Info Lect. 1

Measures of Information

ex.

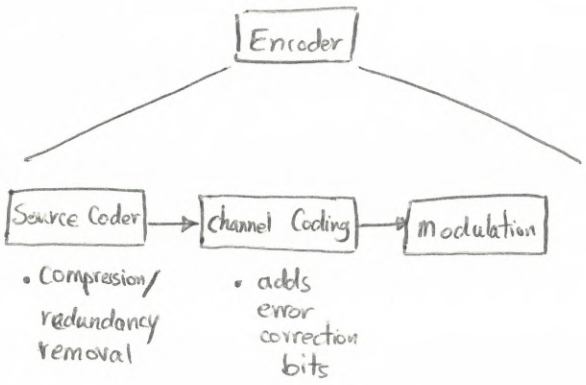
chirp ≡ modulated signal

$s(t) = a(t) \cdot \cos(2\pi f_0 t)$

Received signal

$$r(t) = C\, a(t-t_0) \cos\left(2\pi (f_0 + f_D)(t-t_0)\right\} + \varphi_0) + w(t)$$

$w(t)$: receiver noise + radio interferences



* Joint source_channel coding is an OPEN problem.

3 INFORMATION Measures

1. Entropy — limit of lossless compression

2. Mutual Information — distortions in the channel

3. Relative Entropy — cost of using wrong pmf in encoding, related to Hypothesis Testing

---

Entropy:

x is a discrete R.V.

$\mathcal{X}$: all possible outcomes, finite number of elements

$p(x)$

$-\log_2 p(x)$: + uncertainty of observing x
+ Description Length of outcome x (in bits)

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

---

ex. $X \sim$ Bernoulli (p)

$$\begin{cases} 1, & p \\ 0, & 1-p \end{cases} \Rightarrow H(X) = \underbrace{H(p)}_{\text{Binary Entropy}} = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

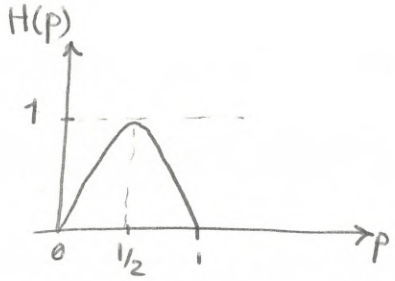
Properties:

1. $H(X) \geq 0$
2. If $X \sim \text{Bernoulli}(p) \Rightarrow 0 \leq H(p) \leq 1$

$0 \leq H(X) \leq \log_2 n$ (number of outcomes)

3. $H_b(X) = \sum p(x) \log_b p(x)$

$= (\log_b a) H_a(x)$



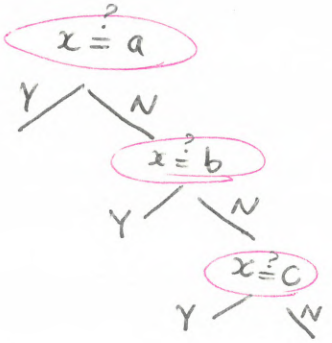
4. Entropy is a functional. (Function of function: $H(p(x))$)

5. $H(X)$ is concave in $p(x)$.

6. n letter alphabet (power of 2)

Ex.

a	b	c	d
1/2	1/4	1/8	1/8



3 questions needed!

$$H(X) = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8}$$

$$= \frac{7}{4} \text{ [bits]}$$

Joint Entropy of $(X, Y) \sim P(X, Y)$

- $-\log p(x, y)$
- $H(X, Y) = \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)}$

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

$$H(X|Y) = \sum_{y} H(X|Y=y)\, p(y)$$

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

$$= +H(X,Y) \stackrel{=}{\text{OR}} H(Y)$$

$$\Rightarrow \boxed{\begin{aligned} H(X,Y) &= H(Y) + H(X|Y) \\ &\stackrel{\text{OR}}{=} H(X) + H(Y|X) \end{aligned}}$$

2. Relative Entropy

X $\quad p_X(x)$, $q_X(x)$

$$D(p\|q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$$

"distance" between p and q

- not symmetric $\quad D(p\|q) \neq D(q\|p)$
- not satisfy (△) triangular inequality $\quad D(p\|q) \not\leq D(p\|r) + D(r\|q)$

---

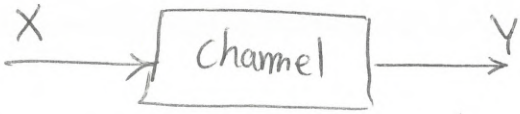
3. Mutual Information (Description of Channel)

X, Y are 2 R.V.'s

$p(x,y)$

$$I(X;Y) \triangleq \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)\cdot p(y)}$$

- How Much Information Y has about X or X has about Y



- $I(X;Y) = D(p(x,y) \| p(x)p(y))$

- $I(X;Y) = \sum_{y} \sum_{x} p(x,y) \log \frac{p(y|x)\,p(x)}{p(x)\,p(y)}$

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

- $D \geq 0 \Rightarrow I(X;Y) \geq 0 \Rightarrow H(Y) \geq H(Y|X)$

Info. Let 3

Entropy: limit of lossless compression

Review:

X : discrete R.V

$\mathcal{X}$: alphabet

$$H(X) = -\sum_{x} p(x) \cdot \log_2 p(x) = \underset{x \sim p(x)}{E}\left[\log_2 \frac{1}{p(x)} \right]$$

- $H(X) \geq 0$
- concave in p $\cap$

Relative Entropy

$\mathcal{X}$

$p(x)$

$q(x)$

(X) symmetry

(X) Δ ineq.

$$D(p\|q) = \sum_{x} p(x) \log_2 \frac{p(x)}{q(x)}$$

(V) symmetry

(X) Δ ineq.

$D(p\|q) + D(q\|p) \equiv$ Kullback - Liebler Divergence

symmetrical (V)

- $D(p\|q) \geq 0$
- concave in p and also in q

Mutual Information: description of channel / noise and distortion

$$I(X;Y) = \sum_{p(x,y)} p(x,y) \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

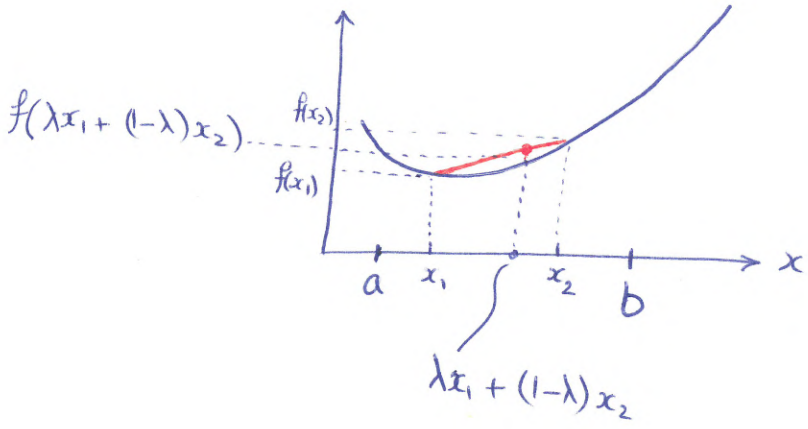
$$I(X;Y) = D\left(p(x,y) \,\|\, p(x)p(y) \right)$$

- $I(X;Y) \geq 0$
- $I(X;Y) = H(X) - H(X|Y) \geq 0$
  - ↳ conditioning reduces entropy

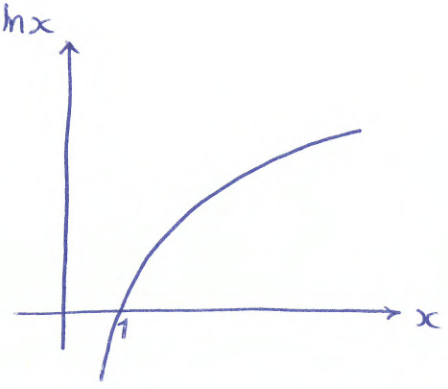
Convexity

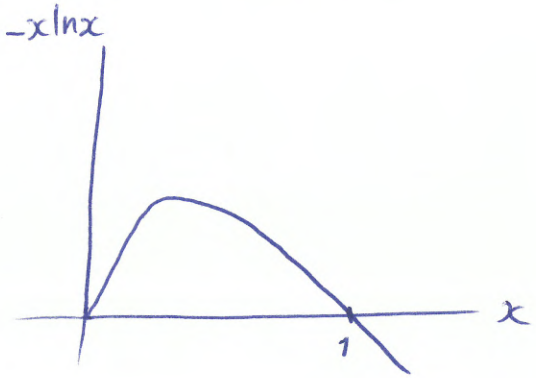
Defn. $f(x)$ is **convex** $\cup$ on (a, b) if for $x_1, x_2 \in (a,b)$ and for $\lambda \in (0, 1)$

$$f(x_1 \lambda + x_2 (1-\lambda)) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



ex.





Thm. $f(x)$ is twice continuous and differentiable, then if $\frac{d^2 f(x)}{dx^2} \geq 0$, then $f(x)$ is convex $\cup$

$x \in (a,b)$

---

proof.

Taylor Expansion

$$f(x) = f(x_0) + f'(x_0)\cdot(x-x_0) + \frac{1}{2}\frac{d^2 f(x^*)}{dx^2}\cdot(x-x_0)^2$$

$$\Rightarrow f(x) \geq f(x_0) + f'(x_0)\cdot(x-x_0)$$

$\times\lambda$ $\quad x_1 \to x$ $\quad f(x_1) \geq f(x_0) + f'(x_0)(x_1-x_0)$

$\times(1-\lambda)$ $\quad x_2 \to x$ $\quad f(x_2) \geq f(x_0) + f'(x_0)(x_2-x_0)$ $\qquad \lambda \in (0,1)$

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq \lambda f(x_0) + (1-\lambda) f(x_0) + \lambda f'(x_0)(x_1-x_0) + (1-\lambda) f'(x_0)(x_2-x_0)$$

replace x_0 with $\lambda x_1 + (1-\lambda)x_2$:

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(x_0) + \lambda f'(x_0)x_1 + f'(x_0)x_2 - f'(x_0)x_0 - \lambda f'(x_0)x_2$$

$$\geq f(x_0) + f'(x_0)(\lambda x_1 + x_2 - \lambda x_1 - x_2 + \lambda x_2 - \lambda x_2)$$

$$\geq f(x_0)$$

~~$f(x_0) \leq$~~ $\boxed{f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)}$ (V)

8/25 - 2

(P 2.2, C.T'06)

X
H(X)

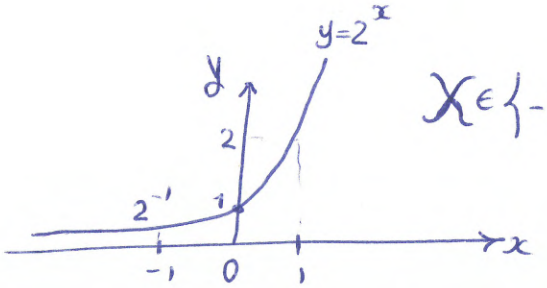
(a) $Y = 2^X$

(b) $Y = \cos(X)$

$H(Y) \overset{?}{\leqslant} H(X)$

(a) $y = 2^x$ is monotonically increasing ↑

one-to-one mapping



$X \in \{-1, 0, 1\}$

$Y \in \{\frac{1}{2}, 1, 2\}$

$P(Y = \frac{1}{2}) = P(X = -1)$

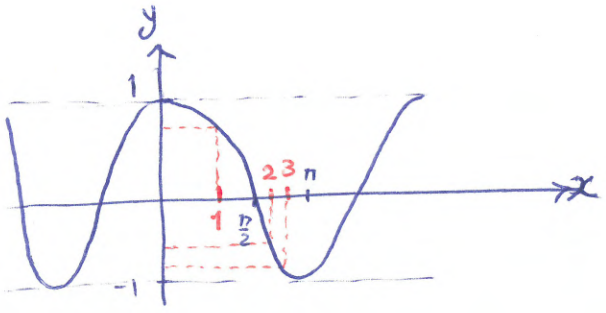
$P(Y = 1) = P(X = 0)$

$P(Y = 2) = P(X = 1)$

$\Rightarrow H(Y) = H(X)$ (Y: monotonic of X)

(b)

$X \in \{1, 2, 3\}$ → monotonic → $H(X) = H(Y)$



$X \in \{0, [redacted], \pi, 2\pi\}$

$P(Y = 1) \overset{M.E.}{=} P(X = 0) + P(X = 2\pi)$

$P(Y = -1) = P(X = \pi)$

$\Rightarrow$ so in general for many-to-one: $H(Y) \leqslant H(X)$

[P. 230, CT '06]

X $\quad p(x)$ is unknown.

X

$H(X)$

$E[X] = \sum_{n=0}^{+\infty} n p(n) = A$ (known)

$\max \quad H(X)$

$p(x)$: $\underbrace{E[X]=A}_{1}$ $\qquad \underbrace{\sum_{n=0}^{+\infty} p(n) = 1}_{2}$

Lagrange Multipliers (λ, μ)

$$\mathcal{L}(\lambda,\mu) = \sum_{n=0}^{+\infty} p(x)\ln p(x) + \lambda\left(\sum_{n=0}^{+\infty} n p(n) - A\right) + \mu\left(\sum_{n=0}^{+\infty} p(n) - 1\right)$$

$$\frac{\partial \mathcal{L}(\lambda,\mu)}{\partial p(n)} = 0$$

$$-\log p(n) - \frac{p(n)}{p(n)} + \lambda n - \mu = 0$$

$$\boxed{p(n) = e^{\mu - 1 + \lambda n}}$$

HW 1 → P 1.4 : starting point

$$\ln x \leq x - 1$$

(came from Taylor series)

Info. Lect. 4

P. 2.30

$X \sim$ discrete R.V.

integer-valued X $\Rightarrow$ maximize $H(X)$

$p(n)$, $E(X) = A$

$$\sum_{n=0}^{+\infty} p(n) = 1$$

$$\mathcal{L}(p, \lambda, \mu) = \underbrace{E_p\left[\log \frac{1}{p(x)}\right]}_{-\sum_{n=0}^{+\infty} p(n) \log p(n)} + \lambda\left(\sum_{n=0}^{+\infty} n\,p(n) - A\right) + \mu\left(\sum_{n=0}^{+\infty} p(n) - 1\right)$$

$$\frac{\partial \mathcal{L}(p, \lambda, \mu)}{\partial p(n)} = -\ln p(n) - 1 + \lambda n + \mu \Big|_{p^*(n)} = 0$$

$$p(n) = e^{\lambda n + (\mu - 1)} = e^{\lambda n} \cdot e^{(\mu - 1)}$$

to find μ:

$$\sum_{n=0}^{+\infty} e^{\lambda n} \cdot e^{\mu - 1} = 1 \Rightarrow e^{\mu - 1} = 1 - e^{\lambda}$$

assume $e^{\lambda} < 1 \Rightarrow \lambda < 0$

to find λ : $\quad p(n) = e^{\lambda n}(1-e^{\lambda})$

$$\sum_{n=0}^{+\infty} n e^{\lambda n}(1-e^{\lambda}) = A$$

$$\sum_{n=0}^{+\infty} n e^{\lambda n} = \frac{d}{d\lambda}\left(\sum_{n=0}^{+\infty} e^{\lambda n}\right) = \frac{d}{d\lambda}\left(\frac{1}{1-e^{\lambda}}\right) = \frac{e^{\lambda}}{(1-e^{\lambda})^2}$$

$$\Rightarrow \frac{e^{\lambda}}{(1-e^{\lambda})^2}(1-e^{\lambda}) = A$$

$$\boxed{e^{\lambda} = \frac{A}{1+A}}$$

$$\Rightarrow \quad p(n) = \left(\frac{A}{1+A}\right)^n \left(\frac{1}{A+1}\right)$$

$$\boxed{p(n) = \left(\frac{A}{A+1}\right)^n \frac{1}{A+1}}$$

Jensen's Inequality

- f is convex over (a, b)
- X is a discrete R.V. with alphabet (support) $\mathcal{X}$ and outcomes in (a,b)

$$E[f(X)] \geq f(E(X))$$

- concave f $\Rightarrow$ invert inequality

ex. 1

$$D(p\|q) \geq 0$$

proof:

$$\sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)}\left[\log \frac{p(x)}{q(x)}\right]$$

$$= -E_{p(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

log is concave.

$$E_{p(x)}\left[\log \frac{q(x)}{p(x)}\right] \leq \log\left[E_{p(x)}\left(\frac{q(x)}{p(x)}\right)\right]$$

$$\Rightarrow \quad -E_{p(x)}\left[\log \frac{q(x)}{p(x)}\right] \geq -\log\underbrace{\left(E_{p(x)}\left[\frac{q(x)}{p(x)}\right]\right)}_{1}$$

$$\underbrace{\qquad\qquad}_{0}$$

---

ex. 2

$$H(X) \leq \log |\mathcal{X}|$$

$|X|$: Cardinality $\equiv$ size of alphabet $\mathcal{X}$

$$H(X) = E_{p(x)}\left[\log \frac{1}{p(X)}\right]$$

$$\leq \log \underbrace{E_{p(x)}\left[\frac{1}{p(X)}\right]}_{\sum_{x\in\mathcal{X}} \underbrace{p(x)\cdot\frac{1}{p(x)}}_{1}} = \log |\mathcal{X}|$$

$$|\mathcal{X}|$$

equality holds when $p(x) = \frac{1}{|\mathcal{X}|}$

(uniform distribution)

(like white noise)

Jensen's Inequality proof: $|\mathcal{X}| = 2$

$$X = \begin{cases} x_1 & \text{with prob. } p(1) \\ x_2 & \text{with prob. } p(2) \end{cases}, \quad p(1) + p(2) = 1$$

$$f(x_1)\underbrace{p(1)}_{p} + f(x_2)\underbrace{p(2)}_{1-p}$$

$$\underset{\text{Convexity}}{\geq} f(px_1 + (1-p)x_2)$$

By Induction

$|\mathcal{X}| = k$ and assume that the inequality is proved for alphabet of size $k-1$.

$$E_{p(x)}[f(X)] = \sum_{n=1}^{k} f(x_n) \cdot p(n)$$

$$= f(x_k)\,p(k) + \sum_{i=1}^{k-1} f(x_i)\,\frac{p(i)}{\sum_{j=1}^{k-1} p(j)} \times \sum_{j=1}^{k-1} p(j)$$

$\frac{p(i)}{\sum_{j=1}^{k-1} p(j)}$: a PMF over a $k-1$ size alphabet

our assumed ↑ proved for k-1 ↑

$$\geq f(x_k)\,P(k) + \left(\sum_{j=1}^{k-1} P(j)\right) f\left(\sum_{i=1}^{k-1} x_i \cdot \frac{P(i)}{\sum_{j=1}^{k-1} P(j)}\right)$$

Sum to ONE!

Convexity ↑

$$\geq f\left(x_k P(k) + \cancel{\sum_{j=1}^{k-1} P(j)} \sum_{i=1}^{k-1} x_i \frac{P(i)}{\cancel{\sum_{j=1}^{k-1} P(j)}}\right)$$

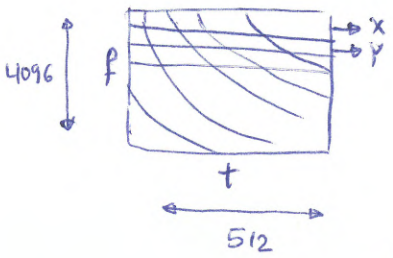
$$f(E(X))$$

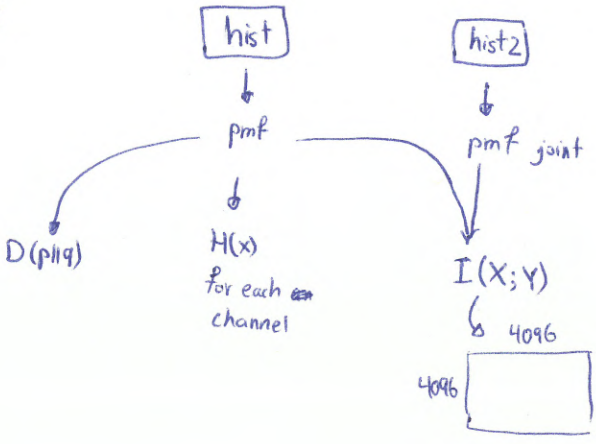
■

---

project assignment:

4096 x 512





- fit Gaussians to each channel?

info. lect.

Project due: 2 weeks from now.

Log-Sum Inequality

$\underline{a} = (a_1, \cdots, a_n) \qquad a_i \geq 0$

$\underline{b} = (b_1, \cdots, b_n) \qquad b_i \geq 0$

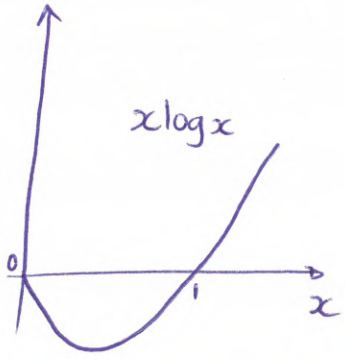
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

proof:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} = \sum_{i=1}^{n} a_i \log \left(\frac{a_i}{b_i} \times \frac{b_i}{b_i} \times \frac{\sum_{j=1}^{n} b_j}{\sum_{j=1}^{n} b_j} \right)$$

x : $\frac{a_i}{b_i} \times \frac{b_i}{b_i}$; $p(x)$: $\frac{\sum_{j=1}^{n} b_j}{\sum_{j=1}^{n} b_j}$

convexity



$E[f(x)] \geq f(E[x])$

$$\geq \left(\sum_{j=1}^{n} b_j \right) \cdot \sum_{i=1}^{n} \underbrace{\frac{a_i}{b_i}}_{x} \underbrace{\frac{b_i}{\sum_{j=1}^{n} b_j}}_{p(x)} \log \sum_{i=1}^{n} \frac{a_i}{b_i} \frac{b_i}{\sum_{j=1}^{n} b_j}$$

$$= \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum a_i}{\sum b_j}$$

prove it myself again!

