ORIE 5530: Modeling Under Uncertainty

Lecture 9 (Continuous Time Markov Chains)

Professor Mark S. Squillante,
Adapted from Professor Itai Gurvich's original notes

We begin our journey into continuous-time Markov chains (CTMCs). Simply put, CTMCs are an extension of discrete-time Markov chains (DTMCs) where, instead of staying in each state for exactly one time unit, the chain remains in each state for a random amount of time that follows an exponential distribution.

There are two **equivalent** definitions of continuous time Markov chains. The first is based on defining state-clocks and transition probabilities and the second is what one can call a "competing-clocks" construction.

It is useful to have both as they can be handy in different settings.

Some of this material follows Chapter 6 in Ross's book. The first construction, for example, follows Section 6.2 there. However, Chapter 6 of Ross takes for granted that you had gone through his chapter 5 that is about the exponential distribution. We will summarize the needed essentials within the notes below.

9.1 First definition

Let's take as a starting point a three-state chain.

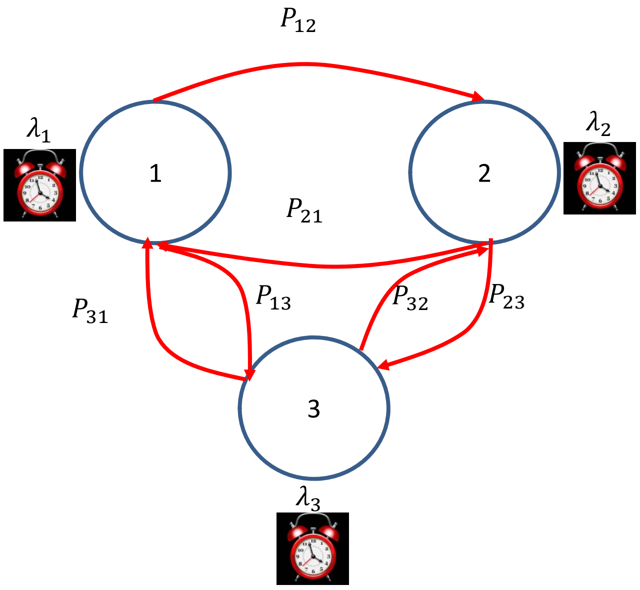

Figure 9.1: A 3-state continuous time Markov chain

Here is how the dynamics work. Suppose you start at time 0 in state 1. That is, $X_0 = 1$. You set up an alarm clock to go off after an exponential amount of time with rate λ_1 and hence expectation $1/\lambda_1$. We will refer frequently to the parameter of the exponential as its rate. When the alarm clock sounds, you move with probability P_{12} to state 2 and with probability $P_{13} = 1 - P_{12}$ to state 3. Say, you moved to state 2. As soon as you enter the state, you activate an $\exp(\lambda_2)$ alarm clock. Once it sounds you move with probability P_{21} to state 1 and with probability $P_{23} = 1 - P_{21}$ to state 3, and so on in this manner.

Notice that this fully specifies the process: if someone asked you to simulate this process, you have enough information to do so, right? So all we need are two ingredients — the rates

$$\lambda = (\lambda_1, \lambda_2, \lambda_3)$$

of the alarm clocks and the transition probability matrix

$$P = \begin{bmatrix} 0 & P_{12} & P_{13} \\ P_{21} & 0 & P_{23} \\ P_{31} & P_{32} & 0 \end{bmatrix}.$$

In summary, a CTMC is a process having the property that each time it enters state i:

(1) the amount of time it spends in that state before making a transition into a different state is exponentially distributed with mean $1/\lambda_i$ and

(2) when the process leaves state i, it next enters state j with probability P_{ij}. The probabilities P_{ij} must satisfy $P_{ii} = 0$ for all i (no return in one step) and $\sum_j P_{ij} = 1$.

What makes this process that we just created appealing? Well, if you tell someone that the chain in Figure 9.1 is in state 1 now (say this is minute 4), then they can compute the probability that it will be in state 2 in 6 minutes from now. We do not need to know how much time has passed since you entered state 1. That is, even if you throw in all the history before minute 4, we do not care about that:

$$\mathbb{P}\{X_{10} = 2 | X_4 = 1, X_s = x_s; s < 4\} = \mathbb{P}\{X_{10} = 2 | X_4 = 1\}.$$

Why is that? This is driven by what is called the memoryless property of the exponential distribution. For the above, it implies that if you tell me that 30 seconds passed since the chain entered state 1, then the distribution of the remaining time until you move is still $\exp(\lambda_1)$ – in other words, the likelihood that is takes another 6 minutes until you move is as if you moved to state 1 just now, hence we can ignore when you entered.

