INSE6300 Jia Yuan Yu

7: Queueing

Concordia March 30, 2016

In this lecture, we study the notion of quality of service in queues. Queues appear supply chain design: communication networks, supermarkets, assembly lines, airports. Whenever you have customers or items arriving at one rate and departing at another rate, you have a queue. For instance, queues arise when stock arrive in a warehouse at a different rate than the demand—inventory is an example of queue. Queues also arise when customers arrive at random time instants and take a nonzero amount of time to serve and depart, which is not capture in the Bass model. Queueing models do capture the interaction between the arrival times and the service times of the supply chain. The arrivals can model jobs, phone calls, inventory items, etc. Sevice can model demand, sales, etc.

1 Deterministic queues

Figure 1: From `http://sgforums.com/`.

Consider the following congestion example[1] as an analogy of supply chains. There are N commuters who must cross a bridge every morning to go to work, with the goal of arriving at time t^*. Only the travel time on the bridge are non-negligible. The

[1] Due to W. Vickrey (cf. `http://www.econ.ucsb.edu/~tedb/Courses/UCSBpf/pflectures/vick4.pdf`).

bridge has a carrying capacity of s cars per minute. It takes $1/s$ minutes for one commuter to cross. Hence, the minimim rush-hour lasts N/s minutes. The decision of each commuter is what time t to leave home.

There are two costs involved:

- arriving τ minutes too early or too late incurs $\beta\tau$;
- waiting in the traffic queue τ minutes incurs $\alpha\tau$, with $\alpha > \beta$ known constants for all commuters.

In contrast to other decision problems seen so far, there are multiple decision makers whose decisions affect each other. In such settings, it is more useful to look for a set of decisions that are stable, i.e., an equilibrium. One notion of equilibrium is to require fairness, or lack of envy from one commuter to another in terms of their total costs. Let's find a set of departure times so that all communters experience the same cost, while keeping the total rush-hour to a minimum of N/s minutes. This can occur only if the first commuter and the last commuter depart before and after t^* respectively. These two do not incur any queueing cost, only late-early costs. By symmetry, they depart at $t^* - \frac{N}{2s}$ and $t^* + \frac{N}{2s}$ both incur a cost of $\frac{N}{2s}$.

Let D_t denote the length of the queue at time t. Next, consider another commuter who leaves at time t, waits in the queue D_t/s minutes, and arrives at time $t + D_t/s$. First, consider the case $t + D_t/s < t^*$, the equilibrium condition requires that

$$\beta\frac{N}{2s} = \alpha D_t/s + \beta(t^* - t - D_t/s).$$

Solving to D_t, we obtain

$$D_t = \frac{\beta N}{2(\alpha - \beta)} + \frac{\beta s}{\alpha - \beta}(t - t^*).$$

Observe that by definition, the rate of arrivals of commuters at the bridge is

$$D_t' + s = \frac{\beta s}{\alpha - \beta} + s = \frac{\alpha s}{\alpha - \beta},$$

which is larger than the rate s of commuters going through the bridge. Hence, at an equilibrium, starting from $t^* - \frac{N}{2s}$, commuters should arrive at a rate of $\frac{\alpha s}{\alpha - \beta}$.

Next, consider the case $t + D_t/s \geq t^*$:

$$\beta\frac{N}{2s} = \alpha D_t/s + \beta(t + D_t/s - t^*).$$

The analysis is similar.

Remark 1. How does this equilibrium solution compare with an optimal solution? Is an optimal solution fair and stable?

Figure 8.1: Rush Hour

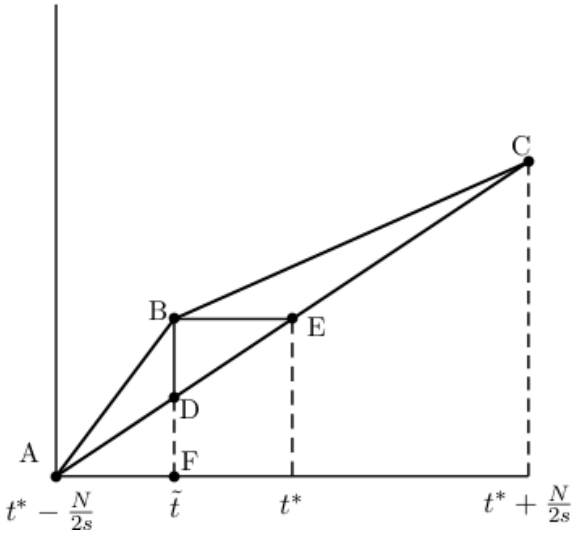

Figure 2: Number of commuters arriving at and departing from the bridge over time.