9, 1, 22-1

Remarks:

1) $H_p(X)$ is concave in p.

2) $I(X;Y) = \sum\sum p(x,y) \log \frac{p(x,y)}{p(x)\cdot p(y)}$

$= \sum\sum q(y|x)p(x) \log \frac{q(y|x)p(x)}{p(x)\cdot p(y)}$ → $\sum_{x'} p(x')q(y|x')$

$\Rightarrow$ M.I. is concave in $p(x)$ and convex in $q(y|x)$.

3) $D(p\|q)$ is convex in both p and q.

$H(\lambda p(x) + (1-\lambda) q(x)) = H_{\lambda p(x)+(1-\lambda)q(x)}(X)$

$= -\sum_x \underbrace{\lambda p(x) + (1-\lambda)q(x)}_{a_1 \quad a_2} \log \frac{\overbrace{\lambda p(x) + (1-\lambda) q(x)}^{\sum a_i}}{\underbrace{\lambda}_{b_1} + \underbrace{(1-\lambda)}_{b_2}}$

$\sum a_i \qquad \sum b_i$

Just forget about the sum for a second:

convex

$\leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$

considering minus

$\geq -\lambda p(x) \log \frac{\lambda p(x)}{\lambda} - (1-\lambda) q(x) \log \frac{(1-\lambda)q(x)}{1-\lambda}$

$$\Rightarrow \quad = \lambda H_p(X) + (1-\lambda) H_q(x)$$

brought back the sum:

$$\Rightarrow \quad H_{\lambda p + (1-\lambda)q}(X) \geq \lambda H_p(X) + (1-\lambda) H_q(x)$$

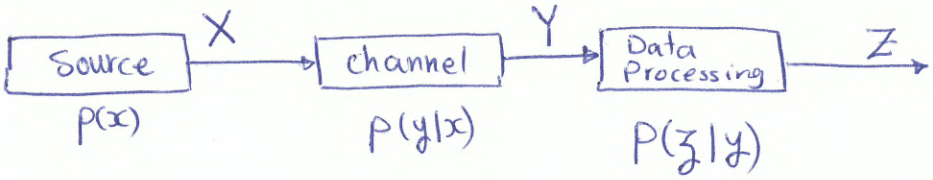
Data Processing Inequality

$X \to Y \to Z$ markov chain (no feedback)

$$P(x,z|y) = \frac{P(x,y,z)}{P(y)} = \frac{P(z|x,y) \cdot P(x,y)}{P(y)} \qquad P(x,y) = P(x|y)\,P(y)$$

$$\boxed{P(x,z|y) = P(z|y) \cdot P(x|y)}$$

Markov * Given present (y), past(x) and future(z) are independent.



- $I(X;Y) \geq I(X;Z)$ → equality: when data processing is perfect.

proof

$$I(X;(Y,Z)) = I(X;Z|Y) + I(X;Y)$$

$I(X;Z|Y)$: ≥ 0? $I(X;Y)$: ≥ 0

$$= I(X;Y|Z) + I(X;Z)$$

$I(X;Y|Z)$: ≥ 0 $I(X;Z)$: ≥ 0

$$I(X;Z|Y) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,z|y)}{p(x|y)\cdot p(z|y)} = 0$$

markov: $p(x|y)\cdot p(z|y)$

■

$$\Rightarrow I(X;Y) \geq I(X;Z)$$

■

$$H(X) - H(X|Y) \geq H(X) - H(X|Z)$$

$$\Rightarrow \boxed{H(X|Z) \geq H(X|Y)}$$

info. lect.

Telescoping Properties

1. $$H(X_1, \dots, X_n) = -\sum_{\underline{x} \in \mathcal{X}^n} p(\underline{x}) \log p(\underline{x})$$

notation
$$x^n = \underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = (x_1, \dots, x_n)$$

$$p(\underline{x}) = p(x_1, \dots, x_n) = p(x_n | x_{n-1}, \dots, x_1) \underbrace{p(x_{n-1}, \dots, x_1)}_{p(x_{n-1} | x_{n-2}, \dots, x_1) p(x_{n-2}, \dots, x_1)}$$

$$= \left[\prod_{k=2}^{n} p(x_k | x_{k-1}, \dots, x_1) \right] p(x_1)$$

$$\Rightarrow H(X_1, \dots, X_n) = -\sum_{\underline{x}} p(\underline{x}) \left\{ \sum_{k=2}^{n} \log p(x_k | x_{k-1}, \dots, x_1) + \log p(x_1) \right\}$$

$$= -\sum_{k=2}^{n} \left\{ \sum_{\underline{x}} p(\underline{x}) \log p(x_k | x_{k-1}, \dots, x_1) + \sum_{\underline{x}} p(x) \log p(x) \right\}$$

$$\boxed{H(X_1, \dots, X_n) = \sum_{k=2}^{n} H(X_k | X_{k-1}, \dots, X_1) + H(X_1)}$$

2. $I(\overbrace{X,Y}^{\text{joint}};Z) = I(X;Z|Y) + I(Y;Z)$

$\hookrightarrow H(X,Y) - H(X,Y|Z) = H(X|Y) + H(Y) - H(X|Y,Z) - H(Y|Z)$

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

$H(Y) - H(Y|Z) = I(Y;Z)$

- AEP: fancy title for usage of law of large numbers.
- chapter 3 of Cover & Thomas

$X_1, \dots, X_n$ iid

$$\begin{cases} S_n = \frac{1}{n}\sum_{i=1}^{n} X_i & \text{a sequence of sample means} \\ V_n = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - S_n)^2 & \text{a sequence of sample variances} \end{cases}$$

X_i: random variable, S_n: R.V., V_n: R.V.

- Do they converge?
- In what sense?
- What is the limit?

Different kinds of convergence:

✓ 1. Convergence in probability (WLLN)

✓ 2. Convergence in distribution (CLTh'm)

✓ 3. Mean-square convergence (continuity and 2nd order statistics of R.P.'s)

x 4. Probability 1 convergence (almost surely)

$$P\left(\lim_{i \to +\infty} X_i = a\right) = 1$$

a → number

→ related to Sandwich theorem

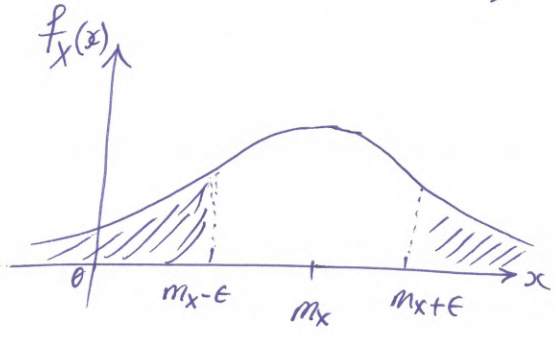
---

we are going to use WLLN, so 2 results need to be recalled: Chebyshev and WLLN.

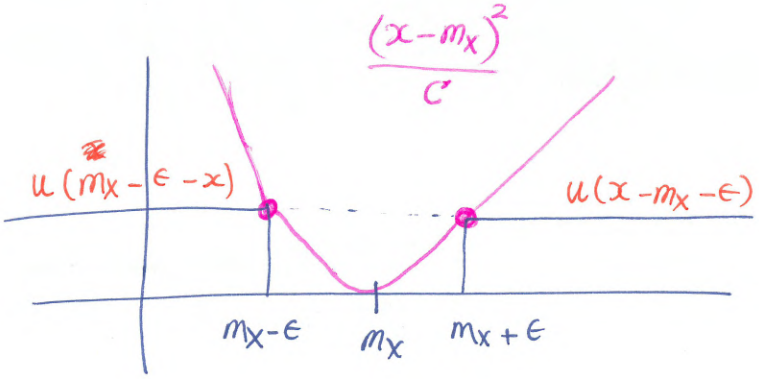
Chebyshev Inequality

X is a R.V. with mean m_X and variance σ_X^2

$$P\left(|X - m_X| > \epsilon\right) \leq \frac{\sigma_X^2}{\epsilon^2}$$



$$P(|X-m_X|>\epsilon) = \int_{-\infty}^{m_X-\epsilon} f_X(x)\,dx + \int_{m_X+\epsilon}^{+\infty} f_X(x)\,dx$$



$$= \int_{-\infty}^{+\infty} \left\{ u(m_X-\epsilon-x) + u(-m_X-\epsilon+x) \right\} \times f_X(x)\,dx$$

2 points chosen to tighten the bound :

$$\frac{(m_X+\epsilon-m_X)^2}{C} = 1 \Rightarrow C = \epsilon^2$$

tightest bound

$$\Rightarrow \leqslant \int_{-\infty}^{+\infty} \frac{(x-m_X)^2}{\epsilon^2} f_X(x)\,dx$$

$X_1, \ldots, X_n, \ldots$

iid

m_X , σ_X^2 : mean and variance of each X_i

R.V.

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$n = 1, 2, 3, \ldots$

$$E[S_n] = \frac{1}{n} \sum E[X_i] = m_X$$

$$Var[S_n] = \frac{1}{n} \sigma_X^2$$

(think how to prove)

$$Var[S_n] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$$

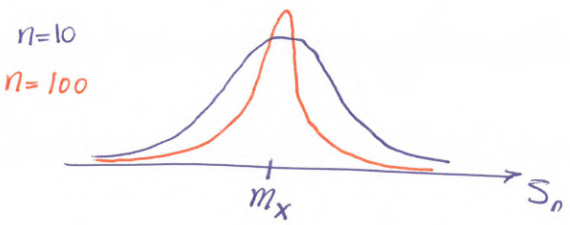
$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma_X^2$$

$$= \frac{n}{n^2} \sigma_X^2 = \frac{1}{n} \sigma_X^2$$

$$P(|S_n - E[S_n]| > \epsilon) \leq \frac{Var[S_n]}{\epsilon^2} = \frac{\sigma_X^2}{n\epsilon^2}$$

as $n \to +\infty \Rightarrow P(|S_n - E[S_n]| > \epsilon) \to 0$

WLLN



WLLN (iid data)

$$S_n \xrightarrow{P} \underbrace{E[S_n]}_{\| \; m_X} \qquad \text{as} \quad n \to +\infty$$

Sample entropy

$$\frac{1}{n} \log \frac{1}{p(x_1, \cdots, x_n)}$$

$$\frac{1}{n} \log \frac{1}{P(X_1, X_2, \cdots, X_n)} \overset{iid}{=\!=} \frac{1}{n} \log \frac{1}{P(X_1) \cdots P(X_n)}$$

($P(X_1, X_2, \cdots, X_n)$: $X_1, X_2, \cdots, X_n$ are iid)

$$= -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \triangleq S_n^e$$

$$S_n^e \xrightarrow{P} ? \quad H(X) \qquad \text{as } n \to +\infty$$

← AEP

Let $Y_i = -\log p(X_i)$ (X_i: iid) $\Rightarrow Y_i$: iid

$$S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{P} E[Y_i]$$

$$E_{P_{Y_i}}[Y_i] = E_{P_{X_i}}[-\log P(X_i)] = -\sum_x p(x) \log p(x) = H(X)$$

info. lect.

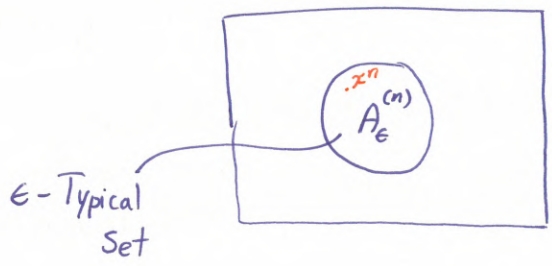
$\mathcal{X}$: $X_1, \dots, X_n$ iid RV

AEP

$$\frac{1}{n} \log \frac{1}{P(X^n)} \xrightarrow[n \to +\infty]{P} H(X)$$

$$P\left(\left| \frac{1}{n} \log \frac{1}{P(X^n)} - H(X) \right| > \epsilon \right) \longrightarrow 0$$

as $n \to +\infty$



$\mathcal{X}^n$

$|\mathcal{X}^n|$ = size = total # sequences

for $x^n \in A_\epsilon^{(n)}$

$$\left| A_\epsilon^{(n)} \right| \cong 2^{nH(X)}$$

$$P(x^n) \cong 2^{-nH(X)}$$

$$P\left(A_\epsilon^{(n)} \right) \approx 1$$

* typical set sequences have their empirical entropy close to $H(X)$.
* $A_\epsilon^{(n)}$ is a small set of high probability with uniformly distributed sequences in it.

9,8,22 -1

Def: $A_\epsilon^{(n)}$ is typical set

$$A_\epsilon^{(n)} = \left\{ x^n : \left| \frac{1}{n} \log \frac{1}{p(x^n)} - H(X) \right| < \epsilon \right\}$$

ex. (ternary alphabet)

$\mathcal{X} = \{1, 2, 3\}$

or ~~$\{x_1, x_2, x_3\}$~~

or $\{a, b, c\}$

$p(a)$

$p(b)$

$p(c) = 1 - p(a) - p(b)$

$(x_1, x_2, \dots, x_n)$ — n

order not important

$(a, \dots, a, b, \dots, b, c, \dots, c)$ — $n(a)$, $n(b)$, $n(c) = n - n(a) - n(b)$

$$p(x^n) = p(a)^{n(a)} \cdot p(b)^{n(b)} \cdot p(c)^{n(c)}$$

x^n → $\underline{x}$

$n \cdot \frac{n(a)}{n} \to f_n(a)$: frequency of observing a

WLLN: $\frac{n(a)}{n} \xrightarrow{p} p(a)$ as $n \to +\infty$

as $n \to +\infty$

$$\sim p(a)^{np(a)} \cdot p(b)^{np(b)} \cdot p(c)^{np(c)}$$

$$= 2^{\log\left(P(a)^{np(a)}\right)} \cdot 2^{\log(\cdot)} \cdot 2^{\log(\cdot)}$$

$$= 2^{np(a)\log p(a)} \cdot 2^{np(b)\log p(b)} \cdot 2^{np(c)\log p(c)}$$

$$\Rightarrow P(x^n) \approx 2^{-nH(X)}$$

Corollaries

1. If $x^n \in A_\epsilon^{(n)}$:

$$\left| \frac{1}{n} \log \frac{1}{P(x^n)} - H(X) \right| < \epsilon$$

$$H(X) - \epsilon < -\frac{1}{n} \log p(x^n) < H(X) + \epsilon$$

↘ $\times (-n)$

$$-nH(X) - n\epsilon < \log p(x^n) < -nH(X) + n\epsilon$$

↘ log: monotonic

$$2^{-n(H(X)+\epsilon)} < P(x^n) < 2^{-n(H(X)-\epsilon)}$$

2/ $$\left| A_\epsilon^{(n)} \right| \leq 2^{n(H(X)+\epsilon)}$$

$$1 = \sum_{x^n \in \mathcal{X}^n} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)$$

$$\geq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)}$$

$$= 2^{-n(H(X)+\epsilon)} \left| A_\epsilon^{(n)} \right|$$

■

cardinality of $A_\epsilon^{(n)}$

3/ $$P(A_\epsilon^{(n)}) > 1 - \epsilon$$

From AEP property:

$$\underbrace{\frac{1}{n} \log \frac{1}{P(x^n)}}_{\text{empirical entropy}} \xrightarrow{P} H(X)$$

$$P\left(\left| \frac{1}{n} \log \frac{1}{P(x^n)} - H(X) \right| > \epsilon \right) \longrightarrow 0$$

$< \epsilon \to 1$ as $n \to +\infty$

we can define $n(\epsilon)$ and find the n that $n > n(\epsilon)$.

4/

$$1-\epsilon < P(A_\epsilon^{(n)}) = \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)$$

$$\leqslant \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)}$$

$$= 2^{-n(H(X)-\epsilon)} \left| A_\epsilon^{(n)} \right|$$

$$\Rightarrow (1-\epsilon)\ 2^{n(H(X)-\epsilon)} \leqslant \left| A_\epsilon^{(n)} \right|$$

$$\left| A_\epsilon^{(n)} \right| \approx 2^{n H(X)}$$

worst case → largest typical set

$$H(X) = \log |\mathcal{X}|$$

$$\Rightarrow \boxed{\left| A_\epsilon^{(n)} \right|_{max} = 2^{n \log |\mathcal{X}|}}$$

$B_\delta^{(n)}$ sets of smallest size and largest possible probability

↓ n can vary

$$P(B_\delta^{(n)}) > 1-\delta$$

$$0 < \delta < 0.5$$

---

ex.

$X \sim \text{Bernoulli}(0.8)$

↳ $P(X=1)$

$\delta = 0.4$

$n=1$

$$B_{0.4}^{(1)} = \{1\}$$

$$P(B_{0.4}^{(1)}) = 0.8$$ greater than $\overbrace{1-\delta}^{0.6}$ ⇒ ✓

$n=2$

$$B_{0.4}^{(2)} = \{11\}$$

$$P(B_{0.4}^{(2)}) = 0.64 > 0.6$$ ✓

$n=3$

$$B_{0.4}^{(3)} = \{111, 110\}$$

$$P^3(X=1) = 0.512 < 0.6 \quad !$$

we need to add more to the typical set.

$$P(110) = 0.128$$

↳ now satisfied $0.512 + 0.128 > 0.6$ ✓

$n=4$

$$B_{0.4}^{(4)} = \{1111, 1110, 1101\}$$

$$P(1111) = (0.64)^2 = 0.4096$$

$$P(1110) = 0.512 \times 0.2 = 0.1024$$

$$H(X) = -0.8 \log 0.8 - 0.2 \log 0.2 = 0.7219 \left[\frac{\text{bits}}{\text{symbol}}\right]$$

Defn:

$$A_\epsilon^{(n)} = \left\{ x^n : \left| \frac{1}{n} \log \frac{1}{p(x^n)} - H(X) \right| < \epsilon \right\}$$

$$H(X) - \epsilon < \frac{1}{n} \log \frac{1}{p(x^n)} < H(X) + \epsilon$$

$n = 1$

$\epsilon = 0.4$ pick $\{1\}$

$$0.7219 - 0.4 < \frac{1}{1} \log \frac{1}{0.8} < 0.7219 + 0.4$$

$$0.3219 \not< 0.2231 < 1.1219$$

$n = 2$

$$\not< \underbrace{\frac{1}{2} \log \frac{1}{(0.8)^2}}_{\log \frac{1}{0.8} = 0.2231} <$$

Sequences of all 1 will not make a typical set.

$n = 2$

$\{10\} \rightarrow$

$0.3219 < \underbrace{\frac{1}{2} \log \frac{1}{0.16}}_{1.3219}$ ~~1.1219~~

$n = 3$

111 is not typical.

let's try 110:

$$\underbrace{\frac{1}{3} \log \frac{1}{0.64 \times 0.2}}$$

$0.9886 = \hat{H}(X)$

$$A_{0.4}^{(3)} = \{110, 101, 011\}$$

$n = 4$

$$A_{0.4}^{(4)} = \{1110, 1101, 1011, 0111\}$$

$\hat{H}(X) = 0.6610$ ✓

also tried: $1100 \rightarrow \hat{H}(X) = 1.3219$ X

info. lect.

Constrained Optimization

- P(error)
- MAP
- max $P_{Detection}$ given $P_{False\ Alarm} < d$
- ML estimate
- MAP estimate, MMSE

---

minimize $f(\underline{x})$ $\quad \underline{x} \in \mathbb{R}^n$

assume m equality constraints

$h_1(\underline{x}) = 0, \ \dots, \ h_m(\underline{x}) = 0$

and p inequality constraints

$g_1(\underline{x}) \leqslant 0, \ \dots, \ g_p(\underline{x}) \leqslant 0$

$\min f(\underline{x})$

$\underline{x} \in \mathbb{R}^n$: $h_1(\underline{x}) = 0, \dots, h_m(\underline{x}) = 0$

$g_1(\underline{x}) \leqslant 0, \dots, g_p(\underline{x}) \leqslant 0$

---

ex.

$$\min f(x_1, x_2) \qquad f(x_1, x_2) = 3x_1^2 + 2x_1x_2 + 3x_2^2 - 20x_1 + 4x_2$$

(x_1, x_2): $x_1 \geqslant 0$

$x_2 \geqslant 0$

$$\left.\begin{aligned} \frac{df}{dx_1} &= 6x_1 + 2x_2 - 20 = 0 \\ \frac{df}{dx_2} &= 2x_1 + 6x_2 + 4 = 0 \end{aligned}\right\} \longrightarrow \left\{\begin{aligned} x_2 &= -2 \\ x_1 &= 4 \end{aligned}\right.$$

$(x_1, x_2) = (4, -2)$

$f(4, -2) = -44$ (Global minimum)

→ violating the constraints!

If both constraints are ON, then set $x_2 = 0$

Cause this is the first point satisfying the constraint on x_2

when set to equality, the constraint is active.

