IEOR 151 – Lecture 20
Queues

1 Generic Queues

Queueing theory is the mathematical study of waiting lines, and here we will discuss models of queues using a stochastic processes approach to this topic. In general, we will be interested in modeling the following aspects of queues:

1. Arrival process – that describes how customers enter the queue
2. Service process – that describes how customers are served.
3. The number of servers
4. Maximum number of customers that can be in line. We will assume this is infinite.
5. The size of the pool of customers. We will assume this is infinite.
6. Service discipline – that describes how customers in line are chosen for service. We will assume first in first out (FIFO).

These models are analyzed to determine

1. The mean number in the system (in the line and in service)
2. The mean number in line
3. The mean time spent in the system or in the line

2 Review of Poisson Processes

A Poisson process is an integer-valued counting process $\{N(t), t \geq 0\}$ (with $N(t) \in \mathbb{N}$) in which the interarrival times are described by an exponential distribution. Recall that an exponential distribution with rate $\lambda > 0$ has distribution function $F(u) = 1 - \exp(\lambda u)$, and its mean is $1/\lambda$. Exponential distributions have an interesting *memoryless* property, meaning that if T has exponential distribution, then

$$\mathbb{P}[T > s + t | T > s] = \mathbb{P}(T > t).$$

What this intuitively means is that if you are waiting for an arrival for s units of time, then the probability of an arrival after t additional units of time does not depend on how long you have been waiting for.

3 Steady-State Distribution of Markov Chain

Suppose we have a Markov chain represented by a graph $G = (V, E)$ with appropriate edge weights to represent transition rates. Let $p(t) \in \mathbb{R}^n$ be a vector of probabilities, in which $p_k(t)$ denotes the probability that the current state is $v_k \in V$. We can represent the change in probability over time by a differential equation:

$$\frac{dp_k}{dt} = \big(w_{kk} + \textstyle\sum_{j:e_{kj}\in E} -w_{kj}\big)p_k + \big(\textstyle\sum_{j:e_{jk}\in E} w_{jk}\big)p_j,$$

which are known as the Kolmogorov forward equations. The interpretation is that probability flows into v_k from $v_j : e_{jk} \in E$ and flows out of v_k into $v_j : e_{kj} \in E$. Because the quantities multiplying p on the right hand side are constants, this is in fact a linear ordinary differential equation, which can be written as

$$\frac{dp}{dt} = Ap.$$

The steady-state distribution is a vector of probabilities such that $dp/dt = 0$, which in this case implies $Ap = 0$ or that p belongs to the null space of A. A Markov chain is not guaranteed to have a steady-state distribution, but it will have such a distribution in many cases. Also, the situation of Markov chains with an infinite-dimensional (but countable) state space is similar but with the change that A must now be an infinite-dimensional linear operator.

4 $M/M/1$ Queues

We will begin by focusing on Markovian queues, in which the arrival process is Poisson and the service times have an exponential distribution. The notation $M/M/1$ indicates a model in which the arrival process is Poisson with rate λ, the service times are exponential with rate μ, and there is one server. First, note that the queue is unstable if $\lambda \geq \mu$; this is a situation in which on average customers arrive at a faster rate than they are served.

Here, we will calculate the average number of customers in line. Let $p_k(t)$ be the probability that there are k customers in line at time t. Then, we have that

$$\begin{aligned} dp_0/dt &= -\lambda p_0 + \mu p_1 \\ dp_k/dt &= \lambda p_{k-1} - (\lambda + \mu)p_k + \mu p_{k+1}. \end{aligned}$$

For p_k and $k > 0$, we can think of $(\lambda + \mu)$ as the rate of outward flow, λ the rate of inward flow from p_{k-1}, and μ the rate of inward flow from p_{k+1}. And at steady state, we must have that $dp_0/dt = 0$ and $dp_k/dt = 0$ for all $k > 0$, which means that

$$\begin{aligned} 0 &= -\lambda p_0 + \mu p_1 \\ 0 &= -\lambda p_{k-1} - (\lambda + \mu)p_k + \mu p_{k+1}. \end{aligned}$$

Additionally, we must have that $\sum_{k\geq 0} p_k = 1$. After some work (there are a number of approaches to show this), we get that $p_k = (1-\rho)\rho^k$, where $\rho = \lambda/\mu$ is the utilization ratio that describes the fraction of time that the server is working.

Given the steady state distribution of the queue, we can now compute the expected number of customers in line. The distribution p_k is a geometric distribution, and so a standard result gives that

$$L = \mathbb{E}(k) = \sum_{k\geq 0} k p_k = \sum_{k\geq 0} k(1-\rho)\rho^k = \rho/(1-\rho).$$

5 More Information and References

The material in these notes follows that of the course textbook "Service Systems" by Mark Daskin.