Information Theory

- Entropy: measure of surprise / amount of information
- Joint Entropy: $H[X,Y]$ info conveyed by X, Y
- Conditional Entropy: $H[X|Y]$ info in X not provided by Y
- Mutual information: $I[X;Y]$ info. shared by X, Y.

Equations:

$$H[X] = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_X[X]}\right] = \sum_x \mathbb{P}[X=x] \log \frac{1}{\mathbb{P}[X=x]}$$

$$H[X,Y] = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{X,Y}[X,Y]}\right] = \sum_{x,y} \mathbb{P}[X=x, Y=y] \log \frac{1}{\mathbb{P}[X=x, Y=y]}$$

$$H[X|Y] = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{X|Y}[X|Y]}\right] = \sum_y \mathbb{P}_Y[y]\, H[X|Y=y]$$

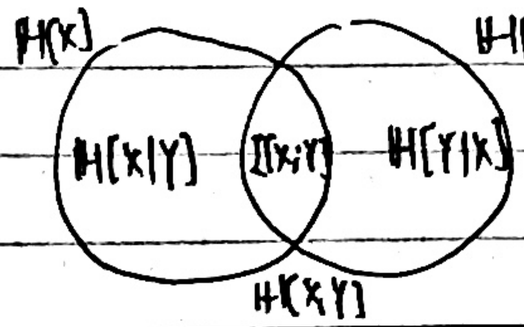
$$= \sum_y \mathbb{P}_Y[y] \sum_x \mathbb{P}[x|y] \log \frac{1}{\mathbb{P}_{X|Y}[x|y]}$$

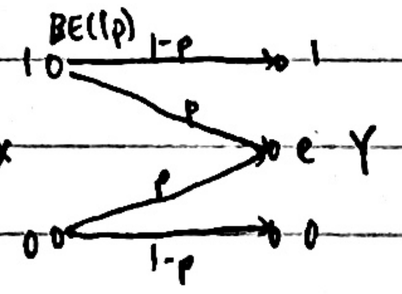
$$I[X;Y] = H[X] - H[X|Y] = H[X] + H[Y] - H[X,Y] = I[Y;X]$$

$$I[X;Y] = -\sum_x \sum_y \mathbb{P}_{X,Y}[x,y] \log \frac{\mathbb{P}_X[x]\mathbb{P}_Y[y]}{\mathbb{P}_{X,Y}[x,y]}$$

Properties:

- $H[X] \geq 0$; H concave (i.e. $-H$ convex)
- $H[X]$ represents minimum number of bits per character to represent a file. $H[X] \leq \log_2 |\mathcal{X}|$ (maximized at uniform dist)
- $\log x$ and $x \log \frac{1}{x}$ are both concave (use Jensen's)
- $H[X,Y] \leq H[X] + H[Y]$ with equality iff X, Y independent
- $H[X,Y] \geq \max(H[X], H[Y])$
- [Chain Rule] $H[X,Y] = H[Y] + H[X|Y] = H[X] + H[Y|X]$
- $H[Y|X] \leq H[Y]$ with equality iff X, Y independent.
- $I[X;Y] \geq 0$ with equality iff X, Y independent.





Fountain Code

- Encoding
  - Divide data into n chunks. Pick $1 \leq d \leq n$ according to a distribution (degree of packet)
  - Select d random chunks of data, combine binary representations of them using XOR ($Y_i = X_{k_1} \oplus X_{k_2} \oplus \cdots \oplus X_{k_d}$)
  - Transmit Y_i
- Decoding
  - If degree of $Y_i = 1$, assign to X_i. Else, XOR known chunks
  - With new info, peel other unfinished packets. Repeat.

Let $f_k(d)$ = expected number of packets of degree $d > 1$ when k chunks have been recovered.

$$f_k(d) = \frac{n-k}{d(d-1)} \qquad P(d) = \begin{cases} \frac{1}{n} & d=1 \\ \frac{1}{d(d-1)} & d=2,\ldots,n. \end{cases}$$

Discrete Time Markov Chain (continued)

- For irreducible MC, some state positive recurrent $\Leftrightarrow$ all states positive recurrent. Expected return time $= \frac{1}{\pi_i}$. Same for null recurrent

Source coding and compression

Source Coding Theorem: Let $(X_n)_n$ be i.i.d and $\varepsilon > 0$. $\exists$ source coding scheme s.t. $\lim_{n\to\infty} \mathbb{E}\left[\frac{\ell(X_1,\ldots,X_n)}{n}\right] \leq H[X] + \varepsilon$

where $\ell(X_1,\ldots,X_n)$ denotes length of binary encryption of $(X_n)_n$. $(X_n)_n$ can be recovered from encoding with probability $1-\varepsilon$

Impossible to compress source $X_1,\ldots,X_n$ below its entropy $H(X_1,\ldots,X_n) = n\,H[X]$ bits, but possible for $n(H[X]+\varepsilon)$.

