Introduction to Queueing Theory

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

Lecture - 43
G/M/1 Queues

Hi and hello, everyone. What we have been seeing in the class of semi-Markovian Queues, where the service time was generally distributed, are the M/G type queues; we just saw $M/G/1$ in detail, much like how one would analyze an $M/M/1$ model. In a similar way, we did the complete analysis of an $M/G/1$ model, whether it is being system size, waiting time, busy period, or everything like we have done in that way. Then extension, we just highlighted what are the things that we can do with respect to that. So, that is the one type of semi-Markovian model where we had all the assumptions of $M/M/1$ except that the service time distribution was exponential. So, that was the idea of this $M/G/1$ type model.

Now, next, what we will consider is basically the another type of semi-Markovian model, where we retain all the assumptions of $M/M/1$, but we relax only the condition of the Poisson process arrivals, meaning that the inter-arrival times are now generally distributed random variable, but they are still IID. We are not relaxing that they are IID, but it is not exponentially distributed; the distribution of the inter-arrival times is any general distribution; obviously, this is a nonnegative distribution is what we have in mind. So, that is what we have. So, such models are called $G/M/1$ queues or $G/M/1$ queueing models. So, the systems would correspondingly call the $G/M/1$ system queueing system. So, we have the setting, we have all the assumptions of $M/M/1$ in place except the assumption of Poisson arrivals, but now inter-arrival times follow a general distribution, and they are all IID. So, we are then dealing with $G/M/1$ type of queueing systems, and again we are looking at the equilibrium analysis.

So for the analysis of this $G/M/1$ queue, the process that we are going to follow is like the one that we had for an $M/G/1$ queue. And the approach that we are going to use is the same as the one we have used for $M/G/1$ model, which is the embedded Markov chain technique. So, the embedded Markov chain approach is what we are going to use here as well. But now, the difference is that to get this embedded Markov chain, in the $M/G/1$, we looked at the departure epochs, but in $G/M/1$, we will look at now arrival epochs to extract an embedded Markov chain. That is the difference, except that the approach will remain the same, the points at which you are looking at the system or arrival epochs rather than the departure epochs. Which is what are the case with $M/G/1$; in $G/M/1$, it will be the arrival epochs is what we are going to consider. So, at those arrival points, at those arrival times, we are looking at the system, and we could define this X_n, $X_n = N(t_n-)$ meaning there is the number of customers in the system just prior to the arrival of the nth customer. As usual, this t_n say t_1, t_2, t_3, t_4 and so on; if you consider that as arrival epochs, then just prior to arrival like what was the system size is what we are calling it is by X_n. Then,

$$X_{n+1} = X_n + 1 - B_n, \quad \text{if } B_n \leq X_n + 1,\ X_n \geq 0,$$

where B_n is the number of customers served during the interarrival time $t_{n+1} - t_n$ between the n^{th} and $(n+1)^{st}$ arrivals. Since $t_{n+1} - t_n$'s are IID, they have a common CDF by $A(t)$ and B_n's are also IID.

So, basically, this is what is the relationship. Now, is this relationship a Markov chain? As you can see, B_n does not depend on the past history of the queue, like what was the system sizes prior to the nth arrival. So, given this X_n at time of nth arrival so and hence $\{X_0, X_1, X_2, \dots\}$ is a Markov chain, you can easily see from this $X_{n+1} = X_n + 1 - B_n$ like $X_n + 1$ depends only on X_n and something that is happening during this interval of length or during the interval t_n to t_{n+1} and which is independent of the past history. So, this is a Markov chain. So, $\{X_0, X_1, X_2, \dots\}$ is a Markov chain, and hence this is the embedded Markov chain for this semi-Markov process. Again, semi Markov process in terms of X_n you can define it much like the $M/G/1$ case and so on. So, there is no difference between them; that is clear. So, we have this semi-Markov process which has $\{X_0, X_1, X_2, \dots\}$ as the embedded Markov chain.