$\Rightarrow$ $x_1 = 10/3$

OR

~~$x_1 = -2$~~

Constrained Solution

$(x_1^*, x_2^*) = (\frac{10}{3}, 0)$

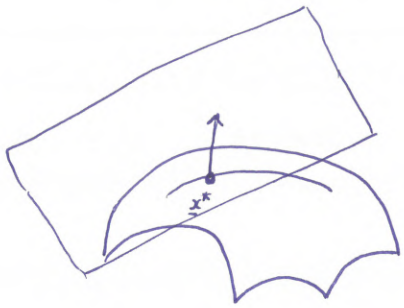
$f(\frac{10}{3}, 0) = -33.33$

$f(-2, 0) = 46$

Geometric Point of View

$\forall i: \quad \nabla f(\underline{x}^*) = \lambda_i \nabla h_i(\underline{x}^*)$

$\Rightarrow \quad \nabla f(\underline{x}^*) = \sum_{i=1}^{m} \lambda_i \nabla h_i(\underline{x}^*)$

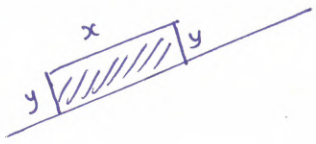


Lagrangian: $\mathcal{L}(\underline{x}, \lambda_1, \dots, \lambda_m) = f(\underline{x}) + \sum_{i=1}^{m} \lambda_i h_i(\underline{x})$

Lagrange multipliers

---

ex.1



$x + 2y = R$

$\text{area} = x \cdot y = f(x,y)$

$\max f$

$(x,y): \quad x + 2y = R$

$\mathcal{L}(x,y,\lambda) = xy + \lambda(x + 2y - R)$

$$\left.\begin{aligned} y + \lambda &= 0 \\ x + 2\lambda &= 0 \end{aligned}\right\} \quad \left.\begin{aligned} y &= -\lambda \\ x &= -2\lambda \end{aligned}\right\} \Rightarrow \quad -2\lambda + (-2\lambda) = R$$

$\lambda = -\frac{R}{4}$

$$\left.\begin{aligned} y &= \frac{R}{4} \\ x &= \frac{R}{2} \end{aligned}\right\}$$

$$\left.\begin{aligned} R &= 100 \\ y &= 25 \\ x &= 50 \end{aligned}\right\}$$

$S^* = 1250$

ex. $f(\underline{x}) = x_1^2 + x_1x_2 + x_2x_3$

$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

maximize such that $x_1 + x_2 = 4$.

$$\mathcal{L}(\underline{x}, \lambda) = x_1^2 + x_1x_2 + x_2x_3 + \lambda(x_1 + x_2 - 4)$$

$\frac{\partial \mathcal{L}}{\partial x_1} = 0 \rightarrow 2x_1 + x_2 + \lambda = 0$ $\qquad x_1 = -\frac{\lambda}{2}$

$\frac{\partial \mathcal{L}}{\partial x_2} = 0 \rightarrow x_1 + x_3 + \lambda = 0$ $\qquad x_3 = -\frac{\lambda}{2}$

$\frac{\partial \mathcal{L}}{\partial x_3} = 0 \rightarrow x_2 = 0$

$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \rightarrow x_1 + x_2 = 4$ $\qquad x_1 = 4$

$x_3 = 4$

Constrained solution:

$(x_1^*, x_2^*, x_3^*) = (4, 0, 4)$

---

ex. 3 (maximum entropy)

$x_1, \ldots, x_n$ outcomes

$p(x_1), \ldots, p(x_n)$ prob. assignments

max $H(X)$ over choices for $p(x_i)$ such that $\sum_{i=1}^{n} p(x_i) = 1$

$$\mathcal{L}\left(p(x_1), p(x_2), \ldots, p(x_n), \lambda\right) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i) + \lambda\left(\sum_{i=1}^{n} p(x_i) - 1\right)$$

n equations $\begin{cases} \dfrac{\partial \mathcal{L}}{\partial p(x_i)} = -\ln p(x_i) - 1 + \lambda = 0 \\ i = 1, \ldots, n \end{cases}$

$$\frac{\partial \mathcal{L}}{\partial \lambda} \overset{=0}{\Longrightarrow} \sum_{i=1}^{n} p(x_i) = 1$$

$$p(x_i) = e^{-1+\lambda}$$

$$\sum_{i=1}^{n} e^{-\lambda+1} = 1 \Rightarrow \boxed{e^{-\lambda+1} = \frac{1}{n}}$$ UNIFORM Distribution!

ex.4

$X = \{1, 2, \ldots, n\}$

$p(1), p(2), \ldots, p(n)$

max $H(X)$ over p. $\sum_{i=1}^{n} p(i) = 1$, $\sum_{i=1}^{n} i\,p(i) = m$

$$\mathcal{L}(p_1, \ldots, p_n, \lambda, \mu) = -\sum_{i=1}^{n} p(i) \ln p(i) + \lambda\left(\sum_{i=1}^{n} p(i) - 1\right) + \mu\left(\sum_{i=1}^{n} i\,p(i) - m\right)$$

$$-\ln p(i) - 1 + \lambda + \mu \cdot i = 0$$

$$\sum_{i=1}^{n} p(i) = 1$$

$$\sum_{i=1}^{n} i\,p(i) = m$$

$$p(i) = e^{-1+\lambda} \cdot e^{\mu \cdot i}$$

$$e^{-1+\lambda} \sum e^{\mu i} = 1 \rightarrow e^{-1+\lambda} = \frac{1}{\sum_{i=1}^{n} e^{\mu i}}$$

$$\Rightarrow \sum q^i - 1 = \frac{1 - q^{i+1}}{1 - q} - 1 \qquad q^i$$

9/13/22-3

$$\sum_{j=1}^{n} j \frac{e^{\mu j}}{\sum_{i=1}^{n} e^{\mu i}} = m$$

$$\frac{d}{d\mu} \underbrace{\left(\sum_{j=1}^{n} e^{\mu j} \right)}_{\frac{e^{\mu}(1-e^{\mu n})}{1-e^{\mu}}} = \sum_{j=1}^{n} j e^{\mu j}$$

info. lect.

Inequality Constraints

ex.1 $\min f(x,y)$ $\qquad f(x,y) = 3x^2 + 2xy + 3y^2 - 20x + 4y$

(x,y) : $x \geq 0$, $y \geq 0$

1. check if any local min or max

$$\frac{\partial f}{\partial x} = 6x + 2y - 20 = 0 \qquad (1)$$

$-3 \times \left(\frac{\partial f}{\partial y} = 2x + 6y + 4 = 0 \right)$ $\qquad (2)$

$$-16y - 32 = 0$$

$$\boxed{y = -2} \Rightarrow \boxed{x = 4}$$ ★ NOT A SOLUTION!

$f(4,-2) = 48 - 16 + 12 - 80 - 8 = 0$

2. set $x = 0$, $y = \frac{-2}{3}$ (from (2))

★ NOT A SOLUTION!

3. set $y = 0$, $x = \frac{10}{3}$ (from (1))

★ POTENTIAL SOLUTION! $\Rightarrow f\left(\frac{10}{3}, 0\right) = -\frac{100}{3}$

4. $x = 0$, $y = 0$

$f(0,0) = 0$

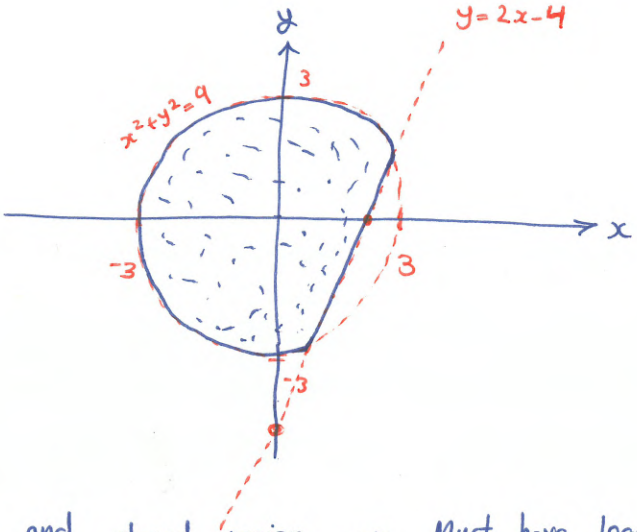
$\Rightarrow$ solution $\begin{cases} (x,y) = \left(\frac{10}{3}, 0\right) \\ f(x,y) = -\frac{100}{3} \end{cases}$

9,20,22-1

ex.2 $\min f(x,y)$

$$(x,y): \begin{cases} x^2+y^2 \leq 9 \\ 2x-y \leq 4 \end{cases}$$

$$f(x,y) = 3x^2 + 4y^2 + 6xy - 6x - 8y$$



continuous f and closed region $\Rightarrow$ Must have local min/max.

1. $$\frac{\partial f}{\partial x} = 6x + 6y - 6 = 0 \qquad (3)$$

$$-1 \times \left(\frac{\partial f}{\partial y} = 8y + 6x - 8 = 0 \right) \qquad (4)$$

$$-2y + 2 = 0$$

$$\boxed{y = 1} \Rightarrow \boxed{x = 0}$$

★ Possible Solution $(0,1)$: $f(0,1) = -4$

2. Both constraints are active.

$$\nabla f_{(x,y)}(x,y) = \lambda \nabla g_1 + \mu \nabla g_2$$

involve $(3,4)$

$$6x + 6y - 6 = \lambda(2x) + \mu(2)$$

$$8y + 6x - 8 = \lambda(2y) + \mu(-1)$$

and $x^2 + y^2 = 9$, $2x - y = 4$

4 equations
4 unknowns

$\Rightarrow$ $(x_1 = 2.6770 \quad , \quad y_1 = 1.3541)$ both ON the

$(x_2 = 0.5230 \quad , \quad y_2 = -2.9541)$ boundary !

$f(x_1, y_1) = 23.6881$

$f(x_2, y_2) = 46.9522$

$\Rightarrow$ Final solution: $(0, 1)$

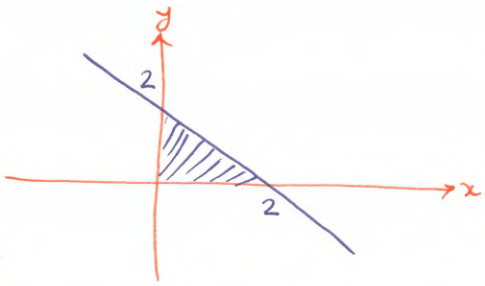
---

ex. 3

$$\max f(x, y)$$

$$\begin{cases} x + y \leq 2 \\ x \geq 0, \quad y \geq 0 \end{cases}$$

$$f(x,y) = \ln\left(1 + \frac{x}{5}\right) + \ln\left(1 + \frac{y}{2}\right)$$



1. $\frac{\partial f}{\partial x} = \frac{1}{5+x}$

$\frac{\partial f}{\partial y} = \frac{1}{2+y}$ $\Big\}$ → no solution!

options:

$x+y=2$, then check $x \geq 0$, $y \geq 0$.

$x=0$ or $y=0$ or both

2. set $x+y=2$ (use Lagrange multipliers with equality constraints)

$$\frac{1}{5+x} = \lambda \longrightarrow x = \frac{1-5\lambda}{\lambda}$$

$$\frac{1}{2+y} = \lambda \qquad = \underbrace{\frac{1}{\lambda}}_{\mu} - 5$$

$$y = \underbrace{\frac{1}{\lambda}}_{\mu} - 2$$

$$\Rightarrow \quad \mu - 5 + \mu - 2 = 2 \Rightarrow \boxed{\mu = 9/2}$$

$$\Longrightarrow \quad x = -1/2, \quad y = 5/2 \quad \text{NOT A SOLUTION!}$$

3. set $x=0$

$f(x,y)$ increasing due to y $\Rightarrow$ $y=2$ (max of $f(0,y)$)

$f(0,2) = \ln(2)$

4. set $y=0$ $\xrightarrow[\text{reasoning}]{\text{the same}}$ $f(2,0) = \ln(7/5)$

5. $f(0,0) = 0$

$\Rightarrow$ Final Solution $(x,y) = (0,2)$

if $x + y \leq 5$

$\Rightarrow$ 2. $f(1,4) = \ln\left(\frac{6}{5}\right) + \ln(3) = 1.28$

3. $f(0,5) = 1.2528$

4. $f(5,0) = 0.6931$

$\Rightarrow$ Final solution $(x,y) = (1,4)$

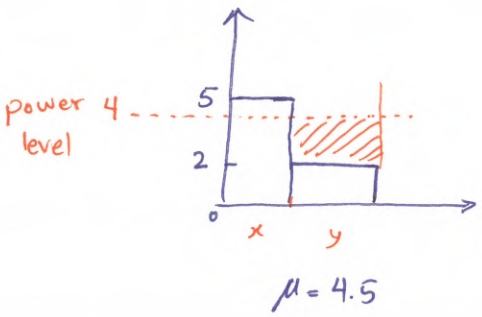
info. lect.

$$\max f(x,y) = \ln\left(1+\frac{x}{5}\right) + \ln\left(1+\frac{y}{2}\right)$$

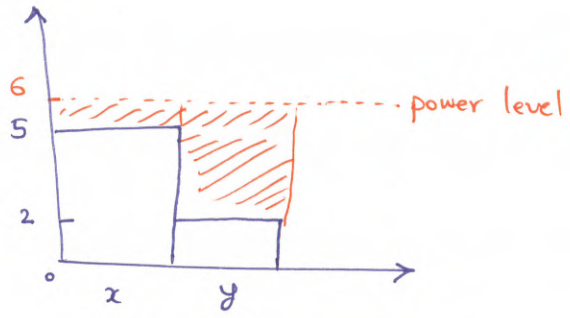
(x,y): $x \geq 0$, $y \geq 0$

$x+y \leq 5$

$x+y \leq 2$



$\mu = 4.5$



$\mu = 6$

Kuhn - Tucker Conditions

$$\min f(\underline{x}) \qquad \underline{x} \in \mathbb{R}^n$$

$\underline{x}$: $h_1(\underline{x}) = 0$ ⋮ $h_m(\underline{x}) = 0$

$g_1(\underline{x}) \leq 0$ ⋮ $g_p(\underline{x}) \leq 0$

Def'n x^* is regular if $\nabla_{\underline{x}} h_i(\underline{x})$ and $\nabla_{\underline{x}} g_j(\underline{x})$ for all i,j are linearly independent.

Theorem: Let $\underline{x}^*$ is a local minimum.

and it is a regular point.

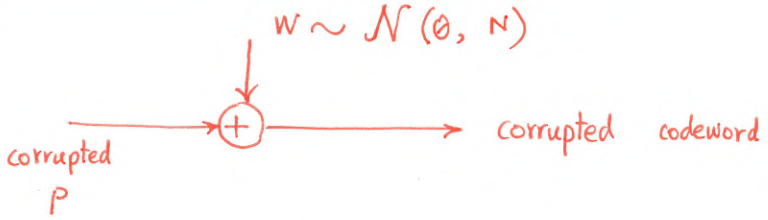
then there exist $\underline{\lambda}$ and $\underline{\mu}$ such that:

$$\nabla_{\underline{x}} f(\underline{x}^*) + \nabla_{\underline{x}} h^T(\underline{x}^*) \cdot \underline{\lambda} + \nabla_{\underline{x}} g^T(\underline{x}^*)\, \underline{\mu} = 0 \quad (1)$$

and $$g^T(\underline{x}^*) \cdot \underline{\mu} = 0 \quad \text{for } \underline{\mu} \leq 0 \quad (2)$$

---

ex.



$$C = \frac{1}{2} \ln\left(1 + \frac{P}{N}\right)$$

Capacity

n channels

$$\sum_{i=1}^{n} P_i = P$$

$$P_i \geq 0$$

$$\max_{\{P_i\}} \sum_i \frac{1}{2} \ln\left(1 + \frac{P_i}{N_i}\right)$$

$\underline{P}$: $\sum P_i = P$

$P_i \geq 0$, $i = 1, \ldots, n$

$$\nabla_{\underline{P}} \left\{ \sum_i \frac{1}{2} \ln\left(1 + \frac{P_i}{N_i}\right) + \lambda \nabla_{\underline{P}} \left\{ \sum_i P_i - P \right\} \right.$$

$$\left. + \nabla_{\underline{P}} \left\{ \sum_i P_i \cdot \mu_i \right\} = 0 \right.$$

$$\sum_i \mu_i P_i = 0$$

1) $P_i > 0$ $\quad (\mu_i = 0)$

$$\frac{1}{2} \frac{1}{N_i + P_i} + \lambda = 0$$

$$P_i = \underbrace{-\frac{1}{2\lambda}}_{\kappa} - N_i$$

$\kappa \rightsquigarrow$ (kappa)

2) $P_i = 0$ $\quad (\mu_i \leq 0)$

$$\underbrace{\frac{1}{2} \frac{1}{N_i + P_i} + \lambda}_{} + \mu_i = 0$$

$$-\frac{1}{2} \frac{1}{N_i + P_i} - \lambda \leq 0$$

$$P_i > \frac{1}{-2\lambda} - N_i \quad \rightarrow \quad P_i > \kappa - N_i$$

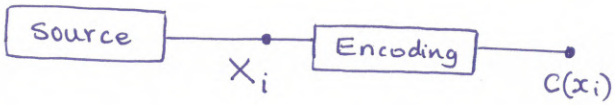
kappa

$(\overset{\parallel}{0})$ assumption

info. lect.

Data Compression (Compaction)

- remove redundancies



X_i : $x_i \in \mathcal{X}$ (finite), $p(x_i)$

$(X_1, X_2, \dots, X_N)$

Code ≡ codeword

D: coder alphabet {ex. {0, 1}}

l_i = length of $C(x_i)$

$|\mathcal{X}| > |D|$

- memoryless source : X_i are independent

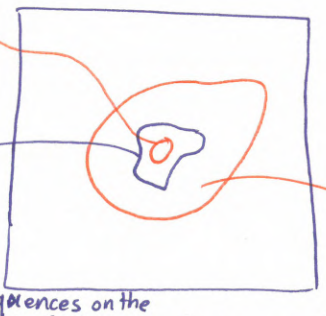
ex. MPEG : encodes difference between two frames

a codeword that is a prefix for another codeword

★ Prefix Codes (Instantaneous Codes)

Uniquely Decodable Codes

have non-singular extension



All Possible codes for sequence $X^n = (X_1, \dots, X_n)$

Nonsingular codes : assign individual code $C(x_i)$ to every x_i (ONE-to-ONE)

ex. $\mathcal{X} = \{1, 2, 3, 4\}$

x	C(x)
1	0
2	010
3	01
4	10

Extension of Code.

code for sequence (1,2,3,4) → C(1).C(2)C(3)C(4)

concatenated symbols 1,2,3,4 (encoded)

0 010 01 10

decoding:

1 1 4 3 4

1 2 3 4

NO ONE-to-ONE Decoding!

ex. Morse code:

$|X| = 28$

$|D| = 4$

$\{ \cdot, —, \text{letter space}, \text{word space} \}$

→ "e" is assigned "$\cdot$".

→ "q" is assigned "$\cdot\cdot — —$".

Kraft Inequality

① A prefix code with alphabet D and codeword's length's $\ell_1, \dots, \ell_{max}$, then $\{\ell_i\}$ satisfy:

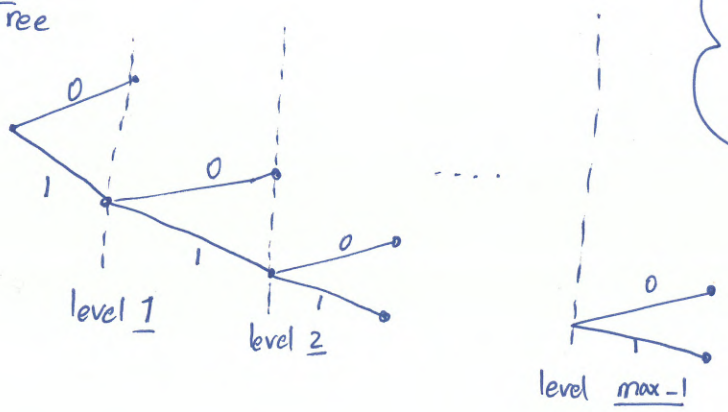
$$\sum_{i=1}^{max} |D|^{-\ell_i} \leq 1$$

(choose $|D|=2$)

② If codeword lengths $\ell_1, \dots, \ell_{max}$ satisfy the inequality, then a prefix code can be constructed with codewords of lengths $\ell_1, \dots, \ell_{max}$.

proof: ① choose $D=2$

Tree



$2^{\ell_{max}}$ ≡ max number of codewords

not necessarily prefix

codewords removed ? (at each level)

$$2^{\ell_{max}-\ell_1} + 2^{\ell_{max}-\ell_2} + \cdots + 2^{\ell_{max}-\ell_k} + \cdots$$

$$= \sum_{k=1}^{\ell_{max}} 2^{\ell_{max}-\ell_k} \leq 2^{\ell_{max}}$$

$$\Rightarrow \sum_{k=1}^{max} 2^{-\ell_k} \leq 1$$

$$\bar{\ell} = \sum_i \ell_i \cdot p_i$$

$p_i = p(x_i)$

average codeword length (description length)

---

minimize $\sum_{i=1}^{n} \ell_i p_i$ (constrained optimization)

$\ell_i : \sum 2^{-\ell_i} \leq 1$

$$\nabla_{\ell_i} \left\{ \sum_{i=1}^{n} \ell_i p_i \right\} = \lambda \nabla_{\ell_i} \left\{ \sum_{i=1}^{n} 2^{-\ell_i} - 1 \right\}$$

$$\sum_{i=1}^{n} 2^{-\ell_i} = 1$$

$2^{-\ell_i} \rightarrow e^{-\ell_i \ln 2}$

$$P_i = \lambda (-\ln 2) \underbrace{e^{-l_i \ln 2}}_{2^{-l_i}}$$

$$\Rightarrow 2^{-l_i} = -\frac{P_i}{\lambda \ln 2}$$

, $\frac{1}{-\lambda \ln 2} \sum_{i=1}^{n} P_i = 1$

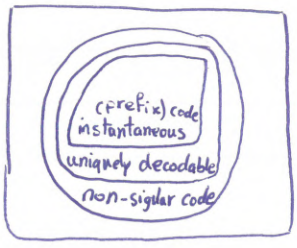
$$\Rightarrow \lambda = -\frac{1}{\ln 2}$$

, $P_i = 2^{-l_i}$

, $\boxed{l_i = -\log_2 P_i}$

Model selection Based on MDL (Minimum Description Length)

info. lect.



