Chapter 1: Source Coding

Gautam Singh

1 Conventions

1.1 X denotes a random variable.
1.2 $\mathcal{X}$ denotes the alphabet.
1.3 x denotes a particular value of the alphabet.
1.4 $p(x) \triangleq \Pr(X = x)$.

2 Uncertaininty and Information

2.1 **Self Information:** The self information of the event $X = x$ is defined as

$$I(x) \triangleq \log\left(\frac{1}{\log p(x)}\right) = -\log(p(x)) \tag{2.1}$$

Clearly, $I(x) = 0$ at $p(x) = 1$, that is, a high probability event conveys lesser information.

2.2 The units are determined by the base of the algorithm

2.2.1 If the base is 2, the units are **bits**.
2.2.2 If the base is e, the units are **nats**.
2.2.3 If the base is 10, the units are **dits**.

2.3 **Mutual Information:** The mutual information between x and y is defined as

$$I(x; y) \triangleq \log\left(\frac{p(x|y)}{p(x)}\right) \tag{2.2}$$

Observe that

$$I(x; y) = \log\left(\frac{p(x|y)}{p(x)}\right) \tag{2.3}$$
$$= \log\left(\frac{p(x|y)p(y)}{p(x)p(y)}\right) \tag{2.4}$$
$$= \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{2.5}$$
$$= \log\left(\frac{p(y|x)}{p(y)}\right) = I(y; x) \tag{2.6}$$

2.4 Can be interpreted as the information event $Y = y$ provides about $X = x$.

2.5 Equation (2.6) can be interpreted as follows

> The amount of information about $X = x$ provided by $Y = y$ is the same as the amount of information about $Y = y$ provided by $X = x$.

2.6 Notice that when X and Y are independent, then $I(x; y) = 0$ as $p(x, y) = p(x)p(y)$. Similarly, if $p(x|y) = 1$, then $I(x; y) = I(x)$.

2.7 **Conditional Self Information:** The conditional self information of the event $X = x$ given $Y = y$ is defined as

$$I(x|y) \triangleq \log\left(\frac{1}{p(x|y)}\right) = -\log p(x|y) \tag{2.7}$$

Notice that

$$I(x; y) = \log\left(\frac{p(x|y)}{p(x)}\right) = I(x) - I(x|y) \tag{2.8}$$

And thus mutual information can be positive, negative or zero.

3 Average Mutual Information and Entropy

3.1 **Average Mutual Information:** The average mutual information between random variables X and Y is defined as

$$I(X; Y) \triangleq \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y)T(x; y) \tag{3.1}$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{3.2}$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y|x) \log\left(\frac{p(y|x)}{p(x)}\right) \tag{3.3}$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y)p(x|y) \log\left(\frac{p(x|y)}{p(y)}\right) \tag{3.4}$$
$$= E\left[\log\left(\frac{p(X, Y)}{p(X)p(Y)}\right)\right] \tag{3.5}$$
$$= E\left[-\log\left(\frac{p(X)p(Y)}{p(X, Y)}\right)\right] \tag{3.6}$$

3.2 When X and Y are independent, (3.5) gives $I(X; Y) = 0$, that is, there is no average information between X and Y.

3.3 In general, $I(X; Y) \geq 0$ with equality iff X and Y are independent.

3.4 **Average Self Information/Entropy:** The average self information or entropy of a random variable X is defined as

$$H(X) \triangleq \sum_{x\in\mathcal{X}} p(x)I(x) \quad (3.7)$$
$$= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(x)}\right) \quad (3.8)$$
$$= E\left[\log\left(\frac{1}{p(X)}\right)\right] \quad (3.9)$$
$$= E\left[-\log p(X)\right] \quad (3.10)$$

3.5 Notice that since $0 \le p(x) \le 1$, (3.7) gives $H(X) \ge 0$.

3.6 The units of $I(X;Y)$ and $H(X)$ are **bits**.

3.7 For a Bernoulli trial with success rate p, the entropy of the outcome X is

$$H(X) = -\left(p\log_2 p + (1-p)\log_2(1-p)\right) \quad (3.11)$$

which is known as the *binary entropy function* and denoted by $h_2(p)$.

