Conditional Entropy

Let Y be a discrete random variable with outcomes, $\{y_1,\ldots,y_m\}$, which occur with probabilities, $p_Y(y_j)$. The avg. information you gain when told the outcome of Y is:

$$H_Y = -\sum_{j=1}^{m} p_Y(y_j)\log p_Y(y_j).$$

Conditional Entropy (contd.)

Let X be a discrete random variable with outcomes, $\{x_1,\ldots,x_n\}$, which occur with probabilities, $p_X(x_i)$. Consider the 1D distribution,

$$p_{Y|X=x_i}(y_j) = p_{Y|X}(y_j\,|\,x_i)$$

i.e., the distribution of Y outcomes given that $X = x_i$. The avg. information you gain when told the outcome of Y is:

$$H_{Y|X=x_i} = -\sum_{j=1}^{m} p_{Y|X}(y_j\,|\,x_i)\log p_{Y|X}(y_j\,|\,x_i).$$

Conditional Entropy (contd.)

The *conditional entropy* is the expected value for the entropy of $p_{Y|X=x_i}$:

$$H_{Y|X} = \left\langle H_{Y|X=x_i} \right\rangle .$$

It follows that:

$$\begin{aligned} H_{Y|X} &= \sum_{i=1}^{n} p_X(x_i) H_{Y|X=x_i} \\ &= \sum_{i=1}^{n} p_X(x_i) \left(- \sum_{j=1}^{m} p_{Y|X}(y_j \,|\, x_i) \log p_{Y|X}(y_j \,|\, x_i) \right) \\ &= - \sum_{i=1}^{n} \sum_{j=1}^{m} p_X(x_i) p_{Y|X}(y_j \,|\, x_i) \log p_{Y|X}(y_j \,|\, x_i). \end{aligned}$$

The entropy, $H_{Y|X}$, of the conditional distribution, $p_{Y|X}$, is therefore:

$$H_{Y|X} = - \sum_{i=1}^{n} \sum_{j=1}^{m} p_{XY}(x_i, y_j) \log p_{Y|X}(y_j \,|\, x_i).$$

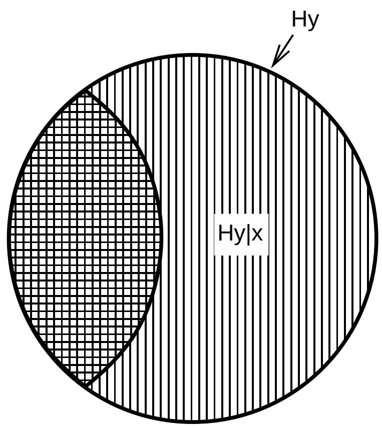

Figure 1: There is less information in the conditional than in the marginal (Theorem 1.2).

Theorem 1.2

There is less information in the conditional, $p_{Y|X}$, than in the marginal, p_Y:

$$H_{Y|X} - H_Y \leq 0.$$

Proof:

$$H_{Y|X} - H_Y =$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j) \log p_{Y|X}(y_j \,|\, x_i)$$

$$+\sum_{j=1}^{m} p_Y(y_j) \log p_Y(y_j).$$

Theorem 1.2 (contd.)

$$H_{Y|X} - H_Y =$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j) \log p_{Y|X}(y_j \,|\, x_i)$$

$$+\sum_{j=1}^{m}\left(\sum_{i=1}^{n} p_{XY}(x_i, y_j)\right) \log p_Y(y_j).$$

$$H_{Y|X} - H_Y =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j) \log\left(\frac{p_Y(y_j)}{p_{Y|X}(y_j \,|\, x_i)}\right).$$

Theorem 1.2 (contd.)

