Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 11
Continuous-Time Markov Chains, Generator Matrix, Kolmogorov Equations

Hi and hello, everyone. What we will do next is the "Continuous-Time Markov Chains" and the basic properties and the basic ideas that we need from this particular concept because this is what is ultimately we need when we start our queueing analysis, like whatever we did in discrete-time Markov chain and Poisson process both of them like will lead to this particular thing that we are doing here is continuous-time Markov chain. And this is what directly like we are going to use as a model for all our queueing systems because things that we are going to deal with are going to be in a continuous-time and state is discrete. So, what we are having here is the continuous-time Markov chain is what then practically will be applied in our queueing analysis. So, that is what we are coming. So, whatever we did in the discrete-time Markov chain, we will see that we are now transferring those results to the continuous-time version of them.

So, what is a continuous-time Markov chain? Here the state space is discrete, whereas the parameter space now becomes continuous. So, we will transfer the results from discrete-time, which is basically DTMC or simply MC Markov chain, to CTMC. Now first, let us define what we mean by continuous-time Markov chain.

Definition.

Define $S_0 = 0$ and let $\{S_n, n \geq 1\}$ denote a sequence of RVs such that $S_n > S_{n-1}$ for all $n \geq 1$ and $S_n \to \infty$ as $n \to \infty$. Further, let $\{X_n, n \geq 0\}$ be a sequence of RVs taking values in a countable state space S. A stochastic process $\{Y_t, t \geq 0\}$ with $Y_t = X_n$ for $S_n \leq t < S_{n+1}$ is said to be a pure jump process. The variable $T_n = S_{n+1} - S_n$ (resp. X_n) is called the nth **holding time** *(resp. the nth* **state***) of the process $\{Y_t\}$.*

If, further, $\{X_n, n \geq 0\}$ is a Markov chain with (stationary) transition probability matrix $P = ((p_{ij}))$ and the variables T_n are independent and distributed exponentially with parameter λ_{X_n} only depending on the state X_n, then $\{Y_t, t \geq 0\}$ is called a (time-homogeneous) continuous-time Markov chain (CTMC). The chain $\{X_n, n \geq 0\}$ is called the embedded Markov chain of the CTMC.

So, we will always assume that this $\sup\{\lambda_i, i \in S\}$ is bounded. Because you remember that T_n are all exponential and λ_{X_n} is the parameter; if this parameter is ∞, if it is not finite, then you know what it means, or you could talk about the parameter λ, mean is $1/\lambda$ and what happens in that particular case. So, to avoid that, we will always assume that this is the maximum of all these λ_i's one for each state of the process; they are all finite. So, the $sup\{\lambda_i, i \in S\} < \infty$, which means each one of them is finite, which is what you are assuming. So, there are all some finite values that we are assuming to avoid any of the trivial cases that we might encounter in this particular case. So, this is what is a continuous-time Markov chain.

- Now, a CTMC moves from state to state just like a DTMC, but the time spent in each state is now an exponential random variable; it is not; otherwise, it is an exponential random variable what then we are making it here.

- Note that if i is not an absorbing state, we can assume that the single-step transition from state i back to itself is not allowed (i.e., $p_{ii} = 0$).
  If i is an absorbing state, then $p_{ii} = 1$ and, in this case, we have $\lambda_i = 0$ because it is not going to go out of the state. So, that is what it would mean. So, it would be in that particular state forever. So, for an absorbing state, $\lambda_i = 0$.

- The process $\{Y_t, t \geq 0\}$ satisfy the Markov property. That is,

$$P\left\{Y_t = j | Y_{t_n} = i,\, Y_{t_{n-1}} = i_{n-1},\, \ldots,\, Y_{t_0} = i_0\right\} = P\left\{Y_t = j | Y_{t_n} = i\right\}$$

  for all states $i_0,\, i_1,\, \ldots,\, i_{n-1},\, i$ and j in S, for all $n \geq 1$, and for all $t_0, t_1, \ldots, t_n, t$ such that $0 \leq t_0 < t_1 < \cdots < t_n < t$.
  Of course, one can define the Markov process straight away as if satisfying this, but for us, it is defining through an embedded Markov chain idea taking DTMC to move to CTMC is more useful from our viewpoint. So, we have taken that route; otherwise, one can define directly as a process satisfying the Markov property is what would be a Markov process in general. So, this is a discrete state continuous time. So, this is a continuous-time Markov chain that then one would define.

