$H_b(X) = -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\log_b \mathbb{P}(X=x)$ $= -\mathbb{E}[\log_b p(X)]$

$D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = -\mathbb{E}[\log q(X)] - H(X)$ (w.r.t. p)

$H(X,Y) = -\sum_{x,y}\mathbb{P}(X=x,Y=y)\log\mathbb{P}(X=x,Y=y)$ $= H(X)+H(Y)$ if $X\perp Y$

$I(X;Y) = D(p_{XY}\|p_X p_Y) = H(X)+H(Y)-H(X,Y)$

$H(X|Y) = -\sum_{x,y}\mathbb{P}(X=x,Y=y)\log\mathbb{P}(X=x|Y=y) = -\mathbb{E}[\log p_{Y|X}(X)]$

$D(p_{Y|X}\|q_{Y|X}|p_X) = \sum_x p(x)D(p_{Y|X=x}\|q_{Y|X=x})$

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

Inequalities

Gibbs: $-\sum p(x)\log p(x) \le -\sum p(x)\log q(x)$, $= \iff p=q$

Jensen: $a_i\ge 0$, f convex: $f\left(\frac{\sum a_i x_i}{\sum a_i}\right)\le\frac{\sum a_i f(x_i)}{\sum a_i}$ OR: $f(\mathbb{E}[X])\le\mathbb{E}[f(X)]$

Log-sum: $\sum_i a_i\log\frac{a_i}{b_i}\ge\left(\sum_i a_i\right)\log\frac{\sum a_i}{\sum b_i}$ for $a_i,b_i\ge 0$, $i\le n$. $=\iff\frac{a_i}{b_i}$ const.

Fano: $H(X|Y)\le H(1_{X\ne Y})+\mathbb{P}(X\ne Y)\log(|\mathcal{X}|-1)\le 1+\mathbb{P}(X\ne Y)\log(|\mathcal{X}|-1)$

entropy
1) $0\le H(X)\le\log(|\mathcal{X}|)$
2) $0\le H(X|Y)\le H(X)$ ($X=f(Y)$; $X\perp Y$)
3) $H(X_1\ldots X_n)=\sum_i^n H(X_i|X_{i-1},\ldots,X_1)\le\sum_i^n H(X_i)$
4) $H(f(X))\le H(X)$, $=\iff f$ bijective
5) X,X' iid $\Rightarrow \mathbb{P}(X=X')\ge 2^{-H(X)}$
6) $H(X)$ concave in p_X

information
1) $I(X;Y)\ge 0$, $=\iff X\perp Y$
2) $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$
3) $I(X_1\ldots X_n;Y)=\sum_i^n I(X_i;Y|X_{i-1},\ldots,X_1)$
4) $I(X;Y)\ge I(X;f(Y))$ $\forall f$
5) $(X\perp Y)|Z \Rightarrow I(X;Z)\ge I(X;Y)$ ← data proc

divergence
1) $D(p\|q)\ge 0$, $=\iff p=q$
2) $D(p_{X,Y}\|p_{\hat X,\hat Y})\ge D(p_X\|p_{\hat X})$
3) $D(p_{XY}\|p_{\hat X\hat Y})=D(p_{Y|X}\|p_{\hat Y|\hat X}|p_X)+D(p_X\|p_{\hat X})$
4) $D(p_{Y|X}\|p_{\hat Y|\hat X}|p_X)=D(p_X p_{Y|X}\|p_X p_{\hat Y|\hat X})$
5) convexity $D(\lambda p_1+(1-\lambda)p_2\|\lambda q_1+(1-\lambda)q_2)\le\lambda D(p_1\|q_1)+(1-\lambda)D(p_2\|q_2)$

Typical seqs

$T_n^\varepsilon := \{(x_1,\ldots x_n)\in\mathcal{X}^n : |-\frac{1}{n}\log p_{X_1\ldots X_n}(x_1,\ldots,x_n)-H(X)|\le\varepsilon\}$

$-\frac{1}{n}\log p_{X_1\ldots X_n}(X_1\ldots X_n)\to H(X)$ in prob — Weak AEP 1

$\forall\varepsilon>0\ \exists N\ \forall n\ge N$:
1) $\forall\bar x\in T_n^\varepsilon$: $2^{-n(H(X)+\varepsilon)}\le p_{\bar X}(\bar x)\le 2^{-n(H(X)-\varepsilon)}$
2) $(1-\varepsilon)2^{n(H(X)-\varepsilon)}\le|T_n^\varepsilon|\le 2^{n(H(X)+\varepsilon)}$
3) $\mathbb{P}(\bar X\in T_n^\varepsilon)\ge 1-\varepsilon$.

