Introduction to Queueing Theory

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

Lecture - 12

Stationary and Limiting Distributions of CTMC, Balance Equations, Birth-Death Processes

Hi and hello, everyone; let us continue our discussion of Continuous-Time Markov Chains that we have seen in the previous lecture. Recall

- 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.
  ► Q uniquely determines all transition matrices.

- 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, as an example, what we consider was a BD Birth-Death process.

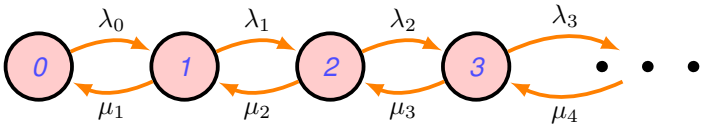
Example. *(Birth-Death Process (BDP))*

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

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

So,

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

is the most important system of equations in the analysis of the birth-death model. If you can obtain a solution to this, it is that is what ideally you would look for; then you can talk about the system or the process, the birth-death process or the which is a continuous-time Markov chain, being in a particular state at a particular time, you can always get these probabilities. So, that is what it is. So, that is what it would mean; this system of equations, the solution to this, is what you ideally look for. Now, whether it is always possible to obtain and if so, how complex is it whether it makes sense that even you are looking for it, that solution is all the question that you ask with respect to the BDP.

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).

So, this is what one would ideally want to do to obtain $p(t)$, whether it is a transition matrix or the state probabilities if one is possible. Just like in the DTMC case, our interest is, what happens in the long run, the long-run behaviour of the continuous-time Markov chain; that is what we are also interested in here as well. So, basically, what you are looking at, much like the DTMC case or the Markov chain case, you are looking at $p_{ij}(t)$, as $t \to \infty$, this is transition probabilities as time tends to ∞ or $p_i(t)$, the state probabilities as $t \to \infty$ is what the quantities that are of interest for you. What you are looking at is limiting probabilities, which is how one can obtain them. This whole idea is exactly similar to what we have done in discrete time.

- As in DTMC, for "nice" CTMCs, a unique stationary distribution exists and equal to the limiting distribution.
- We shall assume the technical assumption $\inf\{\lambda_i, i \in S\} > 0$.

  So, $\lambda_i = 0$ case, which is basically the absorbing case, which we are excluding here by this $\inf\{\lambda_i, i \in S\} > 0$ assumption. Already you know that λ_i is bounded we have made it. Now, we want to assume that $\inf\{\lambda_i, i \in S\}$ is also strictly greater than 0.

- A CTMC is called **irreducible, transient, recurrent** or **positive recurrent** if the defining Markov chain is.

  So, whatever the DTMC that we are using to define the CTMC, whatever the properties that DTMC has, it carries over to that CTMC as well. If the DTMC is irreducible, the CTMC would also be irreducible. Whatever the Markov chain or simply Markov chain or discrete-time Markov chain, if that is positive recurrent, then the corresponding CTMC is also positive recurrent and so on.

- Let $\{Y_t\}$ be a CTMC with transition matrix $P(t)$ and state space $S = \{0, 1, 2, \ldots\}$. A probability distribution $\boldsymbol{p}$ on S, i.e, a vector $\boldsymbol{p} = [p_0, p_1, p_2, \ldots]$, where $p_i \in [0, 1]$ and $\sum_{i \in S} p_i = 1$ is said to be a **stationary distribution** for $\{Y_t\}$ if $\boldsymbol{p} = \boldsymbol{p}P(t)$, for all $t \geq 0$.

  So, in case of discrete also like you have a something similar things that you know, here we want to require that this is for all $t \geq 0$, $\boldsymbol{p} = \boldsymbol{p}P(t)$ be true, and a $\boldsymbol{p}$ which is of $\boldsymbol{p} = \boldsymbol{p}P(t)$ form is what will be called as the Stationary distribution.

