Lecture 20 - Information Theory

[[Scribe - Cover-Thomas, Braverman survey...]]

[Shannon '48] [[Invented everything. Applications in coding thy, combinatorics, comm cxty...]]

Let X be a (discrete) random variable. E.g. $X = \begin{cases} \text{red} & \text{w.p. } 1/2 \\ \text{green} & 1/4 \\ \text{blue} & 1/8 \\ \text{yellow} & 1/8. \end{cases}$

Entropy of X, denoted $H(X)$, is - intuitively -

if you never worried about $\lceil\,\rceil$, $\lfloor\,\rfloor$

- "amount of randomness in X" [[in bits]]
- [[min]] # of rand bits needed to generate a draw of X, on average

e.g.:
H → red
T → H → green
T → H → blue
T → yellow

: w.p. $\frac{1}{2}$ used 1
$\frac{1}{4}$ " 2
$\frac{1}{8}$ " 3
$\frac{1}{8}$ " 3
$= \frac{14}{8} = 1.75.$

- # of bits needed to store a draw from X using best compression scheme, on avg.,

red ⟷ H
green ⟷ TH
blue ⟷ TTH
green ⟷ TTT.

Rem: "prefix-free" encoding [[no codeword is prefix of another]].

⇒ can store n draws w/o "punctuation"
↳ THTTHHH = green, blue, red, red

- # of bits needed (on avg.) for Alice to communicate one draw from X to Bob
[[storing on a hard drive is just like transmitting/comm'ing to your future self]]

- # of YES/no q's Bob needs to ask (on avg.) to learn a draw of X.
[[20 questions]]

[["Is it red? Green? Blue?" Kinda triv....]]

All exactly correct if X's probabilities are 2^{-k}, $k \in \mathbb{N}$. def: $\boxed{H(X) = \sum_{x \in range(X)} p_X(x) \cdot \log_2 \frac{1}{p_X(x)}}$

But...
$$X = \begin{cases} a & \text{w.p. } \frac{1}{2} \\ b & 1/3 \\ c & 1/6 \end{cases}$$

Communicating X: w.p. $\frac{1}{2} \to 1$ bit.

w.p. $\frac{1}{3} \to \log_2 3$ bits ?? (not an integer!)

$\frac{1}{6} \to \log_2 6$

Elegant to just ~~define~~ define $H(X)$ thus anyway. [(*) Skip down]

fact: If $|range(X)| \le N$, $0 \le H(X) \le \log_2 N$ eq. iff X is unif. dist. [on N outcomes]

eq. iff X const.

[Takes "$\log_2 N$" bits to gen unif dist on $[N]$]

[pf: exercise using Jensen]

[How can we get around this slightly annoying gap between H(X) and the actual compression/comm./generation cost? Amortization.]

trad. to write as a string: XY [juxtapos.]

Say X, Y are indep. r.v.'s. The pair (X,Y) can be viewed as a single r.v.

fact: $H(\underset{XY}{(X,Y)}) = H(X) + H(Y)$. [To optimally gen. a draw of (X,Y), optimally do it for X, and Y. No savings poss. if indep.]

[one time we'll do a proof; rest of time follow your intuit.]

pf: $H(XY) = \sum_{(x,y) \in range(X,Y)} p_{XY}(x,y) \cdot \log \frac{1}{p_{XY}(x,y)}$

$$= \sum_x \sum_y \underbrace{p_X(x) p_Y(y)}_{\text{indep.}} \left(\log \frac{1}{p_X(x)} + \log \frac{1}{p_Y(y)}\right) = H(X) \cdot \cancel{\sum_y p_Y(y)} + H(Y) \cancel{\sum_x p_X(x)} \quad \square$$

[(*)] Aside: can store/gen. using $\sum_x p(x) \lceil \log \frac{1}{p(x)} \rceil$ bits: $\le \sum p(x) (\log \frac{1}{p(x)} + 1) = H(X) + 1$.

