Transient analysis of Markovian queueing systems: a survey with focus on closed-forms and uniformization

Gerardo Rubino

Transient analysis of Markovian queueing systems: a survey with focus on closed-forms and uniformization

Gerardo Rubino[1]

[1] Inria, Rennes, France
gerardo.rubino@inria.fr

(to appear as Chapter 7 in Queueing Theory 2, coordinated by Vladimir Anisimov and Nicolaos Limnios, ISTE Editions, 2020)

Abstract

Analyzing the transient behavior of a queueing system is much harder than studying its steady state, the difference being basically that of moving from a linear system to a linear *differential* system. However, a huge amount of efforts has been put on the former problem, from all kinds of points of view: trials to find closed-forms of the main state distributions, algorithms for numerical evaluations, approximations of different types, exploration of other transient metrics than the basic state distributions, etc.

In this survey we focus on the first two elements, the derivation of closed-forms for the main transient state distributions, and the development of numerical techniques. The paper is organized as a survey, and the main goal is to position and to underline the role of the uniformization technique, for both finding closed-forms and for developing efficient numerical evaluation procedures. In some cases, we extend the discussion to other related transient metrics that are relevant for applications.

Keywords: Markovian queueing models; transient analysis; uniformization; closed-forms

1 Introduction

Queueing models are the most used ones when analyzing the performance of a resource sharing system, and they are applied on a daily basis in many different areas.

As for other dynamical systems, most of their analysis is done in steady-state, and for several reasons (see the nice developments in the Introduction of [46]). Studying the system in equilibrium can provide qualitative insight into its structure and its limits, and also about the connections between the different parameters that characterize it. Moreover, in several areas, the related systems get very close to steady-state in short time delays for human standards (often in minutes for computing facilities or telecommunication networks).

However, if we are interested in the behavior of a system long before equilibrium, or if we don't know how far we will be from equilibrium at some specific point in time, only the transient behavior of the system matters. Another situation of interest here is when the parameters' values of a model must be changed, to take into account, for instance, a change in the environment; just after the change, the behavior of the system is naturally captured by a transient analysis. This is not just a theoretical consideration, anyone that needs to evaluate the performance of a system, will in general try to identify time regions where the parameters characterizing the environment don't change too much, and will look at both the transient distributions at the beginning of that time area, and then at steady-state, for the case where we are far enough of the changing zone. The problem is that transient analysis is much harder than studying the asymptotic behavior, and the amount of known results and related algorithmic tools is significantly less than for the steady state case. Of course, when we rely on simulation, analyzing the system in the transient regime is even simpler than in steady-state, since the latter corresponds to look at what happens when time goes to infinity.

In this chapter, we will describe some elements of the transient behavior of queues when they are in their transient phase, limiting the focus on Markovian models, because they are by far the most used ones in performance or dependability evaluation studies. We will also concentrate the effort on analytical results when available, because they are richer from the mathematical point of view and also because they show structural aspects of the behavior of the models, harder to see when applying numerical techniques or simulators. We will also look at algorithmic tools specialized to specific classes of models.

We will only consider here queues and not networks of queues, because even if in steady state there are many results concerning Markovian networks, mainly around the concept of product-form, there is nothing similar for the transient regime. Boucherie and Taylor [9] showed that to have a product-form transient distribution, the queueing network must be composed of $./M/\infty$ nodes, plus supplementary conditions. For material related to this negative result, see also [8] (where the transient behavior of the important Engset model is explored). It is shown that only in particular cases (uninteresting in practice) such a product-form transient distribution of a network of queues exists. The reader can also see [28] for an analysis of Jackson networks in the transient regime, in the positive case (nodes with an infinite number of servers).

There are many related topics that have been excluded in this text, to avoid the immediate increase in its size. These include a large part of the material concerning the numerical analysis of Markovian queueing models, where the idea is to numerically solve the Kolmogorov differential equations, the development of approximations instead of exact analysis, including mean field approaches, the consideration of non-stationary arrival processes or of state-dependent models, the study of other transient metrics than the basic distributions (for instance, different cases of hitting times, or important queueing metrics such as waiting or response times – their moments, their distributions, etc.), the time to reach equilibrium in the transient phase,... Another world we avoided is that of discrete-time queues, simply because of their limited use compared to the continuous-time ones. As stated before, we concentrate on analytical results related to the fundamental state distributions, and, in particular, on closed-forms. From the methodological viewpoint, we will often focus on the use of the uniformization technique that allows attacking the problem in discrete time.

The chapter is organized as follows. Section 2 presents the general framework of the paper, and introduces the uniformization technique that moves a problem specified in continuous time to a discrete time setting. In Section 3 we introduce some queueing models where closed-forms for the standard state distribution are not hard to obtain, and then, the most fundamental system, the $M/M/1$ model, together with its bounded version, the $M/M/1/H$ queue. Here, we basically describe the historical approaches that led to the first closed-form expressions of the associated transient distributions. In Section 4 a more recent way of deriving the transient distribution of the $M/M/1$ model is presented, and some new material is added to the previously published work. This approach is based on uniformization, and it leads to a particularly simple and useful closed-form expression. Then, in Section 5 another technique based on the concept of duality, described following [3], is presented. It allows to obtain again the same kind of expression as in Section 4 (since the initial step is again uniformization), but following a very different path. The method extends to other Markovian queueing systems. Section 6 briefly describes other models, metrics and results related to the main stream of the paper.

2 Basics on Markovian queues

In this section we establish the Markovian framework and the main notation. The goal is to present then the uniformization procedure, which plays an important role in analyzing transient distributions of basic queueing systems.

2.1 Markov models

Let us resume here the main elements we need about Markov models, and choose some global notation. Let us call process a Markov model in continuous time and chain if time is discrete. We will only consider discrete state spaces here.

Since we are dealing with queues, we will focus on homogeneous continuous time Markov models, on a finite or infinite denumerable state space S (often $\mathbb{N}$, or a segment $\{0,1,\ldots,H\}$ of positive integers). If X is such a model, its infinitesimal generator will be denoted by A and the transition rate matrix by Q. The transition rate from state i to state $j \neq i$ will then be $Q_{i,j} = A_{i,j}$. The departure rate from state i is $d_i = -A_{i,i} = \sum_j Q_{i,j}$[1]. For any $i \in S$, $Q_{i,i} = 0$. In matrix terms, if D is the diagonal matrix with $D_{i,i} = d_i$, then $A = Q - D$. When $d_i = 0$, we say that state i is absorbing. If X represents the state of a queueing model, in general it has no such absorption state; process X is then irreducible. In basic queues, the analysis focuses on the number of customers in the system; in such cases, we will use $N(t)$ instead of $X(t)$.

Let us denote by Y the discrete time Markov chain canonically embedded into X at X's jump times. If the departure rate from state i is $d_i > 0$, then the transition probability to move to state j coming from state i, $j \neq i$, is $P_{i,j} = Q_{i,j}/d_i$. If $d_i = 0$, that is, if state i is absorbing then we set $P_{i,j} = 0$ for all $j \neq i$ and $P_{i,i} = 1$. In matrix notation and for an irreducible model, we write $P = D^{-1}Q$, where D^{-1} is the diagonal matrix whose element (i,i) is $\left(D^{-1}\right)_{i,i} = 1/d_i$. Chain Y shares its initial distribution with process X.

Matrix $\left(\mathbb{P}(X(t) = j \mid X(0) = i)\right)_{i,j\in S}$ is the transition function of process X, and we will denote it by $P(t)$. When the state space is finite, matrix e^{At} exists for any $t \in \mathbb{R}_{\geq 0}$ and $\mathrm{e}^{At} = P(t)$. In the infinite case previous matrix exponential doesn't always exists, but if it does, the same equality holds. A sufficient condition for its existence is that the set of departure rates $\{d_i,\ i \in S\}$ is upper-bounded (see below).

In this chapter, we will denote $p_x(t) = \mathbb{P}(X(t) = x)$ or $p_x(t) = \mathbb{P}(N(t) = x)$. The distribution of $X(t)$ (or $N(t)$) will be denoted $p(t)$ and will be seen as a row vector, the usual convention in the Markovian world. The steady-state distribution for the process, assuming its existence, will be π, also a row vector, with the element indexed by state x denoted π_x. The Chapman-Kolmogorov equations are $p'(t) = p(t)A$ and the equilibrium or balance equations can be written $\pi A = 0$. When dealing with queues, and only to simplify the presentation, we will in general assume that the queue is empty at the time origin, that is, that $\mathbb{P}(N(0) = 0) = 1$. In the provided references, the reader can find the details for different initial conditions.

[1] "Pathological" Markov processes where this is not true can be constructed. We ignore these cases, our models will always be stable and conservative. The reader can see [3] or [24] for details.

Some notation for basic queues. When discussing the $M/M/1$ or the $M/M/1/H$ models, the arrival rate will always be $\lambda > 0$ and the service rate will be $\mu > 0$. Then, the transition probability of moving from state i to state $i+1$ is $p = \lambda/(\lambda+\mu)$ and of moving from i to $i-1$, for $i \geq 1$, is $q = \mu/(\lambda+\mu) = 1-p$; the load is $\varrho = \lambda/\mu = p/q$. The canonical Markov process associated with these models is $\{N(t), t \in \mathbb{R}_{\geq 0}\}$, where $N(t)$ is the number of customers in the system at time t.

2.2 Uniformization

If the space state S is finite, or if it is infinite denumerable with bounded departure rates, we say that X is uniform, or uniformizable. In this case, the matrix exponential e^{At} exists. Taking any positive real number Λ satisfying $\sup_i d_i \leq \Lambda$, we can construct a new process as follows. First, we define a new matrix $U = I + A/\Lambda$, where I is the identity matrix. It is easy to see that U is stochastic. We can then build a new discrete time Markov chain on the same state space S, with the same initial distribution as X and whose transition probability matrix is U; let us call Z this new chain, called the *uniformized* chain built from X with *uniformization rate* Λ. We call *uniformization* this procedure, or *randomization*, or *Jensen's method* because it was Arne Jensen who first proposed it as a procedure [22]. Observe that X is irreducible $\iff$ Z is irreducible.

From $U = I + A/\Lambda$ write $A = -\Lambda(I-U)$, multiply by t and take exponentials; we get

$$\mathrm{e}^{At} = \mathrm{e}^{-\Lambda t(I-U)} = \mathrm{e}^{-\Lambda tI}\mathrm{e}^{\Lambda tU} = \mathrm{e}^{-\Lambda t}\mathrm{e}^{\Lambda tU}. \tag{2.1}$$

The before last equality comes from the fact that I commutes with any matrix M and thus, so do matrices aI and bM for any coefficients a and b. The last one is an immediate property of the matrix exponential function. Writing the Taylor series of previous matrix function of t at $t = 0$, we obtain

$$\mathrm{e}^{At} = \sum_{n\geq 0} \mathrm{e}^{-\Lambda t}\frac{(\Lambda tU)^n}{n!} = \sum_{n\geq 0} \mathrm{e}^{-\Lambda t}\frac{(\Lambda t)^n}{n!}U^n. \tag{2.2}$$

Looking at probability distributions as row vectors and denoting by α the initial distribution of X (and thus, also of Z), by $p(t)$ the distribution of $X(t)$, $t \in \mathbb{R}_{\geq 0}$, and by $q(n)$ the distribution of $Z(n)$, $n \in \mathbb{N}$, we get the vectorial representation

$$p(t) = \sum_{n\geq 0} \mathrm{e}^{-\Lambda t}\frac{(\Lambda t)^n}{n!}q(n). \tag{2.3}$$

In scalar form, for any $j \in S$,

$$p_j(t) = \sum_{n\geq 0} \mathrm{e}^{-\Lambda t}\frac{(\Lambda t)^n}{n!}q_j(n), \tag{2.4}$$

with, concerning X, $p(t) = \big(\dots p_j(t)\dots\big)$ where $p_j(t) = \mathbb{P}(X(t) = j)$, and similarly, $q(n) = \big(\dots q_j(n)\dots\big)$, where $q_j(n) = \mathbb{P}(Z(n) = j)$.

