Probability Spaces
(Ω, F, P) defines a random experiment.

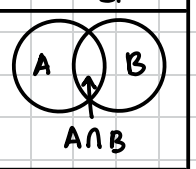
Sample Space (Ω) - the set of all possible outcomes of the experiment.
- Ex: Pick 1 card from deck
- {A spades, A clubs, K diamonds, ...}

Event Space (F) - events in your experiment, subsets of Ω, usually 2^Ω
- $\Omega \in F$, the whole sample space is an event
- Closed under complement (if "all red cards" is an event, so is "all non red cards".
- Closed under countable unions (if $A_1, A_2, \ldots \in F$, then $\cup_i A_i \in F$)
- Ex: all red cards, all black cards, $\emptyset$, F

Probability Function (P) - assigns a probability to each event.
- $P(\emptyset) = 0 \qquad P(E) \geq 0 \ \forall E \in F$ (Nonnegativity)
- $P(\Omega) = 1$
- Countable additivity (probs add up for disjoint events)

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \text{ for disjoint events } E_i$$

Bayes Rule
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$



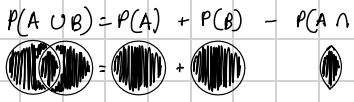
Total Probability
If $B_1, \ldots, B_n$ are disjoint & partition the sample space, $P(A) = \sum_i P(A \cap B_i) = \sum P(A|B_i)P(B_i)$

Independence
Knowing one event occured doesn't change the probability of another event.

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

Inclusion Exclusion
When finding $P(A \cup B)$, adding their individual probabilities double counts the shared outcome.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \ldots < i_k \leq n} |A_{i_1} \cap \ldots \cap A_{i_k}| \right) \text{ and}$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \ldots < i_k \leq n} P(A_{i_1} \cap \ldots \cap A_{i_k}) \right)$$

Pf: $P(A \cup B) = P(A \cup (B|A))$
$= P(A) + P(B|A)$
$= P(A) + P(B|A) + P(A \cap B) - P(A \cap B)$
$= P(A) + P(B|A) \cup (A \cap B)) - P(A \cap B)$
$= P(A) + P(B) - P(A \cap B)$

Union Bound - extension of inclusion exclusion
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

Random Variables - function from sample space (Ω) to a subset of the real numbers ($\mathbb{R}$)
Ex: 2 coin flips, # of heads - $\Omega = \{HH, HT, TH, TT\}$

Let X = # of heads: $X: \Omega \to \{0, 1, 2\}$ $\quad X(HH) = 2 \quad X(HT) = 1 \quad X(TH) = 1 \quad X(TT) = 0$

Discrete RVs have probability mass functions:

$$P_X(x) = P(\{X = x\}) = P(\{\omega \in \Omega \mid X(\omega) = x\}) = \sum_{\omega \in \Omega} 1\{X(\omega) = x\} \cdot P(\omega)$$

(Calculate by looping over every outcome ω in Ω, keep only the ones where $X(\omega) = x$, and add their probabilities.)

$1\{x\}$ is an indicator function which returns 1 if x is true, else 0.

Ex: $P_X(0) = P(\{TT\}) = 1/4$, $P_X(1) = P(\{HT, TH\}) = 1/2$, $P_X(2) = P(\{HH\}) = 1/4$

Expectation - Center of mass of a distribution.
$$E[X] = \sum_{a \in A} a \cdot P[X = a] = \sum_x x \cdot p_X(x)$$

Linearity of Expectation: $E[X+Y] = E[X] + E[Y]$
$E[cX] = cE[X], \ \forall c \in \mathbb{R}$

Law of the Unconscious Statistician (LOTUS):
$$E[f(X)] = \sum_x f(x) P(X = x) \quad \text{For example, } E[X^2] = \sum_x x^2 P[X = x]$$

Conditional Expectation: $E[X|Y=y] = \sum_x x \cdot \underbrace{P_{X|y}(x|y)}_{= \frac{P_{X,Y}(x,y)}{P_Y(y)}}$

Total Expectation Theorem:
$$E[X] = \sum_y = \sum_y p_y(y) \cdot E[X|Y=y]$$

Variance - Average squared distance of a random variable from its mean.
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if independent.

$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

$\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$

$\mathrm{Var}(X+b) = \mathrm{Var}(X)$

$\mathrm{Var}(aX_1 + bX_2) = a^2 \mathrm{var}(X_1) + 2ab \cdot \mathrm{cov}(X_1, X_2) + b^2 \mathrm{var}(X_2)$

Bernoulli Distribution
X_i is a Bernoulli random variable with parameter p. $X \in \{0, 1\}$. $P(X=1) = p$, $P(X=0) = 1-p$.

Let I_A be the Bernoulli variable for event A w/ parameter $P(A)$.

$E[I_A] = P(A) = p$

$\mathrm{Var}(I_A) = p(1-p)$

Binomial Distribution
Models the number of successes in n independent identical trials, where each trial succeeds with probability p. For example, flipping a coin multiple times. Converges to Poisson as $n \to \infty$

Let $X_i = \begin{cases} 1 & \text{if flip } i \text{ is H} \\ 0 & \text{otherwise} \end{cases}$

$X_i \sim \mathrm{Bernoulli}(p)$ $\quad X_i$ is Bernoulli RV with probability p

Let $X = \sum_{i=1}^{n} X_i$ $\quad P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$

$X \sim \mathrm{Bin}(n, p)$ $\quad X$ is Binomial RV with n trials and probability p.

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np \qquad \mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Geometric Distribution
Models how many trials are needed until the first success. Let X = # of flips until first H.

$X \sim \mathrm{Geometric}(p)$ $\quad X$ is a geometric RV with probability p.

$P(X=x) = (1-p)^{x-1} p \qquad E[X] = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$

Memoryless property: If m trials have already failed, the remaining waiting time is distributed the same as if you are starting over. Essentially, the coin doesn't remember its past flips.

$P(X > m+n \mid X > m) = P(X > n) \qquad P(X > x) = (1-p)^x$

Poisson Distribution
Models how many times an event happens in a fixed interval.

λ = average number of events per interval

$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad E[X] = \lambda \qquad \mathrm{Var}(X) = \lambda$

Uniform Distribution
A random variable is uniform on an interval if every value of the RV within the interval is equally likely

$X \sim \mathrm{Uniform}(a, b)$, where a, b is the interval

$P(X=x) = \frac{1}{b-a+1} \qquad E[X] = \frac{a+b}{2} \qquad \mathrm{Var}(X) = \frac{(b-a+1)^2 - 1}{12}$

Indicator Tricks
Useful to decompose RVs into sums of Bernoulli RVs

Ex: Let X = # of H in n coin flips

Let RV $= I_i = \begin{cases} 1 & \text{flip } i \text{ is heads} \\ 0 & \text{otherwise} \end{cases}$, $X = \sum_{i=1}^{n} I_i$

Then, $E[X] = E[\sum_{i=1}^{n} I_i] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} 1/2 = \frac{n}{2}$

$\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^{n} I_i) = \sum_{i=1}^{n} \mathrm{Var}(I_i) = \sum_{i=1}^{n} \frac{1}{4} = \frac{n}{4}$

Calculus
Fundamental Theorem: $\frac{d}{dx} \int_{a(x)}^{b(x)} f(t)\,dt = f(b(x))b'(x) - f(a(x))a'(x)$

$\int_a^b f(t)\,dt = F(b) - F(a)$, where F is antiderivative of f

Chain Rule: $\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$

Discrete vs. Continuous

	Discrete	Continuous
Expectation	$E[X] = \sum_{a \in A} a \cdot P[X=a]$	$E[X] = \int_{-\infty}^{\infty} a \cdot f_X(a)\,da$
LOTUS	$E[g(X)] = \sum_x g(x) P(X=x)$	$E[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t)\,dt$
Total Prob	$P[B] = \sum_{i=1}^{n} P[B \mid A_i] P[A_i]$	$P[B] = \int_{-\infty}^{\infty} P[B \mid A_i] f_X(t)\,dt$
Variance	$\mathrm{Var}(X) = E[X^2] - (E[X])^2$	$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\int_{-\infty}^{\infty} x f(x)\,dx\right)^2$

Law of Total Variance: $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$

Continuous Random Variable
Takes uncountably many values (any real number in an interval)

$P(X=a) = 0$ because a single point has no width. Instead of $P(X=a)$, $P(a \leq X \leq b)$. What is the probability that a RV lies in some interval?

