes is quite heavy to define, and is beyond the scope of this book. We will not get into these technicalities here. In fact, in our case it suffices to check that the inter-points of $\bar{N}^n$ tend in distribution to those of the Poisson process of intensity λ. This is clearly the case since for any $i \in \mathbf{N}$ and any t,

$$\mathbf{P}(\bar{\xi}_i^n > t) = \mathbf{P}(\tilde{\xi}_i^n > nt) = \left(1 - \frac{\lambda}{n}\right)^{\lfloor nt \rfloor},$$

and the latter quantity tends to $e^{-\lambda t}$ as n goes large. □

6.4. Simulation of the Poisson process

Definition 6.1 allows us to simulate trajectories of a Poisson process of intensity λ, by simply making successive draws of random variables of exponential distribution of parameter λ.

Algorithm 6.1. A sample path of a Poisson process (method 1)

```
Data: $\lambda\, T$
Result: a trajectory $(t_n,\ n \geq 1)$ on $[0,\ T]$ of a Poisson process of intensity $\lambda$.
$t \leftarrow 0$; $n \leftarrow 0$;
while $t \leq T$ do
    $x \leftarrow$ drawing of a $\varepsilon(\lambda)$;
    $t \leftarrow t + x$;
    $t_n \leftarrow t$;
    $n \leftarrow n + 1$
end
return $t_1,\ t_2,\ \ldots,\ t_n$
```

Definition 6.2 enables us to simulate a trajectory on $[0, t]$ by making a draw of a Poisson distribution of parameter λt, whose result is denoted k, then to carry out k uniform draws on $[0, t]$. Once arranged in increasing order, the smallest of these draws can be assimilated to the first point T_1, the second smallest to T_2, and so on.

Algorithm 6.2. A sample path of a Poisson process (method 2)

```
Data: $\lambda\, T$
Result: a trajectory $(t_n,\ n \geq 1)$ on $[0,\ T]$ of a Poisson process of intensity $\lambda$.
$n \leftarrow$ A random variable of Poisson distribution of parameter $\lambda T$;
for $i = 1,\ \ldots,\ n$ do
    $u_i \leftarrow$ drawing of a $U([0,\ 1])$;
end
$(t_1,\ \ldots,\ t_n) \leftarrow$ Sorting in increasing order of $(u_1,\ \ldots,\ u_n)$;
return $t_1,\ t_2,\ \ldots,\ t_n$
```

In both methods, if we wish to extend the trajectory on $[t, t+s]$, we restart the same procedure only on $[t, t+s]$, as Definition 6.3 guarantees that the trajectory on $[t, t+s]$ is independent of that on $[0, t]$.

Method 1 is the most advantageous for the "large" values of λ, as to simulate a Poisson distribution for a large λ happens to be impossible as long as $\exp(-\lambda)$ becomes smaller than the numerical precision of the computer. But, as we see in Chapter 10, the 2nd method is the only one which fits to the simulation of a spatial Poisson process.

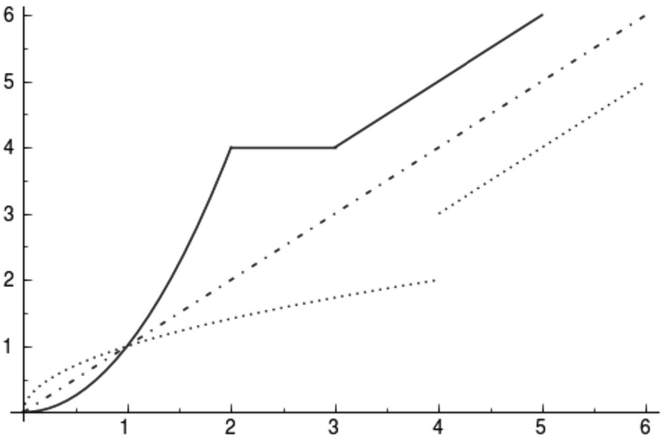

Figure 6.6. *Graph of λ and of its inverse (in dotted line)*

6.5. Non-homogeneous Poisson process

DEFINITION 6.7.– *N is a non-homogeneous Poisson process of intensity $\lambda(s)\,\mathrm{d}\,s$, where s ranges over $\mathbf{R}^+$ if, and only if for any function $f\colon \mathbf{R}^+ \to \mathbf{R}^+$,*

$$\exp\left(-\sum_{n\geq 1} f(T_n)\right) = \exp\left(-\int (1-e^{-f(s)})\,\lambda(s)\,\mathrm{d}\,s\right). \qquad [6.6]$$

This definition should be related to Definition 6.4. If λ is a constant function then we retrieve the definition of a Poisson process of intensity λ. This class of processes models traffics in which the flow varies over time, but in a deterministic manner. We always assume that λ is a rcll function.

Let us set

$$a(t) = \int_0^t \lambda(s)\ \mathrm{d}\,s;$$
$$\tau(t) = \inf\{s \geq 0,\, a(s) \geq t\}.$$

The graph of τ is obtained by taking the symmetric to that of a with respect to the first bisector. By definition, for all t, $a(\tau(t)) = t$. The inverse relation $\tau(a(t)) = t$ holds true only if λ does not vanish on an interval. In fact, in the latter case, a presents an interval of constancy and τ is no longer continuous.

THEOREM 6.6.3.– *Let N be a non-homogeneous Poisson process of intensity $\lambda(s)\,\mathrm{d}\,s$. Then the process N defined by*

$$\tilde{N}(t) = \sum_{s:\ \Delta N(s)=1} \mathbf{1}_{\{a(s)\le t\}}, \tag{6.7}$$

is a homogeneous Poisson process of intensity 1.

In other words, when N jumps at an instant t, $\tilde{N}$ has a jump at the instant $a(t)$.

Proof. It is sufficient to show that $M(t) = N(\tau_t)$ satisfies the properties of Definition 6.3. For all $0 = t_0 < t_1 < \ldots < t_n$, we have

$$M(t_{i+1}) - M(t_i) = \sum_n \mathbf{1}_{[\tau_{t_i},\,\tau_{t_{i+1}}]}(T_n).$$

As τ is deterministic, by definition of the non-homogeneous Poisson process,

$$\begin{aligned}
\mathbf{E}\left[\exp\left(-\sum_j \alpha_j (M(t_{j+1}) - M(t_j))\right)\right] = & \\
&= \mathbf{E}\left[\exp\left(-\sum_n\sum_j \alpha_j\, \mathbf{1}_{\{[\tau_{t_j},\,\tau_{t_{j+1}}]\}}(T_n)\right)\right] \\
&= \exp\left(-\sum_j \int_{\tau_{t_j}}^{\tau_{t_{j+1}}} (1-e^{-\alpha_j})\lambda(s)\ \mathrm{d}\,s\right) \\
&= \exp\left(-\sum_j (1-e^{-\alpha_j})[a(\tau_{t_{j+1}}) - a(\tau_{t_j})]\right) \\
&= \exp\left(-\sum_j \alpha_j (t_{j+1} - t_j)\right),
\end{aligned}$$

hence the result. □

With the previous notations, we deduce from the latter result the algorithm of simulation of a non-homogeneous Poisson process.

Algorithm 6.3. A sample path of a trajectory of a non-homogeneous Poisson process

```
Data: $a, T$
Result: a trajectory $(t_n,\ n \geq 1)$ on $[0,\ T]$ of a Poisson process of intensity
        $\lambda(s)\,\mathrm{d}\,s$.
$s_1,\ \ldots,\ s_n \leftarrow$ simulation of a Poisson process of intensity 1 on $a(T)$;
return $t_i = a(s_i),\ i = 1,\ \ldots,\ n$
```

THEOREM 6.7.– *The point process N is a non-homogeneous Poisson process of intensity $\lambda(s)\,\mathrm{d}\,s$ if, and only if, the process*

$$\tilde{N}\colon t \longmapsto N(t) - \int_0^t \lambda(s)\,\mathrm{d}\,s$$

is a martingale.

Proof. Let f be a function with compact support. According to Itô's Formula A.15,

$$\exp\left(\int_0^t f(r)\,\mathrm{d}\,\tilde{N}(r)\right) = 1 + \int_0^t \exp\left(\int_0^s f(r)\,\mathrm{d}\,\tilde{N}(r)\right) f(s)\,\mathrm{d}\,\tilde{N}(r)$$
$$+ \int_0^t \exp\left(\int_0^s f(r)\,\mathrm{d}\,\tilde{N}(r)\right) \left(e^{f(s)} - 1 - f(s)\right) \mathrm{d}\,N(s).$$

In view of Theorem A.34, the stochastic integral of the term on the right-hand side is a martingale, thus by taking the mean expectation, it remains

$$\begin{aligned}
&\mathbf{E}\left[\exp\left(\int_0^t f(r)\,\mathrm{d}\,\tilde{N}(r)\right)\right] \\
&\qquad = 1 + \mathbf{E}\left[\int_0^t \exp\left(\int_0^s f(r)\,\mathrm{d}\,\tilde{N}(r)\right) (e^{f(s)} - 1 - f(s))\,\mathrm{d}\,N(s)\right] \\
&\qquad = 1 + \mathbf{E}\left[\int_0^t \exp\left(\int_0^s f(r)\,\mathrm{d}\,\tilde{N}(r)\right) (e^{f(s)} - 1 - f(s))\lambda(s)\,\mathrm{d}\,s\right].
\end{aligned}$$

By letting $\phi(t) = \exp(\int_0^t f(r)\,\mathrm{d}\,\tilde{N}(r))$, we have

$$\phi(t) = 1 + \int_0^t \phi(s)u(s)\,\mathrm{d}\,s,$$

where $u(s) = (e^{f(s)} - 1 - f(s))\lambda(s))$. By solving the differential equation, we obtain

$$\phi(t) = \exp\left(\int_0^t \left(e^{f(s)} - 1 - f(s)\right) \lambda(s)\,\mathrm{d}\,s\right),$$

that is, by simplifying both sides by $\exp(-\int_0^t f(s)\lambda(s)\,\mathrm{d}\,s)$,

$$\mathbf{E}\left[\exp\left(\int_0^t f(r)\,\mathrm{d}\,N(r)\right)\right] = \exp\left(\int_0^t \left(e^{f(s)} - 1\right)\lambda(s)\,\mathrm{d}\,s\right).$$

Conversely, by applying [6.6] to

$$f = \sum_{i=0}^{n-1} \alpha_i \mathbf{1}_{(t_i,\,t_{i+1}]},$$

we obtain

$$\mathbf{E}\left[\exp\left(-\sum_i \alpha_i(N(t_{i+1}) - N(t_i))\right)\right]$$
$$= \exp\left(-\lambda\sum_i \int_0^\infty \left(1 - e^{-\alpha_i \mathbf{1}_{(t_i,\,t_{i+1}]}(s)}\right)\mathrm{d}\,s\right).$$

Notice now that the function $s \to (1 - e^{-\alpha_i \mathbf{1}_{(t_i,\,t_{i+1}]}(s)})$ vanishes outside the interval $(t_i, t_{i+1}]$ and equals $1 - e^{-\alpha_i}$ on this interval. We thus have

$$\mathbf{E}\left[\exp\left(-\sum_i \alpha_i(N(t_{i+1}) - N(t_i))\right)\right] = \exp\left(-\sum_i \left(1 - e^{-\alpha_i}\right)\int_{t_i}^{t_{i+1}} \lambda(s)\,\mathrm{d}\,s\right).$$

We thus conclude that the Laplace transform of the random vector $(N(t_i + 1) - N(t_i),\ 1 \le i \le n-1)$ is the product of the Laplace transform of each component (as it is written as a product of functions that depends each only on one of the α_i's), so the random variables are independent. By a monotone class argument, we deduce that $N(t+s) - N(t)$ is independent from $\mathcal{F}_t = \sigma(N(r),\ r \le t)$. In the case where $n = 2,\ t_1 = a,\ t_2 = b$, the previous formula yields

$$\mathbf{E}\left[\exp\left(-\alpha\left(N(b) - N(a)\right)\right)\right] = \exp\left(-(1 - e^{-\alpha})\right)\int_a^b \lambda(s)\,\mathrm{d}\,s.$$

Hence, $N(b) - N(a)$ follows a Poisson distribution of parameter $\int_a^b \lambda(s)\,\mathrm{d}\,s$. Therefore,

$$\mathbf{E}\left[N(t+s) - N(t)\,|\,\mathcal{F}_t\right] = \mathbf{E}\left[N(t+s) - N(t)\right] = \int_t^{t+s} \lambda(s)\,\mathrm{d}\,s,$$

so $\tilde{N}$ is a martingale. □

6.6. Cox processes

The Cox processes are Poisson processes for which the intensity is random. From what is mentioned above, we see that this amounts to putting a probability measure on the set $\mathbb{D}$ of positive and rcll functions.

DEFINITION 6.8.– *Let M be a random variable with values in $\mathbb{D}$. The point process N is a Cox process of intensity M if, and only if, for any function f with compact support,*

$$\mathbf{E}\left[\exp\left(-\sum_n f(T_n)\right) \mid M\right] = \exp\left(-\int (1-e^{-f(s)})M(s)\,\mathrm{d}\,s\right).$$

Consequently, to derive the Laplace transform, it is necessary to un-condition with respect to M, thus

$$\mathbf{E}\left[\exp\left(-\sum_n f(T_n)\right)\right] = \mathbf{E}\left[\exp\left(-\int (1-e^{-f(s)})M(s)\,\mathrm{d}\,s\right)\right].$$

The previous martingale results can be extended without any change, but on the filtration, which must be taken as equal to $\mathcal{F}_0 = \sigma(M)$, $\mathcal{F}_t = \sigma(M)\vee\sigma(N(r),\, r\leq t)$.

The example of Cox process which we will be the most useful to us is that of MMPP; see section 7.6.

6.7. Problems

EXERCISE 10.– Let N be a Poisson process of intensity λ, we denote T_n as the nth instant of jump. By convention, $T_0 = 0$. Let $(Z_n, n \geq 1)$ be a sequence of random variables of same distribution such that, for any n, T_n and Z_n are independent. Let g be the density of the common distribution of the Z_n's.

1) Show that for any function f,

$$E[f(T_n, Z_n)] = \int_0^{+\infty}\int f(t,z)g(z)\lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}\,\mathrm{d}\,z\,\mathrm{d}\,t.$$

2) Deduce that

$$E\left[\sum_{n\geq 1} f(T_n, Z_n)\right] = \lambda\int_0^{+\infty}\int f(t,z)g(z)\,\mathrm{d}\,z\,\mathrm{d}\,t.$$

3) We assume that the telephonic communications of a subscriber lasts for a random time of exponential distribution of about three minutes in average. These durations are

independent of each other. In the last century, the cost of communication was based on its duration t according to the following formula.

$$c(t) = \alpha \text{ if } t \leq t_0, \text{ and } c(t) = \alpha + \beta(t - t_0) \text{ if } t \geq t_0.$$

Deduce from the above that the average cost of one complete hour of communication is given by

$$\lambda \int_0^1 c(t)\lambda e^{-\lambda t} \, \mathrm{d}\, t,$$

with $\lambda = 20$. (*Hint: Consider $Z_n = T_{n+1} - T_n$ and explain why we can apply the previous result*).

Numerical application: for local calls, in 1999, we had the following parameters: $t_0 = 3$ minutes, $\alpha = 0{,}11$ euro and $\beta = 0{,}043$ euro per minute. For national calls, $t_0 = 39$ seconds and $\beta = 0{,}17$ euro per minute. α was the same. For reduced price, divide β by 2. By applying $t_0 = 1$ minute and $\alpha = 0{,}15$ euro, how much is the price of the extra second in mobile telephony in a package, whose amount for 1 hour of communication was 23,62 euros?

EXERCISE 11.– An ATM records the beginning and ending times of queries of the customers, but of course not their arrival times in the queue. A new busy cycle having started at 7: 30, we have recorded the following:

Customer number	Beginning of service	End of service
0	7:30	7:34
1	7:34	7:40
2	7:40	7:42
3	7:45	7:50

Let us assume that the arrivals take place according to a Poisson process, what can we say about the arrival time T_1 of customer 1? In particular, give its mean expectation.

EXERCISE 12.– An insurance company must pay for claims at a rate of 5 per day. We assume that the instants of occurrence of disasters follow a Poisson process, and that the total amounts of damages are independent of each other, of exponential distribution with an average of 500 euros. We introduce $(X_i,\, i \geq 1)$, a sequence of i.i.d. random variables, of exponential distribution of average $1/\mu = 3{,}000$ euros, and a Poisson process N of intensity $\lambda = 5$ days^{-1}, independent of the X_i's.

1) What does $Z = \sum_{i=1}^{N(365)} X_i$ represent?

2) Calculate the average total annual amount spent by the insurance company.

3) Calculate $E[e^{-sZ}]$.

4) Infer the variance of Z.

6.8. Notes and comments

For more detailed results on the point processes in any dimension, we refer the reader to [LAS 95, BRA 81, DAL 03]. For the convergence in distribution of point processes, see [ROB 03, appendix D].

Epitome

– A Poisson process is a process that represents a random flow with an average flow that is constant over time.

– Its intensity λ represents the average number of points per unit of time.

– The superposition of two Poisson processes is a Poisson process. The thinning of a Poisson process form several Poisson processes.

– A Poisson process represents very well the events linked to human activity (telephone calls, arrival times in a store, logging session of a mail, opening of web page, etc.), but not the activity of a machine (sending of packets, signaling messages, etc.).

Chapter 7

Markov Process

The Markovian modeling of a dynamic system often leads to a Markov chain, for which the sojourn time in each state becomes random. In many cases, this description is insufficient to establish interesting mathematical properties. In that purpose, we introduce the formalism of Markov jump processes, with their semi-groups and infinitesimal generators.

To go one step further and, in particular, to prove several crucial results such as PASTA or its avatars, we need to see a Markov process as the solution of a martingale problem. In this chapter, we review these different characterizations, and show that they are in fact equivalent.

Throughout this chapter, E denotes a state space that is at the most countable, and equipped with the discrete topology. We refer the reader to the definitions and notations of Appendix A.1.

7.1. Preliminaries

We start by stating two technical Lemmas on the exponential distribution, which will be useful in the following.

LEMMA 7.1.– *Let U and V be the two independent random variables, of respective distributions $\varepsilon(\lambda)$ and $\varepsilon(\mu)$, where $\lambda, \mu > 0$. Then,*

(i) $\mathbf{P}(U \leq V) = \lambda/(\lambda + \mu)$*;*

(ii) $U \wedge V \sim \varepsilon(\lambda + \mu)$*.*

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

Proof.

(i) The density of the random couple (U, V) is given for all (u, v) by

$$f_{(U,V)}(u, v) = \lambda e^{-\lambda u} \mu e^{-\mu v} \, \mathbf{1}_{\mathbf{R}^+}(u) \, \mathbf{1}_{\mathbf{R}^+}(v).$$

By denoting the subset $A = \{(u, v) \in \mathbf{R}^2; u \leq v\}$, we can write

$$\begin{aligned}
\mathbf{P}(U \leq V) &= \int\int_A f_{(U,V)}(u, v) \, \mathrm{d}\, u \, \mathrm{d}\, v \\
&= \int_0^{+\infty} \lambda e^{-\lambda u} \left\{ \int_u^{+\infty} \mu e^{-\mu v} \, \mathrm{d}\, v \right\} \mathrm{d}\, u \\
&= \int_0^{+\infty} \lambda e^{-(\lambda+\mu) u} \, \mathrm{d}\, u \\
&= \frac{\lambda}{\lambda + \mu}.
\end{aligned}$$

(ii) It suffices to see that for any $x \geq 0$,

$$\mathbf{P}(U \wedge V \geq x) = \mathbf{P}(\{U \geq x\} \cap \{V \geq x\}) = e^{-\lambda x} e^{-\mu x} = e^{-(\lambda+\mu)x}.$$

Hence the result.

□

LEMMA 7.2 (Memoryless property of the exponential distribution).– *Let U be a random variable of distribution $\varepsilon(\lambda)$, where $\lambda > 0$, and $t \geq 0$. Then, conditionally to $\{U \geq t\}$, the random variable $U - t$ has the same distribution $\varepsilon(\lambda)$.*

Proof. It is sufficient to compute, for any $x \geq 0$, the conditional probability

$$\mathbf{P}(U \geq t + x \mid U \geq t) = \frac{\mathbf{P}(U \geq t + x)}{\mathbf{P}(U \geq t)} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}.$$

□

The life duration of a human being is not exponentially distributed. Indeed, if it was so the latter result would imply that the probability of reaching the age of 90 for a human of 80 years old equals the probability that a new-born reach the age of 10 years old! This is, of course, not the case.

7.2. Pathwise construction

DEFINITION 7.1.– *Let ν be a probability measure on E and $Q = \big(q(x,\, y),\, x,\, y \in E\big)$ be a family of real numbers such that*

$$q(x,\, y) \geq 0 \textit{ for any } (x,\, y) \in E \textit{ and } \sum_{y \neq x} q(x,\, y) = 1 \textit{ for any } x \in E. \qquad [7.1]$$

The Markov process X with parameters $(\nu,\, Q)$ is constructed as follows:

– $X(0)$ is a random variable of distribution ν;

– if $X(0) = x_0$, let $\xi_1 = T_1$ be a random variable independent of $X(0)$ and of distribution $\varepsilon(q(x_0,\, x_0))$. Then, we set

$$X(t) = x_0 \ \textit{ for all } \ t < T_1.$$

Then we let $\hat{X}_1$ be a random variable independent of $(X(0),\, T_1)$ and such that

$$\mathbf{P}(\hat{X}_1 = y) = q(x_0, y);$$

– if $\hat{X}_1 = x_1$, we let ξ_2 be a random variable independent of $(X(0),\, \xi_1,\, \hat{X}_1)$ and of distribution $\varepsilon(q(x_1,\, x_1))$. Then, we set

$$X(t) = x_1 \ \textit{ for all } \ T_1 \leq t < T_2.$$

– We continue this construction on and on.

NOTE.– A random variable of exponential distribution with parameter 0 must be understood as almost surely infinite. Hence any point $x \in E$ such that $q(x,\, x) = 0$ is a "graveyard" point: when it has been reached, the process never get out of it.

EXAMPLE 7.1 (M/M/1 queue).– In this model, arrivals occur according to a Poisson process of intensity $\lambda > 0$, and the service times are i.i.d. with exponential distribution of parameter μ. We consider the process $(X(t),\, t \geq 0)$ counting the number of customers in the system at any time. The state space E is that of natural integers.

If there are $i \neq 0$ customers in the system at a given time, the next event is a departure or an arrival. In view of Theorem 6.6, the next arrival will occur after a period of exponential distribution with parameters λ. In addition, Theorem 7.3 below shows us that the next departure will take place after a time exponentially distributed with parameter μ. From (ii) of Lemma 7.1, the sojourn time in state i then follows an exponential distribution with parameter $\lambda + \mu$, thus $q(i,\, i) = \lambda + \mu$ for $i \neq 0$. If $i = 0$, there cannot be any departure, so $q(0,\, 0) = \lambda$. Again, in view of Lemma 7.1 the probability of moving from state i to state $i + 1$ corresponds to the probability that a random variable of exponential distribution with parameter λ be less than a random variable of exponential distribution with parameter μ, thus

$$q(i,\, i+1) = \frac{\lambda}{\lambda + \mu} \ \text{ and } \ q(i,\, i-1) = \frac{\mu}{\lambda + \mu}.$$

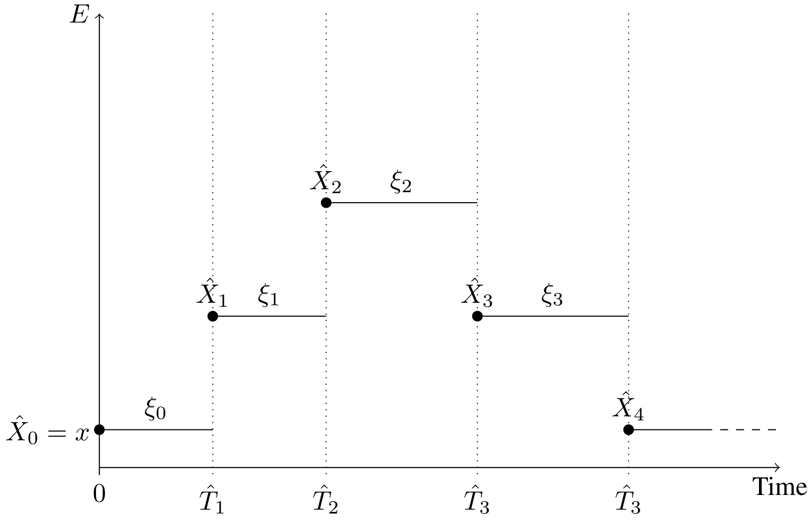

Figure 7.1. *The path of a Markov process with parameters* $(\nu,\, Q)$

If $i = 0$, we obtain

$$q(0,\, 1) = 1,$$

as the only possible move from state 0 is toward state 1.

Hence we have to prove our statement on the distribution of residual service times. It is not clear at this point, whether the departure process is Poisson, since there are no departures when the queue is empty. Hence we cannot *a priori* apply Theorem 6.6 in a straightforward way. However, within a busy period the departure times behave as a Poisson process. We formalize this idea in the following theorem.

THEOREM 7.3.– *F any $t \geq 0$, let $R(t)$ be the residual time at t before the next departure. Then, conditionally to $\{X(t) > 0\}$, the random variable $R(t)$ follows the distribution $\varepsilon(\mu)$.*

Proof. We denote $T'_0 < T'_1 < \cdots$ the successive departure times, and for any $t \geq 0$, $B_t \in \mathbf{N}^*$, the index of the last customer who entered an empty system before t. In other words, T_{B_t} represents the starting time of the last busy period started before t. From there, we mark the departure times of the server by a point process as follows

$$\tilde{T}^t_0 = T_{B_t};$$

$$\tilde{T}^t_k = T'_{B_t+k-1};\; k \in \mathbf{N}^*,$$

the k-th departure time of the server since the beginning of the last busy period started before t. We also denote for any $s \geq T_{B_t}$,

$$M_s^t = \sum_{k \in \mathbf{N}^*} \mathbf{1}_{\{\tilde{T}_k^t \leq s\}}$$

and for any $u \geq 0$,

$$\tilde{M}_u^t = M_{T_{B_t}+u}^t.$$

The notations are a bit complicated, but the idea is simple: the point process $\big(\tilde{M}_u^t,\, u \geq\ 0\big)$ marks the departure times since the beginning of the last busy period started before t, at which we set the origin of the time scale. The number of departures between the beginning of the busy period and t is given by $\tilde{M}_{t-T_{B_t}}^t$. On $\{X(t) > 0\}$, the system is never empty between T_{B_t} and t and therefore, for any $1 \leq k \leq \tilde{M}_{t-T_{B_t}}^t + 1$,

$$\tilde{T}_k^t - \tilde{T}_{k-1}^t = \sigma_{B_t+k-1},$$

the service time of the $k-1$th customer since the beginning of the busy period. These service times are independent of each other, independent of the past before T_{B_t}, and all of distribution $\varepsilon(\mu)$ for $k \leq \tilde{M}_{t-T_{B_t}}^t$. So, conditionally to $\{X(t) > 0\}$, the process $(\tilde{M}_u^t,\, u \geq 0)$ is equal in distribution on the interval $[0, \tilde{T}^t_{\tilde{M}^t_{t-T_{B_t}}}]$ to a a Poisson process with parameter μ, for which $\tilde{T}_0 < \tilde{T}_1 < \cdots$ represent the points and $\tilde{\xi}_0, \tilde{\xi}_1, \ldots$ the sizes of the inter-points. We can hence write that for any x,

$$\begin{aligned}&\mathbf{P}\left(R(t) \geq x \,|\, X(t) > 0\right)\\ &\qquad = \sum_{k \in \mathbf{N}} \mathbf{P}\left(\{\tilde{T}_k + \tilde{\xi}_k - (t - T_{B_t}) \geq x\} \cap \{\tilde{T}_k \leq t - T_{B_t} < \tilde{T}_{k+1}\}\right),\end{aligned}$$

and we can proceed as in the proof of Theorem 6.6, to conclude. □

Embedded Markov Chain

Given the independence assumptions, it is clear that the sequence $(\hat{X}_n, n \geq 0)$ is a Markov chain with transition matrix $\hat{Q}$ defined by

$$\hat{Q}(x,\, y) = \begin{cases} q(x,\, y) & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

The latter is called *embedded Markov chain* of the Markov process. There cannot be any transition from one state to itself since X is observed only when it changes states.

DEFINITION 7.2.– *A Markov process X is called irreducible (respectively recurrent, transient) if and only if the embedded chain $\hat{X}$ is irreducible (respectively recurrent, transient).*

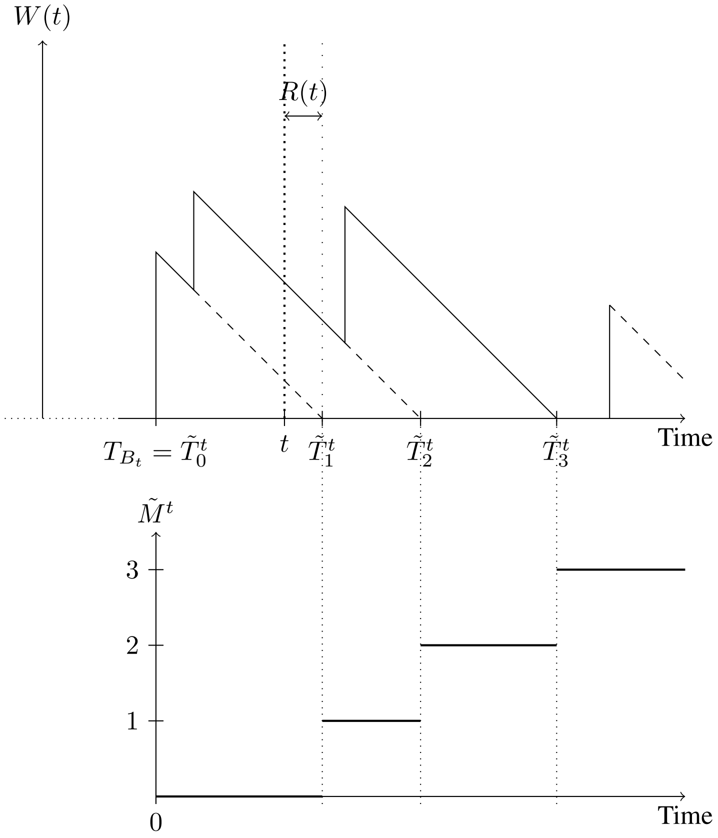

Figure 7.2. *Notations*

Equivalent construction

Another equivalent construction is possible. It is more artificial, however it leads to more simple proofs for several mathematical properties.

DEFINITION 7.3.– *A process X with parameters (ν, Q) is said regular if*

$$\|q\|_\infty = \sup_{x \in E} q(x,\, x) < \infty.$$

Let X be a regular Markov process with parameters (ν, Q). We set

$$\tilde{q}(x,\, y) = \begin{cases} \dfrac{q(x,\, x)}{\|q\|_\infty} q(x,\, y) & \text{if } x \neq y, \\ 1 - \dfrac{q(x,\, x)}{\|q\|_\infty} & \text{if } x = y. \end{cases}$$

Let $\tilde{X}$ be the Markov chain with initial distribution ν and transition matrix $\tilde{Q}$, and N be a Poisson process of intensity $\|q\|_\infty$, independent of $\tilde{X}$. The process $Y(t) = \tilde{X}_{N(t)}$ has same distribution as X. Indeed, the paths of this process may stay for several transitions in the same state. Let $x \in E$ and

$$\tau_{x^c} = \inf\{t > 0,\, Y(t) \neq x\};$$

$$\tilde{\tau}_{x^c} = \inf\{n > 0,\, \tilde{X}_n \neq x\}.$$

Conditionally to $\tilde{X}(0) = x$, $\tilde{\tau}_{x^c}$ is independent of N and follows a geometric distribution with parameter $q(x,\, x)/\|q\|_\infty$. But

$$\mathbf{P}(\tau_{x^c} \geq t \,|\, Y(0) = x) = \mathbf{P}\left(\sum_{j=1}^{\tilde{\tau}_{x^c}} \xi_j \geq t\right),$$

where the ξ_j's are the inter-points of N, so the random variables are independent of each other and independent of $\tilde{\tau}_{x^c}$, and of exponential distribution of parameter $\|q\|_\infty$. From Lemma 7.1, we deduce that the sojourn time in the state x follows an exponential distribution with parameter $q(x,\, x)$. In addition, when $\tilde{X}$ jumps, we observe that

$$\mathbf{P}\left(\tilde{X}_1 = y \,|\, \tilde{X}_1 \neq \tilde{X}_0 = x\right) = \frac{\tilde{q}(x,\, y)}{1 - \tilde{q}(x,\, x)} = q(x,\, y) = \mathbf{P}\left(\hat{X}_1 = y \,|\, \hat{X}_0 = x\right).$$

NOTE.– We also deduce from this construction that at any fixed s, there is no jump almost surely, i.e. $\mathbf{P}(\Delta X(s) > 0) = 0$. Indeed,

$$\begin{aligned}\mathbf{P}(\Delta X(s) > 0 = 0) &\leq \mathbf{P}(\Delta N(s) = 1)\\ &= \mathbf{E}\left[\int \mathbf{1}_s(x)\,\mathrm{d}\,N(x)\right] = \int \mathbf{1}_s(x)\|q\|_\infty\,\mathrm{d}\,x = 0.\end{aligned}$$

Therefore, $\mathbf{P} \otimes \mathrm{d}\, s$-almost surely, $X(s) = X(s^-)$. Indeed, the Lebesgue measure does not see the jumps since they are in quantity at the most countable.

7.3. Markovian semi-group and infinitesimal generator

DEFINITION 7.4.– *Let X be a process with values in E and with rcll (right continuous with left-hand limits) paths, and let $\mathcal{F}_t = \sigma\{X_u,\, u \leq t\}$. The process X satisfies the (simple) Markov property if for any $t,\, s \geq 0$, we have*

$$\mathbf{E}\left[f(X(t+s)) \,|\, \mathcal{F}_t\right] = \mathbf{E}\left[f(X(t+s)) \,|\, X(t)\right]. \qquad [7.2]$$

The process X is called homogeneous when for any $t \geq 0$, for any $x \in E$,

$$\mathbf{E}\left[f(X(t+s)) \,|\, X(t) = x\right] = \mathbf{E}\left[f(X(s)) \,|\, X(0) = x\right]. \qquad [7.3]$$

Let X be a process with rcll paths satisfying [7.2] and [7.3]. For all $f \in l^\infty(E)$ and all $x,\, y \in E$, we set

$$P_t(x,\, y) = \mathbf{P}(X(t) = y \,|\, X(0) = x);$$

$$P_t f(x) = \sum_{y \in E} f(y) P_t(x,\, y) = \mathbf{E}\left[f(X(t)) \,|\, X(0) = x\right].$$

THEOREM 7.4 (KOLMOGOROV'S EQUATION).– *For any $t \geq 0$, P_t is continuous from $l^\infty(E)$ into itself. Moreover, for any $t,\, s \geq 0$, $P_{t+s} = P_t P_s = P_s P_t$.*

NOTE.– The latter is an identity between operators, that is for any $f \in l^\infty(E)$ and any $x \in E$,

$$P_{t+s} f(x) = P_t(P_s f)(x),$$

or between matrices (although the notion of matrix of infinite size remains unclear):

$$P_{t+s}(x,\, y) = \sum_{z \in E} P_t(x,\, z) P_s(z,\, y).$$

The family $(P_t,\, t \geq 0)$ is then called a *semi-group* of operators: the stability property for "$\circ$" is similar to that of a group, but each element of the family does not necessarily admit a symmetric element for "$\circ$".

Proof of Theorem 7.4. Fix $t \geq 0$. First, from the definition of P_t we have

$$P_t \mathbf{1} = \mathbf{1} \text{ and } |P_t f| \leq P_t |f|.$$

Moreover, according to the properties of conditional expectation, $P_t f \geq 0$ for any $f \geq 0$. So in particular, $f \leq g$ implies $P_t f \leq P_t g$. Therefore if f is bounded, so is $P_t f$

$$|P_t f(x)| \leq P_t |f|(x) \leq P_t(\|f\|_\infty \mathbf{1})(x) = \|f\|_\infty P_t \mathbf{1}(x) = \|f\|_\infty.$$

Hence, P_t is continuous from $l^\infty(E)$ to itself. Moreover,

$$\sigma(X(0)) \vee \mathcal{F}_t = \mathcal{F}_t,$$

and according to the interlocking property of conditional expectations, we have for all $s \geq 0$ that

$$\begin{aligned} P_{t+s} f(x) &= \mathbf{E}\left[f(X(t+s)) \,|\, X(0) = x\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[f(X(t+s)) \,|\, \mathcal{F}_t\right] \,|\, X(0) = x\right] \\ &= \mathbf{E}\left[(P_s f)(X(t) \,|\, X(0) = x\right] \\ &= P_t(P_s f)(x). \end{aligned}$$

The second identity is obtained by conditioning to $\mathcal{F}_s$. □

Notice that by dominated convergence, $P_t f(x)$ tends toward $f(x)$ for any $x \in E$, since the process X has rcll paths.

DEFINITION 7.5.– *We say that a semi-group P is strongly continuous if for any $f \in l^\infty(E)$,*

$$\lim_{t \to 0} \sup_{x \in E} |P_t f(x) - f(x)| = 0.$$

DEFINITION 7.6.– *Let P be a strongly continuous semi-group. Let Dom A be the set of functions f for which for all $x \in E$, $t^{-1}(P_t f(x) - f(x))$ has a limit when t tends to 0. The infinitesimal generator A of the semi-group P is defined on Dom A and for all $x \in E$ by*

$$Af(x) = \lim_{t \to 0} \frac{1}{t}(P_t f(x) - f(x)). \qquad [7.4]$$

THEOREM 7.5.– *Let P be a strongly continuous semi-group of infinitesimal generator A.*

(a) For all $f \in$ Dom A and $t \geq 0$, the function

$$x \mapsto \int_0^t P_s f(x) \, \mathrm{d}\, s$$

belongs to Dom A and

$$P_t f - f = A \int_0^t P_s f \, \mathrm{d}\, s. \qquad [7.5]$$

(b) For $f \in$ Dom A and $t \geq 0$, the function $P_t f$ belongs to Dom A and

$$\frac{d}{dt} P_t f = AP_t f = P_t Af. \qquad [7.6]$$

(c) For $f \in$ Dom A and $t \geq 0$, the following identity holds.

$$P_t f - f = \int_0^t AP_s f \, \mathrm{d}\, s = \int_0^t P_s Af \, \mathrm{d}\, s. \qquad [7.7]$$

Proof. Fix $t \geq 0$. By definition of A, we must show that

$$\frac{1}{h}(P_h - \mathrm{Id}) \left(\int_0^t P_s f(x) \, \mathrm{d}\, s \right) \text{ converges as } h \to 0.$$

But

$$\begin{aligned}(P_h - \mathrm{Id})\left(\int_0^t P_s f(x)\,\mathrm{d}\,s\right) &= \int_0^t P_h P_s f(x)\,\mathrm{d}\,s - \int_0^t P_s f(x)\,\mathrm{d}\,s \\ &= \int_0^t P_{s+h} f(x)\,\mathrm{d}\,s - \int_0^t P_s f(x)\,\mathrm{d}\,s \\ &= \int_h^{t+h} P_s f(x)\,\mathrm{d}\,s - \int_0^t P_s f(x)\,\mathrm{d}\,s \\ &= \int_t^{t+h} P_s f(x)\,\mathrm{d}\,s - \int_0^h P_s f(x)\,\mathrm{d}\,s.\end{aligned}$$

As P is a strongly continuous semi-group, the continuity at 0 implies that at any point s, that is to say that for all $f \in l^\infty(E)$, the mapping $(s \mapsto P_s f)$ is continuous from $\mathbf{R}^+$ to $l^\infty(E)$. We thus have the following limit.

$$\lim_{h\to 0} \frac{1}{h}\left(\int_t^{t+h} P_s f(x)\,\mathrm{d}\,s - \int_0^h P_s f(x)\,\mathrm{d}\,s\right) = P_t f(x) - f(x).$$

For $h > 0$, we set $A_h f = h^{-1}(P_h f - f)$. It is easily seen that

$$A_h P_t f = h^{-1}(P_{t+h} f - P_t f) = P_t(h^{-1}(P_h f - f)) = P_t A_h f.$$

As $f \in \mathrm{Dom}\,A$, $A_h f$ tends to Af as h tends to 0, therefore $A_h P_t f$ also converges, which amounts to say that $P_t f$ belongs to $\mathrm{Dom}\,A$ and that [7.6] holds. Finally, the identity [7.7] is an immediate consequence of [7.5] and [7.6]. □

This can be interpreted very easily. Informally, [7.6] implies that

$$P_t f = \exp(tA) f.$$

The sense of the latter expression is well-known if A is a matrix. The fact that A can be an operator (a matrix of "infinite size") requires some mathematical adjustment, but the key point is there. Written in this way, it becomes obvious, for instance, that A and P_t commute (see [7.6]) and that

$$A^{-1} f = \int_0^\infty P_t f\,\mathrm{d}\,t,$$

keeping in mind the value of $\int_0^\infty \exp(at)\,\mathrm{d}\,t$ when a is negative.

THEOREM 7.6.– *Let X be a process with values in E, with rcll trajectories and satisfying properties [7.2] and [7.3]. The pair $(\nu,\, A)$ completely determines the distribution of X. Particularly, for any integer $n \geq 1$, for any bounded functions $f_j,\, j = 1, \dots, n$ defined on E and for any $t_1 < \dots < t_n$,*

$$\begin{aligned}\mathbf{E}\left[\prod_{j=1}^{n} f_j(X(t_j))\right] & \\ = \int_E P_{t_1}(f_1 P_{t_2-t_1}(f_2 \dots P_{t_{n-1}-t_{n-2}}(f_{n-1} P_{t_n - t_{n-1}} f_n)) \dots)(x) \, \mathrm{d}\, \nu(x).\end{aligned} \qquad [7.8]$$

Proof. In view of Theorem 7.5, P fully determines A. Assume for a while that [7.8] holds. By applying it to $f_j = \mathbf{1}_{\mathcal{A}_j}$, where $\mathcal{A}_j$ is any subset of E, we see that the term on the left-hand side equals

$$\mathbf{P}(X(t_1) \in \mathcal{A}_1, \dots, X(t_n) \in \mathcal{A}_n).$$

Finite-dimensional distributions of X are hence completely characterized by $(\nu,\, A)$. According to the extension Theorem, we deduce that the distribution of X is fully determined by $(\nu,\, A)$. It remains to show [7.8]. For $n = 1$, this is the very definition of P. Let us assume that the result holds true for some $n \geq 1$. By conditioning and according to the definition of P, we have

$$\begin{aligned}\mathbf{E}\left[\prod_{j=1}^{n+1} f_j(X(t_j))\right] &= \mathbf{E}\left[\prod_{j=1}^{n} f_j(X(t_j)) \mathbf{E}\left[f_{n+1}(X(t_{n+1})) \,|\, \mathcal{F}_{t_n}\right]\right] \\ &= \mathbf{E}\left[f_1(X(t_1)) \dots f_{n-1}(X(t_{n-1}))(f_n P_{t_{n+1}-t_n} f_{n+1})(X(t_n))\right].\end{aligned}$$

As $f_n P_{t_{n+1}-t_n} f_{n+1}$ is bounded, the result follows by induction. □

THEOREM 7.7.– *Let X be a process with values in E, having rcll paths and satisfying the properties [7.2] and [7.3]. The process X satisfies the strong Markov property: for any stopping time τ, for any bounded function $F \colon D(\mathbf{R}^+,\, E) \to \mathbf{R}$,*

$$\mathbf{E}\left[F(\theta_\tau X) \,|\, \mathcal{F}_\tau\right] = \mathbf{E}\left[F(X) \,|\, X(0) = X(\tau)\right]. \qquad [7.9]$$

Proof. Let us assume at first that τ takes its values in the countable set $\mathfrak{T} = \{t_j,\, j \geq 1\}$. The events $\{\tau \leq t\}$ and $\{\tau > t\}$ belong, by definition of a stopping time, to $\mathcal{F}_t$. Hence,

$$\{\tau = t\} = \left(\{\tau \leq t\} \cap \bigcap_{s<t,\, s \in \mathfrak{T}} \{\tau > s\}\right) \in \mathcal{F}_t.$$

We can then follow the proof of the strong Markov property for Markov chains (see (3.6)), replacing n by t_n, to obtain that for any function $f \in l^\infty(E)$,

$$\mathbf{E}\left[f(X(\tau+s)) \,|\, \mathcal{F}_\tau\right] = \mathbf{E}\left[f(X(\tau+s) \,|\, X(\tau)\right] = P_s f(X(\tau)). \qquad [7.10]$$

Now, for τ an arbitrary stopping time, we consider the sequence of stopping times $\{\tau_n,\, n \geq 1\}$ defined by

$$\tau_n = \sum_{k=0}^{\infty} \frac{k+1}{2^n} \mathbf{1}_{[k2^{-n},\,(k+1)2^{-n}]}(\tau).$$

This sequence converges decreasingly toward τ. As X has right-continuous paths, $X(\tau_n + s)$ tends a.s. to $X(\tau+s)$. Therefore, for any $\mathcal{A} \in \mathcal{F}_\tau \subset \mathcal{F}_{\tau_n}$, by dominated convergence we have that

$$\begin{aligned}\mathbf{E}\left[f(X(\tau+s))\,\mathbf{1}_{\mathcal{A}}\right] &= \lim_{n\to\infty} \mathbf{E}\left[f(X(\tau_n+s))\,\mathbf{1}_{\mathcal{A}}\right] \\ &= \lim_{n\to\infty} \mathbf{E}\left[P_s f(X(\tau_n))\,\mathbf{1}_{\mathcal{A}}\right] = \mathbf{E}\left[P_s f(X(\tau))\,\mathbf{1}_{\mathcal{A}}\right].\end{aligned}$$

Consequently, [7.10] remains true for any stopping time τ.

By successive conditioning, for $0 < s_1 < \cdots < s_k$, we therefore have that

$$\begin{aligned}\mathbf{E}\left[\prod_{j=1}^{k} f_j(X(\tau+s_j)) \,|\, \mathcal{F}_\tau\right] &= (P_{s_k - s_{k-1}} f_k \ldots P_{s_1} f_1)(X(\tau)) \\ &= \mathbf{E}\left[\prod_{j=1}^{k} f_j(X(\tau+s_j)) \,|\, X(\tau)\right].\end{aligned}$$

But in view of [7.8], we also have

$$\mathbf{E}\left[\prod_{j=1}^{k} f_j(X(s_j)) \,|\, X(0) = x\right] = (P_{s_k - s_{k-1}} f_k \ldots P_{s_1} f_1)(x).$$

Therefore,

$$\mathbf{E}\left[F\circ\theta_\tau \,|\, X(\tau)=x\right]=\mathbf{E}\left[F\,|\,X(0)=x\right],$$

for all functions F of the form $\prod_j f_j$. By monotone class, this result remains true for all bounded functions from $D(\mathbf{R}^+,\,E)$ to $\mathbf{R}$, hence [7.9]. □

DEFINITION 7.7.– *Let X be a regular Markov process of parameters $(\nu,\,Q)$. For $f\in l^\infty(E)$, we set*

$$A_Q f(x)=q(x,\,x)\sum_{y\neq x}(f(y)-f(x))q(x,\,y).$$

By identifying f as the column vector $(f(x), x\in E)$ (after rearranging the elements of E, which is possible since there is an injection from E into $\mathbf{N}$), we can rewrite the previous identity as a matrix product, by introducing the matrix A_Q, defined by

$$A_Q(x,\,y)=\begin{cases}-q(x,\,x) & \textit{if } x=y;\\ q(x,\,y)q(x,\,x) & \textit{if } x\neq y.\end{cases}\qquad [7.11]$$

Notice in particular that $A_Q(x,\,y)=(A_Q\,\mathbf{1}_y)(x)$ for any $x,\,y$.

THEOREM 7.8.– *Let X be a regular Markov process with parameters $(\nu,\,Q)$. The process X satisfies the simple Markov property [7.2] and the homogeneity property [7.3]. The associated semi-group is strongly continuous. Its infinitesimal generator is A_Q, and its domain is $l^\infty(E)$.*

Proof. Let us start from the second pathwise construction of X. In this case, the knowledge of $\mathcal{F}_t$ amounts in particular to that of the number of jumps of N before t and the value of X after the last jump of N before t. By the very construction of the paths of X, these two quantities are the only ones that are useful for determining the sequel of the trajectory. Therefore, we have that

$$\begin{aligned}\mathbf{E}\left[f(X(t+s))\,|\,\mathcal{F}_t\right]&=\mathbf{E}\left[f(X(t+s))\,|\,N(t),\,\tilde{X}_{N(t)}\right]\\&=\sum_{n\geq 0}\mathbf{E}\left[f(X(t+s))\,\mathbf{1}_{\{N(t+s)-N(t)=n\}}\,|\,N(t),\,\tilde{X}_{N(t)}\right]\\&=\sum_{n\geq 0}\mathbf{E}\left[f(\tilde{X}_{N(t)+n})\,\mathbf{1}_{\{N(t+s)-N(t)=n\}}\,|\,N(t),\,\tilde{X}_{N(t)}\right].\end{aligned}$$

As $\tilde{X}$ and N are independent and N has independent increments, we have

$$\mathbf{E}\left[f(X(t+s))\,|\,\mathcal{F}_t\right]=\sum_{n\geq 0}\tilde{Q}^{(n)}f(\tilde{X}_{N(t)})e^{-\|q\|_\infty\,s}\frac{(\|q\|_\infty s)^n}{n!}.$$

Moreover,

$$\begin{aligned}
&\mathbf{E}\left[f(X(t+s))\,|\,X(t)\right]\\
&\quad=\sum_{n\geq 0}\mathbf{E}\left[\mathbf{E}\left[f(X(t+s))\,\mathbf{1}_{\{N(t+s)-N(t)=n\}}\,|\,N(t),\,X(t)\right]\,|\,X(t)\right]\\
&\quad=\sum_{n\geq 0}\mathbf{E}\left[\mathbf{E}\left[f(\tilde{X}_{N(t)+n})\,\mathbf{1}_{\{N(t+s)-N(t)=n\}}\,|\,N(t),\,\tilde{X}_{N(t)}\right]\,|\,X(t)\right]\\
&\quad=\mathbf{E}\left[\sum_{n\geq 0}\tilde{Q}^{(n)}f(\tilde{X}_{N(t)})e^{-\|q\|_\infty s}\frac{(\|q\|_\infty s)^n}{n!}\,|\,X(t)\right]\\
&\quad=\sum_{n\geq 0}\tilde{Q}^{(n)}f(\tilde{X}_{N(t)})e^{-\|q\|_\infty s}\frac{(\|q\|_\infty s)^n}{n!},
\end{aligned}$$

in view of the first part of the proof, and the fact that $X(t)=\tilde{X}_{N(t)}$. The simple Markov property is thus satisfied. We can also deduce from the last equation, that

$$\mathbf{E}\left[f(X(t+s))\,|\,X(t)=x\right]=\sum_{n\geq 0}\tilde{Q}^{(n)}f(x)e^{-\|q\|_\infty s}\frac{(\|q\|_\infty s)^n}{n!}.$$

The term on the right-hand side does not depend on t, so the homogeneous property holds as well. □

THEOREM 7.9.– *Let X be a process with values in the E, satisfying [7.2] and [7.3]. Let ν be the distribution of $X(0)$, and A its infinitesimal generator. Then the process X is a Markov process with parameters $(\nu,\,Q_A)$, where*

$$Q_A(x,\,y)=\begin{cases}|A(x,\,x)| & \textit{if } y=x;\\[2ex] \dfrac{A(x,\,y)}{|A(x,\,x)|} & \textit{if } y\neq x.\end{cases}\qquad [7.12]$$

Proof. Let us set $T_0=0$ and for any integer n,

$$\begin{cases}T_{n+1}=\inf\{t>T_n,\,X(t)\neq X(T_n)\}\\ \xi_n=T_{n+1}-T_n,\\ \hat{X}_n=X(T_n),\end{cases}$$

with the usual convention $\inf\emptyset=\infty$.

Let $x\in E$ and for any $u\geq 0$, $g(u)=\mathbf{P}(T_1>u\,|\,X_0=x)$. Let us show that g is a solution of the functional equation characteristic of the exponential function.

According to [7.2] and [7.3], we have

$$\begin{aligned} g(u+v) &= \mathbf{P}(T_1 > u+v \,|\, X_0 = x) \\ &= \mathbf{E}\left[\mathbf{1}_{\{X(s)=x,\, s\in[0,u]\}}\,\mathbf{1}_{\{X(t)=x,\, t\in[u,u+v]\}}\right] \\ &= \mathbf{E}\left[\mathbf{1}_{\{X(s)=x,\, s\in[0,u]\}}\,\mathbf{E}\left[\mathbf{1}_{\{X(t)=x,\, t\in[u,u+v]\}} \,|\, \mathcal{F}_u\right]\right] \\ &= \mathbf{E}\left[\mathbf{1}_{\{X(s)=x,\, s\in[0,u]\}}\,\mathbf{E}\left[\mathbf{1}_{\{X(t)=x,\, t\in[0,v]\}} \,|\, X_0 = x\right]\right] \\ &= \mathbf{E}\left[\mathbf{1}_{\{X(s)=x,\, s\in[0,u]\}}\right] g(v) \\ &= g(u)g(v). \end{aligned}$$

As g is bounded, we deduce from this the existence of a $q(x) \geq 0$, such that

$$g(u) = \exp(-q(x)u).$$

The sojourn time in the initial state thus follows an exponential distribution.

By definition of a stopping time, the event $\{T_1 > u\}$ belongs to $\mathcal{F}_u$. Moreover, on $\{T_1 > u\}$, $X(u) = X(0)$ and thus

$$\begin{aligned} \mathbf{P}(\hat{X}_1 = y,\, T_1 > u \,|\, X(0) = x) &= \mathbf{E}_x\left[\mathbf{1}_{[u,\,\infty)}(T_1)\mathbf{E}_x\left[\mathbf{1}_{\{y\}}(\hat{X}_1) \,|\, \mathcal{F}_u\right]\right] \\ &= \mathbf{E}_x\left[\mathbf{1}_{[u,\,\infty)}(T_1)\mathbf{E}_x\left[\mathbf{1}_{\{y\}}(\hat{X}_1) \,|\, X_0 = x\right]\right]. \end{aligned}$$

As the quantity $\mathbf{P}(\hat{X}_1 = y \,|\, X_0 = x)$ is deterministic, it goes off the expectation. Therefore, setting

$$\mathbf{P}(\hat{X}_1 = y \,|\, X(0) = x) = q(x,\, y),$$

we can write

$$\mathbf{P}(\hat{X}_1 = y,\, T_1 > u \,|\, X(0) = x) = q(x,\, y)\exp\left(-(q(x)u)\right). \qquad [7.13]$$

We deduce from [7.13] that conditionally to $X(0) = \hat{X}_0$, $\hat{X}_1$ and T_1 are independent. Hence, we have obtained what we aimed for, at least until the first jump of X: a sojourn time of exponential distribution of parameter depending on the initial state, then a choice of the new state independently of the sojourn time.

Let us now assume that for a given n, for $j \leq n-1$, for any $y \in E$ and any $u \geq 0$, the following identity holds.

$$\begin{aligned} &\mathbf{P}(\hat{X}_{j+1} = y,\, \xi_j > u \,|\, \hat{X}_0,\, \ldots,\, \hat{X}_j,\, T_1,\, \ldots,\, T_j) \\ &\qquad = q(\hat{X}_j,\, y)\, \exp(-q(\hat{X}_j,\, \hat{X}_j)u). \end{aligned} \qquad [7.14]$$

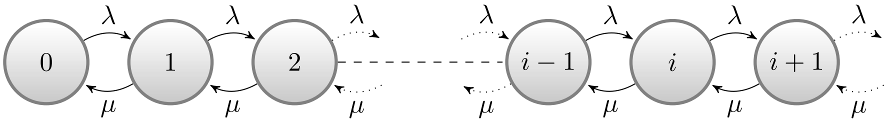

Figure 7.3. *Graphic representation of the transition rates of the M/M/1 queue*

Given that the knowledge of $\hat{X}_0, \ldots, \hat{X}_n, T_1, \ldots, T_n$ amounts to that of the whole past of the process until T_n, we have

$$\begin{aligned}&\mathbf{P}(\hat{X}_{n+1} = y,\, \xi_n > u \,|\, \hat{X}_0, \ldots, \hat{X}_n, T_1, \ldots, T_n)\\ &\qquad = \mathbf{P}(\hat{X}_{n+1} = y,\, \xi_n > u \,|\, \mathcal{F}_{T_n}).\end{aligned}$$

According to the strong Markov property (see Theorem 7.7), the latter quantity can be transformed as follows.

$$\begin{aligned}\mathbf{P}(\hat{X}_{n+1} = y,\, \xi_n > u \,|\, \mathcal{F}_{T_n}) &= \mathbf{P}(\hat{X}_{n+1} = y,\, \xi_n > u \,|\, \hat{X}_n)\\ &= \mathbf{P}_{\hat{X}_n}(\hat{X}_1 = y,\, \xi_0 > u)\\ &= q(\hat{X}_n, y)\, \exp(-q(\hat{X}_n, \hat{X}_n)u),\end{aligned}$$

from [7.13]. Relation [7.14] is verified at rank n. In means in particular that we can construct the trajectories of X as in Definition 7.1. By identification, [7.12] follows from [7.11] and Theorem 7.8. □

EXAMPLE (Example 7.1 continued: M/M/1 queue).– In view of [7.11] and the results of Example 7.1, the infinitesimal generator of the process counting the number of customers in the M/M/1 queue is given for any $i \in \mathbf{N}$ by

$$\begin{cases} A(i,\, i+1) = \lambda,\\ A(i,\, i) = -(\lambda + \mu\, \mathbf{1}_{[1,\,+\infty)}(i)),\\ A(i,\, i-1) = \mu \text{ if } i > 0.\end{cases}$$

We often prefer the following matrix representation:

$$A = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & (0) & \\ & & \ddots & & \\ & (0) & \mu & -(\lambda+\mu) & \lambda \\ & & & & \ddots \end{pmatrix},$$

or the graphic representation of Figure 7.3.

THEOREM 7.10.– *Let X be a regular Markov process of parameters $(\nu,\, Q)$. We denote $\hat{X}$, its embedded chain. The following properties are equivalent:*

(i) $\hat{X}$ is irreducible;

(ii) For any $x,\, y \in E$, there exists $t > 0$ such that $p_t(x,\, y) > 0$;

(iii) For any $x,\, y \in E$, for any $t > 0$, $p_t(x,\, y) > 0$.