Typical set: $A_\varepsilon^{(n)}$ is the set of sequences in $\mathcal{X}^n$ that fits the condition $2^{-n(H[X]+\varepsilon)} < \mathbb{P}[X_1=x_1,\ldots,X_n=x_n] < 2^{n(H[X]-\varepsilon)}$

Size of this set is $\approx 2^{nH[X]} \ll |\mathcal{X}^n| = |\mathcal{X}|^n$

$\Rightarrow \mathbb{P}[(X_1,\ldots,X_n) \in A_\varepsilon^{(n)}] > 1-\varepsilon$ for n large.

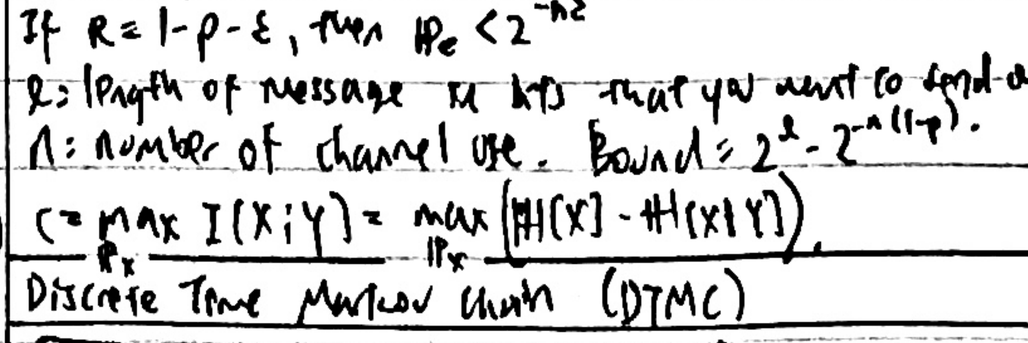
$\Rightarrow (1-\varepsilon)2^{n(H[X]-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H[X]+\varepsilon)}$ for n large

Upshot: Only need to care about encoding of things in $A_\varepsilon^{(n)}$, anything else can be handled one-off. $H[X]$ bits.

Huffman Coding: Expected average length between $H[X]$ and $H[X]+1$

Channels and Capacity

Channel Coding Theorem: Any rate below the channel capacity $C := \max_{p \in P, \text{Xmp}} I(X;Y)$ is achievable. Conversely, any sequence of codes with $\mathbb{P}_e(n) \to 0$ as $n \to \infty$ has rate $R \leq C$. Lossy rate $> C$ with error with probability 1.



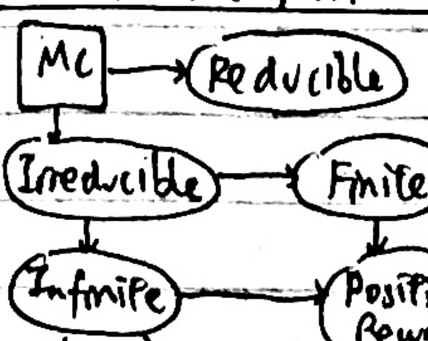
Denote $\mathbb{P}_e[i]$ as probability of error given i bits sent. i.e. $P_{\text{error}}[i] = \max_{m \in \{0,1\}^i} \mathbb{P}[\hat{m} \neq m \mid m \text{ sent}]$

Rate $= \frac{k}{n}$. Say it is achievable if $\mathbb{P}_e[i] \to 0$ as $i \to \infty$

Define capacity = largest achievable rate.

Theorem: Capacity (BEC(p)) $= 1-p$. i.e. $1-p$ bits transmitted per channel use.