Probability Density Function
The PDF is the actual curve of the distribution. Probability is the area under the curve.

Properties:
1. Non-negativity: $f_X(x) \geq 0$ for all x.
2. Interval Probability: $P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$
3. Normalization: $\int_{-\infty}^{\infty} f_X(t)\,dt = 1$ Total probability sums to 1

$$E = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_0^{\infty} P(X \geq x)\,dx$$

Cumulative Distribution Function
$F_X(a) = P(X \leq a)$. Gives the probability that X is to the left of a.

$P(a \leq X \leq b) = P(a < X < b)$

$F_X(a) = \int_{-\infty}^{a} f_X(x)\,dx \qquad \frac{d}{dx} F_X(x) = f_X(x)$

Properties:
1. Non decreasing: $F_X(x)$ never goes down, only add probability as you go right.
2. Limits: $\lim_{x \to -\infty} F_X(x) = 0 \qquad \lim_{x \to \infty} F_X(x) = 1$

$-\infty$ and ∞ refer to the bounds of the domain of X. So if X is only over c to d, use c and d instead of $-\infty$ and ∞.

Uniform Distribution
A random variable is uniform on an interval if every value of the RV within the interval is equally likely

$X \sim \mathrm{Uniform}(a, b)$, where a, b is the interval

$$\mathrm{PDF}: f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \qquad \mathrm{CDF}: F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

$P(c \leq x \leq d) = \frac{d-c}{b-a} = \frac{\text{interval of probability}}{\text{interval of distribution}}$

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

Exponential Distributions
Models waiting time until an event happens

$X \sim \mathrm{Exp}(\lambda) \to X$ = time until first success, λ = rate at which successes happen

$$\mathrm{PDF}: f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \mathrm{CDF}: F_X(t) = P(X \leq t) = 1 - e^{-\lambda t}$$

Probability that X happens by t

$P(a \leq x \leq b) = \int_a^b \lambda e^{-\lambda x}\,dx$

Survival function: Probability that object survives past t

$S_X(t) = P(X > t) = 1 - F_X(t) = e^{-\lambda t}$

$E[X] = \frac{1}{\lambda} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$

Memoryless property: If m time has already past, the additional waiting time does not depend on the past.

$P(X > m+n \mid X > m) = P(X > n)$

pf: $P(X > a+b \mid X > a) = \frac{P(X > a+b, X > a)}{P(X > a)} = \frac{P(X > a+b)}{P(X > a)} = \frac{1 - F_X(a+b)}{1 - F_X(a)}$
$= \frac{e^{-\lambda(b+a)}}{e^{-\lambda a}} = e^{-\lambda b} = P(X > b)$

Normal Distribution
$X \sim N(\mu, \sigma^2)$ means X is normally distributed with mean μ and variance σ^2

$\mathrm{PDF}: f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$

$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$

$E[X] = \mu \qquad \mathrm{Var}(X) = \sigma^2$

$P(W \leq x) = \phi\left(\frac{x-\mu}{\sigma}\right)$

$\mathrm{CDF}: F_X(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}}\,dt$

$X \sim N(\mu, \sigma^2)$

$Z = \frac{X-\mu}{\sigma}$

$Z \sim N(0, 1)$

$P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$

Sum of Independent Normals:
$X \sim N(0, 1)$ and $Y \sim N(0, 1)$
$Z = aX + bY \sim N(0, a^2 + b^2)$

Joint Densities
$P((x,y) \in A) = \iint_A f_{xy}(x,y)\,dx\,dy$

Normalization condition: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1$

Marginal Density: $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$

Law of Total Expectation: $E[X] = E[E[X|Y]]$

Ex: $Z \sim N(\mu, \sigma^2)$ & $A \sim \mathrm{Unif}(0, Z^2)$. What is $E[A]$?

$E[Z^2] = \mathrm{Var}(Z) + (E[Z])^2 = \sigma^2 + \mu^2$

$E[A] = E[E[A|Z]] = E\left[\frac{0+Z^2}{2}\right] = E\left[\frac{Z^2}{2}\right] = \frac{1}{2}E[Z^2] = \frac{1}{2}(\sigma^2 + \mu^2)$

pf: $E[E[X|Y]] = \sum_i E[X|Y=y_i] P(Y=y_i) = \sum_i \left(\sum_j x_j P(X=x_j|Y=y_i)\right) P(Y=y_i)$
$= \sum_j x_j \sum_i P(X=x_j|Y=y_i) P(Y=y_i) = \sum_j x_j P(X=x_j) = E[X]$

Poisson Distribution

Models how many times an event happens in a fixed interval.

λ = average number of events per interval

$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\quad E[X]=\lambda$ $\quad Var(X)=\lambda$

Convolutions

Let $Z = X+Y$ for independent X & Y

Discrete case: $P(Z=z) = \sum_x P(Z=z \mid X=x)P(X=x)$

$= \sum_x P(X+Y=z \mid X=x)P(X=x)$

$= \sum_x P(Y=z-x \mid X=x)P(X=x)$

Convolution Formula: $P(Z=z) = \sum_x P(Y=z-x)P(X=x)$

Continuous case: $f_z(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z-x)\,dx$

Covariance

: Measures the linear association between two random variables

$Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y] = P(X \wedge Y) - P(X)P(Y)$

If X and Y are independent, the $Cov(X,Y)=0$ b/c $E[XY]=E[X]E[Y]$

① $Cov(X,Y) = Cov(Y,X)$ ② $Cov(aX+bY, cZ) = ac\,Cov(X,Z) + bc\,Cov(Y,Z)$

③ $Cov(X,X) = Var(X)$ ④ $Cov(X,X+c) = Var(X)$

Moment Generating Functions (MGF)

$M_X(t) = E[e^{tX}]$

MGF - function of t that encodes all the information about a distribution in one object.

Use 1: Extract Moments

$E[X^k] = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}$: The kth moment of X is the kth derivative of the MGF evaluated at $t=0$.