$\log 2$	H	$\to a$
$\lceil \log 3 \rceil$	TH	$\to b$
$\lceil \log 6 \rceil$	TTH	$\to c$

(ex: even prefix-freely)

"Shannon-Fano code". [not optimal - "Huffman code" is]

[Instantaneous conversion of coin flips to draws... TTT is "wasted".]

["Of course" you can't beat H(X), either. Ex via Kraft...]

$\therefore$ if $X_1, \ldots, X_n$ are indep copies of X, $H(X_1 \cdots X_n) = n\,H(X)$

To communicate/generate, could use $(H(X)+1)\cdot n = nH(X) + n$ bits,
but wasteful. E.g.: if $X = \begin{cases} 0 & \text{w.p. } 1/3 \\ 1 & 1/3 \\ 2 & 1/3 \end{cases}$, $H(X) = \underbrace{\log_2 3}_{} \approx 1.6$,
uses $2.6n$ bits

But also, could use $H(X_1 X_2 \cdots X_n) + 1$ bits
$\| \,\underbrace{\text{It's like, interp as \# betw. } 0 \,\&\, 3^n - 1 \text{ send it in binary}}_{\text{Shannon - Fano.}} \,\|$

$$\begin{aligned} &\| \\ &nH(X) + 1 \\ &= H(X) + \tfrac{1}{n} \text{ amortized bits per draw.} \end{aligned}$$
$\frac{1}{n} \to 0$ as $n \to \infty$.

$\therefore$ $H(X) =$ amortized $\|$over many indep. copies$\|$ bits to gen/comm. draw of X.

Mutual Information: Say X, Y are possibly dep. r.v.'s
How many rand bits to generate? ("Ideally"):
$\underline{H(XY)}$. If X, Y indep., it's $H(X)+H(Y)$.
If not, can get savings.

def: $I(X;Y) =$ savings when gen/comm. X, Y jointly
$= H(X) + H(Y) - H(XY)$ $\quad \left(= \sum_{xy} p_{XY}(x,y) \left(\log\frac{1}{p(x)} + \log\frac{1}{p(y)} - \log\frac{1}{p(x,y)} \right) \right)$ $= \log \frac{p_{XY}(x,y)}{p(x)p(y)}$.

eg.:

X	Y	prob
*	*	1/2
0	a	1/16
0	b	1/16
0	c	.
0	d	.
1	a	'
1	b	'
1	c	'
1	d	1/16.

$H(X) = \underbrace{H\begin{pmatrix} * & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/4 \\ 1 & \text{w.p. } 1/4 \end{pmatrix}}_{} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 = 1.5$,

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

[not indep, b/c $X = * \Rightarrow Y = *$] $\quad H(X)+H(Y) = 3.5$

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

$\therefore I(X;Y) = 3.5 - 2.5 = 1$

$\hookrightarrow$ = # shared bits used to gen. them / reduce to indep. generations $\|$again, amortized over many$\|$.

facts: $0 \le I(X;Y) \le H(X)+H(Y)$

eq. if indep.

eq. if ~~both~~ both const, $H=0$!

better $\le \min\{H(X), H(Y)\}$ [most you could save is everything]

Equality $I(X;Y) = H(Y) \iff Y$ is determined by X (a fcn. of)

e.g.: $I(X;X) = H(X)$

OR: Let X = rand n bit string

Y = parity of all X's bits

$H(X) = n \quad H(Y) = 1 \quad I(X;Y) = 1$

[Y det by X. Savings is 1]

~~[Though mutual info. is nice & symmetric in X, Y, sometimes you need to look @ things asymmetrically.]~~

~~[Say you need to generate X, Y together. You could be like, "first I'll gen X; then I'll see how many more bits I need to gen Y"]~~

~~$H(XY) = H(X) + \underbrace{\text{\# bits (on avg) needed to gen } Y \text{ given } X}$~~