3.8 **Average Conditional Self Information/Conditional Entropy:** The average self information or conditional entropy of a random variable X given a random variable Y is defined as

$$H(X|Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x|y)} \quad (3.12)$$
$$= E\left[\log\frac{1}{p(X|Y)}\right] \quad (3.13)$$
$$= E\left[-\log p(X|Y)\right] \quad (3.14)$$

Clearly,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3.15)$$

3.9 Note that $I(X;Y) \ge 0 \implies H(X) \ge H(X|Y)$. Thus, conditioning can only decrease entropy. In case it does not, X and Y are independent.

3.10 **Joint Entropy:** The joint entropy of a pair of discrete random variables (X,Y) with a joint pmf $p(x,y)$ is defined as

$$H(X,Y) \triangleq \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \quad (3.16)$$
$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \quad (3.17)$$
$$= E\left[\log\frac{1}{p(X,Y)}\right] \quad (3.18)$$
$$= E\left[-\log p(X,Y)\right] \quad (3.19)$$

In general, the joint entropy of an n-tuple of random variables $(X_1, X_2, \ldots, X_n)$ with joint pmf $p(X_1, X_2, \ldots, X_n)$ is

$$H(X_1, X_2, \ldots, X_n) \triangleq E\left[-\log p(X_1, X_2, \ldots, X_n)\right] \quad (3.20)$$

3.11 From (3.10) and (3.14), we get the **chain rule**

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \quad (3.21)$$

In general for n random variables X_i, $1 \le i \le n$, the chain rule is

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, X_2, \ldots, X_{i-1}) \quad (3.22)$$

3.12 From (3.6), we clearly see

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (3.23)$$

4 Information Measures for Continuous Random Variables

4.1 **Average Mutual Information:** The average mutual information between two continuous random variables X and Y with joint pdf $p(x,y)$ and marginal pdfs $p(x)$ and $p(y)$ respectively is defined as

$$I(X;Y) \triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y|x)\log\frac{p(y|x)}{p(y)}\,dxdy \quad (4.1)$$

4.2 Note that while physical interpretation of mutual information can be applied here, such physical interpretations will not work with other quantities. This is because the information in a continuous random variable is infinite. Hence, differential entropy is defined.

4.3 **Differential Entropy:** The differential entropy of a continuous random variable X is defined as

$$h(X) \triangleq -\int_{-\infty}^{\infty} p(x) \log p(x) \, dx \tag{4.2}$$

While there is no physical meaning for this quantity, the units remain bits.

4.4 **Average Conditional Entropy:** The average conditional entropy of a continuous random variable X given Y is defined as

$$h(X|Y) \triangleq -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log p(x|y) \, dxdy \tag{4.3}$$

We can express the *average mutual information* as

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{4.4}$$

4.5 Rules for differential entropy:

4.5.1 The chain rule for differential entropy is given as

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}) \tag{4.5}$$

4.5.2 $h(X + c) = h(X)$, that is, transition does not alter differential entropy.

4.5.3 $h(aX) = h(X) + \log|a|$.

4.5.4 If X and Y are independent, then we have $h(X + Y) \geq h(X + Y|Y) = h(X|Y) = h(X)$.

5 Relative Entropy

5.1 **Relative Entropy or Kullback Leibler (KL) Distance:** The relative entropy of Kullback Leibler Distance between two pmfs $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{5.1}$$

$$= E\left[\log\left(\frac{p(X)}{q(X)}\right)\right] \tag{5.2}$$

It is sometimes denoted by $D_{KL}(p||q)$.

5.2 We can rewrite the mutual information in terms of relative entropy

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \tag{5.3}$$

5.3 **Jensen Shannon Distance:** The Jensen Shannon distance between two pmfs $p(x)$ and $q(x)$ is defined as

$$JSD(p||q) \triangleq \frac{1}{2}\left(D(p||m) + D(q||m)\right) \tag{5.4}$$

It is sometimes denoted by $D_{JS}(p||q)$, and referred to as **Jensen Shannon Divergence** or **Information Radius**.

5.4 **Convex Function:** A function f defined on $[0,1]$ is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{5.5}$$

for all $\lambda \in [0,1]$. On the other hand, if f is **concave**, then the inequality in (5.5) becomes $\geq$. Strict inequalities would make f **strictly convex** or **strictly concave**.

5.5 $D(p||q)$ is convex in the pair (p,q) and $H(p)$ is concave in p.