Key Idea: n i.i.d. last np bits erased. Reliably)



$$\mathbb{P}_e = \mathbb{P}[c_1 = c_1' \text{ OR } c_2 = c_2' \text{ OR } \cdots] \leq \sum_{i=2}^{2^k} \mathbb{P}[c_i = c_i'] < 2^k \cdot 2^{-n(1-p)} = 2^{nR} \cdot 2^{-n(1-p)} = 2^{-n(1-p-R)}$$

If $R = 1-p-\varepsilon$, then $\mathbb{P}_e < 2^{-n\varepsilon}$

k: length of message in bits that you want to send over.

n: number of channel use. Bound: $2^k \cdot 2^{-n(1-p)}$.

$$C = \max_{\mathbb{P}_X} I(X;Y) = \max_{\mathbb{P}_X} \left(H[X] - H[X|Y]\right).$$

Discrete Time Markov Chain (DTMC)

Transient: $\mathbb{P}[X_i < i \text{ for only many } n | X_0 = i] = 1$

$\mathbb{P}[\exists n \in \mathbb{N} \text{ s.t. } X_n = i \mid X_0 = i] < 1$

$\mathbb{P}[\text{never return to } i \mid X_0 = i] > 0$

$\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$

Recurrent:

$\mathbb{P}[X_{n+1} = i \text{ i.o.} \mid X_0 = i] = 1$

$\mathbb{P}[\exists n \in \mathbb{N} \text{ s.t. } X_n = i] = 1$

$\mathbb{P}[\text{never return to } i \mid X_0 = i] = 0$

$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$

Facts:

- Irreducible, possibly a MC is positive recurrent iff π^* exists
- State i is null recurrent if i is recurrent and expected return time $= \infty$
- State i is positive recurrent if i is recurrent and expected return time $< \infty$

Continuous Time Markov Chains (CTMC)

Properties: $\pi Q = 0$; $\sum \pi_i = 1$; $Q \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$, $\sum_j q_{ij} = 0 \ \forall i$

Holding rate of state i = $q_i = \sum_{j \neq i} q_{ij}$ $\Rightarrow$ Transition time $\sim \text{Expo}(q_i)$

$\Rightarrow$ Mean transition time = $\frac{1}{q_i}$. Jump chain stationary distribution p_i

Differential Analysis: For small time interval δ at state i,

$\mathbb{P}[0 \text{ transitions}] = 1 - q_i\delta + o(\delta)$, $\mathbb{P}[1 \text{ transition}] = q_i\delta + o(\delta)$

$\mathbb{P}(\geq 2 \text{ transitions}) = o(\delta)$ $\mathbb{P}[\text{transition to } j (\neq i)] = q_{ij}\delta + o(\delta)$

Relationship to Jump chain / DTMC

- π_i = long term fraction of time CTMC spends in state i
- p_i = long term fraction of number of transitions leading to state i
- $\pi_i q_{ij}$ = expected number of transitions from ij per unit time.
- p may exist even if π does not exist.

$$\pi_i = \frac{\frac{p_i}{q_i}}{\sum_{j \in \mathcal{X}} \frac{p_j}{q_j}} \qquad p_i = \frac{\pi_i q_i}{\sum_{j \in \mathcal{X}} \pi_j q_j}$$

$\therefore [\pi_i q_i]$ eigenvector of P with eigenvalue 1

$\therefore \left[\frac{p_i}{q_i}\right]$ eigenvector of Q with eigenvalue 0

- DTMC with same stationary distribution π
  - Find max holding time q_{max}
  - Divide all edges by q_{max} in CTMC and add self loops
- Jump Chain
  - $P_{ij} = \frac{q_{ij}}{q_i}$, $P_{ii} = 0$ (i.e. no self loops)
  - Same CTMC can correspond to different sets of jump chains and holding times if self loop condition is used.

