16.22: Continuous-Time Queuing Chains

Basic Theory

Introduction

In a *queuing model*, *customers* arrive at a station for *service*. As always, the terms are generic; here are some typical examples:

- The customers are persons and the service station is a store.
- The customers are file requests and the service station is a web server.

Figure 16.22.1: Ten customers and a server

Queuing models can be quite complex, depending on such factors as the probability distribution that governs the arrival of customers, the probability distribution that governs the service of customers, the number of servers, and the behavior of the customers when all servers are busy. Indeed, queuing theory has its own lexicon to indicate some of these factors. In this section, we will discuss a few of the basic, continuous-time queuing chains. In a general sense, the main interest in any queuing model is the number of customers in the system as a function of time, and in particular, whether the servers can adequately handle the flow of customers. This section parallels the section on discrete-time queuing chains.

Our main assumptions are as follows:

1. There are $k \in \mathbb{N}_+ \cup \{\infty\}$ servers.
2. The customers arrive according to a Poisson process with rate $\mu \in (0, \infty)$.
3. If all of the servers are busy, a new customer goes to the end of a single line of customers waiting service.
4. The time required to service a customer has an exponential distribution with parameter $\nu \in (0, \infty)$.
5. The service times are independent from customer to customer, and are independent of the arrival process.

Assumption (b) means that the times between arrivals of customers are independent and exponentially distributed, with parameter μ. Assumption (c) means that we have a *first-in, first-out* model, often abbreviated *FIFO*. Note that there are three parameters in the model: the number of servers k, the exponential parameter μ that governs the arrivals, and the exponential parameter ν that governs the service times. The special cases $k = 1$ (a single server) and $k = \infty$ (infinitely many servers) deserve special attention. As you might guess, the assumptions lead to a continuous-time Markov chain.

Let X_t denote the number of customers in the system (waiting in line or being served) at time $t \in [0, \infty)$. Then $\boldsymbol{X} = \{X_t : t \in [0, \infty)\}$ is a continuous-time Markov chain on $\mathbb{N}$, known as the *M/M/k queuing chain*.

In terms of the basic structure of the chain, the important quantities are the exponential parameters for the states and the transition matrix for the embedded jump chain.

For the M/M/k chain $\boldsymbol{X}$,

1. The exponential parameter function λ is given by $\lambda(x) = \mu + \nu x$ if $x \in \mathbb{N}$ and $x < k$ and $\lambda(x) = \mu + \nu k$ if $x \in \mathbb{N}$ and $x \geq k$.
2. The transition matrix Q for the jump chain is given by

$$Q(x, x-1) = \frac{\nu x}{\mu + \nu x}, \; Q(x, x+1) = \frac{\mu}{\mu + \nu x}, \quad x \in \mathbb{N}, \; x < k$$

$$Q(x, x-1) = \frac{\nu k}{\mu + \nu k}, \; Q(x, x+1) = \frac{\mu}{\mu + \nu k}, \quad x \in \mathbb{N}, \; x \geq k$$

So the M/M/k chain is a birth-death chain with 0 as a reflecting boundary point. That is, in state $x \in \mathbb{N}_+$, the next state is either $x - 1$ or $x + 1$, while in state 0, the next state is 1. When $k = 1$, the single-server queue, the exponential parameter in state $x \in \mathbb{N}_+$ is $\mu + \nu$ and the transition probabilities for the jump chain are

$$Q(x, x-1) = \frac{\nu}{\mu + \nu}, \; Q(x, x+1) = \frac{\mu}{\mu + \nu} \tag{16.22.1}$$

When $k = \infty$, the infinite server queue, the cases above for $x \geq k$ are vacuous, so the exponential parameter in state $x \in \mathbb{N}$ is $\mu + x\nu$ and the transition probabilities are

$$Q(x, x-1) = \frac{\nu x}{\mu + \nu x}, \; Q(x, x+1) = \frac{\mu}{\mu + \nu x} \tag{16.22.2}$$

Infinitesimal Generator

The infinitesimal generator of the chain gives the same information as the exponential parameter function and the jump transition matrix, but in a more compact form.