Now, if you define the probability that there are exactly k service completions between two consecutive arrivals, given that there is at least k just possibly the previous arrival. So, then that would be equal to

$$b_k = \int_0^\infty \frac{e^{-\mu t}(\mu t)^k}{k!} dA(t)$$

because there has to be at least k service completion to happen; that was the idea.

So, b_k suppose, if much like k_i that we defined in $M/G/1$ model, here we are defining it as b_k which is basically; during one arrival time what is the probability that there are k service completion happens. Again, since the service times follow exponential, if you look at an interval of length t and if you are looking at the number of service completion in that particular length, that will be a Poisson process with rate μt, and hence you will get $\frac{e^{-\mu t}(\mu t)^k}{k!}$. So, that is what you get the k service completion in that interval. So, this is what your b_k like; you see here the similarity you can always make this analogy of similar things with what you have done for $M/G/1$. So, that is $b_k = P\{B_n = k | X_n \geq k\}$. Then the one-step transition probability matrix for this embedded Markov chain $\{X_0, X_1, X_2, \dots\}$ you can easily write down.

$$P = ((p_{ij})) = \begin{bmatrix} 1 - b_0 & b_0 & 0 & 0 & 0 & \cdots \\ 1 - \sum_{k=0}^{1} b_k & b_1 & b_0 & 0 & 0 & \cdots \\ 1 - \sum_{k=0}^{2} b_k & b_2 & b_1 & b_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \end{bmatrix}$$

So, here this structure is what is called as $G/M/1$ type matrix.

Because in queueing theory like this is so common to refer to such a structure. We just said we, in the case of $M/G/1$, also we had an $M/G/1$ type of structure. What was that? Like in the $M/G/1$ type, we had this main diagonal and the lower; main diagonal is what is nonnegative upper side was everything was nonnegative, the lower side were all completely 0. So, that structure we call it as $M/G/1$ type structure. Here it is just the reverse. You see here the leading diagonal and the main diagonal above that, is what is nonnegative quantities and above that everything is 0, below that everything is a nonnegative quantity so this is called $G/M/1$ type matrix. The reason is that when you go in a much more general setup, like these kinds of structures, then one can handle it along in a similar way like what we are handling this $M/G/1$ or $G/M/1$ type. So, the single-step transition probabilities are one-step transition probabilities for this Markov chain is simply given by this chain or this P matrix.

Now, this P matrix from this, the chain is immediately you can see that it is irreducible and aperiodic. And it can be shown that it is positive recurrent; when this is greater than 1, means that $\sum_{n=1}^{\infty} nb_n > 1$. And in such a case, the chain is ergodic, and a steady-state solution exists, which is equal to limited distribution, and the stationary distribution, together, is what is generally referred to as steady-state solution. And if you denote it by this vector $\mathbf{a} = \{a_n\}, \quad n = 0, 1, 2, \ldots,$ which is the probability vector that an arrival finds n in the system is what is your a_n. Then this a_n satisfies the usual stationary equations $\mathbf{a}P = \mathbf{a}$ and $\mathbf{ae} = 1$, which if you write explicitly for this matrix P, then you will end up with

$$
\begin{aligned}
a_i &= \sum_{k=0}^{\infty} a_{i+k-1} b_k, \quad i \geq 1, \\
a_0 &= \sum_{j=0}^{\infty} a_j \left(1 - \sum_{k=0}^{j} b_k \right).
\end{aligned}
$$

So, there is a difference between this expression and M/G type expression because here, the summation is all infinite summations, and of course, this works to our advantage in a way when we are trying to analyze this. So, this is what is the stationary equations. Now, we need to get the solution to this to get these a_i's; ultimately, we have to solve this to get a_i's.

