Entropy:

alphabet ↓

↳ discrete random variable: X, $A = \{x_1, \dots, x_n, \dots\}$

↳ probability function of X: $p(x) \geq 0$, $\sum_i p(x_i) = 1$.

* The entropy of X is a measure of the uncertainty of a random variable defined as: $H(X) = -\sum p(x_i) \log p(x_i) \geq 0$

($H(X)$: in bits; $\log$: base 2 log)

* Various possible axiomatic justifications. Shannon:

↳ sequence of symmetric functions $H_m(p_1, \dots, p_m)$ such that i) $H_2(1/2, 1/2) = 1$, ii) $H(p, 1-p)$ continuous in p

iii) $H_m(p_1, \dots, p_m) = H_{m-1}(p_1+p_2, p_3, \dots, p_m) + (p_1+p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$

* The joint entropy of (X, Y) is $H(X,Y) = -\mathbb{E}[\log p(X,Y)]$

* The conditional entropy of $Y|X$ is

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

* The chain rule

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

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

Generalization. (independence bound)

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \leq \sum_{i=1}^n H(X_i)$$

↑ nested conditioning

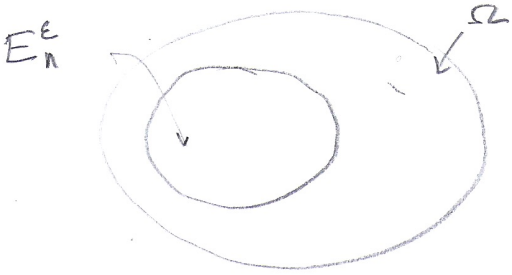
Entropy and description length

$X_1, \dots, X_n$ i.i.d then the asymptotic equipartition property (law of large numbers) states that

$$-\frac{1}{n}\log p(X_1, \dots, X_n) \xrightarrow[n \to +\infty]{} H(X)$$

"Almost all events are almost equally surprising"

Shannon code: typical set: $E_n^\varepsilon : \left\{ x_1, \dots, x_n \mid 2^{-n(H(x)+\varepsilon)} \leq p(\underline{x}) \leq 2^{-n(H(x)-\varepsilon)} \right\}$



$$1 = \sum_{\underline{x}} p(\underline{x}) \geq \sum_{\underline{x} \in E_n^\varepsilon} p(\underline{x}) \geq |E_n^\varepsilon| \, 2^{-n(H(x)+\varepsilon)}$$

$$\Rightarrow |E_n^\varepsilon| \leq 2^{n(H(x)+\varepsilon)}$$

1 bit to ensure integer
1 bit to encode whether $\underline{x} \in E_n^\varepsilon$

Description length of E_n^ε: $n(H(x)+\varepsilon) + 2$ bits

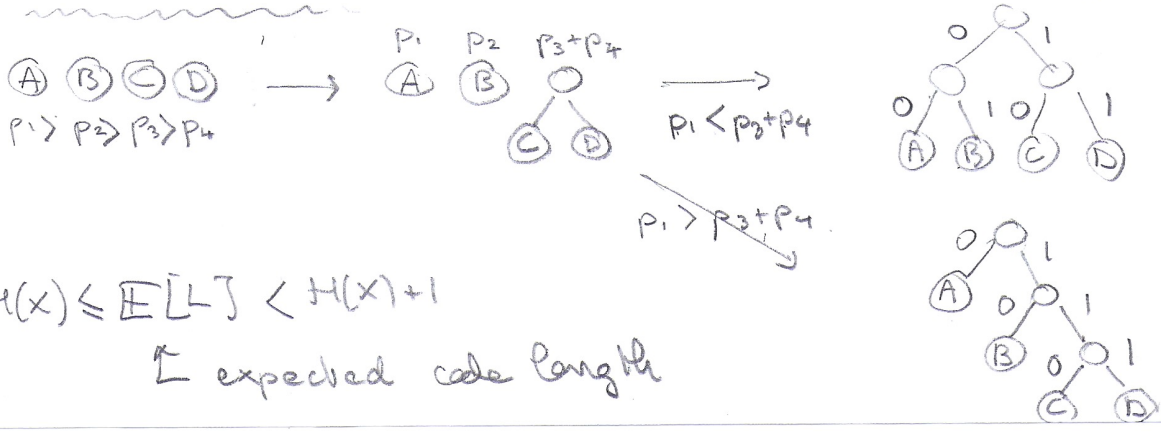
—— of $\Omega \setminus E_n^\varepsilon$: $n(\log A) + 2$ bits