+ same for $S_n^\varepsilon :=$ smallest set w/ $\mathbb{P}(\bar X\in S_n^\varepsilon)\ge 1-\varepsilon$ except $(1-2\varepsilon)$ in $\le|S_n^\varepsilon|$. proof: indep on log, $\mathbb{E}[-\log p_X(X)]=H$, WLLN

Codes

$c:\mathcal{X}\to\mathcal{Y}^*$ is a code, is d-ary if $|\mathcal{Y}|=d$. $\ell_x:=|c(x)|$

unambiguous: c is injective, **uniq. decodable:** $c^*:\mathcal{X}^*\to\mathcal{Y}^*$ by $c(x_1\ldots x_n)\mapsto c(x_1)\ldots c(x_n)$ is injective

prefix: no codeword $c(x)$ is a prefix of another codeword $c(y)$.

optimal for $X\sim p$: uniq. dec. + $\forall$ other uniq dec c', $\mathbb{E}[|c(X)|]\le\mathbb{E}[|c'(X)|]$.

$H_d(X)\le\mathbb{E}[|c(X)|]$ for c uniq-dec + d-ary. "Kraft-Mcmillan $\Leftarrow$:"

Shannon 1: $\forall\varepsilon>0\ \exists n$ $c:\mathcal{X}^n\to\{0,1\}^*$ st $\frac{1}{n}\mathbb{E}[|c(X_1\ldots X_n)|]\le H(X)+\varepsilon$ & c uniq dec [c^* inj].

Kraft-Mcmillan: c uniq dec $\Rightarrow\sum_x|\mathcal{Y}|^{-\ell_x}\le 1$. if $(\ell_x)_{x\in\mathcal{X}}$ st $\sum_x|\mathcal{Y}|^{-\ell_x}\le 1$, then $\exists$ a prefix code c w/ lengths ℓ_x.

$\forall X:\mathcal{X}\to\mathcal{Y}$ $\exists$ opt code c^* st $H_d(X)\le\mathbb{E}[|c^*(X)|]<H_d(X)+1$.

Shannon's code: sort probs $p_1\ge\ldots\ge p_m$, def: $\ell_r=\lceil-\log_d(p_r)\rceil$, $c_S(x_r)$ = first ℓ_r digits of d-ary expansion of n_r. $n_r=\sum_i^{r-1}p_i$, $d=|\mathcal{Y}|$
- not optimal
- $O(k\log k)$ to sort

$H_d(X)+D_d(p\|q)\le\mathbb{E}[|c_q(X)|]\le H_d(X)+D_d(p\|q)+1$ (real dist; $\forall q$; code based on q).

Elias: same setup as Shannon, except $n_r:=\sum_{i<r}p_i+p_r/2$, ℓ_r+1
↳ not opt: $H_d(X)+1\le\mathbb{E}[|c_E(X)|]\le H_d(X)+2$, but no sorting

Huffman: sort $p_1\ge\ldots\ge p_m$, give each a node labelled by prob
- optimal. repeat: join 2 least likely nodes to (tree). read off code: code is path down tree.

optimality proof:
1) a canonical code always exists
2) induct on $|\mathcal{X}|$: ✓ for 2; for >2: given p, p' = w/ 2 smallest merged; c^P: canonical for p, $(c^P)'$ = merged; $e^{P'}$: opt for p', e^P = unmerged, prove $c^P=e^P$

a canonical code:
- optimal
- $p_j>p_k\Rightarrow|c(x_j)|\le|c(x_k)|$
- 2 longest codewords have same len
- " " differ only in last digit

$\Rightarrow$ expansion gives opt code $\Rightarrow$ Huffman opt as based on expansion

Channel codes

an (m,n) code for a DMC $(\mathcal{X},M,\mathcal{Y})$ is (c,d): $c:\{1\ldots m\}\to\mathcal{X}^n$ encoder, $d:\mathcal{Y}^n\to\{1\ldots m\}$ decoder

$\{c(i):i\}$ = code book, message set, $c(i)$ = codeword for i.

rate $\rho(c,d):=\frac{1}{n}\log_{|\mathcal{X}|}(m)$.