Proof. In view of the second pathwise construction of X, the chain $\tilde{X}$ is independent of N, hence

$$\begin{aligned}
p_t(x,\, y) &= \mathbf{P}(X(t) = y \,|\, X(0) = x)\\
&= \sum_{n=0}^{\infty} \mathbf{P}(X(t) = y,\, N(t) = n \,|\, X(0) = x)\\
&= \sum_{n=0}^{\infty} \mathbf{P}(N(t) = n)\mathbf{P}(\tilde{X}_n = y \,|\, \tilde{X}_0 = x)\\
&= \sum_{n=0}^{\infty} e^{-\|q\|_\infty t} \frac{(\|q\|_\infty t)^n}{n!} \tilde{Q}^{(n)}(x,\, y).
\end{aligned}$$

By the definition of irreducibility, the equivalence becomes straightforward. □

Invariant probability

DEFINITION 7.8.– *A measure μ on E is said invariant for the Markov process X, if $X(0) \sim \mu$ implies that $X(t) \sim \mu$ for any $t \geq 0$.*

THEOREM 7.11.– *Let X be a regular Markov process with parameters $(\nu,\, Q)$. A measure μ is invariant, if and only if it satisfies the equations*

$$\int A f \,\mathrm{d}\,\mu = 0, \text{ for any } f \in l^\infty(E). \qquad [7.15]$$

In matrix notation, the latter amounts to $\mu A = \mathbf{0}$, where μ is the row vector $(\mu(x), x \in E)$ and $\mathbf{0} = (0, 0, ...)$.

Proof. Fix $t \geq 0$. That $X(t)$ and $X(0)$ have the same distribution, amounts to

$$\mathbf{E}\,[f(X(t))] = \mathbf{E}\,[f(X(0))],$$

for any $f \in l^\infty(E)$. But

$$\mathbf{E}\,[f(X(t))] = \mathbf{E}\,[\mathbf{E}\,[f(X(t) \,|\, X(0)]] = \mathbf{E}\,[P_t f(X(0))] = \int_E P_t f(x)\,\mathrm{d}\,\mu(x).$$

So μ is an invariant measure if and only if

$$\int_E P_t f(x)\,\mathrm{d}\,\mu(x) = \int_E f(x)\,\mathrm{d}\,\mu(x).$$

Differentiating the latter equation, we deduce [7.15]. Conversely, if [7.15] holds, then in view of [7.7] and Fubini's Theorem,

$$\begin{aligned}\int_E (P_t f(x) - f(x))\,\mathrm{d}\,\mu(x) &= \int_E \int_0^t P_s A f(x)\,\mathrm{d}\,s\,\mathrm{d}\,\mu(x)\\ &= \int_0^t \left(\int_E P_s A f(x)\,\mathrm{d}\,\mu(x)\right)\mathrm{d}\,s\\ &= 0.\end{aligned}$$

As E is discrete, [7.15] can be rewritten as

$$\sum_{x\in E}\left(\sum_{z\in E} A(x,\,z) f(z)\right)\mu(x) = 0,$$

which, by taking $f = \mathbf{1}_y$, yields to

$$\sum_{x\in E}\mu(x)A(x,\,y) = 0 \text{ for all } y \in E.$$

In the matrix language, this exactly means that the product of the row vector μ by the matrix A is zero. □

Similarly to the discrete case, let us set

$$\tau_x^1 = \inf\left\{t > 0,\; X(t) = x\right\},$$

with the convention $\tau_x^1 = \infty$, if $X(t) \neq x$ for all $t > 0$.

THEOREM 7.12.– *Let X be a regular, irreducible and recurrent Markov process. There exists a unique invariant measure up to a multiplicative factor. This measure is proportional to one of the following three measures*:

(i) for any $y \in E$,

$$\mu(y) = \mathbf{E}_x \left[\int_0^{\tau_x^1} \mathbf{1}_{\{X(s)=y\}} \, \mathrm{d}\, s \right], \qquad [7.16]$$

where x is an arbitrary fixed element of E;

(ii) for any $y \in E$,

$$\mu(y) = \hat{\mu}(y)/q(y,\, y),$$

where $\hat{\mu}$ is an invariant measure of the embedded chain $\hat{X}$;

(iii) a solution μ to the matrix equation $\mu\, A = \mathbf{0}$.

Proof. To prove the item (i), we have to check that for any $t > 0$, for any $f \in l^\infty(E)$,

$$\int_E P_t f(x) \, \mathrm{d}\, \mu(x) = \int_E f(x) \, \mathrm{d}\, \mu(x). \qquad [7.17]$$

But according to [7.16], we have that

$$\begin{aligned}
\int_E P_t f(y) \, \mathrm{d}\, \mu(y) &= \sum_{y \in E} \mathbf{E}_x \left[\int_0^{\tau_x^1} P_t f(y)\, \mathbf{1}_{\{X(s)=y\}} \, \mathrm{d}\, s \right] \\
&= \sum_{y \in E} \mathbf{E}_x \left[\int_0^{\infty} P_t f(X(s))\, \mathbf{1}_{\{X(s)=y\}}\, \mathbf{1}_{\{s<\tau_x^1\}} \, \mathrm{d}\, s \right] \\
&= \mathbf{E}_x \left[\int_0^{\infty} \mathbf{E}\left[f(X(s+t)) \,|\, \mathcal{F}_s \right] \mathbf{1}_{\{s<\tau_x^1\}} \, \mathrm{d}\, s \right] \\
&= \int_0^{\infty} \mathbf{E}_x \left[\mathbf{E}\left[f(X(s+t))\, \mathbf{1}_{\{s<\tau_x^1\}} \,|\, \mathcal{F}_s \right] \right] \mathrm{d}\, s \\
&= \int_0^{\infty} \mathbf{E}_x \left[f(X(s+t))\, \mathbf{1}_{\{s<\tau_x^1\}} \right] \mathrm{d}\, s \\
&= \mathbf{E}_x \left[\int_t^{\tau_x^1 + t} f(X(s)) \, \mathrm{d}\, s \right] \\
&= \mathbf{E}_x \left[\int_t^{\tau_x^1} f(X(s)) \, \mathrm{d}\, s \right] + \mathbf{E}_x \left[\int_{\tau_x^1}^{\tau_x^1 + t} f(X(s)) \, \mathrm{d}\, s \right].
\end{aligned}$$

In view of Markov property and noticing that, by definition of τ_x^1, the right-continuity of X implies that $X\left(\tau_x^1\right)=x$, we obtain that

$$\begin{aligned}\int_E P_t f(y)\,\mathrm{d}\,\mu(y) &= \mathbf{E}_x\left[\int_t^{\tau_x^1} f(X(s))\,\mathrm{d}\,s\right]+\mathbf{E}_x\left[\int_0^t f(X(s))\,\mathrm{d}\,s\right]\\ &= \mathbf{E}_x\left[\int_0^{\tau_x^1} f(X(s))\,\mathrm{d}\,s\right]\\ &= \sum_{y\in E}\mathbf{E}_x\left[\int_0^{\tau_x^1} f(X(s))\,\mathbf{1}_{\{X(s)=y\}}\,\mathrm{d}\,s\right]\\ &= \sum_{y\in E}\mathbf{E}_x\left[\int_0^{\tau_x^1} f(y)\,\mathbf{1}_{\{X(s)=y\}}\,\mathrm{d}\,s\right]\\ &= \int_E f(y)\,\mathrm{d}\,\mu(y).\end{aligned}$$

Concerning the item (ii), let us recall that with the notations of Chapter 3,

$$\hat{\tau}_x^1=\inf\{n>0,\,\hat{X}_n=x\}.$$

By using the first pathwise construction of X, we obtain that

$$\mathbf{E}_x\left[\int_0^{\tau_x^1}\mathbf{1}_{\{X(s)=y\}}\,\mathrm{d}\,s\right]=\sum_{n=1}^{\infty}\mathbf{E}_x\left[\xi_n\,\mathbf{1}_y(\hat{X}_{n-1})\,\mathbf{1}_{\{n\le\hat{\tau}_x^1\}}\right].$$

The event $\{n\le\hat{\tau}_x^1\}$ is the complementary of $\{\tau_x^1<n\}=\{\tau_x^1\le n-1\}$, so it is $\hat{\mathcal{F}}_{n-1}$ measurable (with obvious notations). Thus by construction,

$$\mathbf{E}\left[\xi_n\,|\,\hat{\mathcal{F}}_{n-1}\right]=\mathbf{E}\left[\xi_n\,|\,\hat{X}_{n-1}\right]=\frac{1}{q(\hat{X}_{n-1},\,\hat{X}_{n-1})}.$$

From this, we deduce that

$$\begin{aligned}\mathbf{E}_x\left[\xi_n\,\mathbf{1}_y(\hat{X}_{n-1})\,\mathbf{1}_{\{n\le\hat{\tau}_x^1\}}\right] &= \mathbf{E}_x\left[\mathbf{E}\left[\xi_n\,|\,\hat{\mathcal{F}}_{n-1}\right]\mathbf{1}_y(\hat{X}_{n-1})\,\mathbf{1}_{\{n\le\hat{\tau}_x^1\}}\right]\\ &= \frac{1}{q(y,\,y)}\mathbf{P}_x(\hat{X}_{n-1}=y,\,\hat{\tau}_x^1\ge n),\end{aligned}$$

and we get

$$\mathbf{E}_x\left[\int_0^{\tau_x^1}\mathbf{1}_{\{X(s)=y\}}\,\mathrm{d}\,s\right]=\frac{1}{q(y,\,y)}\sum_{n=1}^{\infty}\mathbf{P}_x(\hat{X}_{n-1}=y,\,\hat{\tau}_x^1\ge n)\quad=\frac{1}{q(y,\,y)}\hat{\mu}(y),$$

where $\hat{\mu}$ is a stationary measure of $\hat{X}$. The proof is complete, as item (iii) has already been shown in Theorem 7.11. □

EXAMPLE (EXAMPLE 7.1 CONTINUED: M/M/1 QUEUE).– For this example, the system

$$(\pi(0),\,\pi(1),\,\cdots,\,\pi(i),\,\cdots)\begin{pmatrix} -\lambda & \lambda & & & & \\ \mu & -(\lambda+\mu) & \lambda & (0) & & \\ & & \ddots & & & \\ & (0) & \mu & -(\lambda+\mu) & \lambda & \\ & & & & \ddots & \end{pmatrix} = 0$$

is equivalent to the equations

$$\begin{aligned} -\lambda\pi(0) + \mu\pi(1) &= 0 \\ \lambda\pi(0) - (\lambda+\mu)\pi(1) + \mu\pi(2) &= 0 \\ &\vdots \\ \lambda\pi(i-1) - (\lambda+\mu)\pi(i) + \mu\pi(i+1) &= 0 \\ &\vdots \end{aligned}$$

By adding these equations successively by pairs, we obtain that

$$-\lambda\pi(0) + \mu\pi(1) = 0, -\lambda\pi(1) + \mu\pi(2) = 0, \ldots - \lambda\pi(i) + \mu\pi(i+1) = 0,$$

that is for any integer i,

$$\pi(i+1) = \rho\pi(i) \text{ with } \rho = \lambda/\mu,$$

or in other words

$$\pi(i) = \rho^i\pi(0).$$

We know from the study of the G/G/1 queue in Section 4.1, and from the particular case of M/GI/1 treated in Chapter 5, that the queue is stable, in that the process counting the number of customers is recurrent, if and only if $\rho < 1$. Indeed, provided that the latter holds true, the normalization equation

$$\sum_{i\in\mathbf{N}} \pi(i) = 1$$

implies that the unique invariant probability is given for all $i \in \mathbf{N}$ by

$$\pi(i) = \rho^i(1-\rho).$$

In other words, the size of the system at equilibrium follows a geometric distribution with parameter ρ, shifted from 1.

Given the strong Markov property of Theorem 7.7, the proof of the following result is exactly similar to its analog in the discrete case, i.e. Theorem 3.22.

THEOREM 7.13.– *Let X be a regular, irreducible, and recurrent Markov process. We denote π, its only invariant probability. For any $f \in L^1(\pi)$, we have*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(s))\,\mathrm{d}\,s = \sum_{x\in E} f(x)\pi(x).$$

NOTE.– This Theorem provides an interesting technique for approaching π. In fact, it suffices to simulate a trajectory of the process following, for instance, the pathwise construction presented in the beginning of this chapter, and to compute the proportion of time spent by the process in each state. This will provide an approximation of π in the long run, allowing then to compute the expectations of several more general functions at equilibrium . It remains to determine what "long" means, that is to determine the rate of this convergence. This is beyond the scope of this book, but is however an important topic in current research.

THEOREM 7.14.– *Let X be a regular irreducible Markov process on E and $x \in E$. If X is transient, then*

$$p_t(x,\, x) \xrightarrow[t\to\infty]{} 0.$$

If X is recurrent of invariant probability π, then

$$p_t(x,\, y) \xrightarrow[t\to\infty]{} \pi(y) \textit{ for all } y \in E \textit{ and } \mathbf{E}_x\left[\tau_x^1\right] = \frac{1}{\pi(x)}.$$

Proof. We develop the proof only in the recurrent case, the transient case is addressed similarly. Let $h > 0$ and $X_n^h = X(nh)$. According to Theorem A.12,

$$\begin{aligned}\mathbf{E}\left[f(X_n^h)\,|\,X_j^h,\, j = 0, \cdots, n-1\right] &= \mathbf{E}\left[f(X_n^h)\,|\,X_{n-1}^h\right] \\ &= \mathbf{E}\left[f(X(nh))\,|\,X((n-1)h)\right] = P_h f(X((n-1)h)).\end{aligned}$$

The sequence X^h is a thus a Markov chain with transition operator P_h. In view of Theorem 7.10, X^h is irreducible and according to [7.17], π is an invariant probability for this chain. According to Theorem 3.26, for any $x \in E$,

$$p_{nh}(x,\, y) \xrightarrow{n\to\infty} \pi(y).$$

Since we have

$$\begin{aligned}|p_t(x,\, y) - p_s(x,\, y)| &= \left|\int_s^t Ap_u(x,\, y)\,\mathrm{d}\,u\right| \\ &\le \int_s^t |Ap_u(x,\, y)|\,\mathrm{d}\,u \\ &\le \|q\|_\infty (t-s),\end{aligned}$$

we deduce that

$$|p_t(x,\, y) - \pi(y)| \leq \|q\|_\infty |t - nh| + |p_{nh}(x,\, y) - \pi(y)|.$$

Let $\epsilon > 0$, and fix h such that $h||q||_\infty < \epsilon$ and n_0 such that

$$n \geq n_0 \Longrightarrow |p_{nh}(x,\, y) - \pi(y)| \leq \epsilon.$$

If $t \geq n_0 h$, then for some n such that $|t - nh| \leq h$, we have $|p_t(x,\, y) - \pi(y)| \leq 2\epsilon$. Thus,

$$\lim_{t \to \infty} p_t(x,\, y) = \pi(y).$$

We know that the invariant probability is proportional to the measure defined in [7.16]. But

$$\sum_{y \in E} \mathbf{E}_x \left[\int_0^{\tau_x^1} \mathbf{1}_{\{y\}}(X(s)) \,\mathrm{d}\, s \right] = \mathbf{E}_x \left[\tau_x^1\right],$$

hence by applying [7.16] for $y = x$, we get

$$\pi(x) = \frac{1}{\mathbf{E}_x \left[\tau_x^1\right]},$$

hence the result. □

7.4. Martingale problem

DEFINITION 7.9.– *Let ν be a probability measure on the countable space E, and A a continuous operator from $l^\infty(E)$ in itself, that is for some $c > 0$, for any $f \in l^\infty(E)$,*

$$\|Af\|_\infty = sup_{x \in E} |Af(x)| \leq c \|f\|_\infty.$$

Then, we denote

$$\|A\|_\infty = \sup_{\|f\|_\infty = 1} \|Af\|_\infty.$$

THEOREM 7.15.– *The process X is a regular Markov process with initial distribution ν and infinitesimal generator A, if and only if the distribution of $X(0)$ is ν and for any bounded function f, the process*

$$M^f \colon t \mapsto f(X(t)) - f(X(0)) - \int_0^t Af(X(s)) \,\mathrm{d}\, s$$

is a local martingale, whose quadratic variation is given for all t by

$$\langle M^f, M^f \rangle_t = \int_0^t \big(Af^2(X(s)) - 2f(X(s))Af(X(s)) \big) \,\mathrm{d}\, s. \qquad [7.18]$$

Proof. Let $f \in l^\infty(E)$. Assume at first that the processes

$$t \mapsto \int_0^t Af(X(s))\,\mathrm{d}\,s \text{ and } t \mapsto \int_0^t Af^2(X(s))\,\mathrm{d}\,s$$

are bounded. Notice that the process $(s \mapsto Af(X(s)))$ is adapted. For any $s \geq 0$, we have

$$\begin{aligned}
&\mathbf{E}\left[M^f(t+s)\,|\,\mathcal{F}_t\right] \\
&= \mathbf{E}\left[f(X(t+s))\,|\,\mathcal{F}_t\right] - \int_0^{t+s} \mathbf{E}\left[Af(X(u))\,|\,\mathcal{F}_t\right]\mathrm{d}\,u \\
&= \mathbf{E}\left[f(X(t+s))\,|\,X(t)\right] - \int_0^t Af(X(u))\,\mathrm{d}\,u - \int_t^{t+s} \mathbf{E}\left[Af(X(u))\,|\,\mathcal{F}_t\right]\mathrm{d}\,u \\
&= P_s f(X(t)) - \int_0^t Af(X(u))\,\mathrm{d}\,u - \int_0^s P_u Af(X(t))\,\mathrm{d}\,u.
\end{aligned}$$

From [7.5],

$$P_s f(X(t)) - \int_0^s P_u Af(X(t))\,\mathrm{d}\,u = f(X(t)),$$

thus

$$\mathbf{E}\left[M^f(t+s)\,|\,\mathcal{F}_t\right] = f(X(t)) - \int_0^t Af(X(u))\,\mathrm{d}\,u = M^f(t),$$

which means that M^f is a martingale.

Now, for any $f \in l^\infty(E)$ and any $t \geq 0$, $f^2(X(t))$ is integrable. On the one hand, in view of the argument of the first part of this proof, we can write

$$f^2(X(t)) - f^2(X(0)) = \int_0^t Af^2(X(s))\,\mathrm{d}\,s + M^{f^2}(t). \qquad [7.19]$$

On the other hand, Itô's Formula for rcll martingales with finite variation (Theorem A.23), gives

$$\begin{aligned}
f^2(X(t)) - f^2(X(0)) &= 2\int_0^t f(X(s^-))\,\mathrm{d}\,M^f(s) \\
&\quad + 2\int_0^t f(X(s^-))Af(X(s))\,\mathrm{d}\,s + [f\circ X,\, f\circ X]_t.
\end{aligned} \qquad [7.20]$$

By comparing equations [7.19] and [7.20] and by definition of the quadratic variation, we deduce that

$$[f \circ X,\, f \circ X]_t = \int_0^t Af^2(X(s))\,\mathrm{d}\,s - 2\int_0^t f(X(s^-))Af(X(s))\,\mathrm{d}\,s.$$

As this process is continuous and adapted, it is predictable, hence [7.18]. Notice, that we can replace $f(X(s^-))$ by $f(X(s))$ in the second integral, since there is no fixed jump in s, and hence $\mathrm{d}\,s$-almost surely $X(s^-) = X(s)$ (see the Note at the end of section 7.2).

We now address the general case. Let for any integer n,

$$\tau_n = \inf\{t \geq 0,\ \int_0^t |Af(X(s))|\,\mathrm{d}\,s > n \text{ or } \int_0^t |Af^2(X(s))|\,\mathrm{d}\,s > n\}$$

and $M^f_{\tau_n}(t) = M^f(\tau_n \wedge t)$. Let us notice that by construction, $|M^f_{\tau_n}(t)| \leq 2\|f\|_\infty + n$, therefore $M^f_{\tau_n}$ is uniformly integrable. In addition, since A is continuous from $l^\infty(E)$ into itself, τ_n tends toward infinity, so $\{\tau_n,\, n \geq 1\}$ reduces M^f. We observe finally that for any stopping time τ,

$$\begin{aligned} M^f_\tau(t) &= f(X(t\wedge\tau)) - f(X(0)) - \int_0^{t\wedge\tau} Af(X(s))\,\mathrm{d}\,s \\ &= f(X^\tau(t)) - f(X^\tau(0)) - \int_0^{t\wedge\tau} Af(X^\tau(s))\,\mathrm{d}\,s, \end{aligned}$$

where $X^\tau(s) = X(\tau \wedge s)$. Therefore, we can apply the above argument to $M^f_{\tau_n}$ and X^{τ_n}. As a result, $M^f_{\tau_n}$ is a square integrable martingale whose quadratic variation is given by [7.18].

For the converse, let us take for granted Lemma 7.16 below. We will show that for any $f \in l^\infty(E)$, for any $t,\ s \geq 0$,

$$\mathbf{E}\left[f(X(t+s))\,|\,\mathcal{F}_t\right] = \lim_{n\to\infty}(\mathrm{Id} - n^{-1}A)^{-[ns]}f(X(t)). \qquad [7.21]$$

The conditional expectation given $\mathcal{F}_t$ appears as depending only on $X(t)$ and on s, implying homogeneity and the simple Markov property. Let A' be the infinitesimal generator of X. According to the first part of this proof,

$$t \mapsto \int_0^t (A - A')f(X(s))\,\mathrm{d}\,s$$

is a martingale which is null at 0. In addition, this process is continuous and has finite variations. This implies that the process is null $\mathbf{P} \otimes dt$-almost surely. Moreover,

$$(A - A')f(x) = \lim_{t\to 0}\frac{1}{t}\int_0^t (A - A')f(X(s))\,\mathrm{d}\,s = 0,$$

hence $A = A'$. Finally, as A is continuous from $l^\infty(E)$ into itself and

$$\sup_{x\in E} |a(x,\, x)| = \sup_{x\in E} |A\,\mathbf{1}_{\{x\}}(x)| \leq c,$$

we conclude that X is a regular Markov process.

We now show [7.21]. By possibly studying X locally, we can always assume that $E^{\lambda,\, f}$ is a uniformly integrable martingale. Applying [7.22] for $\lambda = n$, we obtain that

$$f(X(t)) = \mathbf{E}\left[\int_0^\infty e^{-ns}((n\,\mathrm{Id} - A)f)(X(t+s))\,\mathrm{d}\,s \,|\, \mathcal{F}_t\right].$$

Applying this result to $(\mathrm{Id} - \frac{1}{n}A)^{-1}f$ yields

$$\begin{aligned}\left(\mathrm{Id} - \frac{1}{n}A\right)^{-1} f(X(t)) &= n\mathbf{E}\left[\int_0^\infty e^{-ns} f(X(t+s))\,\mathrm{d}\,s \,|\, \mathcal{F}_t\right] \\ &= \mathbf{E}\left[\int_0^\infty e^{-s} f(X(t+s/n))\,\mathrm{d}\,s \,|\, \mathcal{F}_t\right].\end{aligned}$$

It follows by induction that for any integer k,

$$\begin{aligned}&\left(\mathrm{Id} - \frac{1}{n}A\right)^{-k} f(X(t)) \\ &\quad = \mathbf{E}\left[\int_0^\infty \ldots \int_0^\infty e^{-(s_1+\ldots+s_k)} f(X(t+n^{-1}(s_1+\ldots+s_k)))\,\mathrm{d}\,s_1 \ldots \mathrm{d}\,s_k \,|\, \mathcal{F}_t\right].\end{aligned}$$

Let $\{Z_k,\, k \geq 1\}$ be a sequence of independent random variables, independent of X, and of exponential distribution with parameter 1. For any non-zero u, the strong Law of Large Numbers states that

$$u\,\frac{1}{\lfloor nu\rfloor}\sum_{j=1}^{\lfloor nu\rfloor} Z_j \xrightarrow{n\to\infty} u\mathbf{E}\left[Z_1\right] = u.$$

We therefore have

$$\begin{aligned}&\int_0^\infty \cdots \int_0^\infty e^{-(s_1+\cdots+s_{\lfloor nu\rfloor})} f(X(t+n^{-1}(s_1+\cdots+s_k)))\,\mathrm{d}\,s_1 \ldots \mathrm{d}\,s_k \\ &\qquad = \mathbf{E}\left[f\left(X\left(t + u(nu)^{-1}\sum_{j=1}^{\lfloor nu\rfloor} Z_j\right)\right)\right] \xrightarrow{n\to\infty} f(X(t+u)),\end{aligned}$$

and obtain [7.21] by dominated convergence. □

NOTE.– The derivation of the quadratic variation is interesting in itself, as it is one of the keystones of many approximation methods for processes, such as fluid limits, mean fields and diffusion approximations.

LEMMA 7.16.– *Let X be a rcll process. For any $f \in l^\infty(E)$, the process*

$$M^f : t \mapsto f(X(t)) - f(X(0)) - \int_0^t Af(X(s))\,\mathrm{d}\,s$$

is a local martingale if, and only if, for any $\lambda \in \mathbf{R}$, the process

$$E^{\lambda,\,f} : t \mapsto e^{-\lambda t} f(X(t)) + \int_0^t e^{-\lambda s}(\lambda f(X(s)) - Af(X(s)))\,\mathrm{d}\,s$$

is a local martingale. In particular, if the sequence $\{\tau_n,\ n \geq 1\}$ reduces $E^{\lambda,\,f}$, we obtain that

$$f(X^{\tau_n}(t)) = \mathbf{E}\left[\int_0^\infty e^{-\lambda s}(\lambda f(X^{\tau_n}(t+s)) - Af(X^{\tau_n}(t+s)))\,\mathrm{d}\,s\,|\,\mathcal{F}_t\right]. \qquad [7.22]$$

Proof. By possibly studying X locally, we can always assume that the processes $E^{\lambda,\,f}$ and M^f are uniformly integrable. We set for any $t \geq 0$, $U(t) = \exp(-\lambda t)$. This process is continuous and with bounded variation, so according to the integration by parts Formula, we have for any t,

$$U(t)M^f(t) = \int_0^t U(s)\,\mathrm{d}\,M^f(s) - \int_0^t M^f(s)\lambda e^{-\lambda s}\,ds. \qquad [7.23]$$

By assumption, M^f is a martingale, hence so is the case for $\int U\,\mathrm{d}\,M^f$ by the very construction of the stochastic integral. Let us now notice that

$$\begin{aligned}\int_0^t \int_0^s Af(X(u))\,\mathrm{d}\,u\lambda e^{-\lambda s}\,\mathrm{d}\,s &= \int_0^t Af(X(u)) \int_u^t \lambda e^{-\lambda s}\,\mathrm{d}\,s\,\mathrm{d}\,u \\ &= \int_0^t Af(X(u))(e^{-\lambda u} - e^{-\lambda t})\,\mathrm{d}\,u.\end{aligned} \qquad [7.24]$$

Substituting [7.24] into [7.23], we see that the process

$$\begin{aligned} t \mapsto\ & U(t)M^f(t) + \int_0^t M^f(s)\lambda e^{-\lambda s}\,ds \\ &= e^{-\lambda t} f(X(t)) + \int_0^t f(X(s))\lambda e^{-\lambda s}\,\mathrm{d}\,s - \int_0^t Af(X(s))e^{-\lambda s}\,\mathrm{d}\,s\end{aligned}$$

is a martingale, hence the result. □

7.5. Reversibility and applications

In this section, $(X(t), t \geq 0)$ is a homogeneous Markov process, regular, irreducible and recurrent with rcll paths on the countable space E, having transition operator Q and infinitesimal generator A. Hereafter, we will say that $(X(t), t \geq 0)$ is stationary if it admits an invariant probability π and if the distribution of $X(0)$ is π. Without loss of generality, we may assume that $\pi(x) > 0$ for any $x \in E$.

DEFINITION 7.10.– *Assume that* $(X(t),\, t \geq 0)$ *is stationary. For any* $T > 0$*, the reversed process of* $(X(t), t \geq 0)$ *from* T *is the process* $\big(\bar{X}^T(t),\, t \in [0, T]\big)$ *defined for any* $t \in [0, T]$ *by*

$$\bar{X}^T(t) = X((T-t)^-) = \lim_{s \searrow t} X((T-s)).$$

Let us recall (see the Definition A.5 for more details) that $l^2(\mathbf{N},\, \pi)$ is the Hilbert space of square integrable sequences for the measure π.

LEMMA 7.17.– *The generator* A *is a continuous operator from* $l^2(\mathbf{N},\, \pi)$ *into itself. Therefore it admits an adjoint* $\bar{A}$ *in* $l^2(\mathbf{N},\, \pi)$*, defined by*

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

Proof. Let us recall that $A(x,\, x) = -\sum_{y \neq x} A(x,\, y) < 0$, thus we have

$$\sum_{y \in E} |A(x,\, y)| = 2|A(x,\, x)|.$$

Therefore, the measure ν defined by

$$\nu(y) = \frac{1}{2|A(x,\, x)|}\, |A(x,\, y)|, \text{ for any } y \in E,$$

is a probability measure on E. Consequently,

$$\begin{aligned}\left(\sum_{y \in E} A(x,\, y) f(y)\right)^2 &\leq 4A(x,\, x)^2 \left(\sum_{y \in E} \frac{|A(x,\, y)|}{2|A(x,\, x)|} f(y)\right)^2 \\ &\leq 2|A(x,\, x)| \sum_{y \in E} |A(x,\, y)||f(y)|^2 \\ &\leq 2\|A\|_\infty \sum_{y \in E} |A(x,\, y)||f(y)|^2,\end{aligned}$$

from Jensen's inequality and the regularity of A. Therefore,

$$\begin{aligned}\sum_{x\in E}(AF)(x)^2\pi(x) &\le 2\|A\|_\infty \sum_{x\in E}\sum_{y\in E}|A(x,\ y)||f(y)|^2\pi(x)\\ &= 2\|A\|_\infty \sum_{y\in E}(1+|A(y,\ y)|)\pi(y)f(y)^2\\ &\le 2\|A\|_\infty(1+\|A\|_\infty)\|f\|^2_{l^2(\pi)}.\end{aligned}$$

Thus, A is continuous from $l^2(\mathbf{N},\ \pi)$ into itself, hence so is the case for the adjoint $\bar{A}$ of A, defined for any $u, v\ \in l^2(\mathbf{N},\ \pi)$ by

$$\langle Au,\ v\rangle_{l^2(\mathbf{N},\,\pi)} = \langle u,\ Av\rangle_{l^2(\mathbf{N},\,\pi)}. \qquad [7.25]$$

Furthermore,

$$\bar{A}(x,\ y) = A(y,\ x)\frac{\pi(y)}{\pi(x)},$$

since as $\bar{A}(x,\ y) = (\bar{A}\,\mathbf{1}_{\{x\}})(y)$, taking $u = \mathbf{1}_{\{y\}}$ in [7.25] gives

$$\left(A\,\mathbf{1}_{\{y\}}\right)(x)\pi(x) = \left(\bar{A}\,\mathbf{1}_{\{x\}}\right)(y)\pi(y),$$

that is

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

The proof is complete. □

THEOREM 7.18.– *Under the ongoing assumptions, for any $T > 0$ the process $\left(\bar{X}^T(t),\, t\in[0,T]\right)$ is a Markov process having rcll paths, and of infinitesimal generator $\bar{A}$.*

Proof. The paths of $\left(\bar{X}^T(t),\, t\ge 0\right)$ are a.s. right-continuous, since by almost sure existence of a left-hand limit at any point for X, we have a.s.

$$\begin{aligned}\lim_{h\searrow 0}\bar{X}^T(t+h) &= \lim_{h\searrow 0}\lim_{h'\searrow 0} X(T-(t+h)-h')\\ &= \lim_{\epsilon\searrow 0} X(T-(t+\epsilon)) = X(T-t)^- = \bar{X}^T(t).\end{aligned}$$

The existence of a left-hand limit is proven similarly. Hence it suffices to take the left-hand limit at $T-t$ to make the paths of the reversed process rcll.

With Theorem 7.15 in hand, we must prove that for any $f\in l^\infty(E)$,

$$t\mapsto f(\bar{X}^T(t)) - f(\bar{X}^T(0)) - \int_0^t \bar{A}f(\bar{X}^T(r))\,\mathrm{d}\,r$$

is a local martingale. The filtration is of course that generated by the paths of $\bar{X}^T$ and not those of X, that is $\bar{\mathcal{F}}_t = \sigma\{\bar{X}^T(s),\, s \le t\}$. By using a monotone class argument, it is necessary and sufficient to prove that for $0 \le s_1 < \ldots < s_n \le s < t$, for any bounded function $\varphi\colon\ E^n \to \mathbf{R}$, we have

$$\mathbf{E}\Bigg[\bigg(f(\bar{X}^T(t)) - f(\bar{X}^T(s)) \\ - \int_s^t \bar{A}f(\bar{X}^T(r))\,\mathrm{d}\,r\bigg)\,\varphi(\bar{X}^T(s_1),\, \cdots,\, \bar{X}^T(s_n))\Bigg] = 0.$$

As X is Markov, this identity can be rewritten as

$$0 = \mathbf{E}\Bigg[\bigg(f(X(T-t)) - f(X(T-s)) - \int_s^t (\bar{A}f)(X(T-r))\,\mathrm{d}\,r\bigg) \\ \times \mathbf{E}\left[\varphi(\bar{X}^T(s_1),\, \cdots,\, \bar{X}^T(s_n))\,|\,X(T-s)\right]\Bigg],$$

which, according to Theorem A.7, is in turn equivalent to the fact that for any bounded ψ,

$$\mathbf{E}\Bigg[\bigg(f(X(T-t)) - f(X(T-s)) \\ - \int_{T-t}^{T-s} (\bar{A}f)(X(r))\,\mathrm{d}\,r\bigg)\,\psi(X(T-s))\Bigg] = 0. \qquad [7.26]$$

By introducing the semi-group associated with X and recalling that by stationarity, the distribution of $X(T-s)$ is that of $X(0)$, that is to say π, we obtain on the one hand,

$$\mathbf{E}\left[(f(X(T-t)) - f(X(T-s)))\psi(X(T-s))\right] \\ = \int_E \psi P_{t-s} f\,\mathrm{d}\,\pi - \int_E f\psi\,\mathrm{d}\,\pi \qquad [7.27]$$

and on the other hand, from [7.25] and [7.5],

$$\begin{aligned}\mathbf{E}\left[\int_s^t \bar{A}f(X(T-r))\,\mathrm{d}\,r\,\psi(X(T-s))\right] &= \int_E\int_s^t P_{r-s}\psi\bar{A}f\,\mathrm{d}\,r\,\mathrm{d}\,\pi\\ &= \int_0^{t-s}\int_E P_r\psi\bar{A}f\,\mathrm{d}\,\pi\,\mathrm{d}\,r\\ &= \int_0^{t-s}\int_E AP_r\psi f\,\mathrm{d}\,\pi\,\mathrm{d}\,r \qquad [7.28]\\ &= \int_E\int_0^{t-s} AP_r\psi\,\mathrm{d}\,r f\,\mathrm{d}\,\pi\\ &= \int_E (P_{t-s}\psi-\psi)f\,\mathrm{d}\,\pi.\end{aligned}$$

By substracting [7.28] to [7.27], we obtain [7.26]. □

Let us notice in particular, that the generator of $\left(\bar{X}^T(t),\,t\geq 0\right)$ does not depend on T, which will be crucial in the construction hereafter.

DEFINITION 7.11.– *The process $(X(t),\,t\geq 0)$ is said to be reversible, if it is stationary and for any $T>0$, of same distribution as its reversed process on $[0,T]$.*

For any $T>0$, $X(0)$ and $\bar{X}^T(0)$ have the same distribution π. As the distribution of a Markov process is fully determined by its generator and its initial distribution, we deduce from Theorem 7.18 that $(X(t),\,t\geq 0)$ is reversible if and only if for any $x,\,y\in E$,

$$\pi(x)A(x,\,y)=\pi(y)A(y,\,x). \qquad [7.29]$$

This relation, known as *local balance equation*, is a “mirror” property: the transition rate from x to y equals that from y to x.

The following result will be used in Chapter 8.

LEMMA 7.19.– *Let X be a Markov process on E with generator A and let π, a probability on E. Define $\hat{A}(x,\,y)$ for any $x,y\in E$ such that $x\neq y$ by*

$$\hat{A}(x,\,y)=\frac{A(y,\,x)\pi(y)}{\pi(x)}.$$

Then, if for all $x\in E$,

$$\sum_{y\neq x}\hat{A}(x,\,y)=\sum_{y\neq x}A(x,\,y),$$

then π is the stationary probability of X.

Proof. For all $x \in E$,

$$\begin{aligned}
\pi A(x) &= \sum_{y \in E} \pi(y) A(y,\, x) \\
&= \pi(x) A(x,\, x) + \sum_{y \neq x} \pi(y) A(y,\, x) \\
&= \pi(x) A(x,\, x) + \sum_{y \neq x} \pi(x) \hat{A}(x,\, y) \\
&= \pi(x) \left(- \sum_{y \neq x} A(x,\, y) + \sum_{y \neq x} \hat{A}(x,\, y) \right) \\
&= 0.
\end{aligned}$$

□

It is a straightforward corollary of the latter result that any Markov process satisfying to a local balance equation is stationary, and hence reversible.

COROLLARY 7.20.– *For any Markov process X on E, if there is a probability π satisfying [7.29], then X admits π as stationary probability.*

Proof. Lemma 7.19 is satisfied for $\hat{A} = A$. □

Birth and Death processes

An important class of Markov processes, including most of the processes studied in the following, enjoys the reversibility property automatically when there exists an invariant probability: these are the *birth and death* processes.

DEFINITION 7.12.– *A homogeneous Markovian process $(X(t),\, t \geq 0)$ with values in E, where $E = \mathbf{N}$ or $[\![0, n]\!]$, is said to be a birth and death process if its infinitesimal generator is tridiagonal on E, that is for any $x \in E$,*

$$A(x,\, y) = 0 \text{ for any } y \text{ such that } \mid y - x \mid \geq 2.$$

This terminology is inherited from that of population dynamics. If $(X(t),\, t \geq 0)$ represents the size of a given population at any time, the jumps of $(X(t),\, t \geq 0)$ only occur at instants of birth (from x to $x+1$) or death (from x to $x+1$).

THEOREM 7.21.– *Any stationary birth and death process of distribution π is reversible.*

Proof. As A is tridiagonal, it suffices to check that

$$\pi(x-1)A(x-1,\, x) = \pi(x)A(x,\, x-1)$$

for any $x \in E$, which we verify by induction. Initially,

$$\pi(0)A(0,\, 1) = -\pi(0)A(0,\, 0) = \pi(1)A(1,\, 0),$$

since $\pi A(0) = 0$. Then, if

$$\pi(x-1)A(x-1,\, x) = \pi(x)A(x,\, x-1)$$

for an index $x \geq 1$ such that $x+1 \in E$, then

$$\begin{aligned}\pi(x)A(x,\, x+1) &= \pi(x)(-A(x,\, x) - A(x,\, x-1))\\ &= -\pi(x)A(x,\, x) - \pi(x-1)A(x-1,\, x)\\ &= \pi(x+1)A(x+1,\, x),\end{aligned}$$

since $\pi A(x) = 0$. The proof is complete. □

The following theorem allows us to derive easily many stationary distributions in practical cases.

THEOREM 7.22 (Kelly's Theorem).– *Let $(X(t),\, t \geq 0)$ be a reversible Markov process on E, of infinitesimal generator A and of invariant probability π. Let $F \subset E$. We define, for a certain $\alpha \geq 0$, the following matrix $\tilde{A}$ on $E \times E$:*

$$\tilde{A}(x,\, y) = \begin{cases} \alpha A(x,\, y) & \textit{if } x \in F,\ y \in E \setminus F; \\ A(x,\, y) & \textit{if not for } x \neq y; \end{cases}$$

$$\tilde{A}(x,\, x) = -\sum_{y \neq x} \tilde{A}(x,\, y) \ \textit{for all } x \in E.$$

Then, the Markov process $\left(\tilde{X}(t),\, t \geq 0\right)$ of generator $\tilde{A}$ is reversible and of invariant probability $\tilde{\pi}$ given for any $i \in E$ by

$$\pi(x) = \begin{cases} C\pi(x) & \textit{if } x \in F; \\ C\alpha\pi(x) & \textit{if } x \in E \setminus F, \end{cases}$$

where $C = \left(\sum_{k \in F} \pi(k) + \alpha \sum_{k \in E \setminus F} \pi(k)\right)^{-1}$ is the normalization constant.

Proof. That $\tilde{\pi}$ defines a probability measure on E is straightforward. It is thus sufficient to check that it is reversible, which is immediate by observing that

$$\tilde{\pi}(x)\tilde{A}(x,\, y) = \alpha C\pi(x)A(x,\, y) = \tilde{\pi}(y)\tilde{A}(y,\, x),$$

for any $x \in F$ and $y \in E \setminus F$. □

In particular, if we set $\alpha = 0$, we prevent the process from leaving the subset F while keeping the reversibility property. We will see an application of this result in Chapter 9.

7.6. Markov Modulated Poisson Processes

The Markov Modulated Poisson Processes (MMPP) correspond to a class of Markov processes generalizing the Poisson process, while keeping most of its tractable characteristics. They naturally appear in the modeling of overflow systems, see Chapter 9. It was once thought they could serve as models for data streams. Even if this approach seems to become obsolete, it is however interesting; see Chapter 1.

DEFINITION 7.13.– *Let $(J(t),\, t \geq 0)$ be a stationary Markov process with values in $E = [\![1, m]\!]$. Let Q its infinitesimal generator and ν its stationary probability. Let λ be a function from E to $\mathbf{R}^+$. The point process N is an MMPP with parameters Q and Λ if and only if, for any function f with compact support,*

$$\mathbf{E}\left[\exp\left(-\int_0^t f(s)\,\mathrm{d}\,N(s)\right)\right] = \mathbf{E}\left[\exp\left(-\int_0^t \left(1 - e^{-f(s)}\right)\lambda(J(s))\,\mathrm{d}\,s\right)\right].$$

NOTE.– An MMPP is thus nothing but a Cox process whose intensity, varying over time, depends on the evolution of a Markov process with finite state space.

The latter definition means that when the phase process J is in phase j, the points of N follow a Poisson process of intensity $\lambda(j)$. When the phase process changes state, according to the dynamics induced by its infinitesimal generator, the intensity of the Poisson process changes. Figure 7.4 shows a sample path of a 2-phase MMPP.

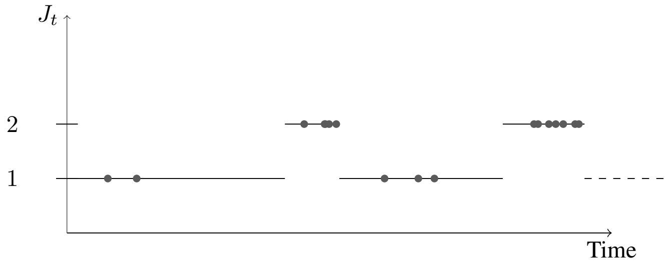

Figure 7.4. *A sample path of a 2-phase MMPP*

The process N alone is not Markov. To see this, remark that at any time t, the remaining time before its next point is exponentially distributed, but with a parameter depending on the phase. However, the couple process $(N,\, J)$ is Markovian and its generator reads as a block matrix. To see this, denote Λ, the diagonal matrix whose coefficient in $(i,\, i)$ is $\lambda(i)$. With these notations, the infinitesimal generator of $(N,\, J)$ is given by

$$\begin{pmatrix} Q-\Lambda & \Lambda & & & \\ 0 & Q-\Lambda & \Lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Indeed, the events are of two types: arrival or phase change. An arrival causes a transition from the state $(i,\, \phi)$ toward $(i+1,\, \phi)$. A phase change leaves the first component unchanged and modifies the second one. We deduce from this a simulation algorithm of a trajectory of an MMPP. We always assume that the state space is $E = [\![1,\, m]\!]$.

Algorithm 7.1. Construction of a trajectory of an MMPP

```
Data: $m,\, Q,\, \lambda, \{i_0\},\, T$
Result: A trajectory $(t_n,\, n \geq 1)$ on $[0,\, T]$ of an MMPP $(m,\, Q,\, \lambda)$ of initial
        state $\{i_0\}$.
for $i = 1,\, \ldots,\, m$ do
    $r_{i0} = \lambda_i$;
    for $j = 1, \ldots,\, m$ do
        $r_{i,\, j} = r_{i,\, j-1} + q(i,\, j)$;
    end
end
Phase $\leftarrow i_0$;
$t \leftarrow 0$;
$n \leftarrow 0$;
while $t \leq T$ do
    $x \leftarrow$ sample of a r.v. of distribution Exp $\Big(\lambda(\text{Phase}) + \sum_j q(\text{Phase},\, j)\Big)$;
    $t \leftarrow t + x$;
    $u \leftarrow$ sample of a r.v. uniformly distributed on $[0,\, 1]$;
    if $u \leq r_{i0}$ then
        $t_n \leftarrow t$
    else
        $j \leftarrow 1$;
        while $u > r_{i,\, j}$ do
            $j \leftarrow j + 1$;
        end
        Phase $\leftarrow j$;
    end
    $n \leftarrow n + 1$
end
return $t_1,\, t_2,\, \ldots,\, t_n$
```

EXAMPLE 7.2.– A typical example of MMPP is given by the overflow process of a M_λ/M_μ/S/S queue. In such a system, if some server is available, there is no overflow. If all servers are busy, new arrivals are rejected and form the overflow process. To represent this system with an MMPP, it suffices to consider as phase process J, the process counting the number of busy servers, and the function λ defined by

$$\lambda(i) = \begin{cases} 0 & \text{if } i < S \\ \lambda & \text{if } i = S. \end{cases}$$

Indeed, the "arrival" process in the overflow system is that of the original queue, since all the corresponding customers find a full system, and are thus lost by the queue, and re-directed toward the overflow system.

Superposition of MMPP

To be applicable in practice (if one think of a network, for instance), a class of processes must be stable by superposition: if we superpose two processes of the same class (i.e. Poisson, Cox, MMPP), it is desirable that the "sum" process be of the same class. We already know that the Poisson processes satisfy this property. We will extend this to the case of MMPP. This requires to introduce new notations.

DEFINITION 7.14.– *Let A and B be two square matrices, respectively of size n and p. The Kronecker product of A and B, denoted $A \otimes B$, is the square matrix of size np given by*

$$A \otimes B = \begin{pmatrix} A(1,1)B & A(1,2)B & \dots & A(1,n)B \\ \vdots & \vdots & & \vdots \\ A(n,1)B & \dots & \dots & A(n,n)B \end{pmatrix}.$$

The Kronecker sum is then defined by

$$A \oplus B = (A \otimes Id_p) + (Id_n \otimes B).$$

In particular, if A is diagonal it can be associated with function

$$a\colon \begin{cases} [\![1, n]\!] & \longrightarrow \mathbf{R} \\ i & \longmapsto A(i,i). \end{cases}$$

For two diagonal matrices A_1 and A_2 corresponding to two functions a_1 and a_2 defined respectively on $[\![1, n_1]\!]$ and $[\![1, n_2]\!]$, the matrix $A_1 \oplus A_2$ corresponds to the function

$$a_1 \oplus a_2\colon \begin{cases} [\![1, n_1]\!] \times [\![1, n_2]\!] & \longrightarrow \mathbf{R} \\ (i, j) & \longmapsto A_1(i,i) + A_2(j,j) = a_1(i) + a_2(j). \end{cases}$$

THEOREM 7.23.– *Let J_1 and J_2 be two independent Markov processes of respective characteristics $(E_1,\, A_1,\, \nu_1)$ and $(E_2,\, A_2,\, \nu_2)$. The "couple" process $J = (J_1,\, J_2)$ is a Markov process of characteristics $(E_1 \times E_2,\, A_1 \oplus A_2,\, \nu_1 \otimes \nu_2)$.*

Proof. First, it is obvious that the state space is $E = E_1 \times E_2$. In view of Theorem 7.15, it suffices to demonstrate that for any function $f \in l^\infty(E)$, the process

$$t \to \Theta^f(t) = f(J_1(t),\, J_2(t)) - \int_0^t (A_1 \oplus A_2) f(J_1(s),\, J_2(s))\,\mathrm{d}\,s$$

is a martingale. Let us first assume that $f = f_1 \otimes f_2$ with f_1 and f_2 bounded. Since J_1 and J_2 are Markov processes for $i = 1,\, 2$, we can write for any $t \geq 0$ that

$$f_i(J_i(t)) = f_i(J_i(0)) + \int_0^t A_i f_i(J_i(s))\,\mathrm{d}\,s + M_i(t),$$

where M_1 and M_2 are martingales. Itô's formula then gives

$$\begin{aligned} &f_1(J_1(t)) f_2(J_2(t)) - f_1(J_1(0)) f_2(J_2(0)) \\ &\quad = \int_0^t f_1(J_1(s)) A_2 f_2(J_2(s))\,\mathrm{d}\,s + \int_0^t f_2(J_2(s)) A_1 f_1(J_1(s))\,\mathrm{d}\,s \\ &\qquad\qquad + \sum_{s \leq t} \Delta f_1(J_1(s)) \Delta f_2(J_2(s)) + M(t), \end{aligned}$$

where M is a martingale. As J_1 and J_2 are independent, they have almost surely no common jumps, hence the second last term is zero. We can therefore write

$$\begin{aligned} &f_1(J_1(t)) f_2(J_2(t)) - f_1(J_1(0)) f_2(J_2(0)) \\ &\qquad = \int_0^t (A_1 \oplus A_2)(f_1 \otimes f_2)(J_1(s),\, J_2(s))\,\mathrm{d}\,s + M(t). \end{aligned}$$

The result is hence proven for $f = f_1 \otimes f_2$. By linearity, for all $f = \sum_{k=1}^n f_{1k} \otimes f_{2k}$ and all bounded $\psi \in \mathcal{F}_t$, we thus have

$$\mathbf{E}\left[\Theta^f(t+s)\psi\right] = \mathbf{E}\left[\Theta^f(t)\psi\right] \text{ for any } s \geq 0. \qquad [7.30]$$

For any bounded function f, there exists a sequence of functions $(f^l,\, l \geq 1)$ of the form $f^l = \sum_{k=1}^{n_l} f^l_{1k} \otimes f^l_{2k}$ and tending uniformly to f. As A_1 and A_2 are Markov kernels, $\|A_i f\|_\infty \leq \|f_i\|_\infty$. Therefore, by dominated convergence, Θ^{f^k} converges in

$l^\infty(E)$ to Θ^f. As ψ is bounded, [7.30] is also true for any bounded f, therefore Θ^f is a martingale. Finally,

$$\begin{aligned}\nu_1 \otimes \nu_2(A_1 \oplus A_2) &= \nu_1 \otimes \nu_2(A_1 \otimes \mathrm{Id}_2 + \mathrm{Id}_1 \otimes A_2)\\ &= \nu_1 A_1 \otimes \nu_2 + \nu_1 \otimes \nu_2 A_2\\ &= 0,\end{aligned}$$

and $\nu_1 \otimes \nu_2$ is thus an invariant measure for the process $(J_1,\ J_2)$. □

It is interesting to write "by hand" the generator of the process product, and obtain the form $A_1 \oplus A_2$ (see exercise 15).

THEOREM 7.24.– *Let N_i, $i = 1, \ldots, K$ be independent MMPP such that for any i, J_i takes values in $E_i = [\![1,\ m_i]\!]$, is of infinitesimal generator A_i and of invariant probability ν_i. We note Λ_i as the diagonal matrix of arrival rates for the MMPP N_i.*

The superposition process N of the N_i's is an MMPP process of J phases, whose infinitesimal generator is given by

$$A = A_1 \oplus A_2 \oplus \cdots \oplus A_K$$

and rate function $\lambda = \lambda_1 \otimes \lambda_2 \otimes \cdots \otimes \lambda_K$. The invariant probability of the phases process is $\nu_1 \otimes \cdots \otimes \nu_K$.

Let us give an insight on the case of two MMPP. As long as neither one of the phases processes changes phase, customers arrive according to the superposition of two independent Poisson processes, i.e. a Poisson process of "sum" intensity. Hence, we have a Poisson process on random intervals, whose intensity is modulated by the couples of underlying phases. All combinations of phases are *a priori* possible, thus there are $m_1 m_2$ possible phases and intensities, which reflects the fact that λ is defined on the product space $E_1 \times E_2$. Then, we have to check that the overall phase process is indeed a Markov process.

The argument using Laplace transforms presented hereafter is more abstract, although it is, by far, easier than this sketch of proof.

Proof. We address only the case $K = 2$, the general case follows by induction. From the definition of an MMPP, for $i = 1, 2$ and for any bounded f_1, f_2, we have for any $t \geq 0$,

$$f_i(N_i(t)) = f_i(N^i(0)) + \int_0^t f_i(N^i(s))\lambda_i(J_i(s))\,\mathrm{d}\,s + M^i(t),$$

where M^i is a martingale. Since the processes are independent, the martingales M^i are independent and thus their mutual quadratic variation is zero. Therefore, according to the integration by parts formula [A.13],

$$\begin{aligned}
&f_1(N^1(t))f_2(N^2(t)) - f_1(N^1(0))f_2(N^2(0)) \\
&\quad = \int_0^t f_2(N_2(s))\,\mathrm{d}\,M^1(s) + \int_0^t f_1(N_1(s))\,\mathrm{d}\,M^2(s) \\
&\qquad + \int_0^t f_1(N_1(s))f_2(N_2(s))(\lambda_1(J_1(s)) + \lambda_2(J_2(s)))\,\mathrm{d}\,s.
\end{aligned}$$

From Theorem A.32, we have thus proven that

$$\begin{aligned}
t \longmapsto (f_1 \otimes f_2)\big(N_1(t),\, N_2(t)\big) - (f_1 \otimes f_2)\big(N_1(0),\, N_2(0)\big) \\
- \int_0^t (f_1 \otimes f_2)\big(N_1(s),\, N_2(s)\big)\big(\lambda_1(J_1(s)) + \lambda_2(J_2(s))\big)\mathrm{d}\,s
\end{aligned}$$

is a martingale. By density, this result holds true for all bounded functions on $\mathbf{N} \times \mathbf{N}$, especially for functions of the form $(n_1,\, n_2) \mapsto f(n_1 + n_2)$ with f bounded from $\mathbf{N}$ to $\mathbf{R}$. It follows that the process $N_1 + N_2$ is a Cox process of compensator $\lambda_1(J_1) + \lambda_2(J_2)$. In view of Theorem 7.22, this corresponds to the intensity process associated with the couple process $(J_1,\, J_2)$. □

PASTA property

We conclude this chapter with a generalization of the PASTA property for MMPP.

THEOREM 7.25 (PASTA modified).– *Let* $N = (E,\, J,\, \nu)$ *be an MMPP and* $(\psi(t),\, t \geq 0)$*, a* $\mathcal{F}^N$*-predictable and bounded process. Then,*

$$\lim_{t\to\infty} \frac{1}{N_t} \int_0^t \psi(s)\,\mathrm{d}\,N_s = \frac{1}{\sum_{j\in E} \lambda(j)\nu(j)} \lim_{t\to\infty} \frac{1}{t} \int_0^t \psi(s)\lambda(J(s))\,\mathrm{d}\,s, \qquad [7.31]$$

provided these two limits exist.

Proof. In view of Theorem A.37, as ψ is bounded, we have that

$$\lim_{t\to\infty} \frac{1}{N(t)} \int_0^t \psi(s)\,\mathrm{d}\,N(s) = \lim_{t\to\infty} \frac{t}{\int_0^t \lambda(J(s))\,\mathrm{d}\,s} \frac{1}{t} \int_0^t \psi(s)\lambda(J(s))\,\mathrm{d}\,s.$$

Since J is an ergodic Markov process, we also have

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \lambda(J(s))\,\mathrm{d}\,s = \sum_{j\in E} \lambda(j)\nu(j),$$

hence the result. □

7.7. Problems

EXERCISE 13.– Let $(J(t),\, t \geq 0)$ be the Markov process with values in $\{A, B\}$ and whose infinitesimal generator is given by

$$\begin{pmatrix} -\sigma_A & \sigma_A \\ \sigma_B & -\sigma_B \end{pmatrix}.$$

1) Determine the stationary probability π of J.

2) What is the distribution of the sojourn time in state A? We now construct an arrival process as follows: there are no arrivals during the periods where $J(t) = A$, and during the periods where $J(t) = B$, the arrivals follow a Poisson process of intensity λ. Let us denote N_t the number of customers arrived up to time t. It is assumed that there cannot be simultaneously a phase change (i.e. a change of state of the process J) and an arrival.

3) Write down the non-zero coefficients of the infinitesimal generator of the Markov process (J, N).

4) Does this process admit a stationary probability?

This process called IPP (Interrupted Poisson Process) is a possible model for human voice (where the phases A represent the silent periods and the phases B the periods of talking). We aim to study the IPP/M/1/1 queue, setting $1/\mu$ as the mean service time. To do so, we study the process (X, J), where X represents the number of customers in the queue. We list the states in lexicographic order, i.e.

$$(0, A),\ (0, B),\ (1, A),\ (1, B).$$

5) Write the infinitesimal generator of (X, J).

6) Does this process admit a stationary probability? Make explicit the values of three out of its components in function of the fourth one.

7) Express the stationary probability of X using that of (X, J).

We now assume that

$$\lambda = 3,\ \mu = 1,\ \sigma_A = 1,\ \sigma_B = 2.$$

8) Calculate the stationary probabilities of (X, J), and then that of X.

9) What is the blocking probability at equilibrium?

10) What is the loss probability at equilibrium?

EXERCISE 14.– *We recall that for any integer n and any positive real number β,*

$$\int_0^{+\infty} x^n \beta e^{-\beta x}\, \mathrm{d}\, x = \frac{n!}{\beta^n}.$$

We consider a queue where the service times are independent and identically distributed. A request has a probability p that its service time follow an exponential distribution with parameter μ_α, and a probability $q = 1 - p$ that its service time follow an exponential distribution with parameter μ_β. Formally, if X_1 and X_2 are two independent random variables of exponential distribution with respective parameters μ_α and μ_β, if Y is a random variable independent of X_1 and X_2 such that

$$P(Y = 1) = p = 1 - P(Y = 0),$$

then the service duration reads as the random variable

$$X = X_Y = X_1 \mathbf{1}_{\{Y=1\}} + X_2 \mathbf{1}_{\{Y=2\}}.$$

The arrivals follow a Poisson process of intensity λ. There is a single server and a buffer of size K.

1) Show that the average service time is given by

$$p\frac{1}{\mu_\alpha} + q\frac{1}{\mu_\beta},$$

which will be denoted $1/\mu$.