$|\mathcal{X}|^n$: # of all possible sequences of length n consisting of alphabet $\mathcal{X}$

$$(x_1, \cdots, x_n) \longrightarrow C(x_1, \cdots, x_n)$$

$\Rightarrow$ block code $\Leftarrow$

Kraft Inequality

- necessary and sufficient conditions to assign prefix codes

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

$$\min \; E[\ell(X)]$$

$$\ell_i : \sum_{i=1}^{n} 2^{-\ell_i} = 1 \quad \Rightarrow \quad \ell_i = -\log_2 p_i$$

expected value $\Rightarrow$ $\sum_{i=1}^{n} p_i \ell_i = -\sum p_i \log p_i = H(X)$

Limit of Compression

If p_i have base 2: $\left\{\frac{1}{2}, \frac{1}{4}, \cdots\right\} \rightarrow \ell_i$: integer-valued

In general $\ell_i = \lceil -\log p_i \rceil$

↳ Is it still prefix code? check Kraft Inequality!

Yes, it's possible! ✓

$$\sum_{i=1}^{n} 2^{-\ell_i} = \sum_{i=1}^{n} 2^{-\lceil -\log p_i \rceil} \leq \sum_{i=1}^{n} 2^{-\log \frac{1}{p_i}} = \sum_{i=1}^{n} p_i = 1$$

9, 29, 22 - 1

- Prefix codes have $\mathcal{L} = E[\ell(X)] \geq H(X)$.

proof:

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

$$\log \sum 2^{-\ell_i} \times \frac{P_i}{P_i} \leq 0$$

$y_i \triangleq \frac{2^{-\ell_i}}{P_i}$

$$\log \left(E_{P_i}(y_i) \right) \leq 0$$

log is concave

$$E_{P_i}(\log y_i) \leq \log \left(E_{P_i}(y_i) \right) \leq 0$$

$$\underbrace{-\ell_i - \log P_i}$$

$$H(X) - \underbrace{E[\ell(X)]}_{L}$$

$$\Rightarrow \quad E[\ell(X)] \geq H(X)$$

$P_i \times$

$$\left(-\log P_i \leq \lceil -\log P_i \rceil \leq -\log P_i + 1 \right)$$

$$\underbrace{-\sum P_i \log P_i}_{H(X)} \leq \mathcal{L} \leq H(X) + 1$$

for symbol-by-symbol coding.

block coding

$$\ell(x_1, \cdots, x_n) = \lceil -\log p(x_1, \cdots, x_n) \rceil$$

$$H(X^n) \leq \underbrace{\sum_{x^n \in \mathcal{X}^n} p(x^n) \lceil -\log p(x_1, \cdots, x_n) \rceil}_{L(X^n)} \leq H(X^n) + 1$$

$$\frac{1}{n} H(X^n) \leq \frac{1}{n} L(X^n) \leq \frac{1}{n} H(X^n) + \frac{1}{n}$$

by increasing n, we can tighten the sandwich bound as far as we need.

Huffman Coding

$X \quad \mathcal{X} = \{x_1, \cdots, x_n\}$

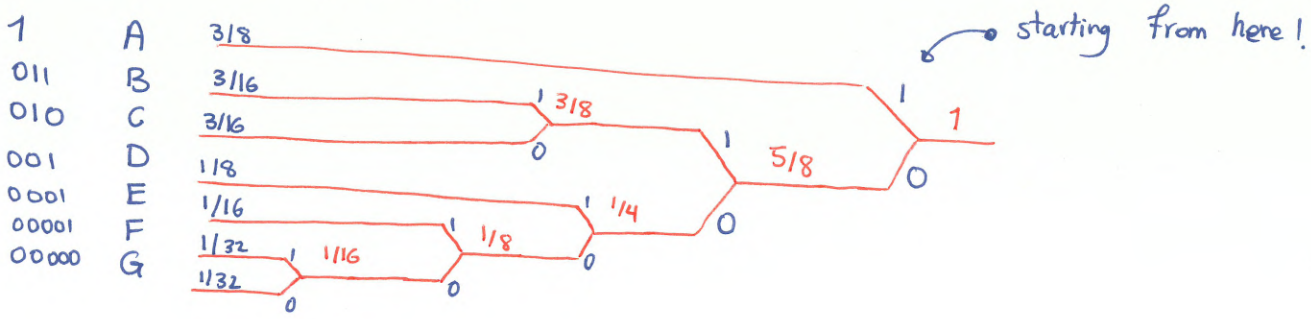
$$P_1 \geq P_2 \geq \cdots \geq P_n$$

- combine two outcomes with smallest probabilities

x^*_{n-1} $\qquad P_{n-1} + P_n = P(x^*_{n-1})$

- this will lead to a new C^* with $\ell^* \leq \ell$

ex.1 $X \in \{A, \dots, G\}$



$H(X) = 2.28 \left[\frac{\text{bits}}{\text{symbol}}\right]$

average codeword length $\equiv L(X) = 1 \cdot \frac{3}{8} + 3\left(\frac{3}{16} + \frac{3}{16} + \frac{1}{8}\right) + 4\,\frac{1}{16} + 5\left(\frac{1}{32} + \frac{1}{32}\right) =$

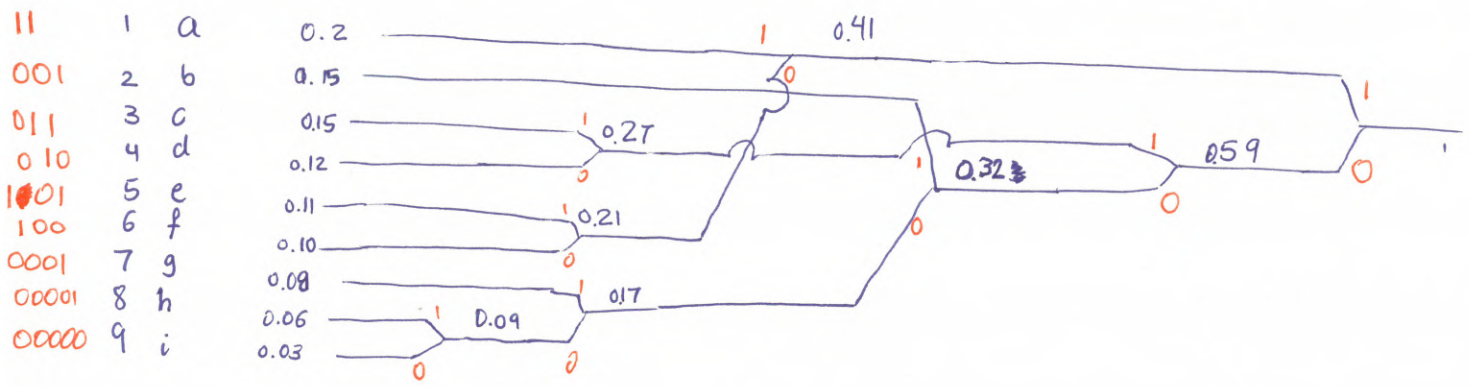
$= \frac{3}{8} + \frac{12}{8} + \frac{2}{8} + \frac{5}{16} = \frac{39}{16} = 2.4375 \left[\frac{\text{bits}}{\text{symbol}}\right]$

[info. lect.]

- if distribution changes, you have to run Huffman from scratch!

but block codes/stream codes like arithmetic coding can deal with this issue easier.

ex.2 Huffman code



$$L = E[\ell(X)] = 2(0.2) + 3(0.63) + 4(0.08) + 5(0.09) = 3.06$$

[bit/symbol]

$$H(X) = 3.02 \quad \text{bits/symbol}$$

---

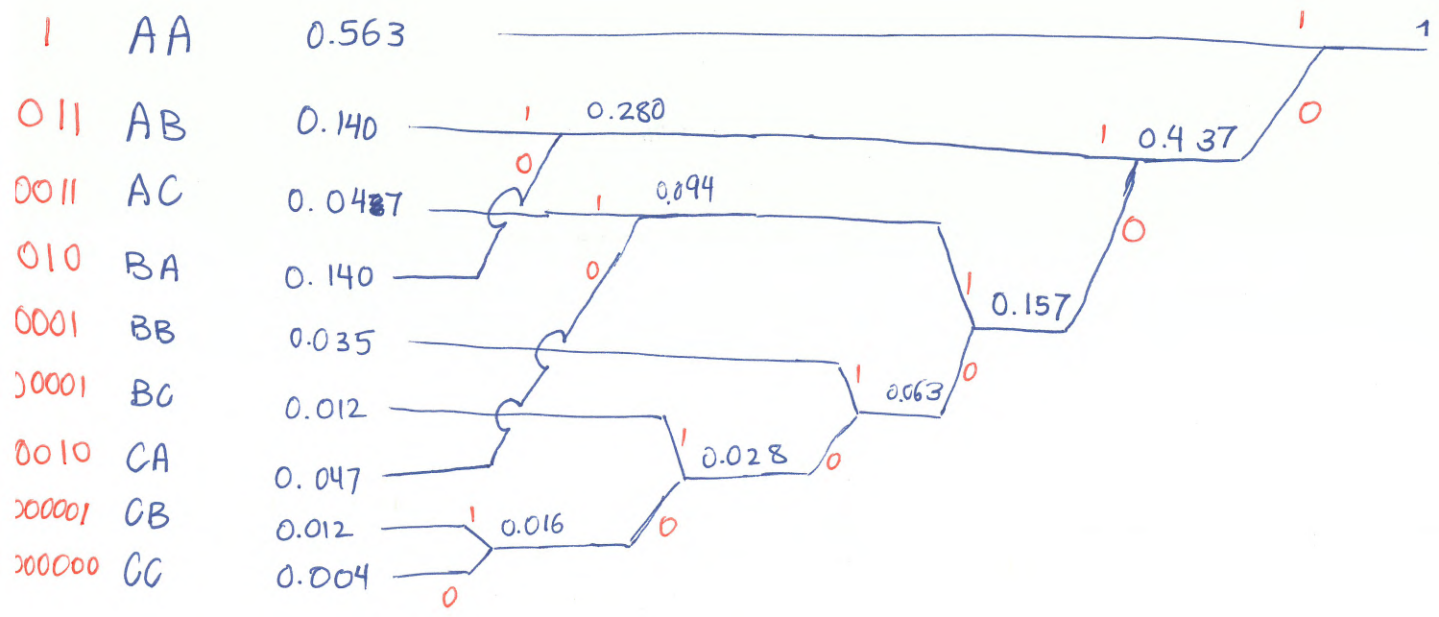
ex. 3 (block code)

A	3/4
B	3/16
C	1/16

Huffman
1
01
00

$$L(X) = E[\ell(X)] = 1.25 \quad , \quad H(X) = 1.012$$

what if we plan to code blocks of length two?



$$\frac{L}{2} = \frac{E[\ell(x)]}{2} = 1.037 \quad \left[\frac{\text{bits}}{\text{symbol}}\right]$$

↳ closer to the Entropy!

Length of alphabet x_i : $\ell(x_i) = \lceil -\log p_i \rceil$

$\Rightarrow -\log p_i \leq \ell(x_i) \leq -\log p_i + 1$

$E_{\sim q_i}$ or $\sum q_i$

q_i: true pmf

p_i: model pmf

$D(q\|p) + H(x) \leq L \leq D(q\|p) + H(x) + 1$

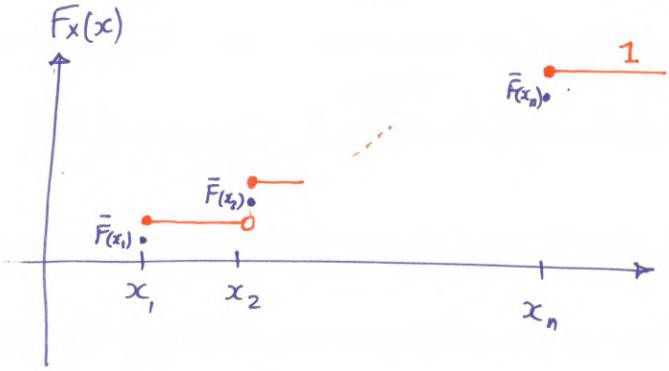
Shannon Codes

$$\ell(x_i) = \lceil -\log p_i \rceil + 1$$

$$H(X) < L < H(X) + 2$$

$x_1, \dots, x_n$

$p_1, \dots, p_n$



$$\bar{F}(x_i) = F(x_{i-1}) + \frac{p_i}{2}$$

in general: real-valued.

$$\lfloor \bar{F}(x_i) \rfloor_{\ell(x_i)}$$

$$\bar{F}(x_i) - \lfloor \bar{F}(x_i) \rfloor_{\ell(x_i)} = 0.b_1 b_2 b_3 \dots = \sum_{k=1}^{+\infty} b_k \left(\frac{1}{2}\right)^k \qquad b_k \in \{0,1\}$$

$$\sum_{k=1}^{\ell(x_i)} b_k \left(\frac{1}{2}\right)^k$$

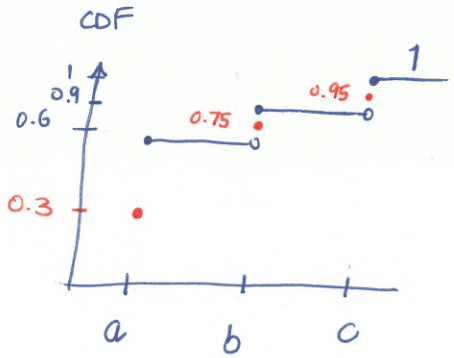
$$= \sum_{k=\ell(x_i)+1}^{+\infty} b_k \frac{1}{2^k} \leq \sum_{m=0}^{+\infty} \frac{1}{2^{m+1-\ell(x_i)}} = \frac{1}{2^{\ell(x_i)}} = \frac{1}{2^{\lceil \log p_i \rceil + 1}}$$

$m = k - \ell(x_i) - 1$

$$\leq \frac{1}{2^{-\log p_i}} \cdot \frac{1}{2} = \frac{p_i}{2}$$

10, 4, 22 – 2

1	a	0.6
01	b	0.3
00	c	0.1



$H(X) = 1.30$ bit/sym

$E(\ell(x))_{\text{Huffman}} = 1.40$ bit/sym

$\bar{F}(a) = 0.3 = 0.\frac{1}{2} + 1\,\frac{1}{2^2} + 0.\frac{1}{2^4} + \cdots$

$\bar{F}(b) = 0.75 = 1.\frac{1}{2} + 1\,\frac{1}{2^2}$

$\bar{F}(c) = 0.95 = 1\,\frac{1}{2} + 1\,\frac{1}{2^2} + 1\,\frac{1}{2^3} + \cdots$

$\ell(a) = \lceil \underbrace{-\log 0.6}_{0.7370} \rceil + 1 = 2$

$\ell(b) = \lceil \underbrace{-\log 0.3}_{1.7370} \rceil + 1 = 3$

$\ell(c) = \lceil \underbrace{-\log 0.1}_{3.3219} \rceil + 1 = 5$

Shannon codes:

a	01
b	110
c	11110

$L = 2(0.6) + 3(0.3) + 5(0.1) = 2.6$ bit/sym $< H(X) + 2$

Arithmetic Coding

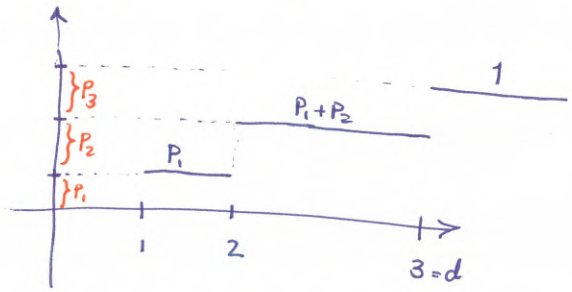
- recursive
- inspired by Shannon Coding

X

$\mathcal{X} = \{a_1, a_2, \dots, a_d\}$

$P_1 \quad P_2 \quad \dots \quad P_d$

$d=3$:



to represent 1 any value between $[0, P_1]$ can be selected.

to represent 2 " " " $[P_1, P_2+P_1]$ " " " .

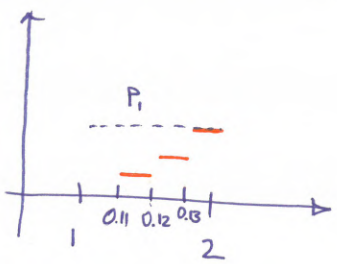
to represent 3 " " " $[1-P_3, 1]$ " " " .

1 → 0.1
2 → 0.2
3 → 0.3

---

pair coding: (x_1, x_2) ⟶ $3^2 = 9$ possible combinations.

0.11 $P_1 \cdot P_1$
0.12 $P_1 \cdot P_2$
0.13
0.21
0.22
0.23
0.31
0.32
0.33



to respresent 0.12 any value between $[P_1P_1, P_1P_1+P_1P_2]$ can be chosen.

$$F_{X^n}(x^n) = P(X_1 < x_1) + P(X_1 = x_1, X_2 < x_2) + \cdots + P(X_1 = x_1, X_2 = x_2, \cdots, X_n < x_n)$$

X^n : sequence of length n.

1. n-length string

$$(x_1, x_2, \cdots, x_n)$$

$$\ell(x_n) = \lceil -\log p(x^n) \rceil + 1$$

2. Given an interval to encode

assign midpoint to x^n, encode the midpoint.

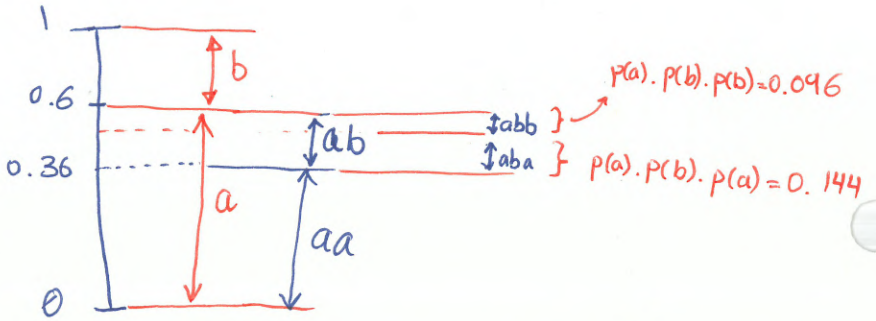
ex.

$a < b$

$$X = \begin{cases} a, & P(a) = 0.6 \\ b, & P(b) = 0.4 \end{cases}$$

$n = 3$

(a, b, b)



midpoint of $[0.504, 0.6) = 0.552$

$$\ell(a,b,b) = \lceil -\log 0.096 \rceil + 1 = 5 \text{ bits}$$

$0.552 = \underbrace{1\ \frac{1}{2} + 0\ \frac{1}{4} + 0\ \frac{1}{8} + 0\ \frac{1}{16} + 1\ \frac{1}{32}}_{\text{length is 5 bits from previous page.}}$

Final code of abb : 10001

Decoding: $(10001)_2 = (0.53125)_{10}$

question 1: Is 0.53125 in [0, 0.6) or [0.6, 1] ?

$\Rightarrow$ a

question 2: Is 0.53125 in [0, 0.36) or [0.36, 0.6] ?

$\Rightarrow$ ab

question 3: Is 0.53125 in [0.36, 0.504) or [0.504, 0.6]

$\Rightarrow$ abb

$$F_{X^3}(a,b,b) = \underbrace{P(X_1<a)}_{0} + \underbrace{P(X_1=a,\ X_2<b)}_{P(X_1=a)\cdot P(X_2<b)} + P(X_1=a, X_2=b, X_3<b)$$

memoryless source

0.5×0.6

$0.6 \times 0.4 \times 0.6 = 0.504$

* Mackay Laplace / Dirichle model

to have adaptive conditional probability model!

arithmetic: $\mathcal{L} \leq H(X^n) + 2$

as $n \to \infty$ arithmetic buys us alot in comparison to Huffman.

~~by having n. at limits arithmetic better.~~

Project. } 2 texts

2 p and q

Chapters 5 and 6.

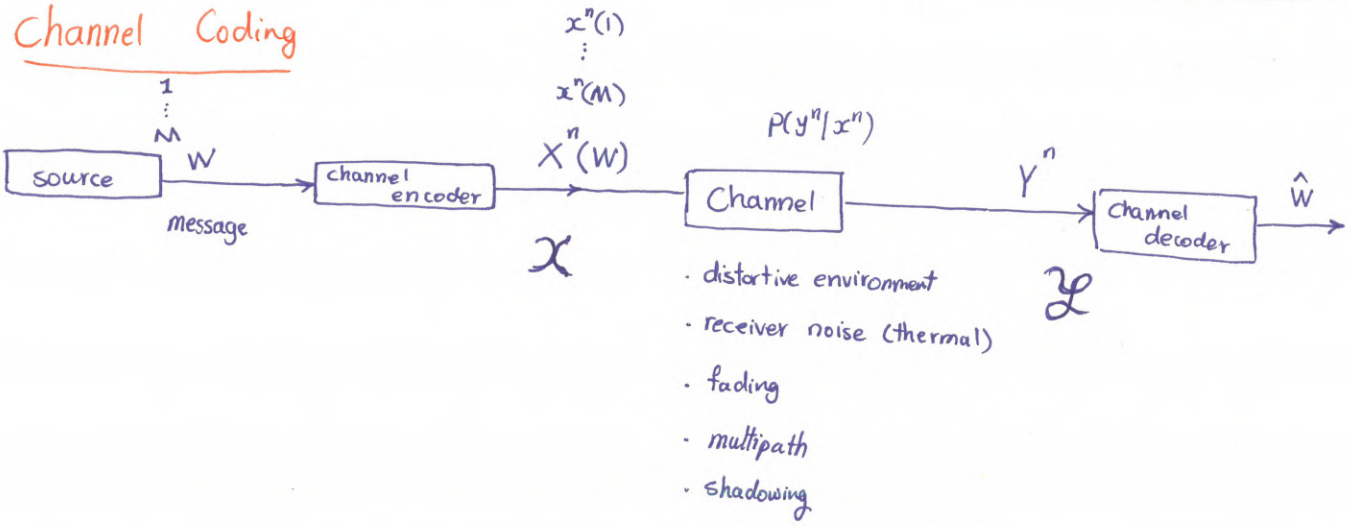
info. lect.

project { 2 texts: Alice in Wonderland (Walrus) + sonnet 8 of Shakespear
Huffman + Arithmetic + Lempel Ziv (universal coding)

area of research: Joint Source/channel Coding.