Equations

$$\pi_j \underbrace{\sum_{k \neq j} q_{jk}}_{\text{flow out}} = \underbrace{\sum_{k \neq j} \pi_k q_{kj}}_{\text{flow in}} \qquad \sum_{k=1}^{|\mathcal{X}|} \pi_k = 1$$

$\pi_j = 0$ for transient/null recurrent; $\pi_j > 0$ for positive recurrent

$\pi_i q_{ij} = \pi_j q_{ji}$ (Detailed Balance) — CTMC reversible $\Leftrightarrow$ jump chain reversible

$\mathbb{E}[S_i] = \frac{1}{q_i} + \sum_j \frac{q_{ij}}{q_i} \mathbb{E}[S_j]$ (First step Equations) $\frac{1}{\pi_i}$ (Expected Return)

$\mathbb{P}[X(t) = j \mid X(s) = i] = \mathbb{P}[X(t-s) = j \mid X(0) = i] \ \forall s \leq t, i, j \in \mathcal{X}$ (Time homogeneous)

Poisson Process PP(λ)

Notation: $N(t) = N([0,t])$, $N([a,b]) = \#$ arrivals in $[a,b]$, $N(0) = 0$

Arrival Times $T_n \sim \text{Erlang}(n;\lambda)$; Interarrival times $S_n \sim \text{Expo}(\lambda)$ i.i.d

$N([a,b]) \sim \text{Poisson}(\lambda(b-a))$ — time homogeneous, memoryless, independency of disjoint intervals.

Differential Analysis

For small time interval δ, $\mathbb{P}[0 \text{ arrivals}] = 1 - \lambda\delta + o(\delta)$

$\mathbb{P}[1 \text{ arrival}] = \lambda\delta + o(\delta)$

$\mathbb{P}[\geq 2 \text{ arrivals}] = o(\delta)$

Facts & Techniques

- Poisson process run backwards is still Poisson
- Given $N(t)$, the $N(t)$ arrivals are uniformly distributed in $[0,t]$ same order statistics as n i.i.d uniform $([0,t])$ RVs
- Poisson splitting (receiving processes independent)
- Poisson merging (arrival from each process is $\frac{\lambda_i}{\sum \lambda_j}$)
- Reduce to first arrival time; interarrival time, arrivals in fixed interval $\Rightarrow$ use binomial / geometric distribution).
- Random Incidence Paradox
- $f_{T_1,\dots,T_n}(t_1,\dots,t_k) = \lambda^k e^{-\lambda t_k}$
- $N(n) = \sum_{i=1}^{n} N(i) - N(i-1) = \sum_{i=1}^{n} Y_i$ where $Y_i \sim \text{Poisson}(\lambda)$

  $\therefore \frac{N(n)}{n} \xrightarrow{a.s.} \lambda$.
- $S_N = X_1 + \dots + X_N$ where $N \sim \text{Geometric}(p)$, $X_i \sim \text{Expo}(\lambda)$ i.i.d

  $\Rightarrow S_N \sim \text{Expo}(p\lambda)$.
- $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, $X \mid X+Y \sim \text{Binomial}\left(n, \frac{\lambda}{\lambda+\mu}\right)$.
- $\mathbb{P}[\tau \leq t+\varepsilon \mid \tau > t] = \lambda\varepsilon + o(\varepsilon)$.

Metropolis Hastings Algorithm

Target distribution $p(x)$. Want to create an aperiodic Markov chain with stationary distribution $\pi(x) = p(x)$.

$\Rightarrow$ allows generation of random sample even if p intract

Assumptions:

- Can compute $f(x)$: a direct proportional estimate of $p(x)$ i.e. $p(x) = \frac{f(x)}{\sum_y f(y)}$ (x is current state)
- Can compute $q(x,\cdot)$: a proposal distribution for next state

Algorithm: For each time step t,

- propose next candidate state y according to $q(x,\cdot)$
- Accept y with probability $A(x,y) = \min\left(1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right)$
- If accept, move to y, else, stay in x.

Revised Extensions:

- Stationary distribution of MC equals $p(x)$
- Taking every k^{th} step state reduces dependence between samples
- Letting chain propagate a bit lets distribution converge to stationary
- Any MHMC can be made aperiodic by giving each state a self loop with some probability.
- Burn in time: how long initial distribution converges to stationary distribution.
- Disadvantage for bimodel distributions

Convergence

General Procedure: • Markov / Chebyshev; reduce to WLLN/SLLN • Convergence in distribution, Levy MGF

Theorems / Tricks: • $Y \xrightarrow{\mathbb{P}} y$, $Z \xrightarrow{\mathbb{P}} z \Rightarrow Y+Z \xrightarrow{\mathbb{P}} y+z$

