"Surprise"

$S(p) =$ "amount of surprise from learning that an event with probability p of happening took place."

Axiom 1. $S(1) = 0$

Axiom 2. $S(p)$ is strictly decreasing as a function of p:

$$p < q \implies S(p) > S(q)$$

Axiom 3. $S(p)$ is a continuous function of p.

Axiom 4. $S(pq) = S(p) + S(q), \qquad p, q \in (0,1]$

Thm. If $S: (0,1] \to \mathbb{R}$ satisfies the above Axioms 1–4, then $S(p) = -C \log_2 p$; where $C > 0$.

Pf. From Axiom 4 with $q = p$, $S(p^2) = S(p) + S(p) = 2S(p)$

with $q = p^2$, $S(p^3) = S(p^2) + S(p) = 3S(p)$

$\vdots$

by induction, we find that $S(p^m) = m \cdot S(p)$ for all $m \in \mathbb{N}$.

More generally, replacing p with $p^{1/n}$ above, we have:

$$S(p) = S(\underbrace{p^{1/n} \cdot p^{1/n} \cdot \ldots \cdot p^{1/n}}_{n}) = n\, S(p^{1/n})$$

Thus $S(p^{1/n}) = \frac{1}{n} S(p)$

Altogether, $S(p^{m/n}) = m\, S(p^{1/n}) = \frac{m}{n} S(p)$, i.e.

$S(p^x) = x\, S(p)$ for all $x \in \mathbb{Q}$.

By density of $\mathbb{Q}$ in $\mathbb{R}$ and continuity of S (Axiom 3), we conclude that

$$S(p^x) = x\, S(p) \quad \text{for all } x \in \mathbb{R}$$

For any $p \in (0,1]$, let $x = -\log_2 p$, so $p = \left(\frac{1}{2}\right)^x$. Therefore

$$S(p) = S\left(\left(\frac{1}{2}\right)^x\right) = x \underbrace{S\left(\frac{1}{2}\right)}_{= C} = -C \log_2 p$$

where $C = S\left(\frac{1}{2}\right) \overset{\text{Axiom 2}}{>} S(1) \overset{\text{Axiom 1}}{=} 0$. $\square$

Q: What does "Surprise" mean?

A: In probability, "surprise" measures uncertainty.

In information theory, "surprise" measures the amount of information that is learnt upon observing an event.

It is customary to normalize $C=1$, in which case

$$I(p) = S(p) = -\log_2 p$$

is called "information content", measured in bits.

Def: The Shannon Entropy of a discrete random variable X is the expected value of the information content of X:

$$H(X) = E(I(X)) = \sum_{i=1}^{n} p_i \underbrace{I(p_i)}_{= -\log_2 p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where:

X	x_1	x_2	$\ldots$	x_n
$P(X=x_i)$	p_1	p_2	$\ldots$	p_n

$p_i = P(X=x_i)$

prob. mass function of X.

Example: Flip of a (possibly biased) coin.

$$X \sim \text{Bernoulli}(p) \qquad P(X=H) = p$$
$$P(X=T) = 1-p$$

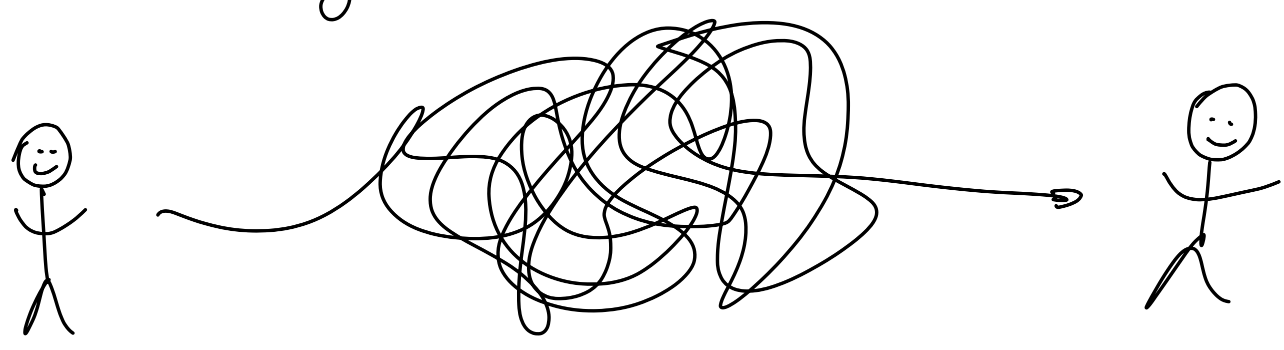
$$H(X) = -p\log_2 p - (1-p)\log_2(1-p)$$

Note that the largest value possible for $H(X)$ is attained when $p = 1/2$ (unbiased coin):

$$H(X) = -\log_2 \frac{1}{2} = +1.$$

while, e.g., with $p = 0.7$, one has $H(X) \cong 0.882 < 1$.