- The probability distribution $\boldsymbol{p} = [p_0, p_1, p_2, \ldots]$ is called the **limiting distribution** of the CTMC $\{Y_t\}$ if

$$p_j = \lim_{t \to \infty} P\{Y_t = j | Y_0 = i\} \quad \text{for all } i, j \in S, \text{ and we have } \sum_{j \in S} p_j = 1.$$

  The limiting probabilities may exist, but the limiting distribution may not much like the DTMC case; that is also the same thing here.

So, this is what, ideally, one would want. We basically want to know the long-term behaviour, which is p_j. But again, the interpretation for p_j is like in the Markov chain case; what is the probability that you would see the system being in state j a long-time from now. Whereas stationary distribution, meaning the long time proportion of time that you would find the system to be in a particular state, the same interpretation which we had for the Markov chain. Now,

for example, if you look at the Two-state continuous-time Markov chain, which we have seen.

Example. *(Two-state (CTMC))*
Recall the transition matrix, for any $t \geq 0$*,*

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

With $\boldsymbol{p} = [p_0, p_1]$*, the equations* $\boldsymbol{p} = \boldsymbol{p}P(t), p_0 + p_1 = 1$ *gives* $p_0 = p_1 = 1/2$ *as the stationary distribution. This is also the limiting distribution.*

A "nice" chain with a unique stationary distribution that equals the limiting distribution!

But here, one thing which you will notice with respect to the discrete-time Markov chain and continuous-time Markov chain, here what is that. In discrete-time Markov chain, what is the defining Markov chain here. The defining Markov chain here is

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

that is what is the defining Markov chain. If I go back and if I look here, this is the defining Markov chain that we had for the two-state Markov CTMC case, and

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

that we obtained earlier. So, $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and in the DTMC case, what was the stationary distribution, what was the limiting distribution idea. If you recall, for $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ we said that there is no limiting distribution in the way that we have defined. Because it is never becoming independent of the initial state, but it has a stationary distribution which was $(1/2, 1/2)$. It has a stationary distribution, but it was not having any limiting distribution. But here, the corresponding continuous-time Markov chain has a limiting distribution which is the same as the stationary distribution. So, what was the difference?

What was the property that we had there, which was not guaranteeing as in the case of discrete-time Markov chain, the existence of limiting distribution was the periodicity property, that aperiodicity property. But here, in continuous time, periodicity does not play a role. That is why this happens, and this is a now becomes a nice chain, when in the continuous-time case, this becomes a nice chain with the unique stationary distribution that equals the limiting distribution.

- In theory, we can find the stationary (and limiting) distribution by solving $\boldsymbol{p}P(t) = \boldsymbol{p}$, or by finding $\lim_{t\to\infty} P(t)$.
- In practice, finding $P(t)$ itself is usually very difficult. Hence, direct determination of the **steady-state solution** is more difficult.
  ▶ We need to find alternative ways!

And that is what this theorem gives us. What this says is the following:

Theorem.
For a CTMC, if the embedded DTMC is irreducible and positive recurrent, then there is a unique stationary distribution given by the solution to the stationary equations

$$\mathbf{0} = \boldsymbol{p}Q \quad \textit{and} \quad \sum_{i \in S} p_i = 1.$$

Further, under our assumption that the mean holding times in all states are bounded, the chain has a limiting distribution equal to the stationary distribution.

So, if that is the case, you do not have an absorbing chain; that is what it would effectively mean; that is what it would mean, an irreducible thing. So, already like, we are imposing this condition. So, in that case, the chain has a limiting distribution that equals the stationary distribution. So, now, if you want to get that distribution which is what we call either stationary distribution or limiting distribution, then you do not need to look for solving $\boldsymbol{p}P(t) = \boldsymbol{p}$ after getting $P(t)$ or taking the limit $t \to \infty$ after getting $P(t)$. What you can do you can simply solve $\mathbf{0} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1$.. Now, you know what Q is; there is no t dependency here; it is a single matrix; you solve $\mathbf{0} = \boldsymbol{p}Q$ with $\sum_{i \in S} p_i = 1$ this. And if the chain has this property, irreducible and positive recurrence and mean holding times are all bounded, which means that it stays only a finite amount of time in a particular state and it makes a move to the other state; that is what it would mean. Then, $\mathbf{0} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1.$ is what you are looking for. As the quantity of interest for you to know about the long-term behaviour of this continuous-time Markov chain.