2) Show that the variance of the service time is given by

$$p\frac{2}{\mu_\alpha^2} + q\frac{2}{\mu_\beta^2} - \frac{1}{\mu^2}.$$

Finally, to study the performances of this queuing system, we consider the Markov process X representing the number of customers in the system and the phase of the customer in service. The state space is hence

$$E = \{0\} \cup \big\{(i, \gamma),\ 1 \leq i \leq K + 1,\ \gamma \in \{\alpha, \beta\}\big\}.$$

For instance, X being in state $(5, \alpha)$ means that there are four customers in the buffer and that the customer in service is in phase α, so that his service time follows an exponential distribution of parameter μ_α.

3) For $K = 0$, write down the infinitesimal generator of X.

4) For $K = 0$, calculate the loss probability in steady state.

5) Compare to the corresponding result for the M/M/1/1 queue having the same load.

From now on, we assume that the buffer size is infinite.

6) What is the condition of existence of a stationary probability? (No computation needed!)

7) What is the mean number of customers in the system in steady state?

8) If $\mu_\alpha = 1$ and μ is fixed, what is the relation between p and μ_β?

9) Represent the changes in the average number of customers when p varies from 0 to $1/\mu$.

10) Write down the non-zero coefficients of the infinitesimal generator of X.

11) Let π be the vector representing the stationary probability. We set $x_0 = \pi(0)$ and $x_i = (\pi(i,\alpha), \pi(i,\beta))$. Write down the equations satisfied by the x_i's, by using products of block matrices.

EXERCISE 15.– Let $J_1 = (E_1, A_1, \nu_1)$ and $J_2 = (E_2, A_2, \nu_2)$ be two independent Markov processes. Write "by hand" the generator of the couple process $J = (J_1, J_2)$, to corroborate the conclusion of Theorem 7.22.

7.8. Notes and comments

There exist in the literature, many books on Markov processes with discrete state spaces and their applications to queuing. More or less in chronological order, let us quote non-exhaustively [KLE 76, CIN 75, KEL 79, ASM 03]. These books do not address the martingale problems, due to their complexity. However, this tool is the basis of many of the fluid limits, and diffusion approximation results for queuing systems. Several wonderful examples can be found in [ROB 03]. Other reference books on this subject, with a strong mathematical background, are [BRÉ 81, DEL 76, ETH 86, JAC 79].

We have deliberately chosen this approach which is a little more formal than that used in the chapter on Markov chains, for its elegance as well as to pave the way for those readers who would want to get into the theory of Markov processes in continuous state spaces, which is more difficult.

To learn more about the MMPP, we refer the reader to [FIS 93, NEU 94] and the references therein. The approach developed here, based on martingales, is original.

Epitome

– A Markov process with values in a discrete state space is a process whose paths are made of jumps regulated by a Markov chain, and of constancy intervals of exponential distributions. The parameter of each exponential distribution depend on the state.

– Its infinitesimal generator A is a (possibly infinite) matrix whose coefficients may be interpreted as follows:

- $A(x,\, x)$ is the opposite of the parameter of the exponential distribution governing the sojourn time in state x. Beware of the minus sign!
- $A(x,\, y)/|A(x,\, x)|$ is the probability that the process goes to y when it leaves x.

– The nature of the states (recurrent, transient) of the Markov process is the same as that of the embedded Markov chain.

– The stationary probability π is obtained by solving the equations $\pi A = \mathbf{0}$ and $\sum_{x \in E} \pi(x) = 1$ where π is written as a row vector.

– The stationary probability is invariant by multiplication of all the coefficients of A by the same positive number. However the dynamics changes: the sojourn times in each state change accordingly, and the transitions remain the same. In the modeling of queuing systems, we use this fact by choosing the mean service length as time unit, that is $\mu = 1$.

– If the process is irreducible, recurrent and of invariant probability π we have

$$\frac{1}{t}\int_0^t f(X(s))\,\mathrm{d}\,s \xrightarrow{t\to\infty} \sum_{x \in E} f(x)\pi(x);$$

$$\mathbf{P}(X(t) = y \,|\, X(0) = x) \xrightarrow{t\to\infty} \pi(y),\ \forall x,\, y \in E.$$

– The excursion duration from a given point x to itself can be deduced from the invariant probability

$$\mathbf{E}_x\left[\tau_x^1\right] = \frac{1}{\pi(x)}.$$

– The paths of a Markov process are simulated using the construction of Definition 7.1.

Chapter 8

Systems with Delay

All actual telecommunication systems are loss systems, since all the buffers have finite, and hence limited capacity. By system with delay, we mean a system in which the dimensioning is such that the loss caused by the overflow is negligible, and for which the relevant criterion for assessing the performances, is the waiting time.

Before we start a detailed study of these systems, we first introduce a well-known, general and very useful relation called *Little's Formula*.

8.1. Little's Formula

We consider a system with delay, in which the customers arrive at times $(T_n,\, n \geq 1)$, spend in the system sojourn times given by $(W_n,\, n \geq 1)$ and leave the system at times $(D_n = T_n + W_n,\, n \geq 1)$. We denote N as the point process of arrivals, D as the departure process and X, the process counting the number of customers in the system. At time 0, the system is assumed to be empty, i.e. $X(0) = 0$. The key point is that the system is *conservative* : all the incoming work is processed by the server(s).

THEOREM 8.1 (Little's Formula).– *We assume that N is asymptotically linear, i.e. there exists $\lambda > 0$ such that*

$$\frac{N(t)}{t} \xrightarrow{t\to\infty} \lambda \ a.s.$$

and that the sequence W is ergodic, i.e.

$$\frac{1}{n}\sum_{j=1}^{n} W_j \xrightarrow{n\to\infty} W \ \ a.s..$$

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

Under these assumptions, we have

$$X = \lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)\,\mathrm{d}\,s = \lambda W.$$

The existence of the latter limit is shown in the following proof.

Proof. Let us fix $t \geq 0$ and apply the integration by parts Formula A.13 to the processes X and $t \mapsto t$. As the second process is continuous, there are no quadratic variation terms and we have

$$tX(t) = \int_0^t X(s)\,\mathrm{d}\,s + \int_0^t s\,\mathrm{d}\,X(s).$$

But the jumps of X are those of N and D, hence

$$\begin{aligned}\int_0^t X(s)\,\mathrm{d}\,s &= \int_0^t (t-s)\,\mathrm{d}\,N(s) - \int_0^t (t-s)\,\mathrm{d}\,D(s)\\ &= \sum_{T_n\leq t}(t-T_n) - \sum_{T_n+W_n\leq t}(t-T_n-W_n).\end{aligned}$$

We must now distinguish between the customers who left the system before t, and those who entered before t, but left the system after t. We have

$$\begin{aligned}\int_0^t X(s)\,\mathrm{d}\,s &= \sum_{T_n+W_n\leq t}(t-T_n-(t-T_n-W_n)) + \sum_{T_n+W_n>t, T_n\leq t}(t-T_n)\\ &= \sum_{T_n+W_n\leq t} W_n + \sum_{T_n+W_n>t, T_n\leq t}(t-T_n).\end{aligned}$$

For all instants T_n such that $(T_n+W_n > t, T_n \leq t)$, we clearly have $0 \leq t-T_n \leq W_n$, therefore

$$\sum_{T_n+W_n\leq t} W_n \leq \int_0^t X(s)\,\mathrm{d}\,s \leq \sum_{T_n+W_n\leq t} W_n + \sum_{T_n+W_n>t, T_n\leq t} W_n = \sum_{T_n\leq t} W_n.$$

By the definition of $N(t)$,

$$\sum_{T_n\leq t} W_n = \sum_{n=1}^{N(t)} W_n,$$

hence

$$\frac{1}{t}\sum_{T_n \le t} W_n = \frac{N(t)}{t}\frac{1}{N(t)}\sum_{n=1}^{N(t)} W_n \xrightarrow{t\to\infty} \lambda W,$$

according to the two assumptions.

It remains to be proved that we have the same limit for the lower bound. This requires controlling the number of customers who entered before t and have not yet left the system at this time. Let us observe that

$$\begin{aligned}\frac{W_n}{n} &= \frac{1}{n}\left(\sum_{j=1}^{n} W_j - \sum_{j=1}^{n-1} W_j\right)\\ &= \frac{1}{n}\sum_{j=1}^{n} W_j - \frac{n-1}{n}\frac{1}{n-1}\sum_{j=1}^{n-1} W_j\\ &\xrightarrow{n\to\infty} 0, \quad \text{a.s..}\end{aligned}$$

Therefore,

$$\frac{W_n}{T_n} = \frac{W_n}{n}\frac{N(T_n)}{T_n} \xrightarrow{n\to\infty} 0.\lambda = 0, \quad \text{a.s..}$$

Let us fix the sample path. For any $\varepsilon > 0$, there exists m (depending on the path) such that $W_n \le \varepsilon T_n$ for $n \ge m$. For $n \ge m$,

$$T_n + W_n \le (1+\varepsilon)T_n.$$

If $n \ge m$ and $T_n \le t(1+\varepsilon)^{-1}$, then customer n has left before time t. Hence,

$$\sum_{T_n+W_n \le t} W_n \ge \sum_{j=m}^{N(t(1+\varepsilon)^{-1})} W_j = \sum_{j=1}^{N(t(1+\varepsilon)^{-1})} W_j - \sum_{j=1}^{m-1} W_j.$$

The same argument as for the upper bound thus gives

$$\liminf_{t\to\infty} \sum_{T_n+W_n \le t} W_n \ge \lambda W(1+\varepsilon)^{-1}.$$

As this result holds true for any $\varepsilon > 0$, we deduce from it that it is still true for $\varepsilon = 0$. The result follows by comparison of limits. □

EXAMPLE.– Take as "system" in Little's Formula, the server of a single server queue. The sojourn time in this "system" hence equals the service time. As there is 0 or 1

customer in service, the mean number of customers correspond to the rate of occupancy of the server, denoted by τ. Little's Formula entails that

$$\tau = \lambda \times 1 / \mu.$$

In other words, the traffic load ρ is the proportion of time where the server is busy.

EXAMPLE.– Let us come back to relation [1.2], associated with Figure 1.3. We view this as a system in which the nth customer arrives at time T_n, and leaves at time $T_n + Y_n$. The sojourn time of customer n is hence Y_n. According to the strong Law of Large Numbers,

$$\frac{T_n}{n} = \frac{1}{n} \sum_{j=1}^{n} (X_j + Y_j) \xrightarrow{n\to\infty} \frac{1}{\mu} + \frac{1}{\tau} \text{ almost surely.} \qquad [8.1]$$

Denoting N as the number of arrivals up to time t, we clearly have

$$T_n \leq t < T_{n+1} \Longleftrightarrow N(t) = n,$$

hence

$$\frac{N(t)}{T_{N(t)+1}} \leq \frac{N(t)}{t} \leq \frac{N(t)}{T_{N(t)}}.$$

As $N(t)$ tends to infinity almost surely, the Theorem of limits by comparison together with [8.1] implies that $N(t) / t$ tends to

$$1 / (1 / \mu + 1 / \tau) = \lambda.$$

By construction, X represents the proportion of time when the server is active, and according to Little's formula we obtain that

$$X = \frac{\lambda}{\mu}.$$

Hence we have shown in this case that the traffic load equals the product of the average number of arrivals with the average processing time.

NOTE.– Notice, that nothing in the assumptions of Little's Formula is mentioned about the service discipline. This means for instance that the average sojourn time is the same in the FIFO discipline as in the LIFO discipline. This paradoxical phenomenon means only that the average sojourn time provides a very poor information on the behavior of system. However, it is often very easy to calculate. In fact, a qualitative difference between the various service disciplines will appear when considering convex functions of the sojourn time (see section 4.1.6). In particular, the latter implies a difference in the variance and more generally, in the distribution of the sojourn time.

8.2. Single server queue

THEOREM 8.2.– *Consider a M_λ / M_μ / $1/\infty$-FIFO queue, and set $\rho = \lambda / \mu$. In the sense of Theorem 4.2, the stability condition is given by*

$$\rho < 1. \qquad [8.2]$$

In that case, the invariant probability π is given by

$$\pi(n) = \rho^n(1-\rho),\ n \in \mathbf{N}.$$

In particular, the average number of customers in the system in steady state is given by $\rho(1-\rho)^{-1}$.

Proof. We saw in example 7.1 that the process $X = (X(t),\, t \geq 0)$ denoting the number of customers in the system (also called *congestion* process of the system) is a Markov process of infinitesimal generator

$$A = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & (0) & \\ & & \ddots & & \\ & (0) & \mu & -(\lambda+\mu) & \lambda \\ & & & & \ddots \end{pmatrix}.$$

The stationary probability π of X has been explicitly given using the tools developed in Chapter 7 (see the derivation after Theorem 7.12). We obtain that for any integer i,

$$\pi(i) = \rho^i \pi(0).$$

It then follows from the normalization constraint that under condition [8.2], we have

$$\pi(0) = \frac{1}{\sum_{i \in \mathbf{N}} \rho^i} = 1 - \rho,$$

and therefore for any $i \in \mathbf{N}$,

$$\pi(i) = \rho^i(1-\rho).$$

This means in other words that under condition [8.2] which is, according to Theorem 4.2, the stability condition of the system (and thus of recurrence of the state 0 for X), the only stationary probability of X is the distribution of a random variable $X_\infty = Z - 1$, where Z follows the geometric distribution of parameter $1 - \rho$.

In particular, the average number of customers in the system at equilibrium is given by

$$\mathbf{E}\left[X_\infty\right] = \mathbf{E}\left[Z-1\right] = \frac{1}{1-\rho} - 1 = \frac{\rho}{1-\rho}. \tag{8.3}$$

□

Throughout the remainder of this section, we assume that the stability condition [8.2] is met.

For this system, an interesting performance measure is the proportion of customers entering an empty system, or (equivalently, as we shall see) the proportion of time during which the server is idle. By recalling that we denote $T_1 < T_2 < \dots$, as the arrival times of the customers in the queue, the asymptotic proportion of customers entering an empty system is given by

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{\{0\}}\left(X(T_n^-)\right) &= \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{\{0\}}(X(t^-))\,dt \\ &= \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{\{0\}}(X(t))\,dt \\ &= \mathbf{E}[\mathbf{1}_{\{0\}}\left(X_\infty\right)] \\ &= \pi(0) = 1-\rho,\end{aligned} \tag{8.4}$$

where the first almost sure equality follows from the PASTA property for the measurable function $\mathbf{1}_{\{0\}}$ (see Theorem A.38), the second one, from the fact that X has a.s. rcll paths, and the third one, from the ergodicity of the process X (Theorem 3.22). For instance, a system of load 1 / 2 is hence busy half of the time, and on a long time interval, half of the customers find an empty system (and are hence immediately attended) upon arrival.

Waiting time - sojourn time

Let us now address the waiting time (i.e. the time spent in the waiting room) and the sojourn time (i.e. the total time spent in the system) of the customers in the system. We adopt again the notation of Chapters 4 and 5, and denote for any $n \geq 1$, TA_n as the waiting time of customer C_n (which coincides in FIFO, with the workload of the server at the arrival of C_n, as studied in Chapter 4) and TS_n as the sojourn time of C_n in the whole system (waiting room + service).

According to Theorem 4.2, provided [8.2] holds there exists, for any admissible service discipline, a stationary waiting time TA and consequently, a stationary sojourn time TS in the system, given by $\text{TS} = \text{TA} + \sigma$. Little's formula entails

$$\mathbf{E}\left[\mathrm{T}\textsc{s}\right] = \frac{\mathbf{E}\left[X_\infty\right]}{\lambda} = \frac{1}{\mu - \lambda}, \qquad [8.5]$$

with [8.3]. We therefore have

$$\mathbf{E}\left[\mathrm{T}\textsc{a}\right] = \mathbf{E}\left[\mathrm{T}\textsc{s}\right] - \mathbf{E}\left[\sigma\right] = \frac{1}{\mu - \lambda} = \frac{\rho}{\mu - \lambda}. \qquad [8.6]$$

Even though the average congestion [8.3] depends only on the traffic load ρ, we can compare the waiting and sojourn times of two systems having the same traffic load $\mathrm{M}_\lambda \,/\, \mathrm{M}_\mu \,/\, 1$ and $\mathrm{M}_{\alpha\lambda} \,/\, \mathrm{M}_{\alpha\mu} \,/\, 1$, where α is a positive parameter. By adding the exponent $^\alpha$ to the parameters of the second system, we deduce from equations [8.5] and [8.6] that

$$\mathbf{E}\left[\mathrm{T}\textsc{s}^\alpha\right] = \frac{1}{\alpha}\mathbf{E}\left[\mathrm{T}\textsc{s}\right] \text{ and } \mathbf{E}\left[\mathrm{T}\textsc{a}^\alpha\right] = \frac{1}{\alpha}\mathbf{E}\left[\mathrm{T}\textsc{a}\right].$$

Thus, the average waiting and sojourn times are, for instance, divided by 2 as soon as the service and inter-arrival times are divided by 2. The output parameters hence keep the time scale factor of the input parameters.

NOTE.– The same volume of information can thus be transmitted by packets of average size m emitted at a rate λ, or by packets of average size αm emitted at a rate $\lambda \,/\, \alpha$. The traffic load ρ, and as we saw above, the average number of customers waiting in line remain the same.

Therefore, the greater the size of the packets, the larger the needed size of the buffer will be (a concrete buffer is not infinite!!). This is not obvious *a priori*, because the total amount of bytes transmitted is the same!

At constant traffic load, the average size of the packets also has an influence on the limit sojourn time, see [8.5]. Transmitting the information in larger packets therefore entails large transmission times. It seems that the solution would be to choose small packets. Unfortunately, even that solution has drawbacks. Indeed, a packet consists of useful material, and control informations (origin, destination, type, etc.). As the control informations are generally of constant size for a given protocol, limiting the packet size then amounts to limiting the proportion of useful material on the whole information. Therefore, smaller the packets, more is the decrease in the efficiency (defined as the ratio of useful information provided out of the size of the packet).

Recall, that in Chapter 5 we derive explicitly the distribution of the stationary waiting and sojourn times through their Laplace transforms. By specializing the results of Theorem 5.6 to the case where the service times follow the distribution $\varepsilon(\mu)$, we obtain the following.

THEOREM 8.3.– *The stationary waiting time* TA *and the stationary sojourn time* TS *in an M/M/1 queue admit, respectively, the Laplace transforms defined for all s by*

$$\mathbf{E}\left[e^{-s\mathrm{T_A}}\right] = (1-\rho)\left(\frac{\mu+s}{\mu(1-\rho)+s}\right); \quad [8.7]$$

$$\mathbf{E}\left[e^{-s\mathrm{T_S}}\right] = (1-\rho)\left(\frac{\mu}{\mu(1-\rho)+s}\right). \quad [8.8]$$

The PASTA property also allows us to show that the stationary distribution of the congestion process coincides with that of its embedded chain.

LEMMA 8.4.– *Under the condition [8.2], the Markov chain $\widehat{X} = (X(T_n^-), n \geq 0)$ is irreducible, positive recurrent and of stationary distribution π.*

Proof. It is obvious that the chain is irreducible and recurrent, so let $\widehat{\pi}$ be its invariant distribution. According to Theorem 7.13, for all $f \in l^\infty(\mathbf{N})$ we have that

$$\frac{1}{N}\sum_{n=1}^{N} f(X(T_n^-)) \xrightarrow{N\to\infty} \int_{\mathbf{N}} f \,\mathrm{d}\,\widehat{\pi}.$$

As $N(t)$ takes values in $\mathbf{N}$ and tends to infinity as t tends to infinity, we have by extraction

$$\frac{1}{N(t)}\sum_{n=1}^{N(t)} f(X(T_n^-)) \xrightarrow{t\to\infty} \int_{\mathbf{N}} f \,\mathrm{d}\,\widehat{\pi}.$$

But in view of Theorem A.38,

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{N(t)}\sum_{n=1}^{N(t)} f(X(T_n^-)) &= \lim_{t\to\infty}\frac{1}{N(t)}\int_0^t f(X(s^-))\,\mathrm{d}\,N(s)\\ &= \lim_{t\to\infty}\frac{1}{t}\int_0^t f(X(s^-))\,\mathrm{d}\,s = \int_{\mathbf{N}} f\,\mathrm{d}\,\pi.\end{aligned}$$

By identification, it follows that $\widehat{\pi} = \pi$. □

NOTE.– Setting up a connection is an investment that must be profitable. Therefore, the connection should be used as much as possible, in other words it should carry on a traffic load as close as possible to 1. Unfortunately, according to [8.6], such a load induces some significant delays, and hence a limited quality of service. Fortunately, in real life the connections are loaded at 10% of their capacity.

8.3. Multiple server queue

We now turn to the Markovian queue with S servers ($S \geq 1$) and waiting room of unlimited capacity, denoted by $\mathrm{M}_\lambda / \mathrm{M}_\mu / S / \infty$-FIFO in the usual nomenclature. This model is represented by the process $\big(X^{(S)}(t),\, t \geq 0\big)$, thereby counting the total number of customers in the system, with values in $\mathbf{N}$. This process is naturally Markov, we determine hereafter its infinitesimal generator:

– for any $i \in [\![1,\, S]\!]$, if the process is in state i there are i customers in the system, who are all in service. So the sojourn time of the process in state i (which will denoted by Y_i for any i) is the minimum of a random variable $U \sim \varepsilon(\lambda)$ counting the current inter-arrival time, and i random variables independent and identically distributed $V_1,\, V_2,\, \ldots,\, V_i$ of distribution $\varepsilon(\mu)$, counting the residual service times of the i customers in service. Then, the process jumps to $i-1$ if $Y_i = V_k$ for some k and to $i+1$ if $Y_i = U$;

– for all $i > S$, in state i all the S servers are busy, so $i - S$ customers are waiting. Thus, Y_i equals the minimum of S random variables $V_1,\, V_2,\, \ldots,\, V_S$ of distribution $\varepsilon(\mu)$ and of the random variable U having distribution $\varepsilon(\lambda)$, keeping the previous notations. The process then jumps to $i-1$ or to $i+1$ as in the previous case;

– The sojourn time Y_0 in 0 has the distribution $\varepsilon(\lambda)$, and the process almost surely leaves 0 to go to 1.

According to Definition 7.1, the transition matrix $Q^{(S)}$ of the process hence reads as follows.

– $q^{(S)}(0,0) = \lambda$ and $q(0,1) = 1$;

– for all $i \in [\![1, S]\!]$, $q^{(S)}(i,i) = i\mu + \lambda$ and the only non-zero transitions are

$$q^{(S)}(i,i+1) = \frac{\lambda}{i\mu + \lambda} \text{ and } q^{(S)}(i,i-1) = \frac{i\mu}{i\mu + \lambda};$$

– for all $i > S$, $q^{(S)}(i,i) = S\mu + \lambda$ and the only non-zero transitions are

$$q^{(S)}(i,i+1) = \frac{\lambda}{S\mu + \lambda} \text{ and } q^{(S)}(i,i-1) = \frac{S\mu}{S\mu + \lambda}.$$

Therefore, in view of [7.11] the process $\big(X^{(S)}(t),\, t \geq 0\big)$ is Markov, with infinitesimal generator given by

$$A^{(S)} = \left(\begin{array}{ccccccc}
-\lambda & \lambda & & & & & \\
\mu & -(\lambda+\mu) & \lambda & & & & \\
0 & 2\mu & -(\lambda+2\mu) & \lambda & & (0) & \\
 & & & \ddots & & & \\
 & (0) & & S\mu & -(\lambda+S\mu) & \lambda & \\
 & & & & S\mu & -(\lambda+S\mu) & \lambda \\
 & & & & & & \ddots
\end{array}\right).$$

The proof of the following Theorem uses the same arguments as that of Theorem 8.2, it is hence left to the reader.

THEOREM 8.5.– *Under the stability condition*

$$\rho < S, \qquad [8.9]$$

the only stationary probability $\pi^{(S)}$ *of the process* $\left(X^{(S)}(t),\, t \geq 0\right)$ *reads*

$$\begin{aligned}
\pi^{(S)}(0) &= \left(\sum_{k=0}^{S-1} \frac{\rho^k}{k!} + \frac{S\rho^S}{S!(S-\rho)}\right)^{-1}; \\
\pi^{(S)}(i) &= \frac{\rho^i}{i!}\pi(0) \textit{ for all } i \in [\![1, S-1]\!]; \\
\pi^{(S)}(i) &= \frac{\rho^i}{S^{i-S}S!}\pi(0) \textit{ for all } i \geq S.
\end{aligned} \qquad [8.10]$$

Under condition [8.9], we can deduce from Theorem 8.5, as in [8.4], the limiting rate of customers who must wait to be attended. It is given by the so-called Erlang-C Formula

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{\{[S,+\infty)\}}\left(X^{(S)}(T_n-)\right) &= \mathbf{E}\left[\mathbf{1}_{\{[S,+\infty)\}}\left(X_\infty^{(S)}\right)\right] \\
&= \sum_{i=S}^{+\infty} \pi^{(S)}(i) = \frac{\pi^{(S)}(0)}{S!S^{-S}} \sum_{i=1}^{+\infty} \left(\frac{\rho}{S}\right)^i \\
&= \left(\sum_{k=0}^{S-1} \frac{\rho^k}{k!} + \frac{S\rho^S}{S!(S-\rho)}\right)^{-1} \frac{S\rho^S}{S!(S-\rho)} \\
&=: C(S,\, \rho).
\end{aligned}$$

We can therefore implement a simple algorithm for dimensioning a multiple server queue, being fixed a quality of service constraint in terms of proportion of customers put on hold.

Algorithm 8.1. Dimensioning of a multi-server system with guaranteed holding rate

```
Data: $\rho$, $p$
Result: The optimal number of servers $S$, given a traffic load $\rho$ and the
        guarantee of a proportion $p$ of users put on hold.
$S \leftarrow 1$;
until
    $p < \left(\sum_{k=0}^{S-1} \frac{\rho^k}{k!} + \frac{S\rho^S}{S!(S-\rho)}\right)^{-1} \frac{S\rho^S}{S!(S-\rho)}$
do
    $S \leftarrow S + 1$;
end
return $S$
```

The average number of customers in the system at equilibrium is given by

$$\begin{aligned}
\mathbf{E}\left[X_\infty^{(S)}\right] &= \sum_{i=0}^{\infty} i\pi^{(S)}(i) \\
&= \pi^{(S)}(0)\left(\sum_{i=0}^{S} \frac{i\rho^i}{i!} + \frac{S^{S-1}\rho}{S!} \sum_{i=S+1}^{\infty} i\left(\frac{\rho}{S}\right)^{i-1}\right) \\
&= \pi^{(S)}(0)\rho\left(\sum_{i=0}^{S-1} \frac{\rho^i}{i!} + \frac{S^{S-1}}{S!}\frac{d}{dx}\left(\sum_{i=S+1}^{\infty} x^i\right)\left(\frac{\rho}{S}\right)\right) \\
&= \pi^{(S)}(0)\rho\left(\sum_{i=0}^{S-1} \frac{\rho^i}{i!} + \frac{\rho^S}{S!}\frac{S^2 - S\rho + S}{(S-\rho)^2}\right).
\end{aligned}$$

According to Little's Formula, the sojourn time $\mathrm{Ts}^{(S)}$ and waiting time $\mathrm{TA}^{(S)}$ in steady state therefore have the following respective mean expectations

$$\mathbf{E}\left[\mathrm{Ts}^{(S)}\right] = \frac{\pi^{(S)}(0)}{\mu}\left(\sum_{i=0}^{S-1} \frac{\rho^i}{i!} + \frac{\rho^S}{S!}\frac{S^2 - S\rho + S}{(S-\rho)^2}\right); \quad [8.11]$$

$$\mathbf{E}\left[\mathrm{TA}^{(S)}\right] = \frac{1}{\mu}\left[\pi^{(S)}(0)\left(\sum_{i=0}^{S-1} \frac{\rho^i}{i!} + \frac{\rho^S}{S!}\frac{S^2 - S\rho + S}{(S-\rho)^2}\right) - 1\right]. \quad [8.12]$$

Relation [8.12] allows one to dimension a system, having guaranteed an average waiting time.

Algorithm 8.2. Dimensioning of a multi-server system with guaranteed waiting time

```
Data: $\rho$, $\mu$, Ts
Result: The optimal number of servers $S$ given the traffic load $\rho$ and a
        guaranteed average waiting time Ts.
$S \leftarrow 1$;
until
    $\text{Ts} < \frac{1}{\mu}\left[\left(\sum_{i=0}^{S-1}\frac{\rho^i}{i!}+\frac{S\rho^S}{S!(S-\rho)}\right)^{-1}\left(\sum_{i=0}^{S-1}\frac{\rho^i}{i!}+\frac{\rho^S}{S!}\frac{S^2-S\rho+S}{(S-\rho)^2}\right)-1\right]$
do
    $S \leftarrow S+1$;
end
return $S$
```

In addition, as in the case of the single server queue, we can derive the distribution of the stationary waiting time through its Laplace transform:

THEOREM 8.6.– *If $\rho < S$, the waiting time* TA_n *converges in distribution to* $\text{TA}^{(S)}$, *whose Laplace transform is given by*

$$\mathbf{E}\left[e^{-s\text{TA}^{(S)}}\right] = 1 - C(S,\,\rho) + \pi^{(S)}(0)\frac{\rho^S}{S!}\frac{S\mu}{s+(S-\rho)\mu}, \tag{8.13}$$

which is equivalent to saying that

$$d\mathbf{P}_{\text{TA}^{(S)}}(x) = (1 - C(S,\rho))\delta_0(x) + \pi_0\frac{\rho^S}{S!}S\mu\exp(-(S-\rho)\mu x)\mathbf{1}_{\mathbf{R}^+}(x)\,\mathrm{d}\,x.$$

NOTE.– The last relation means that the waiting time is zero with probability $1-C(S,\rho)$ (which is obvious) and that conditionally to being positive, $\text{TA}^{(S)}$ is exponentially distributed with the parameter $(S-\rho)\mu$.

Proof. As the service times follow exponential distributions, conditionally to the fact that all servers are busy, the inter-departure times are all exponentially distributed with parameters $S\mu$, and independent of each other. We therefore have that

$$\text{W}_n \overset{\text{law}}{=} \sum_{j=1}^{\hat{X}_n-S+1} \eta_j\,\mathbf{1}_{\{\hat{X}_n\geq S\}},$$

where $(\eta_k,\ k \geq 0)$ is a sequence of random variables independent and identically distributed of exponential distribution with parameter $S\mu$. Therefore,

$$\mathbf{E}\left[e^{-s\mathrm{W}_n} | \hat{X}_n\right] = \left(\frac{S\mu}{s+S\mu}\right)^{\hat{X}_n - S+1} \mathbf{1}_{\{\hat{X}_n \geq S\}} + \mathbf{1}_{\{\hat{X}_n < S\}} \,.$$

From Theorem 8.5, $\hat{X}_n$ converges in distribution to X. As this conditional expectation is a bounded function of $\hat{X}_n$ if $\rho < S$, $\mathbf{E}\left[e^{-s\mathrm{W}_n}\right]$ converges simply for any s toward

$$\mathbf{E}\left[\left(\frac{S\mu}{s+S\mu}\right)^{\hat{X}_n - S+1} \mathbf{1}_{\{\hat{X}_n \geq S\}} + \mathbf{1}_{\{\hat{X}_n < S\}}\right] = \mathbf{E}\left[e^{-s\mathrm{T}\mathrm{A}^{(S)}}\right].$$

We therefore obtain

$$\begin{aligned}
\mathbf{E}\left[e^{-s\mathrm{T}\mathrm{A}^{(S)}}\right] &= 1 - C(S,\rho) + \sum_{j=S}^{\infty} \left(\frac{\rho}{S}\frac{S\mu}{s+S\mu}\right)^{j-S+1} \\
&= 1 - C(S,\rho) + \pi^{(S)}(0)\frac{\rho^S}{S!}\frac{S\mu}{s+S\mu}\sum_{j=0}^{\infty}\left(\frac{\rho\mu}{s+S\mu}\right)^{j} \\
&= 1 - C(S,\rho) + \pi_0\frac{\rho^S}{S!}\frac{S\mu}{s+(S-\rho)\mu}.
\end{aligned}$$

□

8.3.1. *Comparison of systems*

Starting from a simple queue, we compare hereafter qualitatively three types of operations aiming to improve the performances of the system: the *multiplexing* of the resources, the *parallelism*, and the *speed* of execution.

We consider the following four systems, all subject to Poisson arrivals with intensity λ, of customers requesting service times of distribution $\varepsilon(\mu)$. Hence, the traffic load always equals $\rho = \lambda/\,\mu$ Erlang, and it is assumed that $\rho < 2$. We assume further that all considered servers work in FCFS, and that the waiting rooms are always of unlimited capacity:

– system A has one server working at unit speed, and hence corresponds to a simple $\mathrm{M}_\lambda\,/\,\mathrm{M}_\mu\,/\,1$ queue;

– in system B, the customers are redirected with probability $1/2$ (independently from a customer to the other, and independently of all other random variables) toward one or the other of two independent systems operating in parallel, each of those having one server working at unit speed. According to Theorem 6.6, each of the two parallel queues is hence a $\mathrm{M}_{\frac{\lambda}{2}}\,/\,\mathrm{M}_\mu\,/\,1$ queue;

– system C is a single queue with two servers, each of those working at unit speed. C is hence exactly a $M_\lambda\,/\,M_\mu\,/\,2$ queue;

– finally, system D has one server working at double speed: the customer C_n who requires a service time σ_n actually needs only a duration $\sigma_n/\,2$ to be serviced. It is then easy to see that the service times of the customers are i.i.d of distribution $\varepsilon(2\mu)$, and D hence corresponds to a $M_\lambda\,/\,M_{2\mu}\,/\,1$ queue.

In the following, we add respectively the exponents A, B, C, and D to the characteristics of the different systems.

According to [8.5], the average sojourn time of a customer at equilibrium in the system D is given by

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{D}}\right] = \frac{1}{2\mu - \lambda}, \qquad [8.14]$$

whereas in A, provided $\rho < 1$,

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{A}}\right] = \frac{1}{\mu - \lambda}. \qquad [8.15]$$

We now consider the system B. Let X^1 and X^2 denote, respectively, the processes counting the number of customers in each queue and, for any t,

$$X_t^{\mathrm{B}} = X_t^1 + X_t^2$$

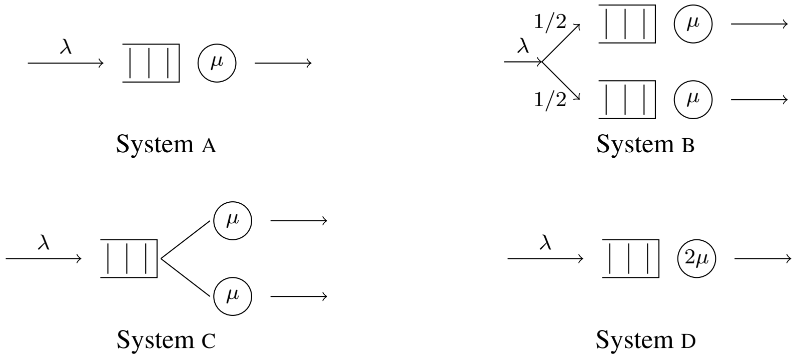

Table 8.1. *Comparison of four systems. The traffic load is ρ in* A *and $\rho\,/\,2$ in the other three*

represent the total number of customers in the system at t. As $\rho < 2$, X^1 and X^2 are ergodic, and by denoting X^1_∞ and X^2_∞ as their limits in distribution, we have

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t}\int_0^t X^{\mathrm{B}}_s\, ds &= \lim_{t\to\infty} \frac{1}{t}\int_0^t X^1_s\, ds + \lim_{t\to\infty} \frac{1}{t}\int_0^t X^2_s\, ds \\
&= \mathbf{E}\left[X^1_\infty\right] + \mathbf{E}\left[X^2_\infty\right] \\
&= \frac{\lambda / 2}{\mu - \lambda / 2} + \frac{\lambda / 2}{\mu - \lambda / 2} \\
&= \frac{2\lambda}{2\mu - \lambda},
\end{aligned} \tag{8.16}$$

in view of [8.3]. According to Theorem 4.20, there exists provided $\rho < 2$ a stationary waiting time, and hence a stationary sojourn time Ts^{B} for this system, defined by the almost sure limit

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathrm{Ts}^{\mathrm{B}}_n = \mathbf{E}\left[\mathrm{Ts}^{\mathrm{B}}\right],$$

where for any n, $\mathrm{Ts}^{\mathrm{B}}_n$ denotes the sojourn time proposed to the nth customer. Little's formula applies to this system. With [8.16] we therefore have that

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{B}}\right] = \frac{2}{2\mu - \lambda}. \tag{8.17}$$

Finally, from [8.11],

$$\begin{aligned}
\mathbf{E}\left[\mathrm{Ts}^{\mathrm{C}}\right] &= \frac{1}{\mu}\pi(0)\left(1 + \rho + \frac{\rho^2}{2}\frac{4 - 2\rho + 2}{(2-\rho)^2}\right) \\
&= \frac{1}{\mu}\frac{2-\rho}{2+\rho}\left(\frac{4}{(2+\rho)(2-\rho)}\right) \\
&= \frac{4\mu}{(2\mu - \lambda)(2\mu + \lambda)}.
\end{aligned} \tag{8.18}$$

Gathering [8.14, 8.15, 8.17] and [8.18], we obtain that provided $\rho < 2$,

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{B}}\right] - \mathbf{E}\left[\mathrm{Ts}^{\mathrm{C}}\right] = \frac{2\lambda}{(2\mu - \lambda)(2\mu + \lambda)};$$

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{C}}\right] - \mathbf{E}\left[\mathrm{Ts}^{\mathrm{D}}\right] = \frac{1}{2\mu + \lambda}.$$

Consequently, we have that

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{D}}\right] < \mathbf{E}\left[\mathrm{Ts}^{\mathrm{C}}\right] < \mathbf{E}\left[\mathrm{Ts}^{\mathrm{B}}\right] < \mathbf{E}\left[\mathrm{Ts}^{\mathrm{A}}\right]; \tag{8.19}$$

$$\mathbf{E}\left[\mathrm{Ts}^{\mathrm{D}}\right] < \frac{1}{2}\mathbf{E}\left[\mathrm{Ts}^{\mathrm{A}}\right], \tag{8.20}$$

where [8.20] and the last inequality of [8.19] make sense only if $\rho < 1$. When considering the time spent in the system, it is then preferable to double the speed of execution than doubling the number of resources, which is in turn more efficient than two parallel systems. In addition, according to [8.20], doubling the speed of service achieves more than the double of the system efficiency.

8.4. Processor sharing queue

We briefly address the case of the Markovian $\mathrm{M}_\lambda\,/\,\mathrm{M}_\mu\,/\,\mathrm{PS}$ queue, studied in the general case in section 4.2. Customers enter the system according to a Poisson process of intensity λ, requesting service times of distribution $\varepsilon(\mu)$, and are immediately attended by a processor sharing server : all are served simultaneously, at a rate that is inversely proportional to the number of customers in service.

Let us denote X^{PS}, the process counting the number of customers in the system. It is easily seen that this process is Markov, and to give its generator. For all $t \geq 0$, we denote as above:

– $W(t)$, the residual time at t before the next arrival;

– $R(t)$, the residual time at t before the next departure.

Let $i \geq 1$. On the event $\left\{X^{\mathrm{PS}}(t) = i\right\}$, let us denote $R_1(t),\ R_2(t), \ldots,\ R_i(t)$, the residual service times of the i customers in service at t, in time unit. According to Theorem 7.3, these service times are independent and of distribution $\varepsilon(\mu)$. As long as there is no new arrival after t (i.e. up to $t + W(t)$), the server works at a speed of $1\,/\,i$, which multiplies the time scale by this factor. Thus, from the perspective of the server, the residual service times of the customers follow the distribution $\varepsilon(\mu\,/\,i)$ and then:

(i) if $\min_{j=1,\,\ldots,\,i} R_j(t) \leq W(t)$, the server works at the same speed until the next departure, which takes place before the next arrival, and therefore $R(t)$ is the minimum of i independent random variables of distribution $\varepsilon(\mu\,/\,i)$;

(ii) if $W(t) < \min_{j=1,\,\ldots,\,i} R_j(t)$, at time $t + W(t)$ a new arrival occurs and hence there are $i+1$ customers in the system. Again, according to Theorem 7.3, the statistics of the system do not change if we draw once again the service times of the $i+1$ customers, according to the distribution $\varepsilon(\mu)$. Therefore $R(t)$ has the same distribution as the minimum of $i+1$ independent random variables with distribution $\varepsilon(\mu\,/\,(i+1))$ unless there is another arrival before the next departure, in which case we draw once again the $i+2$ service times according to the same distribution $\varepsilon(\mu)$, and so on.

In conclusion, in all cases the distribution of $R(t)$ follows, conditionally to $\left\{X^{\mathrm{PS}}(t) = i\right\}$, the distribution $\varepsilon(\mu)$. As $W(t)$ follows for all t the distribution $\varepsilon(\lambda)$, as is the case for the previous systems, we obtain the following result.

THEOREM 8.6.1.– *The process X^{PS} has the same distribution as the congestion process X of the $M_\lambda/M_\mu/1$ queue. It therefore admits the same stationary distribution,* *provided* $\rho < 1$. *Especially, according to Little's formula the average sojourn time in steady state is the same as that of the M/M/1 queue.*

NOTE.– This does not mean that the processes X and X^{PS} have the same paths almost surely! This is obviously not the case, and the latter identity holds true only in distribution.

8.5. The M / M / ∞ queue

The M / M / ∞ queue is obviously a theoretical illusion, because no system can have an infinite number of servers. However, this object is used in several situations (at least for comparison), for example in Theorem 10.14.

Let $(X^\infty(t),\, t \geq 0)$ be the process counting the number of customers in the system (and thus, in service). As above, it is easily checked that for any $i \in \mathbf{N}$,

– the process X^∞ stays in state i during a time of distribution $\varepsilon(i\mu + \lambda)$;

– it leaves state i to go to state $i+1$ with probability $\lambda \,/\, (\lambda + i\mu)$ and provided that $i \geq 1$, to state $i-1$ with probability $\lambda \,/\, (\lambda + i\mu)$.

Consequently, the process $(X^\infty(t),\, t \geq 0)$ is Markov, with infinitesimal generator

$$A^\infty = \begin{pmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & & (0) & \\ & & \ddots & \ddots & \ddots & & \\ & (0) & & i\mu & -(\lambda+i\mu) & \lambda & \\ & & & & (i+1)\mu & -(\lambda+(i+1)\mu) & \lambda \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

The theory developed in Chapter 7 does not apply as such, because the coefficients of the infinitesimal generator are unbounded. By using a widely different representation of this queue (see Example 10.3), we can still deduce that the only stationary probability π^∞ of $(X^\infty(t),\, t \geq 0)$ reads

$$\pi^\infty(i) = \frac{\rho^i e^{-\rho}}{i!} \text{ for all } i \in \mathbf{N},$$

setting as above $\rho = \lambda \,/\, \mu$.

NOTE.– Notice however that π^∞ is as a matter of fact the only solution of the system

$$\begin{cases} \pi A^\infty & = \mathbf{0}, \\ \sum_{i \in \mathbf{N}} \pi(i) & = 1. \end{cases}$$

This result means in other words, that the limiting size of the system X_∞^∞ follows a Poisson distribution with parameter ρ. In particular, the average size of the system at equilibrium equals

$$\mathbf{E}\left[X_\infty^\infty\right] = \rho. \qquad [8.21]$$

8.6. The departure process

Consider one of the three models considered in the previous sections (single, multiple or infinite server queue). We show hereafter, that whenever the stability condition is met, the distribution of the departure process is identifiable.

Let us recall that we denote for any $n \in \mathbf{N}$, T_n as the instant of the nth arrival, and T'_n as the instant of the nth departure. We still denote $(N(t),\, t \geq 0)$ as the arrival process and $(D(t),\, t \geq 0)$ as the departure process, that is for any t,

$$D(t) = \sum_{n \in \mathbf{N}} \mathbf{1}_{T'_n \leq t}\,.$$

At first glance, we might think that the departure process is such that the time between two departures is of distribution $\varepsilon(\mu)$. We will see that this is not the case.

Let us start with a heuristic on the M / M / 1 queue to deny this idea. Let us place ourselves at a departure time, say T'_n, the departure time of C_n. So T'_{n+1} equals

– $T'_n + \sigma_{n+1}$, if C_{n+1} is in the queue at the departure of C_n;

– $T_{n+1} + \sigma_{n+1}$, if C_n leaves an empty system behind.

So, if we assume that $T'_{n+1} - T'_n$ is independent of the past of T'_n (property that shall be demonstrated hereafter), we have that

$$\begin{aligned}\mathbf{E}\left[T'_{n+1} - T'_n\right] &= \mathbf{E}\left[\sigma_{n+1}\right] + \mathbf{E}\left[\left(T_{n+1} - T'_n\right)\mathbf{1}_{X(T'_n)=0}\right]\\ &= \frac{1}{\mu} + \frac{1}{\lambda}\mathbf{P}(X(T'_n) = 0),\end{aligned}$$

according to Theorem 6.7. Thus, at equilibrium the average time between two departures equals

$$\frac{1}{\mu} + \frac{1}{\lambda}\pi(0) = \frac{1}{\lambda},$$

as much as the average time between two arrivals!

The following Theorem makes this result precise, and shows that in steady state, the departure process is in fact, like that of arrivals, a Poisson process of intensity λ.

THEOREM 8.7 (Burke's Theorem).– *In steady state, the departure process is a Poisson process of intensity* λ*. Moreover, for any* $s \geq 0$*, the departures times after* s *are independent of the arrival times before* s*, i.e.* $\sigma(D(u);\, u \geq s)$ *is independent of* $\sigma(N(u);\, u \leq s)$*.*

Proof. In all the cases, the congestion process $(X(t),\, t \geq 0)$ is a birth and death process (see section 7.5). It is therefore reversible, with stationary probability π. In particular, by assuming that X has the initial distribution π, it has for any $U \geq 0$ the same distribution as its reversed process $\big(\bar{X}^U(t),\, t \geq 0\big)$.

It is clear that for any $n \in \mathbf{N}$, T'_n (respectively T_n) is the nth instant of downward (respectively upward) jump of the process X. Moreover, for any $U \geq 0$ the instants of downward jumps from X up to U correspond to the instants of upward jumps of the reversed process: the original process decreases by one when the reversed process increases by one. Specifically, if we denote

$$\tilde{T}'_n = T'_n \wedge U,\ \ n \in \mathbf{N},$$

as the points of the process $(D(t),\, t \geq 0)$ restricted to $[0, U]$, we first have

$$\begin{aligned}\tilde{T}'_1 &= \inf\{t \leq U; X(t) = X(t^-) - 1\}\\ &= S - \sup\{s \leq U; X(S-s) = X\left((S-s)^-\right) - 1\}\\ &= S - \sup\{s \leq U; \bar{X}^U(s) = \bar{X}^U(s^-) + 1\}.\end{aligned}$$

The latter supremum is the last upward jump instant of the reversed process before U. Since X is reversible, this sup equals in distribution the last upward jump instant of X (i.e. the last arrival) before time U, that is $T_{N(U)}$. According to Theorem 6.7, we therefore have the following identity in law

$$\tilde{T}'_1 \overset{\mathcal{L}}{=} U - T_{N(U)} \overset{\mathcal{L}}{=} U \wedge Y_1,$$

where Y_1 is a random variable of distribution $\varepsilon(\lambda)$. Similarly, we can show that

$$\tilde{T}'_2 - \tilde{T}'_1 \overset{\mathcal{L}}{=} \big(U - \tilde{T}'_1\big) \wedge Y_2,$$

where Y_2 is independent of Y_1 and of distribution $\varepsilon(\lambda)$, and so on.

This shows that $(D(t),\, t \geq 0)$ is equal in distribution to a Poisson process on $[0, U]$. As this is true for any $U \geq 0$, $(D(t),\, t \geq 0)$ is a Poisson process. Finally, the independence property results, naturally, from that of the arrival process. □

8.7. Queueing Networks

In this section, we present the main stability results on the simplest model of *queueing networks*, that is a set of queues in which customers leaving a queue may join another one, to receive a new service.

8.7.1. *Open Jackson networks*

We first consider the following system:

– a Poisson process of intensity $\lambda > 0$, which is interpreted as the arrival process from the outside into the system;

– a set of N queues, where the i-th queue is of type $./\mathrm{M}_{\mu_i}/S_i/\infty$-FIFO : the customers entering the i-th queue request for service times that are independent and identically distributed of distribution $\varepsilon(\mu_i)$, to a group of S_i servers. The distribution of inter-arrival times in the queue is not known *a priori* because it depends on the other queues, as we shall see;

– a Markovian matrix P of size $N + 1$: for any $i \in [\![0, N]\!]$, $P_{ij} \in [0, 1]$ and $\sum_{j=0}^{N} P_{ij} = 1$. The matrix P is called the *routing matrix* of the system: as soon as a customer has finished his service in the queue i, he makes a draw that is independent of all the other parameters, in order to decide the next queue in which he will request a service. For any $j \in [\![0, N]\!]$, he then joins the queue j with probability P_{ij}. The "queue 0" represents here the "outside" of the system: customers moving from 0 to j arrive directly from outside into the queue j, and those going from i to 0 leave the system after having visited the queue i. Let us notice, that if $P_{ii} > 0$ for some i, a customer may get back in the same queue i just after having received service in the same queue. Let us assume that P satisfies the following two conditions

$$P_{00} = 0, \tag{8.22}$$

and for any $i \in [\![1,\ N]\!]$, there exists $n \in \mathbf{N}$ and a n-uple $\{i_1,\ i_2,\ \ldots,\ i_n\}$ of elements of $[\![1,\ N]\!]$ containing i and such that

$$P_{0i_1}P_{i_1i_2}\ldots P_{i_{n_1}i_n}P_{i_n0} > 0. \tag{8.23}$$

Condition [8.23] thus ensures that any queue i is part of a possible path, going from the outside to itself.

The system is described by the process

$$(X(t), t \geq 0) = ((X_1(t),\ X_2(t),\ \ldots,\ X_N(t)),\ t \geq 0)\,,$$

where for any i, $X_i(t)$ counts the number of customers in the i-th queue at time t. As the traffic in each queue depends on the other queues, it is easy to see that each process $(X_i(t), t \geq 0)$ alone is not Markov. This is the case, however, for the process $(X(t), t \geq 0)$, as will be shown in the following lemma. We denote (see the notations of appendix A) for any $i \in \mathbf{N}$,

$$\mathbf{e}_i = (0,\ \ldots,\ 0,\ \underbrace{1}_{i},0,\ \ldots,\ 0)$$

and for any $k \in \mathbf{N}$,

$$\mu_i(k) = \mu_i(k \wedge S_i).$$

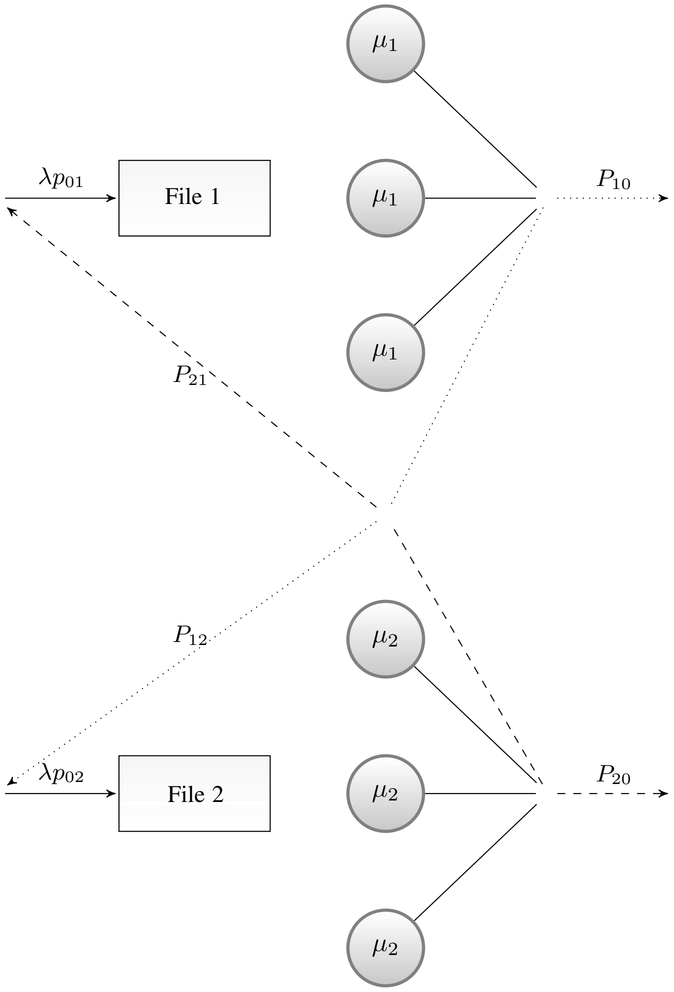

Figure 8.1. *An open Jackson network with two queues*

LEMMA 8.8.– *The process $(X(t), t \geq 0)$ describing the open Jackson's network is Markov, of infinitesimal generator A^{OJ} defined for any $x = (x(1), \ldots, x(N))$ by*

$$\begin{cases} A^{\text{OJ}}(x, x + \mathbf{e}_j - \mathbf{e}_i) = \mu_i(x(i))P_{ij}; \\ A^{\text{OJ}}(x, x + \mathbf{e}_j) = \lambda P_{0j}; \\ A^{\text{OJ}}(x, x - \mathbf{e}_i) = \mu_i(x(i))P_{i0}, \end{cases}$$

all the other coefficients A^{OJ} (x, y) being zero and the diagonal coefficients $A^{\text{OJ}}(x, x)$, equal to the opposite of the sum of the $A^{\text{OJ}}(x, y)$ for $y \neq x$.

Proof. Let us assume that the process $(X(t), t \geq 0)$ is in state x at t. Then, the process may directly leave x only to go to the following states:

1) $x + \mathbf{e}_j - \mathbf{e}_i$ if $x(i) > 0$ and if a customer in service in queue i completes his service, and then goes to the queue j;

2) $x - \mathbf{e}_i$ if $x(i) > 0$ and if a customer in service in queue i completes his service, and then leaves the system;

3) $x + \mathbf{e}_j$ if a customer enters from outside toward the queue j.

According to Theorem 6.6, the process of arrivals from the outside toward the queue j is Poisson of intensity λP_{0j}. The residual time before the next point of this process therefore follows the distribution $\varepsilon(\lambda P_{0j})$. Then, the residual time before the next end of service among the $x(i) \wedge S_i$ services currently in queue i follows, as seen above, the distribution $\varepsilon((x(i) \wedge S_i)\mu_i) = \varepsilon(\mu_i(x(i)))$. By denoting for any i the event

$$B_i = \{\text{The first service that ends is a service of queue } i\},$$

the probability that the process actually leaves x for $x - \mathbf{e}_i + \mathbf{e}_j$ (respectively $x - \mathbf{e}_i$) is given by $\mathbf{P}(B_i \cap \{ \text{ the customer leaves } i \text{ for } j\}) = \mathbf{P}(B_i)P_{ij}$ (respectively $\mathbf{P}(B_i)P_{i0}$). This concludes the proof, in view of the above-mentioned properties of exponential distributions. □

We will need the following technical result in order to characterize the steady state of $(X(t), t \geq 0)$.

LEMMA 8.9.– *The system*

$$\lambda_j = \lambda P_{0j} + \sum_{i=1}^{N} \lambda_i P_{ij}, \qquad [8.24]$$

of unknown $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ *and called* traffic equation*, admits a unique solution in* $(\mathbf{R}+)^N$.

Proof. As the matrix P is Markovian, there exists a unique Markov chain $(M_n,\ n \in \mathbf{N})$ with values in $[\![0, N]\!]$ and with transition matrix P. For any pair (i, j) of elements of $[\![1, N]\!]$, there exists according to [8.23] two finite families $\{i_1, \ldots, i_n\}$ and $\{j_1, \ldots, j_p\}$ of elements of $[\![1, N]\!]$, including i and j, respectively, and such that

$$P_{0i_1} \ldots P_{i_{n-1}i_n} P_{i_n0} > 0 \quad \text{and} \quad P_{0j_1} \ldots P_{j_{p-1}j_p} P_{j_p0} > 0.$$

Therefore, with the notations of Chapter 3, there exists an integer $q < n + p$ such that the probability $p^{(q)}(i, j)$ that $(M_n,\ n \in \mathbf{N})$ goes from i to j in q steps verifies

$$p^{(q)}(i, j) \geq P_{0i_1} \ldots P_{i_{n-1}i_n} P_{i_n0} P_{0j_1} \ldots P_{j_{p-1}j_p} P_{j_p0} > 0.$$

The Markov chain $(M_n,\, n \in \mathbf{N})$ is thus irreducible on the finite state space $[\![0, N]\!]$. Hence it is positive recurrent, and according to Theorem 3.16, there exists up to a multiplicative coefficient, a single stationary measure ν on $[\![0, N]\!]$ which, represented as a row vector, satisfies the matrix equation

$$\nu = \nu P. \tag{[8.25]}$$

On the other hand, if $\{\lambda_1, \ldots, \lambda_N\}$ is a solution of [8.24], as P is a Markov matrix,

$$\begin{aligned}\sum_{i=1}^{N} \lambda_i P_{i0} &= \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{j=1}^{N} P_{ij}\right)\\ &= \sum_{i=1}^{N} \lambda_i - \sum_{j=1}^{N}\sum_{i=1}^{N} \lambda_i P_{ij}\\ &= \sum_{i=1}^{N} \lambda_i - \sum_{j=1}^{N} (\lambda_j - \lambda P_{0j})\\ &= \lambda \sum_{i=1}^{N} P_{0i}\\ &= \lambda,\end{aligned} \tag{[8.26]}$$

in view of [8.22]. This shows that $\{\lambda_1, \ldots, \lambda_N\}$ is a solution to [8.24] if and only if $\Lambda = (\lambda, \lambda_1, \ldots, \lambda_N)$ is a solution to [8.25]. The single solution of equation [8.24] is the only invariant measure having λ as first component. □

For any $i \in [\![1, N]\!]$, we know from the results of section 8.3 that provided $\lambda_i < \mu_i S_i$, the congestion process of the M$_{\lambda_i}$ / M$_{\mu_i}$ / S_i queue admits the invariant probability π^i given by

$$\pi^i(0) = \left(\sum_{k=0}^{\infty} \frac{(\lambda_i)^k}{\prod_{j=1}^{k} \mu_i(j)}\right)^{-1}; \tag{[8.27]}$$

$$\pi^i(k) = \frac{(\lambda_i)^k}{\prod_{j=1}^{k} \mu_i(j)} \pi^i(0),\ i \geq 1. \tag{[8.28]}$$

We can therefore state the following result.

