Poisson and Markov Processes

Objectives

• Review of random variables of interest
• Poisson Process
• Queueing theory and Markov chains

Review of Probability

• A probability space is the triplet \((\Omega, \mathcal{B}, \mathbb{P})\) where \(\Omega\) is the sample space, \(\mathcal{B}\) is an algebra of sets of the sample space and \(\mathbb{P}\) is a probability measure
• A random variable is a function that maps the sample space onto another space where it can be measured
• A random variable \(X\) has cdf \(F_X(x) = P(X \leq x)\) and pdf \(f_X(x)\)
Exponential Distribution

- Probability Distribution Function of random variable $X$
  \[ F_X(t) = P(X < t) = 1 - e^{-\lambda t}, \quad t \geq 0. \]
  - mean value $= E(X) = \frac{1}{\lambda}$
  - second moment $= E(X^2) = \frac{1}{\lambda^2}$
  - variance $= \text{var}(X) = \frac{1}{\lambda^2}$
- Probability Density Function
  \[ f_X(t) = \frac{dF_X(t)}{dt} = \lambda e^{-\lambda t}, \quad t \geq 0. \]

Memoryless Property

- Consider an exponential random variable $T$, with parameter $\lambda$.
- Given an event occurred at time $x$.
  \[
P(T < (x+t) \mid T > x) = \frac{P(x < T < (x+t))}{P(T > x)} = \frac{P(T < (x+t)) - P(T < x)}{1 - P(T < x)}
  = \frac{e^{-\lambda(x+t)} - e^{-\lambda x}}{1 - e^{-\lambda x}} = e^{-\lambda x} \frac{e^{-\lambda t} - 1}{e^{-\lambda t} - 1} = e^{-\lambda x} \frac{e^{-\lambda t} - 1}{e^{-\lambda t} - 1} = (1 - e^{-\lambda t})
\]

Memoryless Property

This property states that the distribution is the same if we consider a point in time $x>0$ or $x=0$.

The event does not remember that it has already lasted for $x$ seconds.
Random Processes

- A random process is a collection of indexed random variables, examples: \( \{X_k, k=0,1,\ldots\} \), \( \{X(t), t\geq 0\} \)
- The random variables can be discrete or continuous
- The index set can be discrete or continuous
- In general, time (discrete or continuous) is considered the index set

Arrival Processes

- Counting process \( \{N(t), t \geq 0\} \) with nonnegative integers as state space such that \( N(0)=0 \) and whose sample functions are nondecreasing
  - \( N(t) \) can be viewed as number of events that have occurred up to time \( t \), i.e., in \( [0,t) \)
  - \( N(t) \geq 0 \), for all \( t \)
  - If \( s < t \) then \( N(s) < N(t) \)
  - The increment \( N(t)-N(s), s < t \), can be viewed as number of arrivals between \( s \) and \( t \)

Counting Processes

- Example of sample function of counting process

\[
\begin{align*}
N(t) & \\
\text{Random variables are discrete} & \\
0 & \\
1 & \\
2 & \\
3 & \\
4 & \\
N(0) = 4 \\
\text{Time is continuous} & \\
\end{align*}
\]
Arrival Process

- The counting process \( \{N(t), t \geq 0\} \) is said to be a Poisson process having rate \( \lambda > 0 \), if
  - \( N(0) = 0 \)
  - The process has independent increments
  - The number of events in any interval of length \( \delta \) is Poisson distributed with mean \( \lambda \delta \), i.e., for all \( s,t \geq 0 \) we have
    \[
    P[N(t) - N(s) = n] = P[N(t - s) = n] = e^{-\lambda s} \frac{(\lambda s)^n}{n!}, \quad n = 0,1,\ldots
    \]
- The third condition implies stationary increments and that the mean value is \( \lambda \).

Poisson Process

- Poisson process \( \{N(t), t \geq 0\} \) with rate \( \lambda \) is a counting process such that
  - increments corresponding to nonoverlapping time intervals are independent (Independent Increments property)
  - the increment \( N(t) - N(s) \) for \( s < t \), is a Poisson random variable with parameter (mean) \( \lambda(t - s) \) (Stationary increments)

Poisson Process

- Properties of Poisson process
  - It can be viewed as a continuous-time Markov chain
  - The interarrival times are iid (independent and identically distributed) random variables with exponential distribution and parameter \( \lambda \).
  - Superposition: If \( N(t) \) and \( M(t) \) are independent Poisson processes with rates \( \lambda \) and \( \mu \), then the process \( Z(t) = N(t) + M(t) \) is a Poisson process with rate \( \lambda + \mu \).
### Poisson Process

- **Properties**
  - **Decomposition**: Let $Z(t)$ be a Poisson process with rate $\lambda$. Each time an arrival occurs, perform a Bernoulli trial where probability of success is $p$. If success, assign arrival to counting process $N(t)$; otherwise to $M(t)$. Then $N(t)$ and $M(t)$ are independent Poisson processes with rates $\lambda = p \lambda$ and $\nu = (1 - p) \lambda$, respectively.