Remark 2.1 ***– The power of uniformization.*** *Uniformization is a very fruitful and powerful way of working with Markovian models. It is fruitful because it moves a problem specified in continuous time into a new one that evolves in discrete time, that is, a linear differential problem into a linear difference one. Moreover, the transformation has a very simple probabilistic interpretation, as briefly described before, that has led to many new results in many published works. It is powerful because when we need numerical values, the obtained series involves only positive terms and the only operators involved are sums and products; this leads to stable algorithms. Next remark emphasizes a last important feature of this transformation, extremely useful for numerical evaluations.*

Remark 2.2 ***– Error control in uniformization.*** *Assume we need to know, say, $p_j(t)$, for some fixed $j \in \mathbb{N}$, with an absolute error less than ε. If we truncate the series at index $n = N$, the absolute error is*

$$err_N(t) := p_j(t) - \sum_{0\le n\le N} e^{-\Lambda t}\frac{(\Lambda t)^n}{n!} q_j(n) = \sum_{n>N} e^{-\Lambda t}\frac{(\Lambda t)^n}{n!} q_j(n) \ge 0. \tag{2.5}$$

Now, since $q_j(n) < 1$,

$$0 \le err_N(t) < \sum_{n>N} e^{-\Lambda t}\frac{(\Lambda t)^n}{n!} < \varepsilon \iff \sum_{0\le n\le N} e^{-\Lambda t}\frac{(\Lambda t)^n}{n!} > 1-\varepsilon.$$

So, given ε, we compute

$$N_{min} = \min\left\{ N \in \mathbb{N} \;\middle|\; \sum_{0\le n\le N} e^{-\Lambda t}\frac{(\Lambda t)^n}{n!} > 1-\varepsilon \right\}. \tag{2.6}$$

Previous observations then say that summing the first N_{min} terms of the series in (2.4) guarantees an approximation of $p_j(t)$ with absolute error less than ε. The key observation here is that this is not only elementary, but (i) computing N_{min} is an extremely fast procedure, and (ii) we can do it beforehand, and thus evaluate the cost of the whole numerical evaluation before starting it.

Just for completeness, another way of connecting chain Z with the original process X is the following. If $\{N(t), t \in \mathbb{R}_{\ge 0}\}$ is the counting process of a Poisson process independent of Z and having rate Λ, then we have $Z(N(.)) \equiv_{\rm st} X(.)$, where ‘ $\equiv_{\rm st}'$ means “stochastically equivalent” (same joint distributions at any number of arbitrary points in $\mathbb{R}_{\ge 0}$).

There are many references concerning this important tool of the Markovian universe. See, for instance, [14] and the many references therein.

3 First examples

In this section we will describe the transient regime of some basic models for which obtaining transient distributions is simple and/or well known. In the infinite state case, the main difficulty in the analysis of these models is the fact that we are dealing with an infinite linear differential system with an infinite number of unknowns.

3.1 The Ehrenfest model in continuous time

Let us start with the very famous and simple model proposed in 1907 by Tatiana and Paul Ehrenfest for the analysis of the second principle of thermodynamics. We have two boxes numbered 1 and 0 and H particules. Initially all particles are in box 1, and the particules move randomly from a box to the other. Let us say that $N(t)$ is the number of particles in box 1 at time t (so, $N(0) = H$). The system's dynamics corresponds to the birth and death process $\{N(t),\ t \in \mathbb{R}_{\geq 0}\}$, with non-null rates $Q_{j,j-1} = j\lambda$ for $1 \leq j \leq H$, and $Q_{j,j+1} = (H-j)\lambda$ for $0 \leq j \leq H-1$.

This process can be also represented as the sum $X_1(t) + \cdots + X_H(t)$, where the $X_i(t)$s are i.i.d., and each term is the state of the two-state Markov process given in Figure 1, with the initial condition $X_1(0) = 1$.

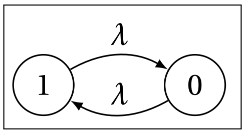

Figure 1: The evolution of an individual particle in the continuous time Ehrenfest model. When the initial state is state 1, the probability for the particle to be in state 1 at time t is $p(t) = \big(1 + \mathrm{e}^{-2\lambda t}\big)/2$.

Observe that the number of particles in box 1 at t is also the number of customers at time t in a finite-source queue with state-dependent arrival rate $\lambda_i = (H-i)\lambda$, $i = 0, 1, \ldots, H-1$ (that is, a queue fed by H Exponential sources, each with same rate λ) and state-dependent service rate $\mu_j = j\lambda$, $j = 1, 2, \ldots, H$ (that is, a queue with H servers, each with, again, the same rate λ), the $M/M/H//H$ model in Kendall's notation.

From $\mathbb{P}(X_i(t) = 1) = p(t)$ for the ith particle, we immediately obtain for any $j \in \{0, 1, 2, \ldots, H\}$, the expression

$$p_j(t) = \frac{1}{2^H}\binom{H}{j}\big(1 + \mathrm{e}^{-2\lambda t}\big)^j\big(1 - \mathrm{e}^{-2\lambda t}\big)^{H-j}.$$

The mean number of units in, say, box 1, is then $H\big(1 + \mathrm{e}^{-2\lambda t}\big)/2$.

3.2 The $M/M/\infty$ model

This is probably the simplest queueing model with an infinite state space. The family of distributions $\{p(t), t \in \mathbb{R}_{\geq 0}\}$ satisfies the system $p'(t) = p(t)A$, that here writes as follows. For $j \geq 1$,

$$p'_j(t) = -(\lambda + j\mu)p_j(t) + \lambda p_{j-1}(t) + (j+1)\mu p_{j+1}(t)$$

and if $j = 0$,

$$p'_0(t) = -\lambda p_0(t) + \mu p_1(t).$$

The derivation of the solution to this infinite differential system is pretty simple. See, for instance, the probabilistic proof in [27, page 183], or see the 2nd edition of [21] (in the References we mention the current 5th edition of the book, but the derivation details are in the 2nd one). Assume, to simplify the discussion, that the system is empty at time 0. The result is that the transient distribution is Poisson: for any $j \in \mathbb{N}$,

$$p_j(t) = \mathrm{e}^{-\psi(t)} \frac{\psi(t)^j}{j!}, \qquad \text{where } \psi(t) = \varrho\big(1 - \mathrm{e}^{-\mu t}\big) \text{ and } \varrho = \frac{\lambda}{\mu}.$$

It follows that the mean number of customers in the system at time t is

$$\mathbb{E}(N(t)) = \psi(t) = \varrho\big(1 - \mathrm{e}^{-\mu t}\big).$$

3.3 A queue with no server and catastrophes

Even if generally speaking, it is hard to attack a system of infinite differential equations looking for closed-forms solutions, the $M/M/\infty$ model is not the only case where this task can be successfully performed and using elementary techniques only. Consider a system where units arrive according to a Poisson process with rate λ, and where the service is exponentially distributed with rate μ, but when it ends, the whole buffer is emptied. Equivalently, think of a queue with no server and catastrophes or breakdowns, also called by some authors a birth process with catastrophes, where the latter arrive according to a new Poisson process independent from the arrival one, having rate μ. It still corresponds to a queueing system storing units, and such that after an exponentially distributed amount of time, once the system occupied with at least one unit, all units staying there are instantaneously sent to some other place, say for being consumed (or destroyed). Assume, as usual, that the system is empty at time 0. Next picture (Figure 2) illustrates this behavior.

In this example, the Chapman-Kolmogorov equations are $p'_0(t) = -\lambda p_0(t) + \mu\big(1 - p_0(t)\big)$ and for $j \geq 1$, $p'_j(t) = -(\lambda + \mu)p_j(t) + \lambda p_{j-1}(t)$. Denote by $\widetilde{f}$ the Laplace transform of function f of the time variable t, and by s the Laplace transform variable.

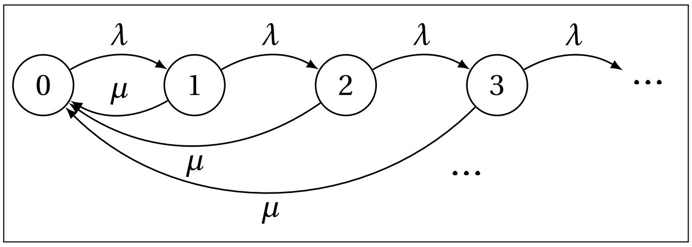

Figure 2: An example where a closed-form for the transient behavior is easy to derive.

Taking Laplace transforms in the Chapman-Kolmogorov equations, the corresponding linear system is easily solved. We get, for any $j \geq 0$,

$$\tilde{p}_j(s) = \left(\frac{\lambda}{\lambda + \mu + s}\right)^j \frac{\mu + s}{s\big(\lambda + \mu + s\big)}.$$

For the inversion of these transforms, let us set first the notation

$$\varrho = \frac{\lambda}{\mu}, \qquad p = \frac{\lambda}{\lambda + \mu}, \qquad \mathscr{P}_M(t) = \sum_{m=0}^{M} \mathrm{e}^{-(\lambda+\mu)t} \frac{(\lambda + \mu)^m t^m}{m!}.$$

We then obtain first

$$p_0(t) = \frac{1 + \varrho \mathrm{e}^{-(\lambda+\mu)t}}{1 + \varrho} = 1 - p + p\mathrm{e}^{-(\lambda+\mu)t},$$

and for $j \geq 1$,

$$p_j(t) = \pi_j - p^j \mathscr{P}_{j-1}(t) + p^{j+1} \mathscr{P}_j(t),$$

where the steady-state distribution π is given, for any $j \in \mathbb{N}$, by

$$\pi_j = \frac{1}{1 + \varrho}\left(\frac{\varrho}{1 + \varrho}\right)^j = (1 - p)p^j.$$

If we want to compute the mean number of units at time t, $m(t) = \mathbb{E}\big(N(t)\big) = \sum_{j \geq 1} j p_j(t)$, it's better to stay in the Laplace world, and compute

$$\tilde{m}(s) = \sum_{j \geq 1} j \tilde{p}_j(s) = \frac{\lambda}{s\big(\mu + s\big)} = \varrho\left(\frac{1}{s} - \frac{1}{\mu + s}\right),$$

leading immediately, by inversion of this transform, to $m(t) = \varrho\big(1 - \mathrm{e}^{-\mu t}\big)$. Observe that this expectation is the same as for the $M/M/\infty$ model.

