CS 3.307

Performance Modeling for Computer Systems

Tejas Bodas

Assistant Professor, IIIT Hyderabad

Logistics

- Feel free to contact me anytime at tejas.bodas@iiit.ac.in.
- Office @ A5304.
- Book– Performance modeling and design of computer systems (Cambridge press) by Mor Harchol-Balter (Professor, CMU)
- Other books: 1) Stochastic processes by Sheldon Ross 2) Probabilistic modeling by Isi Mitrani.
- Assignment 1 : 15%. Midsem exam: 25%. Project: 25% Endsem 35 %.

Course Outline

- Module 1 (2 lectures)
  - Motivation, Probability refresher, Introduction to Stochastic Processes
- Module 2 (4 lectures)
  Poisson Process & Markov Chains
- Module 3 (2 lectures) Elementary Queues
- Module 4 Renewal theorems and Busy period analysis (3 lectures)
- Module 5 (3 lectures) Advanced Queues

Performance modeling for Computer systems

- How do you measure the performance of your computer?
- Speed with which it runs programs. RAM, clock speed, GPU, Cores.
- Storage space ? SSD or not ?
- What is the key word here ? LATENCY!
- Performance metrics?
  - response time (run time, lag, delay, jitter)
  - blocking probability (screen freeze, no disk space, packet loss, buffer full)

Modeling ?

- Design for performance: How many cores or GPU's? which core to use? how to schedule instructions in a core?
- Routing (which core) and scheduling (which program/ instruction to execute)
- How do you know which is a good design? via experimentation?(costly!)
- Performance analysis! via stochastic modeling

Applications Beyond Computers

- Computer systems
  - server farms, cloud computing, distributed storage systems
  - Communication systems, Wifi, Sensor networks.
- Heathcare
  - How many OT? How many Specialists or nurses?
  - Scheduling operations, stocking of medicines, scheduling tests.
- Hospitality industry
  - Designing hotel lobbies for faster checkin
  - Restaurant seating! (How many tables of size 2,4,8?)
- Transportation systems
  - Airline or Railway scheduling
  - Priority scheduling, class differentiation
- Operation Research!
- Henceforth use the term Queueing system!

A single server queue

- One server, one FIFO queue for jobs to wait.
- μ denotes service rate, λ denotes the arrival rate.
- Service requirements S_n and inter-arrival times A_n are typically assumed to be i.i.d.
- In its simplest form, we will assume $S_n \sim Exp(\mu)$ and $A_n \sim Exp(\lambda)$.
- Jobs face queueing delay due to waiting for other jobs.
- This is the most basic $M/M/1$ queue. Modeling this as a Markov chain and solving its stationary distribution gives us mean response time (mean of service time + waiting time).
- $E[T] = \frac{1}{\mu - \lambda}$.

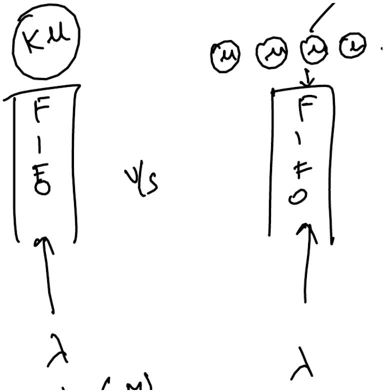
A single server queue

- $E[T] = \frac{1}{\mu - \lambda}$.
- Let N is the number of jobs in the system (Queue + server). Then what is $E[N]$?
- We will see Little's law that says that $E[N] = \lambda E[T]$.
- Mean number of jobs $E[N] = \frac{\lambda}{\mu - \lambda}$.
- This course is about Markov chain analysis to derive such formulas.

Example 1: Doubling the arrival rate

- $E[T] = \frac{1}{\mu - \lambda}$.
- What would happen to $E[T]$ if $\lambda \to 2\lambda$?
- It could blow up if $\mu < 2\lambda$.
- If you want to maintain the same level of response time then do you need to double μ?
- This course is about making such design choices!

Example 2: A fast server versus many slow servers

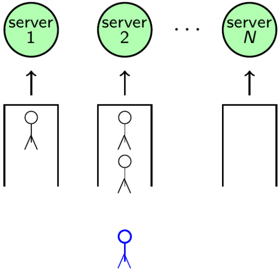


- Which system will have lower $E[T]$?
- Is a fast server $(K\mu)$ better that K normal servers (μ)?
- Does job variability impact this decision? Suppose job sizes were $\cancel{XS}, S, M, L, XL$.
- In the first model, an S, or M job has to possibly wait behind XL. This is avoided in the second scenario.