- We now use the operators method to solve the equations. Let $Da_i = a_{i+1}$. For $i \geq 1$, we can write the above equation as

$$
\begin{aligned}
& a_i - (a_{i-1}b_0 + a_i b_1 + a_{i+1} b_2 + \ldots) = 0 \\
\implies \quad & a_{i-1}(D - b_0 - Db_1 - D^2 b_2 - D^3 b_3 - \ldots) = 0
\end{aligned}
$$

- The characteristic equation for this difference equation is

$$z - b_0 - zb_1 - z^2 b_2 - z^3 b_3 - \cdots = 0, \quad i.e., \sum_{n=0}^{\infty} b_n z^n = z.$$

- Since $\{b_n\}$ is a probability distribution, the LHS is the PGF and hence the above becomes

$$\boxed{\beta(z) = \sum_{n=0}^{\infty} b_n z^n = z.}$$

- As in the case of $M/G/1$, it can be shown that $\beta(z) = A^*[\mu(1-z)]$, where A^* is LST of the interarrival-time CDF, and hence the above equation can be written as

$$\boxed{z = A^*[\mu(1-z)].}$$

- If we can find solutions of the characteristic equation then we can determine $\{a_n\}$.
- Assume that $\rho = \frac{\lambda}{\mu} < 1$. Then it can be shown that there is exactly one real root, say r_0, in $(0, 1)$ (We will prove this below).

  We can show that there is only exactly one real root under the assumption that $\rho < 1$, which is λ/μ, λ is the arrival rate or $1/\lambda$ is the mean of the inter-arrival time distribution is what we are assuming. So, that is what we will show a bit in a moment, but once we if we understand that there is only one real root r_0, then this distribution, so, then this a_n's; what is this a_n's now? a_n's are nothing but the arrival point system size probabilities because we are only talking about the embedded Markov chain, which are embedded at the arrival epochs. So, arrival epochs, then is, basically would be equal to

$$a_n = Cr_0^n, \quad n \geq 0.$$

  because there is only one real root. So, the solution is given by $a_n = Cr_0^n, \quad n \geq 0$ by our usual theory of difference equation approach, operator method, and difference equations.

  Then C can be determined using the usual normalization condition of total probability being equal to 1, which gives us $C = 1 - r_0$, which is basically the steady-state arrival point or arrival epoch system size distribution as in a complete form as

$$a_n = (1 - r_0)r_0^n, \quad n \geq 0, \ \lambda/\mu < 1.$$

- So, before we go further, we will try to prove this fact, which is basically for $\beta(z) = z$, you have only one real root say r_0, that is what we will try to prove. We will now prove that r_0 is the single root of the characteristic equation.
- We consider the two sides of the equation $\beta(z) = \sum_{n=0}^{\infty} b_n z^n = z$ separately as

$$y = \beta(z) \quad \text{and} \quad y = z.$$

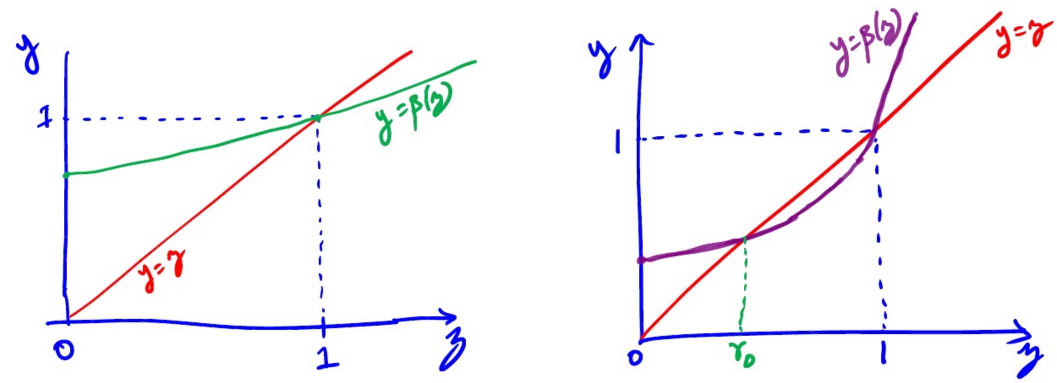
- Observe that $0 < \beta(0) = b_0 < 1 \quad \text{and} \quad \beta(1) = \sum_{n=0}^{\infty} b_n = 1.$
- Also, $\beta(z)$ is monotonically nondecreasing and convex because