3.4 The fundamental $M/M/1$ model

This is the most fundamental queue and one of the most used queueing models. Its transient regime has been the object of a large amount of publications, with many variations around similar approaches and many interesting results. Its derivation is more difficult than for the $M/M/\infty$ model. In this survey, we will first focus on some of the main historical developments. Then, we will describe our own contributions following the approach we believe is the richest, the uniformization-based one.

Start by writing the associated Chapman-Kolmogorov equations: for $j \geq 1$,

$$p_j'(t) = -(\lambda + \mu)p_j(t) + \lambda p_{j-1}(t) + \mu p_{j+1}(t) \tag{3.1}$$

and for the empty queue,

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t). \tag{3.2}$$

Researchers looked for a solution to these equations for years, and the breakthrough arrived in the period 1953–1956, where different works were published, proposing the first expressions of the $p_j(t)$ probabilities, all in terms of Bessel functions.

The first analytical results on the transient regime of the $M/M/1$ models appear to be those of Clarke in 1953, in an unpublished work that only appeared as an internal technical report of the University of Michigan [11]. In that text, the author obtains the classic solution in terms of modified Bessel functions of the first kind[2] making a change of variable and solving the new differential system (now, of the hyperbolic type) in terms of the solutions to Volterra integral equations.

In 1954 appear two papers, [29] and [5], providing the same type of expressions of the transient distribution of the $M/M/1$ queue. In the first one, the authors analyze the spectrum of general birth and death processes (including the case of processes where state 0 is made absorbing, useful for instance to study the busy periods of queues). The second paper follows a completely different approach, by applying z-transforms first, then the Laplace transform, and obtaining a closed-form for the double transform function. The latter is then inverted, leading to the same expression as before based on modified Bessel functions. We will present it in some detail below. Then we have [32], where the same type of representation of the transient distribution appears but not in a constructive manner, basically showing that it satisfies the differential equations. Closing this series of papers of the 50s, we have [10], who follows a purely combinatorial path to derive again the same type of expression.

Let us provide now some elements of the solving procedure in [5]. This is probably the most frequently described method for deriving the transient distribution of the

[2] The modified Bessel function of the first kind, with index or order $k \in \mathbb{N}$, is the function of z

$$z \mapsto I_k(z) = \sum_{n\geq 0} \frac{1}{n!(k+n)!}\left(\frac{z}{2}\right)^{k+2n}.$$

$M/M/1$ queue (for instance, in textbooks). Consider the generating function of the distribution $p(t)$ (or z-transform)

$$G(z,t) \overset{\text{def}}{=} \sum_{j\geq 0} p_j(t) z^j,$$

defined at least on the domain $\{|z| \leq 1\}$ for any $t \in \mathbb{R}_{\geq 0}$. If we multiply the differential equation associated with state j by z^j and sum on j over $\mathbb{N}$, we obtain the equation

$$\frac{\partial G(z,t)}{\partial t} = -(\lambda+\mu)G(z,t) + \mu p_0(t) + \mu \frac{G(z,t) - p_0(t)}{z} + \lambda z G(z,t),$$

with the initial condition $G(z,0) = z^i$ if $p_i(0) = 1$. As before, let us concentrate on the case of an empty queue at time 0, that is, on the border condition $G(z,0) = 1$. Observe that since $|G(z,t)| \leq 1$ when $|z| \leq 1$, for all $t \in \mathbb{R}_{\geq 0}$, the convergence region of the Laplace transform of $\partial G(z,t)/\partial t$ is, at least, $\{\mathrm{Re}(s) > 0\}$, for all $t \geq 0$. Taking then Laplace transforms on both sides of previous equation, we obtain

$$s\widetilde{G}(z,s) - G(z,0) = -(\lambda+\mu)\widetilde{G}(z,s) + \mu\widetilde{p}_0(s) + \mu\frac{\widetilde{G}(z,s) - \widetilde{p}_0(t)}{z} + \lambda z \widetilde{G}(z,s),$$

leading, using $G(z,0) = 1$, to

$$\widetilde{G}(z,s) = \frac{\mu\widetilde{p}_0(s)\big(1-z\big) - z}{\lambda z^2 - (\lambda+\mu+s)z + \mu}.$$

At this point, the difficulty is the unknown transform $\widetilde{p}_0$ appearing in the numerator. But look at the denominator as a polynomial in z. We first observe that $\lambda z^2 - (\lambda+\mu+s)z + \mu = 0$ has the two solutions

$$r_1(s) = \frac{\lambda+\mu+s-\sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}}{2\lambda}, \qquad r_2(s) = \frac{\lambda+\mu+s+\sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}}{2\lambda}.$$

Some elementary algebra then shows that $r_1(s) \neq r_2(s)$ because $\lambda\mu > 0$, and that $|r_1(s)| < 1 < |r_2(s)|$. Since function $\widetilde{G}$ is analytic inside the unit disc in z, the numerator of the obtained expression for this function must also be zero when $z = r_1(s)$ allowing solving for $\widetilde{p}_0$. This is the technical finding[3] that led to obtaining formal results following this direct approach (for instance, in [29, end of page 366] it is said that the generating function method can't provide a path towards the solution, showing that an idea for finding function $\widetilde{p}_0$ was missing, until this breakthrough arrived). The closed-form for $\widetilde{p}_0(s)$ is

$$\widetilde{p}_0(s) = \frac{r_1(s)}{\mu\big(1-r_1(s)\big)},$$

[3] Some authors prefer to use Rouché's theorem to prove that the denominator in the expression of G^* has only one zero inside the unit disc.

from which we get a closed-form for $\widetilde{G}$.

This function is then inverted in s and developed at 0 as a Taylor series, leading to the well-known expression

$$p_j(t) = \mathrm{e}^{-(\lambda+\mu)t}\Big[\varrho^{j/2}I_j + \varrho^{(j+1)/2}I_{j+1} + (1-\varrho)\varrho^j \sum_{k=j+2}^{\infty} \varrho^{-k/2}I_k\Big] \tag{3.3}$$

where all appearing modified Bessel functions (I_j, I_{j+1}, I_k) are all taken at the same point $2t\sqrt{\lambda\mu}$, and where $\varrho = \lambda/\mu$. This is the expression most often found in textbooks, for instance. Numerically speaking, it is not an efficient way of obtaining the transient probabilities [2]. We claim that this position is taken today by uniformization-based approaches (see below).

A different approach. In the book [40], the author develops an approach (used by him and coauthors in several papers) where the idea is to study the joint process $\big(A(t), D(t)\big)_{t\in\mathbb{R}_{\geq 0}}$, with $A(t)$ = number of arrivals to the queue in $[0,t]$ and $D(t)$ = number of departures in the same period (an idea coming from [35]; see also [44]). Assuming again the system empty at the beginning, if $p_{a,d}(t) = \mathbb{P}(A(t) = a, D(t) = d)$, we have, for any $j \in \mathbb{N}$,

$$p_j(t) = \mathbb{P}(N(t) = j) = \sum_{\ell\geq 0} p_{j+\ell,\ell}(t).$$

The author starts by writing the system of differential equation satisfied by the distribution of this bi-dimensional Markov process; for instance, for $a > d \geq 1$,

$$p'_{a,d}(t) = -(\lambda+\mu)p_{a,d}(t) + \lambda p_{a-1,d}(t) + \mu p_{a,d-1}(t).$$

Taking Laplace transforms, solving for them, inverting the transforms, and making usage of some technical methods such as hypergeometric series to obtain integrals, the result is a new type of expression. For any state $j \in \mathbb{N}$,

$$p_j(t) = (1-\varrho)\varrho^j + \varrho^j \mathrm{e}^{-(\lambda+\mu)t} \sum_{\ell=0}^{\infty} \frac{(\lambda t)^\ell}{\ell!} \sum_{m=0}^{j+\ell} (\ell - m)\frac{(\mu t)^{m-1}}{m!}, \tag{3.4}$$

where $\varrho = \lambda/\mu$. As we can see, the obtained representation is simpler than the previous one based on Bessel functions. However, observe that the factor $(\ell - m)$ inside the series at the r.h.s. can be negative, thus leading to numerical instabilities in some cases. See also [13] for related material.

3.5 $M/M/1$ with bounded waiting room: the $M/M/1/H$ model

First of all, observe that for this finite system, since the number of free unknowns is H, when H is very small, we can obtain closed-forms by just asking a tool such as `Maple` or `Mathematica` to symbolically solve the corresponding differential equations. For instance, for the very first values and just solving the equations by hand, we have the following expressions:

- for the $M/M/1/1$,
$$p_0(t) = \frac{\mu}{\lambda+\mu}\Big(1-\mathrm{e}^{-(\lambda+\mu)t}\Big),$$
$$p_1(t) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}\mathrm{e}^{-(\lambda+\mu)t}.$$

- For the $M/M/1/2$,
$$p_0(t) = \frac{1}{D}\left(\mu^2 + \frac{r_2}{2}e^{-r_1 t} + \frac{r_1}{2}e^{-r_2 t}\right),$$
$$p_1(t) = \frac{\lambda}{\mu D}\left(\mu^2 + \frac{\lambda\big(\sqrt{\lambda\mu}-\mu^2\big)}{2}\mathrm{e}^{-r_1 t} - \frac{\lambda\big(\sqrt{\lambda\mu}+\mu^2\big)}{2}\mathrm{e}^{-r_2 t}\right),$$
$$p_2(t) = \frac{\lambda}{\mu D}\left(\lambda\mu - \frac{r_2\sqrt{\lambda\mu}}{2}e^{-r_1 t} + \frac{r_1\sqrt{\lambda\mu}}{2}e^{-r_2 t}\right),$$
where $0 < r_1 = \lambda+\mu-\sqrt{\lambda\mu} < r_2 = \lambda+\mu+\sqrt{\lambda\mu}$ and $D = r_1 r_2 = \lambda^2+\lambda\mu+\mu^2$.

In the $M/M/1/H$ model we deal with a finite Markov process, which suggests to try spectral methods to analyze transients. This has of course been done; we will illustrate the approach following the well-known book by Takács [43], for the historical importance of his contribution. The idea is simply to find the eigenvalues of the transition matrix of the process, having size $H+1$. The $H+1$ eigenvalues values are

$$\{0\} \cup \Big\{2\sqrt{\lambda\mu}\cos(\frac{h\pi}{H+1}) - (\lambda+\mu), \quad h = 1,\dots,H\Big\}.$$

Then, the two matrices B and B^{-1} such that $A = BDB^{-1}$ with D diagonal, are computed, and finally, e^{At} is obtained through the expression $B\mathrm{e}^{Dt}B^{-1}$. In scalar form, the given expression of the transient distribution is the following one (as usual, starting with the queue empty at time 0): if $\varrho \neq 1$,

$$p_j(t) = \frac{1-\varrho}{1-\varrho^{H+1}}\varrho^j - \frac{2\varrho^{j/2}}{H+1}\sum_{1\le h\le H}\Bigg\{\frac{\mathrm{e}^{-\big(\lambda+\mu-2\sqrt{\lambda\mu}\cos(\frac{h\pi}{H+1})\big)t}}{1+\varrho-2\sqrt{\varrho}\cos(\frac{h\pi}{H+1})} \times \\ \times\sqrt{\varrho}\sin(\frac{h\pi}{H+1})\Big[\sin(\frac{jh\pi}{H+1}) - \sqrt{\varrho}\sin(\frac{(j+1)h\pi}{H+1})\Big]\Bigg\}, \quad (3.5)$$

and if $\varrho = 1$,

$$p_j(t) = \frac{1}{H+1} - \frac{1}{H+1} \sum_{1 \le h \le H} \left\{ \frac{e^{-2\lambda(1-t)\cos\left(\frac{h\pi}{H+1}\right)}}{1 - \cos\left(\frac{h\pi}{H+1}\right)} \times \right.$$
$$\left. \times \sin\left(\frac{h\pi}{H+1}\right) \left[\sin\left(\frac{jh\pi}{H+1}\right) - \sin\left(\frac{(j+1)h\pi}{H+1}\right) \right] \right\}. \quad (3.6)$$

It is worth mentioning that in another well known book, [33], the same expression is obtained following a slightly different approach, based on trigonometric representations.

