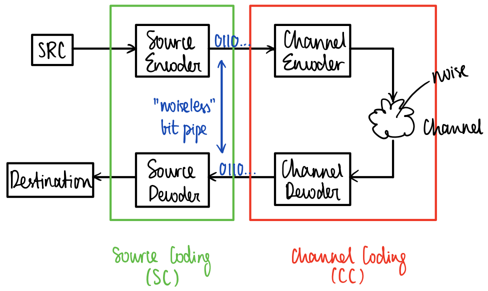
# Information Theory

## Big Picture




Separation Theorem: SC & CC can be done separately w/o loss of optimality from end-to-end.

SC: Impossible to compress a source $X^N$ below its entropy $H(X^N) = NH(X)$ bits.
Possible to compress $X^N$ to $N(H(X)+\epsilon)$ bits $\forall\, \epsilon > 0$ as $N \to \infty$.

CC: Impossible to transmit reliably at rate $R > C$, where $C$ is channel capacity.
Any rate $R < C$ is achievable w/ high reliability (i.e. $P_{error}(N) \to 0$ as $N \to \infty$).

$X \sim B(p)$

If $p = \frac{1}{2}$, how much info does a single toss provide? 1 bit

What if $p = 0.11$? ½ bit

Suppose you have a seq of indep coin tosses by Alice & Bob using fewest # of bits.
How do you do it?

## Entropy

$X \sim$ discrete RV, $x \in X$

$$\boxed{H(X) = \mathbb{E}[-\log p(X)] = \mathbb{E}\left[\log \frac{1}{p(X)}\right] = \sum_{x \in X} p(x) \log \frac{1}{p(x)}} \quad (\log \text{ base } 2)$$

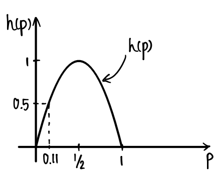
Remark: $H(X)$ measures the avg uncertainty / surprise level / information content.

Intuition: The higher $p(x)$, the lower the surprise / information you gain when you see it.

Ex: $X \sim B(p)$

$$H(X) = \sum_{x=0}^{1} p(x) \log \frac{1}{p(x)} = \underbrace{p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}}_{h(p)}$$

$h(p)$ ↳ Binary entropy function




Some properties of entropy:

If $X \in \{1, 2, \dots, D\}$, $H(X)$ is max when it is uniformly distributed (i.e. $p(1) = p(2) = \dots = p(D)$)

$\Rightarrow H_{max}(X) = \log_2 D$ bits $\Rightarrow \boxed{H(X) \le \log_2 D \text{ bits}}$

## Joint Entropy

$$\boxed{H(X,Y) = \sum_x \sum_y p_{XY}(x,y) \log_2 \frac{1}{p_{XY}(x,y)}}$$

$H(X,Y) = \mathbb{E}\left[\log_2 \frac{1}{p_{XY}(x,y)}\right]$

Note: If $X, Y$ indep, $H(X,Y) = H(X) + H(Y)$.

Proof: $H(X,Y) = -\mathbb{E}[\log(p(x,y))] = -\mathbb{E}[\log(p(x), p(y))] = -\mathbb{E}[\log p(x)] - \mathbb{E}[\log p(y)]$
$= H(X) + H(Y)$

If you observe 2 indep RVs, the info content should be additive in the info-content of each RV.

## Conditional Entropy

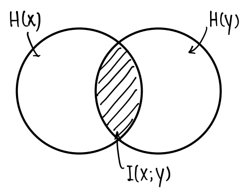
$H(Y|X) = \mathbb{E}\left[\log \frac{1}{p(y|x)}\right]$

can show: $\underbrace{H(Y|X)}_{\text{uncertainty in } Y \text{ after observing } X} = H(X,Y) - H(X)$

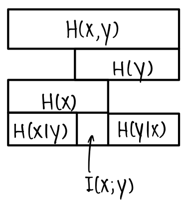
## Mutual Information (MI)

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

↳ intuition: $X, Y$ are coupled, filter out some info after observing the output




Note: If $X, Y$ are indep, $I(X;Y) = 0$




# Asymptotic Equipartition Property (AEP)

↳ Info theory version of WLLN

$$\boxed{\text{Theorem: If } X_1, \dots, X_n \text{ are iid} \sim p(X), \text{ then } \frac{-\log p(X_1, \dots, X_n)}{n} \xrightarrow{P} H(X)}$$

Proof: If $X, Y$ are indep RVs, so are $f(X)$ & $g(Y)$ for any $f(\cdot), g(\cdot)$

$\Rightarrow$ If $X_1, \dots, X_n$ are indep, so are $\log p(X_1), \log p(X_2), \dots, \log p(X_n)$

$\Rightarrow$ WLLN: $-\frac{1}{n}\log p(X_1, \dots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) \xrightarrow{P} \underbrace{-\mathbb{E}[\log p(X)]}_{H(X)}$ in prob

## Intuition of AEP

Ex: You flip a coin w/ bias $p$, $n$ times indep

What's the prob of seeing a typical sequence?