Use 2: Compute Sums

For independent X and Y: $M_{X+Y}(s) = E[e^{s(X+Y)}] = E[e^{sX}]E[e^{sY}] = M_X(s)\cdot M_Y(s)$

Use 3: Affine Transforms

For $aX+b$: $M_{aX+b}(s) = e^{sb}\cdot M_X(as)$

pf $E[e^{s(aX+b)}] = E[e^{sas\cdot X}\cdot e^{sb}] = e^{sb}E[e^{asX}] = e^{sb}\cdot M_X(as)$

RV	MGF
Bernoulli	$1-p+pe^t$
Binomial	$(1-p+pe^t)^n$
Geometric	$\frac{p}{1-(1-p)e^t}$

RV	MGF
Normal	$e^{t\mu+\sigma^2t^2/2}$
Exponential	$\frac{1}{1-\frac{t}{\lambda}}$
Poisson	$e^{\lambda(e^t-1)}$

Bounds

Markov: For a nonnegative RV: $P(X \geq a) \leq \frac{E[X]}{a}$

Chebyshev: For $b>0$: $P(|X-E[X]| \geq b) \leq \frac{Var(X)}{b^2}$

Chernoff: $P(X \geq a) = P(e^{tX} \geq e^{ta}) = \frac{E[e^{tX}]}{e^{ta}} = \frac{M_X(t)}{e^{ta}}$

Convergence in Probability

For every $\epsilon>0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$. The probability that X_n differs from X by more than ϵ shrinks to 0 as $n\to\infty$.

Weak Law of Large Numbers

- Sample mean converges to expected value.

$X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ and a finite variance.

$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right) \to 0$ as $n\to\infty$

$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$ As $n\to\infty$, the probability that the mean of $X_n - \mu$ is $\geq \epsilon$ is 0.

Strong Law of Large Numbers

$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$. The probability of the average mean of all X_n's equalling μ as $n\to\infty$ is 1.

Central Limit Theorem

When n is large, the distribution of the sample mean or sample sum of identical distribution RVs looks Normal, regardless of the original distribution.

$X_1, X_2, X_3, \ldots, X_n$: Independent and identically distributed (i.i.d.) RVs

Each X_i has mean, μ, and standard deviation, σ.

Sample Sum: $S_n = \sum_{i=1}^n X_i$ $\quad E[S_n] = n\mu$ $\quad SD(S_n) = \sqrt{n}\sigma$

$S_n \approx N(\underbrace{n\mu}_{\text{mean}}, \underbrace{n\sigma^2}_{\text{variance}})$ for large n

Sample Mean: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ $\quad E[\bar{X}_n] = \mu$ $\quad SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$

$\bar{X}_n \approx N(\underbrace{\mu}_{\text{mean}}, \underbrace{\frac{\sigma^2}{n}}_{\text{variance}})$ for large n

$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$

You can define $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$, then $E(Z_n)=0$, $Var(Z_n)=1$.

$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$ $\forall x$ when $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt$

Entropy

$H(X) = E[-\log_2 p(X)] = E[\log_2 \frac{1}{p(X)}] = \sum_{x\in\mathcal{X}} p(x)\cdot\log_2\frac{1}{p(x)}$ bits

$\mathcal{X}$: alphabet of X

- average amount of uncertainty/surprise in X.
- $\log\frac{1}{p(x)}$ measures how surprising a single outcome x is.
- If high $p(x)$, low $\log\frac{1}{p(x)}$ → high prob, not surprising
- If low $p(x)$, high $\log\frac{1}{p(x)}$ → low prob, surprising
- High entropy → more uncertainty → learning the outcome tells you more info.

Asymptotic Equipartition Property (AEP)

If $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim p(x)$: $\frac{-\log_2 p(X_1, X_2, \ldots, X_n)}{n} \xrightarrow{p} H(X)$

↓ converges in probability as $n\to\infty$

pf Since X_i are independent:

$-\log_2 p(X_1,\ldots,X_n) = -\sum_{i=1}^n \log_2 p(X_i)$

Divide by n $= \frac{1}{n}\sum_{i=1}^n -\log_2 p(X_i)$, which is the sample mean of $-\log_2 p(X_i)$

By WLLN, sample means converges to expected value so

$\frac{1}{n}\sum_{i=1}^n -\log_2 p(X_i) \xrightarrow{p} E[-\log_2 p(X)] = H(X)$

Typical Sequences

A sequence x^n is typical if $\left|-\frac{1}{n}\log_2 p(x^n) - H(X)\right| \leq \epsilon$ OR $2^{-n(H+\epsilon)} \leq p(x^n) \leq 2^{-n(H-\epsilon)}$

About 2^{nH} typical sequences, each w/ probability about 2^{-nH}.

Mutual Information

- average amount of info X provides about Y and vice versa (symmetric)

$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$

↑ how much uncertainty Y removed about X; ↑ uncertainty of X before seeing Y; ↑ uncertainty of X after seeing Y

If X, Y independent, $I(X;Y)=0$.

Joint Entropy

- How surprising is X given Y?

$H(X,Y) = -\sum_{x,y} p(x,y)\cdot\log(p(x,y))$

$= -\sum_{x,y} p(x,y)\cdot\log(p(y)\cdot p(x|y))$

$= -\sum_{x,y} p(x,y)\cdot\log(p(y)) - \sum_{x,y} p(x,y)\cdot\log(p(x|y))$

$= -\sum_y p(y)\cdot\log(p(y)) - \sum_y p(y)\sum_x p(x|y)\cdot\log p(x|y)$

$H(X,Y) = H(Y) + H(X|Y)$

Channels

Sender chooses an input X with distribution $p(x)$. Sends it thru a noisy channel and the receiver receives Y. (channel behaves according to $p(y|x)$, probability of receiving y given you sent x. Sender controls $p(x)$ but not the channel $p(y|x)$.

Types of channels:

Noiseless: whatever you send arrives perfectly, no distortion $p(y|x)=1$

Noisy: signal is corrupted

BSC - bits flip w/ probability p, $C = 1-H(p)$

BEC - bits are erased w/ probability p, $C = 1-p$

Gaussian - random noise $Y = X+Z$, where $Z\sim N(0,\sigma^2)$

Fundamental Question: How much info can be reliably sent through this channel? Same as how of X can the receiver Y retain? This is mutual information $I(X;Y)$. Maximize this over all possible input distributions $p(x)$.

$C = \max_{p(x)} I(X;Y)$, C is the channel capacity, the max rate at which info can be transmitted reliably, measured in bits per channel use. Reliable iff $m/n \leq C$. Max $m = nC$

KL Divergence

$D(p\|q) = \sum_x p(x)\cdot\log\frac{p(x)}{q(x)}$. This measures how different two distributions p and q are. $D(p\|q)$ is always ≥ 0 and equals 0 only when $p=q$ everywhere.

Interpretation 1: How independent are 2 variables?

$D_{KL}(p(x,y)\|p(x)p(y))$ asks how far is the actual joint probability from the independent joint probability.

If X and Y are independent, $D_{KL}(p(x,y)\|p(x)p(y)) = 0 \to I(X;Y)=0$

If they are highly dependent, $D_{KL}(p(x,y)\|p(x)p(y))$ is large → $I(X;Y)$ is large

Interpretation 2: Cross Entropy

True distribution = P but you don't know it, so you use Q as your best guess for encoding.

If P was known, encoding = H(P) bits, the minimum. Since you are using Q instead, you pay a penalty of $D_{KL}(P\|Q)$.

So Cross entropy $= -\sum_x P(x)\log(Q(x)) = H(P) + D_{KL}(P\|Q)$

Discrete Time Markov Chains

States - represented by circles

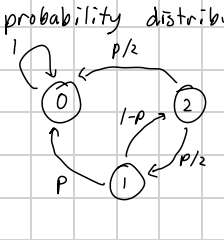
Transition Probabilities going out of any state should add to 1.

Ergodic Thm for finite, irreducible, aperiodic MC

$\Pi = \Pi P$, Π is a vector of the probability distribution of X_n. $\Pi = [\pi_0\ \pi_1\ \ldots\ \pi_n]$

$P =$ From state (→ To)

	0	1	2
0	1	0	0
1	p	0	1-p
2	p/2	p/2	0



All rows add to 0.