$\varepsilon_i:=\mathbb{P}(d(\bar Y)\ne i\,|\,c(i)=\bar X)$ = error on message i ($\bar X$: length n vector of iid copies of X.)

$\varepsilon_{max}:=\max_i\varepsilon_i$, $\bar\varepsilon:=\sum\varepsilon_i/m$.

a rate $R>0$ is **achievable** if $\forall\varepsilon>0\ \exists m,n$ a (m,n) code (c,d) with $\rho(c,d)>R-\varepsilon$ and $\varepsilon_{max}<\varepsilon$.

DMC

$(\mathcal{X},M,\mathcal{Y})$ (in, stochastic matrix, out; possibly empty). (X,Y) realises the DMC if $M=(p_{Y|X}(y|x))_{x,y}$.

Lossless if $\mathcal{Y}$ splits into disjoint $Y_1\ldots Y_{|X|}$ st $\forall i\le|X|$, $\mathbb{P}(Y\in Y_i|X=x_i)=1$. $\Rightarrow H(X|Y)=0$ vs **useless:** $X\perp Y$ so no info in Y about input $X\Rightarrow H(X|Y)=H(X)$

channel capacity: $C:=\sup I(X;Y)=\sup(H(X)-H(X|Y))=\sup(H(Y)-H(Y|X))$.

$0\le C\le\min\{H(X),H(Y)\}$ calculating C: calc $H(Y|X)$ using $\sum_x H(Y|X=x)\mathbb{P}(X=x)$ then want to use $H(Y)\le H(\text{unif})=1$: find Y in terms of X's dist, find dist of X to make Y uniform. Do everything in terms of X's dist.

Shannon 2 (noisy channel coding) for $(\mathcal{X},M,\mathcal{Y})$ with capacity C, $R>0$ is achievable $\iff R\le C$

Joint AEP: $J_\varepsilon^n:=\{(\bar x,\bar y)\in\mathcal{X}^n\times\mathcal{Y}^n:\max\{|-\frac{\log p_{\bar X,\bar Y}(\bar x,\bar y)}{n}-H(X,Y)|, |-\log p_{\bar X}(\bar x)/n-H(X)|, |-\log p_{\bar Y}(\bar y)/n-H(Y)|\}<\varepsilon\}$.
- $\lim_{n\to\infty}\mathbb{P}(\bar X,\bar Y\in J_\varepsilon^n)=1$
- $|J_\varepsilon^n|\le 2^{n(H(X,Y)+\varepsilon)}$
- $X'\perp Y'$, X,Y same marginals as X',Y' (but not indep) $\Rightarrow\exists n_0\ \forall n\ge n_0$ $(1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}\le\mathbb{P}((X',Y')\in J_\varepsilon^n)\le 2^{-n(I(X;Y)-3\varepsilon)}$.

Finding good channel codes: if n fixed, can try all poss codes: there are $\sim|\mathcal{X}|^{mn}$ inj code & for rate close to C need $m \sim |\mathcal{X}|^{nC}$. randomly generate them – lots of codewords, 2^{nR+1}, so exponential storage & decoding time.

non-IID input: disc. stochastic proc: seq of RVs, stationary if $\mathbb{P}(X_1=x_1, \dots X_n=x_n) = \mathbb{P}(X_{1+j}=x_1 \dots X_{n+j}=x_n)$ $\forall n, j$

entropy rate: $\mathcal{H}(X) = \lim_n \frac{1}{n} H(X_1, \dots, X_n)$ $[= H(X_1)$ if iid$]$

for a stationary process, $\mathcal{H}(X)$ exists, $= \lim_{n\to\infty} H(X_n | X_{n-1}, \dots, X_1)$. { $n \mapsto H(X_n|X_{n-1} \dots X_1)$ is (weakly) decr & limit exists.

entropy rate of Markov chain: $\mathcal{H}(X) = \lim H(X_n|X_{n-1} \dots X_1) = \dots = H(X_2|X_1) = \sum_{ij} \mu_i \pi_{ij} \log \pi_{ij}$ Cesaro mean: $a_n \to a \Rightarrow \frac{1}{n}\sum^n a_i \to a$

