Probability Review

Identities

$P(A \cap B) = P(A)P(B)$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ $\quad P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$

$f(x) = \frac{dF(x)}{dx}, \int_{-\infty}^{\infty} f(x) = 1$ $\quad P(a < X < b) = \int_a^b f(x)dx$

$F(y) = \int_{-\infty}^{y} f(x)dx$ $\quad P(X < x) = F(X)$

$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx$ $\quad Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Uniform Distribution: $X \sim uniform(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases} \quad \text{(pdf)}$$

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases} \quad \text{(cdf)}$$

$\mathbb{E}[X] = \frac{a+b}{2}$ $\quad Var[X] = \frac{(b-a)^2}{12}$

Exponential Distribution: $X \sim exp(\lambda)$

$f(x) = \lambda e^{-\lambda x}$ (pdf)

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

$\mathbb{E}[X] = \frac{1}{\lambda}$ $\quad Var[X] = \frac{1}{\lambda^2}$

$P(X > s + t | X > s) = P(X > t)$ [Memoryless]

Poisson Distribution: $X \sim pois(\lambda)$

$P(N(t) = n) = \frac{(\lambda t)^t}{n!} e^{-\lambda t}$ [pmf]

$\mathbb{E}[X] = Var(X)$

Poisson Process

$N(0) = 0$

$f(x) = \lambda e^{-\lambda x}$ [Arrival See Time Average]

$P(X > s + t | X > s) = P(X > t)$ [Memoryless]

$\lambda = \sum_{i=1}^{n} \lambda_i, Y = \left(\sum_{i=1}^{n} X_i \right) \sim \text{pois}(\lambda)$ [Merge Poisson Processes]

$X \sim pois(\lambda), X = [X_1, X_2]$

$\implies X_{1,2} \sim pois(\frac{\lambda}{2})$ [Split Poisson Processes]

Performance Analysis

Identities

$A_i(t)$ [Number of arrivals] $\quad \lambda_i(t) = \frac{A_i(t)}{t}$ [Arrival Rate]

$C_i(t)$ [Completions] $\quad X_i(t) = \frac{C_i(t)}{t}$ [Throughput]

$B_i(t)$ [Busy time] $\quad \rho_i(t) = \frac{B_i(t)}{t}$ [Utilization]

$S_i(t) = \frac{B_i(t)}{C_i(t)}$ [Avg process time] $\quad S_i(t) = \mathbb{E}[S]$

D_i [Processing time of cycle] $\quad \mathbb{E}[D_i] = \mathbb{E}[S_i]\mathbb{E}[Vi]$

$V_i(t)$ [Visits to device] $\quad V_{user} = V_0 = 1$

$\lim_{t \to \infty} \frac{A_i(t)}{t} = \lim_{t \to \infty} \frac{C_i(t)}{t}$ $\quad \lambda_i = X_i$ [Steady state]

$N(t) = A(t) - C(t)$ [Number of jobs in system]

$R(t) \approx \int_0^t \frac{A(s) - C(s)}{A(t)} ds$ [Avg response time]

$\bar{N}(t) \approx \int_0^t \frac{A(s) - C(s)}{t} ds$ [Avg number of jobs in system]

$\bar{N}(t) = \frac{R(t)A(t)}{t}$

Z [Think time]

$\mathbb{E}[N] = N, \lambda = X, R = R + Z$ [Closed System]

Operation Laws

$\mathbb{E}[N] = \lambda \mathbb{E}[R]$ [Little's Law]

$\rho_i = \mathbb{E}[S_i]X_i = \frac{\lambda_i}{\rho_i}$ [Utilization Law]

$\rho_i = \mathbb{E}[S_i]\mathbb{E}[Vi]X = \mathbb{E}[D_i]X$ [Bottleneck Law]

$X_i = \mathbb{E}[V_i]X$ [Forced Flow Law]

$\mathbb{E}[R] = \frac{N}{X} - \mathbb{E}[Z]$ [Closed System Response Time Law]

Bottleneck Analysis

D_{max} [Bottleneck Device] $\quad D = \sum D_i$

$\mathbb{E}[R] \geq D$ $\quad X = \frac{\rho_{max}}{D_{max}}$

$\mathbb{E}[R] \geq max(D, ND_{max} - \mathbb{E}[Z])$ $\quad X \leq min(\frac{1}{D_{max}}, \frac{N}{D + \mathbb{E}[Z]})$

