Info of event

Probability of an event happening is $Pr(A)$. The amount of information we learn from A can be $(information)(A)$, or $\psi(Pr(A))$.

Non-negative. No such thing is negative information.

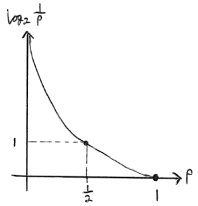
Zero for definite events. If we know an event will happen, we wont learn anything from it.

Monotone. The less likely an event is to happen, the more information we can learn from it.
$p \leq p', \psi(p) \geq \psi(p')$

Continuity. Small changes in probability of event occurring does not cause big changes in info we learn.

Addivity under independence.
$\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$. If events are independent, the information we learn from each event is independent of each other. So the total information learnt is the sum.

Axiom satisfaction. $\psi(p) = log_b \frac{1}{p}$ where $b > 0$ satisfies all 5. b tells us how is information measured. $b = 2$ means measured in bits. $b = e$ means measured in nats.



Entropy

Discrete random variables, then the probability mass function is $P_X(x) = Pr(X = x)$. If $X = X$, then we learnt $\psi(P_X(x))$.

Shannon Entropy. Amount of information (after X) or uncertainty (before X).

$$\psi(P_X(x)) = \sum_x P_X(x) log_2 \frac{1}{P_X(x)}$$

Binary Entropy function. If X is bernoulli, then

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

Uniform Entropy function. If $P_X(x)$ is the same for all x, then

$$H(X) = log_2(|\mathcal{X}|)$$

Continuity property. Small changes in probability dont give large changes in info/uncertainty.

Uniform case. Larger number of possible outcomes means we can gain more info/uncertainty from it.

Successive decisions. First draw from a distribution that doesnt resolve 2 symbols and then draw from another if we need to resolve it.

$$\psi(p_1, \ldots, p_N) = \psi(p_1 + p_2, p_3, \ldots) + (p_1 + p_2)\psi(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$$

Joint Entropy.

$$H(X,Y) = \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\log_2 \frac{1}{P_{XY}(X,Y)}\right] = \sum_{x,y} P_{XY}(x,y) log_2 \frac{1}{P_{XY}(x,y)}$$

May be helpful to build a table for $P_{X,Y}$ to get this value.

Conditional Entropy. Amount of info we get from the next event after observing another. If X uniquely determines Y, then $H(Y|X) = 0$.

$$H(Y|X) = \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right] = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{Y|X}(y|x)} = \sum_x P_X(x) H(Y|X=x)$$

$\sum_x P_X(x)$ is the average over x. $H(Y|X=x)$ is the amount of info we can learn from Y given that we have seen that $X = x$. Its possible to sub in conditional probability to get a value for this directly.

Non-negative.

$$H(X) \geq 0$$

Only when we know that an event will definitely happen, then information gained will be 0.

Proof. $log_2 \frac{1}{p}$ is always non-negative. Since entropy is the average of a quantity that is always non-negative, its also non-negative. Only $p = 1$ gives $H(X) = 0$ iff X is deterministic.

Upper-bound.

$$H(X) \leq log_2(|\mathcal{X}|)$$

Equality only when distribution is uniform. Implies that $H(X|Y) \leq log_2|\mathcal{X}|$.Since uniform is most uncertain since we need to randomly guess.

Proof. Let P be the distribution of X and let Q be the uniform distribution on $\mathcal{X}$ so $Q(x) = \frac{1}{|\mathcal{X}|}$. Then

$$\sum_x P(x) log_2 \frac{P(x)}{Q(x)} = \sum_x P(x) log_2(|\mathcal{X}| \cdot P(x)) = log_2|\mathcal{X}| + \sum_x P(x) log P(x) = log_2|\mathcal{X}| - H(X)$$

Chain Rule.

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

Overall info is the sum of the info of seeing one event and the info of seeing the other given that we saw the first.

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

Proof. Express $P_{XY}(X,Y)$ as $P_X(X)P_{Y|X}(Y|X)$. Then, split the logs to get the expressions.

Conditioning reduces entropy.

$$H(X|Y) \leq H(X)$$

If seeing one event does not affect the information of seeing another, we get equality. This is the upperbound since that is without seeing other events.

Possible for the following: $H(X|Y=y) > H(X)$

Subaddivity.

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$$

Relative Entropy.

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} = \mathbb{E}_{X\sim P}\left[\log_2 \frac{P(X)}{Q(X)}\right]$$

Measures how similar the 2 distributions are. Equality holds when they are the same distribution since no differences. $D(P||Q) \neq D(Q||P)$. No triangle inequality as well.