THEOREM 8.10.– *Let* $\{\lambda_1, \ldots, \lambda_N\}$ *be the unique solution of the traffic equation [8.24]. It is assumed that the stability condition*

$$\lambda_i < \mu_i C_i,\ i \in [\![1, N]\!], \qquad [8.29]$$

holds. Let π^i *be the probability measure on* $\mathbf{N}$ *defined by equations [8.27] and [8.28]. Then the process* $(X(t),\, t \geq 0)$ *describing the open Jackson network admits on* $\mathbf{N}^N$ *the only stationary probability* π^{OJ} *defined for any* $x = (x(1), \ldots, x(N)) \in \mathbf{N}^N$ *by*

$$\pi^{\text{OJ}}(x(1), \ldots, x(N)) = \prod_{i=1}^{N} \pi^i(x(i)). \qquad [8.30]$$

Proof. We aim to apply Lemma 7.19. Define, for any x and $y \in \mathbf{N}^N$,

$$\hat{A}(x, y) = \frac{\pi^{\text{OJ}}(y) A^{\text{OJ}}(y, x)}{\pi^{\text{OJ}}(x)},\ x \neq y,$$

where π^{OJ} is the probability measure defined by [8.30] and A^{OJ} is the infinitesimal generator defined in Lemma 8.8. So, for any i and j such that $i \neq j$ and any x such that $x(j) \geq 1$,

$$\begin{aligned}\hat{A}(x, x - \mathbf{e}_j + \mathbf{e}_i) &= \frac{\pi^{\text{OJ}}(x - \mathbf{e}_j + \mathbf{e}_i) A^{\text{OJ}}(x - \mathbf{e}_j + \mathbf{e}_i, x)}{\pi^{\text{OJ}}(x)} \\ &= \frac{\pi^i(x(j) - 1)\pi^i(x(i) + 1)}{\pi^j(x(j))\pi^i(x(i))} \mu_i(x(i) + 1) P_{ij} \\ &= \frac{\mu_j(x(j))}{\lambda_j} \frac{\lambda_i}{\mu_i(x(i) + 1)} \mu_i(x(i) + 1) P_{ij} \\ &= \frac{\lambda_i}{\lambda_j} \mu_j(x(j)) P_{ij};\end{aligned}$$

and similarly, for any i and j,

$$\begin{aligned}&\hat{A}(x, x + \mathbf{e}_i) = \lambda_i P_{i0}; \\ &\hat{A}(x, x - \mathbf{e}_j) = \frac{\mu_j(x(j))}{\lambda_j} \lambda P_{0j}, \quad \text{for any } x \text{ such that } x(j) \geq 1.\end{aligned}$$

Let us form the following sums for any $x \in \mathbf{N}^N$.

$$\begin{aligned}\sum_{y\neq x}\hat{A}(x,\,y) &= \sum_{j=1}^{N}\sum_{i=1;\,i\neq j}^{N}\hat{A}(x,\,x-\mathbf{e}_j+\mathbf{e}_i)+\sum_{i=1}^{N}\hat{A}(x,\,x+\mathbf{e}_i)+\sum_{j=1}^{N}\hat{A}(x,\,x-\mathbf{e}_j)\\ &= \sum_{j=1}^{N}\left(\frac{\mu_j(x(j))}{\lambda_j}\lambda P_{0j}+\lambda_i\sum_{i=1;\,i\neq j}^{N}\frac{\mu_j(x(j))}{\lambda_j}P_{ij}\right)+\sum_{i=1}^{N}\lambda_i P_{i0}\\ &= \sum_{j=1}^{N}\frac{\mu_j(x(j))}{\lambda_j}(\lambda_j-\lambda_j P_{jj})+\sum_{i=1}^{N}\lambda_i P_{i0}\\ &= \sum_{j=1}^{N}\mu_j(x(j))(1-P_{jj})+\lambda,\end{aligned}$$

where the second last equality is a consequence of the traffic equation, and the last one results from equation [8.26]. On the other hand,

$$\begin{aligned}&\sum_{y\neq x}A^{\text{OJ}}(x,\,y)\\ &= \sum_{i=1}^{N}\sum_{j=1;\,j\neq i}^{N}A^{\text{OJ}}(x,\,x-\mathbf{e}_j+\mathbf{e}_i)+\sum_{i=1}^{N}A^{\text{OJ}}(x,\,x-\mathbf{e}_i)+\sum_{j=1}^{N}A^{\text{OJ}}(x,\,x+\mathbf{e}_j)\\ &= \sum_{i=1}^{N}\mu_i(x(i))(1-P_{ii})+\sum_{j=1}^{N}\lambda P_{0j}\\ &= \sum_{i=1}^{N}\mu_i(x(i))(1-P_{ii})+\lambda\\ &= \sum_{y\neq x}\hat{A}(x,\,y).\end{aligned}$$

We conclude with Lemma 7.19. □

The latter, which is a classical result of queueing theory, is called “Theorem of the Product Form”: if the open Jackson network is stable, it behaves just like a system of N independent queues in equilibrium, where the i-th queue is a $\mathrm{M}_{\lambda_i}\,/\,\mathrm{M}_{\mu_i}\,/\,S_i$ queue. This fairly counterintuitive result has a clear interest in simulation: it indicates that the study of a Jackson network in steady state boils down to that of N multiple server queues. At equilibrium, everything happens just as if the N queues would function independently, and similarly to a classical Markovian queues (although each queue alone is *not* an M / M / queue).

Burke's Theorem applied to each of N queues M_{λ_i} / M_{μ_i} / S_i, $i = 1, \ldots, N$, would entail that the output rate of queue i equal λ_i, and thus that the input rate in queue j be given by

$$\lambda P_{0j} + \sum_{i=1}^{N} \lambda_i,$$

that is λ_j, according to the traffic equation. The result is hence consistent.

At any time t, conditionally to $X(t) = (x(1), x(2), \ldots, x(N))$, we can show by the usual techniques on the exponential distribution, that the residual time to the first output from queue i to the outside after t, follows the distribution $\varepsilon(\mu_i(x(i))P_{i0})$. Consequently, the residual time before the next output from the network taken as a whole to the outside, after t, follows the law $\varepsilon(\sum_{i=1}^{N} \mu_i(x(i))P_{i0})$. It is thus natural to define the instantaneous output rate at t by the random variable

$$D(t) = \sum_{i=1}^{N} \mu_i(X_i(t))P_{i0}.$$

The average output rate at equilibrium is hence given by

$$\begin{aligned}\mathbf{E}\left[D(\infty)\right] &= \mathbf{E}\left[\sum_{i=1}^{N} \mu_i(X_i(\infty))P_{i0}\right] \\ &= \sum_{(x(1),\, \ldots,\, x(N)) \in \mathbf{N}^N} \sum_{i=1}^{N} \mu_i(x(i))P_{i0}\pi^{\mathrm{RJ}}(x(1), \ldots, x(N)),\end{aligned}$$

where $X(\infty) = (X_1(\infty), \ldots, X_N(\infty))$ is a random variable distributed following π^{RJ} on $\mathbf{N}^N$. We then have the following analog of Burke's Theorem.

THEOREM 8.11.– *In an open Jackson network at equilibrium, the average output equals the intensity of the arrival process, i.e.*

$$\mathbf{E}\left[D(\infty)\right] = \lambda.$$

8.7.2. *Closed Jackson networks*

The closed Jackson network is similar to the open network, except that we assume now that the network is not "fed" by an exogenous Poisson process, in that no queue is linked to the outside. Here, K customers (where K is fixed) move forever from queue to queue in a network of N queues . / M_{μ_i} / S_i, which is connected as the previous one. The system is hence fully described by:

– K, the size of the population of the network;

– N queues ./M$_{\mu_i}$/S_i;

– a routing matrix P, Markovian and of size N and satisfying an irreducibility property similar to [8.23]: for any i and $j \in [\![1, N]\!]$, there exists $n \in \mathbf{N}$ and a n-tuple $\{i_1, i_2, \ldots, i_n\}$ of elements of $[\![1, N]\!]$ containing i and j, and such that

$$P_{i_1} P_{i_1 i_2} \ldots P_{i_{n_1} i_n} > 0. \qquad [8.31]$$

Consider the set

$$\mathcal{A} = \left\{ x = (x(1), \ldots, x(N)) \in \mathbf{N}^N;\ \sum_{i=1}^{N} x(i) = K \right\}.$$

The process $(X(t), t \geq 0)$ defined as in the previous section is Markov on $\mathcal{A}$, and of generator A^{CJ} defined for all $x \in \mathcal{A}$ such that $x(i) \geq 1$ and $x(j) \leq K$, by

$$A^{\text{CJ}}(x,\, x + \mathbf{e}_j - \mathbf{e}_i) = \mu_i(x(i)) P_{ij},$$

where all the other terms are zero except the diagonal one, which is the opposite of the sum of the other terms in the same line. The transitions for x such that $x(i) \geq 1$ for some i, or $x(j) = K$ for some j can be obtained similarly.

As above, the introduction of a Markov chain on $[\![1,\, N]\!]$ with transition matrix P allows to conclude that there exists, up to a multiplicative coefficient, a single solution $\Lambda \in (\mathbf{R}^+)^N$ to the matrix equation

$$\Lambda = \Lambda P. \qquad [8.32]$$

We then have the following result.

THEOREM 8.12.– *The process $(X(t), t \geq 0)$ describing the closed Jackson network admits a unique stationary probability π^{CJ}, defined for any $x \in \mathcal{A}$ by*

$$\pi^{\text{CJ}}(x(1), \ldots, x(N)) = C \prod_{i=1}^{N} \prod_{\ell=1}^{x(i)} \frac{\lambda_i}{\mu_i(\ell)},$$

where C is the normalization constant

$$C = \left(\sum_{(x(1), \ldots, x(N)) \in \mathcal{A}} \prod_{i=1}^{N} \prod_{\ell=1}^{x(i)} \frac{\lambda_i}{\mu_i(\ell)} \right)^{-1},$$

and $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ is an arbitrary solution to [8.32].

Proof. We follow the same argument as in the previous proof, i.e. we apply Lemma 7.19. Define $\hat{A}$ as above, i.e. for any $x, y \in \mathcal{A}$,

$$\hat{A}(x,\, y) = \frac{\pi^{\text{CJ}}(y)A^{\text{CJ}}(y,\, x)}{\pi^{\text{CJ}}(x)},\ x \neq y.$$

Then, we form the following sums for any $x \in \mathcal{A}$ such that $x(j) < K$ and $x(1) > 0$ (the other cases can be treated accordingly).

$$\begin{aligned}
\sum_{y\neq x} \hat{A}(x,\, y) &= \sum_{j=1}^{N} \sum_{i=1;\, i\neq j}^{N} \hat{A}(x,\, x - \mathbf{e}_j + \mathbf{e}_i) \\
&= \sum_{j=1}^{N} \sum_{i=1;\, i\neq j}^{N} \frac{\pi^{\text{CJ}}(x - \mathbf{e}_j + \mathbf{e}_i)A^{\text{CJ}}(x - \mathbf{e}_j + \mathbf{e}_i,\, x)}{\pi^{\text{CJ}}(x)} \\
&= \sum_{j=1}^{N} \sum_{i=1;\, i\neq j}^{N} \frac{\lambda_j}{\lambda_i} P_{ji}\mu_i\left(x(i)\right) \\
&= \sum_{i=1}^{N} \frac{1}{\lambda_i}\mu_i\left(x(i)\right) \sum_{j=1;\, j\neq i}^{N} \lambda_j P_{ji} \\
&= \sum_{i=1}^{N} \frac{1}{\lambda_i}\mu_i\left(x(i)\right)\left(\lambda_i - \lambda_i P_{ii}\right) \\
&= \sum_{i=1}^{N} \mu_i\left(x(i)\right)\left(1 - P_{ii}\right),
\end{aligned}$$

where the second last equality is a consequence of the traffic equation [8.32]. On the other hand,

$$\begin{aligned}
\sum_{y\neq x} A^{\text{CJ}}(x,\, y) &= \sum_{i=1}^{N} \sum_{j=1;\, j\neq i}^{N} A^{\text{CJ}}(x,\, x - \mathbf{e}_j + \mathbf{e}_i) \\
&= \sum_{i=1}^{N} \mu_i(x(i))(1 - P_{ii}) \\
&= \sum_{y\neq x} \hat{A}(x,\, y).
\end{aligned}$$

Hence, the result. □

A closed Jackson network is much less comfortable to study than in the open case. One can in fact observe that, unlike the open case, the previous form is not a product form because of the expression of the constant C. In addition, the numerical calculation of this constant for a large network is computationally very expensive.

8.8. Problems

EXERCISE 16.– We consider the M/M/1 queue with the following service discipline: a proportion p of the customers (the "polite" ones) is placed normally in line, while a proportion $q = 1 - p$ of "rude" ones double everyone, and take the first place in the queue. The class ("polite" or "rude") of a given customer is independent of its arrival time. This discipline is non-preemptive, that is to say that one does not interrupt the current service. The intensity of the overall arrival process is λ and the average service time is $1\,/\,\mu$.

We denote $(X(t),\, t \geq 0)$, the number of customers in the system (queue + server) at time t.

1) Give the infinitesimal generator of X. Is it different from that of the M / M / 1 queue with the FIFO discipline (First In First Out)?

2) Deduce the stability condition of the system and the steady-state distribution of X.

3) What is the average waiting time in steady state?

4) What is the nature and intensity of the arrival process of the "rude" customers?

5) Denote W^p, the waiting time of a given polite customer, X as the number of customers in the system upon his arrival and $N(W^p)$, the number of "rude" customers who arrive during his waiting time. Show that we have in law,

$$W^p \overset{\mathcal{L}}{=} \sum_{j=1}^{X} \eta_j + \sum_{l=1}^{N(W^p)} \sigma_l, \qquad [8.33]$$

where $(\eta_j,\, j \geq 0)$ and $(\sigma_l,\, l \geq 0)$ are two sequences independent of each other and of X, of independent r.v., exponentially distributed with parameter μ.

6) Explain why X and $N(W^p)$ are independent conditionally to W^p.

7) We assume now that X has the stationary distribution identified in (2). Prove that

$$\mathbf{E}\left[\sum_{j=1}^{X} \eta_j\right] = \frac{\rho}{(1-\rho)\mu}.$$

8) Show that

$$\mathbf{E}\Big[e^{-s\sum_{l=1}^{N(W^p)} \sigma_l} \,|\, W^p = v\Big] = e^{-\lambda q v \frac{s}{\mu+s}}.$$

9) Show that if X follows the stationary distribution identified in (2), we have

$$\mathbf{E}\left[e^{-\lambda q W^p \frac{s}{\mu+s}} \mathbf{E}\left[e^{-s\sum_{j=1}^{X} \eta_j} \middle| W^p\right]\right].$$

(the explicit form of $\mathbf{E}\left[e^{-sW^p}\right]$ is not requested explicitly).

10) Deduce, by differentiation, the average waiting time of a "polite" customer.

11) We denote W^i, the waiting time of a "rude" customer in steady state and W, the waiting time of an any customer in steady state - we take for granted that the waiting time of the nth rude customer converges in distribution to W^i, and similarly for W. Explain why it holds true that

$$\mathbf{E}[W] = p\mathbf{E}[W^p] + q\mathbf{E}\left[W^i\right].$$

12) Deduce the average waiting time of a "rude" customer.

EXERCISE 17.– Consider a gas station with 2 pumps and a waiting slot:

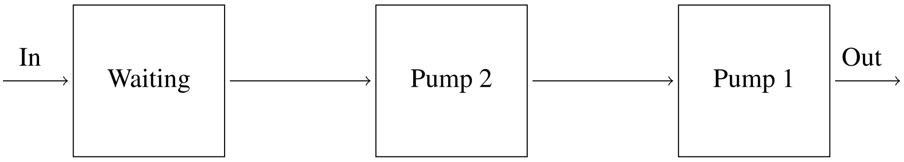

Figure 8.2. *Gas bar with three pumps in series*

Customers arrive at the station according to a Poisson process of intensity λ. If both pumps are free, the customer goes to pump 1. If the pump 1 is taken, the customer goes into pump 2. If both pumps are busy the customer moves into the waiting place. If the waiting room is busy, the customer pass his way. A customer in pump 2 must wait until the pump 1 is free to exit the gas station. In the case, where the customer at pump 2 finishes before the one at pump 1, the two comes out together from the gas station when the customer at pump 1 has finished. If there is, at that time, a customer waiting, he will instantly go into pump 1 and begins filling his container. In the case where the pump 1 is free but the pump 2 is busy, no one can enter into pump 1. The time to fill the containers follow an exponential distribution with parameter μ.

We model this system by a Markov process $X = (X_1, X_2, X_3)$ where X_1 is 0 or 1 and represents the number of customers at pump 1, X_2 represents the number of customers at pump 2 and X_3 is the number of customers in the waiting place (X_3 is thus 0 or 1).

1) What is the state space of X?

2) Write the components of its infinitesimal generator.

3) Is there a stationary probability?

4) If yes, what are the equations that characterize it?

5) What is the percentage of customers who cannot enter the station?

6) What is the percentage of customers who cannot enter the station because it is badly done, i.e. customers who cannot be serviced even though pump 1 is free.

EXERCISE 18.– In a store, customers, on an average of 20 per hour, arrive at the cash desk according to a Poisson process. As long as there are less than two customers in line, there is only one cash desk open. The service times at this cash desk are exponentially distributed and the average service time is 5 minutes.

As soon as three customers are in line, a second cash desk opens. The two cash desks share the same queue. The second cash desk closes when there is not more than two customers waiting. We note X_t as the number of customers in the system at time t.

1) Write the infinitesimal generator of X.

2) Find the stationary probability if there is any.

3) What is the average number of customers in the system?

4) What is the average number of customers in line?

5) What is the average waiting time?

6) What is the percentage of time when the second desk is open?

EXERCISE 19.– A banking agency has $S = 5$ employees. The average number of calls is 20 per hour, the average duration of a call is 6 minutes. Arrivals form a Poisson process of intensity λ and service times are independent and exponentially distributed. In questions 1–5, it is not requested to demonstrate the formulas, just to give them and to perform the numerical computations. Note, that *en route* several partial results are used several times.

1) If there is no possibility of waiting, what is the percentage of those customers whose call fails due to the lack of a free employee?

2) The agency is now equipped with a device to put on hold, that is a record that makes the customers wait until a consultant is free. We assume that all the customers wait as long as necessary. What is the probability that a given customer has to wait?

3) Is it equal to the probability of blocking in the system without waiting? Why?

4) What is the average waiting time?

5) What is the probability that the waiting time exceed 2 minutes?

Now, we remove the stand by system and consider that a percentage $100.p$ of calls require rerouting to a more specialized treatment center. At the level of the PABX (and of the N allocated links), this results in an occupation of two connections: one for the incoming calls, the other one for the rerouted calls. We denote X^1 as the number of direct calls, using a single connection and X^2, the number of calls that need two links. The number N of links is to be determined.

X^1 is limited by S the number of advisers, whereas X^2 is only limited by the condition $X^1+2X^2 \leq N$. We consider, in fact, that the dimensioning of the specialized processing center is such that the calls that are intended for it actually reach it, and we neglect the time needed for the local advisor to decide about the rerouting. We assume that whenever all advisors are busy, the incoming calls are directed to the specialized processing center.

6) Write down the infinitesimal generator of $X = (X^1, X^2)$.

7) Does the process X admit a stationary probability?

8) Write down the equations allowing us to determine it.

9) At fixed N, what is the size of the matrix to invert?

EXERCISE 20.– Derive [8.21] using Little's formula.

8.9. Notes and comments

We have only given a glimpse of the study of Markovian queues with infinite buffer. This subject has been a topic of wide interest for many researchers, and there is a huge amount of literature on this issue.

In particular, we have not addressed the very interesting topic of elastic traffic and processor-sharing queues. This rich framework is very well studied in [BON 11], and provides a formula for the waiting time, which is insensitive to the distribution of the service times.

The study of product form networks is based mainly on Kelly's Lemma, which can be found in [KEL 79]. The local balance equation that it induces is often too restrictive to be satisfied. Many weaker versions exist, that are still less restrictive than the global equations solving $\pi A = 0$. Many examples can be found in [CHA 99].

Epitome

– The stationary probability of the M / M / 1 queue of arrival intensity λ and average service time $1 \,/\, \mu$ is given by

$$\pi(0) = 1 - \rho,\ \pi(n) = \rho^n(1 - \rho),\ \text{where } \rho = \lambda \,/\, \mu.$$

– The waiting time in the steady state equals 0 with probability $1 - \rho$. Conditionally on being strictly positive, the waiting time follows an exponential distribution with parameter $\mu - \lambda$. Its mean expectation is thus $1 \,/\, (\mu - \lambda)$.

– In the M / M / S queue, the stability condition is $\rho < S$. The stationary probability is given by [8.10].

– In an open Jackson network of N queues, the stationary probability of the system is the product of the stationary probabilities of N M / M / 1 queues, taking as input parameters, the components of the solution of the traffic equation [8.24].

Chapter 9

Loss Systems

9.1. General

A loss system is a system that has fewer resources (servers and possibly, buffer) than potential users, and where all the customers arriving at a time when the system resources are taken, are lost. Therefore, we need to find out the adequate number of resources in order to loose as few requests as possible. In the following, we always denote by

– $N(t)$, the number of customers who tried to enter the system up to t (those will be termed *arrived* customers);

– $Y(t)$, the number of customers who have actually entered the system (called *entered* customers) before t. Therefore,

$$X^d(t) = N(t) - Y(t)$$

is the number of customers lost up to t.

– $X(t)$, the number of customers present in the system (server + waiting line) at time t.

– S, the number of servers and C, the number of places in the waiting line.

EXAMPLE 9.1.– The second generation mobile phone network, known as GSM, is technically based on TDMA (Time Division Multiple Access). For a given frequency, we divide time into periods of equal and constant duration, known as *slots*. The slots are gathered in packets of eight to form a *frame*. The voice call is digitized in a way that routing a call amounts in fact to carrying information bits. To route a call, the octets of a call are grouped by packets, and we attribute to a call a slot set during the whole call duration. On Figure 9.1, the fourth slot is assigned to the incoming call.

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

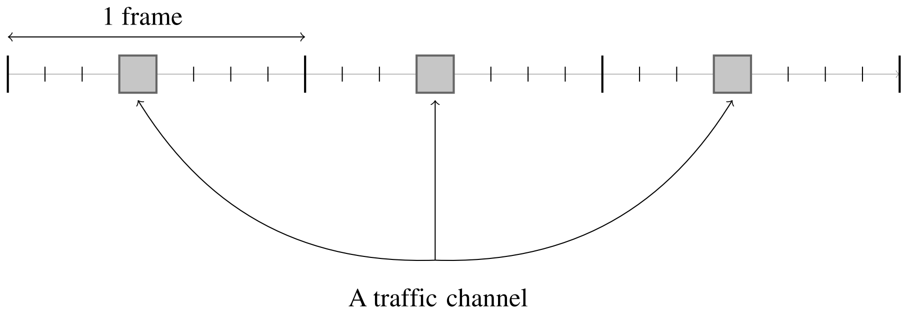

Figure 9.1. *Principle of TDMA*

A transceiver device can route only 8 calls simultaneously. Therefore, what matters to the operator is to determine the number of devices for each base station, that is the antennas that decorate the roofs of our buildings. In fact, to account for the signaling, a single device permits to route 7 simultaneous calls, 14 for a double device and 21 for a triple one, and so on.

EXAMPLE 9.2.– The multiplexer is one of the key elements of data networks. It permits us to route the calls arriving from N input ports to N output ports. Since a single output port can be requested simultaneously by multiple flux, it is important to install a buffer to temporarily store the data until the channel is free. The problem here is to determine the buffer size, so that the loss probability of incoming bits shall be below a certain threshold. This threshold depends on the nature of the data flow. Roughly speaking, voice and video flow can allow the loss of about one per thousand, while the data flow cannot allow any loss. In practice, we consider that something that happens with a probability of less than one per billion does not actually happen.

These two examples are similar but the second raises two major problems: (i) it is not obvious to choose an input traffic model which can reflect the differences (in the throughput, for instance, just to name one of them) existing between the flow of voice, video and speech. In addition, the very small order of the targeted loss probability renders any simulation tedious, and then requires an accurate analytical model.

This also requires to distinguish between three quantities of interest:

DEFINITION 9.1.– *The* congestion *or* blocking probability *of a system with $S + C$ resources is the asymptotic proportion of time when all resources are busy, that is*

$$P_B = \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{\{X(t)=S+C\}}\,\mathrm{d}\,t.$$

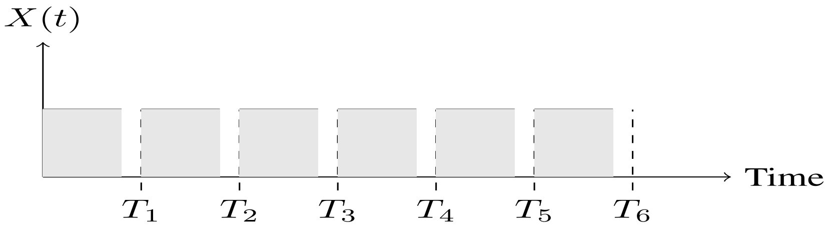

Figure 9.2. *Occupation of the $D/D/1/1$ queue*

The loss probability *is the blocking probability from the customers perspective, i.e.*

$$P_l = \lim_{t\to\infty} \frac{1}{N(t)} \sum_{j=1}^{N(t)} \mathbf{1}_{\{X(T_j^-)=S+C\}} \,. \tag{9.1}$$

The overflow probability *exists in fact only when the number of resources is infinite, and is defined by*

$$P_S = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{X(t)>S\}} \, \mathrm{d}t.$$

In general, these three probabilities are distinct. Indeed, let us consider a system with a single server, with deterministic arrivals and service times, of respective duration $\rho < 1$ and 1.

By its very definitions, the blocking probability equals the percentage of shaded area, that is ρ. But the loss probability is zero, as no customer finds a busy server upon arrival. However, if the arrivals were Poisson, the loss and blocking probabilities would coincide in view of the PASTA property.

THEOREM 9.1.– *If the arrivals form a Poisson process, then $P_B = P_l$.*

Proof. It suffices to apply Theorem A.38 for

$$\psi(s) = \mathbf{1}_{\{X(s)=S+C\}} \,.$$

□

The loss probability is by far, the most important indicator, but also the most difficult to assess. The blocking and the overflow probabilities are much more easily calculable, and are therefore often used as a substitute for the loss probability.

9.2. Erlang model

We assume that the arrivals occur according to a Poisson process of intensity λ, that the service times are independent and identically distributed of distribution $\varepsilon(\mu)$, and that the capacity of the waiting line is zero, that is $C = 0$, so that any customer is attended if and only if some server is free upon his arrival, and lost otherwise. We study in other words the M$_\lambda$/M$_\mu$/S/S queue, and denote as usual the traffic load $\rho = \lambda/\,\mu$.

The system is described by the process $(X^{\text{E}}(t),\, t \geq 0)$ counting the number of customers in the system. On its state-space $[\![0,\, S]\!]$, $(X^{\text{E}}(t),\, t \geq 0)$ has the same transitions as the process $\big(X^{(S)}(t),\, t \geq 0\big)$ describing the M$_\lambda$/M$_\mu$/S/∞ queue (see section 8.3), except for the state S, that the process leaves only to jump into $S - 1$, as no customer is then accepted. The infinitesimal generator of $(X^{\text{E}}(t),\, t \geq 0)$ is thus given by

$$A^{\text{E}} = \begin{pmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & & \\ & & \ddots & & & (0) & \\ & & k\mu & -(\lambda+k\mu) & \lambda & & \\ & (0) & & & \ddots & & \\ & & & & (S-1)\mu & -((S-1)\mu+\lambda) & \lambda \\ & & & & & S\mu & -S\mu \end{pmatrix}.$$

The process is ergodic in that it is valued in a finite state space, and it is easy to compute its stationary probability π^{E}, which satisfies

$$\pi^{\text{E}}(i) = \frac{\lambda}{i\mu}\pi^{\text{E}}(i-1),\ i \in [\![1,\, S]\!];$$

$$\sum_{i=0}^{S} \pi^{\text{E}}(i) = 1,$$

which is equivalent to

$$\pi^{\text{E}}(i) = \frac{\rho^i/i!}{\sum_{k=0}^{S} \rho^k/k!} \text{ for } i \in [\![0,\, S]\!]. \qquad [9.2]$$

Moreover, as $(X^{\text{E}}(t),\, t \geq 0)$ is a birth and death process, it is reversible.

NOTE.– We can also derive the invariant probability by noticing that the $\mathrm{M}_\lambda/\mathrm{M}_\mu/S/S$ queue is nothing but a $\mathrm{M}_\lambda/\mathrm{M}_\mu/\infty$ queue that is constrained not to exceed S customers. In other words, $(X^{\mathrm{E}}(t),\, t \geq 0)$ is the truncated version, forced to stay in $[\![0,\, S]\!]$, of the reversible process $(X^\infty(t),\, t \geq 0)$ representing the number of customers in the $\mathrm{M}_\lambda/\mathrm{M}_\mu/\infty$ system. We thus find π^{E} with Kelly's Lemma (Theorem 7.22): for any $i \in [\![0,\, S]\!]$,

$$\pi^{\mathrm{E}}(i) = \frac{1}{\sum_{i=0}^{S} \pi^\infty(k)} \pi^\infty(i)$$
$$= \frac{\rho^i/i!}{\sum_{k=0}^{S} \rho^k/k!}.$$

According to [9.1], the loss probability for the Erlang model reads

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{S\}} \left(X\left(T_j^-\right)\right) = \lim_{t\to\infty} \frac{1}{t} \int_0^T \mathbf{1}_{\{S\}}(X(s^-))\, ds$$
$$= \pi^{\mathrm{E}}(S).$$

We have therefore established the following result.

THEOREM 9.2.– *The loss probability of the $M_\lambda/M_\mu/S/S$ queue of load ρ reads*

$$Er[\rho,\, S] = \frac{\rho^S/S!}{\sum_{i=0}^{S} \rho^i/i!}.$$

NOTE.– It can be proven, that the loss probability does not depend on the distribution of service times: this expression remains valid for a M/GI/S/S queue.

For the numerical assessment of this probability, we can use the following approximation

$$\mathrm{Er}[\rho, S] \approx \exp S \log \frac{\rho u}{S} + S - \rho u \sqrt{\frac{u + \rho u(1-u)^2}{S}},$$

with u given by

$$u = \frac{(S+\rho+1) - \sqrt{(S+\rho+1)^2 - 4\rho S}}{2\rho}.$$

We can also use the recurrence relation

$$\frac{1}{\mathrm{Er}[\rho, S]} = 1 + \frac{S}{\rho \mathrm{Er}[\rho, S-1]}.$$

In practice, after estimating ρ, we have to find the smallest S such that Er$[\rho,\ S]$ is below the desired threshold, to ensure a given quality of service. It is generally assumed that this threshold is of 0.001 for the telephone network and 0.02 for the GSM network. The algorithm is as follows.

Algorithm 9.1. Calculating the number of servers to ensure a given loss probability

```
**Data**: $\rho,\ \epsilon$
**Result**: $S$ such that Er$[\rho, S] \leq \epsilon$
$S \leftarrow 0$;
$x \leftarrow 1$;
**until** $x < \epsilon^{-1}$ **do**
    $S \leftarrow S + 1$;
    $x \leftarrow 1 + \frac{S}{\rho}x$;
**end**
**return** $S$
```

For a loss of one per thousand, the number of required servers required is (very) approximately equal to $\rho + 3\sqrt{\rho}$. If the arrivals and service times were deterministic, ρ would represent the number of simultaneous calls, therefore also the number of needed servers. The term $3\sqrt{\rho}$ is then interpreted as a guarantee against the random fluctuations of traffic.

9.3. The M/M/$1/1 + C$ queue

We now consider a queue with one server working in FCFS, and whose waiting room has a limited capacity C. As usual, the arrival process is Poisson of intensity λ, and the customers request service times that are independent and identically distributed of distribution $\varepsilon(\mu)$. We denote $\rho = \lambda/\mu$ the traffic load.

The process $(X^{\mathrm{c}}(t),\ t \geq 0)$ counts the number of customers in the system and takes values in $[\![0,\ 1 + C]\!]$. It satisfies the following dynamics:

– for any $i \in [\![0,\ C]\!]$, the sojourn time in i and the probabilities of jumps starting from i are clearly the same as for the M$_\lambda$/M$_\mu$/1/∞ queue;

– once in $1 + C$, the system is full and can no longer accept customers. The process can leave this state only for the state C, after a sojourn time equal to the residual service time of the customer in service, i.e. of distribution $\varepsilon(\mu)$.

It is hence easy to see that $(X^{\mathrm{c}}(t),\ t \geq 0)$ is Markov, and admits in $[\![0,\ 1 + C]\!]$ the infinitesimal generator

$$A^{\mathrm{C}} = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & (0) & \\ & & \ddots & & \\ & (0) & \mu & -(\lambda+\mu) & \lambda \\ & & & \mu & -\mu \end{pmatrix}.$$

As the process $(X^{\mathrm{C}}(t),\, t \geq 0)$ takes values in a finite set, it is ergodic and therefore has a unique stationary probability π^{C}. The process is reversible, as $(X^{\mathrm{C}}(t),\, t \geq 0)$ is a birth and death process. The probability π^{C} can be derived by solving

$$\begin{cases} \pi^{\mathrm{C}} A^{\mathrm{C}} & = \mathbf{0}; \\ \pi^{\mathrm{C}} \mathbf{e} & = 1. \end{cases}$$

We obtain that for any $i \in [\![0,\, 1+C]\!]$,

$$\begin{aligned} \pi^{\mathrm{C}}(i) &= \rho^i \pi^{\mathrm{C}}(0) \\ &= \begin{cases} \dfrac{\rho^i - \rho^{i+1}}{1-\rho^{C+2}} & \text{if } \rho \neq 1; \\ \dfrac{1}{C+2} & \text{if } \rho = 1, \end{cases} \end{aligned} \qquad [9.3]$$

observing by the way, that π^{C} is the uniform probability in $[\![0,\, 1+C]\!]$ in the critical case $\rho = 1$.

NOTE.– In the sub-critical case $\rho < 1$, π^{C} can be obtained as well from Kelly's Lemma (Theorem 7.22), remarking that $(X^{\mathrm{C}}(t),\, t \geq 0)$ is the truncation at $[\![0,\, 1+C]\!]$ of the reversible process X counting the number of customers in the M$_\lambda$/M$_\mu$/1/∞.

THEOREM 9.3.– *In a M$_\lambda$/M$_\mu$/1/1+C queue, the loss probability is given by the formula*

$$P_l = \begin{cases} \frac{\rho^{C+1} - \rho^{C+2}}{1-\rho^{C+2}} & \textit{if } \rho \neq 1; \\ \frac{1}{C+2} & \textit{if } \rho = 1. \end{cases}$$

Proof. The loss probability for this system is given by the asymptotic rate of lost customers, i.e. of customers finding a full system upon arrival. According to the PASTA property and the ergodicity of $(X^{\mathrm{C}}(t),\, t \geq 0)$, it reads

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{\{1+C\}}(X^{\mathrm{C}}(T_n-)) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{1+C\}}(X^{\mathrm{C}}(t))\, \mathrm{d}\, t \\ &= \mathbf{E}\left[\mathbf{1}_{\{1+C\}}\left(X^{\mathrm{C}}_\infty\right)\right] \\ &= \pi^{\mathrm{C}}(1+C). \end{aligned}$$

The result then follows from [9.3]. □

If the traffic load ρ is strictly less than 1, the loss probability can be compared to the overflow probability of the corresponding M$_\lambda$/M$_\mu$/1/∞ queue. The probability that the latter system has a congestion greater than or equal to $1 + C$ in steady state can be written (with the notations of section 8.2) as

$$\mathbf{P}\left(X_\infty \geq 1 + C\right) = \sum_{i=1+C}^{\infty} \pi(i) = (1-\rho) \sum_{i=1+C}^{\infty} \rho^i = \rho^{1+C}, \qquad [9.4]$$

whereas

$$P_l = \pi^{\mathrm{c}}(0)\rho^{1+C} < \rho^{1+C}. \qquad [9.5]$$

Therefore, it is possible to estimate the loss of a finite buffer system by the overflow probability of the corresponding infinite buffer queue. In doing so, we overestimate the loss according to equations [9.4] and [9.5]. However, this may be the only computation that is feasible in practice, since the infinite buffer queue is generally easier to describe using known probability distributions. Moreover, a dimensioning that is pessimistic as long as it is not exaggerated, gives a stronger guarantee against the fluctuations in traffic and the uncertainties in the estimation of the traffic load.

Buffer dimensioning

It is interesting to know what the optimal buffer size is that we should display to guarantee the user a quality of service in terms of packets loss probability. We can already check the intuitively clear result, that the loss probability is a decreasing function of the number of servers. In fact, the function defined for any $\rho \in \mathbf{R} + \backslash\{1\}$ given by

$$f_\rho : \left\{ \begin{array}{ll} \mathbf{R} & \rightarrow \mathbf{R} \\ x & \mapsto \frac{\rho^{x+1} - \rho^{x+2}}{1 - \rho^{x+2}} \end{array} \right.$$

admits as derivative the function defined for any x by

$$f'_\rho(x) = \frac{(\ln \rho)\rho^{x+1}}{(1-\rho^{x+2})^2}(1-\rho) > 0,$$

and the same results holds true, of course, for $\rho = 1$.

The optimal dimensioning of the buffer is hence given by the following algorithm (in the case $\rho \neq 1$).

Algorithm 9.2. Derivating the minimal size of the buffer that guarantees a loss probability $P_l \leq \varepsilon$

```
Data: $\rho$, $\varepsilon$
Result: $K$ such as $P_l \leq \varepsilon$
$K \leftarrow 1$;
until
    $$\varepsilon < \frac{\rho^{K+1} - \rho^{K+2}}{1 - \rho^{K+2}}$$
do
    |
end
return $K \leftarrow K + 1$
```

NOTE.– Similar computations and results can be obtained for several servers, i.e. for a M/M/S/S+C queue (see Exercise 23). We have chosen the single server case to simplify the formulas.

9.4. The "trunk" effect

Let us apply our results to the GSM network. In the language of the protocol, a slot is called a TCH (traffic channel). We usually consider that the traffic load per cell-phone is of 0.025 Erlang and that the admissible loss probability is about 0.02. Based on the above, we obtain the following results:

Number of transceivers	1	2	3	4	5	6	7
Number of TCH	7	14	22	29	37	45	52
Capacity	2.9	8.2	15	21	28	35.5	42
Traffic carried by TCH	0.41	0.59	0.68	0.72	0.76	0.79	0.81
Number of Cell-phones	116	328	596	840	1128	1424	1680

This table calls at least for two comments. First, it is important to have in mind the great variability of the results. A reasonable increase in the number of resources induces a large increase in capacity. For example, if one goes from 1 to 2 transmitters and receivers, the capacity is multiplied by almost 3. To dispose more traffic, we do not need to proportionally scale up the number of TCH: to dispose 42 Erlang instead of 21, we need 52 instead of $2{*}29 = 58$.

The other point is known as "trunk" effect: the greater the number of servers, the more important the charge passed by each server. In economics, this is the same principle as the "economies of scale." Therefore, when the loss threshold is fixed, we will prefer a small number of systems having a large number of servers to a large number of systems with a small number of servers.

9.5. Engset model

In order to use the Erlang model, it is necessary to implicitly assume that the number of potential sources is very large, if not infinite, as is the case, for example, in Telephone networking. However, when the number of sources is small compared to the number of servers, and in the case where a source in service cannot issue another request until the end of service of the previous one for the same source, we can no longer consider that the intensity of arrivals of the customers is independent of the state of the system.

In the model known as Engset, we assume that we have M independent sources, each generating requests according to a Poisson process of intensity λ. The system has S servers (where $S \leq M$), the distributions of service times are always assumed as exponential with parameter μ, and the waiting room is still of size 0. Unlike the Erlang model, when k sources are "in service", only $M - k$ sources are likely to make a request. Therefore the instantaneous rate of arrivals is $(M - k)\lambda$. In fact, X is a Markov process of infinitesimal generator A given by

$$A = \begin{pmatrix}
-\lambda_0 & \lambda_0 & & & & & \\
\mu_1 & -(\mu_1 + \lambda_1) & \lambda_1 & & & & \\
 & & \ddots & & & (0) & \\
 & & \mu_k & -(\lambda_k + \mu_k) & \lambda_k & & \\
 & (0) & & & \ddots & & \\
 & & & & \mu_{S-1} & -(\mu_{S-1} + \lambda_{S-1}) & \lambda_{S-1} \\
 & & & & & \mu_S & -\mu_S
\end{pmatrix},$$

where we have set $\lambda_i = (M - i)\lambda$ and $\mu_i = i\mu$ for all i. The invariant distribution of this process is defined, satisfies as usual the system

$$\nu A = 0 \Longleftrightarrow \left\{ \begin{array}{lcl} -\lambda_0 \nu_0 + \mu \nu_1 & = & 0; \\ \lambda_{i-1}\nu_{i-1} - (\lambda_i + \mu_i)\nu_i + \mu_{i+1}\nu_{i+1} & = & 0, \; i \in [\![1, S-1]\!]. \end{array} \right.$$

Using the normalization constraint $\pi.\mathbf{e} = 1$, this leads to

$$\nu_i = \frac{\rho^i C_M^i}{\sum_{j=0}^S \rho^j C_M^j} \; i \in [\![0, S]\!],$$

setting as usual $\rho = \lambda / \mu$.

THEOREM 9.4.– *In the Engset model, the loss probability is given by*

$$\mathit{Eng}[\rho,\, S,\, M] = \frac{\rho^S C_{M-1}^S}{\sum_{j=0}^S C_{M-1}^j \rho^j}.$$

Proof. From the very construction of the model, the arrival process is a Poisson process whose intensity varies over time as a function of $X(s)$: it is given by

$$\Lambda(t) = \int_0^t (M - X(s))\lambda \, \mathrm{d}\, s.$$

As jumps of $N(t)$ are of height 1, the quadratic variation of the martingale $N - \Lambda$ is also Λ. Consider the bounded adapted process $\psi(s) = \mathbf{1}_{\{X(s)=S\}}$. According to Theorem A.37,

$$\begin{aligned} P_l = \lim_{t\to\infty} \frac{1}{N(t)} \int_0^t \psi(s_-) \, \mathrm{d}\, N(s) &= \lim_{t\to\infty} \frac{1}{\Lambda(t)} \int_0^t \psi(s) \, \mathrm{d}\, \Lambda(s) \\ &= \lim_{t\to\infty} \frac{t}{\Lambda(t)} \, \frac{1}{t} \int_0^t \psi(s)(M - X(s))\lambda \, \mathrm{d}\, s. \end{aligned}$$

As X is ergodic, we have the following almost sure limits

$$\frac{1}{t}\Lambda(t) \xrightarrow{t\to\infty} \left(M - \sum_{j=0}^{S} j\nu(j) \right) \lambda;$$

$$\frac{1}{t} \int_0^t \psi(s)(M - X(s))\lambda \, \mathrm{d}\, s \xrightarrow{t\to\infty} \lambda(M - S)\nu(S).$$

Therefore,

$$\mathrm{Eng}[\rho, \, S, \, M] = \frac{\lambda(M - S)\nu(S)}{\lambda \sum_{j=0}^{S} (M - j)\nu(j)} = \frac{\rho^S C_{M-1}^S}{\sum_{j=0}^{S} \rho^j C_{M-1}^j}.$$

In other words, the loss probability of a system of M machines equals the blocking probability of a system of $M - 1$ machines, the rest remaining unchanged. □

9.6. IPP/M/S/S queue

The above results can make believe that the loss probability depends only on the load. This is not the case in general, as we can see in the following example.

An IPP (Interrupted Poisson Process) is a special case of MMPP process (see 7.6), where the phase process J has two states A and B. The infinitesimal generator of J is of the form

$$Q_J = \begin{pmatrix} -\sigma_A & \sigma_A \\ \sigma_B & -\sigma_B \end{pmatrix},$$

where $1/\sigma_A$ and $1/\sigma_B$ are the average sojourn times in the phases A and B, respectively. Its invariant probability, denoted by ν, is easily calculated. We obtain

$$\nu(A) = \frac{\sigma_B}{\sigma_A + \sigma_B} \text{ and } \nu(B) = \frac{\sigma_A}{\sigma_A + \sigma_B}.$$

The IPP/M/S/S queue is thus the modified version of the Erlang model with S servers, without buffer, and where the arrival process is the IPP described above. The process X counting the number of busy servers (and hence, of customers in the system) is not Markov alone, but the process $(X,\ J)$ is. We number the states in lexicographic order and we denote Λ as the matrix of intensities of arrivals, i.e.

$$\Lambda = \begin{pmatrix} \lambda_A & 0 \\ 0 & \lambda_B \end{pmatrix}.$$

The average intensity of the arrival process is hence given by

$$\lambda = \lambda_A \frac{\sigma_B}{\sigma_A + \sigma_B} + \lambda_B \frac{\sigma_A}{\sigma_A + \sigma_B}\cdot \qquad [9.6]$$

The infinitesimal generator of $(X,\ J)$ thus reads

$$A = \begin{pmatrix} Q_J - \Lambda & \Lambda & & & \\ \mu \,\mathrm{Id} & (Q_J - \Lambda - \mu \,\mathrm{Id}) & \Lambda & & \\ & 2\mu \,\mathrm{Id} & (Q_J - \Lambda - 2\mu \,\mathrm{Id}) & \Lambda & \\ & & \ddots & \ddots \ddots & \\ & & & S\mu \,\mathrm{Id} & (Q_J - S\mu \,\mathrm{Id}) \end{pmatrix}.$$

The Markov process is of finite state space, of course irreducible, therefore it admits as invariant probability π the solution of the usual system $\pi A = 0$ and $\pi.\mathbf{e} = 1$. To simplify the calculations, we introduce the two-component row vectors

$$x_n = \big(\pi(n,\ A),\ \pi(n,\ B)\big),\ n = 0, \cdots, S.$$

The equations corresponding to $\pi A = 0$ thus become S couples of equations

$$x_0(Q_J - \Lambda) + \mu x_1 = 0;$$

$$x_0\Lambda + x_1(Q_J - \Lambda - \mu \,\mathrm{Id}) + 2\mu x_2 = 0;$$

$$\vdots$$

$$x_{S-2}\Lambda + x_{S-1}(Q_J - \lambda - (S-1)\mu \,\mathrm{Id} + S\mu x_S = 0;$$

$$x_{S-1}\Lambda + x_S(Q_J - S\mu \,\mathrm{Id}) = 0.$$

From the first $(S-1)$ equations, we obtain

$$\begin{aligned}
x_1 &= \frac{1}{\mu} x_0 (Q_J - \Lambda);\\
x_2 &= -\frac{1}{2\mu}(x_0 \Lambda + x_1 (Q_J - \Lambda - \mu \,\mathrm{Id}));\\
&\vdots\\
x_S &= -\frac{1}{S\mu}(x_{S-2}\Lambda + x_{S-1}(Q_J - \lambda - (S-1)\mu \,\mathrm{Id}));\\
x_S &= x_{S-1}\Lambda (Q_J - S\mu \,\mathrm{Id})^{-1}.
\end{aligned}$$

Let us set

$$\begin{aligned}
R_0 &= \mathrm{Id},\\
R_1 &= \frac{1}{\mu}(Q_J - \Lambda),\\
R_n &= -\frac{1}{n\mu}\Big(R_{n-2}\Lambda + R_{n-1}(Q_J - \lambda - (n-1)\mu \,\mathrm{Id})\Big), n = 2, \cdots, S.
\end{aligned} \qquad [9.7]$$

We can then write

$$x_n = x_{n-1} R_n,\ n \geq 1$$

and

$$x_0 (R_S - R_{S-1} \Lambda (Q_J - S\mu \,\mathrm{Id})^{-1}) = 0.$$

We thus obtain two equations for the two components of x_0. In fact, only one suffices since if the system was of rank 2, the only solution would be 0, which is excluded. We deduce form this the following resolution algorithm.

Algorithm 9.3. Computation of the invariant probability of the IPP/M/S/S queue.

```
Data: $\Lambda$, $\mu$, $S$, $Q_J$
Result: $pi$ such that $\pi A = 0$ and $\pi.\mathbf{e} = 1$
Compute $R_1, \cdots, R_S$ from [9.7];
$x_0(0, A) \leftarrow 1$;
Find $x_0(0, B)$ such that $x_0$ satisfies $x_0(R_S - R_{S-1}\Lambda(Q_J - S\mu \,\mathrm{Id})^{-1}) = 0$;
Compute $x_n = x_{n-1}R_n$ for $n = 1, \cdots, S$;
Compute $m = \sum_{n=0}^{S} x_n.\mathbf{e}$;
$\pi \longrightarrow \pi$;
return $\pi$
```

THEOREM 9.5.– *In an IPP/M/S/S queue with invariant probability π, the loss probability is given by*

$$\frac{\lambda_A}{\lambda}\pi(S,\, A) + \frac{\lambda_B}{\lambda}\pi(S,\, B), \qquad [9.8]$$

where λ is the average intensity given by [9.6].

Proof. We apply Theorem 7.24 to $\psi(s) = \mathbf{1}_{\{S\}}(X(s))$. It follows that the loss probability is given by

$$\frac{1}{\lambda}\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{S\}}(X(s))\lambda(J(s))\,\mathrm{d}\,s = \frac{1}{\lambda}(\lambda_A\pi(S,\, A) + \lambda_B\pi(S,\, B)).$$

Hence the result. □

Setting the traffic load is equivalent to setting λ, but according to the relative values of λ_A, λ_B, σ_A, and σ_B we get very different loss probabilities, as shown in Table 9.1. We have chosen to set $S = 10$ servers and a traffic load of 5 Erlang. Erlang-B formula would give a loss of 0.018.

Parameters	loss	blocking
$\begin{pmatrix} \sigma_A = 1/2 & \sigma_B = 1/2 \\ \lambda_A = 10 & \lambda_B = 0 \end{pmatrix}$	0.1	0.05
$\begin{pmatrix} \sigma_A = 9/10 & \sigma_B = 1/10 \\ \lambda_A = 50 & \lambda_B = 0 \end{pmatrix}$	0.66	0.07
$\begin{pmatrix} \sigma_A = 99/100 & \sigma_B = 1/100 \\ \lambda_A = 500 & \lambda_B = 0 \end{pmatrix}$	0.96	0.01
$\begin{pmatrix} \sigma_A = 9/10 & \sigma_B = 1/10 \\ \lambda_A = 1 & \lambda_B = 5.4 \end{pmatrix}$	0.02	0.05

Table 9.1. *Loss probability at constant load in an IPP/M/S/S queue*

It appears from reading this table that we can increase the loss probability by keeping a constant load. We might think that the determining factor then becomes the variance of the arrival process, since as we increase the latter, the loss increases. This is certainly a good criterion but it is unfortunately not the only one. Setting the variance and the traffic load amounts to impose two equations satisfied by the four parameters. This leaves two degrees of freedom which can be used to make the loss probability vary in any direction.

Let us also observe that the blocking probability is very far from the loss probability, and hence there is a huge mistake in assimilating these two quantities.

9.7. Generalized Erlang models

9.7.1. *Guard channels*

The mechanism of guard channels is well known by the operators, and allows to define a priority among multiple streams without increasing in a dramatic manner the loss probability of the stream having the lowest priority.

We illustrate this concept by using the management of handover in the GSM network. When a user moves while giving a call from his cell phone, there may come a time when the BTS that managed the call loses the radio connection with this mobile. It is, therefore, necessary to skip the connection to another BTS. This phenomenon, called handover, requires heavy operations on the control plan, and the operator must ensure that there are sufficient resources in the new cell, to handle this new call.

When taking into account the mobility of users, it is necessary to introduce the sojourn duration in the cell of each user. It is preferable, in order to be able to perform the calculations (and quite reasonable statistically), to assume that the crossing time of a cell for a given user follows a an exponential distribution with parameter α. The call duration of a mobile seen from the BTS then follows the minimum between the call duration, and its sojourn time on the cell. Given the properties of the exponential distribution, this call duration hence follows an exponential distribution of parameter $\mu + \alpha$. The probability that a user "leaves" the cell before the end of his call equals the probability that a random variable exponentially distributed with parameter μ is less than a random variable exponentially distributed with parameter α parameter. It is hence given by

$$\theta = \frac{\mu}{\mu + \alpha}.$$

Let us consider a set of cells with identical characteristics, and track the inputs and outputs in a given cell. In Figure 9.3, λ_f is the rate of new calls (also termed "fresh calls"), λ_{Ho} represents the average number of hand-over calls coming into the cell under study, and p is the loss probability in the cell.

As the system is designed such that p is negligible when compared to 1, we have at equilibrium

$$\lambda_{\text{Ho}} \simeq (\lambda_f + \lambda_{\text{Ho}})\theta, \text{ that is } \lambda_{\text{Ho}} \simeq \frac{\theta}{1-\theta}\lambda_f.$$

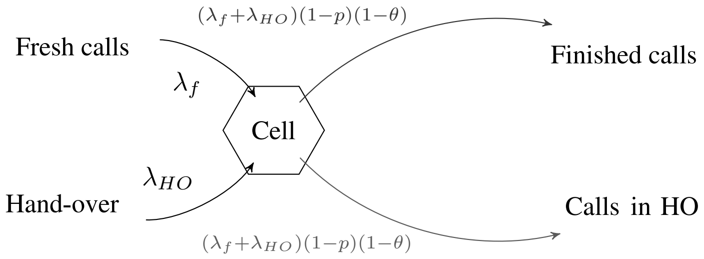
Fresh calls
λ_f
Hand-over
λ_{HO}
Cell
$(\lambda_f+\lambda_{HO})(1-p)(1-\theta)$
Finished calls
Calls in HO
$(\lambda_f+\lambda_{HO})(1-p)(1-\theta)$

Figure 9.3. *Input–output of a cell*

The total traffic load to be handled by the BTS is thus approximately given by

$$\begin{aligned}\rho &= (\lambda_f + \lambda_{\text{Ho}}) \times \frac{1}{\mu+\alpha}\\ &= \lambda_f(1+\frac{\theta}{1-\theta})\frac{1}{\mu+\alpha}\\ &= \lambda_f\frac{1}{1-\theta}\frac{1}{\mu+\alpha}\\ &= \lambda_f\frac{\mu+\alpha}{\mu}\frac{1}{\mu+\alpha}\\ &= \frac{\lambda_f}{\mu}.\end{aligned}$$

However, we cannot apply the Erlang formula for dimensioning the system with hand-overs since the handover calls have a higher requirement in terms of quality of service. It is indeed much more unpleasant to have interrupted the communication, than not being able to initiate one.

To take this difference into account, let us fix two different target loss probabilities: ϵ_{F} for the fresh calls, and ϵ_{Ho} for the hand-over calls. According to the remark above, we consider that

$$\epsilon_{\text{Ho}} < \epsilon_{\text{F}}.$$

The problem arising is to dimension the system in a way that the loss of fresh calls (respectively, of hand-over calls) does not exceed the target loss probabilities, i.e.

$$P_l(\text{Ho}) < \epsilon_{\text{Ho}};\ P_l(\text{F}) < \epsilon_{\text{F}}. \qquad [9.9]$$

First approach: over-dimensioning

A first approach consists of not distinguishing between both types of calls, and dimensioning the system in order to achieve a global loss less than ϵ_{Ho}.

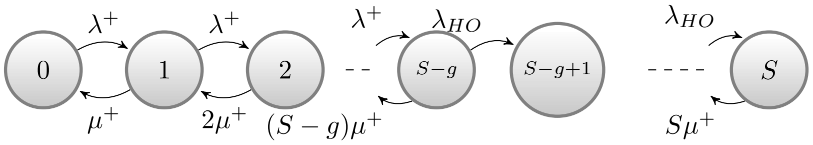

Figure 9.4. *The transition of the process "number of busy servers" with guard channels*

The system clearly is a M/M/S/S queue, with S as the number of channels, $\lambda_f + \lambda_{\text{HO}}$ as intensity of the arrival process and $1/(\alpha + \mu)$ as average call duration.

Set for instance $\epsilon_{\text{HO}} = 10^{-4}$; $\epsilon_{\text{F}} = 10^{-2}$, $\lambda_f = 10$ calls/min, $\lambda_{\text{HO}} = 4$ calls/min, $1/\mu = 2$ min and $1/\alpha$ =5 min. Consequently,

$$\rho = \frac{\lambda_f + \lambda_2 0}{\alpha + \mu} = 20 \text{ Erlang.}$$

Then, Erlang-B formula yields that 39 channels are necessary to achieve the target loss probability. This meets the requirement [9.9], but over-dimension the system, since the fresh calls have a weaker loss constraint.

Second approach: parallel systems

A second idea would be to split the resources into two pools working independently: a first pool of channels is reserved for the fresh calls, and a second one, to the hand-over calls.

So both parallel systems are again Erlang models, where the arrival intensities are given by λ_f and λ_{HO}, respectively and the average service times still equal $1/\left(\alpha + \mu\right)$. By applying Erlang-B formula to both systems, we obtain that 23 channels are necessary for the first system, and 17 for the second one in order to meet condition [9.9]. So 40 channels in total, which is worth than the first approach!

Third approach: guard channels

To give priority to the handover calls, we decide to modify the access control: we choose $g < S$, where the number of free servers is greater than g, we accept new calls and handover calls. As soon as it remains less than g available channels, we no longer accept the fresh calls to give priority to the handover calls. To derive the loss probability, we represent the system by the process X counting the number of busy servers at all times. This is a Markov process whose transitions can be represented by the diagram 9.4 where $\lambda^+ = \lambda_f + \lambda_{\text{HO}}$ and $\mu^+ = \mu + \alpha$.

From there, we deduce the invariant probability ν, then the loss probabilities for the various types of call using the PASTA property. As the arrival process of fresh calls is always Poisson of intensity λ_f, Theorem A.38 implies that the loss probability for fresh calls is given by

$$P_l(f) = \sum_{j=S-g}^{S} \nu(j)$$

and that of hand-over calls, by

$$P_l(\text{Ho}) = \nu(S).$$

We obtain the numerical results of Table 9.2, for the same numerical values as in the first two approaches. It is quite remarkable that for small values of g, we give to handover calls a loss probability which is much less than that of the new calls, without excessively penalizing the latter. We only need 34 channels in total using this access control, that is 5 channels less than with the first approach!