So, $\{Y_t, t \geq 0\}$ satisfies this Markov property which in this particular case is

$$P\left\{Y_t = j | Y_{t_n} = i,\, Y_{t_{n-1}} = i_{n-1},\, \ldots,\, Y_{t_0} = i_0\right\} = P\left\{Y_t = j | Y_{t_n} = i\right\}$$

.

Now, $P\left\{Y_t = j | Y_{t_n} = i\right\}$ is now the transition probabilities of moving from i to j from time t_n to t.

- So, in general, for any $s < t$, define the the **transition probability matrix** $P(s,t)$ from time s to t by its entries $P_{ij}(s,t) = P\{Y_t = j | Y_s = i\}$. And, they satisfy $P_{ij}(s,t) = P_{ij}(0, t-s)$ and we can restrict our attention to $P(t)$ where $P(t) = P(0,t)$.

- With the above notation, the Markov property yields the **Chapman-Kolmogorov equations** $P(s+t) = P(s)P(t)$ for $s, t \geq 0$.

- Note: $P(0) = I$ and the rows of $P(t)$ sum to 1 much like the DTMC case.

One thing that you need to remember here is this part the continuous-time Markov chain and the embedded Markov chain. Embedded Markov chain is a discrete-time Markov chain that we extract out of this continuous-time Markov chain at the times of transitions; is what you would get this embedded Markov chain which is what X_n's are; that is what you need to remember. Now, let us take a couple of examples.

Example. *[Poisson Process]*
Define $X_n = n$ for $n = 0, 1, 2, \ldots$. Then $\{X_n, n \geq 0\}$ is a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ and transition probabilities $p_{n,n+1} = 1$ for all $n \geq 0$. Let the holding times T_n be IID exponential with parameter $\lambda > 0$.

The stochastic process $\{Y_t, t \geq 0\}$ with $Y_t = X_n$ for $S_n \leq t < S_{n+1}$ is a CTMC with state space S and is known as Poisson process with rate (or intensity or parameter) λ.

Example. *[Two-state CTMC]*

Consider a CTMC with two states $S = \{0, 1\}$ and assume the holding time parameters are given by $\lambda_0 = \lambda_1 = \lambda > 0$.

Since none of the states are absorbing (as $\lambda_i > 0$), and we do not allow self-transitions, the TPM of the embedded MC is $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

We will now determine $P(t)$, starting with $P_{00}(t)$. Assuming $Y_0 = 0$, Y_t will be in 0 if and only if we have an even number of transitions in the time interval $[0, t]$. The time between each transition is an $Exp(\lambda)$ RV and thus the transitions occur according to a Poisson process with parameter λ. We have

$$P_{00}(t) = P\{\textit{even number of transitions in}[0,t]\} = \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{2n}}{(2n)!} = \frac{1}{2} + \frac{1}{2}e^{-2\lambda t}.$$

By symmetry, $P_{11}(t) = P_{00}(t)$. We thus have

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} \end{bmatrix}.$$

Observe: $\lim_{t\to\infty} P(t) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

- A CTMC can be parameterized by the quantities $\{\lambda_i\}$ and $\{p_{ij}\}$. Alternatively, a CTMC can be parameterized by a matrix $Q = ((q_{ij}))$, called **generator matrix** or **infinitesimal generator** or **rate matrix**, and is defined as

$$q_{ij} = \begin{cases} -\lambda_i, & i = j \\ \lambda_i p_{ij}, & i \neq j \end{cases} \quad \text{for all } i, j \in S.$$

- If $Y_0 = i$, the chain will move to the next state at time $T_1 \sim Exp(\lambda_i)$. For small $\Delta t > 0$, $P(T_1 < \Delta t) \approx \lambda_i \Delta t$, i.e., the probability of leaving state i in a short interval of length Δt is approximately $\lambda_i \Delta t$. For this reason, λ_i is often called **the transition rate out of state** i (the expected number of transitions per unit of time). Formally, we can write $\lambda_i = \lim_{\Delta t \to 0^+} \left[\frac{P\{Y_{\Delta t} \neq i | Y_0 = i\}}{\Delta t} \right]$.