- Now you can see that a CTMC is said to be **regular** it it satisfies the conditions given in the above theorem.
  ▶ For a regular CTMC, the limit $\lim_{t \to \infty} P_{ij}(t) = p_j$ holds for all $i, j \in S$ and is independent of i.
- Compared to DTMCs, *aperiodicity* is not required for the limiting distribution to exist in a CTMC (as the times between transitions vary continuously). Even if the embedded MC is periodic, the continuous transition times wash out any periodicity that may come from the embedded process.
  We have already seen an example. So, the aperiodicity is not required in this case because there is nothing like a period for a continuous time because time is continuous.

So, that periodicity is getting washed out. So, there is no periodicity concept as such for a continuous-time Markov chain. You can take it that way. So, we are not looking at that. So, basically, that is why irreducibility and positive recurrence is what is; in the discrete-time case, the stationary equations were different. But the solution of that was possible if this (irreducible and positive recurrent) was the case. And in the case of the aperiodic Markov chain, we said that the limiting distribution exists and then equals the stationary distribution. But here, we do not need that aperiodicity condition is what is the difference that you would see with respect to the Ergodic theorem for the continuous-time Markov chain, which is what this result is all about, I mean much like the similar case for the discrete-time Markov chain case. So, if we are looking for long-term behaviour, limiting distribution, or the stationary distribution, they will be then, under such a situation, irreducible and positive recurrence and this boundedness. So, they will be given by the solution of $\mathbf{0} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1.$, which is an alternative way and which is a very simple one to do now; just like in the discrete-time Markov chain case, you have to find P^n to get the limiting probabilities and then let $n \to \infty$. Instead of that, you solved it through $\pi = \pi P$, and then

you obtained that same quantity. Here, this is $\mathbf{0} = \boldsymbol{p}Q$ is; what is the equation that you need to solve to get the distribution. If you get these p_i's, if it forms a distribution, then that is the stationary distribution, and that is what the limiting distribution in this under the conditions given in this Ergodic theorem, and that is what you want to look at it.

Then, this **p**'s whatever you obtained, so obtain as a solution to $\mathbf{0} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1$, is what will give you the probabilities of finding the continuous-time Markov chain to be in a particular state a long time from now; the interpretation is by a limiting distribution or in the long-run fraction of time, you find the process in a particular state, this is basically from the stationary distribution point of view. So, you can have both interpretations for these **p**'s that we have here.

Now, $\mathbf{0} = \boldsymbol{p}Q$ equation, if I look at it, because now, I need to worry about only solving this set of equations, because this is what will give me the equilibrium behaviour or the steady-state behaviour or the stationary behaviour or the limiting behaviour of the continuous-time Markov chain. All these words are used interchangeably in different contexts; you be aware of that.

- The equation $\mathbf{0} = \boldsymbol{p}Q$ is equivalent to an equation system

$$\sum_{i \neq j} p_i q_{ij} = -p_j q_{jj} \quad \Leftrightarrow \quad \sum_{i \neq j} p_i q_{ij} = p_j \sum_{i \neq j} q_{ji} \quad \text{for all } j \in S.$$

  ▶ On LHS, $p_i q_{ij}$ is the rate of transitions from i to j (or stochastic flow from i to j in equilibrium).
  So, p_i is basically the probability of finding the system or the process or the chain in state i, and q_{ij} is the rate at which it is moving from i to j. So, the total product $p_i q_{ij}$ is the rate of transitions from i to j because you are finding it in i, and then it moves from i to j. So, that product $p_i q_{ij}$ will give you the rate of transition from i to j or the stochastic flow from i to j.
  Summing over i gives the overall rate of transitions into state j.