Example 3: Central queue or individual

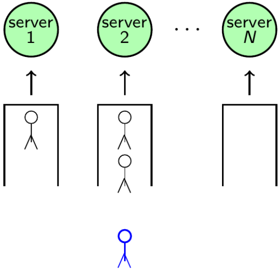
- At Airport immigration, Hotel check-ins you often see central queues.
- But at movie theatres, metro/train ticket counters, you see the second model.
- Which setting has a lower $E[T]$?
- This course will help you answer such performance modeling questions.

Example 4: Supermarket queue and load balancing



- Load balancing concerns the questions which queue to join/assign?
- Popular policy is Join shortest Queue (JSQ).
- What should be ideally done is Join smallest work (JSW).
- N is typically large and the overhead in obtaining queue length information is huge ($2N$).

Example 4: Supermarket queue and load balancing



- In that case, sample d servers randomly and join appropriate queue using $JSQ(d)$ or $JSW(d)$.
- Problem with JSW or $JSW(d)$ is that the workload information is typically unknown. How to implement it then?
- How about replicating jobs on d servers and cancelling copies when one copy starts service ?
- This is redundancy-d with cancel on start.
- We do this at super-markets all the time!

Probability Refresher

Random experiments and Sample space

- Random experiment : Experiment involving randomness
  - Coin toss
  - Roll a dice
  - Pick a number at random from $[0, 1]$.
- Sample space Ω: set of all possible outcomes of the random experiment. It could be a finite or infinite set.
  - $\Omega_c = \{H, T\}$
  - $\Omega_d = \{1, 2, \ldots, 6\}$
  - $\Omega_u = [0, 1]$

Events

- A subset $A \subseteq \Omega$ is called an **event**.
- Examples of events
  - Events in the coin experiment: $C_1 = \{T\}$.
  - Events in the dice experiment: $D_1 = 6, D_2 = \{1, 3, 5\}$
  - Events in $U[0,1]$ experiment: $U_1 = \{0.5\}$, $U_2 = [.25, .75]$.
- Probability of event A is denoted by $\mathbb{P}(A)$.

Probability theory

$\{$Random experiment, Sample space, Events$\}$ are the key ingredients in probability theory.

In probability theory, we are interested in **measuring** the probability of subsets of Ω (events).

Probability measure $\mathbb{P}$ is a **set function**, i.e. it acts on sets and measures the probability of such sets.

sigma-algebra as domain for $\mathbb{P}$

- Event space or *sigma-algebra* $\mathcal{F}$ is a collection of measurable sets that satisfy
  $\bullet\ \emptyset \in \mathcal{F} \quad \bullet\ A \in \mathcal{F} \implies A^c \in \mathcal{F}$
  $\bullet A_1, A_2, \ldots A_n, \ldots \in \Omega \implies \cup_{n=1}^{\infty} A_n \in \Omega$

- The $\sigma-$algebra is said to be closed under formation of compliments and countable unions.

- Is it also closed under the formation of countable intersections?

> When Ω is countable and finite, we will consider power-set $\mathcal{P}(\Omega)$ as the domain.

Formal definition of Probability measure $\mathbb{P}$

Definition

A probability measure $\mathbb{P}$ on the *measurable space* $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ s.t.

1. $\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\Omega) = 1$
2. For a disjoint collection of event sets $A_1, A_2, \ldots$ from $\mathcal{F}$ we have
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$
(countable additivity)

- The trio $(\Omega, \mathcal{F}, \mathbb{P})$ is called as a probability space.

Conditional probability

- Given/If dice rolls odd, what is the probability that the outcome is 1?
- Given/If $\bar{\omega} \in [0, 0.5]$ what is the probability that $\bar{\omega} \in [0, 0.25]$?
- The conditional probability of event B given event A is defined as $\mathbb{P}(B/A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ when $\mathbb{P}(A) > 0$.
- Bayes rule: $P(B/A) = \frac{P(A/B)P(B)}{P(A)}$.

Independence and Mutually exclusive

- Two events A, B are independent iff $P(A/B) = P(A)$ and $P(B/A) = P(B)$.
- Two events A, B are independent iff $P(A \cap B) = P(A)P(B)$.

- If A and B are independent, then so are A^c and B^c.
- What about A and B^c? Are they independent?
- Two events A and B are mutually exclusive if occurrence of one implies that the other event cannot occur. Are they independent?
- If A and B are mutually exclusive, then they are not independent (and vice versa).

