

Quantum information theory

Lecture 2

Classical information and Shannon entropy

Fundamental unit of classical information:

Bit 2 alternatives, 0 or 1

0100101110
10 bits

$$\mathcal{N} = \left(\begin{array}{l} \text{\# of sequences} \\ \text{of length } N \end{array} \right) = 2^N$$

$$I = \log \mathcal{N} = N \text{ bits}$$

Information stored
Information we get (missing info)
of binary questions

↑
Why take logarithm? So
information is additive.

Different units: Dit D alternatives

$$\mathcal{N} = D^N = 2^{N \log D}$$

$$I = \begin{cases} \log_D \mathcal{N} = N \text{ dits} \\ \log_2 \mathcal{N} = N \log_2 D \text{ bits} \end{cases}$$

Convention: $\log = \log_2$

$\ln = \log_e = \left(\begin{array}{l} \text{natural} \\ \text{logarithm} \end{array} \right)$ "nats"

$$\log x = \frac{\ln x}{\ln 2} = \log_e \ln x$$

Different probabilities:

probability for x

X
↑
random variable
alphabet

X
↑
value of random variable
letter of alphabet

↓
p(x)

"Information content" varies from letter to letter. Why?

Consider information I_N conveyed by a long sequence of N letters drawn from the distribution p(x). Extract

information per letter, I_N/N , as $N \rightarrow \infty$.

Heuristics: As N gets large, probability is concentrated on "typical sequences" in which x_j occurs $n_j = N p(x_j) = N p_j$ times, $j=1, \dots, D$

(probability of any typical sequence) = $p_1^{n_1} \dots p_D^{n_D} = p_1^{N p_1} \dots p_D^{N p_D} = 2^{-N H(\vec{p})}$

$-\log(\dots) = N \left(- \sum_j p_j \log p_j \right) = N H(\vec{p})$

$\equiv H(\vec{p}) = \left(\begin{array}{l} \text{Shannon entropy} \\ \text{of probability} \\ \text{distribution} \\ p_1, \dots, p_D \equiv \vec{p} \end{array} \right) \equiv H(X)$

$\mathcal{N} = \left(\begin{array}{l} \# \text{ of typical} \\ \text{sequences} \end{array} \right) = \frac{N!}{n_1! \dots n_D!} = 2^{N H(\vec{p})}$

Use Stirling approximation: $\ln N! \sim N \ln N - N$

$\ln \mathcal{N} = \ln N! - \sum_j \ln n_j!$
 $\sim N \ln N - N - \sum_j (n_j \ln n_j - n_j)$
 $= N \ln N - \sum_j N p_j \ln N p_j$
 $= N \left(- \sum_j p_j \ln p_j \right)$

$\log \mathcal{N} = N \left(- \sum_j p_j \log p_j \right) = N H(\vec{p})$

The typical sequences hog all the probability, and each has the same probability.

$$I_N = \log \mathcal{N} = N H(\vec{p}) \Rightarrow$$

$$\frac{H}{N} = H(\vec{p}) = H(X)$$

③

Rigor:

subscripts here specify which trial, not which letter

Sequence x_1, \dots, x_N in N independent, identically distributed trials with probability

$$p(x_1, \dots, x_N) = p(x_1) \dots p(x_N)$$

ϵ -typical sequences: A sequence is ϵ -typical if

$$\left| -\frac{1}{N} \log p(x_1, \dots, x_N) - H(\vec{p}) \right| \leq \epsilon$$

$T(N, \epsilon)$ is the set of ϵ -typical sequences

$$\Leftrightarrow 2^{-N(H(\vec{p}) + \epsilon)} \leq p(x_1, \dots, x_N) \leq 2^{-N(H(\vec{p}) - \epsilon)}$$

Preliminaries:

$$S = -\frac{1}{N} \log p(x_1, \dots, x_N) = \frac{1}{N} \sum_{x=1}^N -\log p(x_x) \leftarrow \begin{array}{l} \text{Sample mean} \\ \text{of } -\log p(x) \end{array}$$

$$\langle S \rangle = H(\vec{p})$$

$$\langle (DS)^2 \rangle = \frac{1}{N} \langle (\Delta(-\log p(x)))^2 \rangle = \frac{1}{N} \sum_x p(x) (-\log p(x) - H(\vec{p}))^2$$

Asymptotic equipartition (AEP)
 or Typical sequences } Theorem.