- Since the chain moves from state i to state j with probability p_{ij}, we call the quantity $q_{ij} = \lambda_i p_{ij}$, **the transition rate from state** i **to state** j. This is the (i, j)th entry of Q, for $i \neq j$.

- The diagonal elements (i.e, q_{ii}) of Q are such that the rows of Q sum to 0. That is, $q_{ii} = -\sum_{j \neq i} q_{ij} = -\lambda_i \sum_{j \neq i} p_{ij} = -\lambda_i$ holds for all $i \in S$.
  - ▶ If $\lambda_i = 0$, then $\lambda_i \sum_{j \neq i} p_{ij} = \lambda_i = 0$.
  - ▶ If $\lambda_i > 0$, then $p_{ii} = 0$ and so $\sum_{j \neq i} p_{ij} = 1$.

- For small $\Delta t > 0$, based on earlier approximation for λ_i, we can obtain $P_{ii}(\Delta t) \approx 1 + q_{ii}\Delta t$ for $i \in S$ and $P_{ij}(\Delta t) \approx q_{ij}\Delta t$ for $i \neq j$.
  More precisely, we can state $Q = \lim_{\Delta t \to 0^+} \left[\frac{P(\Delta t) - I}{\Delta t} \right]$.

- Q plays a similar role for CTMCs as $P - I$ plays for DTMCs (e.g., stationary equations).

- Q was determined from $\{\lambda_i\}$ and $\{p_{ij}\}$. Alternatively, $\{\lambda_i\}$ and $\{p_{ij}\}$ can be determined from $\{q_{ij}\}$ via $\lambda_i = \sum_{j \neq i} q_{ij}$, basically the off-diagonal, I mean leaving out the diagonal entries the remaining entries, if you sum you will get $\lambda_i = \sum_{j \neq i} q_{ij}$, these quantities are all nonnegative that is what you will see then $p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$.

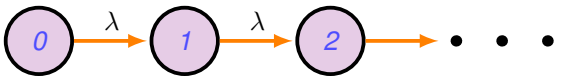
So, this Q is what then you can determine if you know $\{\lambda_i\}$, $\{p_{ij}\}$ you can decide what is Q if you know $\{q_{ij}\}$'s then you can determine what is λ_i and p_{ij} using $\lambda_i = \sum_{j \neq i} q_{ij}, p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$.. Why do we need these? Because, most of the time, we would specify in terms of $\{q_{ij}\}$. If you want the corresponding holding times at a particular state or the embedded Markov chains transition probabilities, then you simply have to use these particular expressions. So, $\lambda_i = \sum_{j \neq i} q_{ij}, \quad p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$. is important. So, you have to remember that you should know exactly how one can move from q_{ij} to λ_i because we will be interested in getting to know what is the embedded Markov chain transition probabilities. So, what do you have to do? You have to simply calculate $p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$., and the holding times in a particular state would be $\lambda_i = \sum_{j \neq i} q_{ij}$ because λ_i is what is coming out here in the denominator of $p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$ when you are computing this transition probability.

So, how can we represent.

- Much like in the discrete-time case the (i, j)th entry of Q is called the **infinitesimal transition rate** from state i to j. A state transition rate diagram (or simply rate diagram) for a CTMC is a directed graph where the nodes represent the states and the edges represent the transition rates q_{ij}. The values of q_{ii} are not shown because they are implied by the other values. As in DTMC, very useful tool here too and we will use extensively.

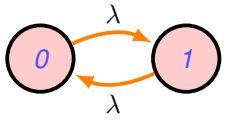
Example. *[Poisson Process]*

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$



Example. *[Two-state CTMC]*

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}$$



So, for $Q = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}$ what are the embedded Markov chain transition probabilities? We have already written down $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Corresponding to $Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$,

it is basically the transition the embedded Markov chains TPM would be $P = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$

Now, these transition probabilities we need to determine $P_{ij})t$ or $P(t)$, and how do they determine?

- The transition probabilities $P_{ij}(t)$ of a CTMC satisfy the systems of differential equations

$$\frac{dP_{ij}(t)}{dt} = \sum_{k \in S} P_{ik}(t) q_{kj} \quad \text{and} \quad \frac{dP_{ij}(t)}{dt} = \sum_{k \in S} q_{ik} P_{kj}(t).$$

  These are called the Kolmogorov forward and backward equations.
  In matrix notations, $P'(t) = P(t)Q$ and $P'(t) = QP(t)$ (with $P(0) = I$).