Represent final states with a self loop of p=1.

Solving:

① Write out balance equations and $\sum_{i\geq 0}\pi_i = 1$

π_i represents fraction of time spent at state i.

$[\pi_0\ \pi_1\ \pi_2] = [\pi_0\ \pi_1\ \pi_2]\begin{bmatrix}1 & 0 & 0\\ p & 0 & 1-p\\ p/2 & p/2 & 0\end{bmatrix}$

$\pi_0 = \pi_0 + \pi_1 p + \pi_2 \frac{p}{2}$

$\pi_1 = \pi_2 \frac{p}{2}$

$\pi_2 = \pi_1(1-p)$

$\pi_0 + \pi_1 + \pi_2 = 1$

Solve for all π_i in terms of p.

Chain always converges to π

Markov Chains Classifications

① Accessibility - State j is accessible from state i if there's a nonzero probability path from i to j in some number of steps.

② Communication - States i and j communicate if you can get from i to j and vice versa.

③ Irreducible - A chain is irreducible if all states communicate with each other.

④ Recurrent - $P(X_n = i \mid X_0 = i) = 1$. Given that state 0 is i, you are guaranteed to return back to i EVENTUALLY

Transient - $P(X_n = i \mid X_0 = i) < 1$. Given that state 0 is i, there is a chance you never return to i.

⑤ Aperiodic - GCD of all paths from a node i to itself is 1.

Period of State i: Find all the possible paths to go from state i back to state i. Record the step count for each. Take the gcd of all the step counts to get the period of State i.

IF A MARKOV CHAIN IS IRREDUCIBLE, ALL STATES HAVE THE SAME PERIOD!!!

Aperiodic

Period of all states is 1.

Fundamental Theorem of Markov Chains

If a Markov Chain is irreducible and aperiodic, then for any starting distribution (i.e. {0.5 0.5} means 50% chance start in state 0 and 50% in state 1), the probability of being in state i at step n converges to the same value π_i from the invariant distribution π.

Formally, if a Markov Chain is irreducible and aperiodic, then for any initial distribution π_0, we have that $\pi_n \to \pi$ as $n\to\infty$, and π is the unique invariant distribution for the Markov chain.

Markov Chain Applications

Let X = # of steps before reaching state A:

$a(i) = 0$ if $i \neq A \Rightarrow$ Already at A

$a(i) = 1 + \sum P(i,j)\,a(j) \Rightarrow$ Step from i to j plus future steps

Probability of Reaching A before B:

$a(i) = 1$ if $i = A \Rightarrow$ Already at A

$a(i) = 0$ if $i = B \Rightarrow$ Already at B

$a(i) = \sum P(i,j)\,a(j) \Rightarrow$ Land in state j with $p = P(i,j)$ and future steps $= a(j)$

Absorption

MC with states $S = \{0, 1, \ldots, r\}$

state 1 is recurrent.

Let $a(i) = P(\text{absorbed in } 1 \mid X_0 = i)$

Then $a(i)=1$, $a(j)=0$ for all other absorbing states

$a_\ell = \sum_k a_k P_{\ell k}$

Mean Return Time: Consider the mean return time to state $\ell \in S$.

$r_\ell = 1 + \sum_k t_k P_{\ell k}$ where t_k is the expected time to hit ℓ given $X_0 = k$.

Then $t_\ell = 0$, $t_k = 1 + \sum_j t_j P_{kj}$

$r_\ell = \frac{1}{\pi_\ell}$

In general, Let B_i be the quantity you care about at the ith state. Decide which subset of states exist such that $B=0$ at those states.

Then $B_k = a(k) + \sum B_\ell P_{k\ell}$. $a(k)$ is the immediate cost of getting to state k. Then for all states, ℓ, you can transition to from k, sum its $B_\ell \cdot P_{k\ell}$, where $P_{k\ell}$ is the probability of going from k to ℓ.

Markovian Reversibility

Running backwards is the same as forwards.

Check: The probability of being in state i and jumping to j equals the probability of being in j and jumping to i

Detailed Balance: $\pi_i P(i,j) = \pi_j P(j,i)$

To guarantee π is stationary distribution:

$\sum_i \pi(i) P(i,j) = \sum_i \pi(j) P(j,i) = \pi(j)\sum_i P(j,i) = \pi(j)$

Ex: 0 ⇄ (λ_i, μ_{i+1}) ⇄ (λ_{i+1}, μ_{i+2}) ⇄ ... ⇄ n

DB: $\pi_n\cdot\lambda_n = \pi_{n+1}\mu_{n+1}$ and $\sum\pi_i = 1 \Rightarrow \pi(n) = \pi_0 \prod_{k=0}^{n-1}\frac{\lambda_k}{\mu_{k+1}}$ FIND THE PATTERN!

$\pi(0)\lambda_0 = \pi(1)\mu_1 \Rightarrow \pi(1) = \pi(0)\frac{\lambda_0}{\mu_1} \Rightarrow \pi(2) = \pi_0\cdot\frac{\lambda_0}{\mu_1}\cdot\frac{\lambda_1}{\mu_2}$

Shannon's Random Coding Scheme

Randomly generate 2^m codewords of length n. To transmit message k, send codeword k through the channel. A random codeword matches on all non erased bits with $p = 2^{-nC}$.

Conditional Entropy

$$H(Z|X_1) = \sum_{x_1} P(X_1 = x_1)\, H(Z|X_1 = x_1)$$

$$H(Z|X_1) = H(Z, X_1) - H(X_1)$$

$\Phi(z) = P(Z \le z)$ shaded area to the left of z

sometimes $Q(z)$ but this is the Gaussian CDF.

Handshake lemma: The sum of degrees of a graph is $2E$.

Detailed Balance is used to verify a proposed stationary distribution π and check reversibility.

If a stationary distribution exists for a finite and irreducible MC, all states are positive recurrent.

Positive recurrent means recurrent and $E[T_i] < \infty$. So for any state i, you will return back to that state in finite time.

$\pi(i) = \frac{1}{E[T_i]}$ Null recurrent is recurrent and $E[T_i] = \infty$

$X_0, X_1, \ldots$ = Markov Chain with transition matrix:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad S = \{0, 1, 2\}$$

Show $Y_0, Y_1, \ldots$ is not a Markov chain.

$Y_i = \begin{cases} 1 & \text{if } X_i = 1 \\ 0 & \text{otherwise} \end{cases}$

Solution:

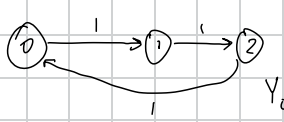
Markov Property: The past gives no extra info about the future beyond the present.

$P(Y_{i+1} = y \mid Y_i) = P(Y_{i+1} = y \mid Y_i, Y_{i-1})$

WTS: Markov Property fails:

Two scenarios where:

① Present is the same
② Past differs
③ Future differs



$X_0 = 0, X_1 = 1, X_2 = 2, X_3 = 0, \ldots$

$Y_0 = 0, Y_1 = 1, Y_2 = 0, Y_3 = 0, Y_4 = 1, \ldots$

		Past	future
let $y=1$	$P(Y_{i+1}=1 \mid Y_i = 0, Y_{i-1}=0) = 1$	$\{0, 0,$	$1\}$
	$P(Y_{i+1}=1 \mid Y_i = 0, Y_{i-1}=1) = 0$	$\{1, 0,$	$0\}$
		present	

The Markov property fails so $Y_0, Y_1, \ldots$ is not a MC.