(i) For any $\epsilon, \delta > 0$, there exists N_0 , such that for all $N \geq N_0$, the probability that a sequence is ϵ -typical is $\geq 1 - \delta$.

Proof:

$$P\left(\left| -\frac{1}{N} \log p(x_1, \dots, x_N) - H(\vec{p}) \right| \leq \epsilon\right)$$

$$= 1 - P\left(\left| -\frac{1}{N} \log p(x_1, \dots, x_N) - H(\vec{p}) \right| > \epsilon\right)$$

$$\leq \frac{\langle (\Delta(-\log p(x)))^2 \rangle}{N\epsilon^2}$$

↑
 Weak law of large numbers

$$\geq 1 - \frac{\langle (\Delta(-\log p(x)))^2 \rangle}{N\epsilon^2}$$

$$\geq 1 - \frac{\langle \quad \rangle}{N_0 \epsilon^2}$$

δ

Choose

$$N_0 = \frac{\langle (\Delta(-\log p(x)))^2 \rangle}{\delta \epsilon^2}$$

$$= 1 - \delta$$

(ii) The number of ϵ -typical sequences, $|T(N, \epsilon)|$, satisfies

$$(1 - \delta) 2^{N(H(\vec{p}) - \epsilon)} \leq |T(N, \epsilon)| \leq 2^{N(H(\vec{p}) + \epsilon)}, \quad N \geq N_0.$$

Proof:

$$1 \geq \sum_{\substack{\epsilon\text{-typical} \\ \text{sequences}}} p(x_1, \dots, x_N) \geq |T(N, \epsilon)| \underbrace{\min p(x_1, \dots, x_N)}_{2^{-N(H(\vec{P}) + \epsilon)}}$$

$$\Rightarrow |T(N, \epsilon)| \leq 2^{N(H(\vec{P}) + \epsilon)}$$

$$1 - \delta \leq \sum_{\substack{\epsilon\text{-typical} \\ \text{sequences}}} p(x_1, \dots, x_N) \leq |T(N, \epsilon)| \underbrace{\max p(x_1, \dots, x_N)}_{2^{-N(H(\vec{P}) - \epsilon)}}$$

$$\Rightarrow |T(N, \epsilon)| \geq (1 - \delta) 2^{N(H(\vec{P}) - \epsilon)}$$

(iii) Let S_N be any set of sequences of length N , containing at most 2^{NR} sequences, where $R < H(\vec{P})$. Given any $\delta > 0$, there exists N_0 such that for all $N \geq N_0$,

$$\sum_{x_1, \dots, x_N \in S_N} p(x_1, \dots, x_N) \leq \delta.$$

Proof: Let $\epsilon < H(\vec{P}) - R$. For part (i), choose $\delta' = \delta/2$, giving rise to N_0' . Now

$$\sum_{x \in S_N} p(x) = \underbrace{\sum_{\substack{\epsilon\text{-typical} \\ x \in S_N}} p(x)} + \underbrace{\sum_{\substack{\epsilon\text{-atypical} \\ x \in S_N}} p(x)}$$

$$N \geq N_0' : \leq 2^{NR} \cdot 2^{-N(H(\vec{P}) - \epsilon)} = 2^{-N(H(\vec{P}) - R - \epsilon)}$$

$$N \geq N'_0: \sum_{\substack{\epsilon\text{-atypical} \\ x}} p(x) = 1 - \sum_{\substack{\epsilon\text{-typical} \\ x}} p(x) \leq \delta' = \delta/2$$

$$\therefore \sum_{x \in S_N} p(x) \leq 2^{-N(H(\vec{p}) - R - \epsilon)} + \delta/2$$

Now choose $N_0 \geq N'_0$ such that $2^{-N_0(H(\vec{p}) - R - \epsilon)} \leq \delta/2$.
 Then for $N \geq N_0$, $\sum_{x \in S_N} p(x) \leq \delta$.

