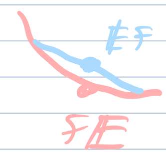
1) Intro to information theory

1.1 Random Vars

Jensens inequality:
For F convex, ie $F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y)$

we have $\mathbb{E}F(X) \geq F(\mathbb{E}X)$

e.g. $\langle x^2 \rangle \geq \langle x \rangle^2$



1.2 Entropy

Entropy $= \mathop{\mathbb{E}}_{x \sim p} \log \frac{1}{p(x)}$

KL Div, $D_{KL}(q \| p) = \sum_x q(x) \log \frac{q(x)}{p(x)} = \mathop{\mathbb{E}}_{q} \log\left(\frac{q}{p}\right)$

D_{KL} is not symmetric

From Rezende's notes:

We generally take q_θ the density of the generative model and p to be the "true" density

generally want to modify θ so that $q_\theta \to p$

$KL(q \| p)$ is # of bits to communicate q given that the receiver knows p

KL is unique divergence satisfying

i) locality:

$$D(q \| p) = \int dx \; F(q, p, x)$$

possible to add dependence on ∇p, ∇q etc

ii) invariance:
under $x \to x' = \varphi(x)$ we have D invariant

$$\Rightarrow \int dx \; F(q, p, x) = \int dx' \, F\left(\frac{q \circ \varphi^{-1}(x')}{\left|\det \frac{\partial x'}{\partial x}\right|}, \frac{p \circ \varphi^{-1}(x')}{\left|\det \frac{\partial x'}{\partial x}\right|}, \varphi^{-1}(x') \right)$$

$\Rightarrow \mathcal{F}$ must take the form $\mathcal{F}\left(\frac{q}{p}\right)\mu(x) \Rightarrow \mathcal{F}\left(\frac{q}{p}\right)p$ or $\mathcal{F}\left(\frac{q}{p}\right)q$

transforming as a measure

iii) Subsystem independence (ie additivity of indep sub-domains)

$\Rightarrow \quad \mathcal{F}\left(\frac{q}{p}\right) = \log \frac{q}{p}$

Back to Montanari

Entropy H satisfies

1) $H_X \geq 0$ proof: $E - \log p \geq -\log E\, p \geq 0$
2) $H_X = 0$ only for $p(x) = \delta_x$
3) Among all distributions $p(x)$ H is maximized for $p = 1/M$

proof: $D_{KL}(p \,|\, \bar{p}) = \log_2 M - H(p) \geq 0$, $\bar{p}$ uniform

Q: can I get stronger bounds from other $\bar{p}$?

4) For X, Y indep $H_{X,Y} = H_X + H_Y$
5) For X, Y generic $H_{X,Y} \leq H_X + H_Y$
6) For X_1, X_2 disjoint take $q_{1,2} = \text{Prob } x \in X_{1,2}$ resp.
then $H_X = H(q) + \tilde{H}(q, r)$

$H(q) = -q_1 \log q_1 - q_2 \log q_2$, $\tilde{H}(q,r) = -q_1 \sum_{x \in X_1} r_1(x) \log r_1(x) - q_2 \sum_{x \in X_2} r_2(x) \log r_2(x)$

1.3 Sequences of random variables

Def entropy rate $h_X = \lim_{N \to \infty} H[X_1 \ldots X_N]/N$

e.g. 1: X_t indep $\Rightarrow P_N(x_1, \ldots, x_N) = \prod_{t=1}^{N} p(x_i) \Rightarrow h_X = H(p)$

e.g. 2: Markov Chain

$\{p_1(x), x \in \mathcal{X}\}$ an initial state

$\{w(x \to y)\}_{x,y \in \mathcal{X}}$ are transition probabilities, $\sum_y w(x \to y) = 1$

$\Rightarrow P_N(x_1, \ldots x_N) = p_1(x_1) \prod_{t=1}^{N-1} w(x_t \to x_{t+1})$ $\quad \lim_{t \to \infty} p_t(x) = p^*(x)$

then $h_X = -\sum_x p^*(x) \sum_y w(x \to y) \log w(x \to y) = H_{Y|X}$ $\leftarrow$ ie sum over all letters weighted by $p(x)$ and use entropy $H(x_{t+1}|x)$

$w(x \to y) = p(y|x)$

1.4 Correlated vars & Mutual info

Conditional entropy

$$H_{Y|X} := -\sum_x p(x) \sum p(y|x) \log p(y|x)$$ no log on $p(x)$!