Using the inequality, $\log a \leq (a-1)\log e$, it follows that:

$$H_{Y|X} - H_Y \leq$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\left(\frac{p_Y(y_j)}{p_{Y|X}(y_j\,|\,x_i)} - 1\right)\log e.$$

$$H_{Y|X} - H_Y \leq$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\left(\frac{p_Y(y_j)}{p_{Y|X}(y_j\,|\,x_i)}\right)\log e$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log e.$$

Theorem 1.2 (contd.)

$$H_{Y|X} - H_Y \leq$$
$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_X(x_i) p_{Y|X}(y_j \,|\, x_i) \left(\frac{p_Y(y_j)}{p_{Y|X}(y_j \,|\, x_i)} \right) \log e - \log e.$$

$$\begin{aligned}
H_{Y|X} - H_Y &\leq \sum_{i=1}^{n} \sum_{j=1}^{m} p_X(x_i) p_Y(y_j) \log e - \log e \\
&\leq \left[\sum_{i=1}^{n} p_X(x_i) \left(\sum_{j=1}^{m} p_Y(y_j) \right) \right] \log e - \log e \\
&\leq \log e - \log e \\
&\leq 0.
\end{aligned}$$

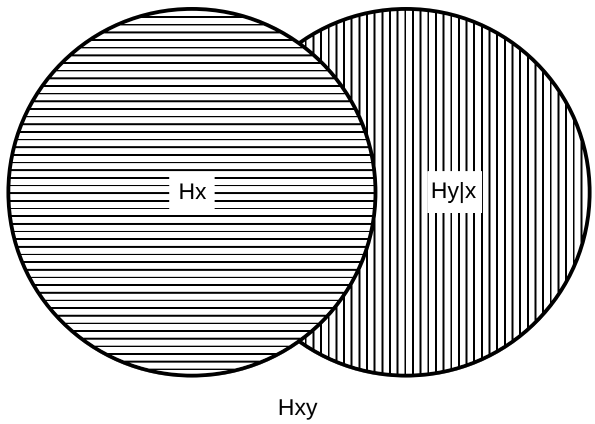

Figure 2: The information in the joint is the sum of the information in the conditional and the marginal (Theorem 1.3).

Theorem 1.3

The information in the joint, p_{XY}, is the sum of the information in the conditional, $p_{Y|X}$, and the marginal, p_X:

$$H_{XY} = H_{Y|X} + H_X.$$

Proof:

$$\begin{aligned}
H_{XY} &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log p_{XY}(x_i, y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log\left[p_X(x_i)\, p_{Y|X}(y_j \,|\, x_i)\right] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\left[\log p_X(x_i) + \log p_{Y|X}(y_j \,|\, x_i)\right].
\end{aligned}$$

Theorem 1.3 (contd.)

$$H_{XY} = -\sum_{i=1}^{n}\left(\sum_{j=1}^{m} p_{XY}(x_i, y_j)\right)\log p_X(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log p_{Y|X}(y_j \,|\, x_i) = H_X + H_{Y|X}.$$

Mutual Information

The *mutual information*, I_{XY}, between X and Y is defined to be:

$$I_{XY} = H_Y - H_{Y|X} = I_{YX} = H_X - H_{X|Y}.$$

The mutual information is a measure of the statistical independence of two random variables.

Mutual Information (contd.)

$$I_{XY} = H_Y - H_{Y|X} = -\sum_{j=1}^{m} p_Y(y_j) \log p_Y(y_j)$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{m} p_{XY}(x_i, y_j) \log p_{Y|X}(y_j \mid x_i).$$

$$I_{XY} = -\sum_{j=1}^{m} \left(\sum_{i=1}^{n} p_{XY}(x_i, y_j) \right) \log p_Y(y_j)$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{m} p_{XY}(x_i, y_j) \log p_{Y|X}(y_j \mid x_i).$$

Mutual Information (contd.)

$$
\begin{aligned}
I_{XY} &= \sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log\left(\frac{p_{Y|X}(y_j \mid x_i)}{p_Y(y_j)}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(x_i, y_j)\log\left(\frac{p_{XY}(x_i, y_j)}{p_X(x_i)p_Y(y_j)}\right) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_{YX}(y_j, x_i)\log\left(\frac{p_{YX}(y_j, x_i)}{p_Y(y_j)p_X(x_i)}\right) \\
&= I_{YX} = H_X - H_{X|Y}.
\end{aligned}
$$