  ▶ The RHS is the rate of transitions out of state j.

  ▶ Thus, $\mathbf{0} = \boldsymbol{p}Q$ means that the rate of transitions out of a state equals the rate of transition into the state, in equilibrium (or in steady-state). These are called the (global) **balance equations**.

Example. *(Poisson Process)*
$\boldsymbol{p}Q = \mathbf{0}$ *gives* $p_0\lambda = 0$ *and* $p_i\lambda = p_{i-1}\lambda$ *for all* $i \geq 1$.
This implies that $p_i = 0$ for all $i \in S$ and there is no stationary distribution.

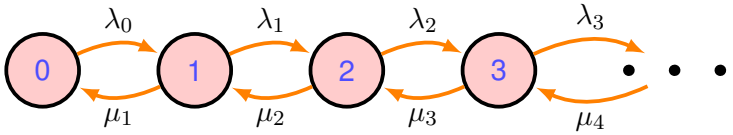
Example. *[Two-state CTMC]*
We have $Q = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}$. *Then,* $\boldsymbol{p}Q = \mathbf{0}$ *gives* $p_0 = p_1$, *and* $p_0 + p_1 = 1$ *would then implies that* $p_0 = p_1 = 1/2$.

Now, let us come back to a Birth-Death Process, the most important one that we are going to use next.

- Recall: A CTMC $\{Y_t, t \geq 0\}$ on $S = \{0, 1, 2, \dots\}$ with the transition rates $q_{i,i+1} = \lambda_i$ for $i \geq 0$, $q_{i,i-1} = \mu_i$ for $i \geq 1$ and $q_{ij} = 0$ for $|i - j| > 1$ is called a **Birth-Death Process (BDP)**.

  ▶ We assume that $\lambda_i > 0$ for $i \geq 0$ and $\mu_i > 0$ for $i \geq 1$ (and are finite).

- Transition rate diagram of BDP:



- Balance Equations (for the state probabilities):

$$\lambda_0 p_0 = \mu_1 p_1$$
$$(\lambda_i + \mu_i) p_i = \lambda_{i-1} p_{i-1} + \mu_{i+1} p_{i+1}, \quad i \geq 1.$$

Let us take the case of state 2 here. Now, what is the rate of flow out of state? So, this the rate of flow out of state is either it can go move to 3 with rate λ_2 or to 1 with rate μ_2. So, the rate of flow out of state 2 is basically $(\lambda_2 + \mu_2)p_2$; that is what you will get on the left-hand side here, which is the rate of flow out of state 2. And, what is the rate of flow into state 2, that it can be in state 1 and with λ_1, it can come to state 2, or it can be in state 3, and with rate μ_3, it can come to state 2. So, $\lambda_1 p_1 + \mu_3 p_3$ is what you will get for $i = 2$ here. $\lambda_1 p_1 + \mu_3 p_3$ is the rate of flow into state i, $(\lambda_2 + \mu_2)p_2$ is the rate of flow out of state i, and in equilibrium, $(\lambda_2 + \mu_2)p_2 = \lambda_1 p_1 + \mu_3 p_3$. So, this is what is the flow balance equation.

Now, this one you can also obtain you have the Q already, and then you can look $pQ = 0$, you will give you this or from this diagram you can immediately write it down, what is the flow balance equations, in which must be satisfied in equilibrium.

- The BDP is irreducible. If it is also positive recurrent, then we will have a unique solution to the above equations and it is called as stationary distribution or limiting distribution or equilibrium distribution or steady-state distribution for the BDP.

Again, you look at the theorem and how we are using it, recall the theorem irreducible positive recurrent implies there is a unique solution to this $\mathbf{o} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1$, set of equations.