S	30	32	34	34
g	0	2	4	3
$P_l(f)$	0.8457%	1.0275%	1.0332%	0.6584%
$P_l(\text{Ho})$	0.8457%	0.0278%	0.0008%	0.0028%

Table 9.2. *Loss of new calls and handover calls according to the number of guard channels.*

9.7.2. *Multi-class system*

In many situations, customers do not request the same amount of resources. In this case, we use the multi-class Erlang formula. Consider a system with K classes of customers and S resources. Class i customers arrive according to a Poisson process of intensity λ_i and their communication lasts an exponential time of parameter μ_i. We set $\rho_i = \lambda_i/\mu_i$. A class i customer consumes s_i resources. The numbers of customers n_i of each class, $i \in [\![1, K]\!]$, are subject to the constraint

$$\sum_{i=1}^{K} n_i s_i \leq S.$$

We study the process X, counting the number of busy resources. Its state space is

$$\mathcal{A} = \left\{ (n_1, \cdots, n_K) \in \mathbf{N}^K, \sum_{j=1}^{k} n_j s_j \leq S \right\}.$$

THEOREM 9.6.– *The invariant probability of this system is given by*

$$\nu(n_1, \cdots, n_K) = \frac{1}{G} \prod_{j=1}^{k} \rho^{n_j}/n_j!$$

for any $n = (n_1, \cdots, n_K) \in \mathcal{A}$*, where* G *is the normalization constant*

$$G = \sum_{(n_1, \cdots, n_K) \in \mathcal{A}} \prod_{j=1}^{k} \rho^{n_j}/n_j!.$$

Proof. If $S = \infty$, the components of X are independent Markov processes evolving, respectively, as the number of busy servers in a $M_{\lambda_i}/M_{\mu_i}/\infty$ queue. So, these are reversible processes (see Definition 7.11) with invariant distribution ν_i, a Poisson distribution with parameter ρ_i (see section 8.5). The process X is hence reversible, with an invariant probability that is the tensor product of these probabilities.

For finite S, X is just the restriction to $\mathcal{A}$ of the previous dynamics. Thus Kelly's Lemma (Theorem 7.22) applied to $E = \mathbf{N}^K$, $F = \mathcal{A}$ and $\alpha = 0$, yields the result. □

Denote as usual, $\mathbf{e}_i$ the ith vector of the canonical basis of $\mathbf{R}^K$. A call of class i is lost whenever $X \in \mathcal{A}$, but $X + \mathbf{e}_i \notin \mathcal{A}$. Therefore, according to the PASTA property, the loss probability of class i reads

$$P_l(i) = \sum_{n \in \mathcal{A}, n + \rceil_i \notin \mathcal{A}} \nu(n_1, \cdots, n_K).$$

EXAMPLE 9.3 (INTERFACE A-BIS).– The recent advances in voice coding imply that, in situations where one TCH per frame was necessary at the time to handle a communication, only half a frame is now necessary. However, some calls always need a full slot. As the latter result applies for integer numbers of resources, it is necessary to count the number of busy half-slots here. So we split the customers in two classes, one with $s_1 = 1$ and the other one with $s_2 = 2$. Let us take a cell with 30 TCH, that is 60 half-slots. The following table shows the loss probability of each class according to the traffic loads

9.8. Hierarchical networks

We have already seen that the greater the number of servers, the lower the loss probability. A good measure of this gain can be the cost of transported Erlang: let us consider a system with S servers and without waiting room, and for which we set an

		Traffic load of calls of class 1			
		10	20	30	40
Traffic load of calls of class 2	10	0.004%	0.24	2.47%	8.34%
		0.1%	0.55%	5.29%	16.71%
	20	2.62%	7.48%	13.9%	20.62%
		5.53%	14.99%	26.54%	37.66%

Table 9.3. *Loss rates of the various classes based on the traffic load*

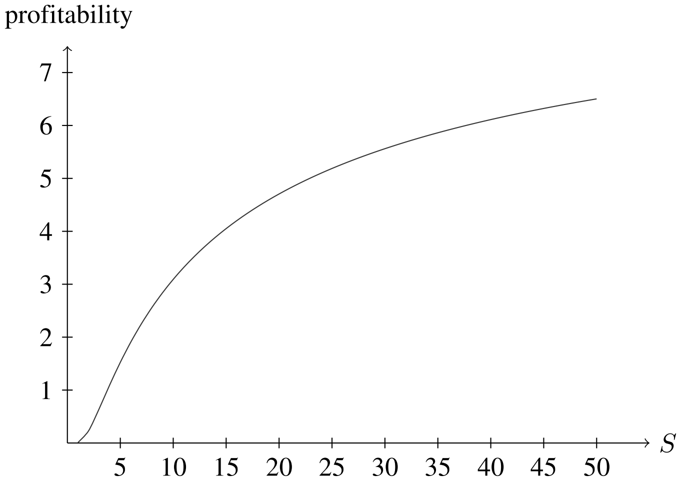

Figure 9.5. *The profitability in function of S*

upper bound α for the loss probability. The maximum traffic load $\rho_{S,\alpha}$ that can pass through this system is defined by the equation

$$\text{Er}[\rho_{S,\alpha}, S] = \alpha.$$

The system profitability is then defined as the ratio of $\rho_{S,\alpha}$ by S, i.e.

$$R = \frac{\rho_{S,\alpha}}{S}.$$

As shown in Figure 9.5, the profitability increases with the number of servers.

However, if we take into account the cost of installing a server, there comes a time when the increase in costs compensates an increase in the profitability gain. The solution implemented in the conventional telephone network consists of organizing the network hierarchically. Subscribers are connected to a switch termed switch of level 0, and then level 0 switches are connected to a level 1 switch, in lesser numbers. We can continue up to level 3 in the hierarchy. For the record, of its golden age, the switched French Telephone network use to include approximately 1500 switches of level 0, some hundreds of switches of levels 1 and 2, and 7 switches of level 3.

The establishment of communication was done by always trying to use junctions of the lowest level possible. If there was a free junction between two switches of level 0, the latter was used. Otherwise, the switch of level 0 was sending the management of the call to the switch of level 1, which itself was trying to route the call by staying at its level, and so on.

If the RTC itself is now obsolete, the fact remains that many of its inventions have been incorporated into modern systems, especially the guard channels and hierarchical networks.

Let us take the case of the GSM network. In urban areas, the cells, i.e. the areas managed by an antenna or base station, are small and the hand-overs become frequent for users as "fast" as motorists, for instance. Hence, we add several more powerful antennas, which cover a wider area and will offer two types of services: (i) support the calls of the fast mobiles, so as to reduce the frequency of hand-overs and (ii) route the overflow calls of the original smaller cells (termed *micro-cells*) that they cover.

One question arises: dimensioning the number of transceivers in the cell of the highest level (termed *macro-cell*). Indeed, the overflow process of a micro-cell is not a Poisson process, but an MMPP (see Example 7.2). For the micro-cell of index i, the phase process of the overflow process has as infinitesimal generator Q_i, that of the number of customers in a $M_{\lambda_i}/M_{\mu_i}/S_i/S_i$ queue, and as rate function λ_i given by

$$\lambda_i(j) = 0 \text{ for } j < S_i \text{ and } \lambda_i(S_i) = \lambda.$$

In view of Theorem 7.23, the phase process of the whole MMPP (consisting of the superposition of the overflow processes of all micro-cells), denoted J, has a generator $Q = Q_1 \oplus \ldots \oplus Q_K$ and a rate function $\lambda = \lambda_1 \otimes \ldots \otimes \lambda_K$.

Therefore, the dimensioning of the macro-cell is equivalent to studying an MMPP/M/S/S queue. Let us denote X as the number of busy servers. This process is not Markov alone, because without knowing the phase, we cannot know how long will it take for the next arrival to occur. However, in this case the couple process $(X,\ J)$ is Markov. Its infinitesimal generator A is written in blocks, as follows

$$\begin{pmatrix}
Q-\Lambda & \Lambda & & & & & \\
\mu I_m & Q-\Lambda-\mu I_m & \Lambda & & & & \\
 & 2\mu I_m & Q-\Lambda-2\mu I_m & \Lambda & & (0) & \\
 & & \cdots & \cdots & \cdots & & \\
 & & S\mu I_m & Q-\Lambda-S\mu I_m & \Lambda & & \\
 & (0) & & S\mu I_m & Q-\Lambda-S\mu I_m & \Lambda & \\
 & & & & \cdots & & \\
 & & & & & S\mu I_m & Q-S\mu I_m
\end{pmatrix}.$$

The invariant probability π satisfies as usual

$$\pi A = 0, \pi \mathbf{e} = 1,$$

where $\mathbf{e}$ is the vector with $(S+1).m$ components all equal to 1. There, m denotes the number of phases of the overflow process, i.e. $m = (S+1)^K$. Set $x_i = (\pi(i,\, 1), \cdots, \pi(i,\, m))$ for $i \in [\![0, S]\!]$. The x_i's are thus solutions to the system

$$\begin{cases} x_0(Q_{mc} - \Lambda_{mc}) + \theta x_1 = 0; \\ x_0\Lambda_{mc} + x_1(Q_{mc} - \Lambda_{mc} - \theta I_m) + 2\theta x_2 = 0; \\ \vdots \\ x_{n-1}\Lambda_{mc} + x_n(Q_{mc} - \Lambda_{mc} - n\theta I_m) + (n+1)\theta x_{n+1} = 0,\ n \geq 1; \\ \vdots \\ \end{cases}$$

The last equation is

$$x_{L-1}\Lambda_{mc} + x_L(Q_{mc} - L\theta n I_m) = 0. \qquad [9.10]$$

From these equations, we obtain

$$\begin{cases} \quad x_n = x_0 R_n \text{ with } R_{-1} = 0,\ R_0 = I_m, \\ R_{n+1} = -\dfrac{1}{\theta(n+1)}(R_{n-1}\Lambda_{mc} + R_n(Q_{mc} - \Lambda_{mc} - n\theta I_m)), \end{cases} \qquad [9.11]$$

which determines the x_i's according to x_0. It remains to determine x_0, and we have two possibilities.

Method 1 The stationary probability of the number of busy servers in each cell is given by the Erlang formula

$$\nu(i) = \frac{\rho^i/i!}{\sum_{j=0}^{S} \rho^j/j!}.$$

Since the cells are independent of each other, the stationary probability of the phase process is the tensor product of K vectors

$$\nu_{mc} = \nu \otimes \nu \otimes \ldots \otimes \nu.$$

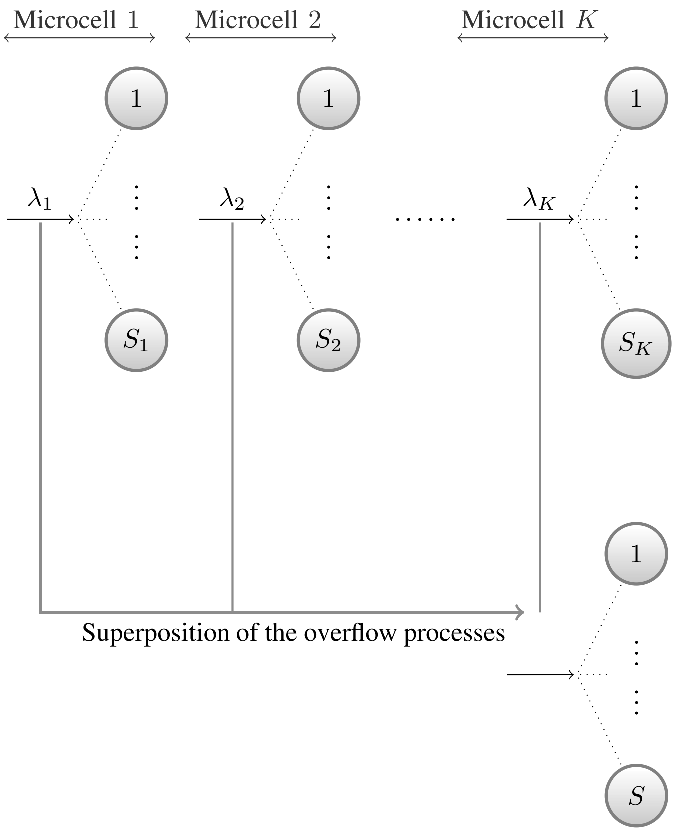

Figure 9.6. *The traffic carried by the macro-cell is the sum of the overflow traffic of micro-cells*

Observe that for all $j \in [\![0, S]\!]^K$,

$$\sum_{n=0}^{S} \pi(n, j) = \nu_{mc}(j).$$

Then, we introduce f_n as the row vector with m components of which only the $n$þis not 0 but 1. We also introduce e_n as the transpose of f_n. The last relation then reads

$$\sum_{n=0}^{L} x_0 R_n . e_j = \nu_{mc}(j).$$

As $e_j.f_j$ is the matrix whose only non-zero coefficient is the coefficient (j, j), we have

$$\sum_{j=1}^{m} e_j.f_j = Id_m.$$

On the other hand,

$$\nu_{mc} = \sum_{j=1}^{m} \nu_{mc}(j) f_j.$$

We thus have

$$x_0 \sum_{l=0}^{L} R_l = \nu_{mc}. \qquad [9.12]$$

Method 2 By putting [9.11] in [9.10], we obtain

$$x_0(R_{L-1}\Lambda_{mc} + R_L(Q_{mc} - L\theta)) = 0.$$

This equation determines all the components of x_0 but one. Then, it is necessary to use the normalization condition to calculate its value.

Finally, from Theorem 7.25, the loss probability is given by

$$P_l = (\nu_{mc}\Lambda_{mc}e)^{-1} x_L \Lambda_{mc} e, \qquad [9.13]$$

where $e = \sum_{n=1}^{m} e_n$.

9.9. A model with balking

In this system, we model a forced limitation of the system workload, by introducing an access control, rather that limiting the system capacity: the more the system is congested, the less chance the customers have of entering into it.

With the usual notation, we consider a M$_\lambda$/M$_\mu$/1/∞ queue, in which the arriving customers make a toss (independent from one customer to another, and from all other parameters) to determine whether they enter the system or not. We denote $(X^{\text{R}}(t), t \geq 0)$ as the process counting the number of customers in the system (the exponent $^{\text{R}}$ will be added to all parameters). For the nth arriving customer, the draw is a Bernoulli experience of probability $p(X^{\text{R}}(T_n^-))$, hence depending on the congestion at the arrival of the customer. If the Bernoulli variable equals 1, then the customer enters the system and will wait until the end of its service. Otherwise, the customer does not even enter the system, and is lost forever. It is thus pure good sense to assume that the function $p(.)$ is decreasing.

NOTE.– The expression of $p(.)$ that is mainly considered in the literature is

$$p(n) = \frac{1}{n+1}. \qquad [9.14]$$

Hereafter, we will call “entering customer”, an arriving customer who actually enters the system. We aim to describe the steady state of the system by studying the process $(X^{\mathrm{R}}(t), t \geq 0)$. Let us begin by noticing the following analog of the bus paradox (Theorem 6.6).

LEMMA 9.7.– *Let for any t, $\tilde{W}(t)$ be the residual time at t before the next arrival of an entering customer . Then, for any $i \in \mathbf{N}$, $\tilde{W}(t)$ follows the distribution $\varepsilon(\lambda p_i)$ conditionally to $X^{\mathrm{R}}(t) = i$, that is for any $x \geq 0$,*

$$\mathbf{P}\left(\tilde{W}(t) \leq x \,|\, X^{\mathrm{R}}(t) = i\right) = 1 - e^{-\lambda p_i x}.$$

Proof. Let us again denote $T_0 = 0$ and $T_1, T_2, \ldots$ the arrival times of the customers, for any $t \geq 0$, $N(t)$ be the number of customers arrived up to t and $Z(t)$, the number of arrivals necessary to see the first actual entering customer after t. It is then easily checked that

$$\tilde{W}(t) = T_{N(t)+Z(t)-x}.$$

On the other hand, $Z(t)$ follows conditionally to $\{X^{\mathrm{R}}(t) = i\}$, a geometric distribution with parameter p_i, as after t, each customer enters with probability p_i independently of the others, up to the first actual entry after t. We can thus write for any x,

$$\begin{aligned}
&\mathbf{P}\left(\tilde{W}(t) \geq x \,|\, X^{\mathrm{R}}(t) = i\right) \\
&\quad = \sum_{k\geq 1} \mathbf{P}\left(\tilde{W}(t) \geq x \,|\, Z(t) = k;\ X^{\mathrm{R}}(t) = i\right) \mathbf{P}\left(Z(t) = k \,|\, X^{\mathrm{R}}(t) = i\right) \\
&\quad = \sum_{k\geq 1} \mathbf{P}\left(\tilde{W}(t) \geq x \,|\, Z(t) = k;\ X^{\mathrm{R}}(t) = i\right) (1-p_i)^{k-1} p_i \\
&\quad = p_i \sum_{k\geq 1} (1-p_i)^{k-1} \mathbf{P}\left(T_{N(t)+k} - t \geq x\right) \\
&\quad = p_i \sum_{k\geq 1} (1-p_i)^{k-1} \sum_{j=1}^{\infty} \mathbf{P}\left(T_{j+k} - t \geq x;\ N(t) = j\right).
\end{aligned} \qquad [9.15]$$

First, according to Theorem 6.6,

$$p_i \sum_{j=1}^{\infty} \mathbf{P}\left(T_{j+1} - t \geq x;\ N(t) = j\right) = p_i e^{-\lambda x}. \qquad [9.16]$$

On the other hand,

$$\begin{aligned}
& p_i \sum_{k \geq 2} (1-p_i)^{k-1} \mathbf{P}\left(T_k - t \geq x;\ N(t) = 0\right) \\
& = p_i \sum_{k \geq 2} (1-p_i)^{k-1} \mathbf{P}\left(T_k \geq t + x;\ \xi_1 > t\right) \\
& = p_i \sum_{k \geq 2} (1-p_i)^{k-1} \{\mathbf{P}\left(\xi_1 > t + x\right) + \mathbf{P}\left(T_k \geq t + x;\ t < \xi_1 < t + x\right)\} \\
& = p_i \sum_{k \geq 2} (1-p_i)^{k-1} e^{-\lambda(t+x)} \\
& \quad + p_i \sum_{k \geq 2} (1-p_i)^{k-1} \int_t^{t+x} \lambda e^{-\lambda u} \int_{t+x-u}^{\infty} \lambda^{k-1} \frac{v^{k-2}}{(k-2)!} e^{-\lambda v} \,\mathrm{d}\, v \,\mathrm{d}\, u \\
& = e^{-\lambda(t+x)}(1-p_i) + \int_t^{t+x} \lambda e^{-\lambda u} (1-p_i) e^{-\lambda p_i (t+x-u)} \,\mathrm{d}\, u \\
& \qquad\qquad = -p_i e^{-\lambda(t+x)} + e^{-\lambda t} e^{-\lambda p_i x}. \qquad [9.17]
\end{aligned}$$

Then,

$$\begin{aligned}
& p_i \sum_{k \geq 2} (1-p_i)^{k-1} \sum_{j=1}^{\infty} \mathbf{P}\left(T_{j+k} - t \geq x;\ N(t) = j\right) \\
& = p_i \sum_{k \geq 2} (1-p_i)^{k-1} \sum_{j=1}^{\infty} \mathbf{P}\left(T_j \leq t; T_j + \xi_{j+1} > t + x\right) \\
& \quad + p_i \sum_{k \geq 2} (1-p_i)^{k-1} \sum_{j=1}^{\infty} \mathbf{P}\left(T_{j+k} \geq t + x;\ T_j \leq t;\ t < T_j + \xi_{j+1} \leq t + x\right).
\end{aligned} \qquad [9.18]$$

But as in the proof of Theorem 6.6,

$$\begin{aligned} p_i\sum_{k\geq 2}(1-p_i)^{k-1}\sum_{j=1}^{\infty}\mathbf{P}\left(T_j\leq t; T_j+\xi_{j+1}>t+x\right) & \\ &= p_i\sum_{k\geq 2}(1-p_i)^{k-1}e^{-\lambda(t+x)}(e^{\lambda t}-1) \\ &= (e^{-\lambda x}-e^{-\lambda(t+x)})(1-p_i), \end{aligned} \qquad [9.19]$$

while

$$\begin{aligned} & p_i\sum_{k\geq 2}(1-p_i)^{k-1}\sum_{j=1}^{\infty}\mathbf{P}\left(T_{j+k}\geq t+x;\ T_j\leq t;\ t<T_j+\xi_{j+1}\leq t+x\right) \\ &= p_i\int_0^t\sum_{j=1}^{\infty}(\lambda^j\frac{u^{j-1}}{(j-1)!})e^{-\lambda u}\int_{t-u}^{t+x-u}\lambda e^{-\lambda v} \\ &\qquad \left(\int_{t+x-u-v}^{\infty}\sum_{k=2}^{\infty}((1-p_i)^{k-1}\lambda^{k-1}\frac{w^{k-2}}{(k-2)!})e^{-\lambda w}\,dw\right)\mathrm{d}\,v\,\mathrm{d}\,u. \\ &\qquad = \int_0^t\lambda e^{-\lambda p_i(t+x-u)}\int_{t-u}^{t+x-u}\lambda(1-p_i)e^{-\lambda(1-p_i)v}\,\mathrm{d}\,v\,\mathrm{d}\,u \\ &\qquad = (e^{-\lambda t}e^{-\lambda p_i x}-e^{-\lambda(t+x)})\int_0^t\lambda e^{\lambda u}\,\mathrm{d}\,u \\ &\qquad = e^{-\lambda p_i x}-e^{-\lambda x}-e^{-\lambda t}e^{-\lambda p_i x}+e^{-\lambda(t+x)}. \end{aligned} \qquad [9.20]$$

By collecting [9.16], [9.17], [9.18], [9.19], and [9.20] in [9.15], we finally obtain that

$$\begin{aligned} & \mathbf{P}\left(\tilde{W}(t)\geq x\,|\,X^{\mathrm{R}}(t)=i\right) \\ &\quad = p_i e^{-\lambda x}-p_i e^{-\lambda(t+x)}+e^{-\lambda t}e^{-\lambda p_i x}+(e^{-\lambda x}-e^{-\lambda(t+x)})(1-p_i) \\ &\qquad + e^{-\lambda p_i x}-e^{-\lambda x}-e^{-\lambda t}e^{-\lambda p_i x}+e^{-\lambda(t+x)} \\ &\quad = e^{-\lambda p_i x}. \end{aligned}$$

□

As we did before, we deduce from Lemma 9.7 that the process $(X^{\mathrm{R}}(t), t \geq 0)$ is Markov of generator A^{R} that is given by

$$\begin{pmatrix}
-\lambda p_0 & \lambda p_0 & & & & & \\
\mu & -(\lambda p_1 + \mu) & \lambda p_1 & & & & \\
0 & \mu & -(\lambda p_2 + \mu) & \lambda p_2 & & (0) & \\
 & & \ddots & \ddots & \ddots & & \\
 & (0) & & \mu & -(\lambda p_i + \mu) & \lambda p_i & \\
 & & & & \mu & -(\lambda p_{i+1} + \mu) & \lambda p_{i+1} \\
 & & & & \ddots & \ddots & \ddots
\end{pmatrix}.$$

We can then verify that whenever the stability condition

$$\sum_{i=1}^{\infty} \rho^i \prod_{j=0}^{i-1} p_j < +\infty \qquad [9.21]$$

holds, the unique stationary probability π^{R} of the congestion process is defined by

$$\pi^{\mathrm{R}}(0) = \left(1 + \sum_{i=1}^{\infty} \rho^i \prod_{j=0}^{i-1} p_j\right)^{-1};$$

$$\pi^{\mathrm{R}}(i) = \rho^i \left(\prod_{j=0}^{i-1} p_i\right) \pi(0),\ i \geq 1. \qquad [9.22]$$

NOTE.– In the classical case where

$$p_i = \frac{1}{i+1},\ i \geq 0,$$

we have $\pi(i) = e^{-\rho}\rho^i/i!$, for any $i \geq 0$. This means that the stationary congestion X_∞ follows a Poisson distribution $\mathcal{P}(\rho)$, just as the infinite server queue.

Loss probability

We need to enrich our model in order to derive the loss probability of the system. So we define the process $(Y(t), t \geq 0)$ by induction on the arrival times as follows: we first set

$$Y(t) = 0;\ t \in [0, T_1],$$

then for any $i \geq 1$,

$$Y(T_i) = \begin{cases} 1 & \text{if the customer } C_i \text{ enters the system;} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$Y(t) = Y(T_i),\ t \in [T_i,\ T_{i+1}].$$

It is then easily seen that the couple process $((X^{\mathrm{R}}(t), Y(t)),\ t \geq 0)$ is Markov on $\mathbf{N} \times \{0,\ 1\}$. Indeed, it is a process with rcll paths (as $(Y(t), t \geq 0)$ is piecewise constant), of which we can write the generator $\tilde{A}^{\mathrm{R}}$ as follows:

– for any $i \geq 0$, the process may leave the state $(i,\ 0)$ for the state $(i+1,\ 1)$ if an arrival of an entering customer actually occurs, and if $i \geq 1$, for state $(i-1,\ 0)$ if the current service ends. So,

$$\tilde{A}^{\mathrm{R}}((i,\ 0), (i+1,\ 1)) = \lambda p_i;$$

$$\tilde{A}^{\mathrm{R}}((i,\ 0), (i-1,\ 0)) = \mu \text{ for } i \geq 1.$$

– The process has the same transitions from $(i,\ 1)$ to $(i+1,\ 1)$ and $(i-1,\ 1)$. Another possible jump is done from $(i,\ 1)$ to $(i,\ 0)$, i.e. when a customer arrives and does not actually enter. Thus,

$$\tilde{A}^{\mathrm{R}}((i,\ 1), (i+1,\ 1)) = \lambda p_i;$$

$$\tilde{A}^{\mathrm{R}}((i,\ 1), (i,\ 0)) = \lambda(1 - p_i);$$

$$\tilde{A}^{\mathrm{R}}((i,\ 1), (i-1,\ 1)) = \mu \text{ for } i \geq 1,$$

where we apply a result similar to Lemma 9.7 for the second transition.

We solve the system

$$\begin{cases} \tilde{\pi}^{\mathrm{R}}\tilde{A}^{\mathrm{R}} & = \mathbf{0}; \\ \tilde{\pi}^{\mathrm{R}}\mathbf{e} & = 1, \end{cases}$$

where $\tilde{A}^{\mathrm{R}}$ is the generator of the process on $\mathbf{N} \times \{0,\ 1\}$ and $\tilde{\pi}^{\mathrm{R}}$ is a probability measure on $\mathbf{N} \times \{0,\ 1\}$.

Let us denote X_∞^{R} and Y_∞^{R} as the limiting r.v.'s for the two processes, if any. So, if it exists, $\tilde{\pi}^{\mathrm{R}}$ should satisfy for any $i \in \mathbf{N}$ to

$$\begin{aligned} \pi^{\mathrm{R}}(i) = \mathbf{P}\left(X_\infty^{\mathrm{R}} = i\right) &= \mathbf{P}\left(X_\infty^{\mathrm{R}} = i;\ Y_\infty^{\mathrm{R}} = 0\right) + \mathbf{P}\left(X_\infty^{\mathrm{R}} = i;\ Y_\infty^{\mathrm{R}} = 1\right) \\ &= \tilde{\pi}^{\mathrm{R}}(i,\ 0) + \tilde{\pi}^{\mathrm{R}}(i,\ 1). \end{aligned} \qquad [9.23]$$

In particular, under the stability condition [9.21], the series involved in the computation of $\tilde{\pi}^{\mathrm{R}}$ are necessarily summable, by [9.23]. Therefore, in that case a unique solution to the previous system necessarily exists, and the process $((X^{\mathrm{R}}(t), Y(t)),\ t \geq 0)$ is ergodic. Let us now observe that for any $i \in \mathbf{N}$,

$$\mu\pi^{\mathrm{R}}(i+1) - \lambda p_i \pi^{\mathrm{R}}(i) = 0,$$

which gives with [9.23] that

$$\mu\tilde{\pi}^{\mathrm{R}}(i+1,\,0)+\mu\tilde{\pi}^{\mathrm{R}}(i+1,\,1)-\lambda p_i\tilde{\pi}^{\mathrm{R}}(i,\,0)-\lambda p_i\tilde{\pi}^{\mathrm{R}}(i,\,1)=0. \qquad [9.24]$$

Further, it appears from the form of the generator $\tilde{A}^{\mathrm{R}}$ that

$$-\lambda p_i\tilde{\pi}^{\mathrm{R}}(i,\,0)-\mu\tilde{\pi}^{\mathrm{R}}(i,\,0)+\lambda\tilde{\pi}^{\mathrm{R}}(i,\,1)-\lambda p_i\tilde{\pi}^{\mathrm{R}}(i,\,1)+\mu\tilde{\pi}^{\mathrm{R}}(i+1,\,0)=0,$$

which, combined with [9.24], implies that for any $i \geq 0$,

$$\begin{aligned}&\tilde{\pi}^{\mathrm{R}}(i+1,\,0)=(\rho p_i+1)\tilde{\pi}^{\mathrm{R}}(i,\,0)+\rho(p_i-1)\tilde{\pi}^{\mathrm{R}}(i,\,1);\\&\tilde{\pi}^{\mathrm{R}}(i+1,\,1)=-\tilde{\pi}^{\mathrm{R}}(i,\,0)+\rho\tilde{\pi}^{\mathrm{R}}(i,\,1).\end{aligned}$$

Therefore, we have the recursive matrix relation

$$\begin{pmatrix}\tilde{\pi}^{\mathrm{R}}(i+1,\,0)\\ \tilde{\pi}^{\mathrm{R}}(i+1,\,1)\end{pmatrix}=\begin{pmatrix}\rho p_i+1 & \rho(p_i-1)\\ -1 & \rho\end{pmatrix}\begin{pmatrix}\tilde{\pi}^{\mathrm{R}}(i,\,0)\\ \tilde{\pi}^{\mathrm{R}}(i,\,1)\end{pmatrix} \qquad [9.25]$$

$$=M_i\begin{pmatrix}\tilde{\pi}^{\mathrm{R}}(i,\,0)\\ \tilde{\pi}^{\mathrm{R}}(i,\,1)\end{pmatrix}. \qquad [9.26]$$

Let us denote for any $i \geq 1$,

$$A_i=\prod_{j=0}^{i-1}M_j$$

and A_i^1 (respectively, A_i^2) the first (respectively, second) row of A_i, and recall relations [9.23] and

$$\sum_{i\in\mathbf{N}}(\tilde{\pi}^{\mathrm{R}}(i,\,0)+\tilde{\pi}^{\mathrm{R}}(i,\,1))=1.$$

The probability $\tilde{\pi}^{\mathrm{R}}$ is hence completely defined by

$$\begin{aligned}\begin{pmatrix}\tilde{\pi}^{\mathrm{R}}(i,\,0)\\ \tilde{\pi}^{\mathrm{R}}(i,\,1)\end{pmatrix}&=A_i\begin{pmatrix}\tilde{\pi}^{\mathrm{R}}(0,\,0)\\ \tilde{\pi}^{\mathrm{R}}(0,\,1)\end{pmatrix}\\&=A_i\left\{\tilde{\pi}^{\mathrm{R}}((0,\,0))\begin{pmatrix}1\\-1\end{pmatrix}+\begin{pmatrix}0\\ \pi^{\mathrm{R}}(0)\end{pmatrix}\right\},\ i\geq 0;\\ \tilde{\pi}^{\mathrm{R}}(0,\,0)&=\left[1-\sum_{i\in\mathbf{N}}\left\{A_i^1\begin{pmatrix}0\\ \pi^{\mathrm{R}}(0)\end{pmatrix}+A_i^2\begin{pmatrix}0\\ \pi^{\mathrm{R}}(0)\end{pmatrix}\right\}\right]\\&\qquad\cdot\left[\sum_{i\in\mathbf{N}}\left\{A_i^1\begin{pmatrix}1\\-1\end{pmatrix}+A_i^2\begin{pmatrix}1\\-1\end{pmatrix}\right\}\right]^{-1}.\end{aligned}$$

Hence, we can assess numerically $\tilde{\pi}^{\text{R}}$ by estimating the values of the series of the previous formula. According to the PASTA property and the ergodicity of the process $((X(t), Y(t)),\, t \geq 0)$, the loss probability is thus given by the formula

$$\begin{aligned}
P_l^{\text{R}} &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{\{0\}}(Y(T_n^-)) \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{i=0}^{\infty} \mathbf{1}_{\{(i,\,0)\}} \big((X^{\text{R}}(T_n^-), Y(T_n^-))\big) \\
&= \sum_{i=0}^{\infty} \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{\{(i,\,0)\}}((X^{\text{R}}(t), Y(t)))\,\mathrm{d}\,t \\
&= \sum_{i=0}^{\infty} \tilde{\pi}^{\text{R}}((i,\,0)).
\end{aligned}$$

9.10. A call center with impatient customers

We have seen that the Erlang model can represent a call center with S servers, where the customer calls are dropped whenever all lines are taken. As we shall see here, we can enrich the previous model in the case where the customers can be put on hold. It is then consistent to assume that customers are likely to become impatient, and hang up before having their call connected.

The model we chose is that of a queue with infinite capacity and impatient customers: more precisely, we consider a M$_\lambda$/M$_\mu$/S/S+M$_\alpha$ queue, that is the inter-arrival times and durations of calls are exponential, and the patience times of the customers $(D_n,\, n \in \mathbf{Z})$ before reaching an operator are independent and identically distributed of distribution $\varepsilon(\alpha)$, where $\alpha > 0$.

Let us stress the fact, that the customers are not impatient anymore as soon as they access an operator, so any customer who has not hung up before this time, continues his call until it ends: with the terminology of section 4.6, the patience runs until the beginning of service. Throughout this section, customers are served in FCFS.

Denote $(X^{\text{I}}(t), t \geq 0)$ as the process counting the number of customers in the system at any time (and add the exponent $^{\text{I}}$ for "impatience" to all parameters of the system). At t, $X^{\text{I}}(t)$ counts the number of customers in service and of waiting customers, irrespective of the fact that they will reach the server or not. This process may leave the state $i \in \mathbf{N}$ to visit the following states:

– the state $i+1$, if an arrival occurs,

– the state $i-1$, if $i \geq 1$ and a service ends, or provided $i > S$, if the patience of some customer in line has expired.

Since the residual patience time of the $i-S$ waiting customers are independent and follow at any times the $\varepsilon(\alpha)$ distribution, it is easily checked that $(X^{\text{I}}(t), t \geq 0)$ is Markov, with infinitesimal generator given by

– $A^{\text{I}}(i,\, i+1) = \lambda$ for any $i \geq 1$;

– $A^{\text{I}}(i,\, i-1) = i\mu$ for any $i \in [\![1,\, S]\!]$;

– $A^{\text{I}}(i,\, i-1) = S\mu + (i-S)\alpha$ for any $i \geq S+1$.

We can then compute, as usual, the stationary probability π^{I} of $(X^{\text{I}},\, t \geq 0)$

$$\pi^{\text{I}}(i) = \frac{\rho^i}{i!}\pi^{\text{I}}(0) \text{ for any } i \in [\![0,\, S]\!];$$

$$\pi^{\text{I}}(i) = \frac{\lambda^{i-S}\rho^S}{\prod_{j=1}^{i-S}(S\mu + j\alpha)S!}\pi^{\text{I}}(0) \text{ for any } i \geq S+1;$$

$$\pi^{\text{I}}(0) = \left(\sum_{i=0}^{S}\frac{\rho^i}{i!} + \sum_{i=1}^{\infty}\frac{\lambda^i\rho^S}{\prod_{j=1}^{i}(S\mu + j\alpha)S!}\right)^{-1}.$$

Estimate of the loss probability

As in the Erlang model, we aim to dimension the system by determining the optimal value of S to ensure a target loss probability. The exact computation of this probability is tedious, and is based on technical arguments which are beyond the scope of this book. We can nevertheless give a heuristic estimate.

In this system, the customer C_n is lost if, and only if, it finds in the system a waiting time W_n greater than its patience time D_n. The loss probability is hence given by

$$P_l^{\text{I}} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}_{\{[0,D_n]\}}(W_n).$$

According to the results of section 4.6, as [4.97] clearly holds, there exists a stationary waiting time TA. The r.v. TA and D are clearly independent (once again, see the constructions of stationary waiting times in Chapter 4), Theorem 2.7 implies that

$$\begin{aligned}P_l^{\mathrm{I}} &= \mathbf{P}\left((\mathrm{T}_{\mathrm{A}}, D) \in \{(x,y);\, x \geq y\}\right)\\ &= \int_{\mathbf{R}+} \int_0^x \alpha e^{-\alpha y}\, dy\, P_{\mathrm{T_A}}(dx)\\ &= 1 - \mathcal{L}_{W_\infty}(\alpha),\end{aligned}$$

where $\mathcal{L}_{W_\infty}$ is the Laplace transform of W_∞.

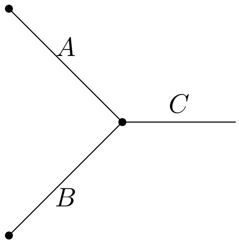

Figure 9.7. *Single hub*

9.11. Problems

EXERCISE 22.– We consider the following concentrator/hub:

Class 1 customers are those who use the connections A and C. Class 2 customers are those using the connections B and C. We denote S_A, S_B, S_C as the capacities of each one of the links. We set $X = (X^1, X^2)$, with $X^i(t)$ the number of class i calls in progress at t. Arrivals of calls of class i form a Poisson process of intensity λ_i, for $i = 1, 2$. The call durations of class i customers are exponentially distributed of average $1/\mu_i$. We set $\rho_i = \lambda_i/\mu_i$.

1) We assume at first that $S_A = S_B = S_C = \infty$. Write the infinitesimal generator of X.

2) What is its stationary probability?

3) Show that it is reversible.

4) Describe the state space S when the capacities are finite.

5) Deduce the stationary probability π of X when all the capacities are finite.

6) Show that the probability P_B^i of blocking (and thus, of loss) of calls of class i is of the form

$$p_i = 1 - \frac{\sum_{(n_1,n_2)\in S_i} \pi(n_1, n_2)}{\sum_{(n_1,n_2)\in S} \pi(n_1, n_2)},$$

where S_i is a subset of S which we will specify.

7) Numerical application: $S_A = S_B = 2,\ S_C = 3,\ \rho_1 = \rho_2 = 2$. Compute p_1.

8) Compare the loss probability to the quantity

$$1 - (1 - E[\rho_1, S_A])(1 - E[\rho_1, S_C]).$$

What does the latter represent?

EXERCISE 23.– Consider the process X counting the number of customers in a M$_\lambda$/M$_\mu$/S/S+C queue.

1) Show that X is Markov, and give its generator.

2) Show that there exists a unique stationary probability π, and express π.

3) Give the loss probability of the system.

9.12. Notes and comments

The explicit computations for the dimensioning of GSM networks with hand-overs can be found in the lectures notes of X. Lagrange, those for the A-bis interface were proposed by N. Dailly. The dimensioning of hierarchical networks is an old problem. The old methods were based on the equivalent trunk method of Kuczura and Wilkinson (see [ITU]). The approach using MMPP processes is due to [MEI 89]. Its application to mobile networks is inspired by [LAG 96].

Epitome

– The loss probability equals the blocking probability, only if the arrival process is Poisson. In other cases, one can refer to Theorem A.34.

– In the M/M/S/S queue, the loss probability is given by Erlang-B formula

$$\mathrm{Er}[\rho,\, S] = \frac{\rho^S/S!}{\sum_{i=0}^{S} \rho^i/i!}.$$

– In the Engset model, the loss probability is given by

$$\mathrm{Eng}[\rho,\, S,\, M] = \frac{\rho^S C_{M-1}^S}{\sum_{j=0}^{S} C_{M-1}^j \rho^j},$$

where S is the number of servers and M represents the number of sources.

– The loss probability depends not only on the load, see the case of the IPP/M/S/S queue.

PART 3

Spatial Modeling

Chapter 10

Spatial Point Processes

10.1. Preliminary

In radio communications, the distance between the transmitter and the receiver plays a crucial role. To evaluate the performance of radio-cellular protocols, it is customary to consider that the access points or base stations are evenly distributed in a hexagonal pattern; see Figure 10.1.

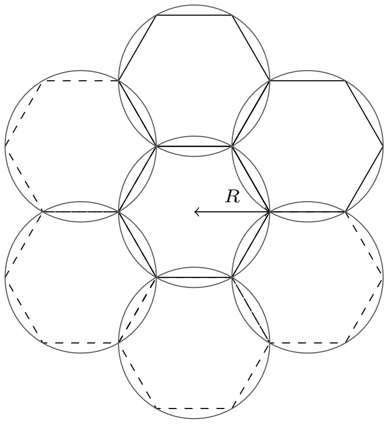

Figure 10.1. *Hexagonal network of base stations*

The mobile phones are often modeled by a continuum: a call can be transmitted from a point x with an infinitesimal probability dx. This approach which is very macroscopic prevents very precise and realistic calculations. For the last few years,

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

under the influence of works of F. Baccelli, the models stemming from stochastic geometry are gaining more and more attention. They enable us to represent the reality more precisely and make calculations more rigorously.

10.2. Stochastic geometry

The concept of configuration is specified in example A.1. Let us recollect the definition, and see section A.1.2 for details.

DEFINITION 10.1.– *A configuration is a locally finite set of points of a set E: there is a finite number of points in any bounded set. We denote $\mathfrak{N}_E$ as the set of configurations of E.*

EXAMPLE 10.1 (BERNOULLI PROCESS).– The Bernoulli point process is a process based on a finite set $E = \{x_1, \cdots, x_n\}$. Each of these points is ON, independently of others and with probability p. If we introduce $A_1, \cdots, A_n$ random independent variables of Bernoulli distribution with p parameter, we can write

$$N = \sum_{i=1}^{n} A_i \delta_{x_i}.$$

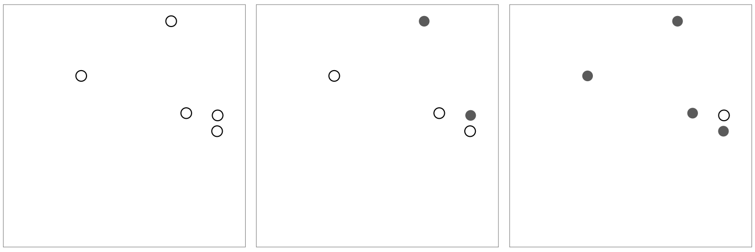

Table 10.1. *On the left, the set E. In the middle and on the right, two possible realisations. In full (red), the ON points*

EXAMPLE 10.2 (BINOMIAL PROCESS).– The number of points is fixed to n and μ, a probability measure on $\mathbf{R}^2$ is given. According to μ, the atoms are drawn randomly independent of each other.

We can easily calculate that

$$\mathbf{P}(N(A) = k) = \binom{n}{k} \mu(A)^k (1 - \mu(A))^{n-k},$$

and for the disjoint sets $A_1, \cdots, A_n$

$$\mathbf{P}(N(A_1) = k_1, \cdots, N(A_n) = k_n) = \frac{(k_1 + \ldots + k_n)!}{k_1! \ldots k_n!} \mu(A_1)^{k_1} \ldots \mu(A_n)^{k_n}. \quad [10.1]$$

10.3. Poisson process

The point process, mathematically the richest, is the spatial Poisson process which we recognise as that which generalizes the Poisson process on the real straight line introduced in Chapter 6.

DEFINITION 10.2.– *Let μ be a Radon measure on a Polish space E that is $\mu(\Lambda) < \infty$ for every compact set $\Lambda \subset E$. The Poisson process with intensity μ is defined by its Laplace transform: for any function $f : E \to \mathbf{R}^+$,*

$$\mathbf{E}\left[\exp(-\int f \, \mathrm{d} N)\right] = \exp\left(-\int_E (1 - e^{-f(s)}) \, \mathrm{d}\, \mu(s)\right).$$

To clarify that the intensity measure is μ, we will often index the expectation by μ.From the definition of a Poisson process, we immediately infer the Campbell formula by differentiation.

THEOREM 10.1 (CAMPBELL FORMULA).– *Let $f \in L^1(E, \mu)$,*

$$\mathbf{E}_\mu\left[\int f \, \mathrm{d} N\right] = \int_E f \, \mathrm{d}\, \mu$$

and if $f \in L^2(E \times E, \mu \otimes \mu)$, then

$$\mathbf{E}_\mu\left[\sum_{x \neq y \in N} f(x, y)\right] = \iint_{E \times E} f(x, y) \, \mathrm{d}\, \mu(x) \, \mathrm{d}\, \mu(y).$$

NOTE.– Particularly, for $f = \mathbf{1}_A$ where A is a compact of E, we notice that $\mathbf{E}\left[N(A)\right] = \mu(A)$. If $\mu = \lambda \, \mathrm{d}\, x$, then λ represents the average number of customers per unit area.

An alternative definition is as follows:

THEOREM 10.2.– *Let μ be a Radon measure on a Polish space E. The Poisson process with intensity μ is the probability measure on $\mathfrak{N}_E$ such that:*

– For every compact set $\Lambda \subset E$, $N(\Lambda)$ follows a Poisson distribution with parameter $\mu(\Lambda)$.

– For Λ_1 and Λ_2 two disjoint subsets of $(E, \mathfrak{B}(E))$, the random variables $N(\Lambda_1)$ and $N(\Lambda_2)$ are independent.

From this second definition, we immediately deduce the result of the following result of uniformity.

THEOREM 10.3.– *Let N be a Poisson process with intensity μ. Let $\Lambda \subset E$ be a compact set. Given that $N(\Lambda) = n$, the atoms are distributed according to a binomial process for $\mu_\Lambda(A) = \mu(A \cap \Lambda)/\mu(\Lambda)$.*

Proof. Let $A_1, \cdots, A_m$ be a partition of Λ or $(k_1, \cdots, k_m)$ such that $k_1 + \ldots + k_m = n$.

$$\begin{aligned}
\mathbf{P}\left(N(A_i) = k_i,\, i = 1, \cdots, m \,\middle|\, N(\Lambda) = n\right) \\
&= \frac{\mathbf{P}(N(A_i) = k_i,\, i = 1, \cdots, m,\, N(\Lambda) = n)}{\mathbf{P}(N(\Lambda) = n)} \\
&= \frac{\mathbf{P}(N(A_i) = k_i,\, i = 1, \cdots, m)}{\mathbf{P}(N(\Lambda) = n)} \\
&= \frac{\exp(-\sum_{i=1}^m \mu(A_i)) \prod_{i=1}^m \dfrac{\mu(A_i)^{k_i}}{k_i!}}{\exp(-\mu(\Lambda)) \frac{\mu(\Lambda)^n}{n!}} \qquad [10.2] \\
&= \frac{n!}{k_1! \ldots k_m!} \prod_{i=1}^m \left(\frac{\mu(A_i)}{\mu(\Lambda)}\right)^{k_i}.
\end{aligned}$$

According to [10.1] for μ_Λ, we see that, given the number of atoms in Λ, they are distributed according to a binomial process. □

From this result, we deduce a way to simulate a Poisson process on any set Λ, such that $\mu(\Lambda)$ is finite.

Algorithm 10.1. Simulation of realisation of a P.P.(μ) on a set Λ.

```
Data: $\mu$, $\Lambda$
Result: $n$= realisation of a random variable with Poisson distribution($\mu(\Lambda)$)
for $i = 1$ to $n$ do
    $X_i$ = draw a point with distribution $\mu/\mu(\Lambda)$
end
return $n$, $X_i$, $i = 1, \cdots, n$
```

EXAMPLE 10.3 (M/M/∞ QUEUE).– The M/M/∞ queue is the queue with Poisson arrivals, independent and identically distributed from exponential distribution service times, and an infinite number of servers (without buffer). It is initially a theoretical object which is particularly simple to analyze and also a model to which we can compare other situations. Due to the independence of the inter-arrivals and service time, according to the second characterization of Poisson processes, the process

$$N = \sum_{n \geq 1} \delta_{(T_n,\, S_n)}$$

where T_n is the instant of nth arrival and S_n the nth service time, is a Poisson process with $d\mu(t,x) = \lambda \,\mathrm{d}\, t \otimes \mu e^{-\mu x} \,\mathrm{d}\, x$ intensity in $E = \mathbf{R}^+ \times \mathbf{R}^+$.

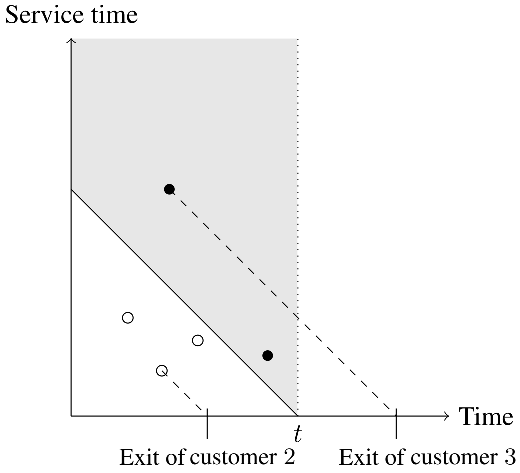

Figure 10.2. *The M/M/∞ queue as a Poisson process in $\mathbf{R}^+ \times \mathbf{R}^+$. Customers still in use at time t are those that correspond to points in the shaded trapezoid*

The customers who are still in service at the time are those who correspond to the points in the shaded trapezium.

We deduce that $X(t)$, the number of busy servers at time t follows a Poisson distribution with parameter

$$\int_0^t \left(\int_{t-s}^{\infty} \mu e^{-\mu x} \,\mathrm{d}\, x \right) \lambda \,\mathrm{d}\, s = \lambda \int_0^t e^{-\mu(t-s)} \,\mathrm{d}\, s = \rho(1 - e^{-\mu t}),$$

where $\rho = \lambda/\mu$. If the system is not empty at time 0, we must add $X(t)$ the number of initial customers still in service at time t. If X_0 follows a Poisson distribution with parameter ρ_0, the number of customers in service at time t follows a Poisson distribution with parameter $\rho_0 e^{-\mu t}$ because each and every customer has a probability $e^{-\mu t}$ of being still in service and the total is thus the thinning of a Poisson random variable. In conclusion, $X(t)$ then follows a Poisson distribution with parameter $\rho + (\rho_0 - \rho)e^{-\mu t}$. Irrespective of the value of ρ_0, the stationary probability of X is a Poisson distribution with parameter ρ.

NOTE.– Let us illustrate some of the differences between a Poissonian model and a hexagonal model for wireless networks. In a hexagonal model, the average number of users per unit of area is the inverse of the area of a hexagon. If the radius of the hexagon is R, this gives an average density of $2/(3\sqrt{3})R^{-2}$. In a Poissonian model, the average number of users per unit of area is λ. In the same way as for comparing two queues, the load must be identical to compare two spatial systems, and the average number of users must be identical. We must therefore choose λ and R^2 such that $\lambda R^2 = 2.3^{-3/2}$.

One of the essential parameters is as we have said, the distance. In particular, a short distance between resources ensures a better coverage, but creates interferences. In a hexagonal model, if the cell radius is R, the distance between nearest neighbours is $R\sqrt{3}$. Let us calculate this quantity in the case of a Poisson process with $\lambda\, \mathrm{d}\, x$ intensity on $\mathbf{R}^2$. Let $x \in \mathbf{R}^2$ be any point in the plane

$$D_x(N) = \mathrm{d}(x,\, N) = \inf\{\|x - y\|,\ y \in N\}.$$

It is clear that we have

$$\begin{aligned}\mathbf{P}(d(x,\, N) \geq \tau) &= \mathbf{P}(N(B(x,\, \tau)) = 0)\\ &= \exp(-\lambda\pi\tau^2).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{E}\left[D_x\right] &= \int_0^{+\infty} \mathbf{P}(D_x \geq \tau)\, \mathrm{d}\, \tau\\ &= \int_0^{+\infty} \exp(-\lambda\pi\tau^2)\, \mathrm{d}\, \tau\\ &= \int_0^{+\infty} \exp(-u^2/2) \frac{1}{\sqrt{2\lambda\pi}}\, \mathrm{d}\, u\\ &= \frac{1}{2\sqrt{\lambda}},\end{aligned}$$

because we recognise the semi-integral of the Gaussian density with variance λ. Using the Palm theory, we could show that this result holds true for the distance at any point of the process and its nearest neighbour. In conclusion, if $\lambda R^2 = 2/3^{3/2}$, we obtain that the average distance in the Poisson model is approximately $0.8R$, and that is much less than the distance in the hexagonal model. For the interferences, which are inversely proportional to the distance, the Poisson model is thus more pessimistic than the hexagonal model.

EXAMPLE 10.4 (MEAN INTERFERENCE).– At a point x of the plane, the interference created by other mobiles is expressed by

$$I(x,\, N) = \sum_{y \in N} h(y) P(y) l(\|y - x\|),$$

where $P(x)$ is the power of the signal emitted by the mobile in y, l is a function of $\mathbf{R}^+$ in $\mathbf{R}^+$ which we generally take from the form

$$l_0(r) = r^{-\gamma} \text{ or } l_1(r) = \min(1,\ r^{-\gamma}). \qquad [10.3]$$

The second formulation gives less elegant formulas but is more realistic (a signal is not going to be amplified on the grounds that the receiver is very close to the transmitter) and avoids indefinite integrals. The random variables $(h(y),\ y \in N)$ are generally identically distributed (the same distribution as that of a random variable H), independent of each other and independent of N. They represent the loss factor induced by the fading (the attenuation due to local movements of the receiver) and shadowing (signal attenuation due to obstacles between the transmitter and receiver). In general, the fading is modeled by a a random variable exponentially distributed with one parameter. The shadowing is represented by a log-normal distribution, i.e. the exponential of a Gaussian variable.

Campbell's formula indicates that

$$\mathbf{E}_\mu\left[I(x)\right] = \mathbf{E}\left[H\right] \int P(y) l(\|y-x\|) \,\mathrm{d}\,\mu(y).$$

Then, assume that power is the same for all the mobiles and that μ is proportional to the Lebesgue measure, that is $\mathrm{d}\,\mu(x) = \lambda \,\mathrm{d}\,x$. We immediately observe that the previous quantity is not dependent on x, hence the result

$$\mathbf{E}_\mu\left[I(0)\right] = P\mathbf{E}\left[H\right] \int l(\|y\|)\lambda \,\mathrm{d}\,y = \lambda \mathbf{E}\left[H\right] \int_0^\infty l(r) r \,\mathrm{d}\,r.$$

If we take path-loss model as in [10.3], we obtain for a cell of radius $R > 1$

$$\mathbf{E}_\lambda\left[I(0)\right] = \mathbf{E}\left[H\right] \lambda \left(\pi + \frac{\pi}{\gamma - 2}(1 - R^{2-\gamma}) \right).$$

For large R, this quantity is approximately equal to $\mathbf{E}\left[H\right] \pi\lambda\gamma$ for $\gamma > 2$.

EXAMPLE 10.5 (INTERFERENCE DISTRIBUTION).– Now, suppose that the random variables $(h(x),\ x \in \mathbf{R}^2)$ are independent of the same distribution. For s real positive

$$\begin{aligned}
\mathbf{E}&\left[\exp(-s \int h(x) l(\|x\|) \,\mathrm{d}\,N(x))\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\exp(-s \int h(x) l(\|x\|) \,\mathrm{d}\,N(x)) \mid N(\Lambda)\right]\right] \\
&= \mathbf{E}\left[\prod_{x \in N} \int \exp(-s l(\|x\|) y) \,\mathrm{d}\,\mathbf{P}_H(y)\right],
\end{aligned}$$

since given the number of points in Λ, the atoms are independent of each other. By denoting $\mathcal{L}_H$ as the Laplace transform of H, we obtain

$$\begin{aligned}\mathbf{E}\left[\exp(-s\int h(x)l(\|x\|)\,\mathrm{d}\,N(x))\right] &= \mathbf{E}\left[\prod_{x\in N}\mathcal{L}_H(sl(\|x\|))\right]\\ &= \mathbf{E}\left[\exp(\int_\Lambda \ln\mathcal{L}_H(sl(\|x\|))\,\mathrm{d}\,N(x))\right]\\ &= \exp(-\int_\Lambda 1-e^{\ln\mathcal{L}_H(sl(\|x\|))}\,\mathrm{d}\,\mu(x))\\ &= \exp(\int_\Lambda(\ln\mathcal{L}_H(sl(\|x\|))-1)\,\mathrm{d}\,\mu(x)).\end{aligned}$$

For Rayleigh fading, H is exponentially distributed with parameter 1. If we assume that $\mu = \lambda\,\mathrm{d}\,x$ and that the path-loss is given by l_0, all calculations are feasible and we obtain the following formula (see [HAE 08, equation (3.21)])

$$\mathcal{L}_{I(0)}(s) = \exp(-\pi\lambda s^\delta\,\frac{\pi\delta}{\sin(\pi\delta)}),$$

where $\delta = 2/\gamma$. Then we know that this corresponds to a stable distribution of characteristic exponent δ; see [SAM 94].

Most of the properties of real Poisson process are transferred to the spatial Poisson process.

THEOREM 10.4 (INTEGRATION).– *Let N^1 and N^2 be two independent Poisson processes with respective intensities μ^1 and μ^2, their superposition N defined by*

$$\int f\,\mathrm{d}\,N = \int f\,\mathrm{d}\,N^1 + \int f\,\mathrm{d}\,N^2$$

is a Poisson process with intensity $\mu^1+\mu^2$.

DEFINITION 10.3.– *Let N be a Poisson process with intensity μ and $p\,:\,E\longrightarrow[0,\,1]$. The $(\mu,\,p)$-thinned Poisson process is the process where an atom of the Poisson process N in x is kept with probability $p(x)$.*

THEOREM 10.5 (THINNING).– *A $(\mu,\,p)$-thinned Poisson process is a Poisson process of intensity μ_p defined by*

$$\mu_p(A) = \int_A p(x)\,\mathrm{d}\,\mu(x).$$

NOTE.– This result is interesting in the framework of modeling. If the users are represented by the points of a point Poisson process, only those that emit at a given

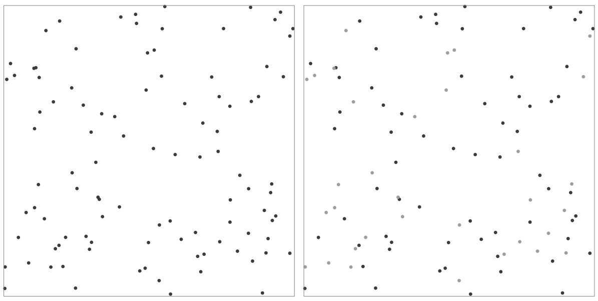

Table 10.2. *A realisation of a Poisson process (on the left) and one of its thinning with $p = 2/3$ (on the right). Filled circles correspond to kept points*

time concern the operator. We can assume that each user has a probability p to emit at any given time irrespective of other mobiles. Theorem 10.5 indicates that the active users are scattered in the plane according to a Poisson process of intensity λp.

Theorem 10.5 is a special case of the displacement theorem.