Codes and coding



measures X which takes 4 possible values:

A

$x_1 \Longleftrightarrow 00$
$x_2 \Longleftrightarrow 01$
$x_3 \Longleftrightarrow 10$
$x_4 \Longleftrightarrow 11$

or

B

$x_1 \Longleftrightarrow 0$
$x_2 \Longleftrightarrow 10$
$x_3 \Longleftrightarrow 110$
$x_4 \Longleftrightarrow 111$

or

C

$\cancel{x_1 \Longleftrightarrow 0}$
$\cancel{x_2 \Longleftrightarrow 1}$
$\cancel{x_3 \Longleftrightarrow 00}$
$\cancel{x_4 \Longleftrightarrow 11}$

To avoid ambiguities, need that none of the sequences (codes) is an extension of another shorter sequence.

Q: How many bits are we going to send on average?

A: For example, if the prob. distr. of X is

X	x_1	x_2	x_3	x_4
$P(X=x_i)$	$1/2$	$1/4$	$1/8$	$1/8$

Then:

$$E(\#\text{ bits using } A) = \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 2 + \frac{1}{8}\cdot 2 = \underline{\underline{2}}$$

$$E(\#\text{ bits using } B) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3$$

$$= 1 + \frac{3}{4} = \underline{\underline{1,75}}$$

Therefore, using B is advantageous: on average, we will only need 1.75 bits to be sent each time, compared with 2 bits on average if we used A.

Q1: What is the "best" possible code?

Q2: How small can the expected # of bits be?

A2: The expected # of bits needed to encode a random variable X is at least $H(X)$.

(Before proving this, need some preliminary work)

Encode X using binary codes as follows:

$x_1 \longleftrightarrow$ word of 0's and 1's of length n_1

$x_2 \longleftrightarrow$ ——— " ——— n_2

$x_3 \longleftrightarrow$ ——— " ——— n_3

$\vdots$

$x_N \longleftrightarrow$ ——— " ——— n_N

Lemma: Let n_j = length of code used for x_j. Such an unambiguous code (with these lengths) exists if and only if $\sum_{j=1}^{N} \left(\frac{1}{2}\right)^{n_j} \leq 1$

Pf: Let $w_j := \#\{i : n_i = j\}$. number of codes with length j

Need $w_1 = \#\{i : n_i = 1\} \leq 2$. only 2 letters are available.

Similarly, $w_2 \leq 2^2 - 2w_1$ b/c words of length 2 cannot be extensions of words of length 1

By induction, one sees that:

$$w_n \leq 2^n - w_1 2^{n-1} - w_2 \cdot 2^{n-2} - \dots - w_{n-1} \cdot 2^1$$

is both necessary and sufficient.

Rewrite the above as:

$$w_n + w_{n-1} \cdot 2 + w_{n-2} \cdot 2^2 + \dots + w_2 \cdot 2^{n-2} + w_1 \cdot 2^{n-1} \leq 2^n$$

Dividing by 2^n, we have:

$$\underbrace{\frac{w_n}{2^n} + \frac{w_{n-1}}{2^{n-1}} + \frac{w_{n-2}}{2^{n-2}} + \dots + \frac{w_2}{2^2} + \frac{w_1}{2^1}}_{\sum_{j=1}^{n} \frac{w_j}{2^j}} \leq 1.$$

for all n.

Since $w_j > 0$, the above holds for all $n \in \mathbb{N}$ if and only if the series with $n \nearrow +\infty$ satisfies

$$\sum_{j=1}^{\infty} w_j \left(\frac{1}{2}\right)^j = \sum_{j=1}^{\infty} \frac{w_j}{2^j} \leq 1$$

Recall $w_j = \#\{i : n_i = j\}$, so $\sum_{j=1}^{\infty} w_j \left(\frac{1}{2}\right)^j = \sum_{i=1}^{N} \left(\frac{1}{2}\right)^{n_i}$

Shannon's (noiseless) Coding Theorem.

