1 QBD Processes

1.1 Birth-death processes

We wish to model the evolution of a queue of people waiting in line. At each moment, there is a probability c that an additional customer arrives and the queue length increases, a probability a that a customer (if there is one) is served and leaves the queue, and a probability b that no one arrives or is served and the queue length stays the same. Clearly one must have $a + b + c = 1$. We assume $a \neq 0$, $c \neq 0$.

This situation is modelled by a Markov chain with state the number of people in the queue; hence the state set is $\mathbb{N}$. It is a Markov chain with an infinite number of states, which is a new concept for us.

We can write the transition probabilities in an infinite matrix

$$\begin{bmatrix} a+b & a & & & \\ c & b & a & & \\ & c & b & a & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{1}$$

1.2 Transient or recurrent?

All the states belong to the same class. Are they transient or recurrent? Note that this is not a moot question: one can have Markov chains with infinite states in which each state is transient: for instance, take model (1) with $a = b = 0, c = 1$.

To answer this question, we need $\mathbb{P}\left[x_k = 0 \text{ for some } k > 0 \mid x_0 = 0\right]$. Clearly, in view of the first transition, this is equal to $a+b+cg$, where $g = \mathbb{P}\left[x_k = 0 \text{ for some } k > 1 \mid x_1 = 1\right]$, that is, the probability of *eventually* returning to level 0 starting from level 1.

Let us divide on the different things that can happen at time 1: we have

$$g = \underbrace{a}_{\text{return immediately to 0, } x_1 = 0} + \underbrace{bg}_{x_1 = 1\text{, then eventually return to 0}} + \underbrace{cg^2}_{\text{go up 1 level, then go down (eventually) twice}}$$

This equation has two solutions. One is 1, for $a + b + c = 1$. The other is $\frac{a}{c}$. The following result holds.

Lemma 1. *The quantity $g = \mathbb{P}\left[x_k = 0 \text{ for some } k > 1 \mid x_1 = 1\right]$ is the smallest solution of the equation $a + bx + cx^2 = x$.*

Proof. (sketch, not a real proof). Let h be this smallest solution, and $g_i = \mathbb{P}\left[x_k = 0 \text{ for some } k \text{ with } 1 \leq k \leq i \mid x_1 = 1\right]$. Clearly, g_i is increasing and its limit is g. One can prove by induction that $g_i \leq h$. □

Hence if $\frac{a}{c} < 1$ ("more people arrive than leave the queue"), then $g < 1$, $a + b + cg < 1$, and the queue is transient. Intuitively, with probability 1 the number of people in the queue grows indefinitely. If $\frac{a}{c} \geq 1$ ("more people leave"), then $g = 1$, and the queue is recurrent. The in-between case, $a = c$, is known as *null recurrent*. The queue returns to lower states with probability 1, but takes infinitely long time to do so. In contrast, when $\frac{a}{c} < 1$ we say that the queue is *positive recurrent.*

1.3 Steady-state probability

When the queue is recurrent, it makes sense to compute an invariant measure, that will be the steady-state limit distribution.

The stationary probability vector π must satisfy

$$(a+b)\pi_0 + a\pi_1 = \pi_0 \tag{2}$$

$$c\pi_k + b\pi_{k+1} + a\pi_{k+2} = \pi_{k+1} \qquad k \geq 0. \tag{3}$$

Equation (3) falls under the general theory for linear recurrence sequences, i.e., all solutions of (3) are of the form $\alpha x_1^k + \beta x_2^k$, where x_1 and x_2 are the solutions of the equation $c + bx + ax^2 = x$. These solutions are $x_1 = 1$ and $x_2 = r := \frac{c}{a}$ (it's the reversal of the equation that we considered above). To get a vector with $\sum_{k=1}^{\infty} \pi_k = 1$, one must have $\alpha = 0$. Then $\beta = (\sum_{k=0}^{\infty} x_2^k)^{-1}$. Luckily for us, this (only) solution satisfies (2) as well.

So we get the following:

Theorem 2. *The steady-state distribution of* (1) *(when it is positive recurrent, i.e.,* $c < a$*) is given by* $\pi_k = (1-r)^{-1} r^k$*, with* $r = \frac{c}{a}$.

1.4 Now with matrices

A very quick description of what happens if we consider a "block case" of this problem.

Suppose that we have a Markov chain that models the arrival rate. For instance (very simple case), we have two states, "congestioned" and "not congestioned". Their transition matrix might look like this:

$$P = \begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix},$$

i.e., $\mathbb{P}[\text{not congestioned} \to \text{congestioned}] = 0.1$, $\mathbb{P}[\text{congestioned} \to \text{not congestioned}] = 0.3$, and so on. Depending on the state, we may have different arrival probabilities. In general, there are three matrices A, B, C, with $A+B+C = P$, with $A_{ij} = \mathbb{P}[\text{one person leaves the queue, and } j \to i]$, $B_{ij} = \mathbb{P}[\text{the queue stays the same length, and } j \to i]$, $C_{ij} = \mathbb{P}[\text{one person joins the queue, and } j \to i]$.

The transition matrix is now the infinite block tridiagonal matrix

$$\begin{bmatrix} A+B & A & & & \\ C & B & A & & \\ & C & B & A & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4}$$

This model is called *QBD (quasi-birth-death).*

An argument similar to the one in scalar case yields $A + BG + CG^2 = G$, where G is the matrix so that $G_{ij} = \mathbb{P}[\text{we return from level 1 to level 0 for the first time, and } j \to i]$ (first return probabilities).

G is the smallest solution (componentwise) to the matrix equation $A+BX+CX^2 = X$. There is no easy closed-form formula, but it can be determined using several matrix iterations (simplest of all: $X_{k+1} = A+BX_k+CX_k^2$, with $X_0 = 0$, converges monotonically to G). The model is positive recurrent if $\rho(G) = 1$.

Similarly, there is a matrix R which is the minimal solution to $AY^2 + BY + C = Y$, and the invariant measure is of the form $\pi_k = w^T R^k$ for some vector w^T (note that π_k, probabilities of being at level k, is a vector with n states, where n is the dimension of the "environment" queue). Matrix analysis is a powerful tool to study these equations. For instance, the $2n$ zeros of $f(z) = \det(A+Bz+Cz^2)$ are the eigenvalues of G and those of R^{-1}.