EECS126 Course Notes [Spring 2021]

Patrick Yin

Updated April 1, 2021

Contents

1 Note

These course notes are my notes from EECS 126 : Probability and Random Processes. The course is linked here. These course notes are in progress.

2 Probability Basics

For this section, since this is mostly review, I will brush over most topics and state them without proof.

2.1 Probability Foundamentals

Definition 2.1 (Probability Space)**.** A probability space is a triple $(\Omega, \mathcal{F}, P)$ where Ω is the sample space, $\mathcal{F}$ is the family of subsets of Ω, and P is the probability measure.

Technical Assumption: $\mathcal{F}$ is a σ-algebra containing Ω itself, meaning that the countable complements/unions/intersections of events are events.

Definition 2.2 (Kolmogorov Axioms)**.** Probability measures must obey the Kolmogorov Axioms:

- $P(A) \geq 0 \quad \forall A \in \mathcal{F}$
- $P(\Omega) = 1$
- If $A_1, A_2, ... \in \mathcal{F}$ and $A_i \cap A_j = \emptyset \quad \forall i \neq j$, then $P(\cup_{i \geq 1} A_i) = \sum_{i \geq 1} P(A_i)$

Theorem 1 (Law of Total Probability)**.** *If A_i are disjoint and $\cup_{i \geq 1} A_i = \Omega$, then $P(B) = \sum_{i \geq 1} P(A_i \cap B)$.*

Definition 2.3 (Conditional Probability)**.** If B is an event with $P(B) > 0$, then conditional probability of A given B is $P(A|B) := \frac{P(A \cup B)}{P(B)}$.

Theorem 2 (Bayes Rule)**.** *If events A and B have positive probability, then $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.*

Definition 2.4 (Independence)**.** Events A, B are independent if $P(A \cap B) = P(A)P(B)$.

Definition 2.5 (Conditional Independence)**.** If events A, B, C with $P(C) > 0$ satisfy $P(A \cup B|C) = P(A|C)P(B|C)$.

2.2 Random Variable

Definition 2.6. A random variable is a function $X : \Omega \to \mathbb{R}$ with the property $\{\omega \in \Omega : X(\omega) \leq \alpha\} \in \mathcal{F} \quad \forall \alpha \in \mathbb{R}$.

This means that $P(X \leq \alpha) := P(\{\omega \in \Omega : X(\omega) \leq \alpha\})$. Technical definition of r.v. implies that

- If X, Y are r.v.s, then so is $X + Y$, XY, X^p where $p \in \mathbb{R}$
- If $X_1, X_2, ...$ are r.v.s, then so is $\lim_{n \to \infty} X_n$

Definition 2.7 (Discrete Random Variables)**.** A discrete r.v. is a r.v. that takes countably many values.

Definition 2.8 (Continuous Random Variables)**.** A continuous r.v. is a r.v. defined via its density $f_X : \mathbb{R} \to [0, \infty)$. So $Pr\{X \in B\} = \int_B f_X(x)dx$ where $f_X \geq 0$ and $\int_{\mathbb{R}} f_X(x)dx = 1$.

2.3 Expectation

Definition 2.9 (Expectation)**.** For a discrete r.v. X, its expectation is $\mathbb{E}[X] = \sum_{x \in X} xp_x(x)$ provided that the series exists. For a continuous r.v., its expectation is $\mathbb{E}[X] = \int_{\mathbb{R}} xf_X(x)dx$. More generally, $\mathbb{E}[g(X_1, ..., X_n)] =$
$\int ... \int_{\mathbb{R}^n} g(x_1, ..., x_n)f_{X_1,...,X_n}(x_1, ..., x_n)dx_1...dx_n$

Theorem 3 (Law of the Unconscious Statistician)**.** *If $Y = g(X)$ and $g : X \to \mathbb{R}$, then Y is a r.v. and $\mathbb{E}[Y] = \sum_{x \in X} g(x)p_X(x)$.*

Theorem 4 (Linearity of Expectation)**.** *$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ where $a, b \in \mathbb{R}$.*

Theorem 5 (Product of Expectation of Independent R.V.s)**.** *If X, Y are independent random variables, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$*

Theorem 6 (Tail Sum Formula for Expectation)**.** *For a discrete r.v., $\mathbb{E}[X] = \sum_{k=1}^{\infty} Pr\{X \geq k\}$*

2.4 Variance, Covariance, and Correlation

Definition 2.10 (Variance)**.** $\text{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Theorem 7 (Sum of Variances of Independent R.V.s)**.** *If X, Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.*

Definition 2.11 (Covariance)**.** $\text{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$

X, Y independent $\implies$ Cov(X, Y) = 0 $\iff$ X, Y are uncorrected

Definition 2.12 (Correlation Coefficient)**.** $\rho(X, Y) := \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Note that $|\rho(X, Y)| \leq 1$.

2.5 Multiple Random Variables

Definition 2.13 (Conditional Expectation)**.** If X is a discrete r.v., $\mathbb{E}[X|Y = y] := \sum_{x \in X} xp_{X|Y}(x|y)$. If Y is a continuous r.v., $\mathbb{E}[X|Y = y] := \int xf_{X|Y}(x|y)dx$.