~~def: $H(Y|X)$ "condit. entropy of Y given X"~~

~~$= H(XY) - H(X)$~~

~~[please remember, that if starts w/ H, it's an expec.] $= \mathop{E}_{x \sim X}[H((Y|X=x))]$~~

~~fact: $H(Y|X) \ge 0$ w/ eq. iff Y determined by X [You gotta gen X to gen XY. Formally, need Jensen.]~~

formula: $I(X;Y) = H(X) + H(Y) - H(XY) = H(Y) - H(Y|X)$

[[One more interp.]]: $I(X;Y) = H(Y) - H(Y|X) \overset{\text{symm}}{=} \underbrace{H(X) - H(X|Y)}$

how much info you gain about X by knowing Y

A $\xrightarrow{\text{noisy channel}}$ B

$X^{(n)} \xrightarrow{\text{corrupts}} Y^{(n)}$. $I(X;Y)$ is info gained by Bob about X.

"Channel capacity" [[Shannon "operationalized"; showed there's a encoding scheme which lets you "transmit info @ this rate"]]

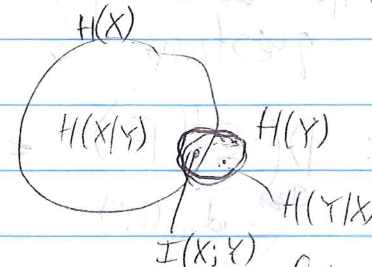
Say X = rand n-bit string, Y = (parity of X's bits, 1 more indep rand bit).

[Say you're jointly generating X, Y. Could first spend the "common part" $I(X;Y)$ bits. Then you still have to do the deficit from $H(X)$ to make X, deficit from Y to make Y.]

[In example, first gen $\oplus X$: 1 bit. Then 1 more to make Y, $n-1$ more for X]

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

<u>Conditional entropy</u> of Y given X. [equiv. is given "mutual info of X,Y"]



[Venn diags misleading if 3+ r.v.s!]

[given I's formula ⇒]

formula: $$H(Y|X) = H(XY) - H(X)$$

⟺ "chainrule": $$H(XY) = H(X) + H(Y|X)$$

[To gen X,Y, first gotta gen X, then gen "Y given X"]

Fact: [Notation is very suggestive/bad. Tho it starts with an H, $H(Y|X)$ really an $E[\cdot]$]

$$H(Y|X) = \mathop{E}_{x\sim X}\left[H((Y|X=x))\right].$$

pf: $$= H(XY) - H(X) = \mathop{E}_{(x,y)\sim XY}\left[\log\frac{1}{p(x,y)}\right] - \mathop{E}_{x\sim X}\left[\log\frac{1}{p(x)}\right]$$

$$= \mathop{E}_{x,y\sim XY}\left[\log\frac{1}{p(x)p(y|x)} - \log\frac{1}{p(x)}\right]$$

$$= \mathop{E}_{x\sim X}\left[\mathop{E}_{y\sim Y|X=x}\left[\log\frac{1}{p(y|x)}\right]\right]. \quad \square$$

def: condit. mut. info: $$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(XY|Z).$$

[One more def, then some apps.] (*) GOTO PAGE 5.5

<u>KL-Divergence</u> [Kullback-Leibler, AKA <u>relative entropy</u>]

[Let] E [denote] rand. letter fr. English [novel]

$H(E)$ = # yes no questions to guess [1st q: is it in {CHUMP GIFTS}?]

≈ 4 [in Latin alpha langs]

[[Pierre comes to America and wants to play
French Scrabble expert, doesn't know English.
Take a slot with his French 20 q's strat.]]

F: rand French letter. ↪ [[1st Q: en "PEU MOL"? "a little bit soft"]]

[[He would pay more than $H(E)$ questions on average. What is the "penalty"?]] [[E.g. long time for Pierre to guess W]]

def: $D(E\|F)$ = avg. # extra bits needed when storing/guessing draw fr. E using scheme for F. [[E.g.: excess per bit if you zip an English novel using encoder designed for French]].