$$\beta'(z) = \sum_{n=1}^{\infty} nb_n z^{n-1} \geq 0,$$
$$\beta''(z) = \sum_{n=1}^{\infty} n(n-1)b_n z^{n-2} \geq 0.$$

- Further, since the service times are exponential, each b_n is strictly positive (i.e., $b_n > 0$ for $n \geq 0$), and this implies that $\beta(z)$ is strictly convex.

So, this $\beta(Z)$ is strictly convex is what we could observe about these functions that we have noticed here. So, these are all; see anything when you want to look at the solution of such an equation. So, you just try to analyze what is the nature of this function, and that is precisely what we are trying to do. So, this is $\beta(Z)$; what is $beta(0)$, $beta(1)$, and β property nondecreasing and strictly convex everything you are observing.

- Now, this gives us two possible cases for the graphs of $y = \beta(z)$ and $y = z$. So, let us draw one by one what are the two possibilities.



- Either there are no intersections in $(0, 1)$, or there is exactly one intersection in $(0, 1)$.
  ▶ The latter case occurs when $\beta'(1) = E[\text{number served during interarrival time}] = \mu/\lambda > 1$.
  ▶ That is, when $\rho = \lambda/\mu < 1$, there is exactly one root r_0 in $(0, 1)$.

So, under the condition of $\rho = \dfrac{\lambda}{\mu} < 1$, whatever we have assumed for, the case where we have this one positive recurrence. Because now you see, under $\rho = \dfrac{\lambda}{\mu} < 1$ condition, it can be shown that we said, but this is exactly the condition for the positive recurrence of the embedded Markov chain. So, under that condition, what happens? This does not happen, and this is what happens, and hence there is exactly one real root in the (01), which is what gives us $a_n = (1 - r_0)r_0^n$ which is the arrival point system size distribution.

- But we can show that r_0 is the only complex root with absolute value less than 1 by using Rouche's theorem, and that is what this argument is or is there available in the text as well.
  You may like to go through if you want to just use that because this equation or the variable is actually a complex variable $\beta(z) = z$. So, if you want to show that this is the only complex root with absolute value less than 1, then you have to use Rouche's theorem again under the assumption

  ▶ $\beta'(1) = 1/\rho > 1$.
  ▶ Let $f(z) = -z$ and $g(z) = \beta(z)$.
  Because $g(1) = 1$ and $g'(1) > 1$, we have $g(1 - \epsilon) < 1 - \epsilon$ for small enough $\epsilon > 0$.
  Consider the set z such that $|z| = 1 - \epsilon$. By the triangle inequality,

$$|g(z)| \leq \sum_{n=0}^{\infty} b_n |z|^n = g(1 - \epsilon) < 1 - \epsilon = |f(z)|.$$

By Rouche's theorem, $f(z) = -z$ and $f(z) + g(z) = -z + \beta(z)$ have the same number of roots within the contour $|z| = 1 - \epsilon$.
Since ϵ can be made arbitrarily small, there is exactly one complex root of $z = \beta(z)$ whose absolute value is less than one.
Thus, it must be the real root r_0 found earlier.

So, this is not that if you want to complete the proof, then that is what you need to understand. But this is not very difficult. Now, how do we find this r_0?

- It is not that at every time like you will find this r_0 in explicit form, but usually, it involves numerical procedures. But you know the advantage is that it is guaranteed that you can obtain it readily in a very easy way.