Theorem 8 (Tower Property)**.** $\mathbb{E}[f(Y)X] = \mathbb{E}\big[f(Y)\mathbb{E}[X|Y]\big]$

If $f(Y) = 1$, then $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

Definition 2.14 (Conditional Variance)**.** $\mathrm{Var}(X|Y=y) = \mathbb{E}[(X - \mathbb{E}[X|Y=y])^2|Y=y] = \mathbb{E}[X^2|Y=y] - (\mathbb{E}[X|Y=y])^2$

Definition 2.15 (Minimum Mean Square Error (MMSE))**.** $\mathbb{E}[\mathrm{Var}(X|Y)] = \mathbb{E}[(X - \mathbb{E}(X|Y))^2]$

Theorem 9 (Law of Total Variance)**.** $Var(X) = \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y])$

2.6 Notes on Distributions

2.6.1 Exponential

Exp(λ) is the unique continuous r.v. with the memoryless property: $P(X > t+s|X > s) = P(X > t) \quad \forall s, t \geq 0$.

2.7 Order Statistics

Let $X_1, .., X_n$ be IID and sort them so that $X^{(1)} \leq ... \leq X^{(n)}$. Then

$$f_{X^{(i)}}(y) = n\binom{n-1}{i-1}F_X(y)^{i-1}(1-F_X(y))^{n-i}f_X(y)$$

2.8 Moment Generating Function

A moment generating function (MGF) encodes moments of a distribution into coefficients of some power series.

$$M_X(t) := \mathbb{E}[e^{tX}] = \mathbb{E}\Big[\sum_{n\geq 0}\frac{(tX)^n}{n!}\Big] = \sum_{n\geq 0}\frac{t^n}{n!}\mathbb{E}[X^n] \quad t \in \mathbb{R}$$

In fact, if an MGF exists, it uniquely determines the distribution of X. To recover the nth moment we simply do

$$\frac{d^n}{dt^n}M_X(t)|_{t=0} = \mathbb{E}[X^n]$$

2.9 Concentration Inequalities

Theorem 10 (Markov Inequality)**.** *If X is non-negative r.v.,* $P(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \quad t > 0$

Theorem 11 (Chebyshev's Inequality)**.** $P(|X - \mathbb{E}[X]| \geq t) \leq \frac{Var(X)}{t^2}$

Theorem 12 (Chernoff Bound)**.** $P(X \geq a) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = e^{-ta}M_X(t) \quad t > 0$

2.10 Convergence of Random Variables

Definition 2.16. Three modes of convergence:

- Almost Sure Convergence: $X_n \to X$ a.s. if $P(\lim_{n\to\infty} X_n = X) = 1$
- Convergence in Probability: $X_n \to X$ i.p if $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$ for $\epsilon > 0$
- Convergence in Distribution: $X_n \to X$ i.d. if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all continuity poiints x of F_X

Note that $a.s. \implies i.p \implies i.d.$

Theorem 13 (Weak Law of Large Numbers (WLLN))**.** $\frac{1}{N}\sum_{i=1}^{N} X_i \to \mathbb{E}[X]$ *in probability if* X_i *are IID and* $\mathbb{E}[|X|] < \infty$*.*

Theorem 14 (Strong Law of Large Numbers (SLLN))**.** $\frac{1}{N}\sum_{i=1}^{N} X_i \to \mathbb{E}[X]$ *almost surely if* X_i *are IID and* $\mathbb{E}[|X|] < \infty$*.*

Theorem 15 (Central Limit Theorem (CLT))**.** *Let* X_i *be IID and* $Var(X) = \sigma^2 < \infty$ *and* $\mathbb{E}[X] = \mu$*. We define* $S_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}$*. Then* $S_n \to \mathcal{N}(0,1)$ *i.d.*

3 Information Theory

3.1 Definitions

Definition 3.1 (Entropy)**.** $\mathcal{H}(X) := \sum_x p_X(x)\text{log}\,\frac{1}{p_X(x)} = \mathbb{E}[\log\frac{1}{p_X(X)}]$

Definition 3.2 (Mutual Information)**.** $I(X;Y) := \sum p_{XY}(x,y)\text{log}\,\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$

3.2 Asymptotic Equipartition Theorem (AEP)

Theorem 16 (AEP)**.** *If* $(X_i)_{i\geq 1} \overset{IID}{\sim} p_X$*, then* $-\frac{1}{n}log\, p(X_1, ..., X_n) \to \mathcal{H}(X)$ *i.p.*

Proof. By WLLN, $-\frac{1}{n}\text{log}\, p(X_1,...,X_n) = \frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_X(X_i)} \to \mathbb{E}[\log\frac{1}{p_X(X)}] = \mathcal{H}(X)$ i.p. □

In other words, with overwhelming probability, we see that $p(X_1,...,X_n) \approx 2^{-nH(X)}$.