Formally, if $Y \sim exp(\lambda)$,

$$\mathbb{P}\{Y \geq s + t | Y \geq t\} = \mathbb{P}\{Y \geq s\} = e^{-\lambda s}.$$

It should also be clear that, because nothing in the properties of the chain (the vector λ and P) changes with time, the probability of moving from state 1 to state 2 in 5 minutes depends only on the time that passes (6 minutes) and not on the running time:

$$\mathbb{P}\{X_{10} = 2 | X_4 = 1\} = \mathbb{P}\{X_6 = 2 | X_0 = 1\}.$$

9.2 Second definition: Competing clocks

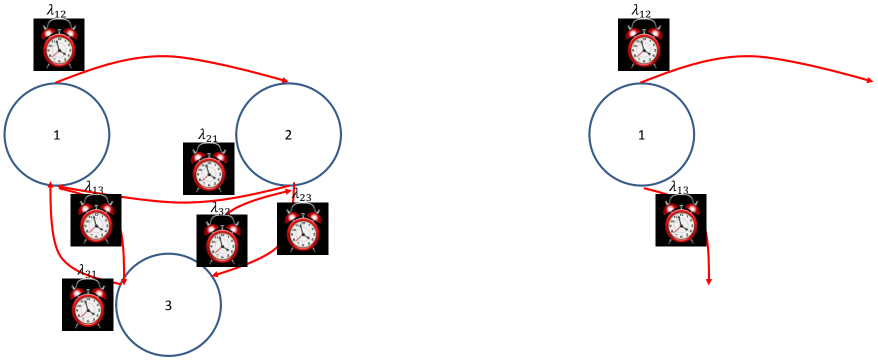

Figure 9.2: A 3-state continuous time Markov chain: competing clocks

Let us focus on the subset of the chain captured on the right-hand side of Figure 9.2. Here are the dynamics:

> Say you just entered state 1. You activate two clocks–the first exponential with rate λ_{12}—clock A—and the second exponential with rate λ_{13}—clock B. The first clock that rings determines where you go. If clock A rings first, you move to state 2 and if clock B rings first you move to state 3. That is, you move after the **minimum of the two exponentials with rates λ_{12} and λ_{13}**.
>
> When in state 2, you have two clocks with competing rates λ_{23} and λ_{21} and you do the same. Analogously for state 3.

Again, this is a complete specification meaning that the above description and the values λ_{ij} are sufficient information to simulate this process.

This is often (as we will see) a more useful construction than the first one. Essentially, however, they are both equivalent:

- How much time until the chain leaves state 1: It is the minimum of two independent exponentials—one, let's call it Y_{13} with rate λ_{13} and the other Y_{12} with rate λ_{12}. The following is known

  $$\min\{Y_{12}, Y_{13}\} \sim exp(\lambda_{12} + \lambda_{13}).$$

  That is the minimum is itself exponential with the sum of the rates.

  Letting $\lambda_i = \sum_j \lambda_{ij}$ we have that after entering state i the chain stays there for an exponential amount of time with rate λ_i.

- What is the likelihood that after leaving state 1, you go to state 2 rather than state 3? This is the likelihood that clock A rings before clock B and it turns out that

  $$\mathbb{P}\{Y_{12} < Y_{13}\} = \frac{\lambda_{12}}{\lambda_{12} + \lambda_{13}},$$

  and the other case is

  $$\mathbb{P}\{Y_{13} < Y_{12}\} = \frac{\lambda_{13}}{\lambda_{12} + \lambda_{13}}$$

  (the probability that $Y_1 = Y_2$ is zero for two independent continuous random variables).

Moreover, this probability is independent of how much time it took until the clocks rang. That is,

$$\mathbb{P}\{Y_{12} < Y_{13} | \min\{Y_{12}, Y_{13}\} = y\} = \frac{\lambda_{12}}{\lambda_{12} + \lambda_{13}}.$$

In summary, say someone gives you a construction with the parameters λ_{ij} as above. Define

$$\lambda_i = \sum_j \lambda_{ij},\ P_{ij} = \frac{\lambda_{ij}}{\lambda_i}.$$

Then, it turns out that both properties we introduce in Section 1 hold:

(1) the amount of time the chain spends in that state before making a transition into a different state is exponentially distributed with mean $1/\lambda_i$ and

(2) when the process leaves state i, it next enters state j with probability P_{ij}. The probabilities P_{ij} must satisfy $P_{ii} = 0$ for all i (no return in one step) and $\sum_j P_{ij} = 1$.

WE ARE MORE OFTEN THAN NOT GOING TO USE THE SECOND CONSTRUCTION.

9.3 The practical approach to constructing rates: An Uber example

The practical approach to construct these CTMCs is to forget about the underlying math and think about the speeds (rates) (the math kicks-in in justifying that this simple way of thinking is correct).

In the single-server queue example we did in class — the speed of arrival was λ and the speed of service completions was μ. In other words, if each service takes an expontential time with expectation $1/\mu$ minutes, it means that we have completions at a speed (rate) of μ per minute, as long as the server remains busy.

Consider the Uber example we started in class but let us limit our attention to two regions A and B. Here are the ingredients:

Customers: Customers arrive to region A at speed λ_A and to region B at speed λ_B. A customer taking a cab in region A wants to get to a point in region A with probability P_{AA} and wants to get to a point in region B with probability P_{BB}. The speed (or rate) at which customers take a cab to go from region A to region B is then $\lambda_A P_{AB}$.