Q/ What's the typical seq?

↳ One having $np$ heads & $n(1-p)$ tails

Q/ What's the prob of seeing a particular typical seq?

$P(\text{typical seq}) = p^{np}(1-p)^{n(1-p)} = 2^{np\log_2 p}\, 2^{nq\log_2 q}$ where $q = 1-p$

$= 2^{n\underbrace{(p\log p + q\log q)}_{-h(p)}}$

$$\boxed{P(\text{typical seq}) = 2^{-nh(p)}}$$

Q/ How many typical seq are there?

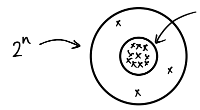
$\binom{n}{np} \approx 2^{nh(p)}$

Use Stirling's approx: $n! \approx \left(\frac{n}{e}\right)^n$

$$\frac{n!}{(np)!\,(nq)!} = \frac{(n/e)^n}{(np/e)^{np}(nq/e)^{nq}} = \frac{(\frac{n}{e})^n}{(\frac{n}{e})^{np+nq}\,p^{np}q^{nq}} = 2^{nh(p)}$$

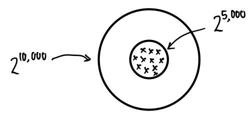
There are about $2^{nh(p)}$ typical seq, each having an equal prob of $2^{-nh(p)}$

$\approx 2^{nh(p)}$ w.p. $2^{-nh(p)}$ for each seq




Ex: $p = 0.11 \Rightarrow p(\text{typ seq}) = 2^{-n/2}$

$n = 10{,}000$




## Typical Set

$A_\epsilon^{(n)}$ w.r.t. $p(x)$ is the set of sequences $(x_1, \dots, x_n) \in X^n$ s.t.

$$\boxed{\Pr\left(2^{-n[H(x)+\epsilon]} \le p(x_1, \dots, x_n) \le 2^{-n[H(x)-\epsilon]}\right) \xrightarrow{n\to\infty} 1 \quad \forall \epsilon > 0}$$

Fact: $\forall \epsilon > 0$, more than $(1-\epsilon)$ of the pmf of $(x_1, \dots, x_n)$ lies in a typical set $A_\epsilon^{(n)}$ having at most $2^{n[H(x)+\epsilon]}$ elems.

Ex: $\epsilon = 0.001$, $n = 10{,}000$, $X \sim B(p = 0.11)$

More than 99.9% of the prob mass lies in a set having no more than $2^{5010}$ elements.

## Huffman Code

Prefix free code: No codeword is prefix of another.

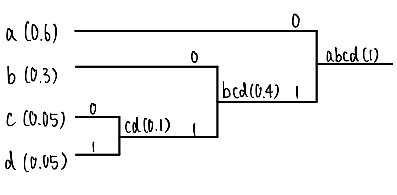
↳ prevent ambiguities from ocurring during decoding

↳ aka "instantaneous codes" since we can decode while streaming

↳ "self punctuating"

$X = \{a, b, c, d\}$, $P_X = \{p_a, p_b, p_c, p_d\}$

Ex: $p_a = 0.6, 0.3, 0.05, 0.05$
(a) (b) (c) (d)




| Sym | Code |
|---|---|
| a | 0 |
| b | 10 |
| c | 110 |
| D | 111 |

$\mathbb{E}[L_{HC}] = \sum p_x l_x = 1.5$

$H(x) = \sum p_x \log \frac{1}{p_x} = 1.395$

Fact: $\boxed{H(x) \le \mathbb{E}[L_{HC}] < H(x) + 1}$

$H(X^n) \le \mathbb{E}[L_{HC}(X^n)] < H(X^n) + 1$

$nH(X) \le \bar{L}_{HC}(X^n) < nH(X) + 1$

$H(X) \le \frac{\bar{L}_{HC}(X^n)}{n} < H(X) + \frac{1}{n}$

↳ Problem: Complexity