Definition 3.3 (Typical Set)**.** Fix $\epsilon > 0$ and for each $n \geq 1$ define the typical set:

$$A_\epsilon^{(n)} = \{(X_1,...,X_n) : p(X_1,...,X_n) \geq 2^{-n(H(X)+\epsilon)}\}$$

- $P((X_1,...,X_n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$ by AEP
- $|A_\epsilon^{(n)}| \leq 2^{n(H(x)+\epsilon)}$ because

$$1 \geq \sum_{(X_1,...,X_n)\in A_\epsilon^{(n)}} p(X_1,...,X_n) \geq \sum_{(X_1,...,X_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = |A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}$$

3.3 Source Coding Theorem

Theorem 17 (Source Coding Theorem)**.** *For any $\epsilon > 0$, IID discrete r.v.s X_i can be losslessly represented using $\leq n(H(x)+\epsilon)$ bits (for all n sufficiently large). Conversely, any representation using $< nH(X)$ bits is impossible without loss of information.*

Proof. We will prove the achievability part of the theorem. Our protocol for source coding will be:

- If I observe $(X_1,...,X_n) \in A_{\frac{\epsilon}{2}}^{(n)}$, I will describe it using $\sim$ log $|A_{\epsilon/2}^{(n)}|$ bits $\leq n(H(X)+\epsilon/2)$
- If I observe $(X_1,...,X_n) \notin A_{\epsilon/2}^{(n)}$, I just describe it brute force using $n\text{log}|X|$ bits.

Then

$$\begin{aligned}\mathbb{E}[\#\text{ bits}] &\leq n(H(X)+\frac{\epsilon}{2})P((X_1,...,X_n) \in A_{\epsilon/2}^{(n)}) + n\text{log}\,|X|P((X_1,...,X_n) \notin A_{\epsilon/2}^{(n)} \\ &\leq n(H(X)+\epsilon) \text{ for all n sufficiently large}\end{aligned}$$

□

3.4 Information Transmission

Fix a rate $R > 0$, send message $M \sim \text{Uniform}(1...2^{nR})$. It takes nR bits to represent $H(M) = nR$. The message is encoded into $X^n(M)$, put through a noisy channel to become Y^n, and then decoded to become $\hat{M}(Y^n)$. The rate $R = \frac{H(M)}{n}$ and the error probability $P_e^{(n)} := P(\hat{M} \neq M)$.

Definition 3.4 (Capacity). $C = \max_{p_X} I(X;Y) =$ max mutual info between channel input and output over all input distributions

Theorem 18 (Shannon's Channel Coding Theorem). *Fix channel $p_{Y|X}$, $\epsilon > 0$, and $R < C$.*

- *For all n sufficiently large, there exists rate-R communication scheme (encoder/decoder) that achieves $P_\epsilon^{(n)} < \epsilon$*
- *If $R > C$, then $P_e^{(n)} \to 1$ for any sequence of communication schemes.*

Definition 3.5 (Binary Symmetric Channel (BSC)). In BSC(p), each input is flipped independently with probability p. $C = 1 - H_2(p)$ where $H_2(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$.

Definition 3.6 (Binary Erasure Channel (BEC)). In BEC(p), each input is erased independently with probability p. $C = 1 - p$.

Proof of Channel Coding Theorem for BEC(p). Suppose we have n channel uses and knew which positions were erased and un-erased. There are then $\leq n(1 - p + \epsilon)$ unerased positions with overwhelming probability for any $\epsilon > 0$ and n sufficiently large. We can only reliably send $\approx n(1-p)$ bits so $R \leq 1 - p$. So we have proved that for $R > C$, this is not possible.
To prove that $R < C$ allows reliable communication, we fix $R < 1 - p - \epsilon$ and generate a random matrix $C \in \mathbb{R}^{(n \times 2^{nR})}$ such that $C_{ij} \overset{IID}{\sim} B(1/2)$. Our protocol is to give C to both the encoder and decoder, send row M of C, and on receiving Y^n look for row in C that matches modulo erasures (error if ≥ 2 rows match what was received).

$$\begin{aligned}
\mathbb{E}_c[P_\epsilon^{(n)}] &= \sum_{E \subset [n]} \mathbb{E}[1\{\hat{M} \neq M\}|E]P(\text{bits erased } = E) \\
&\leq \sum_{E:|E| \leq n(p+\epsilon/2)} \mathbb{E}[1\{\hat{M} \neq M\}|E]P(E) + P(\frac{1}{n}|E| > p + \epsilon/2) \\
&\leq \sum_{E:|E| \leq n(p+\epsilon/2)} P(\cup_{m\geq 2}^{2^{nR}}\{C(1,[n]\backslash E) = C(m,[n]\backslash E)\}|E)P(E) \\
&\leq \sum_{E:|E| \leq n(p+\epsilon/2)} \sum_{m \geq 2}^{2^{nR}} (\frac{1}{2})^{n-|E|}P(E) \\
&\leq \sum_{E:|E| \leq n(p+\epsilon/2)} 2^{-n\epsilon}P(E) \\
&\to 0 \text{ as } n \to \infty
\end{aligned}$$

So there must exist some sufficiently large n such that $P_e^{(n)} < \epsilon$. Note that in line 2, the right hand term goes to zero as n goes to infinity. □

4 Discrete Time Markov Chains (DTMCs)

4.1 Construction

Definition 4.1 (Markov Chain)**.** $(X_n)_{n\geq 0}$ is a MC if each r.v. X_i is a discrete r.v. taking values in discrete set S, and for all $n \geq 0$ and $i, j \in S$

$$P(X_{n+1} = j | X_n = i, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} | X_n = i)$$

For this course, we will be workly only with temporally homogeneous markov chains.