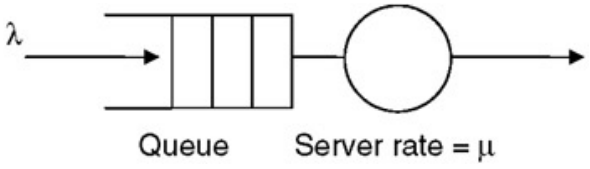

Figure 3: From Flylib

2 Queues as Markov chains

In this section, we use a model where the observations are no longer i.i.d., but Markovian. Many queueing models can be analyzed as Markov chains. For instance, in the $M/M/1$ queue model[2], customers arrive at random time instants, where the time interval between consecutive arrivals is exponential with parameter λ. The customers are served on a first-come first-served basis. The service times are exponential with parameter $1/\mu$. There is no limit on the number of customers in the queue.

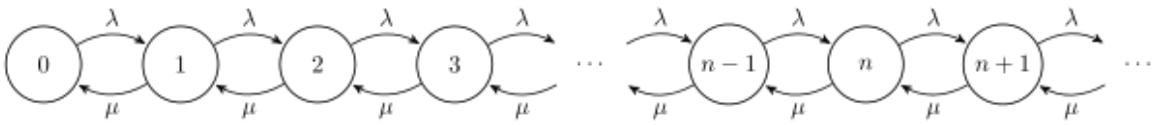

Figure 4: From `https://en.wikipedia.org/wiki/M/M/1_queue`

Example 2.1 ($M/D/1$ (D for deterministic service))**.** Consider deterministic service times of length τ. Let $X_0 = 0$ and let X_k denote the number of customers waiting in the queue when the k-th customer enters service. Let ξ_k denote the number of customers who arrived during the k-th customer's service time. Since the service times are fixed, and the time between arrivals are i.i.d. random variables, $\xi_1, \xi_2, \ldots$ are i.i.d. The queue length evolves as a Markov chain:

$$X_{k+1} = \max\{X_k + \xi_k - 1, 0\}.$$

Remark 2. Queues can also be controlled with actions and analyzed as MDPs (cf. Puterman, Section 3.7).

2.1 Quality of service, waiting time

When queues model customers, an important way of measuring quality of service is through the waiting times for customers. The waiting times are random variables. For instance, in the $M/D/1$ model, the arrival times are $t_1, t_2, \ldots$, where

$$\begin{aligned} t_1 &= 0, \\ t_i &= \sum_{j=1}^{i-1} e_j, \quad \text{for } i = 2, 3, \ldots, \end{aligned}$$

and $e_1, e_2, \ldots$ denote i.i.d. exponential random variables with parameter λ for the times between consecutive arrivals. Observe that the arrival times form a Markov process.

Suppose that it takes τ time units to serve each customer. Let $t_1 + d_1 + \tau, t_2 + d_2 + \tau, \ldots$ denote the departure times of customers, so that $d_1, d_2, \ldots$ are the durations of

[2]M stands for Markovian: the arrival and the service processes are modeled as Poisson processes, which are Markovian.

time that customers spend in the queue—i.e., their waiting times. Waiting times are also described by a Markov process[3]:

$$
\begin{aligned}
&d_1 = 0, \\
&d_i = \max\{d_{i-1} - \underbrace{(t_i - t_{i-1})}_{\text{difference in arrival times}} + \underbrace{\tau}_{\text{service time}}, 0\}, \quad i = 2, 3, \ldots .
\end{aligned}
$$

The probability distribution of the random variables d_i can be derived from first principles or estimated by simulation (by generating many samples of each d_i and counting empirical frequencies, as in Figure 2.1). The decision-maker can control this probability distribution by controling the parameters τ and λ, *e.g.*, by hiring more workers to reduce service time, by limiting the number of customers arriving in the queue through an invitation system, etc.

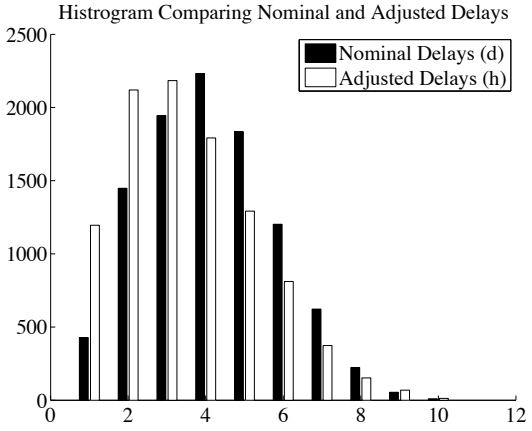

Figure 5: Simulated empirical frequency histogram for d_i given a fixed d_{i-i}.

3 References

- Ted Bergstrom's Lecture Notes on Traffic Congestion for *Theory of Public Goods and Externalities*.
- Chapters 1, 2, and 3 of Markov Decision Processes (Puterman).

[3]We say Markov chain for processes that take a finite number of values.