X stationary markov chain, $Y := f(X)$ ← function of M. chain. note Y not Markov, but stationary

$H(Y_n|Y_{n-1}, \dots Y_1, X_1) \leq \mathcal{H}(Y) \leq H(Y_n|Y_{n-1}, \dots Y_1)$ & limits all the same. ← use $I(X_1; Y_n | Y_{n-1} \dots Y_1) \leq H(X_1)$ & $= \sum^{n \to \infty} I(X_1, Y_i | Y_{i-1} \dots Y_1)$

extend backwards to X_{-k}, Y_{-k} stationarity ↑ is stationary

Proofs

- **Jensen:** f convex $\Rightarrow \forall x_0$ $f(x) \geq f(x_0) + b(x - x_0)$ for some $b>0$ so $f(X) \geq f(\mathbb{E}(X)) + b(X - \mathbb{E}[X])$, take $\mathbb{E}$ of both sides
- **Gibbs** combine to $\sum p \log \frac{p}{q} = \mathbb{E}[-\log \frac{q}{p}] \geq -\log \mathbb{E}\, q/p] = -\log \sum p = 0$. Jensen w/ $-\log$ convex. **log-sum** Jensen [discrete]
- **Divergence:** 2) chain rule: split probs in logs, take out term, 3) from 2, 4)–5) log-sum over x w/ $a_1 = \lambda p_1, a_2 = (1-\lambda)p_2, b_1 = \lambda q_1, b_2 = (1-\lambda)q_2$

mutual info: 1)2) obv 3) using equiv for H 4) $I(Y,Z;X)$: chain rule (3) on Y & on Z sep. 5) use 4 w/ $Z = f(Y)$, else $(X \perp Z) | Y$ as Z known.

entropy 1) ≥ 0: $1/p(x) \geq 0$ so $\log \geq 0$, $\leq$ w/ Gibbs $q = 1/|\mathcal{X}|$ 2) using I 4) $H(X, f(X))$ use twice 5) Jensen w/ 2. 6) $-x\log x$ is concave

Fano $Z = 1_{X \neq Y}$, so $H(Z|X,Y) = 0$. $H(X|Y) = H(X|Y) + H(Z|X,Y) = H(X,Z|Y) = H(Z|Y) + H(X|Y,Z)$ ← split by poss vals of Y, Z: $Z=0$ term disappears, $Z=1 \Rightarrow \leq |\mathcal{X}|-1$ poss X vals

Weak AEP: conv: indep + WLLN $\mathbb{P}(X_1 \dots \in T_n^\epsilon)$: by conv $p_{X_1 \dots} \in [\dots]$ by def size: $1 = \sum_{x^n} p(x_1, \dots x_n) \geq \sum_T p(x_1, \dots x_n) \geq \sum_T 2^{-n(H(X)+\epsilon)}$, lower bd: $1-\epsilon \leq \mathbb{P}(\in T) \leq \sum_T 2^{\dots}$

Shannon 1: let ϵ_0: $c \in 1_n$ get code len $\lceil nH(X) + \epsilon_0 \rceil$ $x \notin T_n^{\epsilon_0}$ gets len $\lceil n \log|\mathcal{X}| \rceil$. $\mathbb{E}[l_n] \leq \dots n(H(X) + \epsilon_0) + 2 + \epsilon_0 n \log|\mathcal{X}|$) choose ϵ_0 st $\epsilon_0(1+\log|\mathcal{X}|) < \epsilon/2$

uniq dec $\Rightarrow \sum d^{-l_x}$. $a(k) :=$ # of words len k which are concats of codewords. $(\sum d^{-l(x)})^n \leq \sum_0^{n l_{max}} a(k) d^{-k}$. $a(k) \leq d^{k}$ as uniq dec, so $\sum \dots \leq (n l_{max})^{1/n} \to 1$. n st $2/n \leq \epsilon/2$

$l_x \in$ given l_k's: $r_m := \sum_1^{m-1} d^{-l_k} \leq 1$, $c(m) =$ first l_m digits in d-ary exp of r_m ← is a pref code as ※ $\Rightarrow r_n - r_m < d^{-l_m}$ but ※ as $r_n - r_m \geq d^{-l_m}$.