$= d_{KL}(E,F)$

$$D(X\|Y) = \underset{x\sim X}{E}\left[\log\frac{1}{p_Y(x)} - \log\frac{1}{p_X(x)}\right]$$

fact: Let X,Y be joint r.v.'s, let $X\times Y$ denote drawing them indep.

$$D(XY\|X\times Y) \overset{?}{=} I(X;Y)$$

[[penalty for not using savings]]

rem: $D(E\|F)$ like a distance between E,F, but NOT SYMMETRIC! $\neq D(F\|E)$.

E.g.: Hawaiian has 12 letters [[I lie, actually 13, one being ʻ, glottal stop, never mind]]

$D(H\|E) \approx 2$, modest

$D(E\|H) = \infty$.

thm: [[moderately annoying to prove]] [[Gibbs–Su]]

[[Lec. 8]]

$$d_{TV}(X,Y) \lesssim \sqrt{d_{\chi^2}(X,Y)}$$

$$d_{TV}(X,Y) \lesssim \tfrac{1}{\sqrt{2}}\sqrt{d_{KL}(X,Y)} \le \sqrt{d_{\chi^2}(X,Y)}$$

$$d_H(X,Y)^2 \lesssim d_{TV}(X,Y) \le \sqrt{2}\, d_H(X,Y)$$

↑ Hellinger.

Can be sharp: $D\left(\mathrm{Bern}\left(\tfrac{1}{2}+\varepsilon\right) \| \mathrm{Bern}\left(\tfrac{1}{2}\right)\right) = \Theta(\varepsilon^2)$

[[one way communic. w/ priors]]

$A \longrightarrow B$

$X \dashrightarrow$

comm. cost : $H(X)$

[[last time]]

$A \xrightarrow{f} B$

$X \rightleftarrows Y$

[[SAT analogy]]

:: distrib'l comm. cxty :: INFORMATION CXTY [[Exploding developing field]]

$\approx$ amt of "info" A,B need to exchange to compute $f(x,y)$ [[maybe with error]]

def: Let Π be a (det.) comm prot.

$$IC_{\Pi}(\mu_{XY}) := I(\Pi; Y \mid X) + I(\Pi; X \mid Y)$$

amt. of info conversation teaches Alice abt. Bob's input; vice versa.

$$IC^{\varepsilon}_{\mu_{XY}}(f) = \inf \{ IC_{\Pi}(\mu_{XY}) : \Pi \text{ s.t. } \Pr_{X,Y \sim \mu}[\Pi(x,y) \neq f(x,y)] \leq \varepsilon \}$$

fact: $D^{\varepsilon}_{\mu}(f) \geq IC^{\varepsilon}_{\mu}(f)$. [[Converse??]]

thm [BYJKS'04]: $I^{\varepsilon}_{\mu^{\otimes n}_{XY}}(DISJ_n) \geq n \cdot I^{\varepsilon}_{\mu_{XY}}(NAND) \Rightarrow R_{\varepsilon}(DISJ_n) \geq \Omega(n)$.

appropriate; $\bigwedge_{i=1}^{n} NAND(x_i y_i)$; > 0? Yes! $\Omega(\varepsilon^2)$, tho slightly tough.

thm: $IC^{\varepsilon}_{\mu^{\otimes n}}(f^n) = n \cdot IC^{\varepsilon}_{\mu}(f)$ (generalize)

$\Rightarrow$ amortized comm cxty, $\lim_{n \to \infty} \frac{D^{\varepsilon}_{\mu^{\otimes n}}(f^{\otimes n})}{n} \geq IC^{\varepsilon}_{\mu}(f)$. must get err $\leq \varepsilon$ on each coord.

thm [BR'10]: $\leq$.

Proof requires compression of conversation : convo w/ little info compressible to short one.

[[Many open probs still]].