Shannon's noiseless channel coding theorem, or block-coding theorem. Essentially a rephrasing of AEP: typical sequences can be coded into a block code of "rate" $H(\vec{p})$, but not smaller.

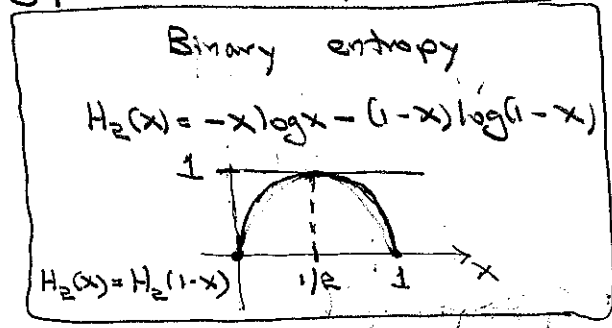
Shannon's noisy channel coding theorem.

These are data-compression results. Other approach to understanding Shannon information is in terms of yes/no questions or variable-length codes (HW).

Properties of Shannon entropy (information)

$$H(\vec{p}) = - \sum_j p_j \log p_j = - \sum_x p(x) \log p(x) = H(X)$$

$$0 \log 0 = \lim_{\epsilon \rightarrow 0} \epsilon \log \epsilon = 0$$

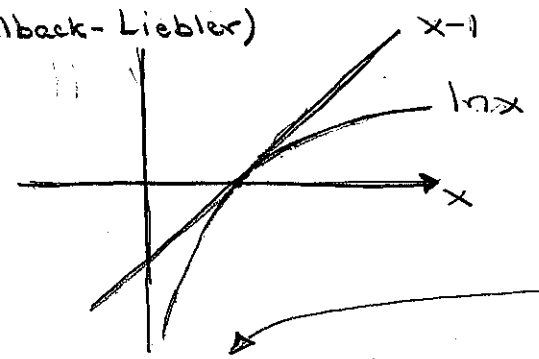


① $0 \leq H(X) \leq \log D$

↑
number of alternatives or
number of values of X

Equality:
 $\vec{p} = \vec{q}$

② Relative entropy: $H(\vec{p} \parallel \vec{q}) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$
(Kullback-Liebler)



$$\ln \log x = \ln x \leq x - 1$$

$$\Rightarrow -\ln \log x = -\ln x \geq 1 - x$$

$$H(\vec{p} \parallel \vec{q}) = \sum_x p(x) \left(-\log \frac{q(x)}{p(x)} \right) \geq \frac{1}{\ln 2} \sum_x p(x) - q(x) = 0$$

$$\geq \frac{1}{\ln 2} \left(1 - \frac{q(x)}{p(x)} \right) \leftarrow \text{Equality iff } q(x) = p(x)$$

If $q(x) = 1/D$ is the uniform distribution, we get

$$0 \leq H(\vec{p} \parallel \vec{q}) = -H(\vec{p}) + \sum_x p(x) \log D = -H(\vec{p}) + \log D$$

$$\Rightarrow H(\vec{p}) = H(X) \leq \log D$$

Alternative: $H(\vec{p} \parallel \vec{q}) = \sum_x p(x) \left(-\log \frac{q(x)}{p(x)} \right) \geq -\log \left(\sum_x p(x) \frac{q(x)}{p(x)} \right) = 0$