- $(X_n)_n$ be nonnegative and integer valued, X too. Then $(X_n)_n \xrightarrow{d} X$ iff $\lim_{n\to\infty} \mathbb{P}[X_n = k] = \mathbb{P}[X = k]$ for $k = 0, 1, 2$
- $|X| \leq |X - \mathbb{E}[X]| + |\mathbb{E}[X]|^{\uparrow a}$ (Δ-inequality)
- If $(X_n)_n \xrightarrow{d} c$ where c constant, then $(X_n)_n \xrightarrow{\mathbb{P}} c$.
- If $\forall \varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}[|X_n - X| > \varepsilon] < \infty$, then $(X_n)_n \xrightarrow{a.s.} X$
- For $\varepsilon > 0$, define $A_m = \{|X_n - X| < \varepsilon \ \forall n \geq m\}$. Then $(X_n)_n \xrightarrow{a.s.} X$ if and only if $\forall \varepsilon > 0 \ \lim_{m\to\infty} \mathbb{P}[A_m] = 1$.
- If $(X_n)_n$, $(Y_n)_n$ defined on same sample space and $(X_n)_n \xrightarrow{a.s.} X$ and $(Y_n)_n \xrightarrow{a.s.} Y$ then $(X_n + Y_n)_n \xrightarrow{a.s.} X+Y$.
- If $(X_n)_n$ and $(Y_n)_n$ defined on same sample space and $(X_n)_n \xrightarrow{\mathbb{P}}$ and $(Y_n)_n \xrightarrow{\mathbb{P}} Y$ then $(X_n + Y_n)_n \xrightarrow{\mathbb{P}} X+Y$.

Continuous Mapping Theorem:

- Let f be continuous function.
  - If $(X_n)_n \xrightarrow{a.s.} X$ then $(f(X_n))_n \xrightarrow{a.s.} f(X)$
  - If $(X_n)_n \xrightarrow{\mathbb{P}} X$ then $(f(X_n))_n \xrightarrow{\mathbb{P}} f(X)$
  - If $(X_n)_n \xrightarrow{d} X$ then $(f(X_n))_n \xrightarrow{d} f(X)$

Borel Cantelli Lemmas:

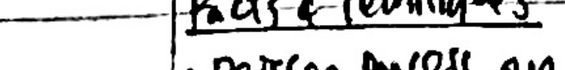
- Let $A_1, A_2, \dots$ be sequence of events in probability space
  - If $\sum_{i=1}^{\infty} \mathbb{P}[A_i] < \infty$ then $\mathbb{P}[A_i \text{ i.o.}] = 0$.
  - If $\sum_{i=1}^{\infty} \mathbb{P}[A_i] = \infty$ and $(A_n)_n$ independent, $\mathbb{P}[A_i \text{ i.o.}] = 1$.

Convergence Theorems: Suppose $(X_n)_n \xrightarrow{a.s.} X$

1. If $0 \leq X_1 \leq \dots \leq X_n \leq \dots$ then $(\mathbb{E}[X_n])_n \to \mathbb{E}[X]$ (Monotone)
2. If $\exists$ RV $Y \geq 0$ with $\mathbb{E}[Y] < \infty$ and $|X_n|, |X| \leq Y \ \forall n$ then $(\mathbb{E}[X_n])_n \to \mathbb{E}[X]$. (Dominated)

Classics

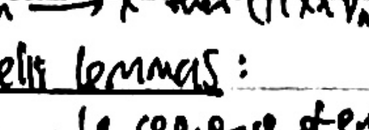
Gambler's ruin: $p = \frac{1}{2} \Rightarrow \mathbb{P}[\text{win}] = \frac{n_1}{n_1+n_2}$; $\mathbb{E}[\text{stopping}] = n_1 \cdot n_2$



$p \neq \frac{1}{2} \Rightarrow \mathbb{P}[\text{win}] = \frac{1-\left(\frac{1-p}{p}\right)^{n_1}}{1-\left(\frac{1-p}{p}\right)^{n_1+n_2}}$; $\mathbb{E}[\text{stopping}] = \frac{n_1}{(1-p)-p} - \frac{n_2}{(1-p)-p}\left(\frac{\left(\frac{1-p}{p}\right)^{n_1}-1}{\left(\frac{1-p}{p}\right)^{n_1+n_2}-1}\right)$

Entropy

$\binom{n}{np} \approx 2^{nH(p)}$



$H(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$

$n! \approx \left(\frac{n}{e}\right)^n$