CE 269
Traffic Engineering

Lecture 17

Introduction to Queueing Theory

Previously on Traffic Engineering

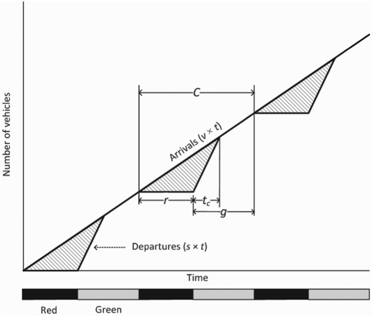
We can easily analyze delays at a junction in the case of deterministic arrivals and departures. This is an example of D/D/1 queues. More general settings will be discussed later in the course.

Suppose the arrival rate of vehicles at an approach be v and saturation flow rate be s. If the cycle time is C,

- The number of vehicles arriving at the junction in one cycle is vC.
- The maximum number of vehicles that can leave is sg, where g is the effective green.

Previously on Traffic Engineering

Suppose $sg > vC$. Calculate the following quantities:



- The time to clear the queues after the start of the effective green.

$$t_c = \frac{vr}{s - v}$$

- The proportion of the cycle time with a queue.

$$t_c = \frac{r + t_c}{C}$$

- Total vehicle delay per cycle and average delay per vehicle.

$$D_t = \frac{vr^2}{2(1 - v/s)} \quad D_{avg} = \frac{0.5C(1 - g/C)^2}{1 - (v/c)(g/C)}$$

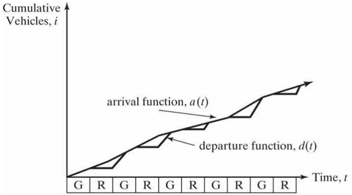
Previously on Traffic Engineering

The earlier examples makes a few implicit assumptions.

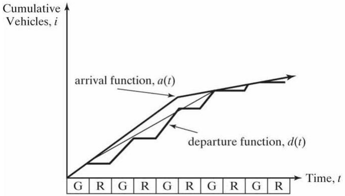
- Arrival rates are uniform, which is not always true. Vehicles at isolated intersection usually follow a random arrival process. If the signal is part of a coordinated network, then arrivals are in batches or platoons.
- The queues are assumed to stack vehicles on top of one another. This is also called a *point-queue* model.

Also, in practice traffic can switch between under-saturated and over-saturated conditions. To address this issue, the delay is usually broken down into *uniform delay* and *overflow delay*.

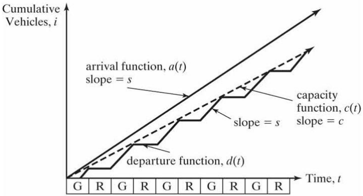
Previously on Traffic Engineering



(a) Stable Flow



(b) Individual Cycle Failures Within a Stable Operation



(c) Demand Exceeds Capacity for a Significant Period

Lecture Outline

1. Review of Probability Distributions
2. Continuous Time Markov Chains
3. Single Server Queues
4. Multiple Servers and Network of Queues

Lecture Outline

Review of Probability Distributions

Review of Probability Distributions

Poisson Distribution

The Possion random variable is used to count the number of random events in a time period.

For example,

- The number of accidents that occur on a highway in an year.
- The number of customers served by a teller at a bank.

It is assumed that occurrences of events are independent of each other. Also, we assume that two or more events cannot happen simultaneously. Further, the average rate of occurrence of events is known and assumed constant.

https://www.randomservices.org/random/apps/PoissonExperiment.html

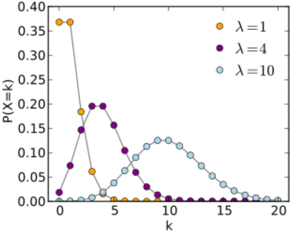
Definition

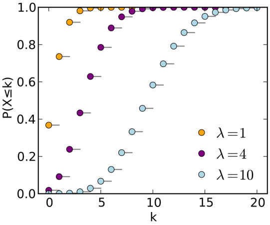
PMF of a Poisson distributed random variable with parameter $\lambda > 0$ is

$$\mathbb{P}(X = x) = p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Claim

Suppose $X \sim Pois(\lambda)$, $\mathbb{E}(X) = V(X) = \lambda$





Exponential distribution is commonly used to model time between consecutive events when the events occur according to Poisson distribution.