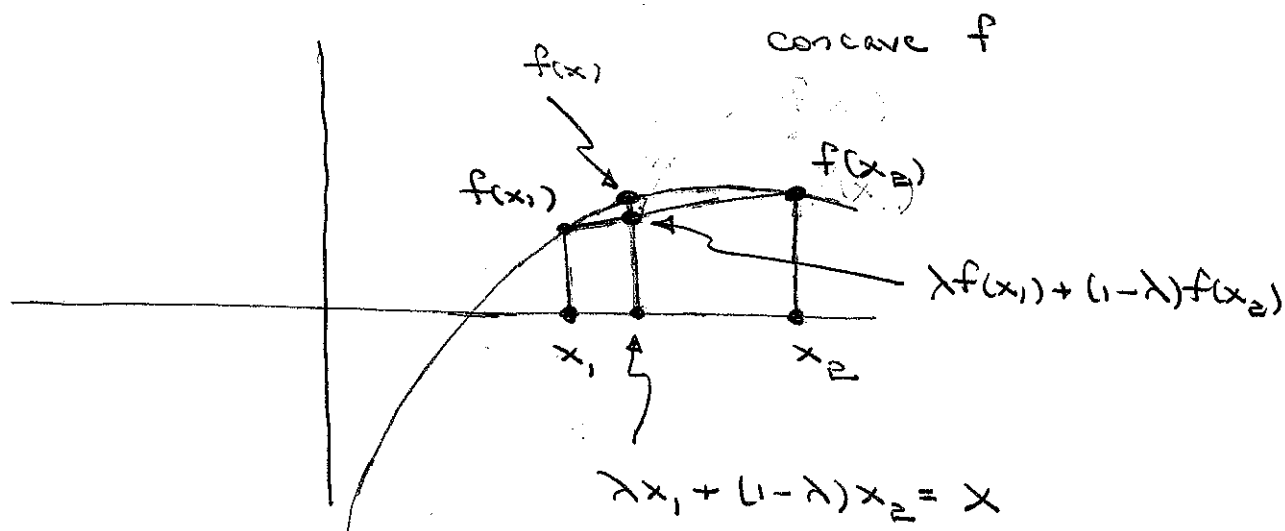
↓
Jensen's inequality

↑
convexity of
-log

$f(x)$ is a concave function if convex

$$f(\lambda x_1 + (1-\lambda)x_2) \begin{matrix} \geq \\ \leq \end{matrix} \lambda f(x_1) + (1-\lambda)f(x_2)$$

for all x_1, x_2 , and $0 \leq \lambda \leq 1$.



Jensen's inequality:

$$\langle f(x) \rangle = \sum_x p(x) f(x) \begin{matrix} \text{concave } f \\ \leq \\ \text{convex } f \end{matrix} f\left(\sum p(x) x\right) = f(\langle x \rangle)$$

Definition of concavity (convexity) interpreted in terms of averages.

Probability simplex is a convex set.

③ Concavity of Shannon entropy: Entropy increases on mixing 2 pds.

$$H(\lambda \vec{p} + (1-\lambda) \vec{q}) \geq \lambda H(\vec{p}) + (1-\lambda) H(\vec{q})$$

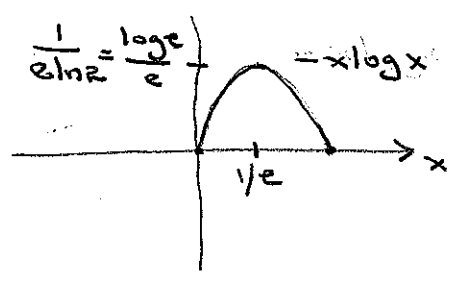
convex combination or mixture of pds \vec{p} and \vec{q}