BEC Channel with erasure probability $p = 0.4$. Use channel $n = 100{,}000$ times to transmit message of length m bits.

a) Max length of m?

Sol: For BEC, $C = 1 - p = 1 - 0.4 = 0.6$

$\max m = nC = 100000(0.6) = \boxed{60{,}000 \text{ bits}}$

b) Max value of m using Shannon's coding scheme if $P(\text{error}) \le 2^{-100}$

$P(\text{error}) = 2^m \cdot 2^{-nC} = 2^{m-nC} \le 2^{-100}$

(2^m: # codewords; 2^{-nC}: P(error for 1 codeword))

$m - nC \le -100$

$m \le nC - 100$

$m \le 100000(0.6) - 100$

$m \le 600000 - 100$

$\boxed{m \le 59900 \text{ bits}}$

$X_1 \sim \text{Bernoulli}(p_1)$, $X_2 \sim \text{Bernoulli}(p_2)$, independent coin flips. $Z = X_1 + X_2$. What is $H(Z|X_1)$?

$H(Z|X_1) = H(Z, X_1) - H(X_1)$

Since $Z = X_1 + X_2$, knowing Z and X_1 means you know X_2 and knowing X_1 and X_2 means you know Z, so $H(Z, X_1) = H(X_1, X_2)$

$H(Z|X_1) = H(X_1, X_2) - H(X_1)$

$H(Z|X_1) = H(X_1) + H(X_2) - H(X_1)$ since $X_1 \perp X_2$ are independent.

$H(Z|X_1) = H(X_2)$

$H(Z|X_1) = p_2 \cdot \log_2\left(\frac{1}{p_2}\right) + (1-p_2) \cdot \log\left(\frac{1}{1-p_2}\right)$

$\lambda > 0$, $n = 100$, $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Use CLT to find b where $P\left(\frac{1}{n}\sum_{i=1}^{n} X_i < b\right) = \frac{1}{10}$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. CLT says $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$

$E[X_i] = \lambda$ $\sigma^2 = \text{Var}(X_i) = \lambda$ so $\frac{\bar{X}_n - \lambda}{\sqrt{\lambda/n}} \xrightarrow{d} N(0,1)$

$P(\bar{X}_n < b) = \frac{1}{10} \rightarrow P\left(\frac{\bar{X}_n - \lambda}{\sqrt{\lambda/n}} < \frac{b-\lambda}{\sqrt{\lambda/n}}\right) = \frac{1}{10}$

Large n and $<$ so $\Phi\left(\frac{b-\lambda}{\sqrt{\lambda/n}}\right) = \frac{1}{10}$

$$\frac{b-\lambda}{\sqrt{\lambda/n}} = \Phi^{-1}\left(\frac{1}{10}\right)$$

$$b - \lambda = \Phi^{-1}\left(\frac{1}{10}\right) \cdot \frac{\sqrt{\lambda}}{\sqrt{n}}$$

$$b = \Phi^{-1}\left(\frac{1}{10}\right) \cdot \frac{\sqrt{\lambda}}{\sqrt{n}} + \lambda$$

$$b = \Phi^{-1}\left(\frac{1}{10}\right) \cdot \frac{\sqrt{\lambda}}{\sqrt{100}} + \lambda$$

$$b = \Phi^{-1}\left(\frac{1}{10}\right) \cdot \frac{\sqrt{\lambda}}{10} + \lambda$$

Connected graph with V vertices and E edges. Consider a random walk on this graph. From any vertex v, move to one of its adjacent vertices with $p = \frac{1}{d(v)}$ where $d(v)$ = degree of v.

a) Find $\pi(v)$

$\pi(v)$ = amt of time spent at v / total time

To get to v, you have to arrive from one of $d(v)$ edges. There are $2|E|$ total edges. So $\pi(v) = \frac{d(v)}{2|E|}$

Detailed Balance: $\pi_i P(i,j) = \pi_j P(j,i)$

$$\pi(i) \cdot \frac{1}{d(i)} = \pi(j) \cdot \frac{1}{d(j)}$$

$$\frac{d(i)}{2|E|} \cdot \frac{1}{d(i)} = \frac{d(j)}{2|E|} \cdot \frac{1}{d(j)}$$

$$\frac{1}{2|E|} = \frac{1}{2|E|} \checkmark$$

Detailed balance holds so $\pi(v) = \frac{d(v)}{2|E|}$.

b) Show that MC is reversible.

Detailed balance holds so the MC is reversible.

c) $P(i,j) = P(j,i)$ for all pairs of states, no self loops. Prove positive recurrent for all states.

Probably spend equal time at each vertex so $\pi(i) = \frac{1}{|V|}$.

Detailed balance: $\pi(i) = \frac{1}{|V|} = \pi(j) \checkmark$

Chain is finite so $E[T_i] = \frac{1}{\pi(i)} = |V| < \infty \checkmark$

Irreducible: The graph is connected and the condition $P(i,j) = P(j,i)$ means you can get from vertex i to vertex j for any $i, j \in V$. So all states communicate and the graph is irreducible. ✓

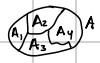
So the MC is finite and irreducible so it is positive recurrent.

Poisson process interarrival times are $\text{Exp}(\lambda)$.

(competing Poisson Processes → merged rate is $\lambda + \mu$ and routing probability is $\frac{\lambda}{\lambda+\mu}$ / $\frac{\mu}{\lambda+\mu}$.

Likelihood = routing of each observation times wait time for each observation.

Conditional Probability



Event Partioning: $A = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1, \ldots, A_n$ are mutually exclusive.

Total Probability Rule: $P[B] = \sum_{i=1}^{n} P[B \cap A_i] = \sum_{i=1}^{n} P[B|A_i]P[A_i]$

Since A_i covers all of the sample space, so you can sum the probabilities of B and A_i happening at the same time.

Independence: Events A and B are independent if knowing one happened doesn't change the probability of the other.

$P(A|B) = P(A)$ OR

$P(B|A) = P(B)$ OR

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

Pairwise Independence: Assume events A, B, C. These are pairwise independent if AB, BC, and AC are all independent

Mutual Independence

Assume events $A_1, A_2, \ldots, A_n$. They are mutually independent if every possible group of them is independent. That means every pair, triple, etc. are independent. Pairwise independence doesn't imply mutual independence.

$$P[\cap_{i \in I} A_i] = \prod_{i \in I} P(A_i)$$

The set of events A_i is mutually independent if the probability that they all happen at the same time is equal to the product of their individual probabilities.

Mutually Exclusive Events

$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$. If events are mutually exclusive, the probability of their union is the sum of their individual probabilities as there is no double counting.

Bayesian Inference

θ = unknown parameter, what you are trying to solve for

Say you have a coin, and don't know if it's fair or biased. Then, θ = fair or biased

Assume for θ = fair: $P(\text{heads}) = 0.5$

Assume for θ = biased: $P(\text{heads}) = 0.9$

① Prior - what you believed before seeing the data.

Ex: 80% chance coin is fair

$P(\theta = \text{fair}) = 0.8$, $P(\theta = \text{biased}) = 0.2$

② Likelihood - probability of the outcome observed given values of θ.

