Lecture 5

Recall: Last time we defined the entropy of a discrete RV X.

Defn: If X is a discrete RV taking on values from

$$R_X = \{x_1, x_2, \ldots, x_n, \ldots\}$$

with probabilities

$$p_i = p_X(x_i) = P(\{X = x_i\}), \ i = 1, 2, \ldots,$$

then $$H(X) = \sum_{i=1}^{\infty} p_i \log \frac{1}{p_i}$$

If X takes on a finite number of values from

$$R_X = \{x_1, x_2, \ldots, x_N\}$$

then

$$H(X) = \sum_{i=1}^{N} P_X(x_i) \log \frac{1}{P_X(x_i)}$$

$$= \sum_{i=1}^{N} p_i \log \frac{1}{p_i}$$

n.b. $H(X)$ is a function of the probability vector $\underline{p} = (p_1, p_2, \ldots, p_N)$.

5.3

A note on Simplexes

A probability simplex of dimension N is an $N-1$ dimensional hyperplane which contains "probability vectors" of the form

$$\underline{p} = (p_1, \dots, p_N)$$

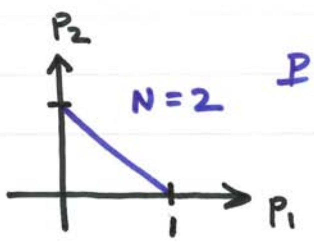
such that

(i) $p_i \geq 0 \quad , \; i = 1, \dots, N$

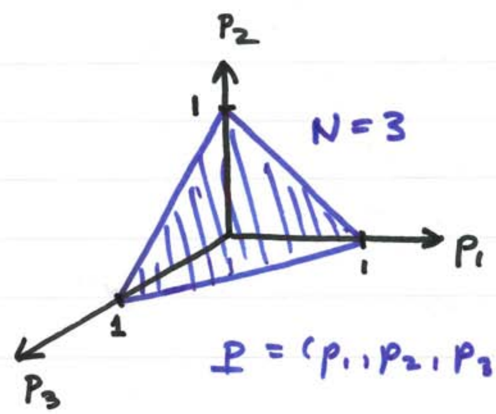
(ii) $p_1 + p_2 + \cdots + p_N = 1$

5.4

Examples of Simplexes:



$N=2$ $\quad \underline{p} = (p_1, p_2)$



$N=3$

$\underline{p} = (p_1, p_2, p_3)$

n.b. A probability simplex is a convex set.

Entropy as a Measure of Uncertainty

5.5

Certain properties of entropy make it a good measure of the uncertainty in the value that a R.V. will take on.

1. For a random variable X with finite $R_X = \{x_1, \ldots, x_N\}$ and

$$p_i = P(\{X = x_i\}) = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases}$$

$$H(X) = 0.$$

5.6

2. For X with finite $R_X = \{x_1, \ldots, x_N\}$, and

$$p_i = P(\{X = x_i\}) = \frac{1}{N}, \quad i = 1, \ldots, N,$$

$$H(X) = \log N$$

n.b. Here the entropy is increasing monotonically in the number of equiprobable values N that X can take on.

3. For a RV X with finite $R_X = \{x_1, \dots, x_N\}$,

$$H(X) \le \log N,$$

with equality iff $p_i = \frac{1}{N}$, $i = 1, \dots, N$.

<u>Proof:</u> $H(X) = \sum_{i=1}^{N} p_i \log \frac{1}{p_i} = \sum_{i=1}^{N} p_X(x_i) \log \frac{1}{p_X(x_i)}$

$$= E\Big[\log \underbrace{\left(\frac{1}{p_X(X)}\right)}_{\text{this is a RV}}\Big]$$

But $\log(\cdot)$ is a strictly convex $\cap$ function.

$$H(X) = E\left[\log\left(\frac{1}{p_X(X)}\right)\right]$$

$$\overset{(J)}{\le} \log\left(E\left[\left(\frac{1}{p_X(X)}\right)\right]\right)$$

$$= \log\left(\sum_x p_X(x) \cdot \frac{1}{p_X(x)}\right)$$

$$= \log\left(\sum_{i=1}^{N} 1\right) = \log N$$

with equality iff $\left(\frac{1}{p_X(x)}\right) = \text{const.}$

$$\Rightarrow p_X(x_i) = \frac{1}{N}, \quad i = 1, 2, \dots, N.$$

Defn: Let X and Y be two jointly distributed discrete RVs with joint pmf $P_{XY}(x,y)$. Then the joint entropy of X and Y is

$$H(X,Y) \triangleq E\left[\log \frac{1}{P_{XY}(x,y)}\right]$$

$$= \sum_{x \in R_X} \sum_{y \in R_Y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)}$$

4. Let X and Y be two jointly-distributed discrete RVs. Assume X and Y are statistically independent. Then $H(X,Y) = H(X) + H(Y)$