Definition 4.2 (Temporally Homogeneous Markov Chains)**.** For temporally homogeneous markov chains, $P(X_{n+1} = j | X_n = i) = p_{ij}$. In other words, transition probabilities don't depend on time.

Theorem 19 (Chapman-Kolmogorov Equations)**.** *n-step transition probabilities can be computed as* $P^n_{ij} = \left[P^n\right]_{ij}$. *Note that* $P(X_n = j | X_0 = i) := P^n_{ij}$.

4.2 Classification of States

If there is a path form i to j, then we say $i \to j$. If there is also path from j to i, then i and j communicate (i.e. $i \leftrightarrow j$). $\leftrightarrow$ is an equivalence relation on S. In other words, it partitions S into classes of communicating states.

Definition 4.3 (Irreducible)**.** A MC is irreducible if it has only one class.

Define $T_i = \min\{n \geq 1 : X_n = i\}$ and $\text{period}(i) := GCD\{n \geq 1 : P^n_{ii} > 0\}$. So aperiodic means period is 1. Below are a list of class properties:

- Recurrent if process starting at start i revisits state i with probability one
- Transient if it is not recurrent
- Positive Recurrent if recurrent and $\mathbb{E}[T_i | X_0 = i] < \infty$
- Null Recurrent if recurrent and $\mathbb{E}[T_i | X_0 = i] = \infty$
- Periodicity (i.e. period is the same in same class)

4.3 Big Theorem

Definition 4.4 (Stationary Distribution)**.** A probability distribution $\pi = (\pi_i)i \in S$ is said to be a stationary distribution if $\pi = \pi P$. In other words, $pi_j = \sum_{i \in S} \pi_i p_{ij} \quad \forall j \in S$.

Theorem 20 (Big Theorem for Markov Chains)**.** *Let* $(X_n)_{n\geq 0}$ *be an irreducible MC. Exactly one of the following is true:*

1. *Either all states are transient, or all are null recurrent. In this case, no stationary distribution exists, and* $\lim_{n\to\infty} P^n_{ij} = 0$ *for all* $i, j \in S$

2. *All states are positive recurrent. In this case, a stationary distribution π exists. It is unique and satisfies*

$$\pi_j = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mathbb{E}[T_j | x_0 = j]}$$

Moreover, if the MC is aperiodic, then

$$\lim_{n\to\infty} P_{ij}^n = \pi_j \quad \forall i, j \in S$$

In fact, every finite-state MC is positive recurrent.

Definition 4.5 (Reversible)**.** An irreducible MC is reversible if there exists a probability vector π satisfying $\pi_j P_{ji} = \pi_i P_{ij} \quad \forall i, j \in S$. These are call the detailed balance equations.

If a MC is reversible, then π is a stationary distribution. (also unique by Big Theorem)

4.4 First Step Equations (FSE)

Consider $A \subset S$, and define hitting time as $T_A = \min\{n \geq 0 : X_n \in A\}$. This is hard to do so we will instead try to compute $t_i = \mathbb{E}[T_A | X_0 = i]$. We can compute this by formulating first step equations:

- For $i \notin A$, $t_i = 1 + \sum_{j\in S} p_{ij} t_j$
- For $i \in A$, $t_i = 0$

5 Poisson Processes

5.1 Construction

A Poisson Process is an example of a counting process. A counting process $(N_t)_{t\geq 0}$ is a non-decreasing continuous-time integer-valued random process, which has right continuous sample paths.

Definition 5.1 (Poisson Process)**.** A rate-λ Poisson Process (i.e. PP(λ)) is a counting process with i.i.d inter-arrival times $S_i \overset{\text{IID}}{\sim} \text{Exp}(\lambda)$. Equivalently, a counting process is PP(λ) iff $N_0 = 0$, $N_t - N_s \sim \text{Poisson}(\lambda(t-s))$ for $0 \leq s \leq t$, and $(N_t)_{t\geq 0}$ has independent increments.

To elaborate on this, we will define T_i to be the arrival times, so $T_i = \min\{t \geq 0 : N_t \geq i\}$, which is the time of ith arrival. We also define the inter-arrival time, $S_i = T_i - T_{i-1}$, for $i \geq 1$.