Four Cases

- X and Y are statistically independent:

$$H_{XY} = H_X + H_Y$$

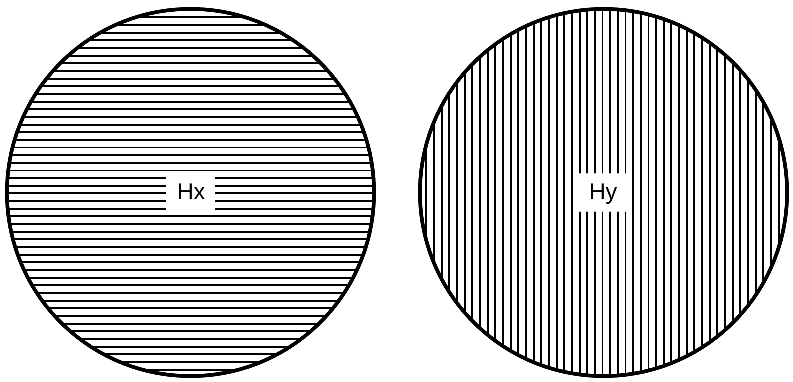

Figure 3: X and Y are statistically independent.

Four Cases (contd.)

- X is completely dependent on Y:

$$H_{XY} = H_Y$$

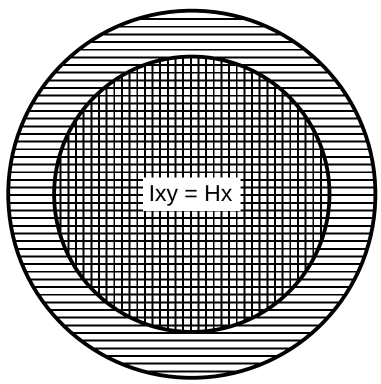

Figure 4: X is a function of Y.

Four Cases (contd.)

- Y is completely dependent on X:

$$H_{XY} = H_X$$

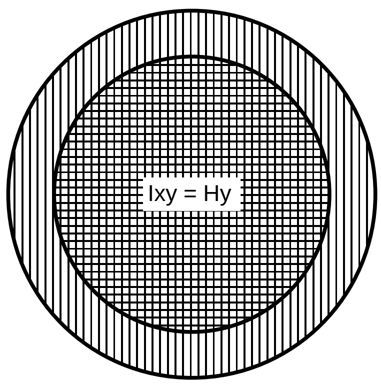

Figure 5: Y is a function of X.

Four Cases (contd.)

- X and Y are not independent but neither is completely dependent on the other:

$$H_{XY} = H_X + H_Y - I_{XY}$$

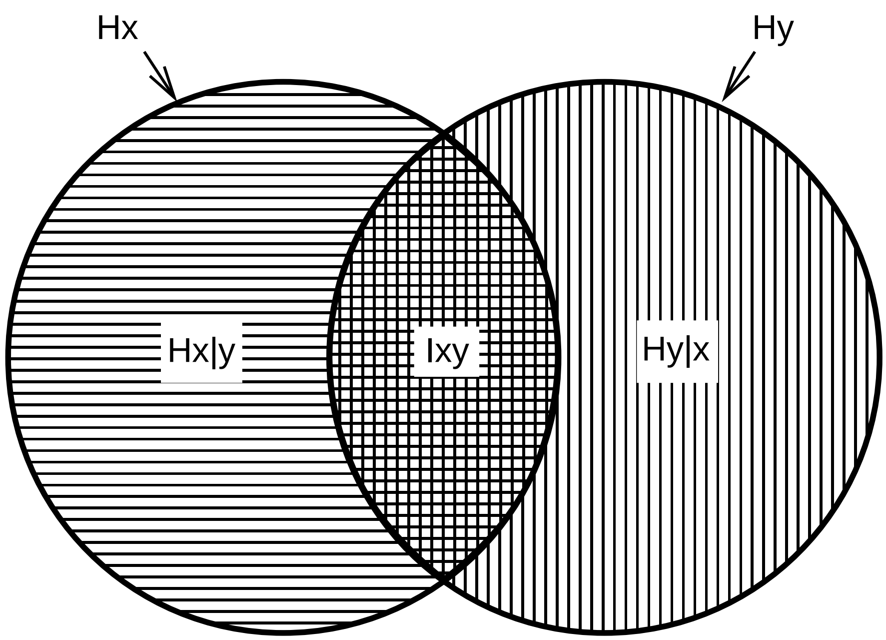

Figure 6: The general case.