Now, we ensured irreducibility; we have ensured the boundedness; we do not know whether it is positive recurrent or not or under what condition this will be positive recurrent. So, we now look at the solution to $\mathbf{o} = \boldsymbol{p}Q$ and $\sum_{i \in S} p_i = 1$. Now, when under the what condition this system has a unique solution and that is the condition for positive recurrent that is how you now we will interpret. So, if it is positive recurrent, then we will have a unique solution to the above equations. And it is then called a stationary distribution or limiting distribution because we know all of them are these two things is what is the main quantities they are one and the same. And whenever one and the same, the additional words terminologies that are used to describe are also called equilibrium distribution or steady-state distribution for the BDPs.

- We now address the existence of steady-state probability distribution.

- Define two sums:

$$S_1 = \sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \qquad \text{and} \qquad S_2 = \sum_{k=0}^{\infty} \left(1 \Big/ \left(\lambda_k \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) \right)$$

Case-1: BDP is positive recurrent if and only if $S_1 < \infty$ and $S_2 = \infty$.
Case-2: BDP is null recurrent if and only if $S_1 = \infty$ and $S_2 = \infty$.
Case-3: BDP is transient if and only if $S_1 = \infty$ and $S_2 < \infty$.

So, if I want only the recurrence or transience of this, then I can take it only through S_2. Because if $S_2 < \infty$, then it is transient; if $S_2 = \infty$, it is recurrent. Now, within this recurrence, if I want the positive or null recurrence to be separated, then I have to look at essentially this S_1, which is what is the most crucial one here. If $S_1 = \infty$, then it is null recurrent, and if $S_1 < \infty$, then it is, positive recurrent. So, mainly, our interest is whether the chain is positive recurrent or not. If that is the scenario, I can look at only S_1 will be finite; S_2 is infinite; obviously, you have to verify. But you know, we will be mostly concerned or will be without saying explicitly we will be concerned with S_1, if S_1 being finite or not, if S_1 is infinite, then I know that it could be one of these cases, but S_2 needs to be checked, but it will satisfy most of the case anyway. So, we will not worry too much about this part. So, this becomes the major or the main one which will contribute to our positive recurrence of this chain.

- It is Case-1 that gives rise to equilibrium probabilities and this is of interest to our studies.

Where both S_1 and S_2 though is needed, we will not say explicitly about S_2, which you can verify by yourself, but it will be in terms of S_1 we will talk.
► Note that the corresponding condition is met whenever the sequence $\{\lambda_n/\mu_n\}$ remains below unity from some n onwards.

- An irreducible BDP on a finite state space $S = \{0, 1, 2, \ldots, N\}$ is said to be a finite-state BDP or finite BDP and is always positive recurrent.

So, this is what is the BDP theory, and then we will come back to this BDP once more when we start the next lectures, which is on the basic queueing models. But this is what we may have; we may not have everything, but this is what you will have;

$$\lambda_0 p_0 = \mu_1 p_1$$
$$(\lambda_i + \mu_i) p_i = \lambda_{i-1} p_{i-1} + \mu_{i+1} p_{i+1}, \quad i \geq 1.$$

is what is called the global balance equation, and a solution to the global balance equation is what you are obtaining. So, this can also be obtained from that the stationary equation $pQ = 0$, like this also gives you these flow-balance equations.

So, we will stop here and about our discussion on this continuous-time Markov chain; what we will do next lecture is we will start the discussion of our basic queueing models, which will be based on this BDP. So, again we will revisit BDP once more when we start the queueing models. So, this is all our basically; our idea is what is a continuous-time Markov chain, what is required to describe a continuous-time Markov chain, and how you will obtain this long-term behaviour. So, the long-term behaviour, the equilibrium behaviour is what we are looking at it for most of the queueing models that we will do, but we will also highlight at one point of time for the simplest of the model, what is the transient solution, which is basically obtaining $P(t)$ the state probabilities at time t explicitly. And you can see, then the complexity of obtaining that $P(t)$ in that particular case, that is what we will be doing, but for all these things, this is what is the main thing that we will be using it. So, this Ergodic theory theorem for this CTMC is what is the main

backbone on which the model's solutions are all obtained for various queueing models. Fine, we will talk about that when we start this queueing model. So, we will end here.
Thank you. Bye.