$\mathbb{E}[|c(x)|] \geq H_d(X)$: use pmf $q(x) = d^{-l_x}/\sum d^{-l_x}$, sub into $\mathbb{E}[|c(X)|]$, move q across to H term to make $D(p\|q)$, get $\overset{\leq 1}{\sum} - \log \sum d^{-l_x} + D(p\|q) \geq 0$.

opt code exists: $l_x := \lceil -\log_d p(x) \rceil$ – by K-M pref code exists, $\mathbb{E}[l_n]$ right. countably many pref codes with $\mathbb{E}[l_n] <$ fin. num, so sort and pick min.

Shannon code bound: $\lceil \epsilon \rceil \leq \epsilon + 1$. | Huffman code: canonical code exists: 1) c opt, $c' := c$ w/ 2 cwds x_j, x_k swapped, $j<k \Leftrightarrow p_k < p_j \Rightarrow |c(x_k)| - |c(x_j)| \geq 0$ so $0 \leq \mathbb{E}[|c'|] - \mathbb{E}[|c|]$ 2) assume contrary, fix, shorter, still pref code 3) pref. code ⟺ rooted tree, by 2) max-len has sibling, now swap codeword til ✓ for 3)

Shannon 2

take pmf p, $m, n \in \mathbb{N}$, $J_\epsilon^{(n)}$ jointly typical set of $P_{XY} = P_{Y|X} P_X$ ← just chosen ↑ the channel

1) generate m random codewords from $\mathcal{X}^n$ by randomly sampling p_X n times
2) randomly assign each message a codeword
3) the decoder: a typical-set decoder: on $\bar{y}$, if $\exists!\, \bar{x}$ st $(\bar{x}, \bar{y}) \in J_{\epsilon/6}^{(n)}$ return $\bar{x}$'s message from 2). else return m.

↑ call this (m–n)-code (C, D). ← it is RANDOM!!

1) sample from (C, D)
2) sample a message $W \sim \mathrm{Unif}\{1 \dots m\}$
3) send $\bar{X} = C(W)$ on the channel
4) decode to $\tilde{W}$ w/ D.

$\forall \epsilon, R$ $m := 2^{n(R - 2\epsilon/3)+1}$: (C, D) has rate $R - \frac{2}{3}\epsilon + 1/n$

choose n st $\mathbb{P}(W \neq \tilde{W}) < \epsilon/2$, by lemma: $W, \tilde{W}$ as def, $\mathbb{P}(W \neq \tilde{W}) \to 0$ as $n \to \infty$: no proof

$\mathbb{P}(W \neq \tilde{W}) = \sum_{c,d} \mathbb{P}(W \neq \tilde{W} | C, D = c, d)\, \mathbb{P}(C, D = c, d) \Rightarrow \exists c^\circ, d^\circ$ st $\mathbb{P}(W \neq \tilde{W} | C, D = c^\circ, d^\circ) < \epsilon/2$

$\dots = \bar{\epsilon}$ so have code w/ rate $R + 1/n - \frac{2}{3}\epsilon$, $\bar{\epsilon} < \epsilon/2$.

$\bar{\epsilon} = \frac{1}{m} \sum^m \epsilon_i < \epsilon/2$. order ϵ_i's, lower half has $\epsilon_i < \epsilon$ [else ※] so throw away rest: now a $(\frac{m}{2}, n)$ code w/ $R - \frac{2}{3} > R - \epsilon$, $\epsilon_{max} < \epsilon$.

can [only] achieve $R \leq C$:

take $\epsilon > 0$, by ↗ $\exists n\ \exists (m,n)$ code st rate $= \log m / n > R - \epsilon$, $\epsilon_{max} < \epsilon$.

$\log(m) = H(W)$ as $W \sim \mathrm{Unif}$

data proc $= H(W|\tilde{W}) + I(W; \tilde{W})$

+split $X \to X_i$'s, $Y \to Y_i$'s $\leq H(W|\tilde{W}) + \sum^n I(X_i; Y_i)$

Fano + cap def $\leq$ " $+ nC \leq 1 + \bar{\epsilon} \log m + nC$

$\Rightarrow \frac{\log m}{n} < \frac{C + 1/n}{1 - \epsilon}$ so $R <$ ← so as $n \to \infty$ $R \leq C$.

entropy rate exists

$n \mapsto H(X_n|X_{n-1} \dots X_1)$ non-incr: $H(X_{n+1}|X_n \dots) \leq H(X_{n+1}|X_n \dots X_2) \leq H(X_n|X_{n-1} \dots)$ by stationarity

↑ limit exists as $n \to \infty$ as bdd below by ≥ 0.

