Birth-Death Processes, Lattice Path Combinatorics, Continued Fractions, and Orthogonal Polynomials

Fabrice Guillemin
France Telecom, CNET, Lannuon, Breizh

February 2, 1998

[summary by Philippe Flajolet and Fabrice Guillemin]

Abstract

Classic works of Karlin-McGregor and Jones-Magnus have established a fully general correspondence between birth-death processes and continued fractions of the Stieltjes-Jacobi type together with their associated orthogonal polynomials. This fundamental correspondence can be revisited in the light of the otherwise known combinatorial correspondence between weighted lattice paths and continued fractions. For birth-death processes, this approach separates clearly the formal apparatus from the analytic-probabilistic machinery and neatly delineates those parameters that are amenable to a treatment by means of continued fractions and orthogonal polynomials.

1. Birth-Death Processes

Consider a particle initially in state 0 that, at any given time, may change to another state 1 (where it stays), with rate λ. This means that the probability of a state change in an interval of time of length dt is λdt. Then, the probability $p_0(t)$ that the particle is still in state 0 at time t satisfies

$$p_0(t+dt) - p_0(t) = -\lambda p_0(t) dt$$

or $p_0'(t) = -\lambda p_0(t)$, whose solution is an exponential distribution,

$$p_0(t) = e^{-\lambda t}.$$

Similarly, a particle initially in state 0 that may change either to state 1 with rate λ or to state -1 with rate μ will satisfy ($p_j(t)$ is the probability of being in state j at time t)

$$p_0(t) = e^{-(\lambda+\mu)t}, \qquad p_1(t) = \frac{\lambda}{\lambda+\mu}(1 - e^{-(\lambda+\mu)t}), \qquad p_{-1}(t) = \frac{\mu}{\lambda+\mu}(1 - e^{-(\lambda+\mu)t}).$$

The interpretation is obvious: the particle stays in state 0 for a random amount of time with an exponential distribution of rate $\lambda+\mu$ and then changes to states $-1, +1$ with probabilities equal to to $\lambda/(\lambda+\mu)$ and $\mu(\lambda+\mu)$.

In a general *birth-death* process a particle can be in any state in $\{0, 1, 2, \ldots\}$ and when in state j, it can only change to state $j+1$ at rate λ_j or to state $j-1$ at rate μ_j. By analogy with the model of an evolving population (whose size is represented by the state), the λ_j are called birth rates and the μ_j death rates. The general problem is to understand the evolution of a process given values (or properties) of its birth and death rates; see [12, Ch. 4] for an excellent introduction.

Let $p_n(t)$ be the probability of being in state n at time t. An essential rôle is played by the coefficients

$$\pi_n = \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}.$$

Indeed, a classical result asserts that the process is ergodic (the expected time to return from each state to itself is finite) if and only if

$$\sum_{n\geq 1}\pi_n < \infty, \qquad \sum_{n\geq 0}\frac{1}{\lambda_n\pi_n} = +\infty.$$

(The first condition ensures the existence of an invariant measure for the embedded discrete-time Markov chain; the second one guarantees that, in the continuous-time process, the particle is not absorbed at infinity in finite time.) In that case, one has

$$p_n := \lim_{t\to\infty} p_n(t) = \frac{\pi_n}{\sum_{n\geq 1}\pi_n},$$

where these quantities represent the long run probability of being in state n.

More puzzling is the nonstationary behaviour of the process that is described by the *infinite-dimensional* differential system

$$(1) \qquad p_j'(t) = \lambda_{j-1}p_{j-1}(t) - (\lambda_j+\mu_j)p_j(t) + \mu_{j+1}p_{j+1}(t), \qquad p_j(0) = \delta_{j,0}.$$

Although finite-dimensional versions are "easy" and reduce to combinations of exponentials, it is precisely the infinite-dimensional character of the system that renders its analysis interesting.

In a series of important papers, Karlin and McGregor [10, 11] have developed a general connection between the fundamental system (1) and an associated family of orthogonal polynomials. Later, Jones and Magnus constructed a direct continued fraction representation; see [8, 9].