10, 11, 22 - 1

Channel Coding



the whole block is ideally an error-free channel!

★ if channel is memoryless: $p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$

ex, $n = 1000$

$P(\text{flip}) = P(Y \neq X) = 0.1$

intuition: about 900 bits will be transmitted error-free.

capacity: $1 - H_b(0.1) = \approx 0.531$ bits/channel use

measure of interest: Mutual Information

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

↳ concave in $P_X(x)$

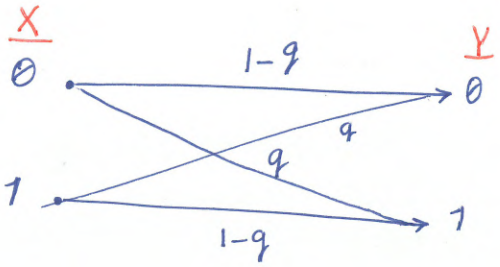
★ we can maximize $I(X;Y)$.

10, 13, 22 - 1

2 important channels : { - Discrete channel
- Gaussian channel

Discrete Channels

1. Binary Symmetric Channel



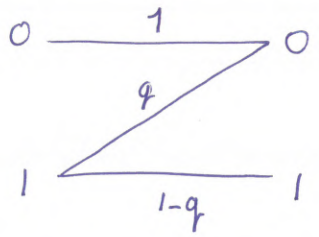
$$\begin{matrix} P_X(0) \\ P_X(1) \end{matrix} \xrightarrow{P_Y(y) = \sum_{x_i \in \mathcal{X}} \overbrace{P(y|x_i)P(x_i)}^{P(y,x_i)}} \begin{bmatrix} P_Y(0) \\ P_Y(1) \end{bmatrix} = \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(0|1) \\ P_{Y|X}(1|0) & P_{Y|X}(1|1) \end{bmatrix} \begin{bmatrix} P_X(0) \\ P_X(1) \end{bmatrix}$$

Transition Matrix

$$\begin{bmatrix} 1-q & q \\ q & 1-q \end{bmatrix}$$

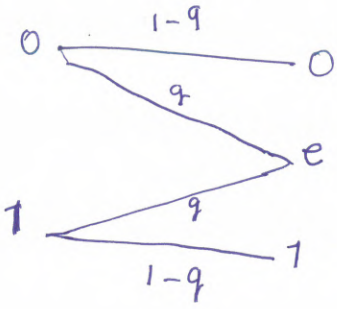
Transition Matrix of Binary Symmetric Channel

2. $\mathbb{Z}$ - Channel



$$\begin{bmatrix} 1 & 0 \\ q & 1-q \end{bmatrix}$$

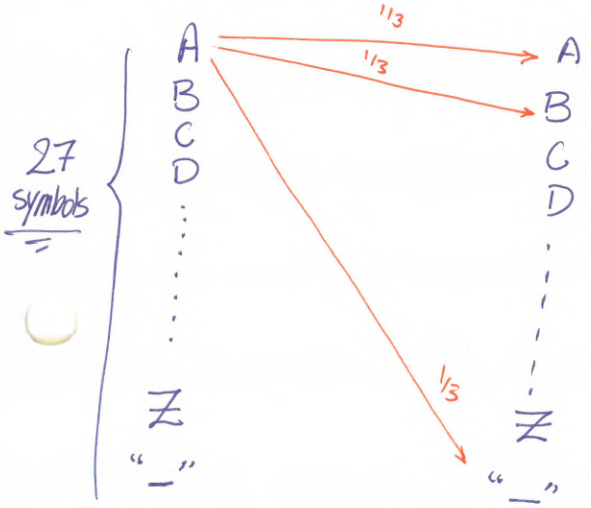
3. Erasure Channel



$$\begin{bmatrix} P_Y(0) \\ P_Y(e) \\ P_Y(1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1-q & 0 \\ q & q \\ 0 & 1-q \end{bmatrix}}_{3\times 2} \begin{bmatrix} P_X(0) \\ P_X(1) \end{bmatrix}$$

4. Type Writer Channel

A-Z but any other alphabet is valid.



$$\begin{bmatrix} P_Y(A) \\ P_Y(B) \\ \vdots \\ P_Y(\text{"-"}) \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 0 & \cdots & 0 & 1/3 \\ & & & & & \\ 1/3 & 0 & \cdots & & 0 & 1/3 \end{bmatrix} \begin{bmatrix} P_X(A) \\ P_X(B) \\ \vdots \\ P_X(\text{"-"}) \end{bmatrix}$$

Information in the Channel

Defn: $C = \max_{P_X(x)} I(X;Y)$

ex.1 binary symmetric

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

$$H(Y|X) = H(Y|X=0)\,P(X=0) + H(Y|X=1)\,P_X(1)$$

$$= \underbrace{\{-\log(1-q)\cdot(1-q) - q\log q\}}_{H_b(q)} P_X(X=0) + H_b(q)\,P_X(X=1)$$

$$= H_b(q)$$

$$I(X;Y) \leq 1 - H_b(q)$$

equality achieved by $P_Y(0) = P_Y(1) = \frac{1}{2}$

ex. 2 ($\mathbb{Z}$-channel)

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

$$H(Y|X) = H(Y|X=0)\overbrace{P(X=0)}^{p} + H(Y|X=1)\overbrace{P(X=1)}^{1-p}$$

$$= \left(-(1-q)\log(1-q) - q\log q\right) p$$

$$= H_b(q)\cdot p$$

$$P_Y(0) = 1-p+qp$$

$$P_Y(1) = (1-q)p$$

$$H(Y) = -(1-p+qp)\log_e(1-p+pq) - (1-q)p\log_e(1-q)p$$

$$\frac{\partial I}{\partial p} = (1-q)\log(1-p+qp) + (1-q) - (1-q)\log(1-q)p + q - H_b(q)$$

$$1-p+qp = p(1-q)\left(e^{(H_b(q)-1)/(1-q)}\right)$$

$$\Rightarrow p^* = \frac{1}{(1-q)\left(1+e^{\frac{H_b(q)-1}{1-q}}\right)}$$

10/13/22 - 3

for $q = 1/2 \Rightarrow C(q = 1/2) = \log 5 - 2$ bits/channel use

look at Cover & Thomas!

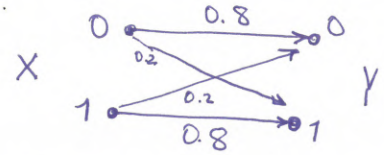
Absent on 10, 18, 22.

info. lect.

Ex. from MacKay

$X \sim$ iid Bernoulli(0.1)

B.S.C. $q = 0.2$



$n = 100$

X^{100}: 11…11 0…0 0…0
10 T2 18

$P(Y=0) = 0.1 \times 0.2 + 0.9 \times 0.8 = 0.74$, $P(Y=1) = 0.26$

10 ones out of 100 symbols: ϵ-typical!
(empirical dist. in agreement with true dist.)

11…1100 0…0 1…1
8 2 T2 18

ϵ-typical! (empirical dist. in agree. with true dist.)

⇒ Jointly Typical Sequence! (chapters 9, 10 of MacKay!)

Channel Coding Theorem

Thm. All rates below capacity, $R < C$ are achievable.

If there is a sequence of (M, n) codes ($M = 2^{nR}$) with $\lambda_{max}^{(n)} \to 0$, as $n \to \infty$,

then the rate $R \leq C$.

Proof.

Random Coding

1. Pick $p(x)$

$2^{nR} = M$ messages/labels

codeword $\rightarrow C =$ iid sequences

$$\begin{bmatrix} x_1(1) & \cdots & x_n(1) \\ \vdots & \ddots & \vdots \\ x_1(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

2. Share C with decoder.

3. Pick at random $W = w$

$$P(W=w) = \frac{1}{2^{nR}} = \frac{1}{M}$$

transmit $X^n(w)$.

4. At the decoder, Y^n

$$P(Y^n = y^n \mid X^n(w)) = \prod_{i=1}^{n} P(Y_i = y_i \mid X(w))$$

5. Decode by joint typicality

$$g(Y^n = y^n) = \hat{w}$$

$$(X^n(\hat{w}), Y^n) \in A_\epsilon^{(n)}$$ this sequences are jointly Typical.

Analysis Generate many codes, iid

$$Error = \{\hat{W} \neq W\}$$

$$P(Error) = \sum_{\substack{C \\ \text{all codes}}} P(C)\, P_e^{(n)}(C) \rightarrow \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

$$= \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \sum_{C} P(C)\ \lambda_w(C)$$

due to symmetry →

$$= \sum_{C} P(C)\ \lambda_1(C)$$

$$= P(\text{Error} \mid W=1)$$

Event

$$E_i = \left\{ (X^n(i), Y^n) \in A_\epsilon^{(n)} \right\}$$

$$\{\text{Error} \mid W=1\} = E_1^{c} \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}=M}$$

(E_1^c: complement)

$$P(\text{Error} \mid W=1) = P(\ ) \leq \underbrace{P(E_1^c)}_{\leq \epsilon \text{ very small}} + \sum_{i=2}^{M=2^{nR}} \underbrace{P(E_i)}_{\leq 2^{-n(I(X;Y)-3\epsilon)}}$$

(Union Bound)

$$P(\text{Error} \mid W=1) \leq \epsilon + 2^{nR - n(I(X;Y)-3\epsilon)}$$

If $R < I(X;Y) - 3\epsilon$ then $2^{\cdots} \leq \epsilon$

$$\Rightarrow \underline{P(\text{Error} \mid W=1) \leq 2\epsilon}$$

$$\lambda_{max}^{(n)} \leq 4\epsilon$$

Pick $p(x) = p^*(x)$

$$R < C - 3\epsilon$$

$$\frac{M}{2} = 2^{nR-1}$$

$$R - \frac{1}{n} < C - 3\epsilon$$

Lemma 1 For the n^{th} extension (using n times) of channel

$$I(X^n; Y^n) \leq nC$$

(for any $p(x^n)$)

Lemma 2 (Fano's Inequality)

$$H(X^n|Y^n) \leq 1 + P_e^{(n)} \cdot nR$$

$$nR = H(W) = H(W|Y^n) + I(W; Y^n)$$

$$I(W; Y^n) = H(W) + H(W|Y^n)$$

data processing inequality

$$\leq \underbrace{H(W|X^n)}_{H(X^n(W)|Y^n)} + I(X^n(W);Y^n)$$

Fano's Inequ.

$$\leq 1 + P_e^{(n)} \cdot nR + nC$$

$$\Rightarrow R \leq \frac{1}{n} + \underbrace{P_e^{(n)}}_{\to 0 \text{ as } n\to\infty} \cdot R + C$$

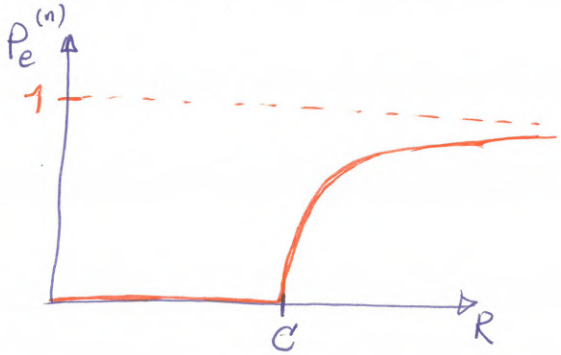
$$\Rightarrow R \leq \frac{1}{n} + C$$

Another Way to Write

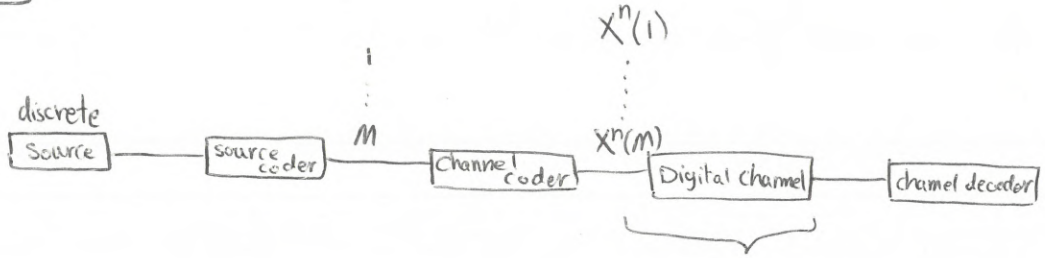
$$P_e^{(n)} \geq 1 - \frac{1}{nR} - \frac{C}{R}$$

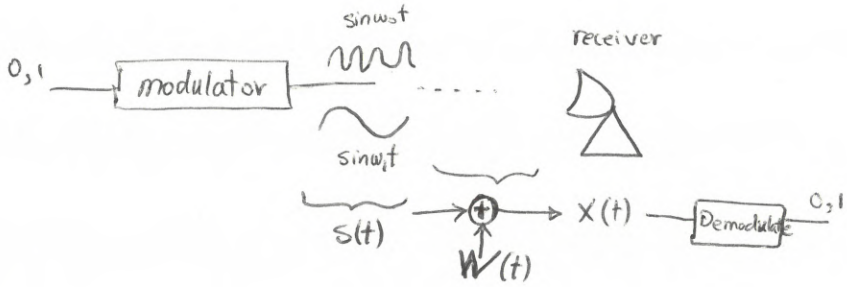
look at MacKay for more illustrations + intuitions!

if $R > C$ then $P_e^{(n)} \to 1$.



info. lect.





Continuous World ← Gaussian Channel

Differential Entropy

X is continuous R.V. with pdf $f_X(x)$

$$f_X(x) = \frac{d}{dx} F_X(x)$$

$F_X(x)$ → cdf

Defn $$h(X) = E\left[\log_2 \frac{1}{f_X(x)}\right] = -\int_{x:\, f_{x(x)}>0} f_X(x) \log_2 f_X(x)\, dx$$

ex.1 $X \sim unif(0, a)$

$$h(X) = +\int_0^a \log_2 a \times \frac{1}{a}\, dx = \log a$$

If $a > 1$, then $h(X) > 0$

If $0 < a < 1$, then $h(X) < 0$

ϵ typical sets: $A_\epsilon^{(n)} \xrightarrow[\text{of set}]{\text{size}} 2^{nH(X)}$

Continuous set $n=1$: $A_\epsilon^{(1)} \xrightarrow{\text{size}} 2^{h(X)}$

Volume of set! (not size)

(always positive)

---

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

$$h(X) = -E[\ln f_X(x)] = E\left[-\frac{1}{2\sigma^2}(X-\mu)^2 - \frac{1}{2}\ln(2\pi\sigma^2)\right]$$

$$= \frac{1}{2}\cancel{\ln e} + \frac{1}{2}\ln(2\pi\sigma^2) = \frac{1}{2}\ln(2\pi e\sigma^2) \quad \text{nats/symbol}$$

AEP

$X_1, \ldots, X_n \sim$ iid from $f_X(x)$,

then $-\frac{1}{n} \log f_{\underline{X}}(x^n) \xrightarrow[\text{converge in probability}]{P} h(X)$

(proof by using WLLN)

Typical Set

$$A_\epsilon^{(n)} = \left\{ x^n : \left| -\frac{1}{n} \log f_{X^n}(x^n) - h(X) \right| < \epsilon \right\}$$

1) $P(A_\epsilon^{(n)}) \longrightarrow 1$ as $n \longrightarrow \infty$

2) $vol(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$

3) $vol(A_\epsilon^{(n)}) \geq (1-\epsilon)\, 2^{n(h(X)-\epsilon)}$

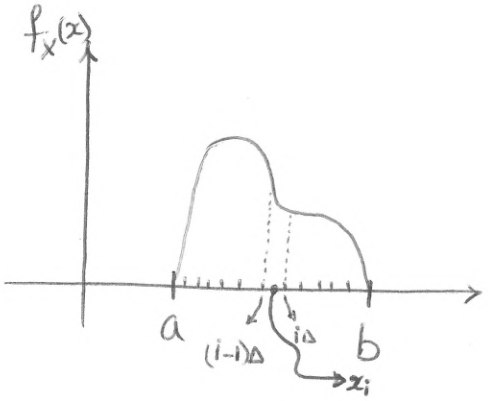
($A_\epsilon^{(n)}$: small set)

★ $vol(A_\epsilon^{(n)}) \approx 2^{nh(X)}$ OR $\doteq$

Volume of a cube $\rightsquigarrow$ $\ell^n \approx 2^{nh(X)}$

$$2^{n \log \ell} \approx 2^{n h(X)}$$

$$\ell = 2^{h(X)} \overset{\text{uniform } X}{=\!=} 2^{\log a} = a$$



n intervals of length Δ

mean value theorem

$$\int_{(i-1)\Delta}^{i\Delta} f_X(x)\,dx = f(x_i) \cdot \Delta$$

$$X^\Delta = \left\{ x_i \quad \text{with} \quad p(x_i) = f(x_i) \cdot \Delta \right.$$

$$H(X^\Delta) = -\sum_i f_X(x_i)\,\Delta \times \underbrace{\log\left(f_X(x_i)\Delta\right)}_{\log f_X(x_i) + \log \Delta}$$

If $\Delta \to 0$, $n \to +\infty$

$$\Rightarrow \quad -\sum_i \underbrace{f_X(x_i)\,\Delta}_{dx} \log f_X(x_i) \longrightarrow h(X)$$

$$-\log\Delta \sum f_X(x_i)\Delta \longrightarrow -\log\Delta$$

as $n \to +\infty$
$\Delta \to 0$

$$H(X^\Delta) + \log\Delta \longrightarrow h(X)$$

$\log\Delta = -n$, 2^n = number of divisions

---

ex. $X \sim \text{unif}(0, 1/2)$

$$h(X) = \log a = -1$$

division step: $\Delta = 2^{-4}$ (log)

$$H(X^\Delta) = -1 + 4 = 3$$

* as we decrease Δ, number of bits/symbol goes to ∞!!!

Defn Relative Entropy $D(f\|g) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} dx$ ($g_X(x) \neq 0$)

Jensen's Inequality $D(f\|g) \geq 0$

$I(X;Y) = \iint f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dx\,dy \Rightarrow I(X;Y) \geq 0$

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

$$h(Y) \geq h(Y|X)$$

lect. info.

ex.

$$\underline{X} = X^n \sim N(\underline{\mu}, K_{n\times n})$$

$$h(\underline{X}) = E\left[-\log_e f_{\underline{X}}(x^n)\right]$$

$$f_{X^n}(x^n) = \frac{1}{\sqrt{(2\pi)^n \det(K)}}\, e^{-\frac{1}{2}(x^n-\underline{\mu})^T K^{-1}(x^n-\underline{\mu})}$$

$$= -E\left[-\tfrac{1}{2}(x^n-\underline{\mu})^T K^{-1}(x^n-\underline{\mu})\right] - \frac{E}{2}\left[-\ln(2\pi)^n(\det(K))\right]$$

$$= \frac{1}{2}\,\mathrm{Tr}\left(K K^{-1}\right) + \frac{1}{2}\ln(2\pi)^n \det(K)$$

$$K K^{-1} = E[X^n (X^n)^T]$$

$$\frac{n}{2} = \frac{1}{2}\ln e^n \qquad I$$

$$= \frac{1}{2}\ln\left((2\pi e)^n \det(K)\right)$$

Proposition

$$X^n$$

$$E[X^n] = \emptyset$$

$$E[X^n (X^n)^T] = K$$

show: $h(X^n) \leq \frac{1}{2}\ln\left((2\pi e)^n \det(K)\right)$

↳ if $X^n \sim N(\emptyset, K)$

proof: $D(f_{X^n} \| g_{X^n})$ ← pdf of jointly Gaussian

$$= E_{f_{X^n}}\left[\ln \frac{f_{X^n}}{g_{X^n}} \right] \geq 0$$

$$-h_{f_{X^n}}(X^n) + \underbrace{E_{f_{X^n}}\left(\ln \frac{1}{g_{X^n}} \right)}_{\frac{1}{2}\ln\left[(2\pi e)^n \det(K)\right]} \geq 0 \quad \checkmark$$

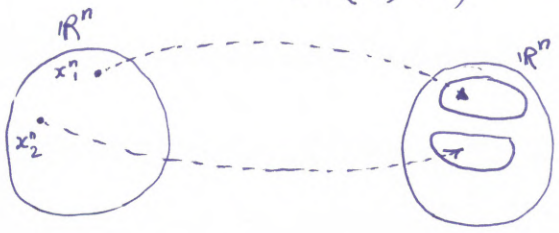
$$E_{f_{X^n}}\left[\frac{1}{2} X^{n^T} K^{-1} X^n + \frac{1}{2} \ln\left((2\pi)^n \det(K)\right) \right]$$

$$= \frac{1}{2} \mathrm{Tr}(K K^{-1}) + \frac{1}{2} \ln (2\pi)^n \det(K)$$

Gaussian Channel

• power constraint: $E[X^2] \leq P$

$Z \sim \mathcal{N}(0, N)$



no overlapping in output space.