  ▶ For example, the method of successive substitution

$$z^{(k+1)} = \beta(z^{(k)}), \quad k = 0, 1, 2, \dots, \ 0 < z^{(0)} < 1$$

  is guaranteed to converge because of the shape of $\beta(z)$.

- In summary, the steady state arrival-point system size distribution is

$$a_n = (1 - r_0) r_0^n, \quad n \geq 0, \ \rho = \lambda/\mu < 1.$$

- Note the analogy between the above and that of $M/M/1$ (keeping in mind the fact the above are arrival-point probabilities, and not general-time probabilities).
  ▶ Therefore, using a similar formula of $M/M/1$, we can get the expected measures only at arrival time.

- Unlike $M/G/1$, here for $G/M/1$, it is not true here that $a_n = p_n$ in general.

  If so, then one can also write a_n as p_n. So, that is the minus with respect to comparison of $M/G/1$ and $G/M/1$ that here you have to obtain these arrival point probabilities. But is there a relationship between them yes, there exists a relationship between them which was proved way back in the 1950s itself, and the relationship later, different kinds of proofs were given. And basically, they can be shown that there exists a relationship between them, and it can be obtained as $p_n = \rho a_{n-1}$ for $n \geq 1$ is what then you would find as the relationship between these two. But we are not going into the proof of this; if you all want to look at it, you can look at it elsewhere, but it is also proved long back anyway.
  ▶ But, also, $a_n = p_n$ here if and only if the arrivals are Poisson, i.e., $G = M$.

In all other cases, this is not equal to p_n, $a_n \neq p_n$, but p_n it can be shown to be equal to ρa_{n-1}. This is what then the relationship. We are not going into the details, but of course, this is what is the relationship that one can understand; you can look at Ross's book, for example, to get the idea of how it can be proved in a probabilistic way very simple arguments he uses to prove this. So, this is what we have to keep in mind. So, in comparison to this $M/G/1$ in $G/M/1$, whatever we are obtaining is all with respect to the arrival point, not at general time points. If you want to relate, then you have to use such a relationship to relate to the general time points.

- We use a superscript (A) to denote the particular measure of effectiveness is taken relative to arrival points only. We have

$$L^{(A)} = \frac{r_0}{1-r_0}, \quad L_q^{(A)} = \frac{r_0^2}{1-r_0}$$

- The line delay and the system-waiting-time distribution functions, $F_{T_q}(t)$ and $F_T(t)$ can also be obtained from $M/M/1$ by replacing ρ with r_0 to yield

$$F_{T_q}(t) = 1 - r_0 e^{-\mu(1-r_0)t}, \quad t \geq 0,$$
$$F_T(t) = 1 - e^{-\mu(1-r_0)t}, \quad t \geq 0,$$

with mean values

$$W_q = \frac{r_0}{\mu(1-r_0)}, \quad \text{and} \quad W = \frac{1}{\mu(1-r_0)}.$$

Note that the probability that a customer does not have to wait is given by $1 - r_0$.

⧫ The above results refer to the distribution of the waiting times as observed by customers arriving to the system.

So, it should not be interpreted as there is some difference between these and something called which is virtual waiting time, but we will not go into the details fine. So, that is what you need to keep in mind.

Example. *[M/M/1]*

For exponentially distributed interarrival times, we have $A^*(s) = \lambda/(s+\lambda)$.
Hence, the equation $z = A^*(\mu - \mu z)$ reduces to $z = \dfrac{\lambda}{\lambda + \mu - \mu z} \Rightarrow (z-1)(\lambda - \mu z) = 0$.
Thus, the desired root is $r_0 = \rho = \lambda/\mu$ and hence the arrival-point distribution is

$$a_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots.$$

Note that this is also the steady state solution of $M/M/1$, meaning that this is also the arbitrary time point distribution in an $M/M/1$ system, as you expect for a Poisson process case. So, that is what you are seeing it here.