Random variable

- Given a random experiment with associated $(\Omega, \mathcal{F}, \mathbb{P})$, it is sometimes difficult to deal directly with $\omega \in \Omega$. eg. rolling a dice ten times.
- Notice that each sample point $\omega \in \Omega$ is not a number but a sequence of numbers.
- Also, we may be interested in functions of these sample points rather than samples themselves. eg: Number of times 6 appears in the 10 rolls.
- In either case, it is often convenient to work in a new *simpler* probability space rather than the original space.
- Random variable is a device which precisely helps us make this mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to a 'simpler' $(\Omega', \mathcal{F}', P_X)$.
- P_X is called as an induced probability measure on Ω'.

Random variable

- If Ω' is countable, then the random variable is called a discrete random variable.
- In this case it is convenient to use $\mathcal{F}'$ as power-set.
- If $\Omega' \subseteq \mathbb{R}$ or uncountable, then the random variable is a continuous random variable.
- In this case, $\mathcal{F}' = \mathcal{B}(\mathbb{R})$.
- Notation: Random variables denoted by capital letters like X, Y, Z etc. anf their realizations by small letters x, y, z..

PMF and CDF of a Discrete r.v.

- Let $X : \Omega \to \Omega'$ be a discrete r.v.
- Let $p_X(x)$ for $x \in \Omega'$ denote the probability that X takes the value x.
- $p_X(x)$ is called as a probability mass function.
- The cumulative distribution function (CDF) $F_X(\cdot)$ is defined as $F_X(x_1) := \sum_{x \leq x_1} p_X(x) = \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x_1\}$.

Expectation, Moments, Variance

- The mean or expectation of a random variable X is denoted by $E[X]$ and is given by $E[X] = \sum_{x \in \Omega'} x p_X(x)$.
- The n^{th} moment of a random variable X is denoted by $E[X^n]$ and is given by $E[X^n] = \sum_{x \in \Omega'} x^n p_X(x)$.
- Functions of random variables are random variables.
- For a function $g(\cdot)$ of a random variable X, its expectation is given by $E[g(X)] := \sum_{x \in \Omega'} g(x) p_X(x)$
- $Var(X) := E[(X - E[X])^2]$
- HW: Prove that $E[(X - E[X])^2] = E[X^2] - E[X]^2$
- For $Y = aX + b$, what is $E[Y]$? $E[Y] = aE[X] + b$. (Linearity of expectation)

Bernoulli random variable

- Bernoulli random variable $X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{otherwise.} \end{cases}$
- Basic models of Multi-arm bandit problem assume Bernoulli Bandits.
- $E[X] = p, E[X^n] = p$.

Binomial $B(n, p)$ random variable.

- Consider a biased coin (head with probability p) and toss it n times.
- Denote head by 1 and tail by 0.
- Let random variable N denote the number of heads in n tosses.
- PMF of N?. $P_N(k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- HW: What is $E[N], E[N^2], Var(X)$?

Geometric random variable

- Consider a biased coin (head with probability p) and suppose you keep tossing it till head appears the first time.
- Let random variable N denote the number of tosses needed for head to appear first time.
- What is the PMF of N? $p_N(k) = (1-p)^{k-1}p$.
- HW: What is $E[N], E[N^2], Var(N)$?

Poisson random variable

- A Poisson random variable X comes with a parameter λ and has $\Omega' = \mathbb{Z}_{\geq 0}$
- PMF: $p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$
- Intuitively its a limiting case of the Binomial distribution with n increasing and p decreasing such that np converges to λ.
- Mean of binomial is np so p should decrease while n increases.

Continuous random variables

- A random variable X is continuous if there exists a non-negative real valued probability density function (PDF) $f_X(\cdot)$ such that $F_X(x) = \int_{u=-\infty}^{x} f_X(u)du$.
- $P_X(a \leq X \leq b) = \int_a^b f_X(u)du$. (Area under the curve)

$$\boxed{\frac{dF_X(x)}{dx} = f_X(x) \text{ or } P_X(x < X \leq x+h) \simeq f_X(x)h.}$$

Mean, Variance, Moments

- $E[X] = \int_{-\infty}^{\infty} u f_X(u) du$
- $E[X^n] = \int_{-\infty}^{\infty} u^n f_X(u) du$
- $E[g(X)] = \int_{-\infty}^{\infty} g(u) f_X(u) du$
- $Var[X] = E[g(X)]$ where $g(x) = (x - E[X])^2$.
- For $Y = aX + b$, $E[Y] = aE[X] + b$.