M codeword @ input

Defn ~~The capacity of~~ (Theoretical Capacity)

$$C = \max_{f_X(x):\, E[X^2] \leq P} I(X;Y)$$

(assumption: Gaussian Channel)

$$I(X;Y) = h(Y) - h(Y|X)$$

$$\int_{\mathbf{X}} f_X(x) \left(\int f_{Y|X}(y|x) \log \frac{1}{f_{Y|X}(y|x)}\, dy \right) dx$$

for $X = x$

$$Y = x + Z$$

$Z \sim N$

$Y \sim N \Leftarrow x$ is fixed

$\Longrightarrow \frac{1}{2} \ln(2\pi e N)$

$$\Rightarrow I(X;Y) = h(Y) - \frac{1}{2} \ln(2\pi e N)$$

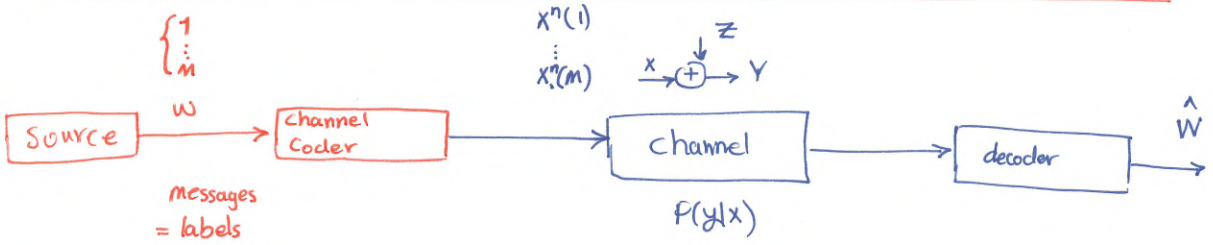
$$h(Y) - \frac{1}{2}\ln(2\pi e N) \leq \frac{1}{2}\ln\left(2\pi e (N+P)\right) - \frac{1}{2}\ln(2\pi e N)$$

to maximize

if $X \sim N(0, P)$

$$\frac{1}{2}\ln\left(1 + \frac{P}{N}\right)$$

$$\Rightarrow C = \frac{1}{2}\ln\left(1 + \frac{P}{N}\right)$$



Defn: (M, n) power constrained code. consist of: messages

1. a set messages / labels

2. an encoder $f: w \longrightarrow x^n(w)$

$w: 1 \ldots M$

3. ensure power constraint

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \leq P$$

4. an encoder $g: Y^n \longrightarrow \hat{W}$

Rate of code:

$$\begin{cases} R = \dfrac{\log M}{n} \\ M = 2^{nR} \end{cases}$$

R is achievable if there is a sequence of (M, n) power contrained codes with

$\lambda^{(n)} \longrightarrow 0$ as $n \longrightarrow 0$

($\lambda^{(n)}$: max. probability of error)

Operational capacity is the maximum of achievable R.

ϵ-typical set on the output of the Gaussian channel

$$A_\epsilon^{(n)} = \left\{ y^n : \left| -\frac{1}{n} \sum \log f_{Y^n}(y^n) - h(Y) \right| < \epsilon \right\}$$

$$Y_i \sim \mathcal{N}(0, P+N)$$

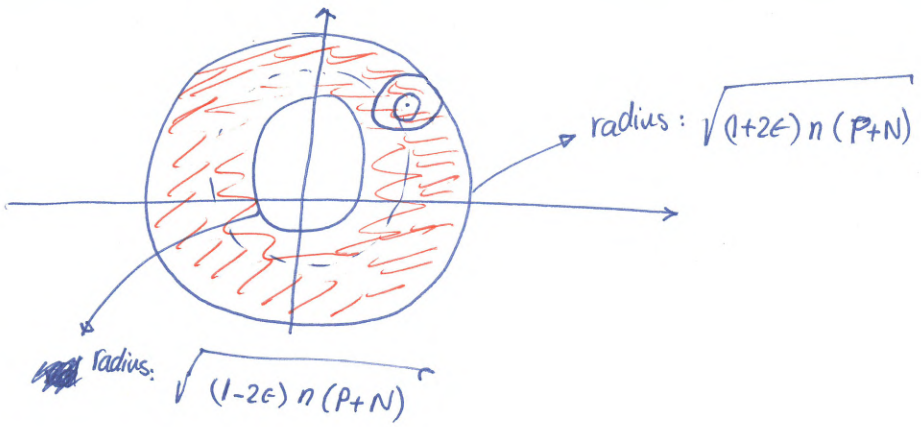
$$\sum_{i=1}^{n} \ln f_{Y_i}(y_i)$$

$$-\sum_{i=1}^{n} \left[\frac{1}{2} \frac{(y_i)^2}{(P+N)} + \frac{1}{2} \ln\left(2\pi (P+N)\right) \right]$$

$$\left| \frac{1}{2n(P+N)} \sum_{i=1}^{n} y_i^2 + \underbrace{\frac{1}{2} \ln(2\pi(P+N)) - \frac{1}{2} \ln(2\pi e(P+N))}_{-\frac{1}{2}} \right| \leq \epsilon$$

$$(1-2\epsilon)\, n(P+N) \leq \underbrace{\sum_{i=1}^{n} y_i^2}_{r^2} \leq (2\epsilon+1)\, n(P+N)$$

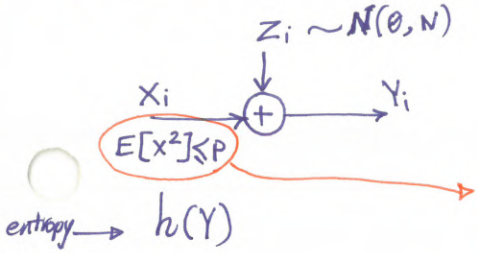
Sphere in n-dim space.
with radius r^2



For a given $x^n(w)$, the output $Y_i \sim N(x_i(w), N)$

↳ much shallower!!!

info. lect.



entropy → $h(Y)$

this will enforce the partition on the output, unless, this problem would be unconstrained with on partitions.

$$h(Y) = \frac{1}{2} \ln(2\pi(N+P))$$

capacity:

$$\max_{f_X: E[x^2] \le P} I(X;Y) = \frac{1}{2} \ln\left(1 + \frac{P}{N}\right)$$

(M, n) - code

with x^n $\quad \frac{1}{n} \sum_{i=1}^{n} x_i^2(w) \le P$

- encoding function

$$f: w \longrightarrow X^n(w)$$

label

- decoding function (by joint typicality)

$$Y^n \quad P(y|x)$$

$$g \rightarrow \theta(Y^n) = \hat{w}$$

$$R = \frac{\log M}{n}$$

$$A_\epsilon^{(n)}(y^n) = \left\{ y^n : \left| -\frac{1}{n} \ln f_{Y^n}(y^n) - h(Y) \right| < \epsilon \right\}$$

$$Y_i \sim N(0, P+N)$$

$$-\frac{1}{n}\left(-\frac{1}{2(P+N)} \sum y_i^2 - \frac{n}{2} \ln(2\pi(P+N))\right) - h(Y)$$

$$\frac{1}{2}\ln(2\pi e(N+P))$$

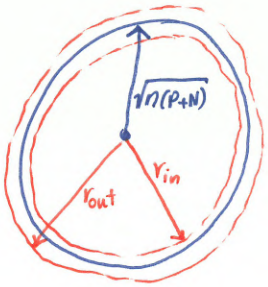
$$\Leftrightarrow \quad \frac{1}{2n} \frac{1}{P+N} \sum_{i=1}^{n} y_i^2 + \frac{1}{2}\ln(2\pi(P+N)) - \frac{1}{2}\ln(2\pi e(P+N))$$

$$-\frac{1}{2}\ln e = \frac{-1}{2}$$

$$(P+N)2n\left(\frac{1}{2} - \epsilon\right) < \sum y_i^2 < \left(\epsilon + \frac{1}{2}\right)2n(P+N) \qquad \epsilon$$

$$\sum_1^n y_i^2 = r^2$$

Sphere of n-dimensional with radius r



$$r_{out} = \sqrt{n(1+2\epsilon)(N+P)}$$

$$r_{in} = \sqrt{n(1-2\epsilon)(N+P)}$$

Fix $X^n = x^n(\omega)$

$$Y^n = x^n(\omega) + Z^n$$

$$Y_i = x_i(\omega) + Z_i$$

$Z_i \sim \mathcal{N}(0, N)$

$Y_i \sim \mathcal{N}(x_i(\omega), N)$

empirical entropy

$$A_\epsilon^{(n)}(y^n \mid x^n(\omega)) = \left\{ y^n : \left| -\frac{1}{n} \log\left(f_{Y^n|X^n}(y^n|x^n(\omega)) \right) - h(Y|X) \right| < \epsilon \right\}$$

$x^n(\omega)$

$$\left| \frac{1}{2n} \frac{1}{N} \sum (y_i - x_i)^2 + \frac{1}{2} \ln(2\pi N) - \frac{1}{2} \ln(2\pi e N) \right| < \epsilon$$

$$\left(\frac{1}{2} - \epsilon\right) 2n N < \sum (y_i - x_i)^2 < \left(\epsilon + \frac{1}{2}\right) 2n N$$

$$r = \sqrt{nN}$$

$$r_{out} = \sqrt{(2\epsilon+1)\, nN}$$

$$r_{in} = \sqrt{(1-2\epsilon) nN}$$

$$\text{Volume}_{\text{Sphere}}(r^n) = C_n \cdot r^n$$ (sphere-packing)

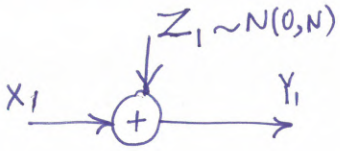
C_n: just a positive number

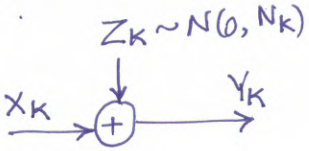
$$\frac{C_n \cdot \{n(P+N)\}^{n/2}}{C_n \cdot \{nN\}^{n/2}} = \left(1+\frac{P}{N}\right)^{n/2} = M = 2^{nR} = 2^{nC}$$

(R: capacity)

$$\Rightarrow C = \frac{1}{2}\ln\left(1+\frac{P}{N}\right)$$

Parallel Gaussian Channels (sth like Eigenvalue Decomposition on Images)





Independent channels

$\sum_{i=1}^{n} P_i \le P$ → all power to be distributed over the channels

P_i: $E[X_i^2]$

$$\max I(X^K; Y^K) = ?$$

$$h(Y^K) - h(Y^K|X^K)$$

$h(Z^K)$

$$= h(Y^K) - \sum_{i=1}^{K} h(Z_i)$$

$h(Z_i) = \frac{1}{2}\ln(2\pi e N_i)$

$$\le \sum_{i=1}^{K} h(Y_i)$$

$h(Y_i)$: $\mathcal{N}(0, P_i+N_i)$

$$= \sum_{i=1}^{K} \frac{1}{2} \ln\left(2\pi e (N_i + P_i)\right) - \sum_{i=1}^{K} \frac{1}{2} \ln\left(2\pi e N_i\right)$$

$$= \frac{1}{2} \sum_{i=1}^{K} \ln\left(1 + \frac{P_i}{N_i}\right)$$

we will use the maximum available power: $\sum_{i=1}^{K} P_i = P$

$$\rightsquigarrow \max \frac{1}{2} \sum \ln\left(1 + \frac{P_i}{N_i}\right) \qquad \text{Lagrange Multiplier}$$

$$P_i : \sum_{i=1}^{K} P_i = P, \; P_i \geq 0$$

$$J(P_1, \dots, P_K, \lambda, \mu_1, \dots, \mu_K)$$

$$= \frac{1}{2} \sum_{i=1}^{K} \ln\left(1 + \frac{P_i}{N_i}\right) + \lambda \left(\sum_{i=1}^{K} P_i - P\right) + \sum_{i=1}^{K} \mu_i P_i$$

Kuhn Tucker
2 cases
$(\sum \mu_i P_i = 0)$

- $P_i > 0, \; \mu_i = 0$
- $P_i = 0, \; \mu_i < 0$

(I) $P_i > 0, \; \mu_i = 0$

$$\frac{\partial J}{\partial P_i} = 0 \Rightarrow \frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0 \rightarrow P_i + N_i = \frac{1}{-2\lambda}$$

why? hard to describe

$\mathcal{K} > 0 \Rightarrow (\lambda < 0)$

$$\Rightarrow \boxed{P_i = \mathcal{K} - N_i}$$

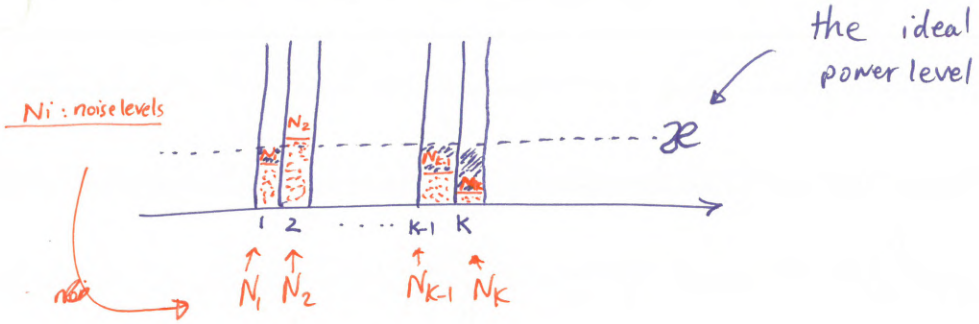
(II) $P_i = 0, \; \mu_i < 0$

$$\frac{\partial J}{\partial P_i} = 0 \Rightarrow \frac{1}{2} \frac{1}{P_i + N_i} + \lambda + \mu_i \Big|_{P_i = 0} = 0 \Rightarrow \mu_i = \underbrace{-\frac{1}{2(P_i + N_i)}\Big|_{P_i = 0}}_{< 0} - \lambda$$
$$(< 0)$$

$$\Rightarrow \mu_i = -\frac{1}{2N_i} - \lambda < 0 \rightarrow N_i > \underbrace{-\frac{1}{2\lambda}}_{\mathcal{K}}$$

So if noise level is above $\mathcal{K}$ we don't put any power into that channel.

$$\sum_{i=1}^{K} (\mathcal{X} - N_i)^+ = P$$



info. lect.

Bounds on Population of Binary Codes

$M = 2^{nR}$

↳ # of distinct codewords

→ $\mathcal{X} = \{0, 1\}$

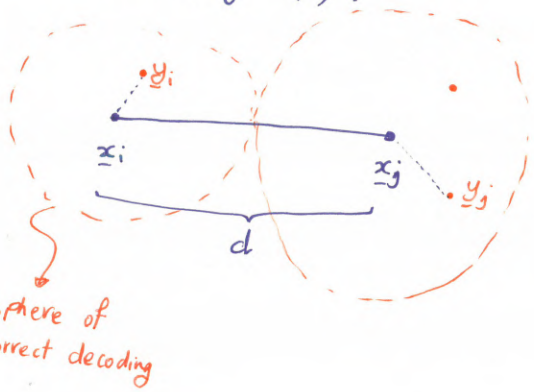
→ length n codewords.

→ 2^n is total number of sequences.

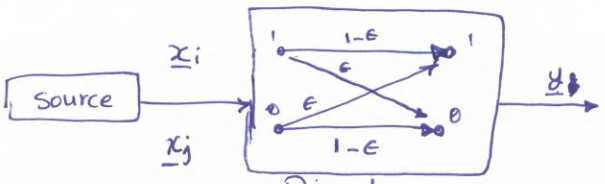
→ Form a code C s.t. $d(\underline{x}_i, \underline{x}_j) \geq d$ (i)

distance between a pair of codewords

Goal: To Find the max population of a binary code C, with code words $\{\underline{x}_i\}$ satisfying (i).



→ channel may introduce distortion.



Discrete Memoryless (DMC) Channel (like binary symmetric)

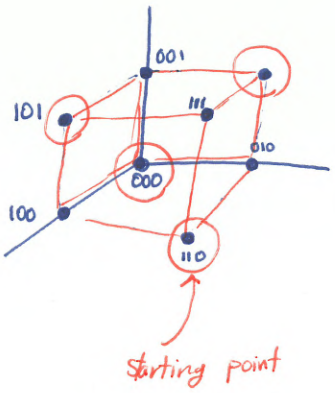
- choose $d(\underline{x}_i, \underline{x}_j)$ as Hamming distance.

ex. 1100 / 1000 → $d_H = 1$

$$d_H = \sum_k x_{i,k} \overset{\text{modular addition}}{\oplus} x_{j,k}$$

- Decode $\underline{y}$ as $\underline{x}_i$ if $d_H(\underline{x}_i, \underline{y}) < \frac{d}{2}$

$n=3$ ⟶ 8 possible codewords.



$d_H \geq 2$

---

upper bound

- Hamming Bound ⟶ C with $d(\underline{x}_i, \underline{x}_j) \geq d$, d integer > 0 (Hamming distance)
- Johnstone "
- Plotkin "
- Singleton "

lower bound

- Gilbert – Varshamov Bound

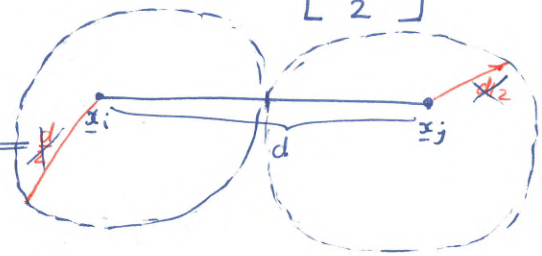
$$M(n,d) \leq \frac{2^n}{\sum_{j=0}^{t} \binom{n}{j}}$$

binary sequences that differ from $\underline{x}_i$ by j bits. : $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} \cdots + \binom{n}{t}$ with $t = \left\lfloor \frac{d-1}{2} \right\rfloor$

$\binom{n}{0}$: $\underline{x}_i$ itself

$\binom{n}{1}$: 1 bit difference to $\underline{x}_i$

$\left\lfloor \frac{d-1}{2} \right\rfloor$



Sphere Packing Bound

Plotkin Bound

1) If d is even and $2d > n$, then

$$M(n,d) \leq 2 \left\lfloor \frac{d}{2d-n} \right\rfloor$$

2) If d is odd and $2d+1 > n$, then

$$M(n,d) \leq 2 \left\lfloor \frac{d+1}{2d+1-n} \right\rfloor$$

3) If d is even and $n = 2d$, then

$$M(n,d) \leq 4d$$

4) If d is odd and $n = 2d+1$, then

$$M(n,d) \leq 4d$$

proof of 1)

$$\sum_{\substack{i \neq j \\ x_i \in C}} d_H(x_i, x_j) \geq M \cdot (M-1) \cdot d \qquad (*)$$

(M-1 comparison)

s_i zeros
$M - s_i$ ones

11/10/22 - 2

$$2 s_i (M - s_i)$$

it is maximized

if $s_i = \frac{M}{2}$ (equal number of zeros and ones)

$$\sum_{\substack{i \neq j \\ \underline{x}_i \in C}} d_H(\underline{x}_i, \underline{x}_j) = \sum_{i=1}^{n} 2 s_i (M - s_i) \leq \sum_{i=1}^{n} \underbrace{2 \frac{M}{2} \left(M - \frac{M}{2}\right)}_{M^2/2}$$

$$(**) \quad \frac{n M^2}{2}$$

combining (*) and (**):

$$M(M-1)d \leq \frac{n M^2}{2}$$

$$M\left(d - \frac{n}{2}\right) \leq d \qquad \blacksquare$$

Sigleton Bound

C with $d(\underline{x}_i, \underline{x}_j) \geq d$

$$M(n, d) \leq 2^{n-d+1}$$

1) remove $d-1$ first bits from each codeword, then the new codewords are still distinct.

2) The number of new sequences of lengths $n-d+1$ to $n-(d-1)$ is 2^{n-d+1} thus it is a upper bound on M.

Gilbert - Varshamov Bound

C with $d(\underline{x}_i, \underline{x}_j) \geq d$

$$M(n, d) \geq \frac{2^n}{\sum_{j=0}^{d-1} \binom{n}{j}}$$

There is at least ONE sequence $\underline{z}$ among 2^n sequence, with property

$$d_H(\underline{x}_i, \underline{z}) \leq d-1$$

→ Then the entire 2^n set of sequences can be covered by $\bigcup_{\underline{x}_i \in C} B(\underline{x}_i, d-1)$ ball with center $\underline{x}_i$ with radius $(d-1)$.

Then

$$2^n = \left| \bigcup_{\underline{x}_i \in C} B(\underline{x}_i, d-1) \right|$$

(cardinality (size))

$$\leq \sum_{\underline{x}_i \in C} \left| B(\underline{x}_i, d-1) \right|$$

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d-1}$$

$$= \left(\sum_{j=0}^{d-1} \binom{n}{j} \right) M \Rightarrow \boxed{M \geq \frac{2^n}{\sum_{j=0}^{d-1} \binom{n}{j}}}$$

info. lect.

- thomas & cover tried to develop an RD theory which covers both source quantization and channel quantization.