Proof.

$$-D(P||Q) = \sum_x P(x) \log_e \frac{Q(x)}{P(x)} \leq \sum_x P(x)(\frac{Q(x)}{P(x)} - 1) = \sum_x (Q(x) - P(x)) = 0$$

$\frac{Q(x)}{P(x)} = 1$ iff $P = Q$. Uses $log_e \alpha \leq \alpha - 1$

Decomposition. Let P be the set that contains a set of probabilities $P = \{p_1, \ldots, p_n\}$

$$H(P) = H(p_1, 1-p_1) = (1-p_1) H(\frac{p_2}{1-p_1}, \ldots, \frac{p_n}{1-p_1})$$

Mutual Information

How much information X gives about Y.

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) = D(P_{XY}||P_X \times P_Y)$$

Joints. $I(X_1, X_2; Y_1, Y_2)$ Similar to above but $X \leftarrow (X_1, X_2)$ and $Y \leftarrow (Y_1, Y_2)$

Conditional. $I(X;Y|Z)$. Conditions **both** X and Y on Z.

$$I(X;Y|Z) = H(Y|X,Z) - \sum_z P_Z(z) I(X;Y|Z=z)$$

Expanded out we get:
$\sum_z \sum_y \sum_x P_{X,Y,Z}(x,y,z) log \frac{P_{X,Y,Z}(x,y,z) P_Z(z)}{P_{X,Z}(x,z) P_{Y,Z}(y,z)}$

Independence. If X, Y are independent, then $H(Y|X) = H(Y)$ so $I(X;Y) = 0$.

Equivalence. If $Y = X$, then $H(Y|X) = 0$ so $I(X;Y) = H(Y)$

Symmetry. $I(X;Y) = I(Y;X)$. X, Y reveal equal amount of information about each other.

Non-negativity. Mutual information is never negative. Equality only when independent.

$$D(P_{XY}||P_X \times P_Y) \geq 0$$

iff $P_{X,Y} = P_X \times P_Y = P_Y$ which also implies independence.

Upper bound.

$$I(X;Y) \leq H(X)$$

equal iff $H(X|Y) = 0$ iff X is deterministic given Y. Vice versa for Y.

Chain Rule.

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X)$$
$$I(X, Y; Z) = I(X; Z) + I(Y; Z|X)$$

Total information of X_1, X_2 is the sum of information gained from X_1 and information gained from X_2 given X_1. Can be extended to be more general.

$$I(X_1, \ldots, X_n; Y) = \sum_n I(X_i; Y|X_1, \ldots X_{i-1})$$

Markov chain. $X \to Y \to Z$

Data processing inequality. If Z depends on (X,Y) **only** via Y, then $I(X;Z) \leq I(X;Y)$. Post processing cannot increase info about X. Iff given Y, Z is independent of X. $X \to Y$ is via a channel and $Y \to Z$ is post processing. Equality holds when $I(X;Y|Z) = 0$. This means that all information that Y reveals about X is revealed by Z alone. Or $X \to Z \to Y$

Partial sub-addivity. $I(X_1, \ldots X_n; Y_1, \ldots Y_n)$ can be smaller or larger than $\sum_n I(X_i; Y_i)$ but typically is $\leq$

Larger or equal. When $X_1 \ldots, X_n$ are mutually independent.

Smaller or equal. When given X_i the rv Y_i is conditionally independent of all remaining variables.

Conditional MI. Might not hold all the time. There is a distribution that makes $I(X;Y|Z) > I(X;Y)$. Let X, Y be 2 bernoulli distributions with $p = \frac{1}{2}$ and $Z = X + Y$.

Symbol Source

Consider the letters in the english alphabet. A possible probabilistic model would be frequency. So a higher frequency leads o shorter encoded length.

$$L(C) = \sum_x P_X(x) l(x)$$

$l(x)$ is the length of a sequence of binary code for x.

First requirement. $C(x) \neq C(x')$ if $x \neq x'$. Not enough. Consider a to 0, b to 1 and c to 01. Impossible to distinguish 01 since it can map to ab or c.

Uniquely Decodable. No 2 sequences (of equal/unequal lengths) of symbols in $\mathcal{X}$ are coded to the same concatenated binary sequence. $x_1, \ldots, x_n$ always uniquely identified from the string $C(x_1) \ldots C(x_n)$.

Prefix-free. No $C(x)$ is a prefix of any $C(x')$ where $x \neq x'$. Its not possible to append more bits to some $C(x)$ in order to produce some other $C(x')$. 2 codes may look to share some common prefix but if that prefix is not valid code, then those codes are unique.

Krafts Inequality. Any prefix-free (any uniquely decodable code) code that maps each $x \in \mathcal{X}$ to a code word of length $l(x)$ must satisfy $\sum_x 2^{-l(x)} \leq 1$.