For the M/M/k queuing chain $\boldsymbol{X}$, the infinitesimal generator G is given by

$$\begin{aligned} G(x, x) &= -(\mu + \nu x), \; G(x, x-1) = \nu x, \; G(x, x+1) = \mu; \quad x \in \mathbb{N}, \; x < k \\ G(x, x) &= -(\mu + \nu k), \; G(x, x-1) = \nu k, \; G(x, x+1) = \mu; \quad x \in \mathbb{N}, \; x \geq k \end{aligned}$$

So for $k = 1$, the single server queue, the generator G is given by $G(0, 0) = -\mu$, $G(0, 1) = \mu$, while for $x \in \mathbb{N}_+$, $G(x, x) = -(\mu + \nu)$, $G(x, x-1) = \nu$, $G(x, x+1) = \mu$. For $k = \infty$, the infinite server case, the generator G is given by $G(x, x) = -(\mu + \nu x)$, $G(x, x-1) = \nu x$, and $G(x, x+1) = \mu$ for all $x \in \mathbb{N}$.

Classification and Limiting Behavior

Again, let $\boldsymbol{X} = \{X_t : t \in [0, \infty)\}$ denote the M/M/k queuing chain with arrival rate μ, service rate ν and with $k \in \mathbb{N}_+ \cup \{\infty\}$ servers. As noted in the introduction, of fundamental importance is the question of whether the servers can handle the flow of customers, so that the queue eventually empties, or whether the length of the queue grows without bound. To understand the limiting behavior, we need to classify the chain as transient, null recurrent, or positive recurrent, and find the invariant functions. This will be easy to do using our results for more general continuous-time birth-death chains. Note first that $\boldsymbol{X}$ is irreducible. It's best to consider the single server and infinite server cases individually.

The single server queuing chain $\boldsymbol{X}$ is

1. Transient if $\nu < \mu$.
2. Null recurrent if $\nu = \mu$.
3. Positive recurrent if $\nu > \mu$. The invariant distribution is the geometric distribution on $\mathbb{N}$ with parameter μ / ν. The invariant probability density function f is given by

$$f(x) = \left(1 - \frac{\mu}{\nu}\right) \left(\frac{\mu}{\nu}\right)^x, \quad x \in \mathbb{N} \tag{16.22.3}$$

Proof

This follows directly from results for the continuous-time birth-death chain, with constant birth rate μ on $\mathbb{N}$ and constant death rate ν on $\mathbb{N}_+$.

The result makes intuitive sense. If the service rate is less than the arrival rate, the chain is transient and the length of the queue grows to infinity. If the service rate is greater than the arrival rate, the chain is positive recurrent. At the boundary between these two cases, when the arrival and service rates are the same, the chain is null recurrent.

The infinite server queuing chain $\boldsymbol{X}$ is positive recurrent. The invariant distribution is the Poisson distribution with parameter μ / ν. The invariant probability density function f is given by

$$f(x) = e^{-\mu/\nu} \frac{(\mu/\nu)^x}{x!}, \quad x \in \mathbb{N} \tag{16.22.4}$$

Proof

This also follows from results for the continuous-time birth-death chain. In the notation of that section, the birth rate is constant, $\mu(x) = \mu$ for $x \in \mathbb{N}$ and the death rate is proportional to the number of customers in the system: $\nu(x) = \nu x$ for $x \in \mathbb{N}_+$. Hence the invariant function (unique up to multiplication by constants) is

$$x \mapsto \frac{\mu(0) \cdots \mu(x-1)}{\nu(1) \cdots \nu(x)} = \frac{\mu^x}{\nu^x x!} \tag{16.22.5}$$

Normalized, this is the Poisson distribution with parameter μ / ν.

This result also makes intuitive sense.