Outcome = heads: $P(\text{heads} \mid \theta = \text{fair}) = 0.5$

$P(\text{heads} \mid \theta = \text{biased}) = 0.9$

③ Posterior - what you believe after seeing the data

θ = fair: $0.8 \times 0.5 = 0.4$ Normalize → $\frac{0.4}{0.4+0.18}$, $\frac{0.18}{0.4+0.18}$

θ = biased: $0.2 \times 0.9 = 0.18$

Posterior: 0.69, 0.31

Posterior = (Prior × Likelihood) Normalized

$P_{\theta|X}(\theta|x) \propto P_\theta(\theta) \cdot P_{X|\theta}(x|\theta)$ AND $P_{\theta|X}(\theta|x) = \frac{P_\theta(\theta) \cdot P_{X|\theta}(x|\theta)}{P_X(x)}$

Unbiased estimator: manipulate $E[\hat{\theta}_{MLE}] = f(\theta)$ to isolate θ, then plug in known values to solve for θ.

German Tank MLE: $\hat{N}_{MLE} = \max(x_1, \ldots, x_n)$

- Unbiased estimator: $\hat{N}_{unbiased} = \frac{n+1}{n}\hat{N}_{MLE} - 1$

Maximum A Posteriori (MAP)

Pick the value of θ that makes the posterior as large as possible.

① Write out prior × likelihood in terms of θ.
② Take the natural log and simplify. Products become sums.
③ Take the derivative with respect to θ or partial derivs if multiple parameters
④ Set the above equal to 0 and solve for θ.

$\arg\max_\theta P_\theta(\theta) \cdot p(x|\theta)$

Maximum Likelihood Estimation (MLE)

Same steps as MAP but no priors.

① Write out likelihood in terms of θ.
② Take the natural log and simplify. (products become sums)
③ Take the derivative with respect to θ. Partial deriv if multiple parameters
④ Set the above equal to 0 and solve for θ.

$\arg\max_\theta p(x|\theta)$

Poisson RVs Tricks

$X \sim \text{Poisson}(\lambda)$

① $P(X \text{ is even})$

$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \rightarrow P(X \text{ even}) = \sum_m \frac{\lambda^{2m} e^{-\lambda}}{(2m)!} = e^{-\lambda} \cdot \sum_m \frac{\lambda^{2m}}{(2m)!}$

Taylor Series: $e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} \cdots$

$e^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} - \cdots$

So $e^{\lambda} + e^{-\lambda} = 2 \sum_m \frac{\lambda^{2m}}{(2m)!} \rightarrow \frac{e^{\lambda} + e^{-\lambda}}{2} = \sum_m \frac{\lambda^{2m}}{(2m)!}$

$e^{-\lambda} \cdot \sum_m \frac{\lambda^{2m}}{(2m)!} = e^{-\lambda} \cdot \left(\frac{e^{\lambda} + e^{-\lambda}}{2} \right)$

$= \frac{e^0 + e^{-2\lambda}}{2}$

$$\boxed{P(X \text{ even}) = \frac{1 + e^{-2\lambda}}{2}}$$

② $E[X^2]$:

$E[X(X-1)] = \lambda^2$

$E[X^2 - X] = \lambda^2$

$E[X^2] = E[X^2 - X] + E[X] = \lambda^2 + \lambda$

$$\boxed{E[X^2] = \lambda^2 + \lambda}$$

Hypothesis Testing

Observed data x comes from either model H_0 or H_1, we're trying to decide which one.

H_0 (null hypothesis): default assumption

H_1 (alternate hypothesis): what you're trying to detect

Design a test by picking an acceptance region A. If $x \in A$, you accept H_0, otherwise reject H_0 and go with H_1.

Errors:

① Type 1 error: H_0 is true, but you reject it.
- Probability of false alarm (PFA) = Significance level = α

② Type 2 error: H_1 is true, but you accept H_0
- Probability = β

Power: Probability you correctly detect H_1
- Probability of correct detection (PCD) = $1 - \beta$

Likelihood Ratio/Neyman Pearson

Making α smaller (being more conservative about rejecting H_0) tends to make β bigger (you miss detecting H_1 more).

Likelihood Ratio: $L(x) = \frac{P_{H_1}(x)}{P_{H_0}(x)}$ Use PMF/PDF depending on distribution

- measures how much more likely is x under H_1 than H_0. If $L(x)$ is large, x is strong evidence for H_1

Likelihood Ratio Test:
- If $L(x) > c$, reject H_0
- If $L(x) < c$, accept H_0
- If $L(x) = c$, reject H_0 with $p = \gamma$ and accept with $p = 1 - \gamma$

Pick c s.t. $PFA = z \Rightarrow P(L(x) > c \mid H_0) = z$, where z is the max PFA you're willing to tolerate $\rightarrow P_{H_0}(L(x) > c) + \gamma \cdot P_{H_0}(L(x) = c)$

Instead of solving $L(x) > c$, you can solve for $x > t$, where t is a simple threshold. Only works if $L(x)$ is always increasing in x.

$P(X > t \mid H_0) = z$

Solve for t using the CDF of X.

Whenever $L(x)$ is a product or exponential, you can take the log. Taking the log should pretty much always be the first step after writing out $L(x)$

Randomization

When X is discrete, $L(x)$ only takes a few specific values. For example, possible PFAs are 0.2 or 0.5, but you need PFA = 0.25 exactly.

Pick the threshold λ that gives you PFA just below z. Reject with probability γ.

$\gamma = \frac{z - P(L(x) > \lambda \mid H_0)}{P(L(x) = \lambda \mid H_0)}$ ← remaining gap (how much more PFA is needed) / ← what full rejection adds

NP Process

① Compute $L(x) = \frac{f(x \mid H_1)}{f(x \mid H_0)}$ for all outcomes x.

② Sort outcomes by $L(x)$, highest to lowest

③ Try each L as threshold λ. For each, compute PFA = sum of $P(x \mid H_0)$ for all outcomes x where $L(x) > \lambda$. If you have an equation for $L(x)$, find $P(x > c \mid H_0) = z$ and do inc/dec

④ Find λ s.t. PFA is just under z without exceeding it.

⑤ Find γ from formula above

⑥ Use Likelihood ratio test w/ λ as c to decide

LLSE/MMSE

You have 2 RVs, X and Y. You can observe X, but want to know Y. You can't observe Y directly, but X and Y are correlated, so knowing X gives you some info about Y. So given X, what is your best guess for Y. Best means minimizing squared error.

LLSE (Linear Least Squares Estimation)

Guesses for Y can only be linear functions of X.
- $Y = a + bX$, find the best a and b.
- error is orthogonal to linear functions of Y

LLSE of Y given X:

$$L[Y \mid X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X])$$

↑ baseline guess if you knew nothing about X

↖ how much you adjust guess per unit of X. Larger if X and Y are strongly correlated and 0 if they are uncorrelated

Squared Error of LLSE:

$E[(Y - L[Y \mid X])^2] = \text{var}(Y) - \frac{\text{cov}(X,Y)^2}{\text{var}(X)}$

If error is small, good guess, X was useful for the prediction and X and Y are more correlated.

If error is large, bad guess, X wasn't useful for the prediction and X and Y are less correlated.

MMSE (Minimum Mean Square Error)

Allow any function of X as your guess. Find the best one with no restrictions.

MMSE of Y given X: Error is orthogonal to all functions of Y.

$E[Y \mid X]$ → conditional expectation of Y given X

Jointly Gaussian Random Variables

$X_1, \ldots, X_n$ are jointly Gaussian if all linear combinations of them follows a normal distribution.
- $5X_1 - 3X_6$, $-7X_2$, $X_4 + 4X_5$, etc. must all be normal
- all weighted sums of the form $a_1X_1 + a_2X_2 + \ldots + a_nX_n$ must be normal

Jointly Gaussian RVs are independent iff they are uncorrelated (cov = 0)

For Jointly Gaussian RVs, MMSE = LLSE:

$$E[Y \mid X] = L[Y \mid X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X])$$

$L[X \mid Y, Z] = L[X, Y] + L[X, Z]$ only when Y, Z orthogonal and X, Y, Z are zero mean.