Proof. In a binary tree, each node is a code word. If there is a codeword that is used at some point in the tree, then there are no codewords further down the tree. The probability of getting any of the codeword is $2^{-l(x)}$. Since total probability of hitting codewords cannot exceed 1, so the sum of them must be ≤ 1.

Existence Property. If there are lengths that satisfies kraft's inequality, then its possible to construct prefix-free code that maps each $x \in (X)$ to a codeword of length $l(x)$.

Proof.

Given codewords, purchase the codewords from shortest to longest. Success if the sum of prices are ≤ 1. **Entropy Bound.** For $X \sim P_X$ and any prefix free code, the expected length satisfies $L(C) \geq H(X)$. Equality iff $P_X(x) = 2^{-l(x)}$ for all $x \in \mathcal{X}$

Proof. First step is to expand by definition and then make both sides have logs by using $l(x) = log_2(2^{l(x)})$. Then let $Z = \sum_x 2^{-l(x)}$ and define a PMF, $Q_X(x) = \frac{2^{-l(x)}}{Z}$. Make $2^{-l(x)}$ the subject and substitute into the equation. Bringing out the constant term, we get $D(P||Q)$. Since by krafts inequality, $Z \leq 1$ and $D(P||Q) \geq 0$, $L(C) - H(X) \geq 0$. For equality, both need to hold with equality.

Optimality. If we can remove some suffix of bits to still get a prefix-free code, then its not optimal.

Shannon-Fano Code.

$$\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$$

Satisfy existence property theorem since

$$\begin{aligned}\sum_x 2^{-l(x)} &= \sum_x 2^{-\left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil} \\ &\leq \sum_x 2^{-\log_2 \frac{1}{P_X(x)}} \\ &= \sum_x P_X(x) = 1\end{aligned}$$

Average length. Average length satisfies $H(X) \leq L(C) \leq H(X) + 1$

Unknown distribution. Apply Shannon-Fano code to Q_x but true distribution is P_x. We get a mismatch case so we have $H(X) + D(P_X||Q_X) \leq L(C) \leq H(X) + D(P_X||Q(X)) + 1$. Relative entropy is the penalty due to mismatch here.

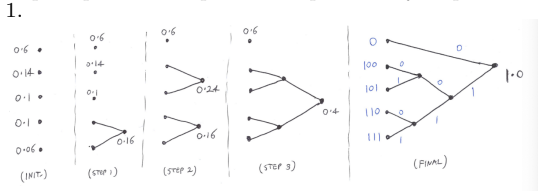
Proof. Similar to above but use

$$\mathbb{E}_p\left[\log_2 \frac{1}{Q_X}\right] = \mathbb{E}_p\left[\log_2 \frac{P_X}{Q_X \times P_X}\right]$$

Fraction obtain from $\frac{D(P_X||Q_X)}{H(X)}$

Huffman Code. Find the shortest code that is uniquely decodable. If a suffix can be removed and the code is still uniquely decodable, then its not optimal which means its not a Huffman Code.

Kraft's inequality. Does not violate Kraft's inequality since its always prefix-free; always satisfies with equality. In the case of $\sum_n 2^{-l(x)} < 1$, we will be able to shorten one or more lengths, to create a new code that satisfies equality. Contradicts the fact that Huffman is shortest. **Method.** List the symbols from highest probability to lowest. Draw a branch connect 2 symbols with the lowest and label the merged point with the sum of the 2 probabilities. Repeat previous step till total probability at point is 1.



Theorem. No uniquely decodable code can achieve a smaller average length $L(C)$ than the Huffman code. Always optimal.

Properties. Don't exploit correlations/memory (dependence between subsequent symbols).

Solution. Code cover the blocks of letters instead. Can exploit statistics of groups of letters. Even if source has independent letters, this can help.

Guessing. Similar to guessing. for each outcome, apply a prefix-free code. Iterate through the codes for 1s. If the outcome is positive, we keep the bit at that position as 1, otherwise make it 0. Repeat for the subsequent bits. When we reach the end of a codeword, we have found the glass.

Shannon-Fano Guarantee.

$$\begin{aligned}H(X_1, X_2, X_3) &= \sum_{i=1}^{3} H(X_i) \\ &= 3H(X) \\ &\leq L(C) < 3H(X) + 1 \\ H(X) < \frac{1}{3}L(C) &< H(X) + \frac{1}{3}\end{aligned}$$

Initial $L(C)$ is code over triplets. After dividing, we get average length per letter.

Disadvantage. Determining the distribution of $P_{X_1} \dots P_{X_n}$ accurately is hard. Sorting the probabilities become computationally difficult.