$N^* = \frac{D + \mathbb{E}[Z]}{D_{max}}$ $\quad \implies$ optimal X and $\mathbb{E}[R]$

Queuing Models

(Arrivals / Service Times / Number of servers / Room in queue)

M/M/1

$\rho = \lambda / \mu$ $\quad \mu > \lambda$ [Stability condition]

$\pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$ $\quad \pi_i = \pi_0 (\frac{\lambda}{\mu})^i = (1 - \rho)\rho^i$

$\mathbb{E}[N] = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}$ $\quad \mathbb{E}[N_Q] = \mathbb{E}[N] - \rho$

$\mathbb{E}[R] = \frac{1}{\mu - \lambda}$ $\quad \mathbb{E}[R_Q] = \frac{1}{\mu - \lambda} - \frac{1}{\mu}$

M/M/c

$\rho = \frac{\lambda}{c\mu}$ $\quad c\mu > \lambda$ [Stability condition]

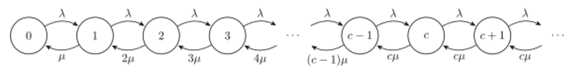
$$\pi_0 = (\frac{\lambda}{\mu})^c \frac{1}{1 - \rho} \quad \pi_i = \begin{cases} \frac{\lambda^i}{i!\mu^i}\pi_0, & \text{if } i < c \\ \frac{\lambda^i}{c!\mu^i c^{i-c}}\pi_0, & \text{if } i \geq c \end{cases}$$

$\mathbb{E}[N] = \lambda \mathbb{E}[R]$ $\quad \mathbb{E}[N_Q] = \lambda \mathbb{E}[R_Q]$

$\mathbb{E}[R] = \mathbb{E}[R_Q] + \mathbb{E}[S] = \mathbb{E}[R_Q] + \frac{1}{\mu}$

$\mathbb{E}[R_Q] = \frac{(\frac{\lambda}{\mu})^c \mu}{(c-1)!(c\mu - \lambda)^2}$

$P(\text{job is queued}) = \sum_{i=0}^{\infty} \pi = \frac{1}{c!}(\frac{\lambda}{\mu})^c \frac{1}{1 - \rho} \pi_0$ [Erlang C Formula]



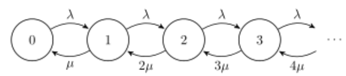
M/M/∞

$\rho = \lambda / \mu$ $\quad \mu > \lambda$ [Always Stable]

$\pi_0 = e^{-\frac{\lambda}{\mu}} = e^{-\rho}$ $\quad \pi_i = \frac{(\lambda/\mu)^i}{i!} e^{-\frac{\lambda}{\mu}} = \frac{\rho^i}{i!} e^{-\rho}$

$\mathbb{E}[N] = \frac{\lambda}{\mu} = \rho$ $\quad \mathbb{E}[N_Q] = 0$

$\mathbb{E}[R] = \frac{1}{\mu} = \mathbb{E}[S]$ $\quad \mathbb{E}[R_Q] = 0$



Birth-Death Process

CTMC where state transitions increase or decrease by a constant factor.

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}}$$

$$\pi_i = \frac{\prod_{j=0}^{i-1} \lambda_j}{\prod_{j=1}^{i} \mu_j} \pi_0$$

Threshold System

$T > 0$, Arrival rate s, processing rate s. If $r > s, N \to 0$. If $s > r, N \to \infty$.

$$\pi_0 = \frac{1}{1 - \frac{r}{s}} (\frac{s}{r})^T - 1$$

$$\pi_i = \begin{cases} (\frac{s}{r})^i \pi_0, & \text{if } i < T \\ (\frac{s}{r})^{i-T} (\frac{r}{s})^2 \pi_0, & \text{if } i \geq T \end{cases}$$

Jackson Networks

1. External arrivals form a Poisson process
2. All service times are exponentially distributed and the service discipline at all queues is first-come, first-served
3. internal routing of jobs between servers is probabilistic
4. The utilization of all of the queues is less than one