The $M/M/1$ as the limit of the $M/M/1/H$ when $H \to \infty$. In [43], Takács obtains also the transient distribution of the unbounded model, that is, the $M/M/1$ queue, basically by making $H \to \infty$ in (3.5) and (3.6). The obtained representation is as follows: for any state $j \in \mathbb{N}$ and starting from the empty queue at time 0,

$$p_j(t) = (1-\varrho)\varrho^j 1\big(\varrho < 1\big) - R_j(t), \quad (3.7)$$

with $1(C) = 1$ iff the condition or predicate C is true, 0 otherwise, and where the transient part $R_j(t)$ has the integral representation

$$R_j(t) = \frac{2e^{-(\lambda+\mu)t}\varrho^{\frac{j}{2}}}{\pi} \int_0^\pi \frac{e^{2t\sqrt{\lambda\mu}\cos(y)}}{1 - 2\sqrt{\lambda\mu}\cos(y) + \varrho} \sqrt{\varrho}\sin(y) \Big[\sin(jy) - \sqrt{\varrho}\sin((j+1)y)\Big] \, \mathrm{d}y. \quad (3.8)$$

3.6 Comments

As said before, a survey-oriented paper on a topic such as this one implies choices, given the impressive amount of publications. Let us add here a few more references, to diminish the set of forgotten papers and works.

Concerning the $M/M/1$ model, in [45] we find integral representations of the first two moments of the basic occupation process at t, with a correction of the formulas given in [43], correction provided by Takács himself to the author. In the 2-pages paper [34] the author derives an expression of the transient distribution in terms of the classic modified Bessel functions of the first kind, but in a very direct way, using a smart change of variable and generating functions.

Concerning the spectral approach in the analysis of the bounded $M/M/1/H$ case, several variants and improvements have been proposed by different authors. See, for instance, the books [12], [6], [40] and the many references therein.

The list of papers working on different aspects of the transient behavior of the $M/M/1$ model and variants is huge, so, it is hard to get close to exhaustivity. Let us just complete a list of book references with abundant material on the topic: we already mentioned [33], [43], [3], or [12], for instance, but we can also add [4], [36], [39], [23], Chapters 5 and 6 in [37], etc.

Considering the model presented in Subsection 3.3, there are many papers dealing with different variants and/or generalizations, some of them dedicated to transient behaviors. For instance, [42], [15]. In Subsection 5.4, the $M/M/1/H$ model with catastrophes is considered.

4 An uniformization-based path for the $M/M/1$ with matrix generating functions

In this section we describe a path we followed in part of our past work for finding the transient distribution of the $M/M/1$ model, published in [30]. The starting point here is uniformization, to move the problem to discrete time. In the paper we reach the following expression for the transient distribution of the $M/M/1$ model, when we start with an empty system: for any $j \in \mathbb{N}$,

$$p_j(t) = \varrho^j \sum_{n \geq j} e^{-(\lambda+\mu)t} \frac{(\lambda+\mu)^n t^n}{n!} C_{n,j}, \tag{4.1}$$

where the coefficient $C_{n,j}$ represents the following *finite* sum:

$$C_{n,j} = \sum_{k=0}^{\left\lfloor \frac{n-j}{2} \right\rfloor} \frac{n+1-2k}{n+1} \binom{n+1}{k} p^k q^{n-k}, \tag{4.2}$$

with $p = \lambda/(\lambda+\mu)$, $q = \mu/(\lambda+\mu) = 1-p$ and $\varrho = \lambda/\mu = p/q$.

This is a typical uniformization-based representation, coming with the associated powerful benefits: a probabilistic interpretation (the number $C_{n,j}$ is the probability that the uniformized discrete time Markov chain is at state j at time n, see Figure 3), a numerically stable procedure (there are only sums and products of positive numbers, furthermore bounded since we basically deal with probabilities) and, for practical purposes, a before-hand error bound on the computation (see Subsection 2.2).

Again, compare (4.2) with the classic expression (3.3) given in terms of Bessel functions, or with (3.4), or with the pair (3.7) and (3.8). The other expressions don't offer the nice properties mentioned in a general way in Remark 2.1.

The approach of [30] goes through the following steps. First of all, the development is done in the algebra of infinite matrices equipped with the norm $||M||_\infty = \sup_i \sum_j |M_{i,j}|$, indexed on $\mathbb{N}$. The set of infinite matrices having a finite norm is a

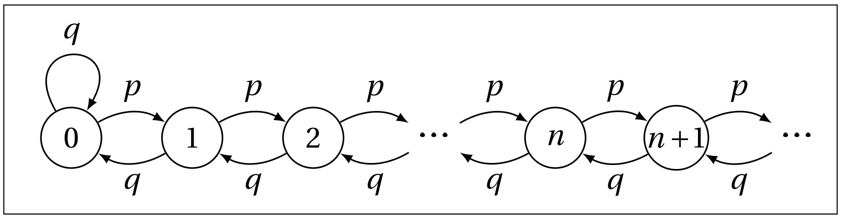

Figure 3: Uniformization of the $M/M/1$ canonical process w.r.t. the uniformization rate $\lambda + \mu$; $p = \lambda/(\lambda+\mu)$ and $q = 1 - p = \mu/(\lambda+\mu)$.

Banach non-commutative algebra, and for any matrix M in this space, its powers M^n, $n \in \mathbb{N}$, and its exponential e^M are well-defined and belong to the same space $\mathscr{M}$. Following the uniformization approach, we reduce the evaluation of $P(t) = \mathrm{e}^{At}$ to that of U^n, where $U = I + A/(\lambda+\mu)$. This evaluation is done through the analysis of the matrix generating function of the sequence of powers of U, exploiting the properties of $\mathscr{M}$ and a sort of product-form appearing in one of the key matrices of the analysis. The matrix generating function is obtained in closed-form, leading immediately to the elements of U^n also in closed-form, first for the case of $i = 0$.

General case. The presentation here completes that of [30]. First of all, for any two integers i, j in the model's state space (this is valid for the $M/M/1$ and for the $M/M/1/H$ models) and any time $t \in \mathbb{R}_{\geq 0}$, we have

$$P_{i,j}(t) = \varrho^{j-i} P_{j,i}(t). \tag{4.3}$$

This is an immediate consequence of reversibility, valid even in the unstable $\varrho \geq 1$ case (see Relation (10) in [1]). In particular, it means that we can limit ourselves to exhibiting the transition function $P_{i,j}(t)$ only for $i \leq j$.

In [30], U is written as the sum $U = pR + qL + qK$, where matrix R is the right-shift operator $R_{i,j} = I_{i+1,j}$, matrix L is the left-shift operator $L_{i,j} = I_{i,j+1}$ and K has only one non-zero element, its corner: $K_{0,0} = 1$ (thus, $K = I - L \cdot R$). Let us denote $\widehat{U} = pR + qL$. If we denote by $G_M(z) = \sum_{n \geq 0} M^n z^n$ the generating function of the sequence of powers of $M \in \mathscr{M}$ at point z, the paper shows ((16) in [30]) that

$$\big(G_U(z)\big)_{i,j} = \big(G_{\widehat{U}}(z)\big)_{i,j} + \frac{1}{\varrho^{i+1}} \big(G_U(z)\big)_{0,i+j+1}.$$

This means that, once the case of $P_{0,h}(t)$ solved, that is, once $\big(G_U(z)\big)_{0,i+j+1}$ known, all we need is to get a closed-form expression of $\big(G_{\widehat{U}}(z)\big)_{i,j}$. This is done in [30] at the end of Section 3.2, for the case of $i \leq j$. Make first the change of notation $j = i + d$, $d \geq 0$.

Then, the only non-null elements of the form $\left(\widehat{U}^m\right)_{i,i+d}$ are

$$\left(\widehat{U}^{2n+d}\right)_{i,j} = \begin{cases} p^{n+d}q^n\binom{2n+d}{n} & \text{if } n \le i, \\ p^{n+d}q^n\left[\binom{2n+d}{n} - \binom{2n+d}{n-i-1}\right] & \text{if } n \ge i+1. \end{cases} \tag{4.4}$$

Let us go here until the end of the process, in order to exhibit an explicit expression of the transient distribution of the model for any $i \le j$. Making the change of variable $2n + d = m$, then looking again for the non-zero elements, and going back to the coefficient $[z^n]G_{\widehat{U}}$ (the coefficient of z^n in the series expansion of $G_{\widehat{U}}$), we obtain, after some algebra, the following uniformization-based expressions, given separately for d even and d odd:

- even case: if $j - i = 2h$, $h \ge 0$,

$$\begin{aligned} P_{i,i+2h}(t) = & \sum_{\ell=h}^{h+i} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{2\ell}}{(2\ell)!}\binom{2\ell}{\ell-h}p^{\ell+h}q^{\ell-h} \\ & + \sum_{\ell\ge h+i+1} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{2\ell}}{(2\ell)!}\left[\binom{2\ell}{\ell-h} - \binom{2\ell}{\ell-h-i-1}\right]p^{\ell+h}q^{\ell-h} \\ & + \varrho^{i+2h}\sum_{n\ge 2i+2h+1} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{n}}{n!}C_{n,2i+2h+1}; \end{aligned} \tag{4.5}$$

- odd case: if $j - i = 2h + 1$, $h \ge 0$,

$$\begin{aligned} P_{i,i+2h+1}(t) = & \sum_{\ell=h}^{h+i} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{2\ell+1}}{(2\ell+1)!}\binom{2\ell+1}{\ell-h}p^{\ell+h}q^{\ell-h} \\ & + \sum_{\ell\ge h+i+1} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{2\ell+1}}{(2\ell+1)!}\left[\binom{2\ell+1}{\ell-h} - \binom{2\ell+1}{\ell-h-i-1}\right]p^{\ell+h+1}q^{\ell-h} \\ & + \varrho^{i+2h+1}\sum_{n\ge 2i+2h+2} \mathrm{e}^{-(\lambda+\mu)t}\frac{\left((\lambda+\mu)t\right)^{n}}{n!}C_{n,2i+2h+2}. \end{aligned} \tag{4.6}$$

Again, this is an uniformization-based expression, that is, a Poissonian sum (this is also called the Poisson generating function of the sequence $(U^n)_{i,i+d}$ in n, for fixed i, d, at point $(\lambda + \mu)t$, and it is also basically what is called an Exponential generating function – the factor $\mathrm{e}^{-(\lambda+\mu)t}$ must be removed to use this terminology). As such, all elements multiplied by the Poisson factor in the three sums is a probability, and everything said before about uniformization matters here.