This summary is an account of Guillemin's lecture (see [5, 6]), as well as of later developments. The point of view that is adopted here consists in relating the combinatorial theory of lattice paths to birth-death processes in the following way: (i) trajectories of birth-death processes are precisely lattice paths; (ii) lattice paths have generating functions expressed as continued fractions; (iii) the Laplace transform expresses the main parameters of birth-death processes as weighted lattice paths to which the combinatorial theory applies.

2. Lattice Paths and Continued Fractions

It is known that the formal theory of continued fraction expansions for power series is identical to the combinatorial theory of weighted lattice paths; see [1, 2, 4]. Define a path $v = (U_0, U_1, \ldots, U_n)$ to be a sequence of points in the lattice $\mathbb{N}\times\mathbb{N}$ such that if $U_j = (x_j, y_j)$, then $x_j = j$ and $|y_{j+1} - y_j| = 1$. If successive points are connected by edges, then an edge can only be an *ascent* ($\underline{a}$: $y_{j+1} - y_j = +1$), a *descent* ($\underline{b}$: $y_{j+1} - y_j = -1$), or a *level step* ($\underline{c}$: $y_{j+1} - y_j = 0$). Thus a path is always nonnegative and by a horizontal translation, one may always assume that $x_0 = 0$. A path can be encoded by a word with a, b, c representing the three types of steps. What we call the *standard encoding* is such a word in which each step a, b, c is subscripted by the value of the y-coordinate of its associated point. For instance,

$$w = a_0a_1a_2b_3c_2c_2a_2b_3b_2b_1a_0c_1$$

encodes a path that connects the source $U_0 = (0,0)$ to the destination $U_{12} = (12,1)$. We freely identify a path v defined as a sequence of points, its word encoding w, and the corresponding monomial.

We consider various geometric conditions that may be imposed on paths: $\mathcal{H}_{k,l}$ is the collection of all paths that connect a source at altitude k to a destination at altitude l, $\mathcal{H}^{[\le h]}$ denotes paths of height (maximal altitude) at most h, etc.

Theorem 1. *The collection $\mathcal{H}_{0,0}$ of all paths has generating function*

$$H_{0,0} = \cfrac{1}{1 - c_0 - \cfrac{a_0 b_1}{1 - c_1 - \cfrac{a_1 b_2}{1 - c_2 - \cfrac{a_2 b_3}{\ddots}}}}$$

Proof. It suffices to observe that $(1-f)^{-1} = 1+f+f^2+\cdots$ generates symbolically all the sequences with components f. For instance, in $\mathcal{H}_{0,0}$, the expressions

$$\text{(2)} \qquad \frac{1}{1-c_0}, \qquad \frac{1}{1-c_0-a_0b_1}, \qquad \cfrac{1}{1-c_0-\cfrac{a_0b_1}{1-c_1}}$$

generate successively paths composed from c_0 level steps only, paths of height at most 1 without c_1 steps, all paths of height at most 1. The complete continued fraction representation is easily built by stages in a similar fashion. ☐

In particular, the collection of all paths from level 0 to level 0 with height at most h is

$$\text{(3)} \qquad H_{0,0}^{[<h]} = \frac{P_h}{Q_h},$$

a rational fraction, whose numerators and denominators, P_h, Q_h, each satisfy the recurrence

$$y_{h+1} = (1-c_h)y_h - a_{h-1}b_h y_{h-1},$$

with $Q_{-1} = P_0 = 0$, $Q_0 = P_1 = 1$. (Linear fractional transformations are 2×2 matrices in disguise!)

Well-known path decompositions, like those based on first or last time at which levels are reached, can then be used provided they are combinatorially "unambiguous". This and simple manipulations on linear fractional transformations give access to many geometric constraints in addition to (2) and (3). We cite here some representative identities from [1, 2],

$$\text{(4)} \qquad H_{0,h-1}^{[<h]} = \frac{a_0a_1\cdots a_{h-1}}{Q_h}, \qquad H_{0,k} = \frac{1}{b_1b_2\cdots b_k}\left(Q_kH_{0,0} - P_k\right),$$

$$\text{(5)} \qquad H_{k,l} = \frac{Q_k}{a_0\cdots a_{k-1}b_1\cdots b_l}\left(Q_lH_{0,0} - P_l\right),$$

where the latter holds provided $k \le l$.