N.B. $H_{X,Y} = -\sum_{x,y} p(x,y) \log(p(x,y)) = -\sum_{x,y} p(x) p(y|x) \log p(x) p(y|x)$

$= H_{Y|X} + \sum_x p(x) \log p(x) \cancel{\sum_y p(y|x)} = \underline{H_{Y|X} + H_X}$

Mutual info:

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

"reduction in uncertainty of X | knowledge of Y"

$= D_{KL}(p(x,y) \| p(x)p(y))$

alt: $H_{X|Y} = H_X - I_{X,Y}$

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

"the decrease in Y's entropy from conditioning on X"

$$= H_X - H_{X|Y}$$

$$I_{X,Y} = \mathbb{E}_{X,Y}\left[-\log \frac{p(x)p(y)}{p(x,y)}\right] \geq -\log \mathbb{E}_{X,Y}\left[\frac{p(x)p(y)}{p(x,y)}\right] \geq 0$$

$\int p(x) \int p(y)$

Data processing inequality: for Markov chain $X \to Y \to Z$

$$\Rightarrow p(x,y,z) = p_1(x)\, w_2(x \to y)\, w_3(y \to z)$$

Lemma: $I_{X,(YZ)} = I_{X,Z} + I_{X,Y|Z}$

$$-\mathbb{E}_{X,Y,Z} \log \frac{p(x)p(y,z)}{p(x,y,z)} = -\mathbb{E}_{X,Z} \log \frac{p(x)p(z)}{p(x,z)} - \mathbb{E}_{X,Y,Z} \log \frac{p(x,z)p(y,z)}{p(x,y,z)p(z)}$$

$= I_{X,Z} + I_{X,Y|Z}$ ✓

here $I_{X,Y|Z} = -\mathbb{E}_{X,Y,Z} \log \frac{p(x|z)p(y|z)}{p(x,y|z)}$

$p(x,y,z)$

$$\Rightarrow I_{X,(YZ)} = I_{X,Z} + \underbrace{I_{X,Y|Z}}_{\geq 0}$$

$$= I_{X,Y} + \cancel{I_{X,Z|Y}}^{\,0} \quad \text{By Markov}$$

$\Leftrightarrow \quad I_{X,Z} \leq I_{X,Y}$

Take $Z = f(Y) \Rightarrow I_{X,Y} \geq I_{X,f(Y)}$

Fano's inequality: Relates the info loss in a noisy channel to the probability of mischaracterization error

take $X \to Y \to \hat{X}$ w/ $\hat{X} = g(Y)$ an estimate of X

let $E = \mathbb{1}_{X \neq \hat{X}}$, $P_e = Pr(X \neq \hat{X}) = \mathbb{E}(E)$

$H_{X,E|Y} = H_{X|Y} + \cancel{H_{E|X,Y}}^{\to 0}$ i) $H_{E|X,Y} = 0$ E is deterministic function of X,Y

$= \underbrace{H_{E|Y}}_{\leq H_E} + H_{X|E,Y}$ ii) $H_{E|Y} \leq H_E = \mathcal{H}(P_e)$

X is g(Y)

iii) $H_{X|E,Y} = (1-P_e) H_{X|E=0,Y} + P_e H_{X|E=1,Y}$

$= P_e H_{X|E=1,Y} \leq P_e \log(|X|-1)$

$H_{X|Y} = H_{E|Y} + H_{X|E,Y} \leq H_E + P_e H_{X|E=1,Y} \leq \mathcal{H}(P_e) + P_e \log(|X|-1)$

bound on uncertainty of X | Y

$\leq$ Uncertainty of $X \neq \hat{X}$ ie P_e

+

$P_{error} \cdot \mathcal{H}($uniform $-1))$

Exercise 1.6 $p(1) = 1-p$ for k values
$p(x) = \frac{p}{k-1}$

take Y indep of X $\Rightarrow$ $H(X|Y) = H(X)$

$\Rightarrow \mathcal{H}(P_e) + P_e \log(k-1) \geq H(X)$

if p small so $1-p > \frac{p}{k-1}$ guess 1 always

$\Rightarrow P_{error} = p \Rightarrow -p \log p - (1-p) \log(1-p) + p \log(k-1) \leq H(X)$

$H(X) = -(1-p)\log(1-p) - \cancel{(k)} \cdot \frac{p}{\cancel{k-1}} \log \frac{p}{k-1}$

