Information

Tuesday, November 3, 2020 5:59 PM

$$I(A) = -\log \mathbb{P}[A]$$

- lower prob. events $\rightarrow$ more info
- if A, B independent, then

$$\begin{aligned} I(A \cap B) &= -\log \mathbb{P}[A \cap B] \\ &= -\log \mathbb{P}[A]\mathbb{P}[B] \\ &= -\log \mathbb{P}[A] - \log \mathbb{P}[B] \\ &= I(A) + I(B) \end{aligned}$$

<u>Info. for a random variable?</u>

X is a discrete r.v. w/ p.m.f. $f(x)$, supp. S

for any $x \in S$, we can write $I(X=x) = -\log f(x)$

<u>Def.</u> The <u>entropy</u> of disc. r.v. X is given by

$$H(X) = \sum_{x \in S} -f(x) \log f(x) = \mathbb{E}[\underbrace{-\log f(x)}_{I(f(x))}]$$

$$\left(= \sum_i -p_i \log p_i \right)$$

the idea: I: event $\to$ real number (information)

H: r.v. $\longrightarrow$ real number (entropy)

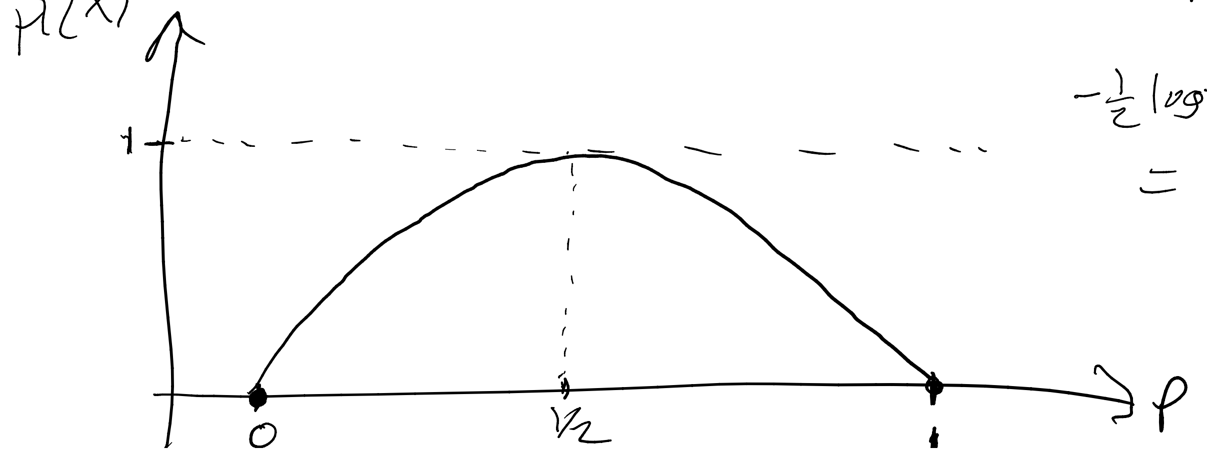
$X \longrightarrow \mathbb{E}[I(f(x))]$

note: If $f(x) = 0$, $0 \log 0 \overset{\text{in this class}}{=} 0$

<u>Ex.</u> What is the <u>entropy</u> of a Bernoulli r.v. $X \sim \text{Bern}(p)$

$$\boxed{H(X) = -p \log p - (1-p) \log(1-p)}$$

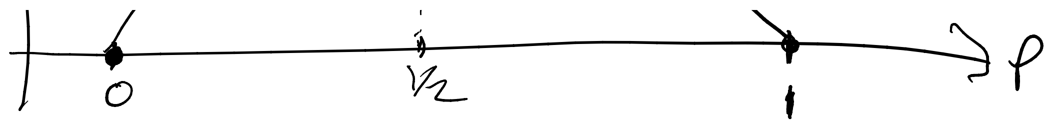
<u>Binary</u> <u>Entropy</u> <u>Function</u> (of p)



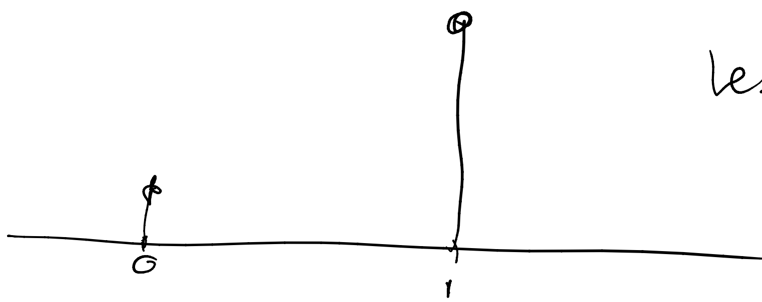
$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$

$= -\log\frac{1}{2}$

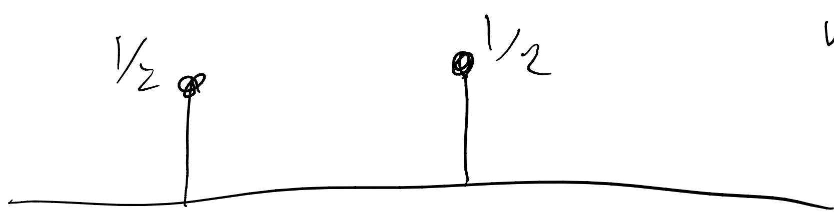
$= 1$



at 0, 1 no disorder (100% chance of same output)



less disorder, because you have a "better guess"



max. disorder

more entropy than Bern(1/2)