Theorem 21. *If $(N_t)_{t\geq 0}$ is a PP(λ), then for $t \geq 0$, $N_t \sim$ Poisson(λt). I.e. $Pr\{N_t = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$*

Proof.

$$\begin{aligned}
Pr\{N_t = n\} &= Pr\{T_n \leq t < T_{n+1}\} \\
&= \mathbb{E}[\mathbb{1}_{\{T_n \leq t\}}\mathbb{1}_{\{t \leq T_n + S_{n+1}\}}] \\
&= \int f_{T_n}(s)\mathbb{1}_{\{s\leq t\}}\mathbb{E}[\mathbb{1}_{\{t\leq s+S_{n+1}\}}]ds \\
&= \int_0^t f_{T_n}(s)\mathbb{E}[\mathbb{1}_{\{t-s\leq S_{n+1}\}}]ds \\
&= \int_0^t f_{T_n}(s)e^{-\lambda(t-s)}ds \\
&= \int_0^t \frac{\lambda e^{-\lambda s}(\lambda s)^{n-1}}{(n-1)!}e^{-\lambda(t-s)}ds \ (f_{T_n}(s) \text{ is Erlang}) \\
&= \frac{\lambda^n e^{-\lambda t}}{(n-1)!}\int_0^t s^{n-1}ds \\
&= \frac{(\lambda t)^n e^{-\lambda t}}{n!}
\end{aligned}$$

□

Remark. By the memoryless property of Exp(λ), if $(N_t)_{t\geq 0} \sim$ PP(λ), then $(N_{t+s} - N_s)_{t\geq 0} \sim$ PP(λ) for all $s \geq 0$. Moreover, $(N_{t+s} - N_s)_{t\geq 0}$ is independent of $(N_\tau)_{0\leq\tau\leq s}$. In particular, Poisson Processes have independent and stationary increments. If $t_0 < ... < t_k$, then $(N_{t_1} - N_{t_0}), ..., (N_{t_k} - N_{t_{k-1}})$ are independent and $(N_{t_i} - N_{t_{i-1}}) \sim \text{Poisson}(\lambda(t_i - t_{i-1}))$ for all i.

5.2 Conditional Distribution of Arrivals

Theorem 22. *Conditioned on $\{N_t = n\}$, $(T_1, ..., T_n) \overset{d}{=} (U_{(0)}, ..., U_{(n)})$ where $U_{(i)}$ are the order statistics of n Uniform(0,t) random variables.*

In other words, given n arrivals occurred up to time t, the arrival times look like i.i.d Unif(0, t) random variables in distribution.

Proof. Let $0 = t_0 \leq t_1 \leq ... \leq t_n \leq t$, then

$$\begin{aligned} f_{T_1 T_2 ... T_n | N_t}(t_1 ... t_n | n) &= \frac{Pr\{N_t = n | T_1 = t_1, ..., T_n = t_n\}}{Pr\{N_t = n\}} f_{T_1 ... T_n}(t_1 ... t_n) \\ &= \frac{Pr\{N_t - N_{t_n} = 0\}}{Pr\{N_t = n\}} \prod_{i=1}^{n} f_{S_i}(t_i - t_{i-1}) \\ &= \frac{e^{-\lambda(t - t_n)}}{e^{-\lambda t} \frac{(\lambda t)^n}{n!}} \prod_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})} \\ &= \frac{n!}{t^n} \text{ (density of uniform random order statistics)} \end{aligned}$$

□

5.3 Merging

Theorem 23. *If $(N_{1,t}) \sim PP(\lambda_1)$ and $(N_{2,t}) \sim PP(\lambda_2)$ are independent, then $(N_{1,t} + N_{2,t}) \sim PP(\lambda_1 + \lambda_2)$.*

Proof. We will show that the sum of the two independent Poisson Processes satisfies the three properties of a PP:

1. $N_{1,0} + N_{2,0} = 0 + 0 = 0$
2. For $0 \leq s \leq t$,

$$\begin{aligned} (N_{1,t} + N_{2,t}) - (N_{1,s} + N_{2,s}) &= (N_{1,t} - N_{1,s}) + (N_{2,t} - N_{2,s}) \\ &\overset{d}{=} \text{Poisson}(\lambda_1(t - s)) * \text{Poisson}(\lambda_2(t - s)) \\ &= \text{Poisson}((\lambda_1 + \lambda_2)(t - s)) \end{aligned}$$

3. $(N_{1,t} + N_{2,t})_{t \geq 0}$ has independent increments since $(N_{1,t})_{t \geq 0}$, $(N_{2,t})_{t \geq 0}$ has independent increments.

□

5.4 Splitting (a.k.a Thinning)

Theorem 24. *Let $p_1, ..., p_k$ be non-negative such that $\sum_{i=1}^{k} p_i = 1$ and $(N_t)_{t \geq 0}$ be a PP(λ). Mark each arrival with label "i" with probability p_i, independently of all other arrivals so that $(N_{i,t})_{t \geq 0}$ be the process that counts arrivals marked with "i". Then $(N_{i,t})_{t \geq 0}$, for $i = 1, ..., k$, are independent Poisson Processes with respective rates $p_i \lambda$ for $i = 1, ..., k$.*

Proof. We will only prove for $k = 2$. This is sufficient because we can simply do induction to get $k > 2$. For $k = 2$, we let $p_1 = p$ and $p_2 = 1 - p$.

$$\begin{aligned} Pr\{N_{1,t} = n, N_{2,t} = m\} &= Pr\{N_{1,t} = n, N_{2,t} = m, N_t = n + m\} \\ &= Pr\{_{1,t}= n, N_{2,t} = m | N_t = n + m\} Pr\{N_t = n + m\} \\ &= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!} \\ &= e^{(-p\lambda)t} \frac{(p\lambda t)^n}{n!} e^{(-(1-p)\lambda)t} \frac{((1-p)\lambda t)^m}{m!} \\ &= \text{Poisson}(p\lambda t)\text{Poisson}((1-p)\lambda t) \end{aligned}$$

□

5.5 Random Incidence Paradox

Consider $(N_t)_{t \geq 0} \sim PP(\lambda)$ and pick a random time t_0. What is the expected length of the inter-arrival interval in which t_0 falls?