---

ex.

$$X_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

with chosen Hamming Distortion.

$$E[d(X,\hat{X})] = D$$

$$R(D) = \min I(X;\hat{X})$$

$$p(\hat{x}|x): \; E[X \oplus \hat{X}] \leq D$$

$$I(X;\hat{X}) = \underbrace{H(X)}_{H_b(p)=1} - \underbrace{H(X|\hat{X})}_{H_b(D)}$$

channel is symmetric and it can be reversed.

$$\Rightarrow R(D) = 1 - H_b(D)$$

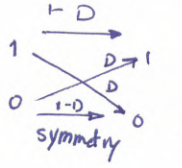
$$E[X \oplus \hat{X}] = D$$

$X \oplus \hat{X} \sim$ { 1 : $P[X \neq \hat{X}]$; 0 : $P[X = \hat{X}]$ }

$$D = 1 \times P[X \neq \hat{X}] + 0 \times P[X = \hat{X}]$$

$$P(X=0,\hat{X}=1) + P(X=1,\hat{X}=0)$$

$$\underbrace{P(\hat{X}=0|X=1)}_{D}\underbrace{P(X=1)}_{1/2} + \underbrace{P(\hat{X}=1|X=0)}_{D}\underbrace{P(X=0)}_{1/2}$$



$$P(\hat{X}=1) = \frac{1}{2}$$

$$P(\hat{X}=0) = \frac{1}{2}$$

$X_i \sim \text{Bernoulli}(p)$

$$H(X|\hat{X}) = H(X \oplus \hat{X}|\hat{X}) = -\sum_{\hat{x}=0}^{1} p(\hat{x}) \cdot \underbrace{\sum_{x=0}^{1} p(x|\hat{x}=\hat{x}) \log p(x|\hat{x}=\hat{x})}$$

$$- P(x=0|\hat{X}=0)\log p(x=0|\hat{X}=0) + \dots$$

$$H(X\oplus\hat{X}\mid\hat{X}) = -\sum_{\hat{x}=0}^{1} p(\hat{x}) \sum_{x=0}^{1} p(x\oplus\hat{x}\mid\hat{x}) \log p(x\oplus\hat{x}\mid\hat{x})$$

$$- p(\underbrace{x=0\oplus\hat{x}=0}_{x=0}\mid\hat{x}=0)\cdot\log(\cdot)$$

$$- p(\underbrace{x=1\oplus\hat{x}=0}_{x=1}\mid\hat{x}=0)\cdot\log$$

we Knew: $H(X\oplus\hat{X}\mid\hat{X}) \leq H(X\oplus\hat{X})$

so we have lowerbound on M.I.: $\boxed{I(X;\hat{X}) \geq H_b(p) - H(X\oplus\hat{X})}$

as previous page. $H_b(D)$

---

ex.2 $X \sim \mathcal{N}(0,\sigma^2)$

$E[(X-\hat{X})^2] \leq D$

$$I(X;\hat{X}) = h(X) - h(X\mid\hat{X})$$

$$= \frac{1}{2}\log(2\pi e\sigma^2) - h(X-\hat{X}\mid\hat{X})$$

$$\leq h(X-\hat{X})$$

$$\leq \frac{1}{2}\ln(2\pi e D)$$

$\Rightarrow I(X;\hat{X}) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$

variance is given.

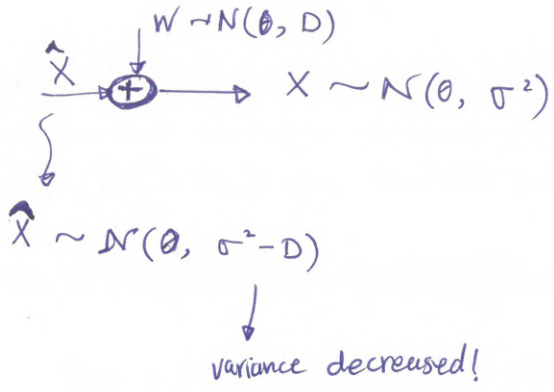
given variance, maximized when Gaussian

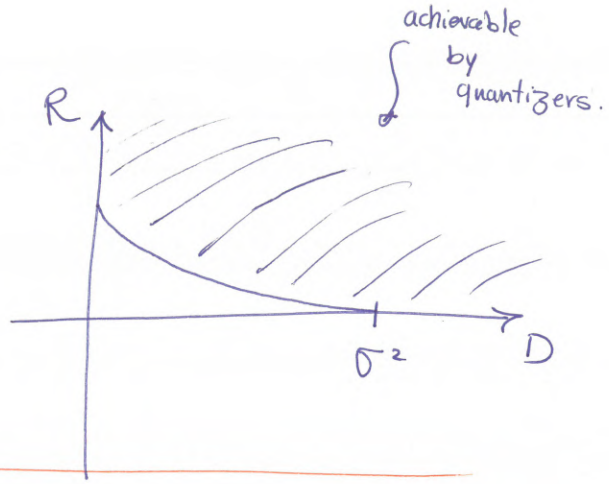
$\log\mathcal{N}(0,D)$

When is it achieved?

let's go for the test channel.

Gaussian channel makes sense as test channel.

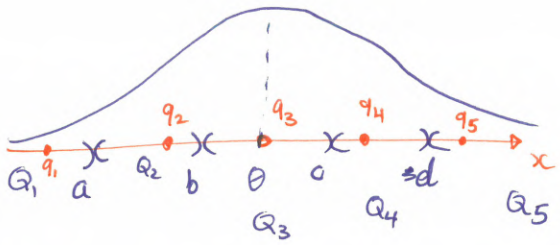




RV $X \sim N(0, \sigma^2)$

Lloyd Quantizer

Least Squares Quantizing (K Means)



$$\min \sum_{i=1}^{5} E\left[(X - q_i)^2\right]$$

Q_i, q_i

q_i : quanta

Q_i : partition

Alternating minimization

1. Given Q_i, $i = 1, 2, \ldots, 5$

calculate

$$q_i = \frac{\int_{x \in Q_i} x \cdot f_X(x)\, dx}{\int_{x \in Q_i} f_X(x)\, dx}$$

conditional expectation

2. Update region

if $|x - q_i|^2$ [illegible] $> |x - q_j|^2$

then [illegible] move x to [illegible] Q_j.

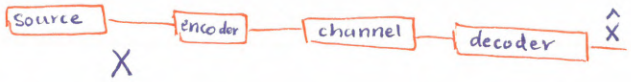
info lect.

Joint source-channel coding is still a problem

rate-distortion

how to quantize

how to do joint source-channel coding



X

discrete → $H(X) < R < C$

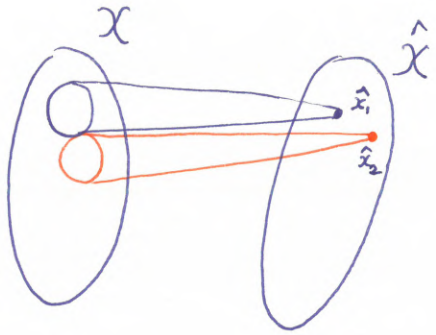
continuous → $R > C \Rightarrow$ error

Rate-Distortion (RD)

- $R(D)$
- $D(R)$

- Theoretical RD function

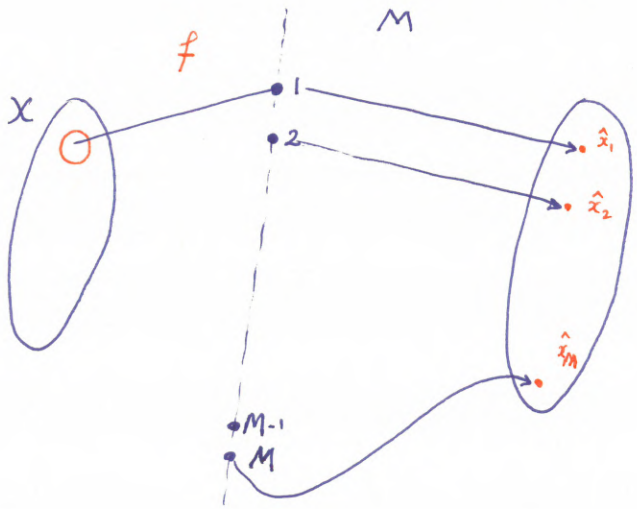
$$R(D) = \min_{P(\hat{x}|x):\ E[d(x,\hat{x})] \leq D} I(X;\hat{X})$$



- Operational RD function

(n, M) codes

$f: x^n \to \{1, \ldots, M\}$



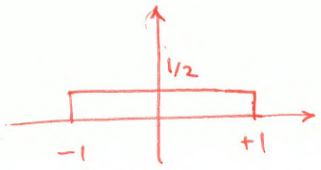
$$R = \frac{\log M}{n}$$

$$\lim E[\bar{d}(x,\hat{x})] = D$$

$$\bar{d}(x,\hat{x}) = \frac{1}{n}\sum_{n=1}^{n} d(x_i,\hat{x}_i)$$

ex. $X \sim \text{unif}(-1, 1)$

Quantize to 1 bit representation.



$R = 1$

encode sign!

$M = 2^R = 2 \longrightarrow \{1, 2\}$

$(-1, 0) \xrightarrow{\text{encode}} 1 \xrightarrow{\text{decode}} x_1$

$(0, 1) \xrightarrow{\text{encode}} 2 \xrightarrow{\text{decode}} x_2$

$$E[d(X, \hat{X})] = E[(X - \hat{X})^2]$$

$$\min_{x_1, x_2} E[(X - \hat{X})^2]$$

$$\int_{-1}^{0} \frac{1}{2} (x - x_1)^2 \, dx + \int_{0}^{1} \frac{1}{2} (x - x_2)^2 \, dx$$

d/dx_1 (

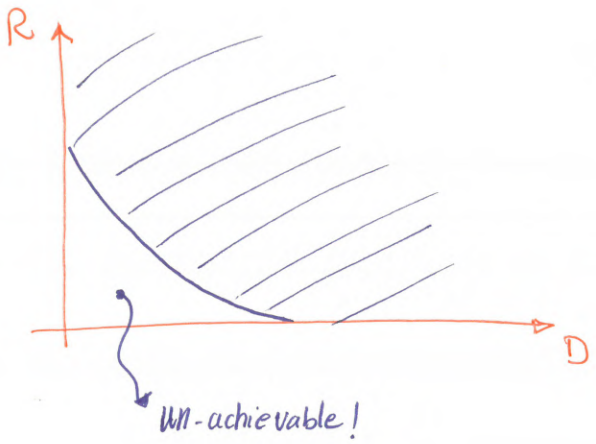
$$-\int_{-1}^{0} (x - x_1) \, dx = 0$$

$$\frac{x^2}{2} - x_1 x \Big|_{-1}^{0} = 0$$

$$-\frac{1}{2} + x_1 = 0$$

Reproduction Points

$x_1 = \frac{1}{2}$ $\xrightarrow{\text{symmetry}}$ $x_2 = \frac{1}{2}$



ex.2 Binary Source

$X_i \sim$ Bernoulli (p)

$$R(D) = \begin{cases} H(p) - H(D), & \text{if } D \leq \min\{p, 1-p\} \\ 0, & \text{else} \end{cases}$$

(for binary source, distortion is always Hanning distance)

$$d(x, \hat{x}) = \begin{cases} 1, & \text{if } x \neq \hat{x} \\ 0, & \text{else} \end{cases}$$

$$R(D) = \min I(x; \hat{x})$$

$$P(\hat{x}|x) : E[d(x, \hat{x})] \leq D$$

$$d(x, \hat{x}) = x \oplus \hat{x}$$

modulo 2 addition

$$I(x; \hat{x}) = \underbrace{H(X)}_{H_b(p)} - \underbrace{H(X|\hat{X})}_{\sum_{\hat{x}} H(X|\hat{x}=\hat{x})\, p(\hat{x})}$$

$H(X|\hat{x}=\hat{x})$ → $H(X \oplus \hat{x} | \hat{x}=\hat{x})$

adding constant to r.v. doesn't chage entropy

we did the same in channel coding notes.

$$H(X \oplus \hat{X} | \hat{X}) \leq H(X \oplus \hat{X})$$

$$\Rightarrow \quad I(X;\hat{X}) \geqslant H_b(p) - \underbrace{H(X \oplus \hat{X})}_{H_b(D)}$$

$$X \oplus \hat{X} = \begin{cases} 1 & \text{if } X \neq \hat{X} \\ 0 & \text{if } X = \hat{X} \end{cases}$$

$$E[X \oplus \hat{X}] = 1 \cdot P[X \neq \hat{X}] + 0 \cdot P(X = \hat{X})$$

we used the upperbound for $E[d(x,\hat{x})]$

$$= D$$

Conditions for the equality to be hold

when $H(X|\hat{X}) \stackrel{?}{=} H_b(D)$?

what if $D > \min\{p, 1-p\}$

pick $\hat{X} = 0$, the $P(\hat{X} = 0) = 1$

$$P[X \oplus \hat{X} \mid \hat{X} = 0] = P[X] = p$$

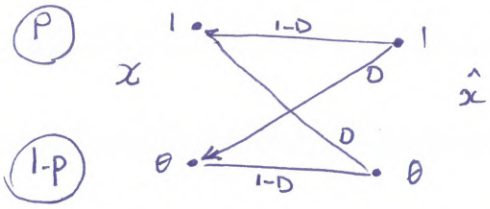
$$I(X;\hat{X}) = 0$$

If $D < p$

Design a test channel

BSC

Prior



$$P = q(1-D) + (1-q)D$$

$$P + D = q(1-2D)$$

$$q_p = \frac{P+D}{1-2D}$$

~~$P[\hat{X}=1]$~~

$$p(\hat{x}|x) = \frac{P(x|\hat{x})\,p(\hat{x})}{P(x)}$$

info. lect.

Introduction to Large Deviations (LD)

Consider a sequence $\{S_n\}$

$$S_n \xrightarrow{P} c \quad , \quad \text{as} \quad n \to \infty$$

same as

$$P(|S_n - c| > \epsilon) \longrightarrow 0 \quad , \quad \text{as} \quad n \to \infty$$

ex.

$\{Y_i\}$

form sample mean.

$$S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$P(|S_n - c| > \epsilon) = \underbrace{k(n, \epsilon, c)}_{\text{slowly varying}} e^{-nI(c,\epsilon)}$$

compared to $e^{-nI(c,\epsilon)}$

↳ Now we can say S_n satisfies LD principle.

2 results

1. Cramer's Theorem

$X_i \sim$ iid R.V.'s

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Assume $E[X_i] = m$

By the LLN $S_n \xrightarrow{P} m$

sth like Chernoff Bound!

Consider function $I(x) = \sup_{\theta} \left[\theta x - \log M(\theta) \right]$

(sup is like max but applies to both open and closed sets)

$$M(\theta) = E_{f_X}\left[e^{\theta x}\right] = \int_{-\infty}^{+\infty} e^{\theta x} f_X(x)\, dx$$

Moment Generating Function → It is CONVEX. ⇒ $\log M(\theta)$ is also CONVEX.

$I(x)$: Large Deviation Rate Function

---

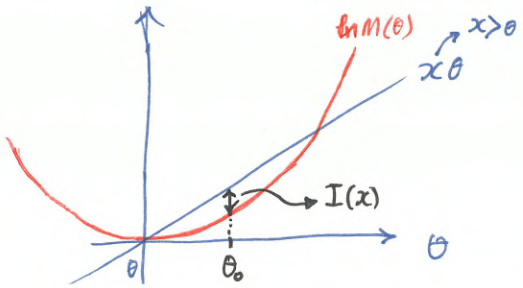
properties of LD rate function

(P.1) $I(x)$ is a convex (U) function.

(P.2) $I(x)$ has minimum at $E[X_i] = m$.

and furthermore $I(m) = 0$.

(P.3) If $M(\theta) < \infty$ $\quad I(x) = x\theta_0 - \ln M(\theta_0)$



θ_0: where the difference is maximum.

ex.1

$$X \sim N(0,1)$$

$$M(\theta) = E[e^{j\theta}] = e^{\theta^2/2}$$

$$I(x) = \sup_{\theta} \; x\theta - \frac{\theta^2}{2}$$

$$\frac{\partial}{\partial \theta} I(x) = \theta \Rightarrow \boxed{x = \theta_0}$$

$$\Rightarrow I(x) = \frac{x^2}{2}$$

ex.2

$$X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

$$M(\theta) = E[e^{\theta x}] = e^{\theta}\frac{1}{2} + \frac{1}{2}$$

$$I(x) = \sup_{\theta} \left(x\theta + \log 2 - \log 1 + e^{\theta}\right)$$

$$x - \frac{e^{\theta_0}}{1 + e^{\theta_0}} = \theta$$

$$x + e^{\theta_0}(x-1) = 0$$

$$\boxed{\theta_0 = \log \frac{x}{1-x}}$$

very like binary entropy...

$$I(x) = x \log \frac{x}{1-x} + \log 2 - \log \frac{1}{1-x} = \overbrace{x \log x + (1-x)\log 1-x}^{} + \log 2$$

11, 29, 22 - 2

Now → **Cramer's Theorem**

Assume $M(\theta) < \infty$ for all θ. Then for every closed set $F \subseteq \mathbb{R}'$

$$\overline{\lim_{n \to +\infty}} \frac{1}{n} \log P(S_n \in F) \leq - \inf_{x \in F} I(x)$$

approaching from above

and for every open set $G \subseteq \mathbb{R}'$

Sandwich

$$\underline{\lim_{n \to +\infty}} \frac{1}{n} \log P(S_n \in G) \geq - \inf_{x \in G} I(x)$$

approaching from below

If F is $[a, b]$ and $G = (a, b)$, then

$$\overline{\lim}(\,.\,) = \underline{\lim}(\,.\,) = - \inf_{x \in [a,b]} I(x)$$

ex.3

$X_i \sim$ iid $N(0,1)$

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$S_n \xrightarrow{P} 1$, as $n \to +\infty$ (since mean is zero, S_n will be sample variance)

$P(S_n > \gamma)$

γ should be larger than 1 ($\gamma > 1$)

$$\frac{S_n - 1}{\sqrt{2/n}} \xrightarrow{\text{in dist.}} Z \sim N(0,1)$$

from chi-squared distribution

$$\Rightarrow S_n \sim N(1, 2/n)$$

The central-limit-theorem.

$$P(S_n > \gamma) \approx \int_{\gamma}^{+\infty} \frac{1}{\sqrt{2\pi \cdot 2/n}} e^{\frac{-(t-1)^2}{2\cdot 2/n}} dt = Q\left(\frac{\gamma - 1}{\sqrt{2/n}}\right)$$

$\gamma > 1$

$$\approx k\, e^{\frac{-n}{4}(\gamma-1)^2}$$

rate function $= \frac{(\gamma-1)^2}{4}$

k: constant

Now using Cramer's Thm.

$$P(S_n > \gamma) \approx k(\gamma, n, 1) \cdot e^{-n \inf_{x>\gamma} I(x)}$$

(1 → mean)

$$I(x) = \sup_{\theta} (x\theta - \log M(\theta))$$

$$M(\theta) = E\left[e^{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \cdot \theta} \right]$$

$$= \left(E\left[e^{\frac{\theta}{n} X_i^2} \right] \right)^n$$

$$\int_{-\infty}^{+\infty} e^{\frac{\theta}{n}x^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(1 - \frac{2\theta}{n}\right)x^2} \times \frac{\sqrt{\frac{1}{1-2\theta/n}}}{\sqrt{\frac{1}{1-2\theta/n}}}$$

→ make it a **Gaussian**.

$$= \left(\frac{1}{1 - \frac{2\theta}{n}}\right)^{\frac{n}{2}}$$

$$I(x) = \sup_{\theta} \left(x\theta + \frac{n}{2} \log\left(1 - \frac{2\theta}{n}\right) \right)$$

$$\frac{\partial}{\partial\theta} I(x) = x + \frac{n}{2} \frac{-2/n}{1 - \frac{2\theta_0}{n}} = 0 \rightarrow x = \frac{1}{1 - \frac{2\theta_0}{n}}$$

$$\rightarrow x - \frac{2\theta_0}{n} x = 1 \rightarrow \boxed{\theta_0 = \frac{x-1}{x}\frac{n}{2}}$$

$$\Rightarrow I(x) = n\left(\frac{x-1}{2}\right) + \frac{n}{2}\log\left(1 - \frac{x-1}{x}\right) = \frac{n}{2}(x - \log x - 1)\Big|_{x=\gamma}$$

$$I(\gamma) = \frac{1}{2}(\gamma - \log\gamma - 1)$$

Compare

$\frac{(\gamma-1)^2}{4}$

set $\gamma = 3$.

The CLT rate func. $= 1$

The LD rate function $I(3) = 0.451$

info lect.

Cramer's Theorem

$X_1, \dots, X_n, \dots$ iid

$E[X_i] = m_X$

$S_n = \sum_{i=1}^{n} X_i$

$$P\left(\underbrace{\frac{S_n}{n} - m_X > \epsilon}_{\frac{S_n}{n} > \overbrace{\epsilon + m_X}^{\gamma}}\right) \overset{\cdot}{=} e^{-n \inf_{x>\gamma} I(x)}$$

asymptotic approximation, $n \to \infty$

log → $x \frac{1}{n}$ → $\lim_{n\to\infty}$

the asymptotic will become equality.

$$I(x) = \sup_{\theta} \left\{ \theta x - \log M_X(\theta) \right\}$$

ex. $X_i \sim N(0,1)$

$$Y_n = \sum_{i=1}^{n} X_i^2$$

$$\frac{Y_n}{n} \xrightarrow[n\to\infty]{P} 1 \quad (\mathcal{LLN})$$

$$P\left(\frac{Y_n}{n} > \underbrace{1 + \epsilon}_{\gamma}\right) \doteq e^{-n I(\gamma)}$$

$\Rightarrow M_{X^2}(\theta) = E\left[e^{\theta X^2}\right]$

$$= \int_{-\infty}^{+\infty} e^{\theta x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \times \frac{\frac{1}{\sqrt{1-2\theta}}}{\frac{1}{\sqrt{1-2\theta}}}$$

$e^{-\frac{1}{2}(1-2\theta)x^2}$

$$= \frac{1}{\sqrt{1-2\theta}}$$

$$\Rightarrow I(x) = \sup_{\theta} \left\{ \theta x + \frac{1}{2} \ln(1-2\theta) \right\}$$

$x - \frac{1}{1-2\theta_0} = \theta$

Gartner and Ellis (1984)

they wanted to remove assumptions.