DEFINITION 10.4.– *Let $(\Omega', \mathcal{A}', \mathbf{P}')$ be a probability space and $(F, \mathcal{F})$ a Polish space. A displacement is a measurable application Θ of $\Omega' \times E \longrightarrow F$ such that the random variables $(\Theta(\omega', x), x \in E)$ are independent. For $A \in \mathcal{F}$, we have*

$$\theta(x, A) = \mathbf{P}'(\omega' : \Theta(\omega', x) \in A).$$

Thus $\theta(x, A)$ represents the probability that the point x is displaced in A. More mathematically, if we denote by $\Theta(\omega', .)^*\mu$ the image measure of μ through the application $\Theta(\omega', .)$, we have

$$\begin{aligned}\mathbf{E}_{\mathbf{P}'}\left[\Theta^*\mu(A)\right] &= \mathbf{E}_{\mathbf{P}'}\left[\int \mathbf{1}_{\{\Theta(\omega', x) \in A\}} \, \mathrm{d}\,\mu(x)\right] \\ &= \int \mathbf{P}'(\Theta(\omega', x) \in A) \, \mathrm{d}\,\mu(x) = \int \theta(x, A) \, \mathrm{d}\,\mu(x).\end{aligned}$$

This means that

$$\mathbf{E}_{\mathbf{P}'}\left[\int \mathbf{1}_A \, \mathrm{d}\,\Theta^*\mu\right] = \int \int_A \theta(x, \mathrm{d}\,y) \, \mathrm{d}\,\mu(x).$$

Therefore, for a non-negative function f, we obtain

$$\mathbf{E}_{\mathbf{P}'}\left[\int f \, \mathrm{d}\,\Theta^*\mu\right] = \int \int f(y)\theta(x, \mathrm{d}\,y) \, \mathrm{d}\,\mu(x). \qquad [10.4]$$

DEFINITION 10.5.– *A displacement is said to be conservative when, for any compact* $\Lambda \subset E$

$$\mathbf{E}_{\mathbf{P}'}\left[\Theta^*\mu(\Lambda)\right] = \int_\Lambda \int_F \theta(x,\,\mathrm{d}\,y)\,\mathrm{d}\,\mu(x) = \mu(A).$$

This signifies that on average, the total mass of the point process is preserved.

Let Θ be a displacement such that $\int_\Lambda \int_F e^{-f(y)}\theta(x,\,\mathrm{d}\,y)\,\mathrm{d}\,\mu(x) = \mu(A)$ and N be a point process, the displaced point process N^Θ is defined by

$$N^\Theta(\omega') = \sum_{x\in N} \delta_{\Theta(\omega',\,x)}.$$

THEOREM 10.6 (DISPLACEMENT).– *Let* N *be a Poisson process with intensity* μ *on* E *and* Θ *be a conservative displacement from* E *to* F*. The process* N^Θ *is a Poisson process with intensity* μ^Θ *defined by*

$$\mu^\Theta(A) = \int_E \theta(x,\,A)\,\mathrm{d}\,\mu(x).$$

Proof. First, assume that f has a compact support denoted by Λ. We know that given $N(\Lambda)$, the atoms of N are independent, distributed according to $\mu/\mu(\Lambda)$. Therefore, we can write

$$\mathbf{E}\left[F\exp(-\int_\Lambda \mathrm{d}\,N)\right] = \sum_{n=0}^{\infty} \frac{e^{-\mu(\Lambda)}\mu(\Lambda)^n}{n!} \int_{E^n} \prod_{j=1}^{n} e^{-f(x_j)}\frac{\mathrm{d}\,\mu(x_j)}{\mu(\Lambda)}.$$

According to the construction of N^Θ, the random displacement is independent of N, thus, we have

$$\begin{aligned}\mathbf{E}\left[\exp(-\int f\,\mathrm{d}\,N^\Theta)\right] &= \mathbf{E}_{\mathbf{P}'}\left[\sum_{n=0}^{\infty} \frac{e^{-\mu(\Lambda)}}{n!} \int_{E^n} \prod_{j=1}^{n} e^{-f(\Theta(\omega',\,x_j))}\,\mathrm{d}\,\mu(x_j)\right] \\ &= \sum_{n=0}^{\infty} \frac{e^{-\mu(\Lambda)}}{n!}\,\mathbf{E}_{\mathbf{P}'}\left[\int_{E^n} \prod_{j=1}^{n} e^{-f(\Theta(\omega',\,x_j))}\,\mathrm{d}\,\mu(x_j)\right].\end{aligned}$$

By definition of a displacement, the random variables $(\Theta(\omega', x_j), j = 1, \cdots, n)$ are independent. By using [10.4],we obtain,

$$\begin{aligned}\mathbf{E}\left[\exp(-\int f \,\mathrm{d}\, N^{\Theta})\right] &= \sum_{n=0}^{\infty} \frac{e^{-\mu(\Lambda)}}{n!} \left(\mathbf{E}_{\mathbf{P}'}\left[\int_E e^{-f(\Theta(\omega', x))} \,\mathrm{d}\,\mu(x)\right]\right)^n \\ &= \sum_{n=0}^{\infty} \frac{e^{-\mu(\Lambda)}}{n!} \left(\int e^{-f} \,\mathrm{d}\,\mu^{\Theta}\right)^n \\ &= \exp\left(-\mu(\Lambda) + \int_\Lambda \int_F e^{-f(y)} \theta(x, \mathrm{d}\, y) \,\mathrm{d}\,\mu(x)\right).\end{aligned}$$

As Θ is conservative, we obtain

$$\mathbf{E}\left[\exp(-\int f \,\mathrm{d}\, N^{\Theta})\right] = \exp\left(-\int_F (1 - e^{-f(y)}) \int_\Lambda \theta(x, \mathrm{d}\, y) \,\mathrm{d}\,\mu(x)\right),$$

so N^{θ} is definitely a Poisson process with intensity μ^{Θ}.

We obtain the general case for f, by truncation (apply the previous result to $f_\Lambda = f\,\mathbf{1}_\Lambda$) and by a limit procedure (consider an increasing sequence of compacts $(\Lambda_n, n \geq 1)$ such that $\cup_n \Lambda_n = E$. Note that the existence of such a sequence is ensured by the Polish character of E.). □

of Theorem 10.5. We consider $F = E \cup \Delta$ where Δ is an external point. With probability $p(x)$, the atom x stays in x, with the complementary probability, it is moved to Δ. This displacement is conservative as we keep the same number of atoms. The restriction at E of the process thus obtained is the thinning of the initial process. Theorem 10.5 is then a direct consequence of Theorem 10.6. □

By applying Theorem 10.6 to the function $(x \in \mathbf{R}^d \mapsto rx)$ where $r \in \mathbf{R}^+$, we obtain a scaling property which is very useful in many applications.

COROLLARY 10.7.– *Let N be a Poisson process with intensity μ on $\mathbf{R}^d$. Let $r > 0$, N^r is the dilation of N process defined by*

$$N^{(r)} = \sum_{x \in N} \delta_{rx}.$$

The process N^r is a Poisson process with intensity $\mu^{(r)}$ where $\mu^{(r)}(A) = \mu(A/r)$ for any $A \in \mathfrak{B}(E)$.

COROLLARY 10.8.– *Let N be a Poisson process with intensity $\lambda \, \mathrm{d}\, x$ on $\mathbf{R}^d$. The process of modules is independent of the process of arguments. The first is a Poisson process with intensity $2\lambda\pi r \, \mathrm{d}\, r$, and the second is a Poisson process of intensity $(2\pi)^{-1} \boldsymbol{1}_{[0,\, 2\pi]}(\theta) \, \mathrm{d}\, \theta$.*

Proof. Theorem 10.6 implies that

$$\hat{N} = \sum_{x \in \mathbf{N}} \delta_{\|x\|,\, \mathrm{Arg}(x)}$$

is a Poisson process with $\lambda r \, \mathbf{1}_{[0,\, 2\pi]}(\theta) \, \mathrm{d}\, r \, \mathrm{d}\, \theta$ intensity. Hence, we have the result. □

EXAMPLE 10.6 (OFDMA PROTOCOL).– Let us illustrate these results in the special case of OFDMA protocol. In this protocol, time and frequency bands are cut into pieces called subcarriers. To a communication, one or more subcarriers are allocated. In practice, the allocation is for a period of a few slots and the assignments are therefore similar to that of Figure 10.3.

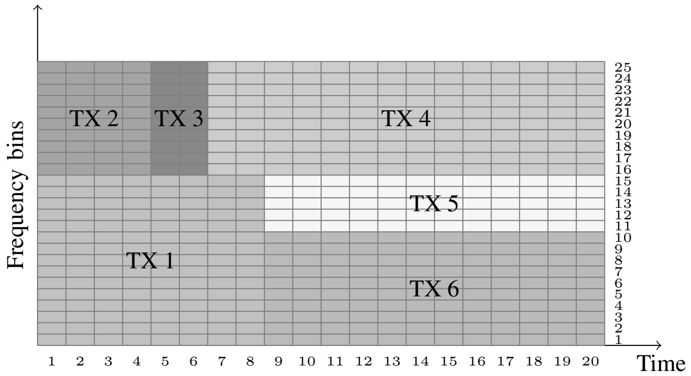

Figure 10.3. *Division and allocation in time-frequency space for the OFDMA protocol*

For radio communications, it is imperative that the signal to noise plus interference ratio (SINR) be large enough, so that communication can be established. Note P_e the transmitted power, A the attenuation factor due to fading and shadowing, γ the coefficient of path-loss, and η the minimal value of admissible SINR to establish a communication. The capacity formula of Shannon stipulates that for a user wishing to transmit at rate C, the required number of subcarriers is given by

$$Q(x) = \left\lceil \frac{C}{W \log_2 \left(1 + \dfrac{A\, P_e}{\|x\|^\gamma} \right)} \right\rceil \mathbf{1}_{\{A\, P_e \|x\|^{-\gamma} > \eta\}},$$

where W is the bandwidth of each subcarrier. The minimal SINR ratio varies between 0.001 (excellent channel) and 0.1 (worse conditions). In order to simplify the analysis, we assume that the assignments are made for each slot. We want to determine the probability that the number of available subcarriers is greater than the number of subcarriers required.

We assume that the access point is located at the origin of the plane and that the users are spatially distributed according to a Poisson process with intensity $\lambda \, \mathrm{d}\, x$. A simple model to represent the active users is to assign to each user, a probability of activity p possibly depending on its position and to consider that the activity or inactivity of a mobile is independent of each others. On the basis of Theorem 10.2, the point process of active users is still a Poisson process with intensity $\lambda \int p(x) \, \mathrm{d}\, x$. The total number of subcarriers required is then

$$M = \int_\Lambda Q(x) \, \mathrm{d}\, N(x),$$

where Λ is the domain that represents the cell covered by the access point in $(0,\, 0)$. By assuming that A and P_e are deterministic and independent of x, based on the Campbell formula, we get

$$\mathbf{E}\,[M] = \lambda \int_\Lambda Q(x)p(x) \, \mathrm{d}\, x.$$

If A and P_e depend on x and on a randomness source independent of N, we have a similar expression

$$\mathbf{E}\,[M] = \lambda \int_\Lambda Q(x)p(x) \, \mathrm{d}\, x.$$

It is often impossible to explicitly calculate the laws of random variables constructed from a point process such as M. The stochastic analysis gives us the tools to obtain some information on these distributions (we refer the reader to section 10.4 for the notations and proofs).

THEOREM 10.9.– *Let N be a Poisson process with intensity μ on E. Let $F \in$ Dom D such that $\mathbf{E}\,[F] = 0$. We, therefore, have the following inequality*

$$d_{TV}(\mathbf{P}_F,\, \mathcal{N}(0,\, 1)) \le \mathbf{E}\left[\left|1 - \int_E D_xFD_xL^{-1}F \, \mathrm{d}\, \mu(x)\right|\right] + \int_E \mathbf{E}\left[|D_xF|^2|D_xL^{-1}F|\right] \mathrm{d}\, \mu(x),$$

where d_{TV}represents the total variation distance between two probability measures, and $\mathcal{N}(0,\, 1)$ is the centered Gaussian measure on $\mathbf{R}$.

COROLLARY 10.10.– *Let $N_1, \cdots, N_K$ be independent spatial Poisson processes with respective intensity $\lambda_i \,\mathrm{d}\, x$, on a bounded domain Λ. Let F be a functional of the form*

$$F = \sum_{i=1}^{K} \sum_{n=1}^{N^i(\Lambda)} Y_n^i$$

where $(Y_n^i,\, i,\, n \in \mathbf{N})$ are independent variables whose distribution does not depend on anything other than the index exponent, that is Y_n^i and Y_m^i have the same distribution but it is not necessarily the case for Y_n^i and Y_n^j. Let

$$c^2 = \sum_{i=1}^{K} \lambda_i(\Lambda) \int_{\mathbf{R}} m^2 \, d\mathbf{P}_{Y_1^i}(m),$$

where $\lambda_i(\Lambda) = \int_\Lambda \lambda_i \,\mathrm{d}\, x$ is the area of Λ multiplied by λ_i. We obtain

$$d_{TV}(\mathit{Dist.}((F - E[F])/c), \mathcal{N}(0,1)) \le \frac{1}{c^3} \left(\sum_{i=1}^{K} \lambda_i(\Lambda) \int_{\mathbf{R}} m^3 \, d\mathbf{P}_{Y_1^i}(m) \right).$$

In particular,

$$\left| \mathbf{P}(\frac{F - E[F]}{c} \ge x) - \int_x^\infty e^{u^2/2} \frac{du}{\sqrt{2\pi}} \right| \le \frac{1}{c^3} \left(\sum_{i=1}^{K} \lambda_i \int_\Lambda dx \int_{\mathbf{R}} m^3 \, d\mathbf{P}_{Y_1^i}(m) \right).$$

Proof. As the individual Poisson processes are independent of each other, the process

$$\sum_{i=1}^{K} \sum_{(x,m) \in N^i} \delta_{x,m}$$

is a Poisson process on $\mathbf{R}^+ \times \mathbf{R}$ with intensity

$$\sum_{i=1}^{K} \lambda_i \, dx \otimes \, d\mathbf{P}_{Y_1^i}(m).$$

In this case, the gradient operator is defined by

$$D_{x,m} F(\omega) = F(\omega + \delta_{x,m}) - F(\omega).$$

The functional F is rewritten

$$F = \int_\Lambda \int_{\mathbf{R}} m \,\mathrm{d}(\sum_{i=1}^{K} \sum_{(x,m) \in N^i} \delta_{x,m})$$

where Λ is the domain of the plane on which we work. Therefore, we have $D_{x,m} F = m$. The result follows from Theorem 10.9. □

EXAMPLE (Continuation of example 10.6).– Initially, we assume that fading is constant. This is an unrealistic hypothesis but serves to illustrate the complexity of the problem. In this case, for each mobile, Q does not depend on anything other than its position. We thus have a set of concentric rings around the base station in each of which the number of subcarriers required is the same. In fact,

$$Q(x) = k \Longleftrightarrow k - 1 < \frac{C}{W \log_2 \left(1 + \dfrac{A\,P_e}{\|x\|^\gamma}\right)} \leq k \Longleftrightarrow R_{k-1} < x < R_k,$$

where $R_0 = 0,\ R_k = \left(\frac{AP_e}{\left(2^{C_0/(kW)}-1\right)I}\right)^{1/\gamma}$. The maximum radius R_M is defined by the relation

$$\frac{A\,P_e}{\|R_M\|^\gamma} = \eta, \text{ thus } R_M = \left(\frac{A\,P_e}{\eta}\right)^{1/\gamma}.$$

Similarly, the maximum number of subcarriers required is $N_M = Q(R_M)$. It is instructive to study the variations of the average number of subcarriers required according to the variations of certain parameters. Let us fix $C = 200$ kb/s, $W = 250$ kHz, $A = 20,000$, and $\lambda = 0.01$ mobile per square meter.

According to the second characterisation of the Poisson process, the number of users in the ring k which is denoted by N_k follows a Poisson distribution with parameter $\lambda\pi(R_k^2 - R_{k-1}^2)$. Moreover, the random variables $(N_k,\ k \geq 1)$ are independent because the rings are disjoint. The number of subcarriers required is thus for each slot $N = \sum_k kN_k$. Therefore,

$$\mathbf{E}\left[N\right] = \lambda\pi \sum_k k(R_k^2 - R_{k-1}^2) \text{ and } \operatorname{Var}(N) = \lambda\pi \sum_k k^2(R_k^2 - R_{k-1}^2).$$

Such a random variable follows a distribution called compound Poisson distribution. We know how to calculate the Laplace transform easily but that does not give the explicit expression of the probability that N is equal to k. We can also numerically calculate its distribution by convolution of the distributions of each of the dilated Poisson random variables which composes it. Nevertheless, the size of the calculations is quickly prohibitive, we therefore resort to approximations. For a Poisson random variable of parameter λ, we know that when λ tends toward infinity, the distribution of this suitably normalised variable tends toward a Gaussian distribution. The disadvantages of this approach are twofold. On the one hand, we do not know what λ big signifies, that is to say, from which values of λ the approximation is valid. On the other hand, the approximation is, according to the values of λ sometimes pessimistic, and sometimes optimistic as shown in Figure 10.4.

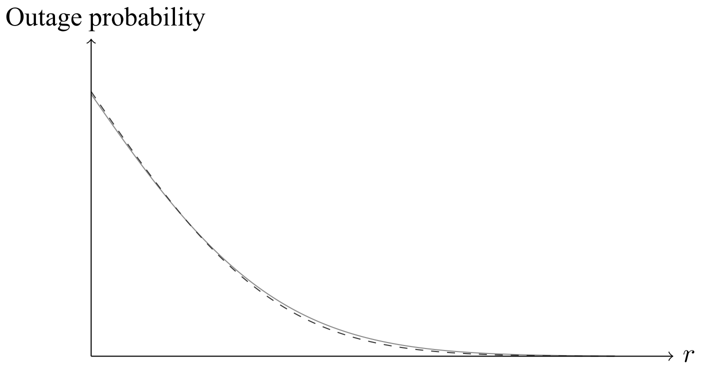

Figure 10.4. *Outage probability with respect to r. The solid line represents the exact value, in dotted line, the Gaussian approximation. Curve obtained for $\gamma = 3.5$ and $\lambda = 0.02$*

To address the first difficulty, one can use the previous results. By taking $Y_n^k = k$ for any n, we notice that the c^2 of Theorem 10.10 is equal to Var(N). We, therefore, have

$$\begin{aligned} \mathrm{d}_{\mathrm{TV}}\left(\mathrm{Dist.}(\frac{N-\mathbf{E}\,[N]}{c}),\,\mathcal{N}(0,\,1)\right) &\le \frac{\lambda\pi\sum_k k^3(R_k^2-R_{k-1}^2)}{(\lambda\pi\sum_k k^2(R_k^2-R_{k-1}^2))^{3/2}} \\ &= \frac{1}{\sqrt{\lambda}}\,\frac{\sum_k k^3(R_k^2-R_{k-1}^2)}{\sqrt{\pi}(\sum_k k^2(R_k^2-R_{k-1}^2))^{3/2}}. \end{aligned}$$

In other words, the convergence toward the Gaussian distribution is of the order of $1/\sqrt{\lambda}$ as expected. This theoretical bound is in fact very pessimistic as proved by Figure 10.4.

However, the advantage of this result is that it provides an explicit bound. Table 10.3 shows a high sensitivity of the average number of required subcarriers with respect to the variations of γ.

The same calculations show such a high sensitivity to variations in λ, the average number of active users per unit of area. However, these two parameters are particularly delicate to estimate, especially the coefficient γ, which largely depends on the environment (rural, urban, semi-urban, suburban, skyscrapers, etc.). It is, therefore, of crucial importance to have dimensioning that is robust to the variations of these parameters.

THEOREM 10.11.– *Let N be a Poisson process with intensity μ on E and Λ be a compact of E. Let $F\,:\,\mathfrak{N}_\Lambda \to \mathbf{R}$ such that*

$$D_xF(N_\Lambda) \le \beta,\ (\mu\otimes\mathbf{P}) -\ \textit{a.e. and} \int_E |D_xF(N_\Lambda)|^2\,\mathrm{d}\,\mu(x) \le \alpha^2,\ \mathbf{P} - \textit{a.e.}.$$

	γ			
η	2.5	3	3.5	4
0.1	2.10^3	334	98	38
0.01	9.10^5	10^5	3.10^3	56
0.001	5.10^6	5.10^5	10^5	555

Table 10.3. *Average number of subcarriers required depending on η (first column) and γ,*

For any $r > 0$, we have the following inequality

$$\mathbf{P}(F(N_\Lambda) - \mathbf{E}\left[F(N_\Lambda)\right] > r) \leq \exp\left(-\frac{r}{2\beta}\ln(1 + \frac{r\beta}{\alpha^2})\right).$$

EXAMPLE (Continuation of example 10.6).– In the case of interest, the maximum number of subcarriers required is limited, and it only depends on the characteristics of the network (transmitting power, bandwidth of the subcarrier, etc.). Therefore, $D_xN \leq N_M$ and if E is the cell, we can choose $\alpha^2 = N_M^2\mu(E) = \lambda\pi R_M^2 N_M^2$ in Theorem 10.11. We thus obtain

$$\mathbf{P}(N - \mathbf{E}\left[N\right] > r) \leq \exp\left(-\frac{r}{2N_M^2}\log(1 + \frac{r}{\lambda\pi R_M^2})\right).$$

With respect to the Gaussian approximation, in this formula we need not calculate the variance of N, just the average is sufficient.

γ	Exact	Gaussian	Concentration	Error
2.5	2.060	2.056	2.357	0.14
2.6	1.420	1.417	1.682	0.18
2.7	1.010	1.006	1.243	0.23
2.8	738	735	948	0.29
2.9	553	549	745	0.35
3	423	420	600	0.42
3.5	146	142	276	0.9
4	69	66	177	1.57

Table 10.4. *Dimensioning by the three methods for different values of γ. For all values of γ, $\lambda = 0.01$. The last column contains the relative over sizing between what is obtained by the concentration inequality and the exact calculation*

We observe that dimensioning by the Gaussian approximation is optimistic, which is absolutely forbidden. The over-dimensioning induced by the inequality of concentration is reasonable in most cases. Although it is clearly large, this can be seen as a guarantee against the inaccuracies in the measurements of λ and γ and against the epistemic errors, i.e. error in the model. If we consider a situation with different classes of users with different rates and integrate considerations about fading and shadowing to the model, exact calculations become impossible but the concentration inequality can still be easily established.

10.4. Stochastic analysis

We now establish the proofs of Theorems 10.9 and 10.11.

LEMMA 10.12 (Girsanov's theorem).– *Let N and N' be two point Poisson processes, with respective intensity μ and μ'. Let us assume that $\mu' \ll \mu$ and let us denote $p = \mathrm{d}\,\mu' / \mathrm{d}\,\mu$. Let Λ be a compact of E. Moreover, if p belongs to $L^1(\mu_\Lambda)$, then for every bounded function F, we have*

$$\mathbf{E}\left[F(N'_\Lambda)\right] = \mathbf{E}\left[F(N_\Lambda)\exp\left(\int \ln p \,\mathrm{d}\,N_\Lambda + \int_\Lambda (1-p)\,\mathrm{d}\,\mu\right)\right].$$

Proof. We verify this identity for the exponential functions F of the form $\exp(-\int f\,\mathrm{d}\,N)$ with f at compact support. On the basis of Definition 10.2

$$\begin{aligned}\mathbf{E}&\left[\exp(-\int f\,\mathrm{d}\,N_\Lambda)\exp\left(\int_\Lambda \ln p\,\mathrm{d}\,N_\Lambda + \int (1-p)\,\mathrm{d}\,\mu\right)\right]\\ &= \mathbf{E}\left[\exp(-\int (f-\ln p)\,\mathrm{d}\,N_\Lambda)\right]\exp(\int_\Lambda (1-p)\,\mathrm{d}\,\mu)\\ &= \exp(-\int (1-\exp(-f+\ln p))\,\mathrm{d}\,\mu + \int_\Lambda (1-p)\,\mathrm{d}\,\mu)\\ &= \exp(-\int_\Lambda (1-e^{-f})p\,\mathrm{d}\,\mu)\\ &= \mathbf{E}\left[F(N'_\Lambda)\right].\end{aligned}$$

As a result, the measures on $\mathfrak{N}_E$, $\mathbf{P}_{N'_\Lambda}$, and $Rd\mathbf{P}_{N_\Lambda}$ where

$$R = \exp\left(\int \ln p\,\mathrm{d}\,N_\Lambda + \int (1-p)\,\mathrm{d}\,\mu\right)$$

have the same Laplace transform. Therefore, they are equal and the result follows for any bounded function F. □

In what follows, for a configuration η

$$\eta + \delta_x = \begin{cases}\eta, \text{ if } x \in \eta,\\ \eta \cup \{x\}, \text{ if } x \notin \eta.\end{cases}$$

Similarly,

$$\eta - \delta_x = \begin{cases} \eta \backslash \{x\}, \text{ if } x \in \eta, \\ \eta, \text{ if } x \notin \eta. \end{cases}$$

As μ is assumed to be diffuse $\mathbf{E}\left[N(\{x\})\right] = \mu(\{x\}) = 0$. Therefore, for fixed x, almost surely, η does not contain x.

DEFINITION 10.6.– *Let N be a Poisson process with intensity μ. Let $F : \mathfrak{N}_E \longrightarrow \mathbf{R}$ be a measurable function such that $\mathbf{E}\left[F^2\right] < \infty$. We define $DomD$ as the set of square integrable random variables such that*

$$\mathbf{E}\left[\int_E |F(N + \delta_x) - F(N)|^2 \, \mathrm{d}\,\mu(x)\right] < \infty.$$

For $F \in$ Dom D, we set

$$D_x F(N) = F(N + \delta_x) - F(N).$$

EXAMPLE.– For example, for f deterministic belonging to $L^2(\mu)$, $F = \int f \, \mathrm{d}\, N$ belongs to Dom D and $D_x F = f(x)$ because

$$F(N + \delta_x) = \sum_{y \in N \cup \{x\}} f(y) = \sum_{y \in N} f(y) + f(x).$$

Similarly, if $F = \max_{y \in N} f(y)$ then

$$D_x F(N) = \begin{cases} 0 & \text{if } f(x) \le F(N), \\ f(x) - F & \text{if } f(x) > F(N). \end{cases}$$

In both cases, if f is bounded, DF is bounded too.

One of the essential formulas for the Poisson process is the following.

THEOREM 10.13.– *Let N be a Poisson process with intensity μ. For any random field $F\colon \mathfrak{N}_E \times E \to \mathbf{R}$ such that*

$$\mathbf{E}\left[\int_E |F(N, x)| \, \mathrm{d}\,\mu(x)\right] < \infty$$

then

$$\mathbf{E}\left[\int_E F(N,\,x)\,\mathrm{d}\,\mu(x)\right] = \mathbf{E}\left[\int_E F(N\backslash x,\,x)\,\mathrm{d}\,N(x)\right]. \qquad [10.5]$$

Proof. According to the first definition of the Poisson process, for f with compact support and Λ a compact E, for any $t > 0$,

$$\mathbf{E}\left[\exp(-\int (f + t\,\mathbf{1}_\Lambda)\,\mathrm{d}\,N)\right] = \exp(-\int_E 1 - e^{-f(x)-t\,\mathbf{1}_\Lambda(x)}\,\mathrm{d}\,\mu(x)).$$

According to the theorem of derivation under the summation sign, on one hand, we have

$$\frac{d}{dt}\mathbf{E}\left[\exp(-\int (f + t\,\mathbf{1}_\Lambda)\,\mathrm{d}\,N)\right]\Bigg|_{t=0} = -\mathbf{E}\left[e^{-\int f\,\mathrm{d}\,N}\int \mathbf{1}_\Lambda\,\mathrm{d}\,N\right],$$

and on the other hand,

$$\frac{d}{dt}\exp(-\int_E 1 - e^{-f(x)-t\,\mathbf{1}_\Lambda(x)}\,\mathrm{d}\,\mu(x))\Bigg|_{t=0}$$
$$= -\mathbf{E}\left[\int e^{-\int f\,\mathrm{d}\,N + f(x)}\,\mathbf{1}_\Lambda(x)\,\mathrm{d}\,\mu(x)\right].$$

As $\int f\,\mathrm{d}\,N - f(x) = \int f\,\mathrm{d}(N - \delta_x)$, [10.5] is true for functions of the form $\mathbf{1}_\Lambda\,e^{-\int f\,\mathrm{d}\,N}$. We admit that this is enough as far as the result is true for all the F functions such that both the members are well defined. □

Let μ be a Radon measure on a Polish space E and Λ be a compact of E. We introduce the Glauber-Poisson process, which is denoted by $\mathfrak{N}^\Lambda$, whose dynamics is as follows:

– $\mathfrak{N}^\Lambda(0) = \eta \in \mathfrak{N}_\Lambda$,

– Each atom of η has a life duration, independent of that of the other atoms, exponentially distributed with parameter 1.

– Atoms are born at moments following a Poisson process with intensity $\mu(\Lambda)$. On its appearance, each atom is localised independently from all the others according to $\mu/\mu(\Lambda)$. It is also assigned in an independent manner, a life duration exponentially distributed with parameter 1.

Algorithm 10.2. The first k transitions of a trajectory of a Glauber process associated with a Poisson process with intensity μ, on a domain Λ, of initial condition N

```
Data: $\mu$, $\Lambda$, $N$, $k$
Result: $T_1, \cdots, T_k$ = birth moments. $N_j = \mathfrak{N}^\Lambda(T_j)$, $j = 1, \cdots, k$
$T_0 \leftarrow 0$;
$N_0 \leftarrow N$;
for $n = 1$ to $k$ do
    $L_n \leftarrow$ drawing of a $\varepsilon(\mu(\Lambda) + N_{n-1}(\Lambda))$;
    $T_n \leftarrow T_{n-1} + L_n$;
    $r \leftarrow \mu(\Lambda)/(\mu(\Lambda) + N_{n-1}(\Lambda))$;
    $U \leftarrow$ drawing of a $U([0, 1])$;
    if $U \leq r$ then
        $X \leftarrow$ drawing of a random variable of distribution $\mu/\mu(\Lambda)$;
        $N_n \leftarrow N_{n-1} + \delta_X$;
    end
    if not $\kappa \leftarrow$ drawing of a $U(\{1, \cdots, N_{n-1}(\Lambda)\})$;
    $N_n \leftarrow N_{n-1} - \delta_{x_\kappa}$;
end
return $T_1, \cdots, T_k$
```

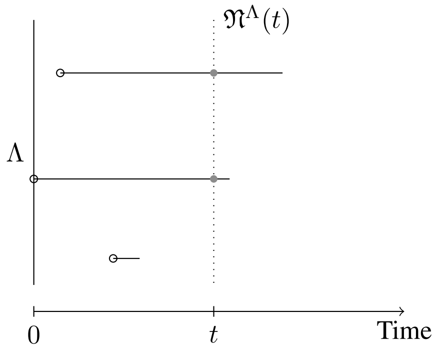

Figure 10.5. *Realisation of a trajectory of* $\mathfrak{N}^\Lambda$

At every instant, $\mathfrak{N}^\Lambda(t)$ is a configuration of E. We first observe that the total number of atoms of $\mathfrak{N}^\Lambda(t)$ follows exactly the same dynamics as the number of busy servers in a M/M/∞ queue with parameters $\mu(\Lambda)$ and 1.

THEOREM 10.14.– *Assume that $\mathfrak{N}^\Lambda(0)$ is a point Poisson process with intensity ν. At each instant t, the distribution of $\mathfrak{N}^\Lambda(t)$ is that of a Poisson process with intensity $e^{-t}\nu_\Lambda + (1 - e^{-t})\mu_\Lambda$ where ν_Λ is the restriction from ν to Λ. Particularly, if $\nu_\Lambda = \mu_\Lambda$,* *the distribution of* $\mathfrak{N}^\Lambda(t)$ *does not depend on* t *and is equal to* μ_Λ. *We denote* $\mathbf{E}_{\mu_\Lambda}[X]$ *as the expectation of a random variable* X *under this induced probability.*

Proof. For two disjoint parts A and B of Λ, by construction, the processes $\mathfrak{G}_A$ and $\mathfrak{G}_B$ are independent and follow the same dynamics as that of a M/M/∞ queue with respective parameters $(\mu(A),\, 1)$ and $(\mu(B),\, 1)$. The result follows from the properties of the M/M/∞ queue established in Example 10.3. □

As all the sojourn time are exponentially distributed, $\mathfrak{N}^\Lambda$ is a Markov process with values in $\mathfrak{N}_E$. Far from the idea of developing the general theory of Markov processes in the space of measures, we can study its infinitesimal generator and its semi group.

THEOREM 10.15.– *Let* Λ *be a compact of* E. *The infinitesimal generator of* $\mathfrak{N}^\Lambda$ *is given by*

$$-\mathfrak{L}_\Lambda F(N) = \int_\Lambda (F(N+\delta_x) - F(N))\,\mathrm{d}\,\mu(x) + \int (F(N-\delta_x) - F(N))\,\mathrm{d}\,N(x), \quad [10.6]$$

for F *bounded from* $\mathfrak{N}_\Lambda$ *into* $\mathbf{R}$.

Proof. We reason in the same way as that of the Markov process. At a time t, there may be a either a death or a birth. At the time of a departure, we choose the uniformly killed atom among the existing atoms. The death rate is thus $\eta(\Lambda)$ and every atom has a probability $\eta(\Lambda)^{-1}$ of being killed. Therefore, the transition η toward $\eta - \delta_x$ take place at rates of 1 for any $x \in \eta$. The birth rate is $\mu(\Lambda)$ and the position of the new atom is distributed according to the measure $\mu_\Lambda/\mu(\Lambda)$ so the transition η toward $\eta + \delta_x$ occurs at a rate $\mathrm{d}\,\mu_\Lambda(x)$ for each $x \in \Lambda$. From it, we deduce [10.6]. □

THEOREM 10.16.– *The semi-group* $\mathfrak{P}^\Lambda$ *is ergodic. Moreover,* $\mathfrak{L}_\Lambda$ *is invertible from* L^2_0 *in* L^2_0 *where* L^2_0 *is the subspace of* L^2 *of the random variables with null expectation and we have*

$$\mathfrak{L}_\Lambda^{-1} F = \int_0^\infty \mathfrak{P}_t^\Lambda F \,\mathrm{d}\,t. \quad [10.7]$$

For any $x \in E$ *and any* $t > 0$,

$$D_x \mathfrak{P}_t^\Lambda F = e^{-t} \mathfrak{P}_t^\Lambda D_x F. \quad [10.8]$$

In addition,

$$\mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda |D_x(\mathfrak{L}_\Lambda^{-1}F(N))|^2 \,\mathrm{d}\,\mu(x)\right] \le \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda |D_xF(N)|^2 \,\mathrm{d}\,\mu(x)\right]. \quad [10.9]$$

Proof. Denote $(x_1, \cdots, x_n)$ the atoms of $\mathfrak{N}^\Lambda(0)$ and $(Y_1, \cdots, Y_n)$ some independent random variables exponentially distributed with parameter 1. We set

$$\mathfrak{N}^\Lambda(0)[t] = \sum_{i=1} \mathbf{1}_{\{Y_i \ge t\}}\,\delta_{x_i},$$

the measure consisting of the atoms of $\mathfrak{N}^\Lambda(0)$ surviving at time t. The distribution of $\mathfrak{N}^\Lambda(t)$ is that of the independent sum of a Poisson process with intensity $(1-e^{-t})\mu_\Lambda$ and of $\mathfrak{N}^\Lambda(0)[t]$. According to Lemma 10.12, we know that for any $F \in L^1$

$$\mathbf{E}_{(1-e^{-t})\mu_\Lambda}\left[F(N_\Lambda)\right] = \mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)\exp(\ln(1-e^{-t})N(\Lambda) + e^{-t}\mu(\Lambda))\right].$$

Therefore, for any bounded function F and any $\eta \in \mathfrak{N}_\Lambda$, we have the following identity

$$\begin{aligned}\mathfrak{P}_t^\Lambda F(\eta) &= \mathbf{E}\left[F(\mathfrak{N}^\Lambda(t)) \,|\, \mathfrak{N}^\Lambda(0) = \eta\right] \\ &= \mathbf{E}\left[F(\eta[t] + N_\Lambda)\exp(\ln(1-e^{-t})N(\Lambda) + e^{-t}\mu(\Lambda))\right]. \quad [10.10]\end{aligned}$$

Set

$$R(t) = \exp(\ln(1-e^{-t})N(\Lambda) + e^{-t}\mu(\Lambda)).$$

On the one hand, we have $R(t) \le e^{\mu(\Lambda)}$ and on the other hand, according to definition 10.2, $\mathbf{E}\left[R(t)\right] = 1$, and this for any $t \ge 0$. As $\mathfrak{N}^\Lambda(0)$ has a finite number of atoms, $\mathfrak{N}^\Lambda(0)[t]$ almost surely tends toward the zero measure when t tends toward infinity. By dominated convergence, we deduce that

$$\mathfrak{P}_t^\Lambda F(\eta) \xrightarrow{t\to\infty} \mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)\right]$$

that is to say, $\mathfrak{P}^\Lambda$ is ergodic. The property [10.7] is a well-known relation between the semi-group and infinitesimal generator. Formally, without worrying about the integral convergence, we have

$$\begin{aligned}\mathfrak{L}_\Lambda\Big(\int_0^\infty \mathfrak{P}_t^\Lambda F\,\mathrm{d}\,t\Big) &= \int_0^\infty \mathfrak{L}_\Lambda\mathfrak{P}_t^\Lambda F\,\mathrm{d}\,t \\ &= -\int_0^\infty \frac{d}{dt}\mathfrak{P}_t^\Lambda F\,\mathrm{d}\,t \\ &= F - \mathbf{E}\left[F\right] = F,\end{aligned}$$

according to ergodicity of $\mathfrak{P}^\Lambda$ and as F is centered. □

Let $\mathfrak{N}^\Lambda(t,\, N_\Lambda)$ denote the value of $\mathfrak{N}^\Lambda(t)$ when the initial condition is N_Λ. We can write

$$D_x \mathfrak{P}^\Lambda F(t) = \mathbf{E}\left[\mathfrak{N}^\Lambda(t,\, N_\Lambda + \delta_x)\right] - \mathbf{E}\left[\mathfrak{N}^\Lambda(t,\, N_\Lambda)\right].$$

Let Y_x be the life duration of the atom located in x. If $Y_x \geq t$ then the atom is still alive at t, thus $\mathfrak{N}^\Lambda(t,\, N_\Lambda + \delta_x) = \mathfrak{N}^\Lambda(t,\, N_\Lambda) + \delta_x$. If the atom is already dead at t then $\mathfrak{N}^\Lambda(t,\, N_\Lambda + \delta_x) = \mathfrak{N}^\Lambda(t,\, N_\Lambda)$. As Y_x is by construction, independent of N_Λ and $\mathfrak{N}^\Lambda$, it is legitimate to write

$$\begin{aligned}
\mathbf{E}&\left[F(\mathfrak{N}^\Lambda(t,\, N_\Lambda + \delta_x) \,|\, N_\Lambda\right] - \mathbf{E}\left[F(\mathfrak{N}^\Lambda(t,\, N_\Lambda) \,|\, N_\Lambda\right] \\
&= \mathbf{E}\left[\mathbf{1}_{\{Y_x \geq t\}}(F(\mathfrak{N}^\Lambda(t,\, N_\Lambda + \delta_x)) - F(\mathfrak{N}^\Lambda(t,\, N_\Lambda)) \,|\, N_\Lambda\right] \\
&\quad + \mathbf{E}\left[\mathbf{1}_{\{Y_x \leq t\}}(F(\mathfrak{N}^\Lambda(t,\, N_\Lambda)) - F(\mathfrak{N}^\Lambda(t,\, N_\Lambda))) \,|\, N_\Lambda\right], \\
&= e^{-t}\mathbf{E}\left[D_x F(\mathfrak{N}^\Lambda(t,\, N_\Lambda))\right],
\end{aligned}$$

hence we have the result. According to the representation [10.10] and Jensen's inequality, we see that

$$\left|\mathfrak{P}_t^\Lambda F\right|^2 \leq \mathfrak{P}_t^\Lambda F^2. \qquad [10.11]$$

Therefore,

$$\begin{aligned}
\int_\Lambda &|D_x(\mathfrak{L}_\Lambda^{-1} F(N_\Lambda))|^2 \, \mathrm{d}\,\mu(x) \\
&= \int_\Lambda |D_x \int_0^\infty \mathfrak{P}_t^\Lambda F(N_\Lambda) \, \mathrm{d}\,t|^2 \, \mathrm{d}\,\mu(x) \\
&= \int_\Lambda |\int_0^\infty e^{-t} \mathfrak{P}_t^\Lambda D_x F(N_\Lambda) \, \mathrm{d}\,t|^2 \, \mathrm{d}\,\mu(x) \\
&\leq \int_\Lambda \int_0^\infty e^{-t} |\mathfrak{P}_t^\Lambda D_x F(N_\Lambda)|^2 \, \mathrm{d}\,t \, \mathrm{d}\,\mu(x) \\
&\leq \int_\Lambda \int_0^\infty e^{-t} \mathfrak{P}_t^\Lambda |D_x F(N_\Lambda)|^2 \, \mathrm{d}\,t \, \mathrm{d}\,\mu(x) \\
&= \int_\Lambda \int_0^\infty e^{-t} \mathbf{E}\left[|D_x F|^2(\mathfrak{N}^\Lambda(t)) \,|\, \mathfrak{N}^\Lambda(0) = N_\Lambda\right] \mathrm{d}\,t \, \mathrm{d}\,\mu(x),
\end{aligned}$$

where we have successively used equations [10.7] and [10.8], Jensen's inequality and [10.11]. As $\mathfrak{N}^\Lambda(t)$ has the same distribution as $\mathfrak{N}^\Lambda(0)$ if this one is chosen as a Poisson process with μ_Λ intensity, when we take expectations of each side, we obtain the following identity

$$\begin{aligned}
&\mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda |D_x(\mathfrak{L}_\Lambda^{-1}F(N_\Lambda))|^2 \,\mathrm{d}\,\mu(x)\right] \\
&\qquad\qquad = \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda \int_0^\infty e^{-t}|D_xF|^2(\mathfrak{N}^\Lambda(t))\,\mathrm{d}\,t\,\mathrm{d}\,\mu(x)\right] \\
&\qquad\qquad = \int_\Lambda \int_0^\infty e^{-t}\mathbf{E}_{\mu_\Lambda}\left[|D_xF|^2(N_\Lambda)\right]\mathrm{d}\,t\,\mathrm{d}\,\mu(x) \\
&\qquad\qquad = \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda |D_xF(N_\Lambda)|^2\,\mathrm{d}\,\mu(x)\right].
\end{aligned}$$

Hence, we have the result.

THEOREM 10.17.– *Let F and G be two functions belonging to Dom D. The following identity is satisfied*

$$\mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda D_xF(N_\Lambda)\,D_xG(N_\Lambda)\,\mathrm{d}\,\mu(x)\right] = \mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)\,\mathfrak{L}_\Lambda G(N_\Lambda)\right].$$

In particular, if G is centered

$$\mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)G(N_\Lambda)\right] = \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda D_xF(N_\Lambda)\,D_x(\mathfrak{L}_\Lambda^{-1}G)(N_\Lambda)\,\mathrm{d}\,\mu(x)\right]. \qquad [10.12]$$

Proof. Let F and G belong to Dom D, according to [10.5] twice and the definition of $\mathfrak{L}_\Lambda$, we have

$$\begin{aligned}
&\mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda D_xF(N_\Lambda)\,D_xG(N_\Lambda)\,\mathrm{d}\,\mu(x)\right] \\
&= \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda (F(N_\Lambda) - F(N_\Lambda - \delta_x))(G(N_\Lambda) - G(N_\Lambda - \delta_x))\,\mathrm{d}\,N_\Lambda(x)\right] \\
&= \mathbf{E}_{\mu_\Lambda}\left[G(N_\Lambda)\mathfrak{L}_\Lambda F\right] + \mathbf{E}_{\mu_\Lambda}\left[\int G(N_\Lambda)(F(N_\Lambda + \delta_x) - F(N_\Lambda))\,\mathrm{d}\,\mu(x)\right] \\
&\qquad - \mathbf{E}_{\mu_\Lambda}\left[\int G(N_\Lambda - \delta_x)(F(N_\Lambda) - F(N_\Lambda - \delta_x))\,\mathrm{d}\,N_\Lambda(x)\right] \\
&\qquad\qquad = \mathbf{E}_{\mu_\Lambda}\left[G(N_\Lambda)\mathfrak{L}_\Lambda F(N_\Lambda)\right].
\end{aligned}$$

The result follows. □

Proof of Theorem 10.11. Let Λ be a compact of E, a bounded function F of zero expectation. According to Theorem 10.17, we can write the following identities

$$\begin{aligned}
&\mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)e^{\theta F(N_\Lambda)}\right]\\
&\qquad = \mathbf{E}_{\mu_\Lambda}\left[\int D_x(\mathfrak{L}_\Lambda^{-1}F(N_\Lambda))\, D_x(e^{\theta F(N_\Lambda)})\,\mathrm{d}\,\mu(x)\right]\\
&\qquad = \mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda D_x(\mathfrak{L}_\Lambda^{-1}F(N_\Lambda))(e^{\theta D_x F(N_\Lambda)}-1)e^{\theta F(N_\Lambda)}\,\mathrm{d}\,\mu(x)\right].
\end{aligned}$$

The function $(x \mapsto (e^x-1)/x)$ is continuously increasing on $\mathbf{R}$; therefore, we have

$$\begin{aligned}
&\mathbf{E}_{\mu_\Lambda}\left[F(N_\Lambda)e^{\theta F(N_\Lambda)}\right]\\
&= \theta\,\mathbf{E}_{\mu_\Lambda}\left[\int_\Lambda D_x(\mathfrak{L}_\Lambda^{-1}F(N_\Lambda))\, D_x F(N_\Lambda)\,\frac{e^{\theta D_x F(N_\Lambda)}-1}{\theta D_x F(N_\Lambda)}\, e^{\theta F(N_\Lambda)}\,\mathrm{d}\,\mu(x)\right]\\
&\le \theta\alpha^2\,\frac{e^{\theta\beta}-1}{\theta\beta}\,\mathbf{E}_{\mu_\Lambda}\left[e^{\theta F(N_\Lambda)}\right].
\end{aligned}$$

This implies that

$$\frac{d}{d\theta}\log \mathbf{E}_{\mu_\Lambda}\left[e^{\theta F(N_\Lambda)}\right] \le \alpha^2\frac{e^{\theta\beta}-1}{\beta}.$$

Therefore,

$$\mathbf{E}_{\mu_\Lambda}\left[e^{\theta F(N_\Lambda)}\right] \le \exp\left(\frac{\alpha^2}{\beta}\int_0^\theta (e^{\beta u}-1)\,\mathrm{d}\,u\right).$$

For $x>0$, for any $\theta>0$,

$$\begin{aligned}
\mathbf{P}(F(N_\Lambda)>x) &= \mathbf{P}(e^{\theta F(N_\Lambda)} > e^{\theta x})\\
&\le e^{-\theta x}\mathbf{E}\left[e^{\theta F(N_\Lambda)}\right] \le e^{-\theta x}\exp\left(\frac{\alpha^2}{\beta}\int_0^\theta (e^{\beta u}-1)\,\mathrm{d}\,u\right). \qquad [10.13]
\end{aligned}$$

This result is true for any θ, so we can optimize with respect to θ. At fixed x, we search the value of θ which cancels the derivative of the right-hand-side with respect to θ. Plugging this value into [10.3], we can obtain the result. □

Proof of Theorem 10.9. We can always assume that $F \in \operatorname{Dom} D$ is of null expectancy. We, respectively, note ϕ and ϕ_c , as the cumulative distribution function (the complementary cumulative distribution function, respectively) of a reduced centered Gaussian random variable, that is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \,\mathrm{d}\, u, \; \phi_c(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-u^2/2} \,\mathrm{d}\, u.$$

For $r > 0$, let us consider ψ_r, as solution of the differential equation

$$x.y(x) - y'(x) = \phi_c(r) - \mathbf{1}_{[r,\,\infty[}(x), \text{ for all } x \in \mathbf{R}.$$

We can choose

$$\psi_r(x) = \begin{cases} \sqrt{2\pi} e^{x^2/2} \phi(x)\phi_c(r) & \text{if } x \le r, \\ \sqrt{2\pi} e^{x^2/2} \phi_c(x)\phi(r) & \text{if } x \ge r. \end{cases}$$

Note that

$$|\psi_r'(x)| \le \phi(r) \text{ and } |\psi_r''(x)| \le 2. \tag{[10.14]}$$

We can then use the Stein method. It is sufficient to note that

$$\begin{aligned} \phi_c(r) - \mathbf{P}(F \ge r) &= \mathbf{E}_{\mu_\Lambda}\left[\mathbf{1}_{[r,\,\infty[}(F) - \phi_c(r)\right] \\ &= \mathbf{E}_{\mu_\Lambda}\left[F\psi_r(F) - \psi_r'(F)\right] \\ &= \mathbf{E}_{\mu_\Lambda}\left[\int D_x\psi_r(F) D_x \mathfrak{L}_\Lambda^{-1} F \,\mathrm{d}\,\mu(x)\right] - \mathbf{E}_{\mu_\Lambda}\left[\psi_r'(F)\right], \end{aligned} \tag{[10.15]}$$

according to [10.12]. According to the Taylor's formula with integral remainder, we have

$$\begin{aligned} D_x\psi_r(F) = \psi_r(F(N_\Lambda + \delta_x)) - \psi_r(F(N_\Lambda)) = \psi_r'(F(N_\Lambda)) D_x F(N_\Lambda) \\ + (D_x F(N_\Lambda))^2 \int_0^1 \psi_r''(F(N_\Lambda) + (1-u)F(N_\Lambda + \delta_x))(1-u) \,\mathrm{d}\, u. \end{aligned} \tag{[10.16]}$$

Plugging [10.16] into [10.15] and using [10.9] and [10.14], we get

$$\begin{aligned} |\phi_c(r) - \mathbf{P}(F \ge r)| \le \phi(r) \mathbf{E}_{\mu_\Lambda}\left[|1 - \int D_x F \, D_x \mathfrak{L}_\Lambda^{-1} F \,\mathrm{d}\,\mu(x)|\right] \\ + 2\,\mathbf{E}_{\mu_\Lambda}\left[\int_0^1 (1-u) \,\mathrm{d}\, u \int |D_x F(N_\Lambda)|^2 |D_x \mathfrak{L}_\Lambda^{-1} F| \,\mathrm{d}\,\mu(x)\right]. \end{aligned}$$

The result follows. □

10.5. Problems

EXERCISE 23.– We continue the study of OFDMA protocol. Now, we consider the Rayleigh fading. This signifies that the number of subcarriers claimed by a user at position x becomes

$$N(x) = \min(N_M, \left\lceil \frac{C}{W \log_2(1 + \frac{K\,F_x}{\|x\|^\gamma})} \right\rceil) \text{ if } KF_x\|x\|^{-\gamma} > \text{SNR}_{min},$$

where F_x is the coefficient of Rayleigh fading. From the standpoint of modeling, this signifies that to each atom of the Poisson process we add a mark representing the Rayleigh fading of that user. We generally assume that

- F_x and F_y are random independent variables, if x and y are different,
- for any $x \in \mathbf{R}^2$, F_x follows an exponential distribution with parameter 1.

The model is now therefore a marked point process where the measures are of the form

$$\tilde{\xi} = \sum_{x \in \xi} \delta_{x,\,m}$$

with $\mathrm{d}\,\mathbf{P}_M(m) = \exp(-m)\,\mathbf{1}_{\mathbf{R}^+}(m)\,\mathrm{d}\,m$.

1) Give the average number of full sub-carriers required for every moment in the cell under integral form. Use the same decomposition principle as before to calculate this integral explicitly. We recall that the Gamma function is defined by

$$\Gamma(z) = \int_0^\infty e^{-m} m^{z-1}\,\mathrm{d}\,z.$$

2) For $k = 1, \cdots, N_M$, what is the distribution of the number of users who require k sub-carriers at a given time?

3) Check this results on the simulations.

4) Calculate through simulation the probability of outage for values of S going from $S_{\min}$ to $S_{\max}$ with a step of 10.

5) What is the value of the inequality of concentration in this new model ?

6) We fix a threshold of loss equal to 0.01, calculate the number of resources necessary to obtain a loss below this threshold by simulation and using the limit obtained by the inequality of concentration.

7) Calculate the probability of outage for λ, $\lambda + 10\%$, $\lambda + 20\%$ through simulation. Compare with the limit of concentration obtained for λ.

8) Same question by varying γ of 10%, 20%.

9) Your conclusions.

Numerical values

C	200 kb/s
W	250 kHz
K	10^6
γ	2.8
R	300 m
λ	0.01 m^{-2}
SINR_{min}	0.1
p	0.01
S_{min}	30
S_{max}	100

EXERCISE 24.– We consider X as the process representing the number of busy servers in a $\text{M}_\lambda/\text{M}_\mu/\infty$ queue. We say that $f : \mathbf{N} \to \mathbf{R}$ is c-Lipschitz if

$$f(n+1) - f(n) \le c, \quad \text{for any } n.$$

1) Express $X(t)$ under an integral form with respect to a Poisson process on $\mathbf{R}^+ \times \mathbf{R}^+$.

We introduce D as the operator of difference associated with the Poisson process. Let f be c-Lipschitz.

2) Show that for any $(s,\, z) \in \mathbf{R}^+ \times \mathbf{R}^+$,

$$|D_{s,z} f(X(t))| \le c\, \mathbf{1}_{[0,\, t]}(s)$$

and that

$$\int |D_{s,z} f(X(t))|^2 \lambda \,\mathrm{d}\, s \,\mathrm{d}\, \mathbf{P}_\sigma(z) \le \lambda c^2 t.$$

3) Deduce a concentration inequality for $f(X(t))$.

10.6. Notes and comments

The basic references for this chapter are [LAC 09a, BAC 09b]. For many results on the interferences in a Poissonian framework, one can consult [EDT 08]. The results of stochastic analysis are inspired by [HOU 02, WU 00]. The introduction of the Glauber's dynamics as an Ornstein-Uhlenbeck process is new. Another approach of Malliavin calculus for the Poisson processes can be found in [NUA 95, MIC 09].

Epitome

– A Poisson point process allows us to represent the mobiles at a given time or access points in a mathematically usable manner.

– Campbell's formula enables us to easily calculate the functional expectations of a Poisson process. It enables us to link to the model known as fluid model.

– In the same way as in dimension 1, the superposition of two independent Poisson processes is a Poisson process.

– The theorem of displacement stipulates that when we move the atoms of a Poisson process independently of each other with the same statistics, the result is still a Poisson process whose intensity we know to calculate.

– The stochastic analysis allows us to establish the inequality of concentration. This identity is useful to bound the overshot probabilities and thus define robust dimensioning.

Appendix A

Mathematical Toolbox

Mathematical concepts and theorems have their existence and their own interests. As regards modeling, mathematics mainly become a toolbox to solve practical problems. To make good food, it still necessary to know the ingredients available. We review in this chapter the mathematical theorems used in the rest of this book.

A.1. Probability spaces and processes

A.1.1. *Countable spaces*

The concept of countability plays an important role in probability if only because the property of measurability is stable only by countable union. It is therefore interesting to clarify some results related to this concept.

DEFINITION A.1.– *A set E is a finite cardinal with n elements if there exists a bijection from E into the set $\{1, \cdots, n\}$.*

DEFINITION A.2.– *A set E is called countable (or countably infinite) if there exists a bijection from E into $\mathbf{N}$, the set of natural integers.*

EXAMPLE.– $2\mathbf{N}$, the set of integers is countable. $\mathcal{P}(\mathbf{N})$, set of subsets of $\mathbf{N}$, is not since it can be bijectively mapped to the set of real numbers $\mathbf{R}$. Indeed, for a subset A of $\mathbf{N}$, we can define a sequence $(\chi_A(n),\ n \geq 0)$ by

$$\chi_A(n) = \begin{cases} 1 \text{ if } n \in A, \\ 0 \text{ otherwise.} \end{cases}$$

Then, with this sequence, we can associate the real number x_A defined by $x_A = \sum_{n\geq 1} 2^{-n}\chi_A(n)$. In the reverse direction, to a real number of $[0,\ 1]$, we can associate

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.

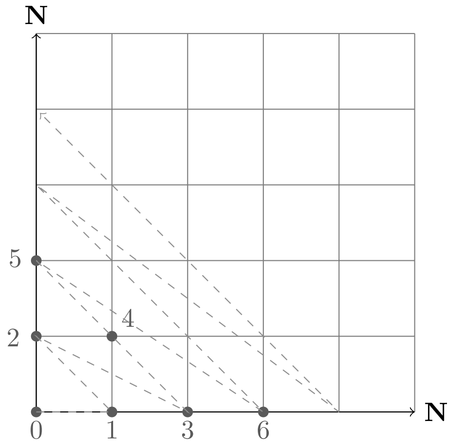

Figure A.1. *Bijection from* $\mathbf{N} \times \mathbf{N}$ *to* $\mathbf{N}$

its proper dyadic development, that is $x = \sum_{n \geq 1} x_n 2^{-n}$ with $x_n \in \{0,\, 1\}$ for any n. This defines a set by taking

$$A = \bigcup_{n\,:\,x_n=1} \{n\}.$$

Therefore, there exists a bijection from $\mathcal{P}(\mathbf{N})$ into $\{0,1\}^{\mathbf{N}}$ and then $\mathcal{P}(\mathbf{N})$ is not countable.

THEOREM A.1.– *If X and Y are countable then $X \times Y$ and $X \cup Y$ are countable.*

These two properties are based on the main result which states that there exists a bijection from $\mathbf{N} \times \mathbf{N}$ into $\mathbf{N}$. The bijection is constructed as shown in Figure A.1. For example, the element $(2, 0)$ is sent to 3 and the element $(0, 2)$ is sent to 5.

COROLLARY A.2.– *The set of relative integers* $\mathbf{Z}$ *is countable.*

The following theorem is far from being trivial.

THEOREM A.3.– *If the set X can be embedded in the set Y and Y can be embedded in X, then there exists a bijection from X into Y.*

COROLLARY A.4.– *The set of rational numbers* $\mathbf{Q}$ *is countable.*

Proof. There exists an injection from $\mathbf{N}$ in $\mathbf{Q}$ and by construction of $\mathbf{Q}$, there exists an injection of $\mathbf{Q}$ in $\mathbf{Z} \times \mathbf{Z}$, which according to the above theorem is in bijection with $\mathbf{N}$. Therefore, $\mathbf{Q}$ is in bijection with $\mathbf{N}$. □

NOTE.– Most of the processes we have to deal with take their values in at most countable spaces. Owing to etymology, this means that we can number the elements of these sets. Let E be an at most countable set and $i_x \in \mathbf{N}$ be the index corresponding to $x \in E$. Conversely, x_i the ith element of E is the element x of E such that $i_x = i$. When E is a discrete subset of $\mathbf{R}$, the numbering may be chosen as the canonical order. It often becomes a little tricky when, for example E is a product space as in the case of the queue MMPP/M/S/S, but this difficulty is more relevant to data representation for computer calculations rather than to mathematics itself.