The forms (2), (3) (4), (5) can be converted into *bona fide* counting generating functions of paths weighted multiplicatively by means of the *combinatorial morphism*,

$$\chi(a_k) = \alpha_k z, \qquad \chi(b_k) = \beta_k z, \qquad \chi(c_k) = \gamma_k z.$$

In that case, the continued fraction (2) becomes the general fraction of the J-type (for Jacobi); see [7, 9, 13].

3. The Connection

We illustrate here in its simplest form the many-faceted connection between birth-death processes and continued fractions. It was apparently first stated explicitly by Jones and Magnus but it is implicit in earlier works of Karlin and McGregor. The connection goes through the probabilities $p_{i,j}(t)$ of being in state j at time t starting from state i and the Laplace transforms,

$$P_{i,j}(s) = \int_0^\infty p_{i,j}(t) e^{-st}\, dt.$$

Theorem 2. *The Laplace transform of the probability of return to the origin satisfies*

$$P_{0,0}(s) = \cfrac{1}{\lambda_0 + s - \cfrac{\lambda_0\mu_1}{\lambda_1 + \mu_1 + s - \cfrac{\lambda_1\mu_2}{\ddots}}}.$$

We offer here two proofs. A third proof that is based on "uniformization of time" can also be given but is omitted in this note.

Proof 1. Take the Laplace transform of the fundamental system (1) (so that $p_j(t) = p_{0,j}(t)$) and use the induced relations on the ratios $P_{0,r}/P_{0,r+1}$. This proof is the most direct but the least illuminating from a structural standpoint. In particular, this proof does not provide an immediate grasp on the question of deciding which parameters are amenable to continued fraction representations. ☐

Proof 2. Examine the times at which the (continuous time) birth-death process $\{\Lambda_t\}$ changes states. This defines an embedded (discrete time) Markov chain $\{Y_n\}$. Then the set of trajectories of the chain $\{Y_n\}$ is exactly the family of lattice paths of Section 2. The method consists in splitting the probabilities by conditioning according to all legal trajectories.

- The first observation is that, given a lattice path $w = w_1 w_2 \cdots w_n$, the probability $p_{0,0}(t \mid w)$ of being back to 0 at time t having followed the path w is

  $$\Pr\{\Lambda_t = 0 \mid w\} = \Pr\left\{S_{q_1} + S_{q_2} + \cdots + S_{q_n} \le t,\ S_{q_1} + S_{q_2} + \cdots + S_{q_n} + S_{q_{n+1}} > t\right\},$$

  where S_{q_j} is the random variable that represents the sojourn time at the state q_j determined by $w_1 \cdots w_j$, while the right-hand side involves q_{n+1} that ranges over all legal "continuations" of w (in the case of $\mathcal{H}_{0,0}$, one has $w_{n+1} = a_0$ and $q_{n+1} = 0$). As seen already, the sojourn time at some state e is exponential with parameter $(\lambda_e + \mu_e)$ so that its Laplace transform is $(\lambda_e + \mu_e)/(s + \mu_e + \lambda_e)$.
- The second observation is that the probability of a path in the embedded chain is the product of the individual transition probabilities, namely $\lambda_j/(\lambda_j + \mu_j)$ and $\mu_j/(\lambda_j + \mu_j)$.