$\Rightarrow$ Equality

1.5 Data Compression

Sequence $\underline{X} = \{X_1 \cdots X_N\}$ for $X_i \in \mathcal{X}$ finite alphabet

$\leftarrow$ source code

assume X_i are random

store a given realization $\underline{x} = \{x_1 \cdots x_N\}$ as compactly as possible

$$w: \mathcal{X}^N \to \{0,1\}^*$$
$$\underline{x} \to w(\underline{x})$$

Often we take a longer stream $\to$ blocks $\underline{x}^1 \cdots \underline{x}^r$

encode each block $w(\underline{x}^1) \cdots w(\underline{x}^r)$

need concatenation of blocks to be uniquely decodable

safe if $\forall x, x'$ $w(x)$ is not prefix of $w(x')$

"instantaneous codes"

$$L(w) = \mathbb{E}_{\underline{x} \in \mathcal{X}^N} \, l_w(\underline{x})$$

$\leftarrow$ length of $w(\underline{x})$

take $N=1$

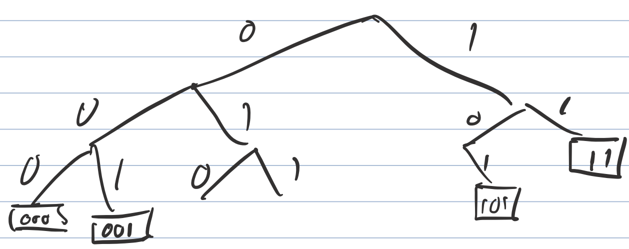
$\mathcal{X} = \{1, \dots, 8\}$

$p(i) = 2^{-i}$ $\quad i = 1 \cdots 7$
$p(8) = 2^{-7}$ $\quad i = 8$

x	$p(x)$	$w_1(x)$	$w_2(x)$
1	1/2	000	0
2	1/4	001	10
3	1/8	010	110
4	1/16	011	1110
5	1/32	100	1110
6	1/64	101	11110
7	1/128	110	111110
8	1/128	111	1111110

both instantaneous

$$L(w_1) = 3$$
$$L(w_2) = \sum_{i=1}^{7} 2^{-i} i + 8 \cdot 2^{-7} \approx 2$$



$\leftarrow$ binary tree where no codeword node has ancestor codewords

What is best w for a given source?

let L_N^* be optimal acheivable instantaneous code length. Then,

1. $H_{\underline{X}} \leq L_N^* \leq H_{\underline{X}} + 1$

2. If the source has finite entropy rate $h = \lim_{N\to\infty} \frac{1}{N} H_{\underline{X}}$

$$\lim_{N\to\infty} \frac{1}{N} L_N^* = h$$

Lemma "Kraft's inequality"

$$\sum_{\underline{x}\in\mathcal{X}^N} 2^{-l_w(\underline{x})} \leq 1$$

Follows from "set of all leaves of binary tree sum to one"

Conversely any set of lengths $\{l_w(\underline{x})\}_{\underline{x}\in\mathcal{X}^N}$ satisfying Kraft have a code

→ start from smallest $l_w(\underline{x})$ and take first binary seq. of that length.

Goal: Find codewords $l_w^*(\underline{x})$ that minimize L subject to Kraft

First, if l could be real-valued

$$\min_{l,\ \alpha\geq 0} \sum_{x} p(x)\, l(x) + \alpha\left(\sum_{x} 2^{-l(x)} - 1\right)$$

$\Rightarrow p(x) - \alpha\, 2^{-l} \log_2 = 0$

$l = -\log_2 p(x) - c$ $\quad c = 0$ from Kraft

$\Rightarrow l = \lceil -\log_2 p(x) \rceil$ also then works

$H_X \leq L \leq H_X + 1$ ✓

"Shannon code", close to optimal for long strings

Not ideal for shorter sequences

may assign super long codeword when shorter ones are available

there Huffman coding is optimal

requires $\Theta(|\mathcal{X}|^N)$ memory to enumerate all $\ell(\underline{x})$

1.6 Data Transmission

$$m \longrightarrow \boxed{\text{Enc}} \xrightarrow{X} \boxed{\text{Channel}} \xrightarrow{Y} \boxed{\text{Dec}} \longrightarrow m'$$