Say it falls between T_i and T_{i+1}. Then the length of the inter-arrival interval is $L = (t_0 - T_i) + (T_{i+1} - t_0)$. We know that $T_{i+1} - t_0 \sim \text{Exp}(\lambda)$ by the memoryless property of the exponential distribution. We also know that

$$Pr(t_0 - T_i > s) = Pr(\text{no arrivals in } (t_0 - s,\, s)) = Pr(N_{t_0} - N_{t_0 - s} = 0) = e^{-\lambda s}$$

so $t_0 - T_i \sim \text{Exp}(\lambda)$. By linearity of expectation, $\mathbb{E}[L] = \frac{2}{\lambda}$. If we arrive at a random time, we are more likely to land in a long interval.

6 Continuous Time Markov Chains (CTMCs)

6.1 Construction

Intuitively, a CTMC is a markov chain where we need to wait for $Exp(\lambda)$ time before transitioning to the next state.

Definition 6.1 (CTMC)**.** Let S be a countable state space. A CTMC is defined in terms of a rate matrix Q satisfying $[Q]_{ij} \geq 0$ for $i \neq j$, $i, j \in S$ and $\sum_{j \in S} [Q]_{ij} = 0$ for all $i \in S$. Specifically, the transition rate for state i is $q_i := [Q]_{ii} = -\sum_{j \neq i} [Q]_{ij}$. We also have $[Q]_{ij} = q_i p_{ij}$ such that $\sum_{j \in S} p_{ij} = 1$ where $p_{ii} = 0$ and $p_{ij} \geq 0$. p_{ij} are the transition probabilities for an associated DTMC called the jump chain.

A CTMC with rate matrix Q works as followed:

1. Start with $X_0 = i$.

2. Hold for $\text{Exp}(q_i)$ amount of time, then jump to state $j \in S$ with probability p_{ij} where $j \in S$.

3. Repeatedly apply the previous line above at next states (starting at state j).

We can equivalently define CTMCs by their jump rates q_{ij}. On entering state i, consider independent random variables $T_j \sim \text{Exp}(q_{ij})$ for $j \in S \backslash \{i\}$ and jump to state $j^* = \text{argmin}_{j \in S}(T_j : j \in S)$ at time T_{j^*}. This valid due to the splitting property of Poisson Processes.

Remark. This is called a markov chain by the memoryless property of the exponential distribution:

$$Pr(X_{t+\tau} = j | X_t = i, X_s = i_s, 0 \leq s < t) = Pr(X_{t+\tau} | X_t = i)$$

6.2 Stationary Distributions

Definition 6.2 (CTMC Stationarity)**.** A probability vector π is (without considering pathological cases) a stationary distribution for a CTMC with rate matrix Q if $\pi Q = 0$. This called the rate conservation principle. This is equivalent to $\pi_j q_j = \sum_{i \in S} \pi_i q_{ij}$ for all $j \in S$. In other words, assuming that $Pr(X_t = i) = \pi_i$, the rate at which transitions are made out of j is equal to the rate at which transitions are made into j.

6.3 Classification of States

Similar to DTMCs, we can classify the states.

- We say i and j communicate (i.e. $i \leftrightarrow j$) iff $i \leftrightarrow j$ is a jump chain iff we can travel $i \rightarrow j$ and back.

- Classes in CTMC are same as those in associated jump chain.
- State j is transient if, given $X_0 = j$, $(X_t)_{t \geq 0}$ re-enters state j finitely many times with probability one. State j is recurrent otherwise.
- For a recurrent state j, define $T_j = \min\{t \geq 0 : X_t = j$ and $X_s \neq j$ for some s $<$ t $\}$.
- State j is positive recurrent if $\mathbb{E}[T_j|X_{0\text{J}}] = \infty$.
- State j is null recurrent if $\mathbb{E}[T_j|X_0 = j] = \infty$.
- Transience/Positive Recurrence/Null Recurrence are class properties
- There is no concept of periodicity.