H_2(avg) vs Avg H_2.

$$\frac{1}{n}\sum_n H_2(p_j) \leq H_2\left(\frac{1}{n}\sum_n p_j\right)$$

$H_2(p_j)$ is entropy of bernoulli with probability p_j. So let $X_j \sim \text{Bern}(p_j)$. For LHS, we define $J \sim \text{Unif}(1, \dots, n)$. Claim that LHS $= H(X_j|J)$.

$$\begin{aligned}H(X_j|J) &= \sum_n P_j(j) H(X_j|J = j) \\ &= \sum_n \frac{1}{n} H_2(p_j)\end{aligned}$$

For RHS, since X_J is binary,

$$\begin{aligned}\mathbb{P}[X_J = 1] &= \sum_n \mathbb{P}[J = j]\mathbb{P}[X_J = 1|J = j] \quad (1) \\ &= \frac{1}{n}\sum_n p_j \quad (2)\end{aligned}$$

Bizarre Balance. Assign numbers for the left scale. If its heavier, assign +1, lighter gets -1 and equal gets 0. We can draw an outcome tree to help us. In general the number of weighs is $\lceil log_{\#\text{outcomes}}(\#\text{balls}) \rceil$.

Algo to weigh. Choose outcome to weight st it provides the most information. Let Y be the test outcome. X be the uniform distribution of the ball.

Minimising $H(X|Y)$ or maximising MI is a common idea. Let entropy function take in 3 variables, left weigh, right weight and remaining. The goal is to maximise $H(Y)$. **Noise.** Argument will hold for symmetric noise.

Optimal Prefix-free code. RV X has $P_X = \{p_1, \dots, p_N\}$ and $p_1 < \dots < p_N$. $C(X)$ is optimal prefix-free code with codeword of lengths $\{l_1, \dots, l_N\}$. Codeword lengths satisfy $l_1 \geq \dots l_N$. Longest 2 lengths are the same.

Proof. Suppose an optimal code has $l_i < l_j$ for i ¡ j. Consider a code that maintains the probabilities but swaps the lengths. Leads to smaller average length which contradicts optimality.

Proof 1. Suppose otherwise. Because of prefix free property, if we remove the final $l_2 - l_1$ bits from the longest code word, we will still have a prefix free code. So now we have a new code, $C'(X)$. This new code is prefix free and the lengths are shorter as well. Contradict the fact that $C(X)$ is optimal.

Smallest const that implies length. Keep the probability as small as possible. Find ways to make the probability as small as possible (even distribution of other probabilities, for eg). Then for each case, express the remaining probabilities as a fraction of what we ar elooking for.

Probability

Expectation. $\mathbb{E}[X] = \sum_x P_X(x)x$

Avg of function. $\mathbb{E}[f(X)] = \sum_x P_X(x)f(x)$

Avg of scaled RV. $\mathbb{E}[cX] = c\mathbb{X}$

Avg of sum. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{Y}$ regardless of whether X and Y are independent.

Avg of product. If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{Y}$

Indicator function. If $\mathbf{1}\{X\}$ denotes indicator function (equaling 1 if event A holds and 0 otherwise), then $\mathbb{E}[\mathbf{1}\{X\}] = \mathbb{P}[A]$

Conditioning. $P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)}$

Total Probability. For an event A and RV X, we have $\mathbb{P}[A] = \sum_x P_X(x)\mathbb{P}[A|X = x]$

Total expectation. $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$ where the outer expectation is over X and the inner is over Y (given X).

Bayes Rule. $\mathbb{P}[A|B] = \frac{\mathbb{P}[A]\mathbb{P}[B|A]}{\mathbb{P}[B]}$

Independence. $P_{XY}(x, y) = P_X(x)P_Y(y)$ for all x, y

Equivalence. $P_{Y|X}(y|x) = P_Y(Y)$

Conditional Independence.

$$\begin{aligned}P_{XY|Z}(x, y|z) &= P_{X|Z}(x|z)P_{Y|Z}(y|z) \\ P_{Y|XZ}(y|x, z) &= P_{Y|Z}(y|z) \\ P_{X|YZ}(x|y, z) &= P_{X|Z}(x|z)\end{aligned}$$

Functions. If X and Y are independent, then $f(X)$ and $g(Y)$ are too for deterministic functions.

Conditional Independence. If X and Y are independent and $Z = X + Y$, then X, Y are not conditionally independent given Z. But if $U = Z + X$ and $V = Z + Y$ then X, Y are conditionally independent given Z but dependent due to common reliance on Z.

Joint independence. A collection of $X_1, \dots, X_n$ of random variables can be defined as $P_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_n P_{X_i}(x_i)$

Logarithms.

$$\begin{aligned}log_a x &= \frac{log_b x}{log_b a} \\ log_e x &\leq x - 1\end{aligned}$$

Equality holds iff $x = 1$