When there is multiple "given" RVs the formula for LLSE is:

$$L[Y \mid X_1, X_2] = E[Y] + \begin{bmatrix} \text{cov}(Y,X_1) & \text{cov}(Y,X_2) \end{bmatrix} \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1,X_2) \\ \text{cov}(X_1,X_2) & \text{var}(X_2) \end{bmatrix}^{-1} \begin{bmatrix} X_1 - E[X_1] \\ X_2 - E[X_2] \end{bmatrix}$$

DTMC Applications

Let α = # of steps before reaching state A:

$\alpha(i) = 0$ if $i = A \Rightarrow$ Already at A

$\alpha(i) = 1 + \sum P(i,j) \alpha(j) \Rightarrow$ Step from i to j plus future steps

Probability of Reaching A before B:

$\alpha(i) = 1$ if $i = A \Rightarrow$ Already at A

$\alpha(i) = 0$ if $i = B \Rightarrow$ Already at B

$\alpha(i) = \sum P(i,j) \alpha(j) \Rightarrow$ Land in state j with $p = P(i,j)$ and future steps = $\alpha(j)$

CTMC - States must be independent

In a DTMC, transitions happen at fixed time steps. In a CTMC, you only know the rate at which each transition occurs. Rates correspond to exponential RVs since they are the only continuous memoryless distribution

CTMC follow the Markov Property: The past gives no extra info about the future beyond the present.

$$P(Y_{i+1} = y \mid Y_i) = P(Y_{i+1} = y \mid Y_i, Y_{i-1})$$

How it works:

View 1:

You're in state i. For each neighboring state j, there's a separate clock ticking at Exponential(q_{ij}) time. All clocks run simultaneously, whichever goes off first is where you jump.

View 2 (Equivalent):

You are essentially leaving state i at the minimum of independent exponentials. This itself is exponential with the rate being the sum of all rates. Thus:

When do I leave? → $\text{Exp}(q_i)$, where $q_i = \sum_{j \neq i} q_{ij}$

Mean time in state $i = \frac{1}{q_i}$

Go to state j with $p = \frac{q_{ij}}{q_i}$

Rate Matrix

Rate Matrix Q:

Non-diagonal entries: $Q(i,j) = q_{ij} \geq 0$ (the rate $i \rightarrow j$)

Diagonal entries: $Q(i,i) = -q_i = -\sum_{i \neq j} q_{ij}$

Rows of Q sum to 0, this is a defining property.

Stationary Distribution (CTMC)

$\pi Q = 0$ and $\sum_i \pi_i = 1$

For each state i: $\underbrace{\pi_i \cdot q_i}_{\text{rate out of } i} = \underbrace{\sum_{j \neq i} \pi_j q_{ji}}_{\text{rate into } i}$

Detailed balance: $\pi(i) Q(i,j) = \pi(j) Q(j,i)$ for all i, j

CTMC Applications

Let α = expected time to reach state A:

$\alpha(i) = 0$ if $i = A \Rightarrow$ Already at A

$\alpha(i) = \frac{1}{q_i} + \sum \frac{q_{ij}}{q_i} \alpha(j) \Rightarrow$ time spent at i then jump to j with $p = \frac{q_{ij}}{q_i}$ and continue.

Probability of Reaching A before B:

$\alpha(i) = 1$ if $i = A \Rightarrow$ Already at A

$\alpha(i) = 0$ if $i = B \Rightarrow$ Already at B

$\alpha(i) = \sum_j \frac{q_{ij}}{q_i} \alpha(j) \Rightarrow$ Land in state j with $p = \frac{q_{ij}}{q_i}$ and future steps = $\alpha(j)$

Jump Chains

The jump chain is the embedded DTMC you get by ignoring holding times and only tracking which states you visit.

Transition probabilities: $P(i,j) = \frac{Q(i,j)}{q_i}$

Jump chain and CTMC don't have the same stationary distribution, π, in general, because the jump chain ignores how long you spend in each state.

Kolmogorov Equations (CTMC)

Transition Probability Matrix $P(t)$: Entry (i,j) is the probability of being in state j at time t given you started in state i.

$P(t) = e^{tQ}$

$P(s+t) = P(s) P(t)$ where s = amt of time already elapsed, and t = additional time after that

Kolmogorov Equations (DTMC)

$P^{n+m} = P^n P^m$

n = # of steps already taken

m = # of additional steps after that

k = n + m = total # of steps

P^k = transition matrix to the kth power where entry (i,j) is the probability of going from i to j in exactly k steps.

Reversibility

A CTMC is reversible if:

① It satisfies detailed balance for some π

② It doesn't have cycles

③ It has cycles, but the product of the rates in one direction is equal to the product of the rates of the other direction.

A DTMC is reversible if:

① It satisfies detailed balance for some π

② It doesn't have cycles

③ It has cycles, but the product of the transition probabilities in one direction is equal to the product of the transition probabilities in the other direction.

Poisson Processes

A Poisson process $PP(\lambda)$ is built from i.i.d. interarrival times $S_1, S_2, \ldots \sim \text{Exp}(\lambda)$

time of the nth arrival, $T_n = S_1 + S_2 + \ldots + S_n$

$N(t)$ = # of arrivals by time t

$N(t_1, t_2) = N(t_2) - N(t_1)$ = # of arrivals in interval $[t_1, t_2]$

Poisson Process Properties

① Stationary increments: The number of arrivals in any interval $[t, t+s]$ depends only on the length s, not where it starts.
- $N(t, t+s) \overset{d}{=} N(s)$

② Independent increments: Arrivals in non-overlapping intervals are independent of each other

③ Poisson Distribution: The # of arrivals by time t follows a Poisson distribution with rate λt.
- $N(t) \sim \text{Poisson}(\lambda t)$
- $P(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$

Erlang Distribution

Time of the nth arrival T_n = sum of n i.i.d. $\text{Exp}(\lambda)$ RVs
- $T_n \sim \text{Erlang}(n, \lambda)$
- PDF: $f_{T_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}$

Merging Poisson Processes

If $N \sim PP(\lambda)$ and $M \sim PP(\mu)$, $N+M \sim PP(\lambda+\mu)$

Splitting Poisson Processes

Let $N \sim PP(\lambda)$.
Each arrival goes to stream 1 with probability p and to stream 2 with probability $1-p$. Then:

$N_1 \sim PP(\lambda p)$ and $N_2 \sim PP(\lambda(1-p))$

N_1 and N_2 are independent.

Random Incidence Property

For a Poisson Process that has been running for a long time:

If you pick an arbitrary time t, the interarrival time containing t has length $\frac{2}{\lambda}$. This is because the time since the last arrival before t has expected value $\frac{1}{\lambda}$, and same for the next arrival after t. This is $\frac{2}{\lambda}$ in total.

Intuitively, t is more likely to land in longer intervals so the expected length of the interval t lands in is larger than the average interval length $\frac{1}{\lambda}$.

Random Graphs

$G(n, p)$ is an undirected graph with n vertices where each of the $\binom{n}{2}$ possible edges appears independently with probability p.