6.4 Big Theorem

Theorem 25. *We define $P^t_{ij} := Pr(X_t = j|X_0 = i)$ and $m_j := \mathbb{E}[T_j|X_0 = j]$. For an irreducible CTMC, exactly one of the following is true:*

1. *Either all states are transient or all states are null recurrent. In this case, no stationary distribution exits, and $\lim_{t\to\infty} P^t_{ij} = 0$ for all $i, j \in S$.*
2. *All states are positive recurrent. In this case, a unique stationary distribution exits and satisfies $\pi_j = \frac{1}{m_j q_j} = \lim_{t\to\infty} P^t_{ij}$ for all $i, j \in S$.*

Remark. Stationary distribution in CTMC is not the same as the stationary distribution in the jump chain. Generally speaking, $\widetilde{\pi_j} = \frac{\pi_j q_j}{\sum_{i\in S} q_i \pi_i}$ given that $\sum_{i\in S} q_i \pi_i < \infty$ where $\widetilde{\pi_j}$ is the stationary distribution of the jump chain.

6.5 Examples

6.5.1 M/M/s queue

Customers arrive to a system with s servers according to PP(λ). If a server is available, the arrival is immediately serviced, which takes $\overset{IID}{\sim}$ Exp(μ). If no server is available, the arrival waits until one becomes available. Let $(X_t)_{t\geq 0}$ denote the number of customers in system at time $t \geq 0$. We can model this with $q_{n,n+1} = \lambda$ and $q_{n,n-1} = \begin{cases} n\mu & 1 \leq n \leq s \\ s\mu & n > s \end{cases}$.

6.5.2 Birth Death Chain

Individuals give birth $\overset{\text{IID}}{\sim}$ $PP(\lambda)$ and have lifetimes $\overset{\text{IID}}{\sim}$ Exp(μ). Let X_t be the number of individuals in population at time t. We can model this with $q_{n,n+1} = n\lambda$ and $q_{n,n-1} = n\mu$. Note that this means $q_n = n(\lambda + \mu)$ so then $p_{n,n+1} = \frac{\lambda}{\lambda+\mu}$ and $p_{n,n-1} = \frac{\mu}{\lambda+\mu}$, so the DT jump chain is also a birth-death chain.

6.5.3 M/M/∞ queue

This is the same of the M/M/s queue problem except there are infinite servers. In this case $q_{n,n+1} = \lambda$ and $q_{n,n-1} = n\mu$. If we solve $\pi Q = 0$, we see that $\pi_n = \frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}$. By the Big Theorem, $X_t \xrightarrow{d}$ Poisson(λ/μ) where X_t is the number of people in the system at time t.

6.6 First Step Equations (FSE)

If $A \subseteq S$, define $T_A = \min_t\{t \geq 0 : X_t \in A\}$. We want to compute $\mathbb{E}[T_A|X_0 = i]$, so we will use FSEs to do this. Define $t_i := \mathbb{E}[T_A|X_0 = i]$ and $t_i = 0\ \forall i \in A$. Then we want to if $t_i = \mathbb{E}[\text{hold time}] + \sum p_{ij}t_j\ \forall i \in S$. Thus our FSEs are

$$t_i = 0 \quad \forall i \in A$$

$$t_i = \frac{1}{q_i} + \sum_{j \in S} p_{ij}t_j \quad \forall i \notin A$$

6.7 Uniformization

Uniformization is an approach to compute CTMC transition probabilities by simulating a DTMC.

6.7.1 Context

For context, consider a CTMC with transition rates $(q_i)_{i \in S}$ and assume $\exists M > 0$ s.t. $q_i \leq M \quad \forall i \in S$. We want to find P_t for some $t \geq 0$. Here,

$$\left[P_t\right]_{ij} := Pr(X_t = j|x_0 = i)$$

Markovity gives the Chapman-Kolmogorov Equations, which is $P^{s+t} = P^s P^t \quad \forall s, t \geq 0$. We can also show that for $h \approx 0$, $P^h \approx I + hQ + O(h)$ so

$$\begin{aligned} P^{t+h} &= P^t P^h \\ &\approx P^t(I + hQ + O(h)) \\ \frac{P^{t+h} - P^t}{h} &= P^t Q + \frac{O(h)}{h} \\ \frac{\partial}{\partial t} P^t &= P^t Q \\ P^t = e^{tQ} &:= \sum_{k \geq 0} \frac{(tQ)^k}{k!} \quad \forall t \geq 0 \end{aligned}$$

So we've found a way to compute P^t, but this becomes intractable for large state spaces. This is where uniformization comes in.

6.7.2 Construction

Let us define a uniformized DTMC which is a DTMC, given $\gamma \geq M$, with transition probabilities

$$p_{ij} = \frac{q_{ij}}{\gamma} \quad p_{ii} = 1 - \frac{q_i}{\gamma} \quad i, j \in S$$

If $P_u =$ transition matrix with uniformed DTMC, then $P_u = I + \frac{1}{\gamma}Q$. So observe $\pi P_u = \pi + \frac{1}{\gamma}\pi Q$. In other words, $\pi P_u = \pi \iff \pi Q = 0 \iff \pi$ is a stationary distribution for both the CTMC and uniformized DTMC. So we're beginning to see that the behavior of the two chains are similar. In fact, we can see that

$$P_u^n = (I + \frac{1}{\gamma}Q)^n \approx e^{\frac{n}{\gamma}Q}$$

So to estimate P^t, we can run the uniformized DTMC for $n \approx \gamma t$ steps because $P^t = e^{tQ} \approx e^{\frac{n}{\gamma}Q} \approx P_u^n$. Notice that with a larger γ, we get a better approximation.