4.1 Mean number of customers at time t in the $M/M/1$

Let us show how to use the obtained result to evaluate the mean number of customers in the $M/M/1$ system at time t. First of all, observe that from a general point of view, if we model a queue using an uniform Markov process $\{N(t), t \in \mathbb{R}_{\geq 0}\}$ where $N(t)$ is the number of customers in the queue at time t, then, from the general uniformization representation

$$p_j(t) = \sum_{n\geq 0} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} q_j(n),$$

we have

$$\begin{aligned}
\mathbb{E}(N(t)) &= \sum_{j\geq 1} j \sum_{n\geq 0} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} q_j(n) \\
&= \sum_{n\geq 0} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \sum_{j\geq 1} j q_j(n) \\
&= \sum_{n\geq 0} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \mathbb{E}(Y(n)).
\end{aligned}$$

This shows that we can use the uniformization approach to obtain $\mathbb{E}(N(t))$ for any model of this type (thus including all the $M/M/$ basic ones, for instance).

Coming now back at the specific case of the $M/M/1$ model, and starting from an empty queue, we saw that

$$\mathbb{P}(N(t) = j) = p_j(t) = \sum_{n\geq j} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \varrho^j c_{n,j},$$

where

$$c_{n,j} = \sum_{k=0}^{\lfloor \frac{n-j}{2} \rfloor} \varphi_{n,k} \qquad \text{and} \qquad \varphi_{n,k} = \frac{n+1-2k}{n+1} \binom{n+1}{k} p^k q^{n-k},$$

with $\Lambda = \lambda + \mu$, $p = \lambda/\Lambda$, $q = \mu/\Lambda = 1 - p$ and $\varrho = \lambda/\mu = p/q$.

Then,

$$\begin{aligned}
\mathbb{E}(N(t)) &= \sum_{j\geq 1} j p_j(t) \\
&= \sum_{j\geq 1} j\varrho^j \sum_{n\geq j} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} c_{n,j} \\
&= \sum_{n\geq 1} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} d_n,
\end{aligned}$$

where d_n is the finite sum

$$d_n = \sum_{1\leq j\leq n} j\varrho^j c_{n,j}.$$

We already saw that we can use this uniformization-based approach to numerically evaluate $\mathbb{E}(N(t))$ in a pretty general setting. Here, we can specialize the truncation level and the effective computations, by observing first that $\varrho^j c_{n,j} < 1$ allows to bound the number d_n by the number $n(n+1)/2$. Then, if we want to numerically evaluate $\mathbb{E}(N(t))$ by truncating previous series at level N, we observe that the absolute error err_N satisfies

$$err_N < \sum_{n>N} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \frac{n(n+1)}{2}.$$

Using the notation $\mathscr{P}_H(x) = \sum_{0\le h\le H} \mathrm{e}^{-x} x^h / h!$ introduced in Subsection 3.3, we can write, after some algebra,

$$\begin{aligned} \sum_{n>N} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \frac{n(n+1)}{2} &= \frac{(\Lambda t)^2}{2} \big[1 - \mathscr{P}_{N-2}(\Lambda t)\big] + \frac{3\Lambda t}{2} \big[1 - \mathscr{P}_{N-1}(\Lambda t)\big] \\ &< \big[1 - \mathscr{P}_{N-2}(\Lambda t)\big] \frac{\Lambda t(3 + \Lambda t)}{2}, \end{aligned}$$

which shows that the standard uniformization approach can be applied as for the probability distributions, given that $\mathscr{P}_\infty(x) = 1$, for instance using N^* defined as

$$N^* = \min\Big\{ N \in \mathbb{N} \,\Big|\, \mathscr{P}_{N-2}(\Lambda t) > 1 - \frac{2\varepsilon}{\Lambda t(3 + \Lambda t)} \Big\}.$$

Let us come back to the general expression of $\mathbb{E}(N(t))$ and to the sequence (d_n). After some algebra, we can write

$$\begin{aligned} d_n &= \sum_{1\le j\le n} j \varrho^j c_{n,j} \\ &= \sum_{1\le j\le n} j \varrho^j \sum_{k=0}^{\lfloor \frac{n-j}{2} \rfloor} \varphi_{n,k} \\ &= \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \varphi_{n,k} \sum_{j=1}^{n-2k} j \varrho^j \\ &= \frac{\varrho}{(1-\varrho)^2} \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \varphi_{n,k} \Big\{ 1 + \varrho^{n-2k} \big[(n-2k)\varrho - n + 2k - 1 \big] \Big\}. \end{aligned}$$

So, we finally obtain the expression

$$\mathbb{E}(N(t)) = \frac{\varrho}{(1-\varrho)^2} \sum_{n\ge1} \mathrm{e}^{-\Lambda t} \frac{(\Lambda t)^n}{n!} \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \varphi_{n,k} \Big\{ 1 + \varrho^{n-2k} \big[(n-2k)\varrho - n + 2k - 1 \big] \Big\}. \tag{4.7}$$

5 An uniformization-based path using duality

First of all, recall that based on uniformization, using the expression

$$P_{i,j}(t) = \sum_{n\geq 0} \mathrm{e}^{-(\lambda+\mu)t} \frac{\big((\lambda+\mu)t\big)^n}{n!} \big(U^n\big)_{i,j},$$

all we need to do is to compute $\big(U^n\big)_{i,j}$. This is a typical combinatorial problem.

Remark 5.1 *The point with computing $\big(U^n\big)_{i,j}$ using combinatorial techniques is the loop at state 0, which makes that different paths from i to j with the same length n may have different probabilities. This is the bottleneck of the direct combinatorial approach. In Section 4 we follow a generating function approach (the most important tool in combinatorics), working with infinite matrices, to find those probabilities. In this specific birth-death topology, the problem is greatly simplified by means of the duality concept (next subsection).*

Concerning the combinatorial side of the discrete time analysis mentioned here and in the sequel of the chapter, there is also an important literature. Related to our topic here, see for instance [7] and the references therein.

5.1 Duality

Let $\big(P_{i,j}(t)\big)_{i,j\in\mathbb{N}}$ be the transition function associated with some Markov process X on $\mathbb{N}$, that is, $P_{i,j}(t) = \mathbb{P}(X(t) = j \mid X(0) = i)$, also denoted $\mathbb{P}_i(X(t) = j)$. The transition function $\big(P_{i,j}(t)\big)_{i,j\in\mathbb{N}}$ is *stochastically increasing* iff for all $i, j \in \mathbb{N}$, for all $t \in \mathbb{R}_{\geq 0}$ and for all $k \in \mathbb{N}$, the inequality $\mathbb{P}_i(X(t) \geq k) \leq \mathbb{P}_{i+1}(X(t) \geq k)$ holds.

In [41] we have the following result: if $\big(P_{i,j}(t)\big)_{i,j\in\mathbb{N}}$ is stochastically increasing, then there exists another transition function $\big(P^*_{i,j}(t)\big)$ associated with another process X^* with values also on $\mathbb{N}$, possibly defective (see Remark 5.3), such that for all $i, j \in \mathbb{N}$ and $t \in \mathbb{R}_{\geq 0}$,

$$\mathbb{P}_i\big(X^*(t) \leq j\big) = \mathbb{P}_j\big(X(t) \geq i\big).^4 \tag{5.1}$$

Moreover, (i) P^* is also stochastically increasing and (ii) the reciprocal is true. We say that $\big(P^*_{i,j}(t)\big)$ (respectively X^*) is the Siegmund-dual of $P(t)$ (respectively of X). When this happens, we can represent the elements of P^* as linear functions of those of P, and vice versa, as follows:

- Given P,

$$P^*_{i,j}(t) = \sum_{k=0}^{i-1} \Big[P_{j-1,k}(t) - P_{j,k}(t)\Big], \tag{5.2}$$

[4] There are slight differences here with [3] in the indexing.

for $i, j = 0, 1, 2, \ldots$. This is [3, (4.2), page 251], with a slight change in notation. In the finite case where the state space of X has n states, say $\{0, 1, \ldots, n-1\}$, (5.2) holds for $j \leq n-1$; when $j = n$, we have $P^*_{i,n-1}(t) = 1 - \sum_{k=i}^{n-1} P_{n-1,k}(t)$. In this case, for the last row of P^* we have $P^*_{n,j}(t) = 0$ if $j < n$ and $P^*_{n,n}(t) = 1$.

- Given P^*,

$$P_{i,j}(t) = \sum_{k=0}^{i} \left[P^*_{j,k}(t) - P^*_{j+1,k}(t) \right]. \tag{5.3}$$

Remark 5.2 *This translates into the same relations but between the infinitesimal generators associated with the transition functions, say A and A^*, that is, replacing in previous expressions $P_{u,v}(t)$ by $A_{u,v}$ and $P^*_{u,v}(t)$ by $A^*_{u,v}$. The same also holds for discrete time chains; in that case, we write $\left(P^n\right)_{u,v}$ in place of $P_{u,v}(t)$ and $\left(P^{*n}\right)_{u,v}$ in place of $P^*_{u,v}(t)$ (see [3]).*

Finally, observe that Relations (5.2) and (5.3) can be also written

$$\mathbb{P}_i(X^*(t) = j) = \mathbb{P}_j(X(t) \leq i) - \mathbb{P}_{j+1}(X(t) \leq i), \tag{5.4}$$

$$\mathbb{P}_i(X(t) = j) = \mathbb{P}_{j-1}(X^*(t) \leq i-1) - \mathbb{P}_j(X^*(t) \leq i-1). \tag{5.5}$$

In [3] this concept of duality is developed and exploited, in particular for birth and death processes, for which the dual always exist, being moreover particularly simple. Graphically, the dual of the standard birth-death process with birth rates λ_i, $i \geq 0$ and μ_j, $j \geq 1$ is depicted in Figure 4. See that the Siegmund-dual of the original process has, at least, one absorbing state.

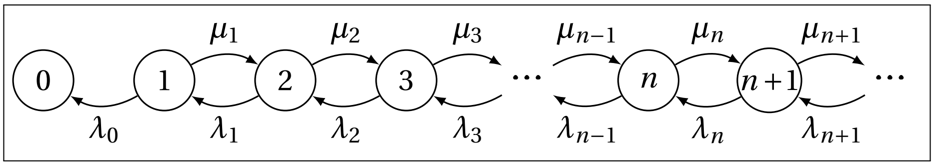

Figure 4: The Siegmund-dual of the standard birth-death process with birth rates λ_i, $i \geq 0$ and μ_j, $j \geq 1$.

Remark 5.3 *There is no room here to provide technical details about the fact that the dual may be a defective process, the meaning of this, and about the nature of the dual of the dual. The reader is sent to [3] for more information about these points. At this stage, and as far as we are concerned by $M/M/*$ queues, we can take the transformation given in Figure 4 as the definition of the dual of a general birth-death process, and Relations (5.2) and (5.3) (or (5.4) and (5.5)) as the main properties of the transformation.*

To complete the introduction of the duality concept, let us show an example using a small Markov process that is not in the birth-death family.