Exponential random variable ($Exp(\lambda)$)

- This is a non-negative r.v. with parameter λ.
- Its pdf $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.
- Its CDF is given by $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.
- $E[X] = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$
- $E[X^n] = \frac{n!}{\lambda^n}$

Summary: Multiple random variables

$p_{XY}(x, y) := \mathbb{P}\{\omega \in \Omega : X(w) = x \text{ and } Y(\omega) = y\}$.
$F_{XY}(x, y) := \mathbb{P}\{\omega \in \Omega : X(w) \leq x \text{ and } Y(\omega) \leq y\}$.

The marginal PMF's p_X and p_Y can be obtained from the joint PMF as follows:
$p_X(x) = \sum_y p_{XY}(x, y)$ and $p_Y(y) = \sum_x p_{XY}(x, y)$.

Two random variables, X and Y are independent if the following is true:
$p_{XY}(x, y) = p_X(x)p_Y(y), F_{XY}(x, y) = F_X(x)F_Y(y)$ and $E[XY] = E[X]E[Y]$.

$E[g(X, Y)] = \sum_{xy} g(xy)p_{XY}(xy)$

The rules for continuous random variables are similar.
Also revise conditioning of variables.

Sums of independent random variable

- Consider $Z = X + Y$. What is the pdf of Z when X and Y?
- What is $p_Z(z)$ or $f_Z(z)$?
- $p_Z(z) = \sum_{(x,y):x+y=z} p_{X,Y}(x,y)$
- $f_Z(z) = \int_{(x,y):x+y=z} f_{X,Y}(x,y)dxdy$.
- Since X and Y are independent $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ and $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. This gives us

Convolution formula

$$p_Z(z) = \sum_x p_X(x)p_Y(z-x)$$
$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

HW: What if X and Y are not independent?

MGF of Sums of independent random variable

- Consider $Z = X + Y$. What is the pdf of Z when X and Y?
- Let $M_X(t)$ and $M_Y(t)$ be their MGF's. What is $M_Z(t)$?
- $M_Z(t) = E[e^{Zt}] = E[e^{(X+Y)t}]$.
- $M_Z(t) = E[e^{Xt}.e^{Yt}]$.
- If X and Y are independent, $E[XY] = E[X]E[Y]$ and $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- $M_Z(t) = E[e^{Xt}].E[e^{Yt}]$.

$$\boxed{M_Z(t) = M_X(t)M_Y(t).}$$

MGF of Sums of independent random variable

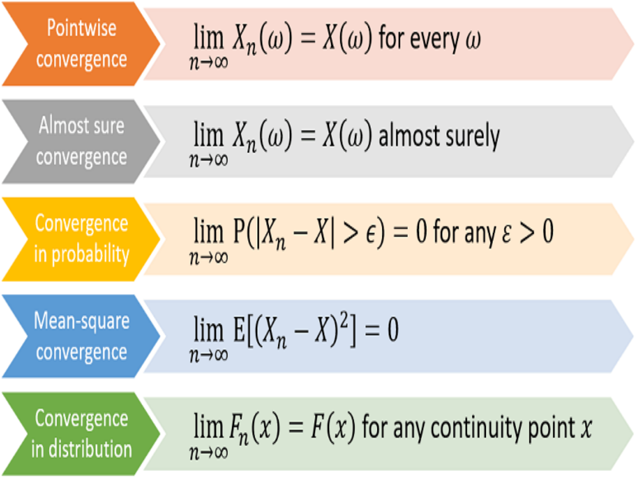
- Consider $Z = X + Y$. What is the MGF of Z when X and Y?

$$M_Z(t) = M_X(t)M_Y(t).$$

- What about $M_Z(t)$ when $Z = X_1 + X_2 + \ldots X_n$ and X_i are iid.?
- $M_Z(t) = (M_X(t))^n$.
- What about $M_Z(t)$ when $Z = X_1 + X_2 + \ldots X_N$ where N is a positive discrete random variable? section 4.5

Convergence of Random Variables

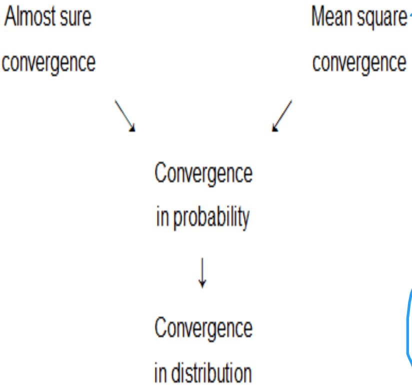
Summary



[1]

[1] Image from probabilitycourse.com

Relation between modes of convergence (no proofs)



https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables