

INFORMATION THEORY

References:

- Elements of Information theory, Cover & Thomas
- Pseudorandomness by Salil Vadhan

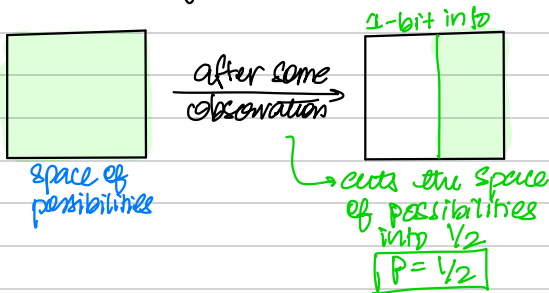
Interesting video (to understand entropy):

Solving wordle using Information theory, [BlueBrown]

ENTROPY

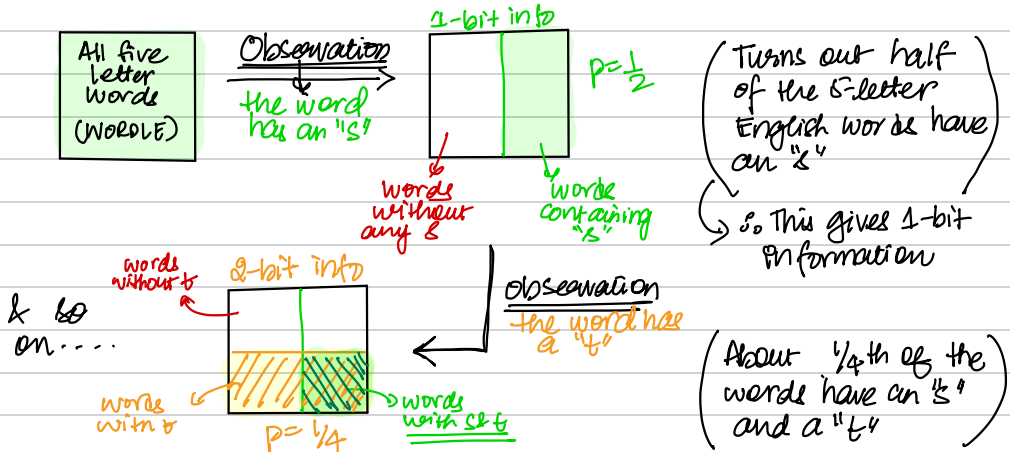
Some informal discussion (Running example: WORDLE)

- How to quantify information bits?



Then, we say that this observation has 1-bit of information.

Example



What is the formula for information in terms of the probability of occurrence?

By above logic, $\left(\frac{1}{2}\right)^I = p \Rightarrow 2^I = \frac{1}{p}$

$$\Rightarrow I = \log_2\left(\frac{1}{p}\right)$$

Shannon entropy is the expected value of this information, [Shannon '48]. Let X be a r.v. with prob mass function $p(x)$

Defⁿ:

$$H(X) = \mathbb{E}[I(X)] = \sum_x p(x) \cdot I(x), \text{ where } I(x) = \log_2\left(\frac{1}{p(x)}\right)$$

In other words, it measures the # bits needed on average to describe the r.v. X

EXAMPLE OF WORDLE

W	E	A	R	Y

(First guess)

If you get this in response from the wordle
 $p(\text{■■■■Y}) = \frac{\text{\# five letter words ending with Y}}{\text{total \# five letter words}}$

$$= \frac{429}{12,972} = 0.0377 \quad (\text{3Blue1Brown})$$

$$I(\text{■■■■Y}) = \log_2\left(\frac{1}{p}\right) = 4.73 \quad (\text{3Blue1Brown})$$

For "WEARY" as 1st guess: $\mathbb{E}[I] = \sum_x p(x) \cdot I(x) = 4.90 \quad (\text{3Blue1Brown})$

A high probability event (e.g. ■■■■Y) gives less information (indicated by the example, as well as by formula for I)

Watch 3Blue1Brown for full fledged analysis to find the best first guess for winning Wordle $\rightarrow$ using just entropy!

Another way to view entropy: $H(X)$ denotes the min # bits needed to store a string drawn from X , on average, using a "compression" scheme.

Example

1		
2	3	3

$X = \begin{cases} \text{red} & \text{w.p. } 1/2 \\ \text{green} & \text{w.p. } 1/4 \\ \text{blue} & \text{w.p. } 1/8 \\ \text{yellow} & \text{w.p. } 1/8 \end{cases}$

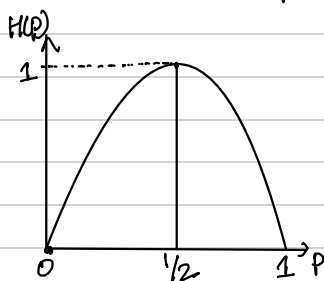
Encoding
 $\rightarrow$ red = TH (1-bit)
 green = TH (2-bits)
 blue = TTH (3-bits)
 yellow = TTT (3-bits)

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{14}{8} = 1.75$$

Example

Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

$$H(X) = p \cdot \log_2\left(\frac{1}{p}\right) + (1-p) \cdot \log_2\left(\frac{1}{1-p}\right) \stackrel{\text{def}}{=} H(p) \quad (\text{Binary entropy function})$$



This graph depicts a few properties of entropy function:

- It's a concave function of distribution.
- $H(p) = 0$ when $p = 0$ or 1 (no uncertainty)
- max uncertainty when $p = 1/2$ ($H(1/2) = 1$)

Properties of $H(X)$

- $H(X) \geq 0$ (easy to prove)
- If $|\text{range}(X)| \leq N$, then $H(X) \leq \log_2 N$. Moreover
 - $H(X) = 0$ iff X is constant (no uncertainty)
 - $H(X) = \log_2 N$ iff X is uniform r.v. (max uncertainty)
- For independent r.v.s X & Y , $H(XY) = H(X) + H(Y)$

Intuitively, take wordle example if a first word guess gave 1-bit information ("s" contained in your word) & a second independent guess gave you another bit of info then combined the two observations gave 2-bits of information!

Proof:

$$H(X,Y) = \sum_{x,y} P_{XY}(x,y) \cdot \log_2 \left(\frac{1}{P_{XY}(x,y)} \right)$$

(using independence)

$$= \sum_x \sum_y P_X(x) \cdot P_Y(y) \cdot \left(\log_2 \left(\frac{1}{P_X(x)} \right) + \log_2 \left(\frac{1}{P_Y(y)} \right) \right)$$

$$= H(X) \cdot \sum_y P_Y(y) + H(Y) \cdot \sum_x P_X(x)$$

$$= H(X) + H(Y)$$

MUTUAL INFORMATION

If X & Y are not independent then # bits (on avg) needed to describe a string from the joint variable is $H(X,Y)$ & this will be less than $H(X) + H(Y)$

Intuitively, again take wordle. If your second guess was not independent & you used "s" (the letter you got from the first guess) then you will get less information.

Defn: Mutual Information

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$= \sum_{x,y} P_{XY}(x,y) \left(\log_2 \frac{1}{P_X(x)} + \log_2 \frac{1}{P_Y(y)} - \log_2 \frac{1}{P_{XY}(x,y)} \right)$$

$$= \sum_{x,y} P_{XY}(x,y) \log_2 \left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \right)$$

Exercise: Prove that $I(X;Y) \geq 0$

EXAMPLE: Suppose (X, Y) have following joint distribution

$Y \backslash X$	♥	♦	♠
♥	$\frac{1}{2}$	0	0
1	0	$\frac{1}{16}$	$\frac{1}{16}$
2	0	$\frac{1}{16}$	$\frac{1}{16}$
3	0	$\frac{1}{16}$	$\frac{1}{16}$
4	0	$\frac{1}{16}$	$\frac{1}{16}$

marginal distributions:

• of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$

• of Y is $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 1.5$$

$$H(Y) = \frac{1}{2} \cdot 1 + 4 \cdot \frac{1}{8} \cdot 3 = 2$$

$$H(X, Y) = \frac{1}{2} \cdot 1 + 8 \cdot \frac{1}{16} \cdot 4 = 2.5$$

$$\therefore I(X; Y) = H(X) + H(Y) - H(X, Y) = 1$$

Properties of Mutual Information:

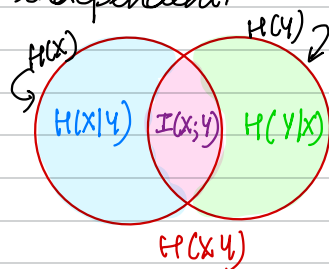
- ① $I(X; Y) = 0$ if and only if X & Y are independent.
- ② Conditional Entropy:

• conditional entropy of Y given X

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

• conditional entropy of X given Y

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



Intuitively, what the conditional entropy captures is: to communicate (X, Y) , think of first communicating shared info ($I(X; Y)$), & then the deficit from $H(X)$ to make X and the deficit from $H(Y)$ to make Y .

E.g. X = random n -bit string $Y = (\bigoplus_{i=1}^n x_i, \text{one random bit})$
 then $I(X; Y) = 1$, $H(X|Y) = n-1$, $H(Y|X) = 1$

$$\textcircled{3} \quad I(X; X) = H(X) - H(X|X) = H(X)$$

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

$$I(X; Y) = I(Y; X)$$

(reduction in uncertainty of X , knowing Y)

(reduction in uncertainty of Y , knowing X)

Symmetric (X says as much about Y as Y says about X)

$$\textcircled{4} \quad \text{CHAIN RULE: } H(X, Y) = H(X) + H(Y|X)$$

$$\left(\begin{aligned} \text{By def}^n \quad H(Y|X) &= H(Y) - I(X; Y) = H(Y) - (H(X) + H(Y) - H(X, Y)) \\ &= H(X, Y) - H(X) \end{aligned} \right)$$

(Exercise) Prove generalized chain rule: $X_1, X_2, \dots, X_n$ be n r.v.s
 $H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \dots + H(X_n|X_1, \dots, X_{n-1})$

$$\textcircled{5} \quad H(Y|X) = \mathbb{E}_{x \sim X} [H(Y|X=x)]$$

(Exercise: try to prove this using chain rule.)

$\textcircled{6}$ Conditional Mutual Information of X, Y , given Z

$$I(X; Y|Z) \stackrel{\text{def}}{=} H(X|Z) + H(Y|Z) - H(X, Y|Z) \\ \text{prove this} \rightarrow = \mathbb{E}_{z \sim Z} [I(X|Z=z; Y|Z=z)]$$

Example

X, Y be independent random bits

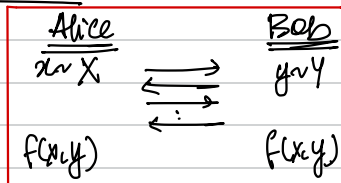
$$Z = X \oplus Y$$

$$I(X; Y) = 0$$

$$I(X; Y|Z) = 1$$

INFORMATION COMPLEXITY

Recall from last lecture: Distributional communication complexity

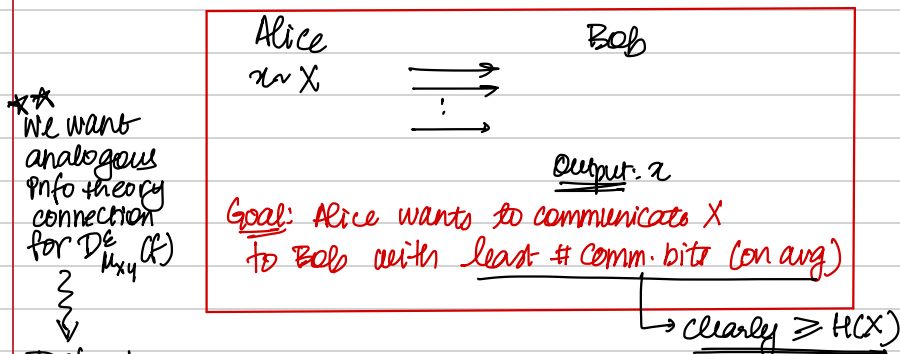


X & Y are r.v.s

$D_{\mu_{xy}}^{\epsilon}(f)$: minimum communication cost over all deterministic protocols π s.t.

$$\text{err}_{\mu}(\pi) = \Pr_{(x,y) \sim \mu_{xy}} [\pi(x,y) \neq f(x,y)] \leq \epsilon$$

You can view entropy in a different way: let X be a r.v. and consider the one-way communication:



Definition:

Information complexity of π under μ_{xy} is:

$$IC_{\mu}(\pi) = I(T_{\pi}; y | x) + I(T_{\pi}; x | y)$$

where T_{π} is a random variable capturing the transcript of protocol π (all messages communicated) on inputs x, y

the amount of info the conversation teaches Alice about Bob's input

the amount of info the conversation teaches Bob about Alice's input.

$$IC_{\epsilon}^{\mu}(f) = \min_{\pi} \{ IC_{\mu}(\pi) : \text{err}_{\mu}(\pi) \leq \epsilon \}$$

information complexity of f under distribution μ

FACT: $D_{\epsilon}^{\mu}(f) \geq IC_{\epsilon}^{\mu}(f)$

Next class: continue looking at this $\rightarrow$ lower bound $R_{\epsilon, \text{pub}}(DIS)$