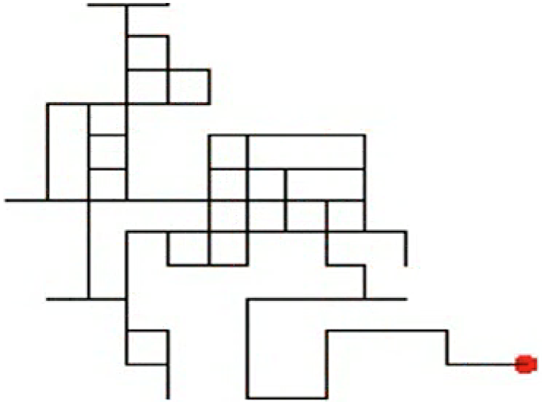
Stochastic Simulations

Discrete event simulations & queuing theory



Dr. Gabor Zavodszky
2020

g.zavodszky@uva.nl

www.zavodszky.com

Discrete Event Simulations (DES)

We often encounter systems, where the dynamics of the system can be characterized in terms of a set of events.

For instance, look at a hard sphere approximation of dilute gas.

The overall state of the gas changes only when collisions happen. However, since it is dilute, it is relatively rare.

So why would we compute the behavior of the hard particle the rest of the time when they are just traveling straight?

We can regard collisions as events of interest, and compute only those, and we can always just progress time to the next event.

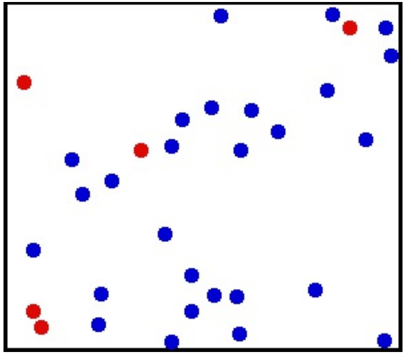


Image: Wikipedia

Discrete Event Simulations

Main assumption: The system stays in the same state between the events.

Often desired: These events are separate and adhere to local causality. I.e. a collision will only change the state of the colliding particles and will not influence the others.

This latter is fundamental for parallel discrete event simulations, and is often the most complicated part to realize, usually requiring some smart trick or approximation.

Two typical numerical methods of the discrete event type are

- Agent based simulations (separate course)

- Queueing models

We will discuss the latter, however, a short detour is needed first.

The Poisson process

It is most often used to model or characterize distribution of random events in time, such as:

- Customers arriving to the store.
- Occurrence of earthquakes.
- Incoming download requests to a file server.

These are scenarios where events appear to happen at a given rate, but completely at random (without any known structure).

The distribution of such events can be modeled using the Poisson distribution.

Remember the *binomial distribution*? Recall that it was the distribution of repeated trials with either positive or negative outcome (p and $1-p$ probability, respectively). It was demonstrated with a Galton table, where every level is a decision for the ball to bounce either left or right.

The *Poisson distribution* is the limiting case of the *binomial distribution* where $p \to 0$ and $n \to \infty$.

Poisson distribution

If X is a discrete RV from the Poisson distribution with $\lambda > 0$ (average number of events), and $k = 0, 1, 2, \ldots$ is the number of events, then the *pmf* is:

$$f(k, \lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Note that $\lambda = E(X) = \mathrm{Var}(X)$!

Another often used formulation is when λ denotes the *mean rate* of events occurring in a given time period $t \geqslant 0$, and k is then the number of event occurring during t time. In that case the above formula looks:

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

Connection to the exponential distribution

Let's assume the first event occurred exactly at time t_1, and the second t_2 later at $t_1 + t_2$. We can look at these t_x intervals as RVs.

Since the first event occurs after t time if and only if there were no other events in $[0, t]$, looking at the previous result: $P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$, we can conclude:

$$P(\text{Event}_1 > t) = P(k = 0) = e^{-\lambda t}$$

Therefore, the probability, that the first event occurs in $[0, t]$ is given by:

$$P(\text{Event}_1 \leqslant t) = 1 - e^{-\lambda t}$$

This is the cumulative distribution function of the *exponential distribution*. It's mean is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$.

The memoryless property

A very important property of the exponential distribution is that the occurrence of the next event does not depend on any past or future information: if $P(X > t) = e^{-\lambda t}$, then for any $\Delta t \geqslant 0$:

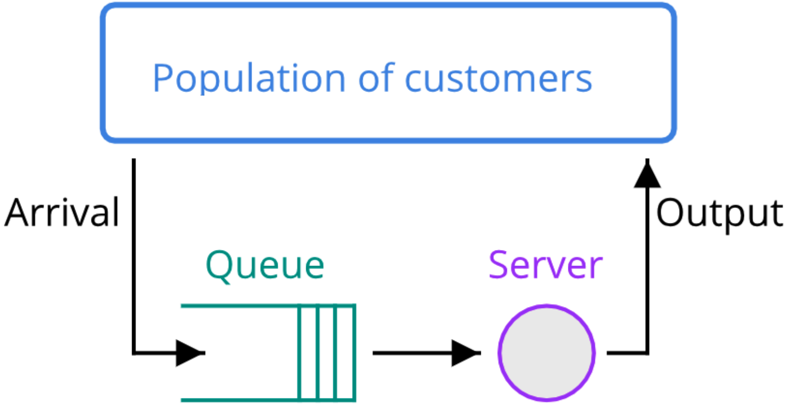
$$P(X > t + \Delta t | X > t) = \frac{P(X > t + \Delta t, X > t)}{P(X > t)} = \frac{P(X > t + \Delta t)}{P(X > t)} = \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} = e^{-\lambda \Delta t}$$

This memoryless-ness(?) will be a very important feature later on.