Remark 5.4 *As it should be clear from the definition, the dual of a process doesn't exist in general, because this needs that specific monotonicity properties must be satisfied. This is illustrated in the following example (taken from [17]), where on the left side of Figure 5 we depict the general 3-states Markov process. Its dual exists under specific conditions on the transition rates given in the figure.*

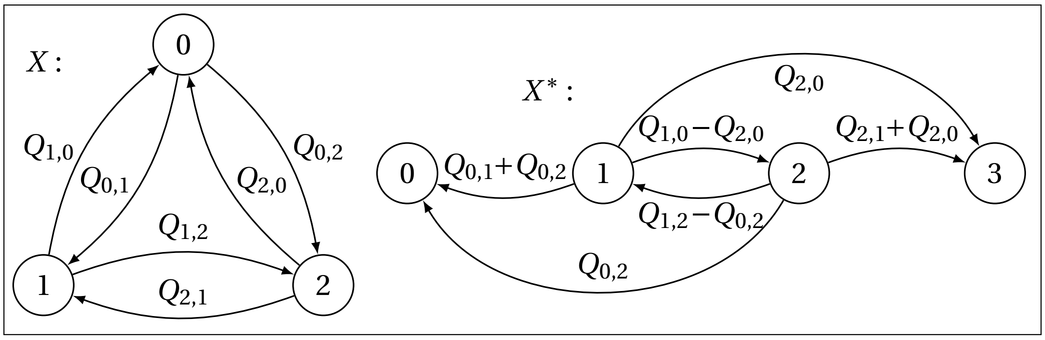

Figure 5: This is a pair (X, X^*) where the dual exists under specific conditions. On the left, the general 3-states process. On the right, its dual, that exists iff $Q_{2,0} < Q_{1,0}$ and $Q_{0,2} < Q_{1,2}$. Observe that the existence conditions depend on the numbering of the nodes.

As a particular case of Figure 5, an example in Figure 6 where the dual never exists.

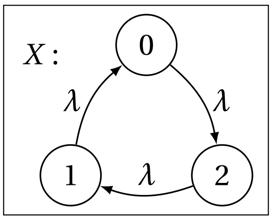

Figure 6: Process X has no dual, whatever the value of λ (and, obviously, whatever the numbering of the states).

The path towards the transient state distributions using duality. The idea is to start by the uniformization approach, that is, from Relations (2.4) or (2.3). We have model X, and we move to the uniformized discrete time Markov chain Z, whose transition probability matrix is U. Then, we build its Siegmond-dual Z^*, with transition probability matrix U^*, we compute the probabilities $\left(U^{*n}\right)_{i,j}$ by any method, and we obtain $\left(U^n\right)_{i,j}$ using (5.3) (recall that the linear relationships between X and X^* are the same in continuous and in discrete time, as developed in Remark 5.2), that is, using

$$\left(U^n\right)_{i,j} = \sum_{k=0}^{i} \left[\left(U^{*n}\right)_{j,k} - \left(U^{*n}\right)_{j+1,k}\right]. \tag{5.6}$$

This path is useful if evaluating $\left(U^{*n}\right)_{i,j}$, for instance using combinatorial techniques, is easier than directly attacking the evaluation of $\left(U^n\right)_{i,j}$. A good example of this situation is the case of the $M/M/1$ model (see next Subsection). The method has allowed to obtain formulas (closed-forms) in cases where this was considered a non-trivial task in the literature.

Before going to some examples, a remark underlining the fact that we can uniformize first and then go to the dual, or first compute the dual and then uniformize. That is, we can follow the path $X \to Z \to Z^*$ or $X \to X^*$ first, and then, uniformize the latter. The result is again Z^*.

Remark 5.5 *The operators "uniformization" and "dual" commute. This is immediate from the definitions:*

$$\begin{tikzcd} X \arrow[r, "uniformization"] \arrow[d, "dual"] & Z \arrow[d, "dual"] \\ X^* \arrow[r, "uniformization"'] & Z^* \end{tikzcd}$$

Let us now briefly describe this Siegmund dual-based approach in some fundamental cases. The goal is to provide details on the corresponding solving processes.

5.2 Application to the $M/M/1$ queueing system

You want to derive $p_j(t)$, the probability that there are j customers in an $M/M/1$ model at time t. As usual, to simplify the description, we assume that the $M/M/1$ is empty at time 0 ($X(0) = 0$). We then have

$$p_j(t) = \sum_{n=j}^{\infty} \mathrm{e}^{-(\lambda+\mu)t} \frac{(\lambda+\mu)^n t^n}{n!} (U^n)_{0,j},$$

where U is the transition probability matrix of the uniformization Z of X, whose graph is given in Figure 3.

Now, the dual of Z is Z^*, whose graph is shown in Figure 7. Given the homogeneity in the structure of this last model and the fact that the process is absorbing, with

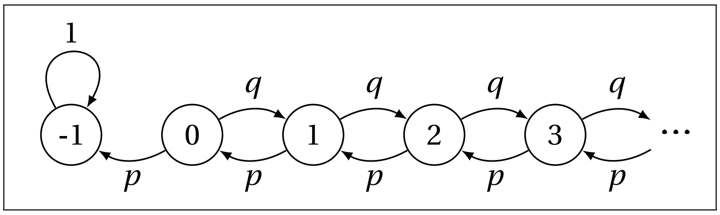

Figure 7: The dual Z^* of the uniformization Z of the $M/M/1$ process, which is also the uniformization of the dual of the $M/M/1$ process with respect to the same uniformization rate $\lambda+\mu$.

no loop in its transient class, the problem finally reduces to particularly simple path counting. Standard lattice path combinatorial techniques can then be used for this task. From the general formula connecting a process and its dual, and more precisely, using Remark 5.2 and Relation (5.6) for $i=0$, knowing the transient distribution of Z^* we recover that of Z through

$$\big(U^n\big)_{0,j} = \Big(U^{*n}\Big)_{j,0} - \Big(U^{*n}\Big)_{j+1,}, \tag{5.7}$$

where U^* is the transition probability matrix of chain Z^*. Obtaining the probabilities $\big(U^{*n}\big)_{\ell,0}$ is an immediate application of the well-known *reflection principle* in the analysis of random walks (see for instance [16]; see also [44]). This is the path we followed in [26]; we refer the reader to this paper for the details, and for the references therein. The symbolic expressions we get are a bit more complex than those obtained in [30], for instance than (4.1), but of course, they are equivalent, and it is a matter of simple combinatorial simplifications to verify it.

5.3 Application to the $M/M/1/H$ queueing system

The application of the same procedure to this finite case is also done in [26], and it needs more involved developments, but the global path is the same. In the following sequence of pictures, we show the graphs of the $M/M/1/H$ model (Figure 8), its uniformization with respect to the rate $\lambda+\mu$ (Figure 9), its dual process (Figure 10), and the uniformization of the dual (Figure 11), which is also the dual of the uniformized chain.

As a check, the reader is invited to verify the duality connexions given in Relations (5.2) and (5.3) on the simplest case of $H=1$, that is, on the smallest and non-trivial irreducible process with two states. In Figure 12 we depict the process and its corresponding dual, having a single transient state ‘1’ and two absorbing states ‘0’

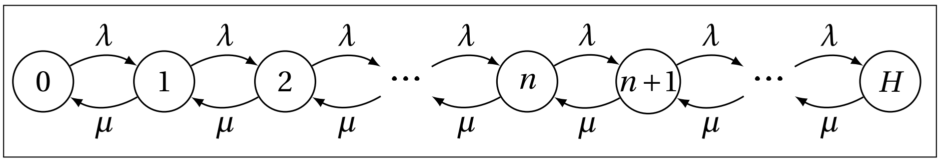

Figure 8: The $M/M/1/H$ model, parameters λ and μ.

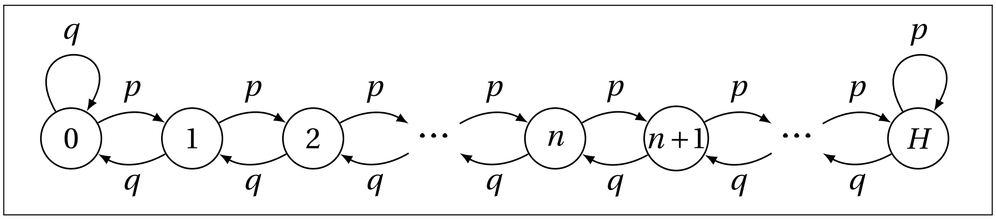

Figure 9: The uniformized chain of the $M/M/1/H$ canonical process depicted in Figure 8, with uniformization rate $\Lambda = \lambda + \mu$, $p = \lambda/\Lambda$ and $q = \mu/\Lambda = 1 - p$.

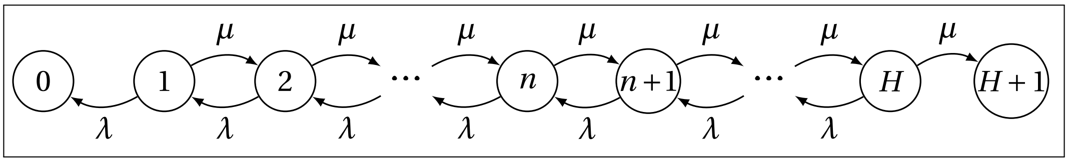

Figure 10: The dual process of the $M/M/1/H$ given in Figure 8.

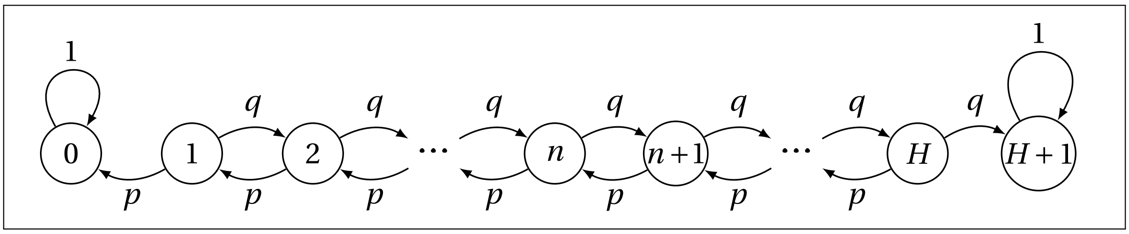

Figure 11: The dual of the uniformized chain shown in Figure 9, which is also the uniformization of the process shown in Figure 10 (with respect to the same uniformization rate $\lambda + \mu$), as explained in Remark 5.5.

and ‘2’. On process X, we have

$$P_{0,0}(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)t}.$$

On X^*,

$$P^*_{1,0}(t) = \frac{\lambda}{\lambda+\mu}\left(1 - e^{-(\lambda+\mu)t}\right).$$

The connection between X and X^* translates here into the relation $P_{0,0}(t) + P^*_{1,0}(t) = 1$, immediate to check.

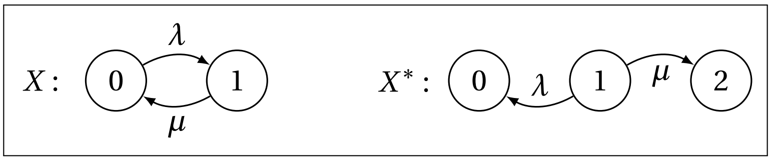

Figure 12: The smallest non-trivial irreducible Markov process with 2 states (left), and its dual (right); $X(0) = 0$ and $X^*(0) = 1$.