For example,

- The time duration between two accidents on a highway
- The amount of time taken by a bank teller to serve a customer
- The time between two arrivals at a checkout queue

Exponential Distribution

Probability Density Function

Definition

Suppose $X \sim \exp(\lambda)$, its probability density function is defined as

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Claim

Suppose $X \sim \exp(\lambda)$, *its CDF is*

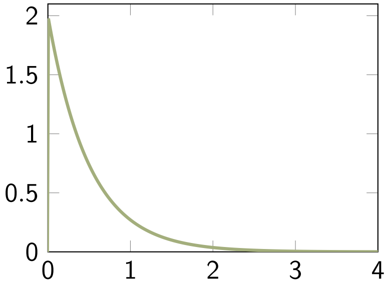
$$F_X(x) = 1 - e^{-\lambda x}$$

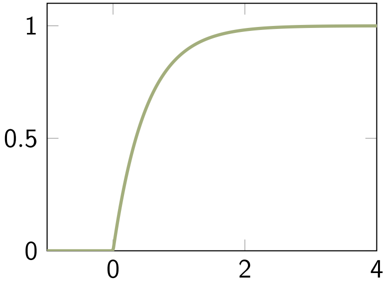
Check if $dF_X(x)/dx = f_X(x)$

Exponential Distribution

PMF and CDF

PDF and CDF of an exponentially distributed random variable with $\lambda = 2$ are shown below.





Claim

If $X \sim \exp(\lambda)$*, then* $\mathbb{E}(X) = 1/\lambda$ *and* $V(X) = 1/\lambda^2$

Exponential Distribution

Connections with Poisson Distribution

To see why inter-arrival times of an Poisson distributed random variable is exponentially distributed, let $X \sim \text{Pois}(\lambda)$.

Consider a time window t. The probability that there are zero arrivals in t is given by

$$\mathbb{P}(X = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

If Y is the inter-arrival time, then $\mathbb{P}(X = 0) = \mathbb{P}(Y > t)$.

Hence, $\mathbb{P}(Y \leq t) = 1 - e^{-\lambda t}$, which is the CDF of the exponential random variable.

Exponential Distribution

Memoryless property

Suppose that the inter-arrival times of buses at a bus stop are exponentially distributed with rate λ. Let X be the arrival time of the next bus.

Assuming, that you have been waiting for t minutes (right after the passing of the previous bus), what is the probability that you will have to wait at least another s minutes.

$$\mathbb{P}(X > s + t | X > t) =?$$

For this reason, exponential random variable is said to exhibit a **memoryless property**.

Lecture Outline

Continuous Time Markov Chains

Definition (Markov Property)

A stochastic process $\{X_n, n \geq 0\}$ with a countable state space S is called a DTMC if $\forall\, n \geq 0, i, j \in S$,

$$\mathbb{P}\big[X_{n+1} = j | X_n = i, X_{n-1}, X_{n-1}, \ldots, X_0\big] = \mathbb{P}\big[X_{n+1} = j | X_n = i\big]$$

Definition (Time Homogeneity)

A DTMC $\{X_n, n \geq 0\}$ is said to be time homogeneous if $\forall\, n \geq 0, i, j \in S$,

$$\mathbb{P}\big[X_{n+1} = j | X_n = i\big] = p_{ij}$$

i.e., RHS does not depend on n or $p_{ij}(n) = p_{ij} \,\forall\, n \geq 0$

The probability with which the system moves from i to j, p_{ij}, is called the **transition probability** and the matrix of p_{ij} values is called the **one-step transition probability matrix**.

$$P = \left[p_{ij}\right]_{|S|\times|S|}$$

Note that P can have countably infinite rows and columns.