M bits — N bits — may not be a string of bits — M bits

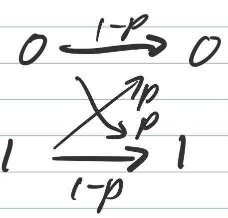
$N > M$

Can have a channel with insertions

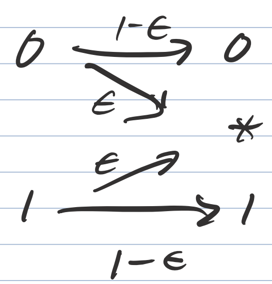
Consider a memoryless channel (noise acts indep on each bit)

$$Q(\underline{y}\,|\,\underline{x}) = \prod_{i=1}^{N} Q(y_i\,|\,x_i)$$

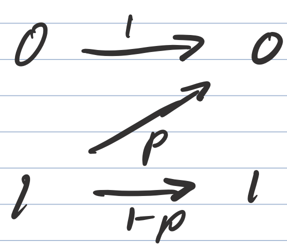
BSC:



BEC:



Z:



Channel capacity C:

$$\boxed{\max_{p(x)} I_{X,Y}}$$

reduction in uncertainty of Y | knowledge of X, vice versa

We will see C characterizes amount of info that can be transmitted faithfully through the channel

E.g. BSC, send a bit drawn from Bern(q)

$$\max_q I_{X,Y} = \sum_{x=\{0,1\}} p(x) \sum_{y=\{0,1\}} p(y|x) \log \frac{p(y|x)}{p(y)}$$

$$p(y=1) = p(y=1|x=1)p(x=1) + p(y=1|x=0)p(x=0)$$
$$= (1-p)(1-q) + pq$$
$$p(y=0) = (1-p)q + p(1-q)$$

$$\Rightarrow I_{X,Y} = q\cdot\left[(1-p)\log\frac{1-p}{(1-p)q+p(1-q)} + p\log\frac{p}{(1-p)(1-q)+pq}\right]$$
$$+(1-q)\cdot\left[p\log\frac{p}{(1-p)q+p(1-q)} + (1-p)\log\frac{1-p}{(1-p)(1-q)+pq}\right]$$

We see $\partial_q I_{X,Y} = 0$ when $q = 1/2$

Faster way

$$H(Y) - H(Y|X) = \mathcal{H}(\underbrace{(1-p)(1-\alpha)+p\alpha}_{p}) - \mathcal{H}(p)$$

N.B. $\partial_p \mathcal{H}(p) = \log_2 \frac{1-p}{p}$

$$\partial_\alpha = 0 \Rightarrow (2p-1)\log\frac{1-p}{p} \Rightarrow p = 0$$

$$\Rightarrow (2p-1)\alpha = p-1 \Rightarrow \alpha = 1/2$$

$$\Rightarrow \boxed{C = \mathcal{H}(1/2) - \mathcal{H}(p) = 1 - \mathcal{H}(p)}$$

E.g. BEC

$p(y=0) = q(1-\epsilon)$
$p(y=1) = (1-q)(1-\epsilon)$
$p(y=*) = \epsilon$

$$I_{X;Y} = q\left[(1-\epsilon)\log\frac{1-\epsilon}{q(1-\epsilon)} + \cancel{\epsilon\log\frac{\epsilon}{\epsilon}}\right] + (1-q)\left[(1-\epsilon)\log\frac{1-\epsilon}{(1-q)(1-\epsilon)} + \cancel{\epsilon\log\frac{\epsilon}{\epsilon}}\right]$$

$D_q I_{X;Y} = 0$ when $q = 1/2$

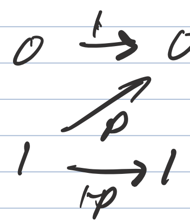
Faster way: $H(Y) - H(Y|X) = \cancel{\mathcal{H}(\epsilon)} + (1-\epsilon)\mathcal{H}(\alpha) - \cancel{\mathcal{H}(\epsilon)}$

$H(Y) = H(Y \overset{?}{\in} *) + \sum P(*?)\,\mathcal{H}(Y|\overset{?}{*})$ $\quad \partial_\alpha = 0 \Rightarrow (1-\alpha)\log_2\frac{1-\alpha}{\alpha} \Rightarrow \alpha = 1/2$

$= \mathcal{H}(\epsilon) + (1-\epsilon)\mathcal{H}(q)$

$\Rightarrow \boxed{C = 1-\epsilon}$

Eg Z-channel



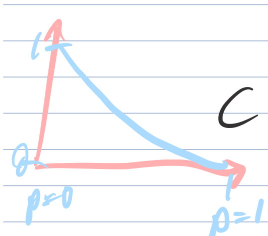
$\alpha = P(0)$ $\quad P(Y=0) = p + \alpha(1-p) \;\; [\alpha + p(1-\alpha)]$ $\quad P(Y=1) = (1-\alpha)(1-p)$