↑ each event has higher Info

Ex. Uniform on $1 \le x \le N$ "alphabet of N equiprobable events"

$$H(X) = \sum_{x=1}^{N} -\frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N} = \boxed{\log N}$$

$N \nearrow \Rightarrow H(X) \nearrow$

$\boxed{\begin{array}{l}\underline{\text{THM}} \text{ for any discrete r.v. w/ support containing } N \text{ pts,} \\ \qquad\qquad 0 \le H(X) \le \log N\end{array}}$

Uniform is entropy-maximizing

Def. The joint entropy of X and Y is

$$H(X,Y) = -\sum_y \sum_x f(x,y) \log(f(x,y))$$

Def. The conditional entropy of X given Y

$$H(X|Y) = -\sum_x \sum_y f(x,y) \log(f(x|y))$$

explanation at $Y=y$, $H(X|Y=y) = -\sum_x f(x|y) \log f(x|y)$

entropy of one r.v.

more generally

$$H(X_n \mid X_1, \dots, X_{n-1}) = \sum_{x_1,\dots,x_n} -f(x_1,\dots,x_n) \log f(x_n \mid x_1,\dots,x_{n-1})$$

"TPT for entropy" (directly from tpt)

$$H(X \mid Y) = \sum_i P[Y = y_i] \, H(X \mid Y = y_i)$$

Theorem $H(X,Y) = H(Y) + H(X \mid Y)$

Proof. $$H(X,Y) = -\sum_{x,y} f(x,y) \log f(x,y)$$

$$= -\sum_{x,y} f(x,y) \log \left(f(y) f(x \mid y) \right)$$

$$= -\sum_{x,y} f(x,y) \log f(y) \underbrace{- \sum_{x,y} f(x,y) \log f(x \mid y)}_{H(X \mid Y)}$$

$$= -\sum_y \log f(y) \sum_x f(x,y) + H(X \mid Y)$$

$$= -\sum_y \log f(y) \sum_x f(x,y) + H(X|Y)$$

$$= -\sum_y f(y) \log f(y) + H(X|Y)$$

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

manipulating is nice b/c of logarithms

Generally: If $X_1, \ldots, X_n$ indep

$$H(X_1, \ldots, X_n) = \sum_i H(X_i)$$

Corollary: $H(Y) = H(X,Y) - H(X|Y)$

Def. The mutual information between X and Y is

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

17

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

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

For cts. r.v. X def. differential entropy

as
$$h(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

Ex. Entropy of $U(a,b)$

$$h(X) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a}\, dx$$

$$= \log(b-a)$$ $\leftarrow$ entropy-maximizing on (a,b)

Ex. $X = N(0, \sigma^2)$

$\infty \quad {}^{-x^2/2\sigma^2} \quad {}^{-x^2/2\sigma^2}$

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}\right) dx$$

$$= \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}}_{\text{a pdf}} \underbrace{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)}_{x\text{-indep}} dx + \int_{-\infty}^{\infty} \frac{x^2/2\sigma^2}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \overbrace{\int_{-\infty}^{\infty} (x-0)^2 f(x)\, dx}^{\sigma^2}$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log(e)$$

$$= \boxed{\frac{1}{2}\log(2\pi e\sigma^2)}$$

(logs here are base e)

DMS

Tuesday, November 3, 2020 6:44 PM

Discrete Memoryless Source

memoryless

A transmitter which sends n independent signals from a discrete dictionary of N symbols.

We call the symbols $a_1, \dots, a_N$ and say they have probs. $p_1, \dots, p_N$.

Over n transmissions, if n large

we see symbol a_k approximately np_k times

So for large n, a sequence has Probability

$$P = \prod_{i=1}^{N} p_i^{np_i}$$

all independent, symbol a_i occurs np_i times

$$= \prod_{i=1}^{N} 2^{n p_i \log p_i}$$

$$= 2^{n \sum p_i \log p_i}$$

$$= 2^{-nH(X)}$$

prob. of any sequence of this type is $2^{-nH(X)}$

Model: DMS output is equiprobable sequences w/ np_i instances of a_i each w/ prob. $2^{-nH(X)}$.

$\rightarrow 2^{nH(X)}$ probable sequences (sequences of non-negligible probability)

$\rightarrow N^n$ possible sequences (ex. $\underbrace{a_1, a_1, a_1, \ldots, a_1}_{n \text{ times}}$)

of which only a small subset are
<u>probable</u>

real sequence space: big, not uniform in prob.

approximation: much smaller, uniform in prob.

valid w prob. $1-\varepsilon$
where ε can be made arbitrarily small by increasing n

To represent the output of a DMS transmitting n symbols from dictionary of size N, need to rep $2^{nH(X)}$ seq., need

$$* \quad * \quad * \quad \boxed{nH(X) \text{ bits}} \quad * \quad * \quad *$$

$P_a = 0.75$, $P_b = 0.25$

$n=4$
a a a b
a a b a
a b a a
b a a a

→ 2 bits is ok

$P_a = .5 = P_b$

a a b b
a b a b
b a a b
b a b a
b b a a

$n=4$

→ need 3 bits

more uniform → more distinct probable sequences

→ higher H