Definition (Stochastic Matrix)

A square matrix $P = \left[p_{ij}\right]_{|S|\times|S|}$ is called right stochastic if

1. $p_{ij} \geq 0 \, \forall \, i, j \in S$
2. $\sum_{j\in S} p_{ij} = 1 \, \forall \, i \in S$

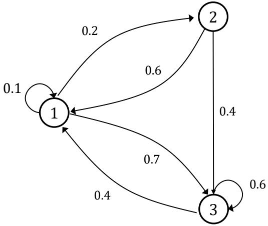
Transition matrices of a Markov chain are right stochastic matrices.

Continuous Time Markov Chains

Discrete-Time Markov Chains

The transition probability matrix can also be visualized as a directed graph in which the states are nodes and an arc (i, j) exists only if $p_{ij} > 0$.

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.6 & 0 & 0.4 \\ 0.4 & 0 & 0.6 \end{bmatrix} \end{array}$$



http://setosa.io/ev/markov-chains/

The P matrix alone doesn't fully describe a DTMC. We'd also need to know the initial distribution.

$$a_i = \mathbb{P}\left[X_0 = i\right] \forall\, i \in S$$

Let a be row vector of a_i's. A Markov chain can thus be fully specified using (S, P, a).

Continuous Time Markov Chains

Definition

Definition (CTMC)

Suppose $\{X(t), t \geq 0\}$ is a continuous time stochastic process with a countable state space $S = \{0, 1, \ldots\}$. Then $X(.)$ is a CTMC if the transition probabilities have the following Markov property

$$\mathbb{P}[X(s+t) = j | X(r) : 0 \leq r \leq s] = \mathbb{P}[X(s+t) = j | X(s)]$$

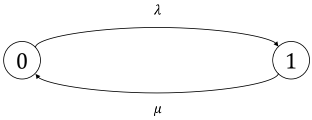
We can think of transitions in CTMCs in two steps. Imagine that the system spends an exponential amount of time in each state and when it transitions, it does like a DTMC.

Instead of dealing with a transition matrix, we define a rate matrix whose (i,j)th element indicates the product of the rate at which we leave state i and the probability of transitioning to j.

Continuous Time Markov Chains

Example

Consider a two-state CTMC with the following transition rates. Imagine that it represents a machine replacement problem.



The rate matrix is written as $\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$

One metric of interest is the steady state probability p_j that the system is in state j.

Definition (Fundamental Theorem of CTMC)

Let $\{X(t), t \geq 0\}$ be an irreducible CTMC. The steady state probabilities are a solution to the following equations

$$pQ = 0$$

$$\sum_{j \in S} p_j = 1$$

These are also called global balance equations as they can be written as the rate out = rate in for every state.

Find the steady state probabilities for the machine reliability example.

Lecture Outline

Single Server Queues

Single Server Queues

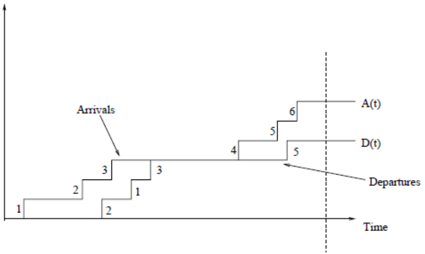
Queuing Theory involves studying quantities such as queue lengths and waiting times in a single queue or a network of queues. Application areas include:

- Checkout lines, ATMs
- Call centers
- Traffic signals and Toll plazas
- Airports
- Communication and Internet
- Manufacturing

Single Server Queues

Queuing Theory

Consider a single queue with an arrival process $A(t)$ and departure process $D(t)$.



The horizontal and vertical cross-sections tell us the amount of time spent in the queue and the queue length.

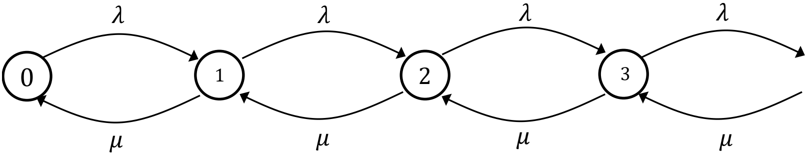
The area represents the overall time spent by everyone in the queue.