Cesaro mean $|\frac{1}{n}\sum^n a_i - a| \leq \frac{1}{n}\sum^n |a_i - a| \leq \frac{1}{n}\sum^{n_0} |a_i - a| + \overbrace{\frac{n - n_0}{n}}^{\leq 1} \epsilon \leq \epsilon$ as $n \to \infty$ ↑ n_0 st $|a_i - a| < \epsilon$.

$\frac{H(X_1 \dots X_n)}{n} = \frac{1}{n} \sum H(X_i | X_{i-1} \dots X_1)$

cond exps converge, + Cesaro mean so conv to lim of cond exps

Distributions

- Uniform$\{1 \dots n\}$: $f(x) = 1/n$, $\mu = \frac{n+1}{2}$, $\sigma^2 = \frac{n^2-1}{12}$
- Ber(p): $\mathbb{P}(X=1) = p = 1 - \mathbb{P}(X=0)$, $\mu = p$, $\sigma^2 = p(1-p)$
- Bin(n,p): $\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $\mu = np$, $\sigma^2 = npq$, $= \sum_1^n \mathrm{Ber}(p)$ indep
- Geometric(p): $\mathbb{P}(X=k) = (1-p)^{k/k-1} p$, $\mu = \frac{1-p}{p}$ or $1/p$, $\sigma^2 = \frac{1-p}{p^2}$, same for both
- Normal(μ, σ^2): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)$
- Poisson(λ): $f(k) = \lambda^k e^{-\lambda}/k!$, $\mu = \lambda$, $\sigma^2 = \lambda$. discrete
- Exp(λ) $f(x) = \lambda e^{-\lambda x}$, $\mathbb{P}(X > x) = e^{-\lambda x}$. $\mu = 1/\lambda$, $\sigma^2 = 1/\lambda^2$
- Gamma(r, λ). $f(x) = 1/\Gamma(r)\, \lambda^r x^{r-1} e^{-\lambda x}$. $[\sum^n X_i \sim \mathrm{Exp}(\lambda)$ indep$] \sim \mathrm{Gamma}(n, \lambda)$
- Gamma func: in general complicated, $\Gamma(n) = (n-1)!$, $\Gamma(1) = 1$

maximising entropy: if have target dist p^*, use it: $X \sim p \in$ class, $Z \sim p^*$. nb $D(p^*\|p) \geq 0$

$H(Y) = -\sum p \log p = -D(p^*\|p) - \sum p \log p^*$

$\leq -\sum p \log p^* = \dots$ use def of p^*,

class bound: e.g. $\mathbb{E}[Y^2] \leq \alpha =: \mathbb{E}[Z^2]$

Taylor series:

$1/(1-x) = \sum_0^\infty x^n$

$e^x = \sum_0^\infty x^n/n!$

$\ln(1+x) = \sum_1^\infty (-1)^{n-1} x^n/n$

$1/(1-x)^2 = \sum_0^\infty (n+1) x^n$

Lagrangian: $g_i \geq a_i$ $h_i = b_i$

$\Lambda(x, \lambda_i, \mu_i) = f(x) + \lambda_i(g_i - a_i) + \mu_i(h_i - b_i)$

$\frac{d\Lambda}{dx, \lambda_i, \mu_i}$ set to 0. simplify consts: $e^c \to K \dots$

Simple note

$\log_2 2 = 1$, $\log 1 = 0$

$H(1/2) = -\frac{1}{2}\log \frac{1}{2} - \frac{1}{2}\log \frac{1}{2}$

$= \log 2 = 1$. from matrix ↴

BSC capacity: [can quote] $= 1 - H(Y|X)$

| code length proofs: try simple ideas, like STOP symbol, | otherwise extra 0 bit to make even...

$\boxed{\sum_0^{k-1} 2^i = 2^k - 1}$ $c(X)$ is an RV, $H(c(X)) = H(X)$ as inj [say on Im], consider RV $c(X)$, $l(c(X))$.