$\Rightarrow \{Y_n\}$ possibly dependent

Define $\varphi_n(\theta) = \frac{1}{n} \log E[e^{\theta Y_n}]$

Assumptions.1

$$\lim_{n \to +\infty} \varphi_n(\theta) = \varphi(\theta)$$

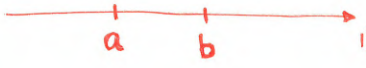
$$D_\varphi = \{\theta : \varphi(\theta) < \infty\}$$

Assumption.2

$\varphi_n(\theta)$ is differentiable on D_φ.

$$I(x) = \sup_\theta \{\theta x - \varphi(\theta)\}$$

Theorem (Gartner - Ellis)



$$\overline{\lim_{n \to +\infty}} \frac{1}{n} \log P(Y_n/n \in [a,b]) \leq -\inf_{x \in [a,b]} I(x)$$

$$\underline{\lim_{n \to +\infty}} \frac{1}{n} \log P(Y_n/n \in (a,b)) \leq -\inf_{x \in (a,b)} I(x)$$

ex. (AutoRegressive Process of order 1)

$$X_i = \alpha X_{i-1} + W_i \qquad (\alpha < 1)$$

$$X_0 = 0$$

This R.P. is stationary the sandwich will be equal from both sides.

W_i are iid over $\left[-\frac{1}{2}, \frac{1}{2}\right]$ → can have any kind of distribution.

$$E[W_i] = 0$$

Form $Y_n = \sum_{i=1}^{n} X_i$

NOT IID → Not Cramer Usage!

Approximate / Analyze $P\left(\frac{Y_n}{n} > \gamma\right)$

$$X_1 = W_1$$

$$X_2 = \alpha W_1 + W_2$$

$$X_3 = \alpha^2 W_1 + \alpha W_2 + W_3$$

$$\vdots$$

$$X_n = \alpha^{(n-1)} W_1 + \alpha^{(n-2)} W_2 + \cdots + \alpha W_{n-1} + W_n = \sum_{i=1}^{n} \alpha^{n-i} W_i$$

$$Y_n = \sum_{i=1}^{n} X_i = \left(1+\alpha+\cdots+\alpha^{n-1}\right) W_1 + \left(1+\alpha+\cdots+\alpha^{n-2}\right) W_2 + \cdots + W_n$$

$$= \frac{1-\alpha^n}{1-\alpha} W_1 + \frac{1-\alpha^{n-1}}{1-\alpha} W_2 + \cdots + W_n$$

$$\varphi_n(\theta) = \frac{1}{n} \log E[e^{\theta Y_n}] \qquad Y_n \text{ is a sum of independent R.V.'s.}$$

$$= \frac{1}{n} \log \prod_{i=1}^{n} M_{W_i}\left(\theta \frac{1-\alpha^{n-i}}{1-\alpha}\right)$$

as $n \to \infty$, $\alpha^{n-i} \to 0$

$$= \frac{1}{n} \cdot n \log M_W\left(\frac{\theta}{1-\alpha}\right)$$

Assume $W_i = \begin{cases} \frac{1}{2}, & \text{with prob. } 1/2 \\ -1/2, & \text{" " "} \end{cases}$

$$= \log M_W\left(\frac{\theta}{1-\alpha}\right)$$

$$= \log \; e^{\frac{\theta}{1-\alpha}\frac{1}{2}} \frac{1}{2} + e^{-\frac{\theta}{1-\alpha}\frac{1}{2}} \frac{1}{2}$$

$$= \log \cosh\left(\frac{\theta}{1-\alpha}\right)$$

$$I(x) = \sup_{\theta} \left\{ \theta x - \log \cosh\left(\frac{\theta}{1-\alpha}\right) \right\}$$

$$x = \frac{\frac{1}{2} e^{\frac{\theta}{1-\alpha}\frac{1}{2}} \left(\frac{1}{2(1-\alpha)}\right) - \frac{1}{2}\left(\frac{1}{2(1-\alpha)}\right) e^{\frac{-1}{2}\frac{\theta}{1-\alpha}}}{\cosh\left(\frac{\theta}{1-\alpha}\right)}$$

$$= \frac{1}{2(1-\alpha)} \frac{\sinh\left(\frac{\theta}{1-\alpha}\right)}{\cosh\left(\frac{\theta}{1-\alpha}\right)} \to \tanh\left(\frac{\theta}{1-\alpha}\right)$$

Sanov's Theorem

Let $X_1, \dots, X_n$ iid with ~~cdf/pdf~~ $Q(x)$

and $\mathcal{E}$ is a set of cdfs $P(x)$

↓ does not include Q.

Form empirical cdf:

$$L_n(x) = \frac{1}{n} \sum \delta_{X_i}(x)$$

like what the objective of MLE is!

$$P(L_n \in \mathcal{E}) \doteq e^{-n \inf_{P \in \mathcal{E}} D(P \| Q)}$$

if D_{KL} ↘ $\Rightarrow$ possibility that L_n lies in $\mathcal{E}$ ↗

Ex. 3 (Cover, p. 365)

A fair coin toss $X_i \in \{0, 1\}$

$n = 1000$

find $P(\sum_{i=1}^{1000} X_i > 700) = P(\frac{\sum X_i}{1000} > 0.7)$

converges → $\frac{1}{2}$

set $\mathcal{E}$ constist of all $p(x)$ which $E_{x \sim p(x)}[X_i] > 0.7$. ($P(X_i = 1) = p$)

$$e^{-n \inf_{p \geq 0.7} D((p, 1-p) \,/\!/\, (0.5, 0.5))}$$

$$\underbrace{D((0.7, 0.3) \,/\!/\, (0.5, 0.5))}_{0.119}$$

$$\Rightarrow \doteq e^{-1000 \times 0.119} = e^{-119}$$ ⇝ such low probability!! ☺

---

ex.24 $X_i \sim$ iid $\mathcal{N}(0,1)$

Sample average power ← $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i^2$ → 1

isn't it energy?

Approximate $P\left(\frac{Y_n}{n} > 5\right)$

$\frac{Y_n}{n}$: sample average power

first define $\mathcal{E}_\epsilon$.

average power

$$\mathcal{E}_\epsilon = \left\{ P(x) : \overbrace{E[X_i^2]} \geq 5 \right\}$$ ✓

$$\inf_{P_X(x) \in \mathcal{E}_\epsilon} D(P_X \,/\!/\, \mathcal{N}(0,1))$$

we have no idea about the family of P_X. → Using Lagrange Multipliers.

$$\mathcal{L}(P_X, \lambda, \mu) = \int_{-\infty}^{+\infty} P_X(x) \log \frac{P_X(x)}{q_X(x)} \, dx$$

$$+ \lambda \left[\int_{-\infty}^{+\infty} x^2 P_X(x) \, dx - 5 \right]$$

$$+ \mu \left[\int_{-\infty}^{+\infty} P_X(x) \, dx - 1 \right]$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial P_X} = \log \frac{P_X(x)}{q_X(x)} + 1 + \lambda x^2 + \mu$$

info lect.

project • just several bounds on an application

$$P(T(X^n) > \gamma) \doteq e^{-n D(P^* \| Q)}$$

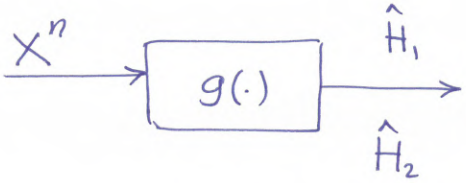
Hypothesis Testing

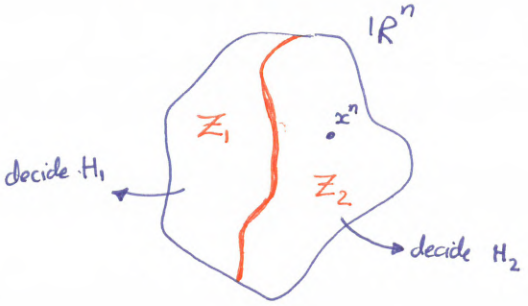
$X_1, \dots, X_n$ iid

H_1: $X_i \sim P_1(x)$

H_2: $X_i \sim P_2(x)$

decision rule → $g(X^n)$





4 outcomes

decide H_1	H_1 is true	
H_2	H_2	correct decision
H_1	H_2	wrong decision
H_2	H_1	wrong decision

in pattern recognition P_{FA} and P_{miss} are trade-off for each other.

$$P_{FA} = P(\text{decide } H_1 \mid H_2 \text{ is true}) = P_2(X \in Z_1)$$

↳ actually there is no signal out there, but we wrongly decide that there is!

$$P_{miss} = P(\text{decide } H_2 \mid H_1 \text{ is true}) = P_1(X \in Z_2)$$

↳ Actually there were a signal but we missed it.

Neyman – Pearson Criterion

$$\min P_{miss}$$

$$(Z_1, Z_2): P_{FA} \leq \epsilon \quad (\epsilon > 0)$$

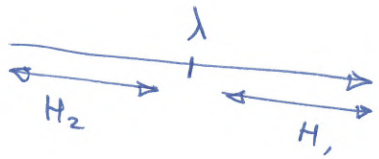
$$J(Z_1, Z_2, \lambda) = P_1(\underline{x} \in Z_2) + \lambda\left(P_2(x \in Z_1) - \epsilon\right)$$

$$= \sum_{x \in Z_2} P_1(\underline{x}) + \lambda\left(\sum_{x \in Z_1} P_2(x) - \epsilon\right)$$

$$\sum_{x \in Z_2} P_1(\underline{x}) = 1 - \sum_{x \in Z_1} P_1(\underline{x})$$

Decide H_1 if $P_1(\underline{x}) > \lambda P_2(\underline{x})$

" H_2 if $P_1(\underline{x}) < \lambda P_2(\underline{x})$



$$P_1(\underline{x}) \underset{H_2}{\overset{H_1}{\gtrless}} \lambda \underbrace{P_2(\underline{x})}_{>0}$$

a R.V.

$$\Lambda \triangleq \underbrace{\frac{P_1(\underline{x})}{P_2(\underline{x})}}_{\text{Likelihood Ratio}} \underset{H_2}{\overset{H_1}{\lessgtr}} \lambda$$

λ → Lagrange Multiplier = Threshold

use the constant $P_2(\underline{x} \in Z_1) = \epsilon$

LRT (LR Test)

If X_i's are iid:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\log\frac{P_1(x_i)}{P_2(x_i)}}_{\frac{1}{n}\mathcal{L}(x)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{\log\lambda}{n}$$

If X_i are from H_1, then

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{P_1(x_i)}{P_2(x_i)} \xrightarrow[\text{on long run}]{\text{Converge in } P} D_{KL}(P_1\|P_2)$$

If X_i are from H_2, then

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{P_1(x_i)}{P_2(x_i)} \xrightarrow[\text{on long run}]{\text{in } P} -D_{KL}(P_2\|P_1)$$

Using AEP (LLN in Information Theory)

$$P_{FA} = P_2\left(\frac{1}{n}\mathcal{L}(x) > \frac{1}{n}\log\lambda\right) \overset{\cdot}{=} e^{-n} \overset{?}{\to} \min_{P\in\mathcal{E}} D(P\|P_2)$$

($\overset{\cdot}{=}$: asymptotically; P_2: true probability)

Using Sanov's Theorem

$$\mathcal{E} = \left\{ P(x) : E_P\left[\underbrace{\frac{1}{n}\mathcal{L}(x)}_{\frac{1}{n}\sum_{i=1}^{n}\log\frac{P_1(x_i)}{P_2(x_i)}}\right] > \frac{\ln\lambda}{n} \right\}$$

$$\sum_x P_{(x)} \log \frac{P(x)}{P_2(x)} - P_{(x)} \log \frac{P(x)}{P_1(x)} > \frac{\log \lambda}{n}$$

i.i.d

n will be ommitted

$$D(P\|P_2) - D(P\|P_1) > T$$

$$\mathcal{E} = \left\{ P(x) : D(P\|P_2) - D(P\|P_1) > T \right\}$$

So:

$$\min \; D(P\|P_2)$$

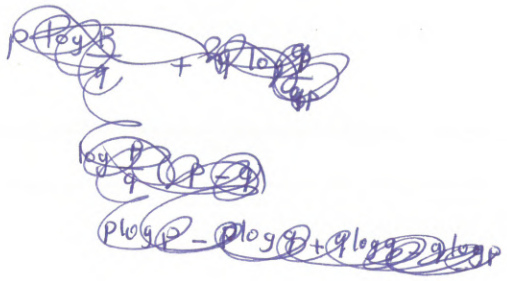
$$P: \; D(P\|P_2) - D(P\|P_1) \geq T$$

$$\Rightarrow J(p(x), \lambda, \mu) = \sum_x p(x) \log \frac{P(x)}{P_2(x)} + \lambda \left(\sum p(x) \log \frac{P_1(x)}{P_2(x)} - T \right) + \mu \left(\sum_x p(x) - 1 \right)$$

$$\frac{\partial J}{\partial p(x)} = \log \frac{P(x)}{P_2(x)} + 1 + \lambda \log \frac{P_1(x)}{P_2(x)} + \mu = 0$$

$$p(x) = P_2(x) \; e^{-1-\mu-\lambda \log \frac{P_1(x)}{P_2(x)}}$$

$$\sum p(x) = 1$$



$$P^*(x) = \frac{P_2^{1+\lambda}(x) \cdot P_1^{-\lambda}(x)}{\sum_t P_2^{1+\lambda}(t)\, P_1^{-\lambda}(t)}$$

- λ is found from the constraint.

$$D(P \| P_2) - D(P \| P_1) = T$$

info. lect.

Hypothesis Testing Problems

$X_i \sim$ iid $\quad i = 1 \dots n$

$H_1: \; P_1(x)$

$H_2: \; P_2(x)$

$\ell(\underline{x})$ — a R.V.

Neyman Pearson Criterion → $\frac{1}{n} \log \frac{P_1(\underline{X})}{P_2(\underline{X})} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{1}{n} \log \frac{\lambda}{1}$

$$P_{FA} = P_2\left(\underline{X} \in Z_1\right) = P_2\left(\frac{1}{n} \log \frac{P_1(\underline{X})}{P_2(\underline{X})} > \frac{\log \lambda}{n}\right) \doteq e^{-n D(P_\lambda^* \| P_2)}$$

$$P_{miss} = P_1\left(\underline{X} \in Z_2\right) = P_1\left(\ell(\underline{X}) < \frac{\log \lambda}{n}\right) \doteq e^{-n D(P_\lambda^* \| P_1)}$$

Sanov's Theorem

where $P_\lambda^*(x) = \dfrac{P_1^{\lambda}(x)\, P_2^{1-\lambda}(x)}{\sum_x P_1^{\lambda}(t)\, P_2^{1-\lambda}(t)}$

with λ such that $D(P_\lambda^* \| P_1) - D(P_\lambda^* \| P_2) = \frac{\log \lambda}{n}$

$D(P_\lambda^* \| P_1) - D(P_\lambda^* \| P_2) = \frac{\log \lambda}{n}$

if $\lambda = \theta \Rightarrow P_\lambda^*(x) = P_2(x) \Rightarrow \begin{cases} P_{FA} = 1 \\ P_{miss} \doteq e^{-n D(P_2 \| P_1)} \end{cases}$

if $\lambda = 1 \Rightarrow P_\lambda^* = P_1(x) \Rightarrow \begin{cases} P_{FA} \doteq e^{-n D(P_1 \| P_2)} \\ P_{miss} = 1 \end{cases}$

Chernoff - Stein Lemma

(Large Deviation extension of Nayman Pearson Lemma)

$$P_{FA} = P_2\left(\ell(\underline{X}) > \frac{\log \lambda}{n} \right) \leq \epsilon \quad , \quad 0 < \epsilon < 1/2$$

Chernoff-stein →

$$P_{miss} = P\left(\ell(\underline{X}) < \frac{\log \lambda}{n} \right) \doteq e^{-n D(P_2 \| P_1)}$$

Likelihood Ratio Test

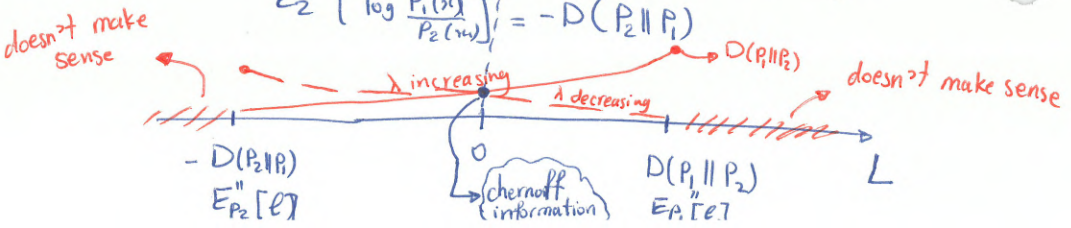
$$\frac{1}{n} \log \frac{P_1(\underline{X})}{P_2(\underline{X})} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{\log \lambda}{n} \longrightarrow E_1\left[\log \frac{P_1(X)}{P_2(X)} \right] = D(P_1 \| P_2) > 0$$

large deviation rate function (I)

$$E_2\left[\log \frac{P_1(x)}{P_2(x)} \right] = -D(P_2 \| P_1)$$

$H_1: \ell \sim P_1(L)$

$H_2: \ell \sim P_2(L)$



Bayesian criterion

π_1 , π_2 priors

$$P(\text{error}) = P_2\left(\ell(\underline{x}) > \frac{\log\lambda}{n}\right)\pi_2 + \pi_1 P_1\left(\ell(\underline{x}) < \frac{\log\lambda}{n}\right)$$

$$\doteq \pi_2 e^{-nD(P_\lambda^* \| P_2)} + \pi_1 e^{-nD(P_\lambda^* \| P_1)}$$

dominate? dominate?

$$\doteq e^{-n \min\{D(P_\lambda^* \| P_2),\ D(P_\lambda^* \| P_1)\}}$$

to maximize minimum of the two : @ Chernoff Information

$$\doteq e^{-n C(P_1, P_2)}$$

where $D(P_\lambda^* \| P_1) = D(P_\lambda^* \| P_2)$

Fully Bayesian Criterion

$$\log \frac{P_1(\underline{x})}{P_2(\underline{x})} \underset{H_2}{\overset{H_1}{\gtrless}} \log \frac{\pi_2}{\pi_1}$$

$\ell(x)$

LLRT

→ Chernoff Bound

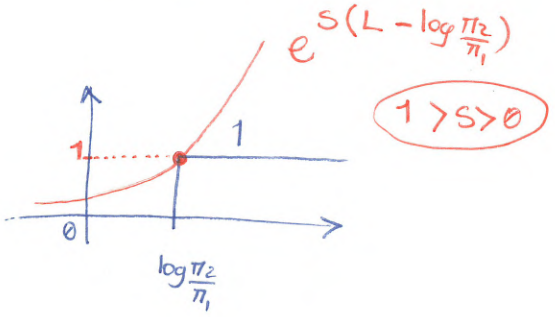
$$P_2\left(\underbrace{\ell(\underline{x})}_{\text{RV.}} > \log\frac{\pi_2}{\pi_1}\right)$$

H_1: $\ell \sim P_1(L)$

H_2: $\ell \sim P_2(L)$

$$P_2\left(\ell > \log\frac{\pi_2}{\pi_1}\right) = \sum_{L > \log\frac{\pi_2}{\pi_1}} P_2(L)$$

$$= \sum_{L \in \mathbb{R}'} P_2(L)\; u\left(L - \log\frac{\pi_2}{\pi_1}\right)$$



$$\leq \sum P_2(L)\, e^{s\left(L - \log\frac{\pi_2}{\pi_1}\right)}$$

$$= e^{-s\log\frac{\pi_2}{\pi_1}} \underbrace{\sum e^{sL} P_2(L)}_{\text{MGF } \varphi_2(s)}$$

Laplace Transform of $P_2(\cdot)$

$$= e^{-s\log\frac{\pi_2}{\pi_1}} + \varphi_2(s)$$

$\log \varphi_2(s)$

$$\varphi_2(s) = E_2\left[e^{sL}\right] \qquad \sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_2(x_i)}$$

$$= E_{P_2(x)}\left[e^{s\log P_1(x)/P_2(x)}\right] \qquad 0 \leq s \leq 1$$

$$\overset{x_i \sim \text{iid}}{=} \left\{ E_{P_2(x)}\left[e^{s\log\frac{P_1(x)}{P_2(x)}} \right] \right\}^n = \left\{ \sum_x P_1^{s}(x) P_2^{1-s}(x) \right\}^n$$

$$\Sigma_{P_2} \log\left(\frac{P_1}{P_2}\right)^s =$$

$$= e^{-s \log \frac{\pi_2}{\pi_1} + \underbrace{\varphi_2(s)}_{\log P_2(s)}}$$

$$= e^{-s \log \frac{\pi_2}{\pi_1} + n \log \left\{ \sum P_1^s(x) P_2^{1-s}(x) \right\}}$$

→ minimize wrt to $0 \le s \le 1$

Let's Tighten the Bound!

$$P_2\left(\ell(\underline{X}) > \log \frac{\pi_2}{\pi_1} \right) \le e^{-n \max_{0 \le s \le 1} \left\{ \frac{s}{n} \log \frac{\pi_2}{\pi_1} - \log \left\{ \sum P_1^s(x) P_2^{1-s}(x) \right\} \right\}}$$

if $n \to \infty$ goes to zero!

by letting #samples to ∞, priors effects wash away!

$$\rightsquigarrow e^{-n \left\{ - \min_{0 \le s \le 1} \log \left(\sum P_1^s(x) P_2^{1-s}(x) \right) \right\}}$$

Chernoff Information

agrees with

$$D(P_\lambda^* \| P_1) = D(P_\lambda^* \| P_2)$$