Equality $\Leftrightarrow \lambda=0$ or 1 or $\vec{p}=\vec{q}$

$$H(\lambda \vec{p} + (1-\lambda) \vec{q}) = \sum_x -(\lambda p(x) + (1-\lambda) q(x)) \log(\lambda p(x) + (1-\lambda) q(x))$$

$$\geq \sum_x -\lambda p(x) \log p(x) - (1-\lambda) q(x) \log q(x)$$

concavity of $-x \log x$



$$\geq \lambda H(\vec{p}) + (1-\lambda) H(\vec{q})$$

Two random variables: X, Y $p(x,y)$

Joint entropy $H(X, Y) = - \sum_{x,y} p(x,y) \log p(x,y)$

Conditional entropy

$$H(X|Y) = \sum_y p(y) \left(- \sum_x p(x|y) \log p(x|y) \right)$$

information in X , given that $Y=y$, averaged over Y

$$\begin{aligned}
 H(X|Y) &= - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \\
 &= - \sum_{x,y} p(x,y) \log p(x,y) + \sum_y p(y) \log p(y) \\
 &= H(X,Y) - H(Y)
 \end{aligned}$$

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

Mutual information:

$$\begin{aligned}
 H(X:Y) &= H(X) - H(X|Y) \\
 &= H(Y) - H(Y|X) \\
 &= H(X:Y)
 \end{aligned}$$

Reduction in information in X by determining Y; Information shared by X and Y

Symmetric in X and Y

$$\begin{aligned}
 H(X:Y) &= - \underbrace{\sum_x p(x,y) \log p(x)}_{H(X)} + \underbrace{\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)}}_{-H(X|Y)} \\
 &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
 &= H(p(x,y) || p(x)p(y))
 \end{aligned}$$

$$\geq 0$$

Equality $\iff p(x,y) = p(x)p(y)$

Properties of Shannon entropy of two random variables:

$$\textcircled{1} \quad H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$\Rightarrow H(X), H(Y) \leq H(X, Y)$$

$$\textcircled{2} \quad H(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y:X) \geq 0$$

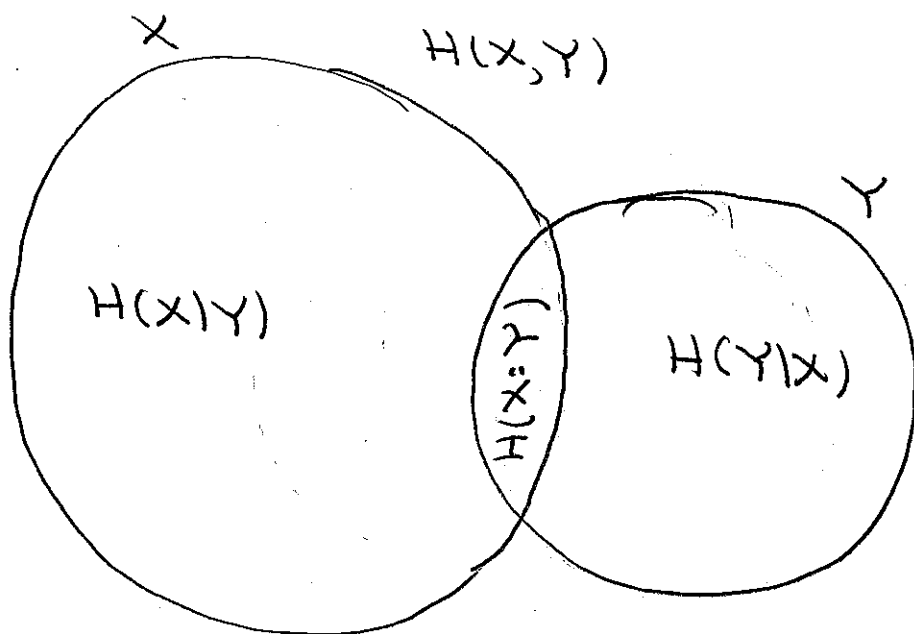
$$\Rightarrow H(X) \geq H(X|Y), \quad H(Y) \geq H(Y|X)$$

$$H(X:Y) \leq H(X), H(Y)$$

$$\textcircled{3} \quad H(X, Y) = H(X) + H(Y) - H(X:Y) \leq H(X) + H(Y)$$

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

↑
Subadditivity



Information

Classical

vs.

quantum



role of probabilities
manipulate realistic
alternatives which
encode information



role of probabilities

States

[quantum dynamics]
[measurements]

no realistic alternatives

no joint probabilities

for noncommuting observables