EXAMPLE.– Let $E = \{0,\, 1\} \times \{a,\, b,\, c\}$. This set contains six elements which can be ordered in the lexicographic order

$$(0,\, a) \prec (0,\, b) \prec (0,\, c) \prec (1,\, a) \prec (1,\, b) \prec (1,\, c).$$

This induces $i_{(0,\, c)} = 3$ and $x_4 = (1,\, a)$.

A function from E to $\mathbf{R}$ is characterized by its "graph" $((x,\, f(x)),\, x \in E)$ but as E is countable, we can consider this set as the column vector whose i_xth component is $f(x)$. A measure on a countable space is characterized by its value on singletons therefore identified as a vector $\pi = (\pi(i_x),\, x \in E)$. This vector is usually written as a line vector since measures and functions are in duality: the integral of f with respect to the measure π is written as

$$\int f \,\mathrm{d}\pi = \sum_{x \in E} f(x)\pi(x) = (\pi(x_1),\, \cdots,\, \pi(x_n), \cdots) \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \\ \vdots \end{pmatrix},$$

where the row vector and column vector are multiplied according to the rules of matrix multiplication. It is sometimes convenient to think in terms of functions, particularly for mathematical proofs, sometimes more convenient to use the vector notation, especially for computations. When we handle functions, it is more natural to denote the states by x, y, and so on. When one refers to vectors, it is customary to number the states by i, j, and so on.

A.1.2. *Polish spaces*

Polish spaces are a class of topological spaces general enough to make probability theory rigorous without superfluous difficulty.

DEFINITION A.3.– *A Polish space E is a space with a distance d which satisfies the following two properties:*

– It is complete: every Cauchy sequence converges.

– It is separable: there exists a sequence $(x_n,\, n \geq 1)$ of elements of E dense in E, that is to say, such that for any $x \in E$, for any $\epsilon > 0$, there exists some x_n such that $d(x_n,\, x) < \epsilon$.

EXAMPLE.– The spaces $\mathbf{R}$, $\mathbf{R}^n$, $\mathbf{C}^n$ are Polish spaces for the common distance. Countable dense families are $\mathbf{Q}$, $(\mathbf{Q})^n$, $(\mathbf{Q} + i\mathbf{Q})^n$.

EXAMPLE.– More generally, we shall have to handle infinite sequences of real numbers. When the size of a "vector" becomes infinite, it creates problems of summability or "boundedness". If E is a countable space, we often consider the sub-spaces of the set of real sequences indexed by E, set denoted by $\mathbf{R}^E$. The two subsets of $\mathbf{R}^E$ which are the most important for us, are $l^\infty(E)$ and $l^2(E)$.

DEFINITION A.4.– *Let E be countable. The set $l^\infty(E)$ is the set of real valued sequences indexed by E which are bounded, that is*

$$u = (u(x),\, x \in E) \in l^\infty(E) \iff \sup_{x \in E} |u(x)| < \infty.$$

The norm on $l^\infty(E)$ is denoted by $\|u\|_\infty = \sup_{x \in E} |u(x)|$. The space $(l^\infty(E),\, \|\,\|_\infty)$ is a complete normed vector space, hence what is called a Banach space. Let ϕ be a bijection between $\mathbf{N}$ and E and $E_n = \phi(\{0,\, \cdots,\, n\})$. The set $\bigcup_{n \geq 0} \mathbf{Q}^{E_n}$ is countable and dense, therefore $(l^\infty(E),\, \|\,\|_\infty)$ is Polish.

DEFINITION A.5.– *Let E be a countable set and π a measure on E. The set $l^2(E,\, \pi)$ is the set of real-valued sequences indexed by E which are square integrable for π, that is*

$$u = (u(x),\, x \in E) \in l^2(E) \iff \sum_{x \in E} |u(x)|^2 \pi(x) < \infty.$$

The norm on $l^2(E,\, \pi)$ is denoted by $\|u\|_2 = (\sum_{x \in E} |u(x)|^2 \pi(x))^{1/2}$. The space $(l^2(E,\, \pi),\, \|\,\|_2)$ is a vector space with a scalar product

$$\langle u,\, v\rangle_{l^2(E,\, \pi)} = \sum_{x \in E} u(x)v(x)\pi(x),$$

which generalizes the scalar product of vectors in $\mathbf{R}^n$. In addition, $(l^2(E),\, \|\,\|_2)$ is complete for this scalar product, so it is an Hilbert space. The set $\bigcup_{n \geq 0} \mathbf{Q}^{E_n}$ is dense in $(l^2(E,\, \pi),\, \|\,\|_2)$, hence $(l^2(E,\, \pi),\, \|\,\|_2)$ is Polish.

EXAMPLE.– The space of continuous functions on $[0,\, T]$ with values in $\mathbf{R}$ is also a Polish space for the distance induced by the uniform norm

$$d(f,\, g) = \sup_{0 \leq t \leq T} |f(t) - g(t)|.$$

It is well known that this space is complete. According to the Weierstrass theorem, polynomials with rational coefficients form a countable dense family. Indeed, the set of polynomials with coefficients of degree at most k is in bijection with $\mathbf{Q}^{k+1}$ therefore the union of these sets is a countable union of countable sets, so it is countable.

EXAMPLE.– The space $D([0, T], \mathbf{R})$ of right-continuous with left-limits functions (rcll for short) is usually equipped with the distance

$$d(f, g) = \inf_{\phi \in \mathfrak{H}_T} \sup_{0 \leq t \leq T} |f(t) - g(\phi(t))|,$$

where $\mathfrak{H}_T$ is the set of homeomorphisms (continuous bijections with a continuous inverse) of $[0, T]$ into itself. This space is complete and the family of polynomials with rational coefficients is still dense.

The last example is that of configurations on a Polish space E.

DEFINITION A.6.– *A configuration on E is a locally finite set of points of E. Locally finite means that there is a finite number of points in any bounded subset of E. We denote by $\mathfrak{N}_E$ the set of configurations on E.*

The set $\{(n, 0),\ n \in \mathbf{N}^*\}$ is a configuration. But, the set $\{(1/n, 0),\ n \in \mathbf{N}^*\}$ is not a configuration because there is an infinite number of points in the interval $[0, 1]$ (see Figure A.1).

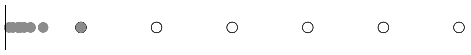

Figure A.2. *Example and counter-example of configurations*

We denote ξ as the generic element of the set of configurations. Each ξ is a set, that is $\xi = \{x_1, x_2, \cdots\}$, but we also identify it with a point measure: $\xi = \sum_{x \in \xi} \delta_x$. Depending on the context, one or the other representation is the most convenient. The empty configuration denoted by $\varnothing$ corresponds to the zero measure, this should not be confused with the Dirac measure at 0. Since any configuration has at most a countable numbers of elements, we can index them by the integers but there is no preferential order: x_1 can represent any of the atoms of ξ. Therefore, it is best to keep the set notation as often as possible. For $f : E \to \mathbf{R}$, we set (regardless of the convergence of the series for the time being)

$$\int f \, \mathrm{d}\xi = \sum_{x \in \xi} f(x).$$

For A a set of E, $\xi(A)$ is the number of points in $\xi \cap A$. We also denote $|\xi| = \xi(E)$ as the total number of points in ξ. It may happen that $|\xi|$ is infinite. Finally, ξ_B denotes the restriction of ξ to B: $\xi_B(A) = \xi(A \cap B)$.

DEFINITION A.7.– *Let E be a Polish space, a point process N on E is a random variable with values in the set $\mathfrak{N}_E$ of configurations on E.*

DEFINITION A.8.– *Let E be a Polish space and $\mathfrak{N}_E$ be the set of configurations on E. The configurations sequence $(\xi_n,\, n \geq 1)$ converges to the configuration ξ in a vague sense if for any continuous function f from E into $\mathbf{R}$ with compact support,*

$$\int f \,\mathrm{d}\,\xi_n \xrightarrow{n\to+\infty} \int f \,\mathrm{d}\,\xi.$$

This almost means that the positions of the atoms of ξ_n tend to those of atoms of ξ but some "mass" may escape to infinity

$$\delta_n \xrightarrow{n\to+\infty} \varnothing \text{ because for } f \text{ with compact support } f(n) \xrightarrow{n\to+\infty} 0.$$

THEOREM A.5.– *The space $\mathfrak{N}_E$ equipped with the vague convergence is Polish.*

In order to construct the distance on $\mathfrak{N}_E$, we consider a dense sequence $(x_n,\, n \geq 1)$ of elements in E. We denote $B_{n,q} = \{x \in E,\, d_E(x,\, x_n) < 1/q\}$ the open ball centered at x_n with radius $1/q$. The countable family of functions $(f_{n,\,q} = \mathbf{1}_{B_{n,\,q}},\, n \geq 1,\, q \geq 1)$ generates the σ-field $\mathcal{B}(E)$. For two configurations ω and η of $\mathfrak{N}_E$, we set

$$d(\omega,\, \eta) = \sum_{n\geq 1}\sum_{q\geq 1} 2^{-(n+q)}\zeta\Big(\int f_{n,\,q}(\mathrm{d}\,\omega - \mathrm{d}\,\eta)\Big),$$

where $\zeta(x) = |x|/(1+|x|)$. This distance induces the same topology as the vague topology: convergent sequences are the same in both cases. To construct a dense countable family, we can mimic what we did for continuous functions. Since E is Polish, there exists a dense countable family $\mathcal{Q} = (x_n,\, n \in \mathbf{N})$. The set of finite simple point measures whose atoms belong to $\mathcal{Q}$ plays the same role as polynomials with rational coefficients.

A.1.3. *Stochastic processes*

A stochastic process is a family of random variables indexed by $\mathfrak{T} = \mathbf{N}$ or $\mathfrak{T} = \mathbf{R}^+$, more rarely by $\mathbf{Z}$ or $\mathbf{R}$. These random variables are assumed to be defined on the same space Ω and take their value in the same Polish space E, which means that we have an application X

$$\begin{aligned} X \,:\, \Omega \times \mathfrak{T} &\longrightarrow E \\ (\omega,\, t) &\longmapsto X(\omega,\, t). \end{aligned}$$

We can consider X as a random variable with values in the space of applications from $\mathfrak{T}$ in E, a set commonly denoted by $E^{\mathfrak{T}}$

$$\begin{aligned} X \,:\, \Omega &\longrightarrow E^{\mathfrak{T}} \\ \omega &\longmapsto (t \mapsto X(\omega,\, t)). \end{aligned}$$

In this description, the value of $X(\omega)$ is called the trajectory of X, it is the set of elements of E obtained for fixed ω by varying t in $\mathfrak{T}$.

The first description is the most natural since it corresponds to the idea we have of a dynamical system: a sequence of values indexed by time. The second and more abstract description is used to build the rigorous mathematical framework.

It is well known that in probability theory, the space Ω is often loosely defined but what really counts is the value space of the random variables. To make life easier, it is common to consider that Ω is already the value space. Then, the random object under consideration is itself a trajectory. The value at time t is then represented by the coordinate application

$$\begin{aligned} X_t \,:\, E^{\mathfrak{T}} &\longrightarrow E \\ \omega &\longmapsto X_t(\omega) = \omega(t). \end{aligned}$$

When the index set is countable, $E^{\mathbf{N}}$ naturally inherits the product topology from that of E and that makes it a Polish space. The distribution of X is then determined by the finite-dimensional distributions. This means that two processes X and Y such that

$$\text{Law}(X_0, \cdots, X_n) = \text{Law}(Y_0, \cdots, Y_n) \text{ for any } n \geq 0,$$

have the same distribution. However, it is not enough that for any n, X_n has the same distribution as Y_n. For instance, if the Y_n are independent copies of X and $X_n = X$ for any n then $\mathbf{P}(X_1 \neq X_2) = 0$ and $\mathbf{P}(Y_1 \neq Y_2) > 0$, which prevents equality in distribution of these two processes.

The situation is more complex if the index space is uncountable, $\mathcal{E}^{\mathbf{R}^+}$ is not naturally equipped with a topology. There are two essential examples whose properties are presented below: the space of continuous functions and the space of rcll functions. In both cases, the distribution of a process is determined by the finite-dimensional distributions: two processes X and Y have the same distribution if and only if

$$\text{Law}(X_{t_1}, \cdots, X_{t_n}) = \text{Law}(Y_{t_1}, \cdots, Y_{t_n}) \text{ for any } t_1, \cdots, t_n \in \mathfrak{T}.$$

A.1.4. *σ-fields*

DEFINITION A.9.– *A σ-field $\mathcal{E}$ of a set E is a set of subsets of E which satisfies the following three properties:*

– $\varnothing \in \mathcal{E}$,

– *if $A \in \mathcal{E}$ then $A^c \in \mathcal{E}$,*

– *if $(A_n,\, n \geq 1)$ is a countable family of elements of $\mathcal{E}$ then $\bigcup_{n\geq 1} A_n$ is an element of $\mathcal{E}$.*

Sets of a σ-field are often called measurable sets.

EXAMPLE.– The most simple examples of σ-fields are the coarse σ-fields: $\mathcal{E} = \{\varnothing, E\}$ and the σ-field of all subsets of E, that is $\mathcal{E} = \mathcal{P}(E)$.

NOTE.– For a discrete random variable, we want to calculate quantities such as $\mathbf{P}(X = i)$, that is to say to calculate the probability of the singleton $\{i\}$ under the distribution of X. It is then necessary that all singletons are measurable sets. As in a countable space, a set is the at-most countable union of its singletons, by stability of a σ-field by countable unions, we see that if the singletons are measurable, all sets are.

The situation is totally different for uncountable spaces. We still want singletons to be measurable but they are not sufficient to describe any set as a countable union. In addition, it may be shown that unless we abandon the axiom of choice, we cannot reasonably build a measure on the set of all subsets of an uncountable space. That is why the notion of generated σ-filed plays a key role in the following.

DEFINITION A.10.– *Let $\mathcal{C} \subset \mathcal{P}(E)$, $\sigma(\mathcal{C})$, called the σ-filed generated by $\mathcal{C}$, denotes the smallest σ-field containing $\mathcal{C}$.*

EXAMPLE A.1.– For a set of $A \in E$, $\sigma(A) = \{\varnothing, A, A^c, E\}$.

DEFINITION A.11.– *For a Polish set E, we denote $\mathfrak{B}(E)$ the Borelian σ-field of E generated by the open sets of E.*

DEFINITION A.12.– *For an application X of E in $\mathbf{R}^d$, we denote $\sigma(X)$ as the smallest σ-field such that X is measurable from $(\mathbf{R}^d, \sigma(X))$ into $(\mathbf{R}^d, \mathfrak{B}(\mathbf{R}^d))$.*

LEMMA A.6.– *Let X be measurable from $(E, \{\varnothing, A, A^c, E\})$ into $(F, \mathcal{F})$, another Polish space. Then there exists f_1 and f_2 in F such that*

$$X(\omega) = \begin{cases} f_1 & \textit{if } \omega \in A, \\ f_2 & \textit{if } \omega \in A^c. \end{cases} \tag{A.1}$$

Proof. Let $\omega_1 \in E$. As X is measurable $B = X^{-1}(\{X(\omega_1)\})$ is one of the four following sets $\{\varnothing, A, A^c, E\}$. As ω_1 belongs to B, B cannot be empty. If $B = E$, this means that for any $\omega \in B = E$, $X(\omega) = X(\omega_1)$ therefore X is constant. The random variable X is of the form [A.1] with $f_1 = f_2 = X(\omega_1)$. If $B = A$ with $X(\omega) = X(\omega_1)$ for any $\omega \in A$. Let $\omega_2 \in A^c$, the set $C = X^{-1}(\{X(\omega_2)\})$ cannot be equal to A^c and therefore $X(\omega) = X(\omega_2)$ for any $\omega \in A^c$. Therefore, X is of the form [A.1] with $f_i = X(\omega_i)$. □

More generally, we have the following theorem.

THEOREM A.7.– *Let $X : E \longrightarrow (F, \mathcal{F})$ and $Y : E \longrightarrow (\mathbf{R}, \mathfrak{B}(\mathbf{R}))$ where E and F are two Polish spaces. The random variable Y is $\sigma(X)$ measurable if and only if there exists $\psi : (F, \mathcal{F}) \longrightarrow (\mathbf{R}, \mathfrak{B}(\mathbf{R}))$ measurable such that $Y = \psi(X)$.*

Proof. The σ-field generated by X necessarily contains the sets of the form $X^{-1}(A)$ for $A \in \mathcal{F}$. As this set of sets constitutes a σ-field, $\sigma(X) = \{X^{-1}(A),\, A \in \mathcal{F}\}$. If Y is of the form $\mathbf{1}_B$ then $Y^{-1}(\{1\}) = B$ belongs to $\sigma(X)$. Thus, there exists $C \in \mathcal{F}$ such that $B = X^{-1}(C)$. Therefore, we have $Y = \mathbf{1}_C(X)$.

Now let Y be a simple function, that is $Y = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$ with $A_i \cap A_j = \varnothing$ and $\alpha_i - \alpha_j \neq 0$ for $i \neq j$. Since $A_i = Y^{-1}(\{\alpha_i\})$, by the same reasoning, we construct $C_1, \cdots, C_n$ $\mathcal{F}$-measurable sets such that $Y = (\sum_{i=1}^n \alpha_i \mathbf{1}_{C_i})(X)$. Let Y be an $\mathcal{F}$-measurable non-negative random variable, we know that there exists a sequence of simple function $(Y_n,\, n \geq 1)$ which converges almost surely to Y. We have previously built ψ_n such that $Y_n = \psi_n(X)$. As Y_n converges to Y, ψ_n converges on the image of E by X. Unfortunately, there is no guarantee that this set is measurable. To avoid this problem, we set $\psi = \limsup_n \psi_n$. As any upper limit of measurable functions is measurable, this function is measurable. In addition, when ψ_n converges to ψ, the upper limit too. In conclusion, we have built ψ measurable such that $Y = \psi(X)$. □

DEFINITION A.13.– *A π-system is a set of subsets stable by finite intersection.*

DEFINITION A.14.– *A λ-system $\mathcal{D}$ is a set of subsets stable by monotone limits:*

- *If $A_n \subset A_{n+1} \in \mathcal{D}$ then $\cup_n A_n \in \mathcal{D}$;*
- *if $B \subset A$ with $A,\, B \in \mathcal{D}$ then $A \backslash B \in \mathcal{D}$.*

We finish this section by a theorem known as monotone class theorem which is thoroughly used to establish some formulas.

THEOREM A.8.– *Let $\mathcal{C}$ be a π-system and $\mathcal{D}$ a λ-system of a Polish space E. If $\mathcal{C} \subset \mathcal{D}$ then $\sigma(\mathcal{C}) \subset \mathcal{D}$.*

A.2. Conditional expectation

For a Polish space $(E,\, \mathcal{E})$, $L^2(E, \mathcal{E},\, \mathbf{P})$ denotes the space of square integrable random variables: the random variables such that $\mathbf{E}\big[X^2\big] < \infty$. For $\mathcal{F}$ sub-σ-field of $\mathcal{E}$, the theory of Hilbert spaces states that we can define the orthogonal projection of $L^2(E,\, \mathcal{E},\, \mathbf{P})$ onto $L^2(E,\, \mathcal{F},\, \mathbf{P})$.

DEFINITION A.15.– *Let $X \in L^2(E,\, \mathcal{E},\, \mathbf{P})$, we denote $\mathbf{E}\big[X \,|\, \mathcal{F}\big]$, the so-called conditional expectation of X given $\mathcal{F}$, defined as the orthogonal projection of X on the Hilbert space $L^2(E,\, \mathcal{F},\, \mathbf{P})$.*

By definition of an orthogonal projection, this means that

$$\mathbf{E}\big[X \,|\, \mathcal{F}\big] \text{ is } \mathcal{F}\text{-measurable and } \mathbf{E}\big[ZX\big] = \mathbf{E}\big[Z\mathbf{E}\big[X \,|\, \mathcal{F}\big]\big], \qquad [A.2]$$

for any bounded, $\mathcal{F}$-measurable, random variable Z. By analogy with the non-conditional case, we introduce the conditional probability given (a σ-field) $\mathcal{F}$, by

$$\mathbf{P}(A \,|\, \mathcal{F}) = \mathbf{E}\big[\mathbf{1}_A \,|\, \mathcal{F}\big].$$

Note that the conditional expectation and the conditional probability are random variables.

EXAMPLE.– Let $\mathcal{F} = \{\varnothing, A, A^c, \Omega\}$ where $A \in \mathcal{E}$. $\mathcal{F}$ is a sub-σ-field of $\mathcal{E}$. Let us calculate $\mathbf{E}\big[X \,|\, \mathcal{F}\big]$ for $X \in L^2(E, \mathcal{E}, \mathbf{P})$. According to Lemma A.6, if Z is an $\mathcal{F}$-measurable random variable then it can be written as

$$Z = a\,\mathbf{1}_A + b\,\mathbf{1}_{A^c} \text{ for some } a, b \in \mathbf{R}.$$

Therefore $\mathbf{E}\big[X \,|\, \mathcal{F}\big]$ can also be written as $c\,\mathbf{1}_A + d\,\mathbf{1}_{A^c}$, and it is sufficient to determine the constants c and d. By replacing $\mathbf{E}\big[X \,|\, \mathcal{F}\big]$ by $c\,\mathbf{1}_A + d\,\mathbf{1}_{A^c}$ in [A.2] we get for any a and any b,

$$\mathbf{E}\big[(a\,\mathbf{1}_A + b\,\mathbf{1}_{A^c})(c\,\mathbf{1}_A + d\,\mathbf{1}_{A^c})\big] = ac\,\mathbf{P}(A) + bd\,\mathbf{P}(A^c).$$

Therefore

$$\mathbf{E}\big[X(a\,\mathbf{1}_A + b\,\mathbf{1}_{A^c})\big] = ac\,\mathbf{P}(A) + bd\,\mathbf{P}(A^c).$$

This must be true for any a and b, thus

$$\mathbf{E}\big[X \,|\, \mathcal{F}\big] = \frac{\mathbf{E}\big[X\,\mathbf{1}_A\big]}{\mathbf{P}(A)}\,\mathbf{1}_A + \frac{\mathbf{E}\big[X\,\mathbf{1}_{A^c}\big]}{\mathbf{P}(A^c)}\,\mathbf{1}_{A^c}\,.$$

In particular, for $X = \mathbf{1}_C, C \in E$

$$\mathbf{P}(C \,|\, \mathcal{F}) = \mathbf{P}(C \,|\, A)\,\mathbf{1}_A + \mathbf{P}(C \,|\, A^c)\,\mathbf{1}_{A^c}\,.$$

The following results are easy to prove.

THEOREM A.9.– *Let X be a random variable of $L^2(E, \mathcal{E}, \mathbf{P})$ and $\mathcal{F}$ a sub-σ-field of $\mathcal{E}$. We have the following properties:*

1) If $X \geq 0$ a.s. then $\mathbf{E}\big[X \,|\, \mathcal{F}\big] \geq 0$;

2) $\mathbf{E}\big[|\mathbf{E}\big[X \,|\, \mathcal{F}\big]|\big] \leq \mathbf{E}\big[|X|\big]$;

3) for any $X \in L^1(E, \mathcal{E}, \mathbf{P})$, there exists a unique random variable, denoted by $\mathbf{E}\big[X \,|\, \mathcal{F}\big]$, that satisfies [A.2];

4) if X is $\mathcal{F}$-measurable and Y $\mathcal{E}$ measurable such that $XY \in L^1$ then $\mathbf{E}\big[XY \,|\, \mathcal{F}\big] = X\mathbf{E}\big[Y \,|\, \mathcal{F}\big]$;

5) if $X \in L^1$ is independent of $\mathcal{F}$ then $\mathbf{E}\big[X \,|\, \mathcal{F}\big] = \mathbf{E}\big[X\big]$;

6) if $\mathcal{F} \subset \mathcal{G} \subset \mathcal{E}$ then for any $X \in L^1$,

$$\mathbf{E}\big[\mathbf{E}\big[X \,|\, \mathcal{G}\big] \,|\, \mathcal{F}\big] = \mathbf{E}\big[\mathbf{E}\big[X \,|\, \mathcal{F}\big] \,|\, \mathcal{G}\big] = \mathbf{E}\big[X \,|\, \mathcal{F}\big].$$

A.2.1. *Independence and conditioning*

The following results on conditional independence are less known but absolutely essential to establish economically different forms of Markov property.

DEFINITION A.16.– *Let $\mathcal{F}_1, \cdots, \mathcal{F}_n,\, \mathcal{G} \subset \mathcal{E}$ be n σ-fields. The σ-fields $\mathcal{F}_1, \cdots, \mathcal{F}_n$ are conditionally independent given $\mathcal{G}$ when*

$$\mathbf{P}\big(\cap_{j=1}^{n} B_j \,|\, \mathcal{G}\big) = \prod_{j=1}^{n} \mathbf{P}(\, B_j \,|\, \mathcal{G})\ a.s. \qquad [A.3]$$

for any $B_j \in \mathcal{F}_j$, $j = 1, \cdots, n$.

An infinite family of σ-fields $(\mathcal{F}_r,\, r \in T)$ is conditionally independent given $\mathcal{G}$ if [A.3] is true for all finite subfamily.

THEOREM A.10.– *Let $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$ be three σ-fields of $(E,\, \mathcal{E})$. The following three properties are equivalent:*

1) $\mathcal{F}$ and $\mathcal{H}$ are conditionally independent given $\mathcal{G}$;

2) For any $H \in \mathcal{H}$, $\mathbf{P}(H \,|\, \mathcal{F} \vee \mathcal{G}) = \mathbf{P}(H \,|\, \mathcal{G})$, a.s.;

3) $\mathcal{H}$ and $\mathcal{F} \vee \mathcal{G}$ are conditionally independent given $\mathcal{G}$.

Proof. Assume $\mathcal{F}$ and $\mathcal{H}$ are conditionally independent given $\mathcal{G}$. For $F \in \mathcal{F},\ G \in \mathcal{G}$ and $H \in \mathcal{H}$, we have

$$\begin{aligned} \mathbf{E}\big[\mathbf{P}(H \,|\, \mathcal{G})\, \mathbf{1}_F\, \mathbf{1}_G\big] &= \mathbf{E}\big[\mathbf{P}(H \,|\, \mathcal{G})\mathbf{P}(F \,|\, \mathcal{G})\, \mathbf{1}_G\big] \\ &= \mathbf{E}\big[\mathbf{P}(H \cap F \,|\, \mathcal{G})\, \mathbf{1}_G\big] \end{aligned}$$

by definition of conditional independence. According to Property 4 of Theorem A.9

$$\mathbf{E}\big[\mathbf{P}(H \cap F \,|\, \mathcal{G})\, \mathbf{1}_G\big] = \mathbf{E}\big[\, \mathbf{1}_H\, \mathbf{1}_{F \cap G}\big].$$

Note that

$$\mathcal{D} = \big\{\mathcal{F} \vee \mathcal{G},\ \mathbf{E}\big[\mathbf{P}(H \,|\, \mathcal{G})\, \mathbf{1}_M\big] = \mathbf{E}\big[\, \mathbf{1}_H\, \mathbf{1}_M\big]\big\}\,.$$

Hence, $\mathcal{C} = \{M = F \cap G,\ F \in \mathcal{F},\ G \in \mathcal{C}\} \subset \mathcal{D}$. It is obvious that $\mathcal{C}$ is a π-system. By linearity and monotone convergence, it appears that $\mathcal{D}$ is a λ-system. According to Theorem A.8, $\mathcal{D}$ contains the σ-field generated by $\mathcal{C}$. Moreover, $\mathcal{F} \subset \mathcal{C}$ and $\mathcal{G} \subset \mathcal{G}$, therefore $\mathcal{C}$ contains $\mathcal{F} \vee \mathcal{G}$. This means that for any $M \in \mathcal{F} \vee \mathcal{G}$,

$$\mathbf{E}\big[\mathbf{P}(H \,|\, \mathcal{G})\, \mathbf{1}_M\big] = \mathbf{E}\big[\, \mathbf{1}_H\, \mathbf{1}_M\big].$$

As $\mathbf{P}(H \,|\, \mathcal{G})$ is $\mathcal{F} \vee \mathcal{G}$ measurable, Point 2 is satisfied.

If Point 2 is satisfied, for any $F \in \mathcal{F}$ and any $H \in \mathcal{H}$, we get

$$\begin{aligned}\mathbf{P}(F \cap H \,|\, \mathcal{G}) &= \mathbf{E}\big[\mathbf{P}(F \cap H \,|\, \mathcal{F} \vee \mathcal{G}) \,|\, \mathcal{G}\big] \\ &= \mathbf{E}\big[\mathbf{1}_F\, \mathbf{P}(H \,|\, \mathcal{F} \vee \mathcal{G}) \,|\, \mathcal{G}\big] \\ &= \mathbf{E}\big[\mathbf{1}_F\, \mathbf{P}(H \,|\, G) \,|\, \mathcal{G}\big] \\ &= \mathbf{P}(H \,|\, G)\mathbf{P}(F \,|\, \mathcal{G}).\end{aligned}$$

This proves the independence of $\mathcal{F}$ and $\mathcal{H}$ given $\mathcal{G}$.

According to Point 2, $\mathcal{H}$ and $\mathcal{F} \vee \mathcal{G}$ are conditionally independent given $\mathcal{G}$ if and only if $\mathbf{P}(H \,|\, \mathcal{F} \vee \mathcal{G}) = \mathbf{P}(H \,|\, \mathcal{G})$ for any $H \in \mathcal{H}$. This is exactly the same condition as that which induces $\mathcal{F}$ and $\mathcal{H}$ are conditionally independent given $\mathcal{G}$. The equivalence of Point 1 and Point 3 follows. □

The reasoning of the first stage is called a “monotone class argument”. It won’t be detailed any more since the principle is always the same.

THEOREM A.11.– *Let $\mathcal{G}$, $\mathcal{H}$, $\mathcal{F}_1, \cdots, \mathcal{F}_n, \cdots$ be σ-fields. The following statements are equivalent.*

1) The σ-fields $\mathcal{H}$ and $\bigvee_n \mathcal{F}_n$ are conditionally independent given $\mathcal{G}$.

2) For any integer n, the σ-fields $\mathcal{H}$ and $\mathcal{F}_{n+1}$ are conditionally independent given $\mathcal{G} \vee \mathcal{F}_1 \vee \ldots \mathcal{F}_n$.

Proof. If $\mathcal{H}$ and $\bigvee_n \mathcal{F}_n$ are conditionally independent given $\mathcal{G}$ then $\mathcal{H}$ and any σ-field generated by a finite subfamily of $\mathcal{F}_j$ are conditionally independent given $\mathcal{G}$. Apply Theorem A.10 with $\mathcal{F} = \bigvee_{j=1}^{n} \mathcal{F}_j$ then $\mathcal{F} = \bigvee_{j=1}^{n+1} \mathcal{F}_j$, we get

$$\mathbf{P}(H \,|\, \mathcal{G}) = \mathbf{P}\left(H \,|\, \mathcal{G} \vee \bigvee_{j=1}^{n} \mathcal{F}_j\right) \text{ and } \mathbf{P}(H \,|\, \mathcal{G}) = \mathbf{P}\left(H \,|\, \mathcal{G} \vee \bigvee_{j=1}^{n+1} \mathcal{F}_j\right).$$

By applying again Theorem A.10 with $\mathcal{F} = \mathcal{F}_{n+1}$, we deduce the point 2.

In the reverse direction, assume that for any $n \geq 0$, for any $H \in \mathcal{H}$, we have

$$\mathbf{P}\left(H \,|\, \mathcal{G} \vee \bigvee_{j=1}^{n} \mathcal{F}_j\right) = \mathbf{P}\left(H \,|\, \mathcal{G} \vee \bigvee_{j=1}^{n+1} \mathcal{F}_j\right).$$

By transitivity of the relation of equality, we then have

$$\mathbf{P}(H \,|\, \mathcal{G}) = \mathbf{P}\left(H \,|\, \mathcal{G} \vee \bigvee_{j=1}^{m} \mathcal{F}_j\right) \text{ for all } m.$$

According to Theorem A.10, $\mathcal{H}$ and $\bigvee_{j=1}^{m} \mathcal{F}_j$ are conditionally independent given $\mathcal{G}$. By the definition of conditional independence for an infinite number of σ-fields, Point 1 follows. □

A.2.2. *Markov property*

We now introduce the shift operator $(\theta_t,\, t \in \mathfrak{T})$. Assume that we have a family of random variables with values in E, indexed by $\mathfrak{T}$ countable or not. Aside from topological considerations, the trajectories of this process are elements of $E^{\mathfrak{T}}$, the set of applications of $\mathfrak{T}$ in E.

DEFINITION A.17.– *For any $t \in \mathfrak{T}$, the shift operator θ_t is defined by*

$$\begin{aligned} \theta_t \,:\, E^{\mathfrak{T}} &\longrightarrow E^{\mathfrak{T}} \\ (\omega(s),\, s \in \mathfrak{T}) &\longmapsto (\omega(t+s),\, s \in \mathfrak{T}). \end{aligned}$$

✍ The trajectory $\theta_t\omega$ is the trajectory that begins at time t and is indexed by the elapsed time between the present instant and t. A random variable $\sigma(X(u),\, t \le u)$-measurable is then written as $F(\omega) = \psi(\theta_t\omega)$ with ψ-measurable. A set of σ-fields $(\mathcal{F}_t,\, t \in \mathfrak{T})$ of a Polish space $(E,\, \mathcal{E})$ is a filtration whenever $t \le t' \Longrightarrow \mathcal{F}_t \subset \mathcal{F}_{t'}$.

DEFINITION A.18.– *A family of random variables $X = (X(t),\, t \in \mathfrak{T})$ with values in E, is Markov process when the following conditions hold.*

– $X(t)$ is $\mathcal{F}_t$ measurable for any $t \in \mathfrak{T}$.

– The σ-fields $\mathcal{F}_t$ and $\sigma(\{X(s)\})$ are conditionally independent given $X(t)$ for any $t \le s \in \mathfrak{T}$.

THEOREM A.12.– *If X is a Markov process then for any $t \in \mathfrak{T}$, the σ-fields $\mathcal{F}_t$ and $\sigma(\{X(s),\, s \ge t\})$ are conditionally independent given $X(t)$. Moreover, for any bounded function ψ from $(E,\, \mathcal{E})$ in $\mathbf{R}$, we have the following identity.*

$$\mathbf{E}\big[\psi \circ \theta_t \,|\, \mathcal{F}_t\big] = \mathbf{E}\big[\psi \circ \theta_t \,|\, X_t\big]. \qquad [A.4]$$

✍ This property means that past and future are conditionally independent given the present. At time t, to determine the evolution of X, it suffices to know the value of X at t, no matter how it got there.

Proof. By a monotone class argument in the case where $\mathfrak{T} = \mathbf{R}^+$, by definition in the case $\mathfrak{T} = \mathbf{N}$, it is necessary and sufficient to show that the σ-fields $\mathcal{F}_t$ and $\sigma\{X(s),\, s \in \{t = t_0 < t_1 < \ldots < t_n\}\}$ are conditionally independent given $X(t)$.

We know that for any j, $\sigma(X(t_n))$ and $\mathcal{F}_{t_{n-1}}$ are conditionally independent given $X(t_n)$. Now, $\mathcal{F}_{t_0} \vee \bigvee_{j=1}^{n-1} \sigma(X(t_j))$ is a sub σ-field of $\mathcal{F}_{t_{n-1}}$ since $X(t_j)$ is $\mathcal{F}_{t_j}$

(therefore $\mathcal{F}_{t_{n-1}}$)-measurable. According to Theorem A.11 with $\mathcal{F} = \sigma(X(t_n))$ and $\mathcal{G} = \sigma(X(t_{n-1}))$, the σ-fields $\mathcal{F}_t$ and $\sigma(X(t_n))$ are conditionally independent given $\sigma(X(t), \cdots, X(t_{n-1}))$. By applying again Theorem A.11 with $\mathcal{F} = \mathcal{F}_t$ and $\mathcal{G} = \sigma(X(t))$, we see that the σ-fields $\mathcal{F}_t$ and $\sigma\{X(s),\ s \in \{t = t_0 < t_1 < \ldots < t_n\}\}$ are conditionally independent given $X(t)$.

To prove the second point, observe that $\psi \circ \theta_t$ is $\bigvee_{s \geq t} \mathcal{F}_s$-measurable. On the other hand, as $\mathbf{E}\big[\psi \circ \theta_t \mid X_(t)\big]$ is $\mathcal{F}_t$-measurable, it is enough to show that for any $F \in \mathcal{F}_t$,

$$\mathbf{E}\big[\mathbf{E}\big[\psi \circ \theta_t \,|\, X(t)\big]\, \mathbf{1}_F\,\big] = \mathbf{E}\big[\psi \circ \theta_t\ \mathbf{1}_F\,\big].$$

However, according to the properties of conditional expectation and the first part of the proof, we have

$$\begin{aligned}\mathbf{E}\big[\mathbf{E}\big[\psi \circ \theta_t \,|\, X(t)\big]\, \mathbf{1}_F\,\big] &= \mathbf{E}\big[\mathbf{E}\big[\psi \circ \theta_t \,|\, X(t)\big]\mathbf{E}\big[\,\mathbf{1}_F\ \,|\, X(t)\big]\big]\\ &= \mathbf{E}\big[\mathbf{E}\big[\psi \circ \theta_t\ \mathbf{1}_F\ \,|\, X(t)\big]\big]\\ &= \mathbf{E}\big[\psi \circ \theta_t\ \mathbf{1}_F\,\big],\end{aligned}$$

hence the result. □

A.3. Vector spaces and orders

In the Euclidean space $\mathbf{R}^p$, for any $x,\ y \in \mathbf{R}^p$ and $\lambda \in \mathbf{R}$,

$$\begin{aligned}x &= (x(1),\ x(2), \cdots,\ x(p)),\\ \lambda.x &= (\lambda x(1),\ \cdots,\ \lambda x(p)),\\ x + y &= (x(1) + y(1),\ \cdots,\ x(p) + y(p)).\end{aligned}$$

We also denote

– the first vector of the canonical basis of $\mathbf{R}^p$ by $\mathfrak{e}_1 = (1,\ 0,\ 0,\ \cdots,\ 0)$;

– the 1-vector by $\mathbf{1} = (1,\ 1,\ \cdots,\ 1)$;

– the positive part of $X \in \mathbf{R}^p$ as $X^+ = \Big(X(1)^+,\ X(2)^+,\ \cdots,\ X(p)^+\Big)$;

– $\bar{X}$ the vector whose coordinates are the coordinates of X sorted in increasing order.

We then note $\overline{(\mathbf{R}_+)^p}$, as the set of vectors with $(\mathbf{R}_+)^p$ whose coordinates are arranged in increasing order.

Definition A.19.– *A relation "$\leq$" on a set E defines a ordering if any:*

(i) reflexive: for any $x \in E$, $x \leq x$;

(ii) transitive: for all $x,\ y,\ z \in E$, $x \leq y$ and $y \leq z$ implies $x \leq z$;

(iii) anti-symmetric: for all $x,\ y \in E$, $x \leq y$ and $y \leq x$ implies $y = x$.

The lag is then said to be total if $x \leq y$ or $y \leq x$ for all $x,\ y \in E$, partial otherwise.

We define a first partial ordering on $\mathbf{R}^p$, denoted by "$\prec$", by the following relation

$$X \prec Y \Longleftrightarrow X(i) \leq Y(i) \text{ for any } i = 1,\ \cdots,\ p. \qquad \text{[A.5]}$$

In particular, $(\mathbf{R}_+)^p$ admits of course the point $\prec$-minimal $\mathbf{0} = (0,\ \cdots,\ 0)$. Moreover, reasoning coordinate by coordinate, it is easy to see that all $\prec$-increasing and bounded sequences converge.

The Schur-convex ordering, denoted by $\prec_c$ is another (quasi-) partial ordering on $\mathbf{R}^p$, especially used in economics.

Definition A.20.– *Let u and v be two vectors of $\mathbf{R}^p$. We say that $u \prec_c v$ if*

$$\begin{cases} \displaystyle\sum_{i=1}^{p} u(i) = \sum_{i=1}^{p} v(i), \\ \displaystyle\sum_{i=k}^{p} \bar{u}(i) \leq \sum_{i=k}^{p} \bar{v}(i), \quad k = 2,\ \cdots,\ p. \end{cases}$$

✍ We can hence compare u and v by the Schur-convex ordering if u and v represent two distributions of p components of the same total mass. Then $u \prec_c v$ means that u divides a larger mass on more components than v. The relation "$\prec_c$" is almost a partial ordering in the sense that where $u \prec_c v$ and $v \prec_c u$ imply that u and v are equal up to a permutation of their coordinates.

Let $\mathfrak{S}_p$ be the set of permutations of $[\![1,\ p]\!]$. For any element $x \in \mathbf{R}^p$ and any $\gamma \in \mathfrak{S}_p$, we set

$$x_\gamma = (x(\gamma(1)),\ x(\gamma(2)),\ \cdots,\ x(\gamma(p))).$$

Definition A.21.– *Let E be a set. A function $F : \mathbf{R}^p \to E$ is called symmetric if $F(x) = F(x_\gamma)$ for all $x \in \mathbf{R}^p$ and $\gamma \in \mathfrak{S}_p$. Let $\gamma \in \mathfrak{S}_p$ and $x \in \mathbf{R}^p$. We say that γ is an ordering permutation of x if x is not totally ordered and if $\gamma(i) = j$ and $\gamma(j) = i$ for a certain couple $(i,\ j)$ such that $i < j$ and $x(i) > x(j)$, and $\gamma(k) = k$ for any $k \notin \{i,\ j\}$.*

The following proof is left to the reader.

LEMMA A.13.– *For any $x,\ y \in \mathbf{R}^p$*

$$x \prec_c y \Longleftrightarrow -x \prec_c -y. \qquad [A.6]$$

LEMMA A.14.– *(i) For any $x,\ y \in \mathbf{R}^p$,*

$$x \prec_c y \Longleftrightarrow F(x) \leq F(y),$$
$$\text{for any convex and symmetric function} F : \mathbf{R}^p \to \mathbf{R}; \qquad [A.7]$$

(ii) For any $x,\ y \in \mathbf{R}^p$, for any permutation γ ordering x,

$$x_\gamma - \bar{y} \prec_c x - \bar{y}. \qquad [A.8]$$

In particular

$$\bar{x} - \bar{y} \prec_c x - \bar{y}. \qquad [A.9]$$

We now introduce the ordering $\prec_*$ on $\overline{(\mathbf{R}_+)^p}$, which is a variant of the Schur-convex ordering.

DEFINITION A.22.– *Let u and $v \in \overline{(\mathbf{R}_+)^p}$. We denote $u \prec_* v$ if*

$$\sum_{i=k}^{p} u(i) \leq \sum_{i=k}^{p} v(i), \text{ for any } k \in [\![1,\ p]\!].$$

We can verify practically that $\prec_*$ defines a partial ordering on $\overline{(\mathbf{R}_+)^p}$. In addition,

LEMMA A.15.– *Let $u,\ v \in \overline{(\mathbf{R}_+)^p}$ such that $u \prec_* v$. Then,*

(i) for any $x \in \mathbf{R}$,

$$[u - x.\mathbf{1}]^+ \prec_* [v - x.\mathbf{1}]^+;$$

(ii) for any $i \in [\![1,\ p]\!]$ such that $u(j) \leq v(j)$, for any $y \in \mathbf{R}_+$,

$$\overline{u + y.\boldsymbol{e}_j} \prec_* \overline{v + y.\boldsymbol{e}_j}.$$

Proof. (i) The result is trivial if $u(p) \leq x$. Otherwise, for any $k \in [\![1,\ p]\!]$, there exists $\ell \geq k$ such that

$$\sum_{i=k}^{p}(u(i) - x)^+ = \sum_{i=\ell}^{p}(u(i) - x) \leq \sum_{i=\ell}^{p}(v(i) - x) \leq \sum_{i=k}^{p}(v(i) - x)^+.$$

(ii) For any $k > j$,

$$\begin{aligned}\sum_{i=k}^{p}(\overline{u + y.\mathfrak{e}_j})(i) &= \left(\sum_{i=k}^{p} u(i)\right) \vee \left(u(j) + y + \sum_{i=k+1}^{p} u(i)\right)\\ &\leq \left(\sum_{i=k}^{p} v(i)\right) \vee \left(v(j) + y + \sum_{i=k+1}^{p} v(i)\right)\\ &= \sum_{i=k}^{p}(\overline{v + y.\mathfrak{e}_j})(i),\end{aligned}$$

while for any $k \leq j$,

$$\sum_{i=k}^{p}(\overline{u + y.\mathfrak{e}_j})(i) = \sum_{i=k}^{p} u(i) + y \leq \sum_{i=k}^{p} v(i) + y = \sum_{i=k}^{p}(\overline{v + y.\mathfrak{e}_j})(i).$$

□

Several SRS that are discussed in Chapter 4 take vector-values with an unbounded number of components. So, we introduce the space $\mathcal{S}$ of almost-null positive sequences, that is to say that all components are null from a certain rank and positive before this index

$$\mathcal{S} := \{u \in (\mathbf{R}_+)^{\mathbf{N}^*},\ \exists N(u) \in \mathbf{N}, u(i) = 0\,\forall\, i > N(u)$$
$$\text{and } u(i) > 0\,\forall\, i \leq N(u)\}. \quad \text{[A.10]}$$

In other words, $N(u)$ is the number of non-zero coordinates of u. For any $u \in \mathcal{S}$, we note $\underline{u}$, the version of u arranged in descending ordering, that is

$$\underline{u}(i+1) \leq \underline{u}(i) \text{ for all } i \in \mathbf{N}.$$

We generalize the definition of the partial ordering "$\prec$" and "$\prec_c$" to $\mathcal{S}$.

DEFINITION A.23.– *Let $u, v \in \mathcal{S}$, $\tilde{u}$ and $\tilde{v}$ the restrictions of $\underline{u}$ and $\underline{v}$ to their $N(u) \vee N(v)$ first coordinates.*

(i) We say that $u \prec_c v$ if $\tilde{u} \prec_c \tilde{v}$, in other words,

$$u \prec_c v \iff \begin{cases} \displaystyle\sum_{i \in \mathbf{N}^*} u(i) & = \displaystyle\sum_{i \in \mathbf{N}^*} v(i) \\ \displaystyle\sum_{i=j}^{\infty} \underline{u}(i) & \geq \displaystyle\sum_{i=j}^{\infty} \underline{v}(i) \textit{ for all } k \in \mathbf{N}^*. \end{cases}$$

(ii) We say that $u \prec v$ if $\tilde{u} \prec \tilde{v}$, in other words,

$$u \prec_c v \Longleftrightarrow \underline{u}(i) \prec \underline{v}(i) \text{ for any } i \in \mathbf{N}^*.$$

We define similarly the set of sequences with values in $(\mathbf{R}_+)^2$. Any element u of $(\mathbf{R}^2)^{\mathbf{N}^*}$ is denoted as

$$u := ((u^1(1), u^2(1)), (u^1(2), u^2(2)), \ldots).$$

Equivalently, we can write u as

$$u = (u^1, u^2),$$

where

$$u^1 = (u^1(1), u^1(2), \ldots) \text{ and } u^2 = (u^2(1), u^2(2), \ldots)$$

are two elements of $\mathbf{R}^{\mathbf{N}^*}$.

We then define

$$\mathcal{S}^2 := \{u \in \left((\mathbf{R}_+)^2\right)^{\mathbf{N}^*}, \exists N(u) \in \mathbf{N}, \left(u^1(i), u^2(i)\right) = (0,0) \, \forall \, i > N(u) \\ \text{and } u^1(i) > 0,\ u^2(i) > 0 \, \forall \, i < N(u) \text{ and } u^1\left(N(u)\right) > 0,\ u^2\left(N(u)\right) = 0\}, \quad \text{[A.11]}$$

the set of sequences whose coefficients are equal to $(0,\ 0)$ from a certain rank $N(u)+1$, with values in $\left(\mathbf{R}_+^*\right)^2$ up to rank $N(u) - 1$ and the first component is positive up to rank $N(u)$ (the second being zero).

A.4. Bounded variation processes

In stochastic calculus, we often meet functions of bounded variation. It turns out that each such function may be written as the difference of two increasing functions and it is known that monotone functions have nice differentiability properties. We can thus write down a change of variable formula (a precursor of the celebrated Itô formula) for bounded variation processes (i.e. random processes whose sample-paths are a.s. of bounded variation) which is of constant use in stochastic calculus.

DEFINITION A.24.– *Let $[a, b] \subset \mathbf{R}$, a subdivision π of $[a, b]$ is a finite set of points $\pi = \{t_0, \cdots, t_n\}$ such that*

$$a = t_0 < t_1 < \ldots < t_n = b.$$

We denote $|\pi|$ as the step of the subdivision defined by $|\pi| = \sup_{t_i \in \pi} |t_{i+1} - t_i|$. The set of subdivisions of $[a, b]$ is $\Pi_{[a,b]}$.

DEFINITION A.25.– *Let* $f : [a, b] \to \mathbf{R}$, f *is of bounded variation when*

$$Var_{[a,b]}(f) = \sup_{\pi \in \Pi_{[a,b]}} \sum_{t_i \in \pi} |f(t_{i+1}) - f(t_i)| \text{ is finite.}$$

NOTE.– Increasing functions are of bounded variation, so are Lipschitz functions.

THEOREM A.16.– *If* $[c, d] \subset [a, b]$ *and* f *is of bounded variation on* $[a, b]$ *then* f *is of bounded variation on* $[c, d]$. *Moreover,*

$$Var_{[a,c]}(f) + Var_{[c,b]}(f) = Var_{[a,b]}(f).$$

Proof. Left to the reader. □

THEOREM A.17 (JORDAN DECOMPOSITION).– *Let* f *be of bounded variation on* $[a, b]$, *there exists one and only one pair of functions* (g, h) *such that*

1) $f = g - h + f(a)$;

2) g *and* h *are increasing;*

3) $g(a) = h(a) = 0$;

4) $Var_{[a,b]}(f) = Var_{[a,b]}(g) + Var_{[a,b]}(h)$.

Proof. Set

$$g(x) = \frac{1}{2}(f(x) - f(a) + \mathrm{Var}_{[a,x]}(f)) \text{ and } h(x) = \frac{1}{2}(f(a) - f(x) + \mathrm{Var}_{[a,x]}(f)).$$

It is immediate that

$$g(a) = h(a) = 0, \ f = g - h + f(a).$$

Since

$$|f(x) - f(y)| \leq \mathrm{Var}_{[x,y]}(f) = \mathrm{Var}_{[a,y]}(f) - \mathrm{Var}_{[a,x]}(f),$$

g and h are increasing. Therefore,

$$\mathrm{Var}_{[a,b]}(g) = g(b) - g(a) \text{ and } \mathrm{Var}_{[a,b]}(h) = h(b) - h(a).$$

Hence $\mathrm{Var}_{[a,b]}(f) = \mathrm{Var}_{[a,b]}(g) + \mathrm{Var}_{[a,b]}(h)$.

It remains to show uniqueness. Assume that there exists another pair (g_1, h_1) satisfying the same properties. Let $x < y$, since g_1 is increasing, $g_1(y) - g_1(x) \geq 0$. On the other hand,

$$g_1(y) - g_1(x) = f(y) - f(x) + h_1(y) - h_1(x) \geq f(y) - f(x),$$

since h_1 is increasing. Thus, we can say that

$$\begin{aligned} g_1(y) - g_1(x) &\geq \max(0, f(y) - f(x)) \\ &= \frac{1}{2}(f(y) - f(x) + |f(y) - f(x)|). \end{aligned}$$

For any subdivision of $[x, y]$, we then have

$$\begin{aligned} g_1(y) - g_1(x) &\geq \frac{1}{2}(f(y) - f(x) + \sum_i |f(t_{i+1}) - f(t_i)|) \\ &= \frac{1}{2}(f(y) - f(x) + \text{Var}_{[x,y]}(f)) \\ &= \frac{1}{2}(f(y) - f(x) + \text{Var}_{[a,y]}(f) - \text{Var}_{[a,x]}(f)) \\ &= g(y) - g(x). \end{aligned}$$

The function $\beta \equiv g_1 - g$ is increasing. In addition, the relation

$$f \equiv g - h + f(a) \equiv g_1 - h_1 + f(a)$$

implies that $h_1 \equiv h + \beta$ and $\beta(a) = 0$. Finally, the constraint

$$\text{Var}_{[a,b]}(f) = \text{Var}_{[a,b]}(g_1) + \text{Var}_{[a,b]}(h_1),$$

implies that

$$g_1(b) - g_1(a) + h_1(b) - h_1(a) = g(b) + h(b).$$

Now $g_1(b) - g_1(a) + h_1(b) - h_1(a) = g(b) + h(b) + 2\beta(b)$. Hence β is the null function, which means that $g \equiv g_1$ and $h \equiv h_1$. □

Let us turn now to the differentiability properties of monotone functions. The main theorem is the Lebesgue differentiation theorem.

THEOREM A.18 (LEBESGUE DIFFERENTIATION THEOREM).– *Let $[a, b]$ be a compact interval of $\mathbf{R}$ and f be an increasing function on $[a, b]$. The function f is differentiable almost everywhere (with respect to the Lebesgue measure) on $[a, b]$, f' is an integrable function and*

$$\int_a^b f'(t)\,\mathrm{d}\,t \leq f(b) - f(a). \qquad [A.12]$$

NOTE.– The derivative is defined in the usual sense as a limit of growth rates but the usual theorem which says that $f(b) - f(a) = \int_a^b f'(t)\,\mathrm{d}\,t$ is no longer true. To explain from where we get this apparent deficit, let us use a modified version of the Radon-Nikodym theorem.

THEOREM A.19.– *Let μ and ν be two σ-finite measures. There exist two measures ν_a and ν_d such that*

1) $\nu \equiv \nu_a + \nu_d$,

2) ν_a is absolutely continuous with respect to μ, i.e. there exists $h \in L^1(\mu)$ such that

$$\nu_a(B) = \int_B h \,\mathrm{d}\, \mu,$$

3) ν_d is singular with respect to μ, that is to say there exists a measurable set N such that $\nu_d(N^c) = 0$ and $\mu(N) = 0$. In other words, the support of ν_d is included in a μ-negligible set.

The second ingredient is the Stieltjes integral.

DEFINITION A.26.– *For $g \in \mathcal{C}_c(\mathbf{R})$ and f right continuous and increasing, we define the integral of g with respect to f, denoted by $\int g \,\mathrm{d}\, f$, as the limit (if it exists) of*

$$\sum_i g(t_i)(f(t_{i+1}) - f(t_i)),$$

when the step of subdivision tends to 0.

NOTE.– The existence of the limit is ensured under the hypothesis that f and g does not have the same points of discontinuity.

Finally, the Riesz theorem ensures that there exists a measure λ_f such that

$$\int g \,\mathrm{d}\, f = \int g \,\mathrm{d}\, \lambda_f \text{ for all } g \in \mathcal{C}_c(\mathbf{R}).$$

THEOREM A.20.– *For f and λ_f thus defined, we have*

$$\lambda_f([a,b]) = f(b) - f(a) \textit{ and } \lambda_f([a,b]) = f(b) - f(a^-),$$

where $f(x^-) = \lim_{y \uparrow x} f(y)$ and $f(x_+) = \lim_{y \downarrow x} f(y) = f(x)$ since f is right continuous.

Proof. We consider the sequence of functions $(h_n,\, n \geq 1)$ (see Figure A.3) defined by

$$h_n(x) = 1 \text{ for } a + \frac{1}{n} \leq x \leq b,$$

$$h_n(x) = n(x - a) \text{ for } a \leq x \leq a + \frac{1}{n},$$

$$h_n(x) = 1 - n(x - b) \text{ for } b \leq x \leq b + \frac{1}{n}\cdot$$

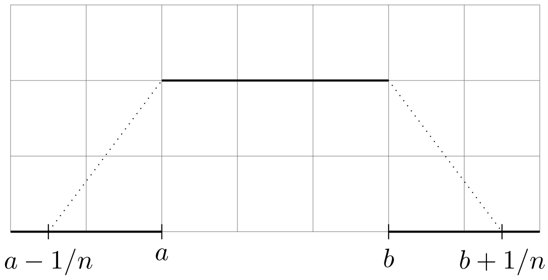

Figure A.3. *Continuous function approaching* $\mathbf{1}_{]a,b]}$

The sequence $(h_n,\ n \geq 1)$ converges a.s. to $\mathbf{1}_{[a,b]}$ and $|h_n| \leq 1$, thus by dominated convergence, we have

$$\int h_n \,\mathrm{d}\lambda_f \xrightarrow{n\to\infty} \lambda_f([a,b]).$$

By choosing the particular subdivision $\{t_0 = a, t_1 = a + 1/n, t_2 = b, t_3 = b + 1/n\}$, we obtain

$$\left(f\left(a+\frac{1}{n}\right)-f(a)\right)+\left(f(b)-f\left(a+\frac{1}{n}\right)\right)+\left(f\left(b+\frac{1}{n}\right)-f(b)\right) \\ \leq \int h_n \,\mathrm{d} f.$$

Since $0 \leq h_n \leq 1$, for every subdivision, we have

$$\sum h_n(t_i)(f(t_{i+1})-f(t_i)) \leq \left(f\left(a+\frac{1}{n}+|\pi|\right)-f(a)\right) \\ +\left(f(b+|\pi|)-f\left(a+\frac{1}{n}\right)\right)+\left(f\left(b+\frac{1}{n}\right)-f(b)\right).$$

Therefore,

$$\int h_n \,\mathrm{d} f \leq f(b+|\pi|)-f(b)+f\left(b+\frac{1}{n}\right)-f(a).$$

For any n, h_n is increasing and non-negative, therefore by monotone convergence, $\int h_n \,\mathrm{d} f$ converges to $\int h \,\mathrm{d} f$ on one hand. On the other hand, the above bounds show that, given the right continuity of f,

$$\lim_{n\to\infty} \int h_n \,\mathrm{d} f = f(b)-f(a).$$

Hence, we have shown that $f(b) - f(a) = \lambda_f[a, b]$. The other identity is proved similarly. □

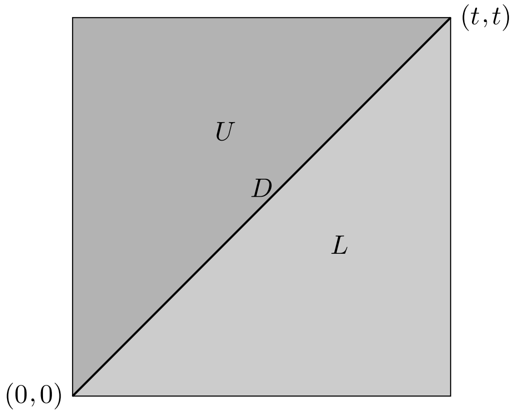

Figure A.4. *Decomposition of* $[0,t]^2$

By combining the last two theorems, we obtain the following description.