Note that there are only two memoryless distributions:

- the geometric (the distribution of the number of Bernoulli trials needed to get a success)
- the exponential distribution

Queuing theory



Introduced by Erlang in the early 20th century. He worked in telephone center in Copenhagen as part of improving the telephone service by finding the optimal amount of circuits and operators to reduce average waiting time.

The relevance of this description comes from its applicability to a wide array of situations, which fit the figure above.

Basic model elements

Queue type

- FIFO (First In, First Out)
- LIFO (Last In, First Out)
- Round Robin - predefined time-slice for every customer. If the service is not completed, the customer rejoins the queue.
- Priority - Every customer has a predefined priority, and the server always selects the one with highest priority.
- Random - Customers in the queue are served in a random order.

Time distributions

- $A(t) = P(t_n \leqslant t)$ - describes the distribution of durations between arrival times
- $B(t) = P(t_n \leqslant t)$ - describes the distribution of service time durations per customer

Kendall notation

$$A/B/m/N-S$$

- A denotes the inter-arrival time distribution.
- B denotes the service time distribution.
- m is the number of servers.
- N is the maximum size of the queue (if it is finite)
- S is the service queuing disciple (e.g. FIFO, LIFO,...etc.). In the default case of FIFO it is omitted.

Common time distributions

For A and B the following abbreviations are the most common:

- M (Markov) - actually M for memoryless, as this is the exponential distribution $A(t) = 1 - e^{-\lambda t}$.
- D (Deterministic) - deterministic distribution, i.e. constant value (hence no randomness).
- E (Erlang-k) - the distribution of the sum of k independent exponential variables.
- H (Hyper-exponential) - Summation of exponential distribution each weighted with a probability. I.e.: the probability that X will take the form of the associated exponential distribution.
- G (General) - General case when the distribution is not specified (however, often at least the mean and the variance are known).

General case

This system gives a formal framework to handle a wide range of problems using the same tools. E.g. results derived for $M/M/1$ are reusable for other problems that map to this configuration.

Typically with A or B (or both) becoming G, our work becomes more difficult, what can we say in a fully general case ($G/G/1$)?

- Let λ denote the arrival rate and $E(B)$ denote the mean service time on our 1 server.
- Intuitively, we want $\lambda E(B) < 1$, otherwise out queue will start to grow. (Note: due to randomness $\lambda E(B) = 1$ will also explode, except for $D/D/1$).
- $\lambda E(B) = \rho$ is often called *server utilization*, since it gives the fraction of time the server is busy.
- If we have multiple servers: $G/G/c$, then $\lambda E(B) < c$, which per server means $\rho = \frac{\lambda E(B)}{c}$.

Little's law

Little's law creates a connection among:

- $E(L)$ - The mean number of customers in the system
- $E(S)$ - The mean sojourn time (time spent in the system: waiting in queue + service time)
- λ - rate of arrivals (average number of customers entering the system per unit time)

Note we assume that the queue does not grow to infinity (see prev. slide).

$$E(L) = \lambda E(S)$$

This is a general results that holds for every distribution type (even G) and every queue discipline (not just FIFO).

Little's law - relations among the sub-components

We can also apply it to the components of the system:

- Applying it to the queue only (omitting the server) it yields a relation between the length of the queue L^q and the waiting time in the queue W: $E(L^q) = \lambda E(W)$.
- If we apply it solely to the server we get back the server utilization: $\rho = \lambda E(B)$

Remark: Little's law only applies to the steady state of the system (when the average number of people arriving to the system equals the average number of people leaving).

PASTA property

Queuing systems with Poisson arrivals (i.e. $M/./.$ systems) the fraction of customers that find the system in some state S on arrival is exactly the fraction of time the system spends in state S.

OK, what does that mean? And by the way isn't it trivial?

Look at a $D/D/1$ system with arrivals at $1, 3, 5, \ldots$ and $E(B) = B = 1$ service time. Every customer will find the system empty, however, it is only empty $\frac{1}{2}$ of the time.

This will not happen in the Poisson case, since due to its randomness, *Poisson Arrivals See Time Averages*.

Remarks on the practical use

- We only consider stationary/steady state systems.

- Using Little's law, PASTA, and often the combination of the two we can devise several important characteristic measures of the system: *performance measures*

- Measures to consider:
  - Mean number of customers in the system
  - Mean sojurn time (also called response time: waiting time + service time)
  - Mean waiting time
  - Distribution of waiting times

Self-study

You can find two background document on the Canvas page of the lecture. Based on what was discussed in this lecture it will be easy to process them. Use them as information source for the next Assignment!

Quick Recap

- Discrete Event Simulations in general
- Poisson distribution, exponential distribution
- Memoryless property
- Queuing theory
- Kendall's notation
- Little's law
- PASTA property

Q&A Section