The different sojourn times are independent by the nature of the process (the strong Markov property satisfied by $\{\Lambda_t\}$). Also, sums of independent random variables correspond to products of Laplace transforms. Thus, the Laplace transform of the probability in the continuous model of following a path w has a product form; for instance, to $w = a_0 a_1 b_2 a_1$, there corresponds the transform

$$\left(\frac{\lambda_0}{\lambda_0 + \mu_0}\frac{\lambda_1}{\lambda_1 + \mu_1}\frac{\mu_2}{\lambda_2 + \mu_2}\frac{\lambda_1}{\lambda_1 + \mu_1}\right) \cdot \left(\frac{\lambda_0 + \mu_0}{s + \lambda_0 + \mu_0}\frac{\lambda_1 + \mu_1}{s + \lambda_1 + \mu_1}\frac{\lambda_2 + \mu_2}{s + \lambda_2 + \mu_2}\frac{\lambda_1 + \mu_1}{s + \lambda_1 + \mu_1}\right).$$

Thus, the Laplace transform $P_{0,0}(s)$ is, apart from a fudge factor of $1/(s+\lambda_0)$, a sum over all paths lattice from zero to zero weighted multiplicatively by the *probabilistic morphism*,

$$\chi'(a_j) = \frac{\lambda_j}{s+\lambda_j+\mu_j}, \qquad \chi'(b_j) = \frac{\mu_j}{s+\lambda_j+\mu_j}, \tag{6}$$

with $\chi'(c_j)=0$. In other words, one has $P_{0,0}(s) = \chi'(H_{0,0})\frac{1}{s+\lambda_0}$, and the statement follows. □

The same method applies to the computation of transition probabilities, the analysis of maximum height, and so on. For instance, the probability of reaching state k has

$$P_{0,k}(s) = \frac{1}{\mu_1\mu_2\cdots\mu_k}\left(A_k(s)P_{0,0}(s) - B_k(s)\right),$$

where A_k/B_k is the kth convergent of the continued fraction that represents $P_{0,0}$, so that A_k, B_k are simple variants of $\chi'(P_k), \chi'(Q_k)$.

Orthogonality. In the case of paths, the reciprocals of the Q_h polynomials, $\overline{Q}_h(z) = z^h\chi(Q)(z^{-1})$ are *formally orthogonal* with respect to a measure defined its moments,

$$\mathcal{L}[z^n] \equiv \int z^n d\mu(z) = H_{0,0,n}. \tag{7}$$

Formal aspects of paths and orthogonality are detailed in Godsil's book [3].

A similar orthogonality property then holds for the probabilistic counterparts A_h, B_h of the P_k, Q_k polynomials. This provides alternative expressions of various probabilistic quantities in terms of scalar products involving the measure μ of (7). One can rederive in this way, via the combinatorial theory, a number of formulæ originally discovered by Karlin and McGregor. For instance, one has

$$p_{m,n}(t) = \pi_n \int_0^\infty e^{-tx}\theta_m(x)\theta_n(x)\, d\mu(x),$$

where the θ_k polynomials (closely related to the B_k and Q_k) satisfy the recurrence $\lambda_n\theta_{n+1} + (x - \lambda_n - \mu_n)\theta_n + \mu_n\theta_{n-1} = 0$.

4. So What?

The original motivation for the talk comes from the need to elucidate the behaviour of certain *queueing systems* in the context of telecommunication applications. For instance, the single server queue ($M/M/1$) is modelled by $\lambda_j = \rho$, $\mu_j = 1$, while the infinite server queue ($M/M/\infty$) corresponds to $\lambda_j = \rho$, $\mu_j = j$. (Models of population growth lead to considering different types of weights, like $\lambda_j = (j+1)\rho$, $\mu_j = j$.) More specifically, the problem is to quantify parameters of some simple statistical multiplexing scheme that describe the quality of service on an ATM link. The relevant model is that of the $M/M/\infty$ queue and parameters are to be analysed, like the duration θ of an excursion above some level c, the volume V of lost information, or the number of bursts C in a busy period.

Each parameter leads to a specific continued fraction representation. By Theorem 2, the basic continued fraction of the $M/M/\infty$ process is

$$\cfrac{1}{s+\rho-\cfrac{1\rho}{s+1+\rho-\cfrac{2\rho}{\ddots}}}.$$

This is recognizable as an instance of Gauß's continued fraction associated to a quotient of *contiguous hypergeometric functions*. The numerator and denominator polynomials are the Poisson-Charlier polynomials that are orthogonal with respect to the Poisson measure.