Proof: $H(X,Y) = -\sum_m \sum_n P_{XY}(x_m, y_n) \log P_{XY}(x_m, y_n)$

$$= -\sum_m \sum_n P_{XY}(x_m, y_n) \log \left(P_X(x_m) \cdot P_Y(y_n)\right)$$

$$= -\sum_m \sum_n P_X(x_m) P_Y(y_n) \left(\log P_X(x_m) + \log P_Y(y_n)\right)$$

$$- \sum_m \sum_n P_X(x_m) P_Y(y_n) \log P_X(x_m)$$

$$- \sum_m \sum_n P_X(x_m) P_Y(y_n) \log P_Y(y_n)$$

$$= -\sum_m p_X(x_m) \log p_X(x_m)$$

$$- \sum_n p_Y(y_n) \log p_Y(y_n)$$

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

5. **Grouping Property** Let X be a discrete RV with $R_X = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_n\}$, and let

$$A \triangleq \{x_1, \ldots, x_m\}$$

$$B \triangleq \{x_{m+1}, \ldots, x_n\}.$$

Then

$$P(A) = \sum_{i=1}^{m} P_X(x_i) \quad \text{and} \quad P(B) = \sum_{i=m+1}^{n} P_X(x_i).$$

Let

$$P_X(x_i|A) \triangleq \frac{P(\{X=x_i\} \cap A)}{P(A)} = \begin{cases} \dfrac{P_X(x_i)}{P(A)}, & i=1,\ldots,m \\ 0, & i=m+1,\ldots,n \end{cases}$$

and

$$P_X(x_i|B) \triangleq \frac{P(\{X=x_i\} \cap B)}{P(B)} = \begin{cases} 0, & i=1,\ldots,m \\ \dfrac{P_X(x_i)}{P(B)}, & i=m+1,\ldots,n. \end{cases}$$

Now define the "conditional entropies"

5.11

$$H_A(X) = -\sum_{i=1}^{m} p_X(x_i|A) \log p_X(x_i|A)$$

$$H_B(X) = -\sum_{i=m+1}^{n} p_X(x_i|B) \log p_X(x_i|B)$$

and define the indicator RV

$$Q \triangleq 1_A(X) = \begin{cases} 1, & X \in A \\ 0, & X \notin A \end{cases}$$

Then

$$H(X) = P(A)\, H_A(X) + P(B)\, H_B(X) + H(Q)$$

Proof:

5.12

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i = -\sum_{i=1}^{m} p_i \log p_i - \sum_{i=m+1}^{n} p_i \log p_i$$

$$= -P(A) \sum_{i=1}^{m} p_X(x_i|A) \left[\log p_X(x_i|A) + \log P(A) \right]$$

$$- P(B) \sum_{i=m+1}^{n} p_X(x_i|B) \left[\log p_X(x_i|B) + \log P(B) \right]$$

$$= \boxed{-P(A)\log P(A) - P(B)\log P(B)} \quad H(Q)$$

$$\boxed{- P(A) \sum_{i=1}^{m} p_X(x_i|A) \log p_X(x_i|A)} \quad P(A)\cdot H_A(X)$$

$$\boxed{- P(B) \sum_{i=m+1}^{n} p_X(x_i|B) \log p_X(x_i|B)} \quad P(B)\cdot H_B(X)$$

$$= P(A)\, H_A(X) + P(B)\, H_B(X) + H(Q)$$

Example: Problem 3 of Homework

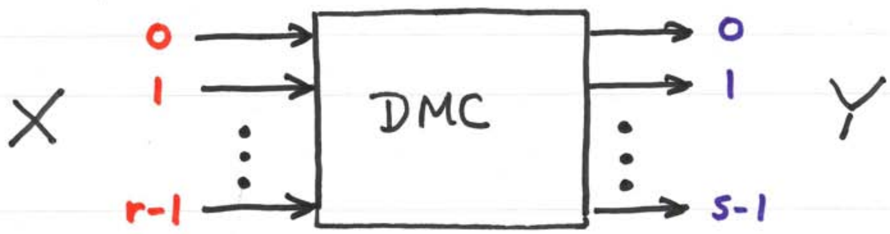
Note: Entropy as defined is the only function that satisfies properties 2, 4 and 5 above and is also continuous in $\underline{P} = (p_1, \ldots, p_n)$.

Proof: Outlined in ...
Khinchin, *Mathematical Foundations of Information Theory*, Dover 1959.

So how does entropy allow us to talk about the amount of information in a system?

Motivating Example: Discrete Memoryless Channel (DMC)

A DMC accepts one of r input symbols and outputs one of s output symbols



Discrete - Finitely many inputs and outputs

Memoryless - Current output depends only on current input, not the past.

n.b In general, $r \neq s$.

The mapping of the input to the output is stochastic and governed by a set of conditional transition probabilities.