$\Rightarrow \max_\alpha \{H(Y) - H(Y|X)\} = \max_\alpha H(Y) - \sum_x H(Y|X=x)P(x)$

$= \max_\alpha \mathcal{H}((1-\alpha)(1-p)) - \cancel{\alpha \cdot H(Y|X=0)} - (1-\alpha)\mathcal{H}(p)$

$\frac{\partial}{\partial\alpha} = 0 \Rightarrow -(1-p)\log\frac{1-(1-\alpha)(1-p)}{(1-\alpha)(1-p)} + H(p) = 0 \Rightarrow \frac{1}{p} - 1 = 2^{H(p)/1-p}$

$\Rightarrow \alpha = 1 - \frac{1}{(1-p)(1+2^{H(p)/1-p})}$



$$C = \mathcal{H}\left(\frac{1}{1+2^{s(p)}}\right) - \frac{s(p)}{1+2^{s(p)}} = \log(1+2^{-s(p)}), \quad s(p) = \frac{\mathcal{H}(p)}{1-p}$$
$$= \log(1+(1-p)p^{p/1-p})$$

Assume each bit is random — surprisingly, Shannon's theorem shows that there is no loss in generality

$\{0,1\}^M \ni m \to \underline{x}(m) \in \{0,1\}^N$

2^M codewords in $\mathbb{F}_2^N$

$Q(\underline{y}|\underline{x}) = \prod_i Q(y_i|x_i)$

$R = \frac{M}{N}$ is the rate

$$P_B(m) = \sum_{\underline{y}} Q(\underline{y} \mid \underline{x}(m)) \, \mathbb{1}(d(\underline{y}) \neq m)$$

$$P_B^{max} = \max_m P_B(m) \qquad \text{"worst case"}$$

$$P_B^{av} = \frac{1}{2^m} \sum_{m \in \{0,1\}^m} P_B(m) \quad \leftarrow \text{more common}$$

Eg. 1 Repetition k (odd) times + majority

$$R = \frac{1}{k}$$

Exercise:

$$P_B^{av} = \sum_{r=\lceil \frac{k}{2} \rceil}^{k} \binom{k}{r} p^r (1-p)^{k-r}$$

Shannon, 1948

For every rate $R<C$, there is a sequence of codes C_N of length N

s.t. : $R_N \to R$ $\quad P_B^{avg} \to 0$ as $N \to \infty$

conversely, any such sequence has $R<C$

Intuition for the role of capacity

$$H_{\underline{y}|\underline{x}} = N H_{y|x} \Rightarrow 2^{N H_{y|x}} \text{ outputs}$$

need $d(y)$ to map all of them to m

possible outputs is NH_y

$\Rightarrow$ can distinguish $2^{N H_y} / 2^{N H_{y|x}}$ codewords

$$= 2^{N(H_y - H_{y|x})} = 2^{N I_{x,y}}$$

one needs to be able to send all 2^M codewords

$$\Rightarrow 2^M = 2^{NR} < 2^{N I_{x,y}}$$

$$\Rightarrow R < I_{x,y} \leq C$$

This also gives another interp of $I_{X,Y}$

can distinguish $2^{N I_{X,Y}}$ codewords

Facts about channel coding:

For p_1, p_2 indep channels

$$(p_1 \times p_2)((y_1,y_2)|(x_1,x_2)) = p_1(y_1|x_1)p_2(y_2|x_2)$$

$$C(p_1 \times p_2) = \sup_{p_{X_1,X_2}} I(X_1 X_2 ; Y_1 Y_2) = \sup_{p_{X_1,X_2}} I(X_1, Y_1) + I(X_2, Y_2)$$

$$\geq C(p_1) + C(p_2)$$

Also

$$\sup I(X_1 X_2 ; Y_1 Y_2) = H(Y_1 Y_2) - H(Y_1 Y_2 | X_1 X_2)$$

$$= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2)$$

$$\leq H(Y_1) + H(Y_2) - \text{''}$$

$$\sup = I(X_1 : Y_1) + I(X_2 : Y_2)$$

$$\Rightarrow C(p_1 \times p_2) \leq C(p_1) + C(p_2)$$

$$\Rightarrow C(p_1 \times p_2) = C(p_1) \times C(p_2)$$