(Source:Willig (1999) A Short Introduction to Queuing Theory)

Single Server Queues

Poisson Arrivals and Exponential Service Times

Let us now analyze a special case where the arrival process is a Poisson process with rate λ and the service times are exponential with rate μ. The fraction $\rho = \lambda/\mu$ is called the traffic intensity.



Can you write the global balance equations and find the steady state probabilities?

Thus, the steady state probabilities are

$$p_0 = 1 - \rho$$
$$p_i = \rho^i (1 - \rho)$$

How do we find the expected length of the queue?

$$\mathbb{E}(L) = \sum_{j=0}^{\infty} j p_j = \sum_{j=0}^{\infty} j \rho^j (1 - \rho) = \frac{\rho}{1 - \rho}$$

We can also find the expected waiting time by defining new random variables and taking their expectations.

$$\mathbb{E}(T) = \frac{1}{\mu - \lambda}$$

How are $\mathbb{E}(L)$ and $\mathbb{E}(T)$ related? $\mathbb{E}(L) = \lambda \mathbb{E}(T)$. This is called Little's law and holds in more general settings.

Multiple Servers and Network of Queues

The example that we discussed so far is referred to as M/M/1 queues. In general, the assumptions used in modelling a queue are represented using Kendall's notation

$$A/S/c/K$$

- A Arrival process
- S Service time distribution
- c Number of servers
- K Capacity of the queue

M indicates Markovian or memoryless and is used to indicate Poisson processes and exponential service times.Other options include D for deterministic and G for general distribution.

Multiple Servers and Network of Queues

Queue Discipline

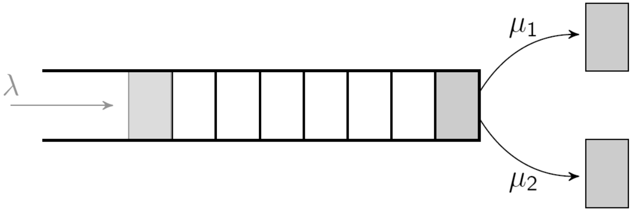
Additionally, we may also specify the service discipline.

- FIFO
- LIFO
- Random Service
- Round Robin
- Priority Order

Multiple Servers and Network of Queues

M/M/2 queue

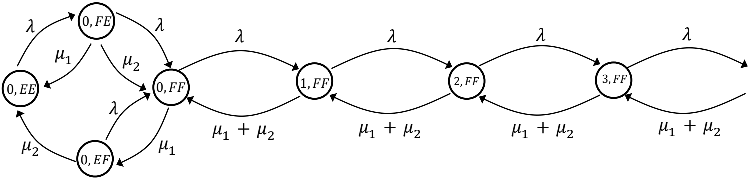
Imagine a single queue but two servers with difference service rates.



Systems like these can also be analysed using CTMCs

Multiple Servers and Network of Queues

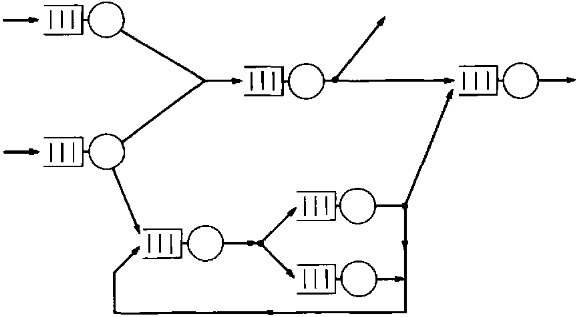
M/M/2 queue



The expressions are relatively simpler when the service rates are the same. For an $M/M/c$ queue, let $\rho = \frac{\lambda}{c\mu}$, then p_0 is

$$\left[\left(\sum_{k=0}^{c-1} \frac{(cp)^k}{k!} + \frac{(cp)^c}{c!}\frac{1}{1-p}\right)\right]^{-1}$$

For a network of queues, we have to resort to simulation. Software packages such as Anylogic, Arena can be used for this purpose.



https://www.supositorio.com/rcalc/rcalclite.htm

Your Moment of Zen