For the analysis of the $M/M/1/H$ using duality, we need a generalization of the reflection principle, because in this case, we have two boundaries instead of a single one in the $M/M/1$ situation. The other needed tool to simplify the combinatorial expressions that appear for this model is a technical trick given as Lemma 1 in [26]. It says that if we define

$$\binom{u}{v}_+ = \begin{cases} \binom{u}{v} & \text{if } u \ge v, \\ 0 & \text{if } v < 0 \text{ or } v > u, \end{cases}$$

then,

$$\sum_{g:\ g=a \bmod m} \binom{\ell}{g} = \frac{1}{m} \sum_{u=1}^{m} \omega_m^{-ua} (1 + \omega_m^u)^\ell,$$

where positive integers a, m, ℓ satisfy $a < m$ and where $\omega_m = e^{2\pi i/m}$ is the mth root of unity.

The obtained form of the transient distribution of this model is trigonometric, formally similar to the one mentioned before, coming back to Takács [43] or Morse [33]. Using basic trigonometric identities and some algebra, we can go from one to the other (see [26] and the references therein).

5.4 Application to an $M/M/1/H$ model with catastrophes

As a last example, consider the $M/M/1/H$ model where the server is subject to breakdowns that happen with rate γ. When a breakdown occurs, the system is emptied, that is, the Markov process jumps to state 0. This is another model where the use of duality allows to derive the transient state distribution (see [26]). The model is depicted in Figure 13.

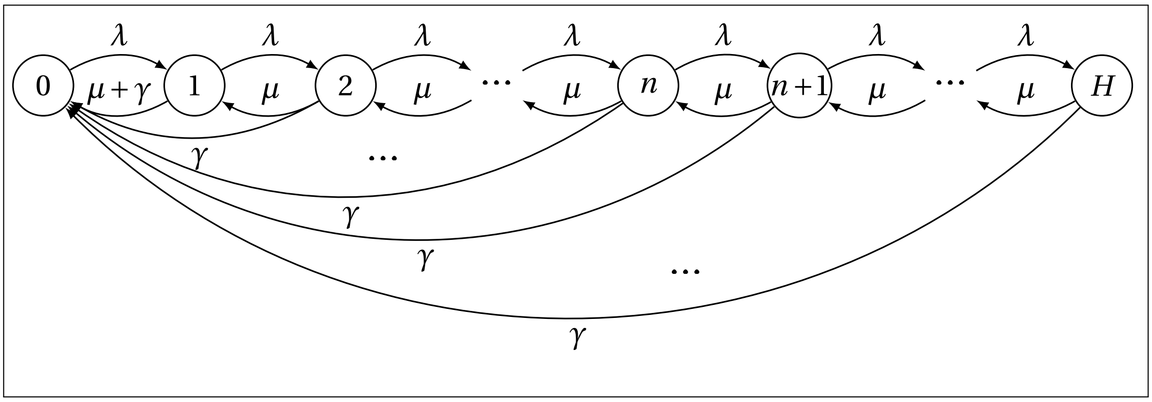

Figure 13: The $M/M/1/H$ model with catastrophes; parameters: arrival rate λ, service rate μ and catastrophe rate γ.

The dual of this process is given in Figure 14. As for the $M/M/1/H$ model, the dual has two absorbing states. The uniformization of the dual is depicted in Figure 15.

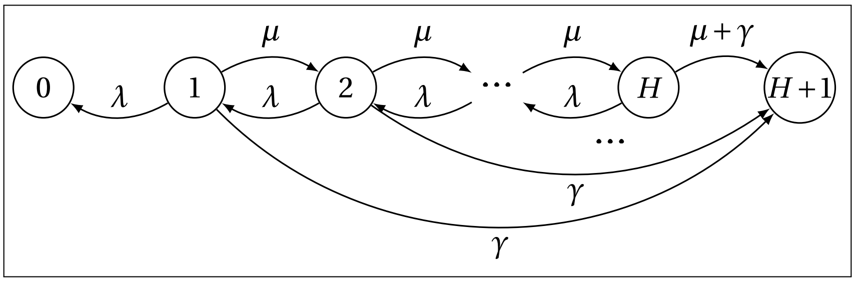

Figure 14: The dual of the $M/M/1/H$ model with catastrophes depicted in Figure 13.

The derivation process follows similar lines as for the $M/M/1/H$ model. Starting with an empty system, the transient distribution of the model is given by the following expression:

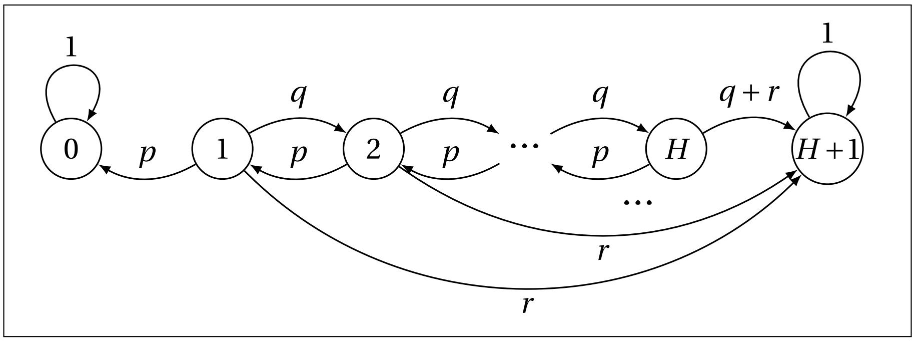

Figure 15: The uniformization of the dual of the $M/M/1/H$ model with catastrophes. See Figure 13 for the model and Figure 14 for its dual. The uniformization rate is $\Lambda = \lambda + \mu + \gamma$ and the notation is $p = \lambda/\Lambda$, $q = \mu/\Lambda$ and $r = \gamma/\Lambda$. Recall that this model is also the dual of the uniformization of the initial model (using obviously the same uniformization rate).

$$p_j(t) = \pi_j - \frac{2\mu\varrho^{j/2}}{H+1} \sum_{1\le h\le H} \left\{ \frac{\mathrm{e}^{-\left(\lambda+\mu+\gamma-2\sqrt{\lambda\mu}\cos(\frac{h\pi}{H+1})\right)t}}{\lambda+\mu+\gamma-2\sqrt{\lambda\mu}\cos(\frac{h\pi}{H+1})} \times \right.$$
$$\left. \times \sqrt{\varrho}\sin(\frac{h\pi}{H+1})\Big[\sin(\frac{jh\pi}{H+1}) - \sqrt{\varrho}\sin(\frac{(j+1)h\pi}{H+1})\Big]\right\}, \quad (5.8)$$

Compare this expression with the result for the $M/M/1/H$ given in (3.5). See the details in the paper, as well as the general $P_{i,j}(t)$ expression, the steady-state one $(\pi_j)_{j\ge 0}$, etc.

6 Other transient results

In this section, we briefly refer to other related developments.

6.1 Busy period of the $M/M/1$

The distribution of the busy period of a queue is another typical transient metric, even if the busy period can be extremely long depending on the values of λ and μ. First of all, it is well known from the analysis of the Markov process $(N(t))$ that the busy period is a.s. finite iff $\lambda \le \mu$. If $\lambda > \mu$, then the probability that the busy period is finite is μ/λ. Assuming $\lambda < \mu$, the case of a positive recurrent Markov process, and denoting by *BP* the busy period, we have, from our paper [30], the following closed-form of the cdf of *BP*:

$$\mathbb{P}(BP \le t) = \sum_{n\ge 1} \mathrm{e}^{-(\lambda+\mu)t} \frac{(\lambda+\mu)^n t^n}{n!} \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} c_k p^k q^{k+1}, \tag{6.1}$$

with c_k the kth Catalan number $c_k = \binom{2k}{k}/(k+1)$ and, as usual, $p = \lambda/(\lambda+\mu)$ and $q = \mu/(\lambda+\mu) = 1-p$.

Observe that the busy period is closely related to the absorption time of the dual of the $M/M/1$ depicted in Figure 4, replacing λ_i by λ and μ_j by μ. This provides another way of deriving (6.1); see also Section 3 in [25]. The busy period of the $M/M/1$ has obviously the same distribution than the *congestion period* over level $\ell \ge 1$, the time spent by the process above $\ell \ge 1$ (that is, on the set of states $\{j \in \mathbb{N} : \ j \ge \ell\}$, another interesting transient metric. Of course, this distribution is different and dependent on ℓ if we consider other processes such that other Markovian queues, including the bounded $M/M/1/H$ one. We omit, for keeping space moderate, developing this point further.

6.2 Max backlog of the $M/M/1$ over a finite time interval

A fine tuning transient metric for a queueing system is given by the random variable defined at a fixed time t

$$M(t) = \max\{N(s),\ s \le t\},$$

that is, the maximum level reached by the number of customers in the queue on the finite interval $[0, t]$. This process was analyzed for the $M/M/1$ model in [38] using uniformization. The result is a numerical scheme allowing to evaluate the distribution of $M(t)$. The idea is to work on the auxiliary process $Y = Y(t)$ where $Y(t) = \big(N(t), M(t)\big)$, depicted in Figure 16. To do so, we derive an evaluation scheme using the uniformization of this process (Figure 17) and then, we exploit the regularities that it exhibits. Again, uniformization is the key for the analysis. See a numerical example in Figure 18, for a heavy-loaded $M/M/1$ with load $\varrho = 0.95$.

6.3 $M/E/1$

To give an example of transient analysis of queueing systems leading to closed-forms and going beyond $M/M/$ structures, let us consider the case of Erlangian services, that is, of the model $M/E/1$. The arrival rate is, as usual, λ, and the service distribution has K phases each with rate $K\mu$. The approach followed in [19] is based on generating functions and an extension to the modified Bessel functions of the second kind

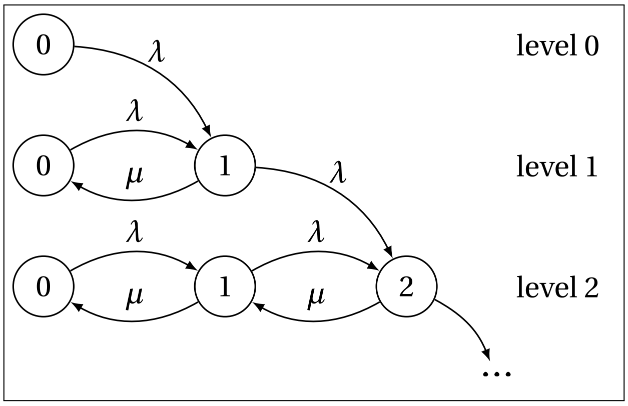

Figure 16: Auxiliary 2-dimensional process Y.