Average length: $\mathbb{E}[L(X^n)] = P[E_n^\varepsilon]\,[n(H(X)+\varepsilon)+2] + (1 - P[E_n^\varepsilon])\,[n(\log A)+2]$

n large enough $\leq n(H(X)+\varepsilon) + 2 + \varepsilon\, n(\log A) + 2$

$= n(H(X)+\varepsilon') + o(1)$

Huffman code: $H(X)$: optimal average code length.



$H(X) \leq \mathbb{E}[L] < H(X)+1$

$\mathbb{E}[L]$ expected code length

Kullback - Leibler divergence (relative entropy)

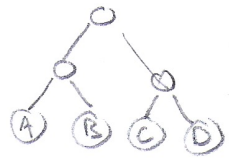
↳ KL divergence measure how much two distributions p and q differ from one another. Informally, it quantifies the inefficiency of coding p while using an optimal code for q.

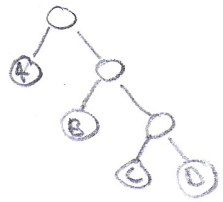
* The KL divergence between p and q is $D[p\|q] = \mathbb{E}_p\left[\log \frac{p}{q}\right] \geq 0$

* $D[p\|q]$ is convex in the pair (p,q):

$$D[\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2] \leq \lambda D[p_1\|q_1] + (1-\lambda) D[p_2\|q_2]$$

This follows from Jensen inequality + convexity of $-\log$

* Huffman trees interpretation: p: 

q: 

$$D_{KL}(p\|q) = \underbrace{-\mathbb{E}_p[\log q]}_{\text{average code length using optimal encoding for } q} - \underbrace{H(p)}_{\text{optimal description length of } p}$$

* The chain rule: $D[p(x,y)\|q(x,y)] = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x))$

* "Second law of thermodynamics":

$$p(x_n, x_{n+1}) = K(x_{n+1}|x_n)\,p(x_n)$$
$$q(x_n, x_{n+1}) = K(x_{n+1}|x_n)\,p(x_n)$$

↳ Markov kernel

$$D[p(x_n, x_{n+1})\|q(x_n, x_{n+1})]$$
$$= D[p(x_n)\|q(x_n)] + \underbrace{D[K(x_{n+1}|x_n)\|K(x_{n+1}|x_n)]}_{=0}$$
$$= D[p(x_{n+1})\|q(x_{n+1})] + D[p(x_n|x_{n+1})\|q(x_n|x_{n+1})] \geq 0$$

$$\Rightarrow D[p(x_{n+1})\|q(x_{n+1})] \leq D[p(x_n)\|q(x_n)]$$

$\Rightarrow$ KL divergence to stationary distribution decreases with time

Mutual information

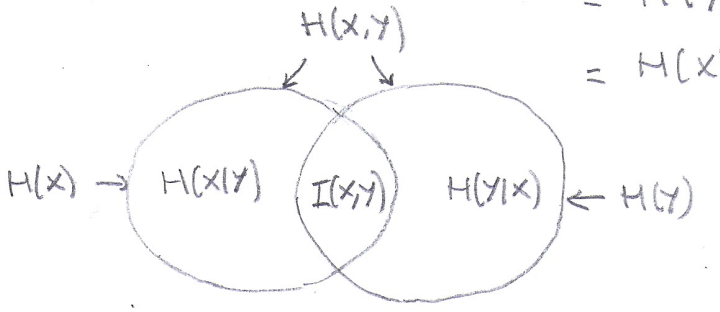
The mutual information between X and Y, denoted $I(X,Y)$ is the KL divergence between the joint distribution and the product distribution:

$$I(X,Y) = D[p(x,y) \,\|\, p(x)p(y)] = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \geq 0$$