- The transition probability matrices can be expressed in terms of the generator by $P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n$, for all $t \geq 0$, with Q^n denoting the nth power of Q.

  So, this is called matrix exponential now like there is a whole lot of theory of how to compute a matrix exponential for the particular case, but whatever it is in our case, the theory tells that $P(t)$ is given e^{Qt}. Now, remember what we wanted to compute; we want to compute $P(t)$ for a Markov chain; if you want to know completely about the Markov chain, then you want to know about $P(t)$. Now, $P(t)$ as supposed to get it for every t you have to compute it; the computation can be done if you know Q in terms of this. If you know Q, then you take $P = e^{Qt}$ to get the $P(t)$; that is what is the advantage of using with Q. ▶ So Q uniquely determines all transition matrices.

So, you can describe in terms of $P(t)$ or equivalently in terms of Q without any time dependency there.

- A CTMC is completely determined (i.e., FDDs are detemined) by the transition matrices and the initial distribution.

- Define $p_i(t) = P\{Y_t = i\}$ for $i \in S$ as the probability that the CTMC is in state i at time t. Denote $\boldsymbol{p}(t)$ as the vector with entries $p_i(t)$ (state probabilities).

- We can characterize the state probabilities via systems of differential equations which are also forms of forward and backward Kolmogorov equations

$$\boldsymbol{p}'(t) = \boldsymbol{p}(t)Q \quad \text{and} \quad \boldsymbol{p}'(t) = Q\boldsymbol{p}(t).$$

  ▶ **Given Q and $\boldsymbol{p}(0)$, we can solve for $\boldsymbol{p}(t)$ from the above.**

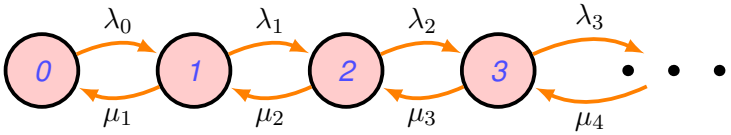
Now quickly, we will go through this example which we call a birth-death process.

Example. *(Birth-Death Process (BDP))*
A *birth-death process (BDP)* is a CTMC $\{Y_t\}$ with $S = \{0, 1, 2, \dots\}$ in which state transitions either increase the system state by 1 (a birth) or decrease the system state by 1 (a death). The generator matrix (or rate matrix) for a BDP is

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

When the system is in state i, births occur with rate λ_i (for $i \geq 0$) and deaths occur with rate μ_i (for $i \geq 1$). In other words, $q_{01} = \lambda_0 = -q_{00}$, and for $i \geq 1$, $q_{i,i+1} = \lambda_i$, $q_{i,i-1} = \mu_i$ and $q_{ii} = -(\lambda_i + \mu_i)$. Also, $q_{ij} = 0$ for $|i - j| > 1$.

Transition rate diagram of BDP:



The system of differential-difference equations (forward Kolmogorov equations) for the system state probabilities for a BDP are given by

$$p_0'(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t)$$
$$p_i'(t) = \lambda_{i-1} p_{i-1}(t) - (\lambda_i + \mu_i) p_i(t) + \mu_{i+1} p_{i+1}(t), \quad i \geq 1.$$

Example. *(BDP (contd. . .))*
Many queueing systems (where customers arrive/depart one at a time) can be represented as BDPs, where the system state Y_t denotes the number of customers in the system at time t.
► *An M/M/1 queue can be modelled by a BDP with $\lambda_i = \lambda$ and $\mu_i = \mu$ for all i.*

A BDP is a *pure-birth process* if $\mu_i = 0$ for all i. And, a BDP is a *pure-death process* if $\lambda_i = 0$ for all i.

A Poisson process is also a special case of a BDP with $\lambda_i = \lambda$ and $\mu_i = 0$. Recall that we derived the forward Kolmogorov equations for the system state probabilities from basic principles. It can alternatively be derived from the forward Kolmogorov equations of CTMC by noting that that $q_{ii} = -\lambda, q_{i,i+1} = \lambda$ (for $i \geq 0$), and $q_{ij} = 0$ elsewhere (or equivalently from the forward Kolmogorov equations of the BDP).

Anyway, we will continue more in the next lecture. So, let me stop here at this point of time.
Thank you, bye.