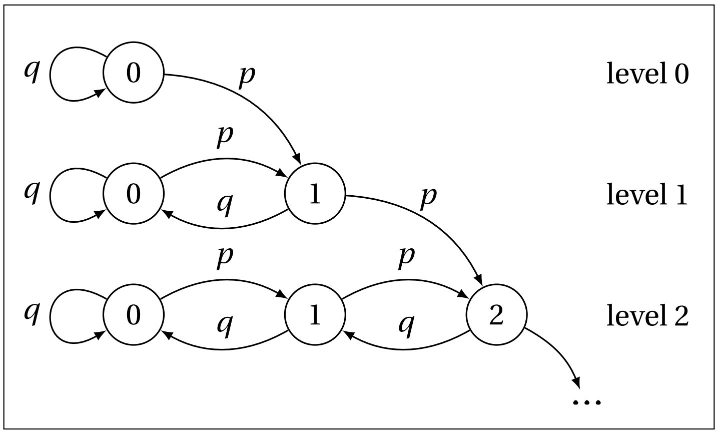

Figure 17: The uniformization of process Y depicted in Figure 16 with respect to $\lambda + \mu$, with $p = \lambda/(\lambda + \mu)$ and $q = 1 - p = \mu/(\lambda + \mu)$.

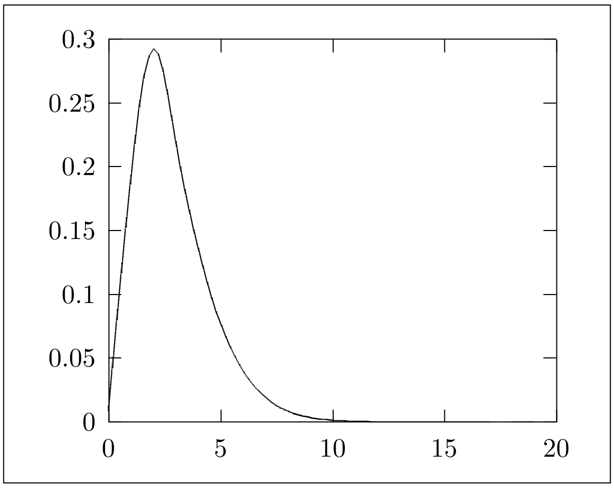

Figure 18: A numerical example where $\lambda = 0.95$ and $\mu = 1$; we plot $\mathbb{P}(M(5) = \ell)$, for level ℓ from 0 to 20 (the curve appears as "continuous" just for a better presentation)

(see [18]), defined by

$$\widetilde{I}_n^{u,v} = \left(\frac{z}{2}\right)^{n+u-v} \sum_{r=0}^{\infty} \frac{(z/2)^{r(u+1)}}{\big(u(r+1)-v\big)!\,\Gamma(n+r+1)}, \tag{6.2}$$

where $z \in \mathbb{C}$, $n \in \mathbb{Z}$, $u \in \mathbb{N}_{>0}$, $v \in \{1,2,\dots,u\}$, and $\Gamma()$ is the classic Gamma function. In the paper they use another extension to modified Bessel functions, of the first kind now, proposed in [31], defined by

$$I_n^u = \left(\frac{z}{2}\right)^n \sum_{r=0}^{\infty} \frac{(z/2)^{r(u+1)}}{r!\,\Gamma(n+ru+1)}, \tag{6.3}$$

where $z \in \mathbb{C}$, $n \in \mathbb{N}$ and $u \in \mathbb{N}_{>0}$.

The analysis is on the queue initially empty, and leads to closed-forms for the function $p_{j,s}(t)$ where j is the number of customers in the system and s is the server's phase, at time t, for $j \geq 1$, plus function $p_0(t)$ for the empty queue at time t. See [19] for the details. The general case (system not necessarily empty at $t = 0$) is a non-trivial extension of the one described here; it has been discussed in [20].

7 Conclusions

This chapter discussed the main efforts done to find closed-forms representations of the transient state distributions of some basic Markovian queueing models. Many approaches have been followed since the first results in this direction obtained in the 50s: combinatorial (including the use of transforms), based on spectral analysis, etc. The focus of the chapter is on the uniformization method that translates the continuous time problem into a discrete time one, a fruitful idea offering several nice properties and dealing to new expressions, that we believe have advantages over traditional ones. In some cases where closed-forms have not been obtained yet, this path leads to efficient numerical procedures. In the chapter another tool that we call *duality* or *Sigmund duality* is enphasized; it changes the topology of the Markovian target, and in some cases this greatly simplifies the analysis.

On a topic like this one, it is almost sure that there are more papers not cited here and relevant for the analyzed problems than those appearing referenced in the text. But remember we are biased by a few central ideas, which made that some very interesting works are not mentioned. Sorry for this to the concerned authors.

References

[1] J. Abate, M. Kijima, and W. Whitt. Decompositions of the $M/M/1$ transition function. *Queueing Systems*, 9:323–336, 1991.

[2] J. Abate and W. Whitt. Calculating time dependent performance measures for the $M/M/1$ queue. *IEEE Transactions on Communications*, 37(10):1102–1104, 1989.

[3] W.J. Anderson. *Continuous-time Markov chains: an applications-oriented approach.* Springer, New York, 1991.

[4] N. Bailey. *The Elements of Stochastic Processes.* John Wilet & Sons, Inc., 1964.

[5] N. T. J. Bailey. A continuous time treatment of a single queue using generating functions. *Journal of the Royal Statistical Society: Series B*, 16(12):288–291, 1954.

[6] R.N. Bhattacharya and E.C. Waymire. *Stochastic Processes with Applications.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2009.

[7] W. Böhm, A. Krinik, and S.G. Mohanty. The combinatorics of birth-death processes and applications to queues. *Queueing Systems*, 26(3-4):255–267, 1997.

[8] Richard J. Boucherie. A note on the transient behavior of the Engset loss model. *Communications in Statistics. Stochastic Models*, 9(1):145–156, 1993.

[9] Richard J. Boucherie and P. G. Taylor. Transient product from distributions in queueing networks. *Discrete Event Dynamic Systems*, 3(4):375–396, Sep 1993.

[10] D. G. Champernowne. An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society. Series B*, 3(3):125–128, 1956.

[11] A. B. Clarke. The time dependent waiting line problem. *University of Michigan: Ann. Arbor Report*, M720–1R39, 1953.

[12] J. W. Cohen. *The Single Server Queue*. North-Holland, 2 sub edition, 1982.

[13] B. W. Conolly and C. Langaris. On a new formula for the transient state probabilities for $M/M/1$ queues and computational implications. *Management Science*, 30(1):237–246, 1993.

[14] E. de Souza e Silva and R. Gail. The uniformization method in performability analysis. In B. Haverkort, R. Marie, G. Rubino, and K. Trivedi, editors, *Performability Modelling: Techniques and Tools*, pages 31–58. John Wiley &amp; Sons, 2001.

[15] A. Economou and D. Fakinos. Alternative Approaches for the Transient Analysis of Markov Chains with Catastrophes. *Journal of Statistical Theory and Practice*, 2(2):183–197, Jun 2008.

[16] W. Feller. *An Introduction of Probability Theory and its Applications, Vol. I.* Wiley, New York, 1968.

[17] M. L. Green, A. Krinik, C. Mortensen, G. Rubino, and R. J. Swift. Transient probability functions: A sample path approach. *Discrete Mathematics and Theoretical Computer Science*, pages 127–136, 2003.

[18] J. D. Griffiths, G. M. Leonenko, and J. E. Williams. New generalization of the modified Bessel function and its generating function. *Fractional Calculus and Applied Analysis*, 8(3):267–276, 2006.

[19] J. D. Griffiths, G. M. Leonenko, and J. E. Williams. The transient solution to $M/E_k/1$ queue. *Operations Research Letters*, 34:349–354, 2006.

[20] J. D. Griffiths, G. M. Leonenko, and J. E. Williams. Time-dependent analysis of non-empty $M/E_k/1$ queue. *Quality Technology & Quantitative Management*, 5(3):309–320, 2008.

[21] D. Gross, C. M. Harris, J. F. Shortle, and J. M. Thompson. *Fundamentals of queueing theory.* Wiley series in probability and statistics. John Wiley & Sons, 5th ed. edition, 2018.

[22] A. Jensen. Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal*, 1953(suo1):87–91, 1953.

[23] S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, 1981.

[24] M. Kijima. *Markov processes for stochastic modeling.* Chapman & Hall, 1997.

[25] A. Krinik, C. Mortensen, and G. Rubino. Connections between birth-death processes. In Alan C. Krinik and Randall J. Swift, editors, *Stochastic Processes and Functional Analysis*, pages 219–240. Marcel Dekker, 2004.

[26] A. Krinik, G. Rubino, D. Marcus, R. J. Swift, H. Kasfy, and H. Lam. Dual processes to solve single server systems. *Journal of Statistical Planning and Inference*, 135(1):702–713, 2005.

[27] Vidyadhar G. Kulkarni. *Modeling and analysis of stochastic systems.* Texts in statistical science. Chapman and Hall/CRC, 3rd ed. edition, 2017.

[28] M. Lebah and J. Pellaumail. Transient behavior for some jackson networks. *Performance Evaluation*, 17(2):115–122, 1993.

[29] W. Ledermann and G. E. Reuter. Spectral theory for the differential equations of simple birth and death process. *Philosophical Transactions of the Royal Society of London Series A*, 246(914):321–369, 1954.

[30] P. Leguesdron, J. Pellaumail, G. Rubino, and B. Sericola. Transient analysis of the $M/M/1$ queue. *Advances in Applied Probability*, 25(3):702–713, 1993.

[31] G. Luchak. The solution of the single-channel queuing equations characterized by a time-dependent poisson-distributed arrival rate and a general class of holding times. *Operations Research*, 4(6):711–732, 1956.

[32] P. Morse. Stochastic properties of waiting lines. *Journal of the Operations Research Society of America*, 3(3):255–261, 1955.

[33] P. Morse. *Queues, Inventories and Maintenance.* Wiley, New York, 1958.

[34] P. R. Parthasarathy. A transient solution to an $M/M/1$ queue: a simple approach. *Adv. Appl. Prob*, 19:997–998, 1987.

[35] C. D. Pegden and M. Rosenshine. Some new results for the $M/M/1$ queue. *Management Science*, 28:821–828, 1982.

[36] N.U. Prabhu. *Queues and Inventories.* Wiley, New York, 1965.

[37] Philippe Robert. *Stochastic networks and queues.* Applications of mathematics. Springer, Berlin, 2003.

[38] G. Rubino. Evaluation of the maximum level reached by a queue over a finite period. In *International Conference on Dependable Systems and Networks (DSN 2002), 23-26 June 2002, Bethesda, MD, USA, Proceedings*, pages 735–744, 2002.

[39] T.L. Saaty. *Elements of Queueing theory with applications.* Dover Publications Inc., 1983.

[40] O.P. Sharma. *Markovian queues.* Mathematics and its applications. Ellis Horwood, 1990.

[41] D. Siegmund. The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Ann. Prob.*, 6:914–924, 1976.

[42] R.J. Swift. Transient probabilities for a simple birth-death-immigration process under the influence of total catastrophes. *International Journal of Mathematics and Mathematical Sciences*, 25(10):689–692, 2001.

[43] L. Takács. *Introduction to the Theory of Queues.* Oxford University Press, New York, 1962.

[44] D. Towsley. An application of the reflection principle to the transient analysis of the $M/M/1$ queue. *Naval Research Logistics*, 34:451–456, 1987.

[45] M. C. T. van de Coevering. Computing transient performance measures for the $M/M/1$ queue. *OR Spektrum*, 17:19–22, 1995.

[46] P. Whittle. *Systems in stochastic equilibrium.* Probability and Mathematical Statistics. John Wiley & Sons, 1986.