Ride time: A trip from a point in region A to a point in region B takes an exponential amount of time with mean $1/\mu_{AB}$. Similarly, we define μ_{AA}, μ_{BA} and μ_{BB}.

Driver self movement: When a ride is complete in region B the driver can choose to stay in this region or relocate. The driver will stay in region B with probability Q_{BB} after completing a ride or move to region A with probability Q_{BA}. We similarly define Q_{AA} and Q_{AB}.

————————————-

The state is 8 dimensional as follows:

- Empty cabs in a region:
  - E_{AA} is the number of empty cabs in region A
  - E_{BB} is the number of empty cabs in region B
- Empty cabs in motion:
  - E_{AB} is the number of empty cabs in transition from A to B

  - E_{BA} is the number of empty cabs in transition from B to A

- Taken cabs in motion:
  - F_{AA} is the number of busy (i.e., taking a customer) cabs in transition from one point in A to another point in A
  - F_{BB} is the number of busy (i.e., taking a customer) cabs in transition from one point in B to another point in B
  - F_{AB} is the number of busy (i.e., taking a customer) cabs in transition from A to B
  - F_{BA} is the number of busy (i.e., taking a customer) cabs in transition from B to A

Our process is the **8-dimensional** chain

$$X(t) = (E_{AA}(t), E_{BB}(t), E_{AB}(t), E_{BA}(t), F_{AA}(t), F_{BB}(t), F_{AB}(t), F_{BA}(t)),$$

where the states are of the form

$$e = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}).$$

What are the transition speeds from state e^1 to state e^2. There are only certain types of transition possible

- Customers taking cabs:
  - A customer takes a cab in region A and goes to region B. In this case e_{AA} decreases by one (a cab is taken from region A) and f_{AB} increases by 1 (a cab is driving to region B). The speed at which customers arrive and are interested in such a ride is $\lambda_A P_{AB}$. So, if $e_{AA} > 0$ (there are cabs available), we have a transition

    $$\begin{aligned} \text{From} \quad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{BA}, f_{BA}) \\ \text{To} \quad & e^2 = (e_{AA} - 1, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB} + 1, f_{BA}, \end{aligned}$$

    with speed (rate) $\lambda_A P_{AB}$. That is $\lambda_{e^1 e^2} = \lambda_A P_{AB}$
    Similarly for a customer that takes a cab from region A and stays in A we have with rate $\lambda_A P_{AA}$ a transition

    $$\begin{aligned} \text{From} \quad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}) \\ \text{To} \quad & e^2 = (e_{AA} - 1, e_{BB}, e_{AB}, e_{BA}, f_{AA} + 1, f_{BB}, f_{AB}, f_{BA}, \end{aligned}$$

    and for transitions from B to A and from B to itself we respectively have $\lambda_B P_{BA}$ and $\lambda_B P_{BB}$ with the corresponding state transitions.

- Customers completing rides and cabs moving:
  - Since the ride from A to B takes an exponential amount of time with rate μ_{AB}, when there are f_{AB} cabs driving from A to B the first of these cabs will be done with rate $\mu_{AB} f_{AB}$ (this is like in the multiserver queueing model). When this ride is done, the driver will stay in region B with probability Q_{BB} or will move back to A with probability Q_{BA}. This will increase by one the number of empty cabs going from B to A (this is e_{BA}).
    Thus, we have the transitions that move

    $$\begin{aligned} \text{From} \quad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}) \\ \text{To} \quad & e^2 = (e_{AA}, e_{BB} + 1, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB} - 1, f_{BA}), \end{aligned}$$

    at a speed of $\mu_{AB} f_{AB} Q_{BB}$ and we move

    $$\begin{aligned} \text{From} \quad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}) \\ \text{To} \quad & e^2 = (e_{AA}, e_{BB}, e_{AB}, e_{BA} + 1, f_{AA}, f_{BB}, f_{AB} - 1, f_{BA}), \end{aligned}$$

    at a speed of $\mu_{AB} f_{AB} Q_{BA}$.

- Empty cabs completing an empty transition:
  - The ride from A to B takes an exponential amount of time with rate μ_{AB}. By the same argument as above, when there are e_{AB} cabs transitioning from A to B, completions happen at rate $\mu_{AB} e_{AB}$ and hence we have transitions that move

    $$\begin{aligned} \text{From} \qquad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}) \\ \text{To} \qquad & e^2 = (e_{AA}, e_{BB}+1, e_{AB}-1, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}), \end{aligned}$$

    at a speed of $\mu_{AB} e_{AB}$, and similarly moving

    $$\begin{aligned} \text{From} \qquad & e^1 = (e_{AA}, e_{BB}, e_{AB}, e_{BA}, f_{AA}, f_{BB}, f_{AB}, f_{BA}) \\ \text{To} \qquad & e^2 = (e_{AA}+1, e_{BB}, e_{AB}, e_{BA}-1, f_{AA}, f_{BB}, f_{AB}, f_{BA}), \end{aligned}$$

    with rate $\mu_{BA} e_{BA}$.