($p(x)$, $p(y)$: marginals)

$$I(X,Y) = 0 \iff X, Y \text{ independent}$$

* Alternatively: $I(X,Y) = H(X) - H(X|Y)$ (conditioning reduces entropy; reduction of uncertainty of X (resp Y) due to knowledge of Y (resp X))
$= H(Y) - H(Y|X)$
$= H(X) + H(Y) - H(X,Y)$



* Conditional mutual information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$
$$= \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)\,p(y|z)}\right]$$

$I(X;Y|Z) = 0 \iff X|Z, Y|Z$ independent

* The chain rule: $I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i, Y \mid X_{i-1},\dots,X_1)$

* $I(X,Y)$ is concave in $p(x)$ and convex in $p(y|x)$

↳ concavity: $I(X,Y) = H(Y) - \sum_x p(x) H(Y|X=x)$

(H concave in $p(Y)$ ← linear in $p(x)$; $p(x)$ fixed)

↳ convexity: consider $X \xrightarrow{p_1(y|x)} Y_1$, $X \xrightarrow{p_2(y|x)} Y_2$ for fixed $p(x)$

consider $Y^{(\lambda)} = Y_\theta$ where θ (independent) $= \begin{cases} 1 & \text{with probability } \lambda \\ 2 & \text{otherwise} \end{cases}$

$X \longrightarrow Y^{(\lambda)}$ with conditional probability

$$p_\lambda(y|x) = \lambda\, p_1(y|x) + (1-\lambda)\, p_2(y|x)$$

we can check that: $p_\lambda(x,y) = \lambda\, p_1(x,y) + (1-\lambda)\, p_2(x,y)$

$$p(x)\, p_\lambda(y) = \lambda\, p(x)\, p_1(y) + (1-\lambda)\, p(x)\, p_2(y)$$

$$I(Y^{(\lambda)}, X) = D\left[\lambda p_1(x,y) + (1-\lambda) p_2(x,y) \,\|\, \lambda p(x) p_1(y) + (1-\lambda) p(x) p_2(y)\right]$$

($Y^{(\lambda)}$ with mixture kernel p_λ)

Kullback-Leibler convex ✓

* $X \longrightarrow Y$ encoding channel

$p(x)$ ↑ $p(y|x)$

$\min_{p(y|x)} I(X,Y) = 0$ ← uninteresting without constraints

$\max_{p(x)} I(X,Y) = C_{p(y|x)}$ ← channel capacity

Interest for neuroscience: optimizing over $p(y|x)$

not necessarily well-posed!

Data processing inequality

Markov chain: $X \to Y \to Z$; $p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$

Remark: • $X|Y$ and $Z|Y$ are independent

$$p(x,z|y) = \frac{p(x,y)\,p(z|y)}{p(y)} = p(x|y)\,p(z|y)$$

• $X \to Y \to Z \Rightarrow Z \to Y \to X$

No processing, deterministic or random, can increase the information that Y contains about X:

$$X \to Y \to Z \Rightarrow I(X,Y) \geq I(X,Z)$$

Proof: $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$

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

conditional independence $\Rightarrow I(X;Z|Y) = 0$

$I(X;Y|Z) \geq 0 \Rightarrow I(X;Y) \geq I(X,Z)$ ✓

Moreover: $I(X;Y|Z) \leq I(X;Y)$ if $X \to Y \to Z$

Observation of "downstream" variable reduces the dependence between "upstream" variables.