### Why Poisson arrivals?

- A renewal process is a counting process whose interarrival times are iid.
  - interarrival time distribution need not be exponential
- **FACT**: Given $n$ independent renewal processes with the same interarrival distribution, each with rate $\lambda/n$. For large $n$, the superposition is approximately Poisson.
- Consequence: for networks with dense topology, assuming Poisson arrivals may be an accurate approximation (Kleinrock independence assumption)
  - recent experiments suggest other models may be better (e.g., self-similar traffic)

### Examples Poisson Process
Queue Model

**System**
- $T$ = Total average service time,
- $N$ = Total average number of users in the system
- Service rate $\mu$ = pkts/sec

**Queue or Buffer**
- $W$ = average waiting time in queue,
- $N_Q$ = Average number of users in queue

**Example**
- $T_k$ = Total time in the system of user $k$ (waiting and service)
- $D_k$ = Departure time of user $k$
- $S_k$ = Service time in the system of user $k$
- $A_k$ = Arrival time of user $k$

**Little's Theorem**
A facility with an arrival process with intensity $\lambda$ and an average service time of $1/\mu$ has an average number of users in the system, $N$, and an average number of users in the queue, $N_Q$. The average time spent in the system per user is $T$, and the average time spent in queue per user is $W$. By ergodicity arguments, these satisfy the following:
- $N = \lambda T$
- $N_Q = \lambda W$
- $T = W + (1/\mu)$
Examples Little's theorem

Queueing Theory and Markov Chains

Objectives

- The M/M/1 queue
- Birth-Death processes
Kendall Notation for Queues

\[ A|S|m|C|K/SD \]

- \( A \): Interarrival time distribution,
  - \( M \): Exponential
  - \( G \): General
  - \( E_k \): Erlang-\( k \)
  - \( H_k \): Hyperexponential-\( k \)
  - \( D \): Deterministic
- \( S \): Service time distribution (same as \( A \))
- \( m \): Number of servers (it could be infinite)
- \( C \): Capacity of system (includes customers in service and customers waiting in line, it could be infinite)
- \( K \): Population size (it could be infinite)
- \( SD \): Service Discipline (FCFS, FCJS, Random, Priority, etc., when not used, FCFS is assumed)

Multiserver Queue

\( M/M/N \)

![Multiserver Queue Diagram](multiserver_queue_diagram.png)

Multiple Single-server queues

\( N \) systems \( M/M/1 \)

![Multiple Single-server Queues Diagram](multiple_single_server_queues_diagram.png)
Kleinrock Independence Assumption (KIA)

- Suppose packet-switched network satisfies
  - Poisson arrivals at entry points
  - Packet lengths that are (approximately) exponentially distributed
  - Densely connected topology
  - Moderate-to-heavy traffic loads
- Then it is “reasonable” to treat interarrival times to be independent of packet length
- Intuition behind KIA is that merging of multiple packet streams has effect of restoring independence of interarrival times and packet lengths

Queueing Theory

- M/M/1 (Bertsekas, D., and Gallager, R. Data Networks)
**Poisson Process**

\[ P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, 2, \ldots \]

- Probability of no arrivals
  \[ P(N(t) = 0) = e^{-\lambda t} = 1 - \lambda t + o(t). \]
- Probability of one arrival
  \[ P(N(t) = 1) = \lambda te^{-\lambda t} = \lambda t + o(t). \]

**Poisson Process**

- Probability of more than one arrival
  \[ P(N(t) \geq 2) = 1 - e^{-\lambda t} - \lambda te^{-\lambda t} = o(t). \]
- Where \( o(t) \) is a function such that
  \[ \frac{o(t)}{t} \xrightarrow{t \to 0} 0 \]
- Example, \( o(t) = t^2 \)

**Birth-Death Processes**

- Assume arrivals with rate \( \lambda \)
- Assume service rate \( \mu \)
- System is in state \( E_k \) when there are \( k \) customers present
- **Birth**: If system is in state \( E_k \), transition from \( E_k \) to \( E_{k+1} \) occurs with probability \( \lambda e^{-kt} + o(k) \) in \( (t, t+h] \)