The quantity V (area) leads to challenging asymptotics questions both for the $M/M/\infty$ queue and for the $M/M/1$ queue. A simple modification of the basic techniques of this note shows that the bivariate Laplace transform with (s,u) "marking" (t,V) is obtained by the modified morphism,

$$\chi''(a_j) = \frac{\lambda_j}{s + ju + \lambda_j + \mu_j}, \qquad \chi''(b_j) = \frac{\mu_j}{s + ju + \lambda_j + \mu_j}.$$

In the case of area under the $M/M/1$ queue, quotients of continuous Bessel functions make an appearance. Stripped of its probabilistic context, the corresponding problem of tail estimation then admits a purely analytic formulation:

Problem. *Let $A(x)$ be a function whose Laplace transform is*

$$\widetilde{A}(s) = \frac{1}{\sqrt{s}} \frac{J_{\nu(s)+1}\left(\frac{2\sqrt{\rho}}{s}\right)}{J_{\nu(s)}\left(\frac{2\sqrt{\rho}}{s}\right)}, \qquad \nu(s) = (1+\rho)/s,$$

with J_ν a Bessel function, and $\rho > 0$ a parameter. Show that, for some constants c_1, c_2, one has

$$\int_x^\infty A(y)\,dy \sim c_1 x^{-1/4} e^{-c_2\sqrt{x}}, \qquad (x \to +\infty).$$

Under plausible analytic or probabilistic conjectures, precise (and useful!) quantitative conclusions can be drawn. See the papers by Guillemin and Pinchon [5, 6] for full developments.

Bibliography

[1] Flajolet (P.). – Combinatorial aspects of continued fractions. *Discrete Mathematics*, vol. 32, 1980, pp. 125–161.

[2] Flajolet (P.), Françon (J.), and Vuillemin (J.). – Sequence of operations analysis for dynamic data structures. *Journal of Algorithms*, vol. 1, 1980, pp. 111–141.

[3] Godsil (C. D.). – *Algebraic Combinatorics*. – Chapman and Hall, 1993.

[4] Goulden (Ian P.) and Jackson (David M.). – *Combinatorial Enumeration*. – John Wiley, New York, 1983.

[5] Guillemin (Fabrice) and Pinchon (Didier). – Continued fraction analysis of the duration of an excursion in an $M/M/\infty$ system. *Journal of Applied Probability*, vol. 35, n° 1, 1998, pp. 165–183.

[6] Guillemin (Fabrice) and Pinchon (Didier). – Excursions of birth and death processes, orthogonal polynomials, and continued fractions. – Preprint, 1998. To appear in *Journal of Applied Probability*. 21 pages.

[7] Henrici (Peter). – *Applied and Computational Complex Analysis*. – John Wiley, New York, 1977. 3 volumes.

[8] Jones (William B.) and Magnus (Arne). – Application of Stieltjes fractions to birth-death processes. In Saff (E. B.) and Varga (Richard S.) (editors), *Padé and rational approximation*. pp. 173–179. – New York, 1977. Proceedings of an International Symposium held at the University of South Florida, Tampa, Fla., December 15-17, 1976.

[9] Jones (William B.) and Thron (W. J.). – *Continued Fractions: Analytic Theory and Applications*. – Addison–Wesley, 1990, *Encyclopedia of Mathematics and its Applications*, vol. 11.

[10] Karlin (S.) and McGregor (J. L.). – The differential equations of birth-and-death processes, and the Stieltjes moment problem. *Transactions of the American Mathematical Society*, vol. 85, 1957, pp. 489–546.

[11] Karlin (Samuel) and McGregor (James). – The classification of birth and death processes. *Transactions of the American Mathematical Society*, vol. 86, 1957, pp. 366–400.

[12] Karlin (Samuel) and Taylor (Howard). – *A First Course in Stochastic Processes*. – Academic Press, 1975, second edition.

[13] Perron (Oskar). – *Die Lehre von der Kettenbrüchen*. – Teubner, 1954, vol. 2.