Solved via markov model

1. **Markov Chain:** We may solve the corresponding Discrete Time Markov Chain to find its steady state distribution, $\mathbb{E}[N]$, and $\mathbb{E}[R]$. If there are N jobs and k nodes, we will have a lower bound of $\Omega(\binom{N+k-1}{k-1}^2)$ when solving the system of equations.
2. **Product form:** Using a temporary value for each node's arrival rate, $\bar{\lambda}$, determine the ratios between the balance equations and then recover the real values using the actual arrival rate, λ, finding the steady-state distribution, $\mathbb{E}[N]$, and $\mathbb{E}[R]$. Still suffers from a combinatorial explosion in complexity with a lower bound of $\Omega(\binom{N+k-1}{k-1})$.
3. **Mean Value Analysis:** Uses the Arrival Theorem in a recursive algorithm to analyse specific nodes when there are N jobs in the system. We only have access to expectations and utilization of specific nodes, i.e. $\mathbb{E}[R_i]$, $\mathbb{E}[N_i]$, ρ_i but is more performant with an upper bound of $\mathcal{O}(Nk)$.

M/G/1

- Markovian (modulated by a Poisson process), service times have a General distribution and there is a single server
- $\mathbb{E}[S] = \frac{1}{\mu}$
- high variance in service distribution $\implies$ high response time
- Has equal $\mathbb{E}[N]$ for all blind non-pre-emptive service policies

Pollaczek–Khinchine formula

$$\mathbb{E}[N] = \rho + \frac{\rho^2 + \lambda^2 \sigma_s^2}{2(1-\rho)}$$

Service Policies

Blind and non-blind policy relates to knowledge of job size on arrival. If service times that jobs require are known, then the optimal scheduling policy is shortest remaining processing time (SRPT).

- first-come, first-served (FCFS)
- processor sharing (PS) where all jobs in the queue share the service capacity between them equally
- last-come, first served (LCFS) with/without preemption where a job in service may or may not be interrupted with work being conserved
- generalized foreground-background (FB) scheduling also known as least-attained-service where the jobs which have received least processing time so far are served first and jobs which have received equal service time share service capacity using processor sharing
- shortest job first (SJF) with/without preemption, where the job with the smallest size receives service
- shortest remaining processing time (SRPT) where the next job to serve is that with the smallest remaining processing requirement

Failure/Hazard Rate

- Increasing Failure Rate (IFR): h(t) is non-decreasing in t, the expected remaining work is decreasing, non pre-emptive is preferable.
- Decreasing Failure Rate (DFR): h(t) is non-increasing in t, the expected remaining work is increasing, pre-emptive policy is preferable.

$$h(t) = \frac{f(t)}{1 - F(t)} \qquad \mathbb{E}[\text{Remaining time}] = \frac{1}{h(t)}$$

$$X \sim uniform(a, b) \text{ (IFR)} \implies \qquad h(t) = \frac{1}{b - t}$$

$$X \sim exp(\lambda) \text{ (IFR and DFR)} \implies \qquad h(t) = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})}$$

$$\text{Time average Excess} = \mathbb{E}[S_c] = \frac{\mathbb{E}[S_c]}{2\mathbb{E}[S]}$$

$$\mathbb{E}[R_Q] = \frac{\rho}{1-\rho} \mathbb{E}[S_c]$$

Pareto Distribution

- popular DFR, "80-20 rule", pre-emptive policy is preferable
- 50% of the load on the system comes from 1% of the jobs
- α shape parameter, $\alpha = 1, X > t \implies P(X > 2t) = \frac{1}{2}$
- $0 < alpha < 1$, $Var(X) = \infty$, $\mathbb{E}[X] = \infty$
- Survival Function:

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} (\frac{x_{\mathrm{m}}}{x})^{\alpha} & x \geq x_{\mathrm{m}}, \\ 1 & x < x_{\mathrm{m}}, \end{cases}$$

Misc

- $\sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}, |\alpha| < 1.$
- $h = \frac{f}{g} \implies h' = \frac{f'g - fg'}{g^2}$
- Max system utilization $\implies$ only bottleneck utilization is 100%
- Want to minimize $\mathbb{E}[R]$ and maximize X.
- Operation Laws work regardless of distributions of random variables
- exponential distributions are a very good assumption for modeling arrivals, but only moderately good for modelling processing times