![Birth-Death Processes Diagram]

\[ \ldots E_{k-1} \quad E_k \quad E_{k+1} \quad \ldots \]
Birth-Death Processes

• Death: If system is in state $E_k$, transition from $E_k$ to $E_{k-1}$ occurs with probability $\mu_k h + o(h)$ in $(t, t+h)$
• Probability that the state changes by more than one unit is $o(h)$ during $(t, t+h)$
• Let $N(t)$ be the number of customers present at time $t$, then the state distribution is $P\{N(t)=k\} = p_k(t)$

---

Birth-Death Process

$$p_k(t+h) = p_{k-1}(t)\lambda_{k-1}h + p_{k+1}(t)\mu_k h + p_k(t)(1 - \lambda_k h - \mu_k h),$$

$$p_k(t+h) - p_k(t) = p_{k-1}(t)\lambda_{k-1}h + p_{k+1}(t)\mu_k h - p_k(t)(\lambda_k h + \mu_k h) + o(h),$$

$$\frac{dp_k(t)}{dt} = \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t) - (\lambda_k h + \mu_k h)p_k(t), \quad k = 1, 2, \ldots$$

Birth-Death Process

• In steady state solution we have

$$p_k(t) \xrightarrow{t \to \infty} p_k,$$

then

$$0 = \mu_k p_k - \lambda_k p_k,$$

and

$$0 = \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} - (\lambda_k + \mu_k) p_k, \quad k = 1, 2, \ldots$$

which gives

$$\mu_k p_k = \lambda_k p_k,$$

and

$$\lambda_k p_k - p_{k+1} p_{k+1} = \lambda_{k-1} p_{k-1} - p_k p_k, \quad k = 1, 2, \ldots$$
Birth-Death Process

Define $f(k) = \lambda_k P_k - \mu_{k+1} P_{k+1}$, $k = 0, 1, \ldots$, and $f(-1) = 0$, then

$$\lambda_k P_k - \mu_{k+1} P_{k+1} = \lambda_{k+1} P_{k+1} - \mu_k P_k, \quad k = 1, 2, \ldots$$

will give

$$f(k) = f(k-1), \text{ and } f(0) = \lambda_0 P_0 - \mu_1 P_1 = 0 = f(-1).$$

hence $f(0) = f(-1) = 0, f(1) = f(0) = 0, f(2) = f(1) = 0, \ldots$.

in general $f(k) = f(k-1) = 0$, i.e., $f(k) = 0, \quad k = 0, 1, 2, \ldots$.

Therefore $\lambda_k P_k - \mu_{k+1} P_{k+1} = 0, \quad k = 0, 1, 2, \ldots$

Detailed balance equation

Birth-Death Process

Steady state solution

$$P_k = \left[ \frac{\lambda_{k+1}}{\mu_{k+1}} \right] P_0, \quad k = 1, 2, \ldots$$

$$P_0 = \left[ 1 + \sum_{j=0}^{\infty} \prod_{l=j}^{\infty} \left( \frac{\lambda_{l+1}}{\mu_{l+1}} \right) \right]^{-1}.$$
**M/M/1 Queue**

Global balance equations: Flow out = Flow in

- $k$: state of the system representing number of users present in the system (in queue and in service)

$p_k$: steady state probability that system is in state $k$

$$p_k = \frac{\lambda}{\mu} p_{k-1} = \frac{\lambda}{\mu} p_{k-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^k p_0.$$  
$$\sum_{k=0}^{\infty} p_k = 1.$$

**Detailed balance equations**

$$\mu p_k = \lambda p_{k-1}.$$  

**Steady State Distribution**

- $p_0 + \sum_{k=1}^{\infty} p_k = 1,$
- $p_0 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k p_0 = 1,$
- $p_0 = \left[\sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k\right]^{-1},$
- $p_0 = \left(1 - \frac{\lambda}{\mu}\right) = (1 - \rho),$  
- $p_k = \rho^k (1 - \rho).$

**Server Utilization**

- $\rho = \frac{\lambda}{\mu}$

**Throughput**

- $N = E(K) = \sum_{k=0}^{\infty} k p_k,$
- $N = \frac{\rho}{1 - \rho},$
- $T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda},$
- $N_Q = \frac{\rho^2}{1 - \rho},$
- $W = \frac{\rho}{\mu - \lambda}.$