7 Random Graphs

7.1 Definition

Definition 7.1 (Erdős–Rényi Random Graphs)**.** Fix $n \geq 1$ and $p \in [0, 1]$. A random graph G(n, p) is an undirected graph on n vertices, where each edge is placed independently with probability p.

7.2 Thresholds

The flavor of results around random graphs are threshold results (aka phase transitions).

7.2.1 Existence of Edges

An example is if $p >> \frac{1}{n^2}$ then graph has edges with high probability but if $p << \frac{1}{n^2}$ then graph doesn't have edges with high probability. To prove this, let $X = \#$ of edges in $G(n, p_n)$ and take $p_n = \frac{c}{n^2}$. Note that $X \sim$ Binomial$(\binom{n}{2}, p_n)$. So $P(X = 0) = (1 - p_n)^{\binom{n}{2}} \to e^{-c/2}$. So if $c >> 1$, $P(X = 0) \approx 0$ and if $c << 1$, $P(X = 0) \approx 1$.

7.2.2 Existence of Cycles

If $p >> \frac{1}{n}$, there exists a cycle whp. If $p << \frac{1}{n}$, there doesn't exist a cycle whp.

7.2.3 Largest Connected Components

If $p >> \frac{1}{n}$, the largest connected component is of size $\Theta(n)$. If $p << \frac{1}{n}$, the largest connected component is of size $O(\log n)$.

7.2.4 Connectivity

Lemma 1. $P(X = 0) \leq \frac{Var(X)}{(\mathbb{E}[X])^2}$

Proof.

$$\begin{aligned} Var(X) &= P(X = 0)\mathbb{E}[(X - \mathbb{E}(X))^2 | X = 0] + P(X \neq 0)\mathbb{E}[(X - \mathbb{E}(X))^2 | X \neq 0] \\ &\geq P(X = 0)(\mathbb{E}[X])^2 \end{aligned}$$

□

Theorem 26 (Erdős–Rényi)**.** *Fix $\lambda > 0$, and let $p_n = \lambda \frac{\log n}{n}$. If $\lambda > 1$, then P(G(n, p_n) is connected) $\to 1$. If $\lambda < 1$, then P(G(n, p_n) is connected) $\to 0$.*

Proof. When $\lambda < 1$, let X = number of isolated vertices. We want to show that $P(X = 0) \to 0$. Let I_i be the indicator that vertex i is isolated. Then

$$\mathbb{E}[X] = n\mathbb{E}[I_i] = n(1 - p)^{n-1} := nq$$

$$Var(X) = \sum Var(I_i) + \sum_{i \neq j} Cov(I_i, I_j) = nq(1-q) + n(n-1)\frac{pq^2}{1-p}$$

So $P(X=0) \leq \frac{nq(1-q)+n(n-1)\frac{pq^2}{1-p}}{n^2q^2} = \frac{1-q}{nq} + \frac{p}{1-p} \leq \frac{1}{nq} + \frac{p}{1-p} \to 0$ as p goes to 0 and n goes to infinity. $\frac{1}{nq}$ converges to 0 because $log(nq) = log(n) + (n-1)log(1-p) \approx log(n) - (n-1)p = log(n^{1-\lambda}) \to \infty$ since $\lambda < 1$.

When $\lambda > 1$,

$$\begin{aligned} P(\text{G(n, p) is disconnected}) &= P(\cup_{k=1}^{n/2}\{\text{exists set of k disconnected vertices}\}) \\ &\leq \sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)} \\ &\to 0 \text{ for } \lambda > 1 \end{aligned}$$

□

8 Statistical Inference

8.1 MAP/MLE

Let X be the state of nature that takes values in $\{0, ..., M-1\}$ (i.e. M hypotheses to consider), and Y be the observation. Then the model is represented by likelihoods $p_{Y|X}(y|x)$. We will be doing bayesian inference, so we are assuming that X is a random variable with a known distribution $P(X = i) = \pi_i$. We call π the prior.

8.1.1 Maximum A Posteriori (MAP)

If we observe $\{Y = y\}$, then the a posteriori probability of $\{X = x\}$ is given by:

$$P(X = x|Y = y) = \frac{p_{Y|X}(y|x)\pi(x)}{\sum_{\tilde{x}} p_{Y|X}(y|\tilde{x})\pi(\tilde{x})} \propto p_{Y|X}(y|x)\pi(x)$$

The idea is that our prior has been updated given observations. This motivates MAP.

$$\hat{X}_{MAP}(y) = \underset{x}{\operatorname{argmax}}\, p_{Y|X}(y|x)\pi(x) = \operatorname{argmax} p_{X|Y}(x|y)$$

MAP esteimate depends on likelihoods and prior. What if we don't have a prior though? Then let us assume that π is uniform over all x. This gives rise to maximum likelihood estimate (ML):

$$\hat{X}_{ML}(y) = \underset{x}{\operatorname{argmax}}\, p_{Y|X}(y|x)$$