Facts:

① $E[\#\text{ of edges}] = \binom{n}{2} p = \frac{n(n-1)}{2} \cdot p$

② Degree of a vertex: $D \sim \text{Binomial}(n-1, p)$
- $E[D] = (n-1)p$

③ Probability a vertex is isolated: $(1-p)^{n-1}$

When $p = \frac{\lambda}{n}$, the degree $D \sim \text{Binomial}(n-1, p) = \text{Binomial}(n-1, \frac{\lambda}{n})$

This is a binomial with large n and small p, so by the Poisson approximation to the binomial, it's approximately Poisson(λ). The mean is $(n-1)\frac{\lambda}{n} \approx \lambda$. In general: $\text{Binomial}(n, p) \approx \text{Poisson}(np)$ when n is large and p is small.

Sharp Threshold for Connectivity

As you increase p from 0 to 1, the graph goes from empty to fully connected. At what value p does it suddenly become connected?

$$p \approx \frac{\ln n}{n}$$

For the graph to be fully connected, there must not be any isolated nodes:

The probability a given node is isolated is $(1-p)^{n-1} \approx e^{-p(n-1)}$. The expected number of isolated nodes is $n \cdot e^{-p(n-1)}$. Plugging in $p = \frac{\ln n}{n}$:

$$n \cdot e^{-p(n-1)} \to n \cdot e^{-\frac{\ln n}{n} \cdot (n-1)} \approx n \cdot e^{-\ln n} = n \cdot e^{\ln(\frac{1}{n})} = n \cdot \frac{1}{n} = 1.$$

So at $p = \frac{\ln n}{n}$, the expected # of isolated nodes is 1. So if you scale p up, it goes to 0, and if you scale it down, it goes to ∞.

Basics

Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(B) > 0$

Total Probability: $P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i)$

Bayes Rule: $P(A|B) = \frac{P(A) P(B|A)}{P(B)}$

Union Bound: $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$

Independence: $P(A|B) = P(A)$ or $P(A \cap B) = P(A) P(B)$

Pairwise Independence: every pair of RVs is independent (does not imply mutual)

Conditional Independence: every subset of of RVs is independent (implies pairwise)

Random Variables

Discrete	PMF / PDF / CDF	$\mathbb{E}[X]$	$Var(X)$	MGF	
Uniform (a, b)	$Pr(X=k) = \frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{1}{b-a+1} \frac{e^{at}(1-e^{(b-a+1)t})}{1-e^t}$	each # between $[a, b]$ equally likely
Bernoulli (p)	$Pr(X=0) = 1-p$; $Pr(X=1) = p$	p	$p(1-p)$	$1-p+pe^s$	one coin flip w/ prob p
Binomial (n, p)	$Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	$(1-p+pe^s)^n$	# of successes in n coin flips
Geometric (p)	$Pr(X=k) = (1-p)^{k-1} p$	$1/p$	$\frac{1-p}{p^2}$	$\frac{pe^s}{1-(1-p)e^s}$	# of trials before first success
Poisson (λ)	$Pr(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ	$e^{\lambda(e^s-1)}$	# of events in window
Exponential (λ)	$f_X(x) = \lambda e^{-\lambda x}$ $F_X(x) = 1-e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda - s}$	waiting time to next event
Gaussian (μ, σ)	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2	$e^{s^2/2}$	shape everything converges to
Uniform (a, b)	$f_X(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$s \neq 0$? $\frac{e^{bs}-e^{as}}{s(b-a)}$: 0	equally likely anywhere in range
Erlang (k, λ)	$Pr(X=x) = (\lambda^k x^{k-1} e^{-\lambda x})/(k-1)!$	k/λ	k/λ^2	$(\lambda/(\lambda-t))^k$	waiting time until k^{th} event; sum of k exp

MGFs

$M_X(s) = \mathbb{E}[e^{sX}] = \sum_k e^{sk} P(X=k) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx$

Derivative of an MGF: $\left.\frac{d^n M(s)}{ds^n}\right|_{s=0} = \int x^n f(x)\, dx = \mathbb{E}[X^n]$

Uniqueness of MGF: $M_X(s) = M_Y(s) \implies X = Y$

Convergence

$\{X_n\} = X_1, X_2, \ldots, X_n$ (iid RVs)

Sample/Empirical Mean: $M_n = \frac{1}{n} \sum X_i$

True/Population Mean: $\mathbb{E}[M_n] = \frac{1}{n} \sum \mathbb{E}[X_i] = \mathbb{E}[X_i]$

Variance: $Var(M_n) = \frac{1}{n^2} \sum Var(X_i) = \frac{n\, Var(X_i)}{n^2} = \frac{Var(X_i)}{n}$

- as $n \to \infty$, $var(M_n) = 0$

WLLN: $\lim_{n\to\infty} P(|\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X]| \geq \varepsilon) = 0$

CLT: $\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$

Concentration Bounds

Markov Bound: $P(X \geq a) \leq \frac{E[X]}{a}$ for non-neg X

Chebyshev: $P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$ for $c > 0$

Chernoff: $P(X > a) \leq \min_{s>0} \frac{M_X(s)}{e^{sa}}$ (right tail)

$P(X < a) \leq \min_{s>0} \frac{\mathbb{E}[e^{-sx}]}{e^{sa}}$ (right tail)

almost surely $\implies$ in probability $\implies$ in distribution

almost surely: $(X_n)_{n=1}^{\infty}$ converges a.s. to X if
$P(\{\omega \in \Omega : \lim_{n\to\infty} X_n = X\}) = 1$ or
$P(\lim_{n\to\infty} X_n \neq X) = 0$ or
$X_n \xrightarrow{a.s.} X$

SLLN: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}(X_i)$

Borel-Cantelli: $\sum_{n=1}^{\infty} P(A_n) < \infty \implies P(A_n \text{ i.o.}) = 0$

$\sum_{n=1}^{\infty} P(A_n) = \infty$ and $(A_n)_{n=1}^{\infty}$ independent $\implies P(A_n \text{ i.o.}) = 1$

in probability: $(X_n)_{n=1}^{\infty}$ converges i.p. to X if
$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$ or
$X_n \xrightarrow{p} X$

in distribution: $(X_n)_{n=1}^{\infty}$ converges i.d. to X if
$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$
$X_n \xrightarrow{d} X$

SLLN/Borel-Cantelli/Big $\implies$ almost surely | Chebyshev $\implies$ in probability | CLT $\implies$ in distribution

Information Theory

Entropy

$H(X) = \mathbb{E}[\log \frac{1}{p(X)}] = \sum_x P(X=x) \log(\frac{1}{P(X=x)}) = -\sum_x P(X=x) \log P(X=x)$; average surprise/info from observing X

Conditional Entropy: $H(X|Y) = \sum_y P(Y=y) H(X|Y=y)$

$= \sum_y P(Y=y) \sum_x P(X|Y=y) \frac{1}{\log P(X|Y=y)}$

$= H(X,Y) - H(Y)$

; average remaining uncertainty in X after Y

Chain Rule of Entropy: $H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$

Mutual Information: $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$

- how much information Y gives about X

Channel Capacity: $C = \max_{f_X} I(X;Y)$; mutual information maximized over input distribution, X

↳ $I(X;Y) = H(Y) - H(Y|X)$ ← fixed

↑ maximize this

Erasure Channels & Coding

Binary Erasure Channel: marks bits as erased w/ probability p

Capacity $= 1-p$

Binary Symmetric Channel: flips bits w/ probability p

Capacity $= 1 - H(p) = 1 - [-p \log p - (1-p) \log(1-p)]$

Shannon's random coding

message of L bits

encode each 2^L possible messages to an n bit codeword

rate $R = L/n$

BSC: prob of a wrong codeword matching on noisy bits $= \frac{1}{2^{n(1-H(p))}}$

BEC: prob of a wrong codeword matching on unerased bits $= \frac{1}{2^{n(1-p)}}$