Any binary code that unambiguously encodes a discrete random variable X satisfies

$$\underbrace{E\left(\begin{matrix}\#\text{ bits that need}\\ \text{to be sent}\end{matrix}\right)}_{\sum_{i=1}^{N} n_i\, p(x_i)} \geq \underbrace{H(X)}_{-\sum_{i=1}^{N} p(x_i)\log_2 p(x_i)}.$$

Pf. Let $p_i = p(x_i)$, $q_i = \dfrac{2^{-n_i}}{\sum_{j=1}^{N} 2^{-n_j}}$. Note that

$$\sum_{i=1}^{N} p_i = 1 \quad \text{and}$$

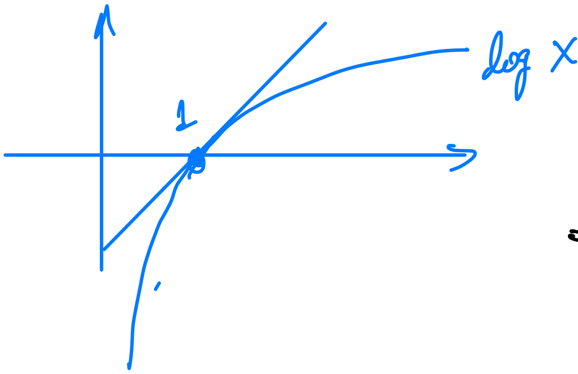
$$\sum_{i=1}^{N} q_i = \sum_{i=1}^{N} \frac{2^{-n_i}}{\sum_{j=1}^{N} 2^{-n_j}} = \frac{\sum_{i=1}^{N} 2^{-n_i}}{\sum_{j=1}^{N} 2^{-n_j}} = 1.$$

Consider

$$-\sum_{i=1}^{N} p_i \log_2\left(\frac{p_i}{q_i}\right) = +\log_2 e \sum_{i=1}^{N} p_i \log\left(\frac{q_i}{p_i}\right)$$

$$\log_2 x = \log_2 e \log x$$

$\log x \leq x - 1$
$\forall x > 0$



$$\leq \log_2 e \sum_{i=1}^{N} p_i \left(\frac{q_i}{p_i} - 1 \right)$$

$$= \log_2 e \left(\sum_{i=1}^{N} (q_i - p_i) \right)$$

$$= \log_2 e \left(\underbrace{\sum_{i=1}^{N} q_i}_{=1} - \underbrace{\sum_{i=1}^{N} p_i}_{=1} \right) = 0$$

Thus

$$- \sum_{i=1}^{N} p_i \log_2 \left(\frac{p_i}{q_i} \right) \leq 0$$

Since $\log_2 \frac{p_i}{q_i} = \log_2 p_i - \log_2 q_i$, we have:

$$H(X) = - \sum_{i=1}^{N} p_i \log_2 p_i \leq - \sum_{i=1}^{N} p_i \log_2 q_i$$

$$= - \sum_{i=1}^{N} p_i \log_2 \left(\frac{2^{-n_i}}{\sum_{j=1}^{N} 2^{-n_j}} \right)$$

$$= -\sum_{i=1}^{N} p_i \underbrace{\log_2\left(2^{-n_i}\right)}_{-n_i} - \underbrace{p_i}_{>0} \underbrace{\log_2\Big(\underbrace{\sum_{j=1}^{N} 2^{-n_j}}_{\leq 1}\Big)}_{<0}$$

$$= \sum_{i=1}^{N} p_i \cdot n_i + \sum_{i=1}^{N} p_i \underbrace{\log_2\left(\sum_{j=1}^{N} 2^{-n_j}\right)}_{<0}$$

$$\leq \sum_{i=1}^{N} p_i n_i = E(\#\text{bits needed}).$$

$\square$

This gives a satisfactory answer to Q2. Regarding Q1, for general random variables, there does not exist a code realizing equality in Shannon's bound; i.e., typically $E(\#\text{bits needed}) > H(X)$. However, it is always possible to devise a code with $E(\#\text{bits needed}) < H(X)+1$.