THEOREM A.21.– *Let f be right continuous and increasing, null at $-\infty$. There exists h locally integrable and ν_f a measure singular with respect to the Lebesgue measure such that*

$$f(x) = \int_{-\infty}^{x} h(u)\,\mathrm{d}\,u + \nu_f([-\infty, x]).$$

NOTE.– Hence, the subset of differentiability of f does not exhaust all the mass, that is why we get relation [A.12]. On the other hand, since h is locally integrable, the dominated convergence theorem ensures that the function $x \mapsto \int_{-\infty}^{x} h(u)\,\mathrm{d}\,u$ is continuous thus if f is discontinuous at x_0, we necessarily have $\Delta f(x_0) = \nu_f(\{x_0\})$.

THEOREM A.22 (INTEGRATION BY PARTS).– *Let f and g be two right continuous functions of bounded variation on $[0,t]$. Then,*

$$f(t)g(t) - f(0)g(0) = \int_0^t f(s^-)\,\mathrm{d}\,g(s) + \int_0^t g(s^-)\,\mathrm{d}\,f(s) + [f,g]_t, \quad \text{[A.13]}$$

where

$$[f,g]_t = \sum_{0\le s\le t} \Delta f(s)\Delta g(s) \textit{ and } \Delta f(s) = f(s) - f(s^-).$$

Proof. We compute the integral of $\mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g$ on the square $[0,t] \times [0,t]$ in two different ways. First, it comes from Theorem A.20 that

$$\int_{[0,t]^2} \mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g = (f(t) - f(0))(g(t) - g(0)).$$

Second, we decompose the square in upper and lower triangles (U and L, respectively) and the diagonal D. We apply to both triangles Fubini's theorem. In the case of the triangle L, we get

$$\begin{aligned}\int_L \mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g &= \int_0^t \left(\int_0^{s^-} \mathrm{d}\,\lambda_g\right) \mathrm{d}\,\lambda_f(s)\\ &= \int_0^t (g(s^-) - g(0))\,\mathrm{d}\,\lambda_f(s)\\ &= \int_0^t g(s^-)\,\mathrm{d}\,\lambda_f(s) - g(0)(f(t) - f(0)).\end{aligned}$$

Likewise, we obtain

$$\int_U \mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g = \int_0^t f(s^-)\,\mathrm{d}\,\lambda_g(s) - f(0)(g(t) - g(0)).$$

The diagonal D is of Lebesgue measure zero thus the integral on D of the measure $\mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g$ is reduced to the integral of its singular part

$$\int_D \mathrm{d}\,\lambda_f \otimes \mathrm{d}\,\lambda_g = \int_D \mathrm{d}\,\nu_f \otimes \mathrm{d}\,\nu_g = \sum_{0 \le s \le t} \Delta f(s) \Delta g(s),$$

according to remark A.4. By summing these three equalities, we obtain [A.13]. □

The next theorem is a change of variables theorem. In one dimension, the common change of variables theorem states that

$$F(g(t)) - F(g(0)) = \int_0^t \mathrm{d}(F \circ g)(s) = \int_0^t F'(g(s))g'(s)\,\mathrm{d}\,s. \qquad \text{[A.14]}$$

This is actually the application of the relation $f(t) - f(0) = \int_0^t f'(s)\,\mathrm{d}\,s$ to the identity $(F \circ g)' = F' \circ g.g'$. In the case of functions of bounded variation, we do not have the first relation and the second relation seems difficult to verify. However, we obtain a result similar to [A.14].

THEOREM A.23.– *Let g be a right continuous function of bounded variation and F be a function of class $\mathcal{C}^1$, we have,*

$$\begin{aligned}F(g(t)) - F(g(0)) &= \int_0^t F'(g(s^-))\,\mathrm{d}\,g(s)\\ &\quad + \sum_{0 \le s \le t} F(g(s)) - F(g(s^-)) - F'(g(s^-))\Delta g(s). \qquad \text{[A.15]}\end{aligned}$$

NOTE.– In particular, if g is continuous we obtain the usual result.

Proof. We prove [A.15] for $F(x) = x^n$, by induction on n using the integration by parts formula. By linearity, [A.15] is true for polynomials. We then approach any C^1 function by a sequence of polynomials and pass to the limit in both sides to obtain the result for F of class $\mathcal{C}^1$. □

A.5. Martingales

A.5.1. *Discrete time martingales*

DEFINITION A.27.– *A* $(X_n,\, n \geq 0)$ *sequence of real random variables, integrable is called a martingale (respectively, a sub-martingale or supermartingale) if:*

1) for any n*,* X_n *is* $\mathcal{F}_n$*-measurable and integrable;*

2) $\mathbf{E}\big[X_{n+1}\,|\,\mathcal{F}_n\big] = X_n$ *a.s. (respectively,* $\mathbf{E}\big[X_{n+1}\,|\,\mathcal{F}_n\big] \geq X_n$ *or* $\mathbf{E}\big[X_{n+1}\,|\,\mathcal{F}_n\big] \leq X_n$*).*

EXAMPLE.– A typical example is that of a sequence of independent random variables. Let $(\xi_i,\, i \geq 1)$ be i.i.d. random variables with $\mathbf{E}\big[\xi_i\big] = 0$, then the sequence defined by $X_0 = 0$, $X_n = \sum_{i=1}^n \xi_i$, is an $\mathcal{F}_n = \sigma(\xi_0, \cdots, \xi_n)$ martingale: Since ξ_{n+1} is independent of $\mathcal{F}_n$, we have

$$\mathbf{E}\Big[\sum_{i=1}^{n+1} \xi_i\,|\,\mathcal{F}_n\Big] = \mathbf{E}\big[X_n + \xi_{n+1}\,|\,\mathcal{F}_n\big] = X_n + \mathbf{E}\big[\xi_{n+1}\big] = X_n.$$

DEFINITION A.28.– *Let* τ *be a random variable with integer values* τ*, it is a stopping time when for any* $n \in \mathbf{N}$*,*

$$\{\omega : \tau(\omega) = n\} \in \mathcal{F}_n.$$

THEOREM A.24.– *Let* $(X_n,\, n \geq 0)$ *be a martingale and* τ *be a stopping time. Then* $(X_n^\tau,\, n \geq 1)$*, where* X_n^τ *is defined by* $X_{\tau\wedge n}$ *is a martingale.*

Proof. We have

$$X_{\tau\wedge n} = \sum_{m=0}^{n-1} X_m\, \mathbf{1}_{\{\tau=m\}} + X_n\, \mathbf{1}_{\{\tau\geq n\}}\,.$$

Therefore, $X_{\tau\wedge n}$ is $\mathcal{F}_n$-measurable. Since $|X_n|$ is a sub-martingale, for $m \leq n$, for any $\mathcal{F}_m$-measurable, non-negative, random variable Y_m, we have

$$\mathbf{E}\big[|X_m|Y_m\big] \leq \mathbf{E}\big[|X_n|Y_m\big].$$

Therefore,

$$\begin{aligned}\mathbf{E}\big[|X_{\tau\wedge n}|\big] &\leq \sum_{m=0}^{n-1} \mathbf{E}\big[|X_m|\,\mathbf{1}_{\{\tau=m\}}\big] + \mathbf{E}\big[|X_n|\,\mathbf{1}_{\{\tau\geq n\}}\big] \\ &\leq \sum_{m=0}^{n-1} \mathbf{E}\big[|X_n|\,\mathbf{1}_{\{\tau=m\}}\big] + \mathbf{E}\big[|X_n|\,\mathbf{1}_{\{\tau\geq n\}}\big] \leq \mathbf{E}\big[|X_n|\big] < +\infty.\end{aligned}$$

Moreover,

$$X_{\tau\wedge(n+1)} - X_{\tau\wedge n} = X_{n+1}\,\mathbf{1}_{\{\tau\geq n+1\}} - X_n\,\mathbf{1}_{\{\tau>n\}} = \mathbf{1}_{\{\tau>n\}}(X_{n+1} - X_n).$$

Hence,

$$\mathbf{E}\big[X_{\tau\wedge(n+1)} - X_{\tau\wedge n}|\mathcal{F}_n\big] = \mathbf{1}_{\{\tau>n\}}\,\mathbf{E}\big[X_{n+1} - X_n\,|\,\mathcal{F}_n\big] = 0.$$

The result follows. □

DEFINITION A.29.– *If τ is a stopping time, $\mathcal{F}_\tau$ denotes the σ-field defined by*

$$A \in \mathcal{F}_\tau \iff A \cap (\tau \leq n) \in \mathcal{F}_n,$$

for any $n \geq 0$.

LEMMA A.25.– *Let τ be a stopping time and X be an $\mathcal{F}_\tau$-measurable random variable. Then,*

$$X\ \mathbf{1}_{\{\tau\leq n\}}$$

is $\mathcal{F}_n$-measurable for any $n \in \mathbf{N}$.

Proof. Let $A \in \mathcal{F}_\tau$ and $X = \mathbf{1}_A$. Then the conclusion is trivially true. By linearity, it is also true for finitely valued random variables (i.e. random variables with a finite number of outcomes). If X is arbitrary, then there exists a sequence of finitely valued $\mathcal{F}_\tau$-measurable random variables $(X_k,\ k \geq 1)$, which converges to X almost surely. Therefore,

$$X\ \mathbf{1}_{\{\tau\leq n\}} = \lim_{k\to\infty} X_k\ \mathbf{1}_{\{\tau\leq n\}}$$

belongs to $\mathcal{F}_n$. □

THEOREM A.26.– *Let $(X_n,\ n \geq 0)$ be a martingale (respectively, a sub-martingale) and $\tau_1 \leq \tau_2$ be the two stopping times such that for $i = 1,\ 2$,*

$$\mathbf{E}\big[|X_{\tau_i}|\big] < \infty \text{ and } \liminf_{n\longrightarrow\infty} \int_{\{\tau_i>n\}} |X_n|\,\mathrm{d}\,\mathbf{P} = 0.$$

Then, we have

$$\mathbf{E}\big[X_{\tau_2}\,|\,\mathcal{F}_{\tau_1}\big] = X_{\tau_1},\ \mathbf{P}\ \textit{a.e.}$$

Proof. Let us first assume that $\tau_2 \leq k$ where $k \in \mathbf{N}$ is a constant. We have $|X_{\tau_2}| \leq |X_1| + \cdots + |X_k|$, therefore X_{τ_2} is integrable. For $n \leq k$ and $A \in \mathcal{F}_{\tau_1}$, we have

$$\mathbf{E}\big[X_k \mathbf{1}_A\big] = \sum_{j=0}^{k} \mathbf{E}\big[X_k \mathbf{1}_{A \cap \{\tau_1 = j\}}\big] = \sum_{j=0}^{k} \mathbf{E}\big[X_j \mathbf{1}_{A \cap \{\tau_1 = j\}}\big] = \mathbf{E}\big[X_{\tau_1} \mathbf{1}_A\big].$$

This entails that

$$\mathbf{E}\big[X_k \,|\, \mathcal{F}_{\tau_1}\big] = X_{\tau_1}.$$

Let us define the martingale $X_m^{\tau_2}$ by $X_{\tau_2 \wedge m}$. Then $X_k^{\tau_2} = X_{\tau_2 \wedge k} = X_{\tau_2}$. Reasoning along the same lives yields to

$$\mathbf{E}\big[X_{\tau_2} \,|\, \mathcal{F}_{\tau_1}\big] = X_{\tau_1}$$

almost-surely. For the general case, according to what we have just seen, for any $m \geq n$, we have

$$\mathbf{E}\big[X_{\tau_2 \wedge m} \,|\, \mathcal{F}_{\tau_1 \wedge n}\big] = X_{\tau_1 \wedge n}.$$

Hence, for $B \in \mathcal{F}_{\tau_1}$,

$$\mathbf{E}\big[X_{\tau_2 \wedge m} \mathbf{1}_{B \cap \{\tau_1 \leq m\}}\big] = \mathbf{E}\big[X_{\tau_1 \wedge m} \mathbf{1}_{B \cap \{\tau_1 \leq m\}}\big].$$

By dominated convergence, the right hand side of this equality converges to $\mathbf{E}\big[X_{\tau_1} \mathbf{1}_B\big]$. Therefore, the left hand side also converges. Furthermore, we can write

$$\mathbf{E}\big[X_{\tau_2 \wedge m} \mathbf{1}_{B \cap \{\tau_1 \leq m\}}\big] = \mathbf{E}\big[X_{\tau_2} \mathbf{1}_{B \cap \{\tau_2 \leq m, \tau_1 \leq m\}}\big] + \mathbf{E}\big[X_m \mathbf{1}_{B \cap \{\tau_1 \leq m\} \cap \{\tau_2 > m\}}\big].$$

When m goes to infinity, the first term of the right-hand-side converges (dominated convergence) thus the second term must also converge. In this case, we can replace the limit of this term with the $\liminf$ that is zero according to the hypothesis and the result follows. □

DEFINITION A.30.– *A process $(A_n,\, n \geq 0)$ is said to be predictable when A_n is $\mathcal{F}_{n-1}$ measurable for any $n \geq 1$.*

THEOREM A.27 (DOOB DECOMPOSITION).– *Let $(X_n,\, n \geq 0)$ be a sub-martingale. Then there exists a martingale $(M_n,\, n \geq 0)$ and $(A_n,\, n \geq 0)$ sequence predictable and increasing such that*

$$X_n = M_n + A_n,\ A_0 = 0.$$

Furthermore, this decomposition is unique.

Proof. Assume that there exist two decompositions as above, denoted by $(M,\ A)$ and $(M',\ A')$. Then, $(A_n - A'_n,,\ n \geq 0)$ is a predictable martingale, so it is constant. By hypothesis we have $A_0 = A'_0$, hence the uniqueness.

We define the sequence A_n by its increments

$$\Delta A_n = A_{n+1} - A_n = \mathbf{E}\big[X_{n+1} - X_n \,|\, \mathcal{F}_n\big] \text{ and } A_0 = 0.$$

Since X is a sub-martingale, ΔA_n is non-negative and therefore A_n is non-negative and increasing. By definition of conditional expectation, ΔA_n is $\mathcal{F}_n$ measurable hence A_{n+1} also. Thus, A is a predictable process. It results from the chain of inequalities

$$0 < \mathbf{E}\big[A_n\big] = \sum_{p=1}^{n-1} \mathbf{E}\big[\mathbf{E}\big[\Delta X_p \,|\, \mathcal{F}_p\big]\big] = \mathbf{E}\big[X_n - X_0\big] \leq \mathbf{E}\big[|X_n| + |X_0|\big],$$

that A is integrable. Finally, we define M_n by $M_n = X_n - A_n$. The integrability of M follows from those of A and X. In addition, a simple calculation

$$\mathbf{E}\big[\Delta M_n \,|\, \mathcal{F}_n\big] = \mathbf{E}\big[\Delta X_n \,|\, \mathcal{F}_n\big] - \mathbf{E}\big[\Delta A_n \,|\, \mathcal{F}_n\big] = 0,$$

shows that M_n is a martingale. □

Let a and b be two real numbers such that $a < b$, we define the following stopping times:

$$\begin{aligned}
\tau_1 &= \inf\{k > 0,\ X_k \leq a\},\\
\tau_2 &= \inf\{k > \tau_1,\ X_k \geq b\},\\
&\vdots\\
\tau_{2m-1} &= \inf\{k > \tau_{2m-2},\ X_k \leq a\},\\
\tau_{2m} &= \inf\{k > \tau_{2m-1},\ X_k \geq b\},
\end{aligned}$$

where any one of these variables is infinite as soon as the set on which the minimum index is calculated, is empty. Then, for a $n \in \mathbf{N}$, we consider the sets

$$\beta_n([a,b]) = \begin{cases} 0 & \text{if } \tau_2 > n,\\ \max\{m, \tau_{2m} \leq n\} & \text{otherwise.}\end{cases}$$

That is to say β_n is the number of $[a,\ b]$-crossings of X up to time n.

LEMMA A.28.– *Let (X_n) be a sub-martingale,*

$$\mathbf{E}\big[\beta_n([a,b])\big] \leq \frac{1}{b-a}\mathbf{E}\big[(X_n - a)^+\big].$$

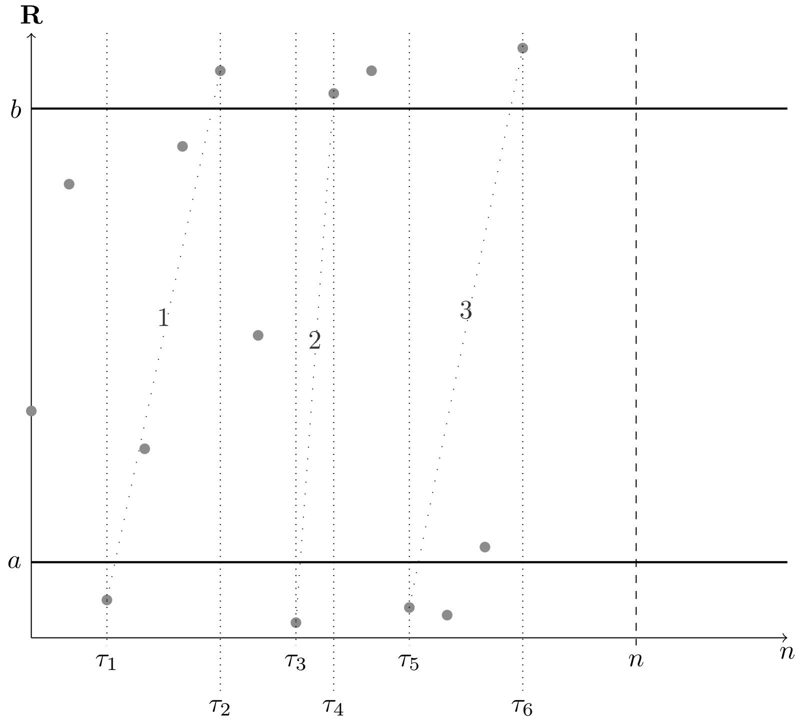

Figure A.5. *Upcrossings of a martingale*

Proof. The number of crossings of X through $[a, b]$ is the same as that of $(X_n - a)^+$ through $[0, b - a]$. Hence, we assume that $(X_n,\, n \geq 0)$ is positive sub-martingale, $a = 0$ and we want to show that

$$\mathbf{E}\big[\beta_n([0, b])\big] \leq \frac{\mathbf{E}\big[X_n\big]}{b}. \tag{A.16}$$

Set

$$\phi_i = \begin{cases} 1 & \text{if } \tau_{2k+1} < i \leq \tau_{2k+2}, \\ 0 & \text{if } \tau_{2k+2} < i \leq \tau_{2k+3}. \end{cases}$$

We easily see that

$$\beta_n([0, b]) \leq \sum_{i=1}^{n} \frac{1}{b}(X_i - X_{i-1})\phi_i$$

and

$$\{\phi_i = 1\} = \bigcup_{\text{m even}} [\{\tau_m < i\} \backslash \{\tau_{m+1} < i\}] \in \mathcal{F}_{i-1}.$$

Therefore, we have

$$\begin{aligned} b\mathbf{E}\big[\beta_n([0,b])\big] &\leq \sum_{i=1}^{n} \mathbf{E}\big[\,\mathbf{1}_{\{\phi_i=1\}}(X_i - X_{i-1})\big] \\ &\leq \sum_{i=1}^{n} \mathbf{E}\big[\,\mathbf{1}_{\{\phi_i=1\}}(\mathbf{E}\big[X_i \,|\, \mathcal{F}_{i-1}\big] - X_{i-1})\big] \\ &\leq \sum_{i=1}^{n} \mathbf{E}\big[\mathbf{E}\big[X_i - X_{i-1} \,|\, \mathcal{F}_{i-1}\big]\big] \end{aligned}$$

because $\mathbf{E}\big[X_i \,|\, \mathcal{F}_{i-1}\big] - X_{i-1} \geq 0$. Since $\mathbf{E}\big[X_0\big] \geq 0$, we have

$$b\mathbf{E}\big[\beta_n([0,b])\big] \leq \mathbf{E}\big[X_n\big] - \mathbf{E}\big[X_0\big] \leq \mathbf{E}\big[X_n\big]$$

Hence, the result. □

THEOREM A.29.– *Let X be a sub-martingale such that* $\sup_n \mathbf{E}\big[X_n^+\big] < +\infty$*, then* $(X_n,\ n \geq 0)$ *converges almost surely to a random variable* X_∞ *such that* $\mathbf{E}\big[|X_\infty|\big] < +\infty$.

Proof. If X_n does not converge a.s. then

$$\mathbf{P}(\limsup X_n > \liminf X_n) > 0.$$

Moreover,

$$(\limsup X_n > \liminf X_n) = \bigcup_{a,b\in\mathbf{Q}} (\limsup X_n > b > a > \liminf X_n),$$

thus there exists a couple of rational numbers $(a,\ b)$ such that

$$\mathbf{P}(\limsup X_n > b > a > \liminf X_n) > 0.$$

Therefore, there exists a A measurable of positive probability such that

$$\text{for all } \omega \in A,\ \limsup X_n > b \text{ and } \liminf X_n < a.$$

This means that for $\omega \in A$, $\lim_{n\to+\infty} \beta_n([a,b]) = +\infty$ hence, according to the monotone convergence theorem,

$$\lim_{n\to+\infty} \mathbf{E}\big[\beta_n([a,b])\big] = \mathbf{E}\big[\lim_{n\to+\infty} \beta_n([a,b])\big] = \infty.$$

We also know that

$$\begin{aligned}\mathbf{E}\big[\beta_n([a,b])\big] &\leq \frac{1}{b-a}\mathbf{E}\big[(X_n-a)^+\big]\big] \\ &\leq \frac{1}{b-a}(\mathbf{E}\big[X_n^+\big]+|a|) \\ &\leq \frac{1}{b-a}(\sup_n \mathbf{E}\big[X_n^+\big]+|a|) < \infty.\end{aligned}$$

We thus obtain a contradiction when n tends to infinity. This entails that $(X_n,\, n\geq 1)$ converges almost surely to a random variable denoted by X_∞. In addition,

$$\begin{aligned}\mathbf{E}\big[|X_n|\big] &= \mathbf{E}\big[X_n^+\big]+\mathbf{E}\big[X_n^-\big] = \mathbf{E}\big[X_n^+\big]+\mathbf{E}\big[X_n^+ - X_n\big] \\ &= 2\mathbf{E}\big[X_n^+\big]-\mathbf{E}\big[X_n\big] \leq 2\mathbf{E}\big[X_n^+\big]-\mathbf{E}\big[X_1\big],\end{aligned}$$

because $(X_n,\, n\geq 0)$ is a sub-martingale. We deduce that $\sup_n \mathbf{E}\big[|X_n|\big]$ is finite, which by Fatou's lemma allows us to conclude that

$$\mathbf{E}\big[|X_\infty|\big] = \mathbf{E}\big[|\liminf X_n|\big] \leq \liminf \mathbf{E}\big[|X_n|\big] \leq \sup_n \mathbf{E}\big[|X_n|\big] < \infty.$$

Hence the limit random variable X_∞ is integrable. □

COROLLARY A.30.– *A non-negative supermartingale converges almost surely.*

Proof. If X is a non-negative supermartingale then $-X$ is a sub-martingale with $(-X_n)^+=0$, thus we can apply the previous theorem to $-X$. □

A.5.2. *Continuous time martingales*

DEFINITION A.31.– *Filtration $(\mathcal{F}_t,\, t\in\mathbf{R}^+)$ is an increasing family of σ-fields. It is said to be right-continuous when*

$$\bigcap_{s>t}\mathcal{F}_s = \mathcal{F}_t, \textit{ for any } t\in\mathbf{R}^+.$$

It is said to be complete when all the negligible sets belong to $\mathcal{F}_0$.

It is assumed that in the following, all the filtrations encountered are complete and right-continuous.

DEFINITION A.32.– *Let $(\Omega, \mathcal{F}=(\mathcal{F}_t,\, t\geq 0),\, \mathbf{P})$ be a filtered space. A process $M=(M(t),\, t\geq 0)$ is an $\mathcal{F}$-martingale (respectively, sub-martingale, supermartingale) when for any $0\leq s\leq t$, $M(t)\in L^1(\mathbf{P})$*

$$\mathbf{E}\big[M(t)\,|\,\mathcal{F}_s\big] = M(s) \textit{ (respectively } \geq M(s) \textit{ and } \leq M(s)\textit{)}. \qquad [A.17]$$

We admit that any martingale admits a version with rcll trajectories. There are two basic types of martingales: the continuous martingales and the purely discontinuous martingales. The continuous martingales whose archetype is the Brownian motion are not of finite variation. Martingales on which we will focus are martingales of finite variation hence discontinuous.

EXAMPLE A.2 (POISSON PROCESS).– Let N be a Poisson process on $\mathbf{R}^+$ with intensity measure μ. The process $M(t) = N(t) - \mu([0, t])$ is a martingale for the filtration generated by the trajectories of $N : \mathcal{F}_t = \sigma(N(u), u \leq t)$. Indeed, for any Poisson process (see Chapter 6), $N(t) - N(s)$ is independent of $\mathcal{F}_s$ hence

$$\mathbf{E}\big[N(t) - N(s) \,|\, \mathcal{F}_s\big] = \mathbf{E}\big[N(t) - N(s)\big] = \mu([s, t]).$$

The result follows.

The concept of stopping time requires a slight adaptation.

DEFINITION A.33.– *A random variable τ with values in $\mathbf{R}^+ \cup \{\infty\}$ is an $\mathcal{F}$-stopping time when for any $t \geq 0$, the event $(\tau \leq t)$ belongs to $\mathcal{F}_t$.*

The σ-field $\mathcal{F}_\tau$ is the σ-field of events A of $\mathcal{F}_\infty$ such that for any $t \geq 0$, $A \cap (\tau \leq t)$ belongs to $\mathcal{F}_t$.

The stopping and convergence theorems mentioned in the section about discrete martingales stay unchanged for $\mathbf{R}^+$-indexed martingales. In particular, if M is a martingale and T a stopping time, the process

$$M^T = \{M(t \wedge T), t \geq 0\}$$

is a martingale. The martingale property is often formally verified, but it is possible that the random variables manipulated are not integrable. To circumvent this problem, we introduce the concept of local martingale.

DEFINITION A.34.– *An $\mathcal{F}$-martingale M is called closed if there exists $M_\infty \in L^1$ such that for any $t \geq 0$, we have*

$$M(t) = \mathbf{E}\big[M_\infty \,|\, \mathcal{F}_t\big].$$

DEFINITION A.35.– *An adapted process with rcll trajectories is a local martingale if there exists an increasing sequence of stopping time $(T_n, n \geq 1)$ a.s. tending to infinity such that for any n, M^{T_n} is a closed martingale. It is said that the sequence of stopping time $(T_n, n \geq 1)$ reduces M.*

THEOREM A.31.– *Let M be a local martingale. If there exists $Z \in L^1$ such that $M(t) \leq Z$ for any $t \geq 0$ then M is a martingale.*

Proof. Let $(T_n,\, n \geq 1)$ be a sequence of stopping times which reduces M. For $0 \leq s \leq t$, we have

$$M(s \wedge T_n) = \mathbf{E}\big[M(t \wedge T_n) \,|\, \mathcal{F}_s\big].$$

As T_n tends to infinity, $s \wedge T_n = s$ for n sufficiently large (depending on ω) then a.s., $M(s \wedge T_n)$ tends to $M(s)$. By dominated convergence, we get $M(s) = \mathbf{E}\big[M(t) \,|\, \mathcal{F}_s\big]$. □

DEFINITION A.36.– *The predictable σ-field is the σ-field on $\Omega \times \mathbf{R}^+$ generated by the adapted and continuous process. It is also generated by adapted and left-continuous processes as well as the processes of the form*

$$u(\omega,\, t) = \boldsymbol{1}_{[a,\, b]}(t)\alpha(\omega) \text{ with } \alpha \in \mathcal{F}_a.$$

THEOREM A.32.– *Let M be a martingale of finite variation. Let u be a predictable process such that*

$$\mathbf{E}\Big[\int_0^\infty |u(s)| \,\mathrm{d}\, \mathit{Var}(M)(s)\Big] < \infty.$$

The process

$$M^u(t) = \int_0^t u(s) \,\mathrm{d}\, M(s)$$

is a martingale.

The integral with respect to M is a Stieltjes integral as defined in section A.4.

Proof. According to the hypothesis on u and the properties of the Stieltjes integral, the integrability of $M^u(t)$ is guaranteed. It remains to prove [A.17]. For u simple predictable, that is $u(\omega,\, t) = \mathbf{1}_{[a,\, b]}(t)\alpha(\omega)$ with $\alpha \in \mathcal{F}_a$, for $0 \leq s \leq a \leq t \leq b$, we have

$$M^u(t) = \alpha(M(t) - M(a)) \text{ and } M^u(s) = 0.$$

Hence,

$$\begin{aligned}
\mathbf{E}\big[M^u(t) - M^u(s) \,|\, \mathcal{F}_s\big] &= \mathbf{E}\big[\alpha(M(t) - M(a)) \,|\, \mathcal{F}_s\big] \\
&= \mathbf{E}\big[\mathbf{E}\big[\alpha(M(t) - M(a)) \,|\, \mathcal{F}_a\big] \,|\, \mathcal{F}_s\big] \\
&= \mathbf{E}\big[\alpha\mathbf{E}\big[(M(t) - M(a)) \,|\, \mathcal{F}_a\big] \,|\, \mathcal{F}_s\big] \\
&= 0,
\end{aligned}$$

since M is a martingale. It is then sufficient, on the same principle, to discuss other cases according to the relative positions of $a,\, s,\, t,\, b$. By passing to the limit, the result is valid for all predictable processes satisfying the integrability property. □

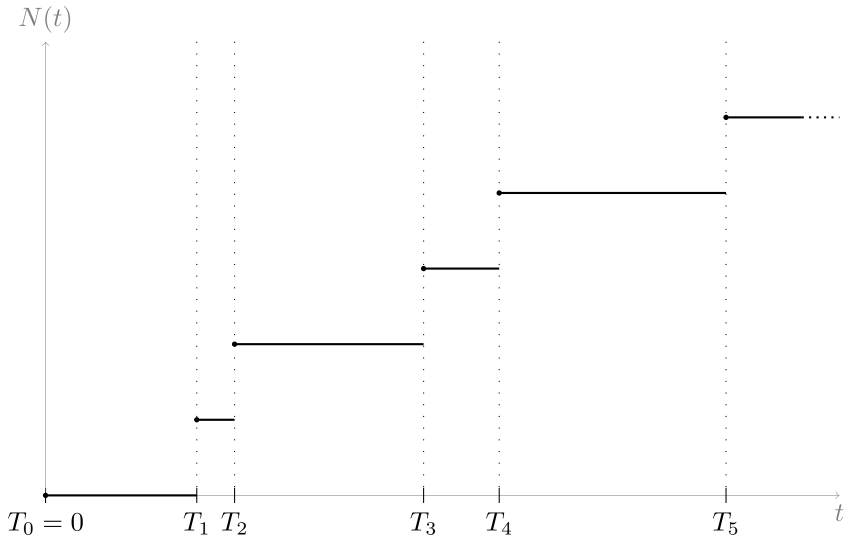

Figure A.6. *A sample-path of a point process*

Of all the martingales, those arising from point processes particularly interests us.

DEFINITION A.37.– *A point process indexed by* $\mathbf{R}^+$ *is a strictly increasing sequence* $(T_n,\, n \geq 1)$ *of non-negative random variables. We set*

$$N(t) = \sum_{n\geq 1} \boldsymbol{1}_{[0,t]}(T_n),$$

the number of points before t*. We say that the point process is integrable if* $\mathbf{E}\big[N(t)\big] < \infty$ *for any* $t \geq 0$*. In particular, this implies that* T_n *tends to infinity almost surely.*

From the knowledge of T_n, we find easily the trajectory of N. From the trajectory of N, the times T_n are nothing but its instants of jumps. Thus, there is equivalence between a purely atomic measure on $\mathbf{R}^+$ and the associated process N. We use the term of point process interchangeably to one or the other of these objects.

DEFINITION A.38.– *Let* N *be an integrable point process. We call the compensator of* N *an increasing predictable process* A*, null at time* 0 *such that* $N - A$ *is a local martingale.*

EXAMPLE.– It comes from Example A.3 that the compensator of the Poisson process is the process $A(t) = \mu([0,\, t])$.

THEOREM A.33.– *Let* N *be a point process such that* $\sup_t \mathbf{E}\big[N(t)^2\big] < \infty$ *and* A *be its compensator. The process* $((N - A)^2(t) - A(t),\, t \geq 0)$ *is a martingale.*

Proof. According to Theorem A.22

$$(N(t) - A(t))^2 = 2\int_0^t N(s^-)(\mathrm{d}\,N(s) - \mathrm{d}\,A(s)) + \sum_{s\le t} \Delta N(s)^2.$$

As a point process has jumps of height 1,

$$\Delta N(s)^2 = \Delta N(s) \text{ and } \sum_{s\le t} \Delta N(s) = N(t).$$

Therefore,

$$\begin{aligned}(N(t) - A(t))^2 - A(t) &= 2\int_0^t N(s^-)(\mathrm{d}\,N(s) - \mathrm{d}\,A(s)) + (N(t) - A(t))\\ &= 2U^1(t) + U^2(t).\end{aligned}$$

According to Theorem A.32, U^1 is a martingale, and according to the definition of A, U^2 as well. □

DEFINITION A.39.– *A marked point process R with values in Polish E is a sequence of random variables $((T_n,\, Z_n),\, n \ge 1)$, where $0 < T_n \le T_{n+1}$ and $Z_n \in E$ for any n. It is said to be integrable when $\mathbf{E}\big[\sum_{n=1}^{\infty} \boldsymbol{1}_{[0,\,t]}(T_n)\big] < \infty$. We use the notation*

$$\sum_{n\ge 1} \psi(T_n,\, Z_n) = \iint_{\mathbf{R}^+\times E} \psi(s,\, z)\,\mathrm{d}\,R(s,\, z).$$

NOTE.– A point process is nothing but a point process marked with E reduced to a singleton.

DEFINITION A.40.– *The filtration canonically associated with a marked point process R is defined by*

$$\mathcal{F}_t = \sigma\{R([0,\, s] \times B),\, s \le t,\, B \in \mathcal{B}(E)\}.$$

The predictable σ-field associated with a marked point process R is the σ-field on $\Omega \times \mathbf{R}^+ \times E$ generated by the processes of the form

$$\psi(\omega,\, s,\, z) = \alpha(\omega)\,\boldsymbol{1}_{[a,\,b]}(s)g(z),$$

with g bounded measurable function $(E,\, \mathcal{B}(E))$ in $(\mathbf{R},\, \mathcal{B}(\mathbf{R})$, $\alpha \in \mathcal{F}_a$.

DEFINITION A.41.– *A random measure on $\mathbf{R}^+ \times E$ is called predictable if for any $B \in \mathcal{B}(E)$, process*

$$t \mapsto R([0,\, t] \times B)$$

is $\mathcal{F}$-predictable.

DEFINITION A.42.– *Let R be a marked point process. We note $\mathbf{Q}_n$ the distribution of (T_{n+1}, Z_{n+1}) given $\mathcal{H}_n = (T_j, Z_j, j = 1, \cdots, n)$. For ψ non-negative and predictable, we define*

$$\int_0^t \int_E \psi(s, z)\, \mathrm{d}\,\nu(s, z)$$
$$= \sum_{n \geq 0} \int_0^t \int_E \psi(s, z) \frac{1}{\mathbf{Q}_n([s, \infty] \times E)} \mathbf{1}_{[T_n, T_{n+1}]}(s)\, \mathrm{d}\,\mathbf{Q}_n(s, z). \quad [A.18]$$

THEOREM A.34.– *Let R be a marked point process. For any predictable process ψ such that*

$$\sup_t \mathbf{E}\Big[\int_0^t \int_E \psi^2(s, z)\, \mathrm{d}\,\nu(s, z) \Big] < \infty, \quad [A.19]$$

the process

$$M^\psi : t \longmapsto \int_0^t \int_E \psi(s, z)\, \mathrm{d}\,R(s, z) - \int_0^t \int_E \psi(s, z)\, \mathrm{d}\,\nu(s, z)$$

is a local martingale. In addition, ν is the only predictable measure that satisfies this property. Moreover, in this case,

$$\langle M^\psi, M^\psi \rangle(t) = \int_0^t \int_E \psi^2(s, z)\, \mathrm{d}\,\nu(s, z). \quad [A.20]$$

Proof. On the interval $[T_n, T_{n+1}]$, the process

$$t \mapsto \int_0^t \int_E \psi(s, z)\, \mathrm{d}\,\nu(s, z)$$

is $\mathcal{H}_n$ measurable, therefore it is predictable. Let us now show that for ψ non-negative and predictable, we have $\mathbf{E}\big[M^\psi(t)\big] = 0$. Since ψ is predictable

$$\mathbf{E}\big[\psi(t, z)\, \mathbf{1}_{[T_n, T_{n+1}]}(t) \,|\, \mathcal{H}_n\big] = \psi(t, z) \mathbf{E}\big[\mathbf{1}_{[T_n, T_{n+1}]}(t) \,|\, \mathcal{H}_n\big].$$

On the other hand, the distribution of T_{n+1} given $\mathcal{H}_n$ is by definition the marginal distribution on $\mathbf{R}^+$ of $\mathbf{Q}_n$, hence

$$\mathbf{E}\big[\mathbf{1}_{[s, \infty]}(T_{n+1}) \,|\, \mathcal{H}_n\big] = \int_s^\infty \int_E \mathrm{d}\,\mathbf{Q}_n(r, \tau) = \mathbf{Q}_n([s, \infty] \times E).$$

Therefore,

$$
\begin{aligned}
&\mathbf{E}\Big[\sum_{n\geq 1}\int_0^t\int_E \psi(s,\, z)\frac{1}{\mathbf{Q}_n([s,\,\infty]\times E)}\,\mathbf{1}_{[T_n,\,T_{n+1}]}(s)\,\mathrm{d}\,\mathbf{Q}_n(s,\, z)\Big]\\
&= \sum_{n\geq 1}\mathbf{E}\Big[\int_0^t\int_E \psi(s,\, z)\frac{1}{\mathbf{Q}_n([s,\,\infty]\times E)}\mathbf{E}\big[\,\mathbf{1}_{[T_n,\,T_{n+1}]}(s)\,|\,\mathcal{H}_n\big]\,\mathrm{d}\,\mathbf{Q}_n(s,\, z)\Big]\\
&= \sum_{n\geq 1}\mathbf{E}\Big[\int_0^t\int_E \psi(s,\, z)\frac{1}{\mathbf{Q}_n([s,\,\infty]\times E)}\int_s^\infty\int_E \mathrm{d}\,\mathbf{Q}_n(r,\,\tau)\,\mathrm{d}\,\mathbf{Q}_n(s,\, z)\Big]\\
&= \sum_{n\geq 1}\mathbf{E}\Big[\int_0^t\int_E \psi(s,\, z)\,\mathrm{d}\,\mathbf{Q}_n(s,\, z)\Big]\\
&= \sum_{n\geq 1}\mathbf{E}\Big[\int_0^t\int_E \mathbf{E}\big[\psi(T_{n+1},\, Z_{n+1})\,|\,\mathcal{H}_n\big]\Big]\\
&= \mathbf{E}\Big[\int_0^t\int_E \psi(s,\, z)\,\mathrm{d}\,R(s,\, z)\Big].
\end{aligned}
\tag{A.21}
$$

For ψ simple, non-negative and predictable, for $t > r > 0$ and Y non-negative and $\mathcal{F}_r$-measurable positive, the process

$$s \mapsto \psi(s,\, z)Y\,\mathbf{1}_{[r,\,t]}(s)$$

is still predictable. By passage to the limit, this remains true for ψ non-negative and predictable. Therefore,

$$
\begin{aligned}
&\mathbf{E}\big[M^\psi(t) - M^\psi(r)\,|\,\mathcal{F}_r\big]\\
&\qquad = \mathbf{E}\Big[\int_0^t \mathbf{1}_{[r,\,t]}(s)\int_E \psi(s,\, z)(\mathrm{d}\,R(s,\, z) - \mathrm{d}\,\nu(s,\, z))\,|\,\mathcal{F}_r\Big].
\end{aligned}
\tag{A.22}
$$

Moreover, for $Y \in \mathcal{F}_r$, according to [A.21]

$$\mathbf{E}\Big[\int_0^t\int_E Y\,\mathbf{1}_{]r,\,t]}(s)\psi(s,\, z)(\mathrm{d}\,R(s,\, z) - \mathrm{d}\,\nu(s,\, z))\Big] = 0.$$

Therefore, $\mathbf{E}\big[M^\psi(t) - M^\psi(r)\,|\,\mathcal{F}_r\big] = 0$ and M^ψ is a martingale. By arguing as in Theorem A.33, we obtain [A.20]. All the above makes sense only if the expectations are finite, of which one knows nothing *a priori* for any predictable ψ. Finally to make sense of these calculations, we consider the sequence

$$\tau_n = \inf\{t,\, \int_0^t\int_E \psi(s,\, z)(\mathrm{d}\,R(s,\, z) + \mathrm{d}\,\nu(s,\, z)) > n\},$$

and we apply the reasoning of [A.21] on $[0,\, t \wedge \tau_k]$ instead of $[0,\, t]$. All expectations are finite and we have a perfectly rigorous calculation. Equation [A.22] becomes

$$\begin{aligned}&\mathbf{E}\big[M^{\psi}(t\wedge\tau_k) - M^{\psi}(r\wedge\tau_k)\,|\,\mathcal{F}_r\big]\\&\qquad = \mathbf{E}\Big[\int_0^{t\wedge\tau_k} \mathbf{1}_{[r,\,t\wedge\tau_k]}(s)\int_E \psi(s,\,z)(\mathrm{d}\,R(s,\,z) - \mathrm{d}\,\nu(s,\,z))\,|\,\mathcal{F}_r\Big].\end{aligned}$$

We deduce that M^{ψ} is a local martingale. Condition [A.19] allows us to pass to the limit in the expectations of the final result. □

We admit the following result whose proof is based on Theorem A.27 but is much more technical.

THEOREM A.35.– *A sub-local martingale M admits a decomposition as*

$$M(t) = X(t) + A(t),$$

where X is a local martingale and A an increasing predictable process null at 0. *The process A is often denoted by $\langle M,\, M\rangle$.*

For properties of PASTA type, we need the following theorem whose very technical proof is omitted.

THEOREM A.36.– *Let M be a martingale. If $\langle M,\, M\rangle(t)$ tends to infinity when t tends to infinity then*

$$\frac{M(t)}{\langle M,\, M\rangle(t)} \xrightarrow{t\to\infty} 0.$$

COROLLARY A.37.– *Let R be a marked point process on E and ν its compensator. We denote by N the associated point process $N(t) = R([0,\, t]\times E)$. Let $\psi\,:\,\Omega\times\mathbf{R}^+\times E \to \mathbf{R}$ be a predictable process. Assume that there exists $c > 0$ such that almost-surely for any $t \geq 0$,*

$$\int_0^t \int_E (1+\psi^2(s,\,z))\,\mathrm{d}\,\nu(s,\,z) \leq c\,\nu([0,t]\times E). \qquad [A.23]$$

Then, almost-surely, we have

$$\lim_{t\to\infty}\left(\frac{1}{N(t)}\int_0^t \psi(s,\,z)\,\mathrm{d}\,R(s,\,z) - \frac{1}{\nu([0,t]\times E)}\int_0^t \psi(s,\,z)\,\mathrm{d}\,\nu(s,\,z)\right) = 0. \qquad [A.24]$$

✍ This formula is the basis for all PASTA type properties that appear in this book. It links the averages computed over large populations of users and the averages computed over long periods.

Proof. To simplify the notations, we set

$$\nu(t) = \nu([0,\, t] \times E) \text{ and } \nu^{\phi}(t) = \int_0^t \phi(s,\, z)\,\mathrm{d}\,\nu(s,\, z).$$

First observe that according to Theorem A.36

$$\frac{1}{\nu(t)}(N(t) - \nu(t)) \xrightarrow{t\to\infty} 0, \qquad [A.25]$$

therefore,

$$\frac{N(t)}{\nu(t)} \xrightarrow{t\to\infty} 1. \qquad [A.26]$$

This induces that

$$\frac{\nu^{\psi}(t)}{N(t)} = \frac{\nu(t)}{N(t)}\,\frac{\nu^{\psi}(t)}{\nu(t)} \le \frac{\nu(t)}{N(t)}\,\frac{1}{\nu(t)}\int_0^t\int_E (1+\psi^2)(s,\, z)\,\mathrm{d}\,\nu(s,\, z).$$

According to equations [A.23] and [A.26], this quantity is bounded uniformly with respect to time. By writing,

$$\begin{aligned}\frac{1}{N(t)}\int_0^t\int_E \psi(s,\, z)\,\mathrm{d}\,R(s,\, z) &= \frac{\nu(t)}{N(t)}\,\frac{\nu^{\psi^2}(t)}{\nu(t)}\\ &\quad\times\frac{1}{\nu^{\psi^2}(t)}\int_0^t\int_E \psi(s,\, z)(\mathrm{d}\,R(s,\, z) - \mathrm{d}\,\nu(s,\, z)) + \frac{\nu^{\psi}(t)}{N(t)}.\end{aligned}$$

We deduce from the above that there exists $r > 0$ such that

$$\limsup_{t\to\infty} \frac{1}{N(t)}\int_0^t\int_E \psi(s,\, z)\,\mathrm{d}\,R(s,\, z) \le r. \qquad [A.27]$$

With these results of domination, we can now calculate the limit we're really interested in.

$$\begin{aligned}&\frac{1}{N(t)}\int_0^t \psi\,\mathrm{d}\,R - \frac{1}{\nu(t)}\int_0^t \psi\,\mathrm{d}\,\nu\\ &= \frac{1}{N(t)}\left(\int_0^t\int_E \psi\,\mathrm{d}\,R\right)\left(\frac{\nu(t) - N(t)}{\nu(t)}\right) + \frac{\nu^{\psi^2}(t)}{\nu(t)}\,\frac{1}{\nu^{\psi^2}(t)}\int \psi(\mathrm{d}\,R - \mathrm{d}\,\nu).\end{aligned}$$

Thus [A.37] follows from Theorem A.36, [A.25], [A.26], and [A.27]. □

NOTE.– We observe that if ψ is bounded then [A.23] is automatically satisfied.

The following theorem is a direct application of Theorem A.34. It states that the averages calculated in terms of customers are equal to the time averages when the arrival process is a Poisson process (hence the name of this property: Poisson Arrivals See Time Averages, PASTA for short). We will look at Figure 9.2 and its useful comments to see that this is not always the case.

THEOREM A.38 (PASTA PROPERTY).– *If N is a Poisson process with λ intensity then*

$$\lim_{t\to\infty} \frac{1}{N(t)} \sum_{n\, T_n \le t} \psi(T_n) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \psi(s)\,\mathrm{d}\,s,$$

as soon as one of the two limits exists.

A.6. Laplace transform

DEFINITION A.43.– *Let Y be a random variable with values in $\mathbf{R}^+$. The Laplace transform of the distribution of Y is a function with value in $\mathbf{R}^+$, defined for any $s \in \mathbf{R}^+$ by*

$$\mathcal{L}_Y(s) = \mathbf{E}\big[e^{-sY}\big].$$

LEMMA A.39.– *The Laplace transform has the following properties:*

1) it characterizes the distribution: if Y and Z are two random variables such that $\mathcal{L}_Y(s) = \mathcal{L}_Z(s)$ for any s, then Y and Z have the same distribution.

2) in particular, it characterizes every moment of the distribution: if Y has moments of order n then $\mathcal{L}_Y$ is n-times differentiable at 0, and for any $k = 1, \cdots, n$,

$$\mathbf{E}\big[Y^k\big] = (-1)^k \mathcal{L}_Y^{(k)}(0).$$

3) if $Y_1, Y_2, \cdots, Y_n$ are n independent random variables admitting a Laplace transform in s, then the Laplace transform of $\sum_{i=1} Y_i$ is defined in s and equals

$$\mathcal{L}_{\sum_{i=1}^n Y_i}(s) = \Pi_{i=1}^n \mathcal{L}_{Y_i}(s).$$

4) Let $(Y_n,\ n \in \mathbf{N})$ and Y, random variables admitting Laplace transform on a common open set. We then have the following equivalence:

$$Y_n \xrightarrow{n\to\infty} Y \textit{ in distribution} \iff \mathcal{L}_{X_n} \xrightarrow{n\to\infty} \mathcal{L}_X \textit{ simply}.$$

A.7. Notes and comments

The conditional expectation is a classic topic in the teaching of advanced probability. There are many references. We were particularly inspired by [KAL 98, CHU 01]. A presentation of the measure theory and Hilbert spaces may be found in [RUD 80, YOS 95]. Everything concerning the compensators of point processes appear completely in [JAC 79] under the disguise of “multivariate point processes”. We can also refer to [LAS 95]. The general theory of integration with respect to a martingale can be found in many books for the continuous case, books that deal with the general case, that is to say rcll trajectories are much rarer and more difficult to access, see for instance [BRA 81, LED 76, JAC 79].

Bibliography

[ARC 10] ARCEP, La qualité des services de voix et de données des réseaux mobiles (2G et 3G) en France métropolitaine, ARCEP, 2010.

[ASM 03] ASMUSSEN S., *Applied Probability and Queues*, vol. 51, Springer-Verlag, New York, 2003.

[BAC 84] BACCELLI F., BOYER P., HÉBUTERNE G., "Single-server queues with impatient customers", *Advances in Applied Probability*, vol. 16, no. 4, p. 887-905, 1984.

[BAC 89] BACCELLI F., MAKOWSKI A., "Multidimensional stochastic ordering and associated random variables", *Operations Research*, vol. 37, no. 3, p. 478-487, 1989.

[BAC 02] BACCELLI F., BRÉMAUD P., *Elements of Queueing Theory*, Springer-Verlag, Berlin, 2002.

[BAC 09a] BACCELLI F., BLASZCZYSZYN B., *Stochastic Geometry and Wireless Networks*, vol. I, Now Publishers, Hanover, 2009.

[BAC 09b] BACCELLI F., BLASZCZYSZYN B., *Stochastic Geometry and Wireless Networks*, vol. II, Now Publishers, Hanover, 2009.

[BAL 01] BALDI P., MAZLIAK L., PRIOURET P., *Martingales et chaînes de Markov*, Hermann, Paris, 2001.

[BON 11] BONALD T., FEUILLET T., *Performances des réseaux et des systèmes informatiques*, Hermès Science, Paris, 2011.

[BOR 84] BOROVKOV A.A., *Asymptotic Methods in Queuing Theory*, John Wiley & Sons, Chichester, 1984.

[BOR 92] BOROVKOV A.A., FOSS S.G., "Stochastically recursive sequences and their generalizations", *Siberian Advances in Mathematics*, vol. 2, no. 1, p. 16-81, 1992.

[BOR 94] BOROVKOV A.A., FOSS S.G., "Two ergodicity criteria for stochastically recursive sequences", *Acta Applicandae Mathematicae*, vol. 34, no. 1-2, p. 125-134, 1994.

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal

[BOR 98] BOROVKOV A.A., *Ergodicity and Stability of Stochastic Processes*, John Wiley & Sons, Chichester, 1998.

[BRA 90] BRANDT A., FRANKEN P., LISEK B., *Stationary Stochastic Models*, vol. 78, Akademie-Verlag, Berlin, 1990.

[BRÉ 81] BRÉMAUD P., *Point Processes and Queues, Martingale Dynamics*, Springer-Verlag, New York, 1981.

[CHA 99] CHAO X., MIYAZAWA M., PINEDO M., *Queueing Networks: Customers, Signals, and Product Form Solutions*, John Wiley & Sons, Chichester, 1999.

[CHU 01] CHUNG K., *A Course in Probability Theory*, Academic Press, San Diego, CA, 2001.

[ÇIN 75] ÇINLAR E., *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[CRO 96] CROVELLA M., BESTAVROS A., "Self-similarity in world wide web traffic: evidence and causes", *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, 1996.

[DAL 03] DALEY D.J., VERE-JONES D., *An Introduction to the Theory of Point Processes*, vol. I, Springer-Verlag, New York, 2003.

[DEL 76] DELLACHERIE C., MEYER P., *Probabilités et Potentiel*, vol. 2, Masson, Paris, 1976.

[DOU 02] DOUKHAN P., OPPENHEIM G., TAQQU M. (eds), *Long Range Dependence: Theory and Applications*, Birkhäuser, Boston, 2002.

[ETH 86] ETHIER S., KURTZ T., *Markov Processes: Characterizations and Convergence*, John Wiley & Sons, New York, 1986.

[FIS 93] FISCHER W., MEIER-HELLSTERN K., "The Markov-modulated Poisson process (MMPP) cookbook", *Performance Evaluation*, vol. 18, no. 2, p. 149-171, 1993.

[FLI 81] FLIPO D., "Comparaison des disciplines de service des files d'attente $G/G/1$", *Annales de l'Institut Henri Poincaré. Section B*, vol. 17, no. 2, p. 191-212, 1981.

[FOS 81] FOSS S.G., "Comparison of service disciplines in multichannel systems with waiting", *Siberian Mathematical Journal*, vol. 22, no. 1, p. 190-197, 1981.

[GAR 65] GARSIA A., "A simple proof of Eberhard Hopf's Maximal Ergodic Theorem", *Journal of Applied Mathematics and Mechanics*, vol. 14, p. 381-382, 1965.

[GRA 08] GRAHAM C., *Chaînes de Markov*, Dunod, Paris, 2008.

[HAE 08] HAENGGI M., GANTI R., "Interference in large wireless networks", *Foundations and Trends in Networking*, vol. 3, no. 2, p. 127-248, 2008.

[HOU 02] HOUDRÉ C., PRIVAULT N., "Concentration and deviation inequalities in infinite dimensions via covariance representations", *Bernoulli*, vol. 8, no. 6, p. 697-720, 2002.

[IVE 01] IVERSEN V., Teletraffic engineering and network planning, Technical University of Denmark, 2001.

[JAC 79] JACOD J., *Calcul stochastique et problèmes de martingales*, Springer-Verlag, Berlin, 1979.

[KAL 98] KALLENBERG O., *Foundations of Modern Probability*, Springer-Verlag, New York, 1998.

[KEL 79] KELLY F., *Reversibility and Stochastic Networks*, John Wiley & Sons, Chichester, 1979.

[KLE 76] KLEINROCK L., *Queueing Systems*, John Wiley & Sons, Chichester, 1976.

[LAG 96] LAGRANGE X., GODLEWSKI P., "Performance of a hierarchical cellular network with mobility dependent hand-over strategies", *Vehicular Technology Conference*, p. 1868-1872, 1996.

[LAS 95] LAST G., BRANDT A., *Marked Point Processes on the Real Line*, Springer-Verlag, New York, 1995.

[LEL 94] LELAND W., TAQQU M., WILLINGER W., WILSON D., "On the self-similar nature of ethernet traffic", *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, p. 1-15, 1994.

[LOY 62] LOYNES R.M., "Stationary waiting-time distributions for single-server queues", *Annals of Mathematical Statistics*, vol. 33, p. 1323-1339, 1962.

[MEI 89] MEIER-HELLSTERN K., "The analysis of a queue arising in overflow models", *IEEE Transactions on Communications*, vol. 37, no. 4, p. 367-372, 1989.

[MOY 08a] MOYAL P., "Convex comparison of service disciplines in real time queues", *Operations Research Letters*, vol. 36, no. 4, p. 496-499, 2008.

[MOY 08b] MOYAL P., "Stability of a processor-sharing queue with varying throughput", *Journal of Applied Probability*, vol. 45, no. 4, p. 953-962, 2008.

[MOY 10] MOYAL P., "The queue with impatience: construction of the stationary workload under FIFO", *Journal of Applied Probability*, vol. 47, no. 2, p. 498-512, 2010.

[NEU 94] NEUTS M., *Matrix-geometric Solutions in Stochastic Models*, Dover Publications, New York, 1994.

[NEV 84] NEVEU J., "Construction de files d'attente stationnaires", *Modelling and Performance Evaluation Methodology (Paris, 1983)*, vol. 60 of *Lecture Notes in Control and Information Science*, p. 31-41, Springer, Berlin, 1984.

[NOR 94] NORROS I., "A storage model with self-similar inputs", *Queuing Systems*, vol. 16, p. 387-396, 1994.

[PAN 88] PANWAR S., TOWSLEY D., WOLF J., "Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service", *Journal of the Association for Computing Machinery*, vol. 35, no. 4, p. 832-844, 1988.

[PRI 09] PRIVAULT N., *Stochastic Analysis in Discrete and Continuous Settings with Normal Martingales*, vol. 1982, Springer-Verlag, Berlin, 2009.

[RIG 98] RIGAULT C., *Principes de commutation numérique*, Hermès, Paris, 1998.

[ROB 03] ROBERT P., *Stochastic Networks and Queues*, vol. 52, Springer-Verlag, Berlin, 2003.

[RUD 80] RUDIN W., *Analyse réelle et complexe*, Masson, Paris, 1980.

[SAM 94] SAMORODNITSKY G., TAQQU M., *Stable Non-Gaussian Random Processes*, Chapman & Hall, New York, 1994.

[SHE 97] SHERMAN R., TAQQU M., WILLINGER W., "Proof of a fundamental result in self-similar traffic modeling", *Computer Communication Review*, vol. 27, no. 2, 1997.

[STO 83] STOYAN D., *Comparison Methods for Queues and Other Stochastic Models*, John Wiley & Sons, Chichester, 1983.

[THO 00] THORISSON H., *Coupling, Stationarity, and Regeneration*, Springer-Verlag, New York, 2000.

[WU 00] WU L., "A new modified logarithmic Sobolev inequality for Poisson point processes and several applications", *Probability Theory and Related Fields*, vol. 118, no. 3, p. 427-438, 2000.

[YOS 95] YOSIDA K., *Functional Analysis*, Springer-Verlag, Berlin, 1995.

Index

Stochastic Modeling and Analysis of Telecom Networks Laurent Decreusefond and Pascal Moyal
© 2012 ISTE Ltd. Published 2012 by ISTE Ltd.