Example.
Suppose that the interarrival time consist of two exponential phases, the first phase with parameter μ and the second one with parameter 2μ, where μ is also the parameter of the exponential service time. The Laplace-Stieltjes transform of the interarrival time is then

$$A^*(s) = \frac{2\mu^2}{(s+\mu)(s+2\mu)}.$$

Hence, the equation $z = A^*(\mu - \mu z)$ *reduces to* $z = \dfrac{2\mu^2}{(2\mu - \mu z)(3\mu - \mu z)} = \dfrac{2}{(2-z)(3-z)} \Rightarrow (z-1)(z-2-\sqrt{2})(z-2+\sqrt{2}) = 0$.

Thus, the desired root is $r_0 = 2 - \sqrt{2} = 0.5858$ and hence the arrival-point distribution is

$$a_n = (\sqrt{2}-1)(2-\sqrt{2})^n, \quad n = 0, 1, 2, \ldots.$$

There is another example.

Example.

Suppose that there is a single-server queueing system with exponentially distributed service times, but there is no basis for assuming either exponential or E_k as interarrival times.

From data, it was observed that a k-point distribution fits well the interarrival times. That is,

$$P\{\textit{interarrival time } = t_i\} = a(t_i) = a_i, \quad 1 \le i \le k.$$

We must find the root r_0 of the characteristic equation $z = A^*[\mu(1-z)] = \sum_{i=1}^{k} a_i e^{-\mu t_i(1-z)}$.

For illustration, assume a 3-point distribution $a_1 = a(2) = 0.2, a_2 = a(3) = 0.7, a_3 = a(4) = 0.1$ (assuming time in minutes).

One can then find the root r_0 of $z = A^*[\mu(1-z)]$ using successive substitution and then obtain the performance measures.

Exercise: Complete the exercise for the above assuming that $1/\mu = 2$ minutes. The root will be $r_0 = 0.467$.

These are some examples of $G/M/1$; it gives a bit more flexibility here. Now, we are not going to go into detail, but we just look at what will happen or whether what can all can happen in the case of the $G/M/c$ queue, which is a multi-server case. But here, see the one disadvantage we in I mean you can say a disadvantage with respect to $G/M/1$ as compared to $M/G/1$ was the arrival point probabilities were not equal to the arbitrary time probabilities, but there is a relationship between them. But the advantage is here that, unlike this $M/G/c$ case where the embedded Markov chain was not available to model this $M/G/c$ system, in case of $G/M/c$, the embedded Markov chain approach will still be applicable. So, whatever we have done for $G/M/1$ theoretically in principle, you can apply a similar process to do the analysis of this $G/M/c$ model, the multi-server G/M models. So, what will the complexity be in terms of the b_n's, how does it affect the embedded matrix, how does the embedded Markov chain stand in probability matrix, and how does it affect the root-finding problem. The mean service time would be either $n\mu$ or $c\mu$ depending upon the state. So, and hence b_n will now depend on both i and j.

And after some lengthier steps, one can obtain the system size probabilities, and hence you can also obtain the line delay distributions; possible. It is possible that you can use the point here is; this embedded Markov chain approach is still applicable in the case of multi-server G/M type models, which was not the case in M/G type multi-server models; that is the difference that you can find. Of course, this is available in the text; we are not going to go into that because of time, but it is along similar lines; you can do that. Similarly, the other extension of $G/M/1$ much like the M/G models that we have considered. So, it can be obtained especially for single server models like this Cohen; it is a thick book on single server queues, is a very classical book. For more details of various other features, one can refer to other general extensions for this $G/M/1$ models. So, this is all about our discussion of this $G/M/1$ model. With that, we are also closing our discussion of this semi-Markovian queueing systems, wherein we consider mainly $M/G/1$ its analysis, $G/M/1$ and its analysis. Some extension with respect to the M/G type models and G/M pretty much we did not do anything except this slide where we have hinted how one can handle this multi-server G/M model. So, with that, we close the discussion of this semi-Markovian queueing systems.

Thank you, bye.