Informational Channel

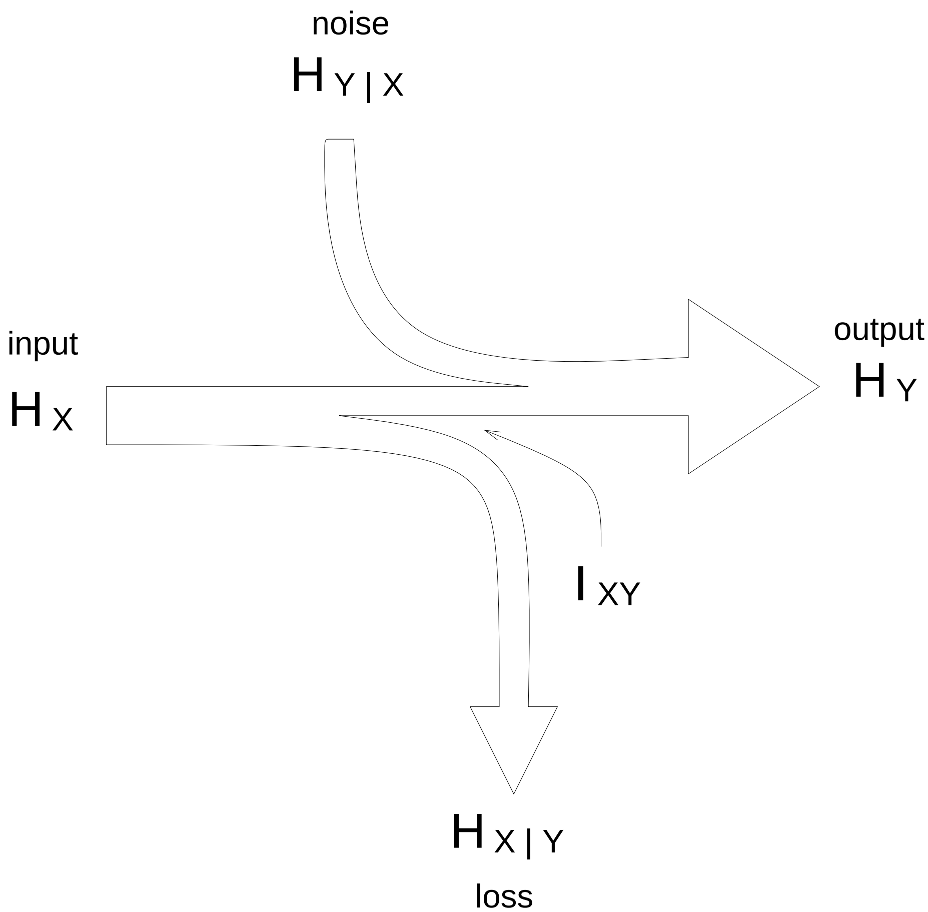

Figure 7: Information channel.

Kullback-Liebler Distance

Let p_X and q_X be probability mass functions for two discrete r.v.'s over the same set of events, $\{x_1, \ldots, x_N\}$. The *Kullback-Liebler distance* (or *KL divergence*), is defined as follows:

$$KL(p_X||q_X) = \sum_{i=1}^{N} p_X(x_i) \log\left(\frac{p_X(x_i)}{q_X(x_i)}\right).$$

The KL divergence is a measure of how different two probability distributions are. Note that in general

$$KL(p_X||q_X) \neq KL(q_X||p_X).$$

Kullback-Liebler Distance (contd.)

Let p_{XY} be the joint p.m.f. for discrete r.v.'s X and Y and let p_X and p_Y be the corresponding marginal distributions:

$$p_X(x_i) = \sum_{j=1}^{M} p_{XY}(x_i, y_j)$$
$$p_Y(y_j) = \sum_{i=1}^{N} p_{XY}(x_i, y_j).$$

We observe that

$$I_{XY} = KL(p_{XY} || q_{XY})$$

where $q_{XY}(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j)$. In other words, mutual information is the KL divergence between a joint distribution, p_{XY}, and the product of its marginal distributions, p_X and p_Y.