- Throughput = utilization / service time = $\frac{\rho}{T}$
- For $\rho = 0.5$ and $T = 1$ ms, Throughput is 500 packets/sec
**M/M/m Queue**

Global balance equations: Flow out = Flow in
for state $k$ of the system representing number of users present in the system (in queue and in service)

$$p_k = \text{steady state probability} \sum_{k=0}^{\infty} p_k = 1,$$
that system is in state $k$.

$$p_k = \frac{\lambda}{k\mu} p_{k-1} = \frac{\lambda}{k\mu(k-1)\mu} p_{k-2} = \cdots = \frac{1}{k!\mu} \left(\frac{\lambda}{\mu}\right)^k p_0, \quad 0 \leq k \leq m.$$

$$p_k = \frac{\lambda}{m\mu} p_{k-1} = \frac{\lambda}{m\mu m\mu} p_{k-2} = \cdots = \left(\frac{\lambda}{m\mu}\right)^k p_0, \quad k \geq m.$$

**Steady State Distribution**

$$p_0 = \frac{\sum_{k=0}^{m-1} (m\rho)^k}{k!} + \frac{(m\rho)^m}{m! (1 - \rho)}^{-1},$$

$$p_k = \begin{cases} p_0 \left(\frac{m\rho}{k!} \right), & k \leq m, \\ \frac{m^m p^m}{m!}, & k > m. \end{cases}$$

where $\rho = \frac{\lambda}{m\mu} < 1$.

**Erlang C Formula**

$$N = E(K) = \sum_{k=0}^{\infty} kp_k,$$

$$N = mp \cdot \frac{\rho P \rho_0}{1 - \rho},$$

$$W = \frac{\rho P_0}{\lambda(1 - \rho)},$$

$$N_Q = \frac{\rho P_0}{(1 - \rho)},$$

$$T = W + \frac{1}{\mu}.$$
**Detailed balance equations**

\[ k\mu p_k = \lambda p_{k-1}. \]

\[ p_k = \frac{\lambda}{k\mu} p_{k-1} = \frac{\lambda}{k\mu(k-1)\mu} p_{k-2} = \cdots = \frac{1}{k!} \frac{\lambda^k}{\mu^k} p_0. \]

\[ \sum_{k=0}^{\infty} p_k = 1. \]

---

**Steady State Distribution**

\[ p_0 = \left( \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right)^{-1}, \]

\[ p_k = \frac{\lambda^k}{k!} p_0, \quad k = 1, 2, \ldots, m. \]

\[ p_m = \frac{(\lambda/\mu)^m}{\sum_{k=0}^{\infty} (\lambda/\mu)^k / k!}. \]

\( p_m \) is the probability that an arriving customer will be denied service.

---

**Erlang B Formula**

- Poisson arrivals with rate \( \kappa \) into a link with \( C \) channels
- A call is blocked if all \( C \) channels are busy
- A call occupies a single channel for holding period
- *Call holding periods* are independent of each other and of the arrival times, and are identically distributed with mean \( 1/\mu \) (insensitivity property)
- Erlang B formula gives proportion of calls that are lost

\[ \frac{E}{C} = \frac{\lambda/\mu}{\sum_{k=0}^{\infty} (\lambda/\mu)^k / k!}. \]
\( M/M/\infty \) Queue

Detailed balance equations:

\[
k\mu p_k = \lambda p_{k-1}.
\]

\[
P_k = \frac{\lambda}{k\mu} P_{k-1} = \frac{\lambda}{k\mu(k-1)\mu} P_{k-2} = \cdots = \frac{1}{k!} \left( \frac{\lambda}{\mu} \right)^k P_0.
\]

\[
\sum_{k=0}^{\infty} p_k = 1.
\]

\( M/M/\infty \) Queue

Steady State Distribution

\[
p_0 = \left[ \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mu \right]^{-1} = e^{-\frac{\lambda}{\mu}},
\]

\[
p_k = \left( \frac{\lambda^k}{k!} \right) \mu^{-1} e^{-\frac{\lambda}{\mu}}, \quad k = 0, 1, 2, \ldots
\]

Poisson distribution

Mean \( \lambda/\mu \), Variance \( \lambda/\mu \)

Peakedness \( Z = V/M-1 \)

Time/Call Congestion

- Time Congestion, \( E \): Fraction of time when all servers are busy. We say that the system is in the blocked state, and calls finding the system in that state are said to be blocked. Calls blocked are lost and do not return or may be sent to other groups as overflow.
- Call Congestion, \( B \): Probability that an arriving call will find the system fully occupied. In general \( E \neq B \).