1 Entropy

Let X be a discrete random variable on a space $\mathcal{X}$ with $|\mathcal{X}| = k < \infty$. We define the entropy of a random variable as

$$H(X) \equiv -\sum_{i=1}^{k} P(x_i) \log P(x_i) \tag{1.1}$$

a) $H(X) \geq 0$

Proof. Suppose $P(X)$ is a probability distribution over the finite sample space $\mathcal{X}$ of size $|\mathcal{X}| = k < \infty$. By definition, it is a probability measure define on some σ-algebra satisfying the non-negativiy property

$$P(X) \geq 0 \,\forall\, X \in \mathcal{X},$$

and the normalization condition

$$P(\mathcal{X}) = \sum_{i=1}^{k} P(x_i) = 1$$

This second condition can be rephrased as the second Kolmogorov axiom:

$$\sup_{X \in \mathcal{X}} P(X) \leq 1$$

Therefore, we have 3 cases to check to estimate the image of $-P(X)\log P(X)$:

1. $0 < P(X) < 0 \implies -P(X)\log P(X) > 0$.
2. Since $\lim_{x \to 0^+} x \log x = 0$ by l'Hôpital's rule, then $P(X) = 0 \implies -P(X)\log P(X) = 0$.
3. $P(X) = 1 \implies -P(X)\log P(X) = 0$

Since all terms are strictly non-negative, then $H(X) \geq 0$. □

We show that $\lim_{x\to 0^+} x \log x = 0$:

$$\begin{aligned}
\lim_{x\to 0^+} x \log x &= \lim_{x\to 0^+} \frac{\log x}{1/x} \\
&\overset{\text{l'H}}{=} \lim_{x\to 0^+} \frac{1/x}{-1/x^2} \\
&= -\lim_{x\to 0^+} x = 0
\end{aligned}$$

b) Relation between $D(p||U)$ and $H(X)$

Suppose that p is some pmf of X and q is the uniform distribution over $\mathcal{X}$. By definition of the KL divergence,

$$D(p||q) = \sum_{i=1}^{k} p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

Since $q(x_i) = \dfrac{1}{|\mathcal{X}|}\ \forall x_i \in \mathcal{X}$,

$$D(p||q) = \sum_{i=1}^{k} p(x_i) \log p(x_i) + \log|\mathcal{X}| \underbrace{\sum_{i=1}^{k} p(x_i)}_{=1}$$

Which we can rewrite as

$$\boxed{D(p||q) = -H(X) + \log|\mathcal{X}|}$$

c) Upper bound on $H(X)$

Since the Kullback-Leibler divergence is strictly non-negative, then the last result impose an upper bound on the entropy:

$$\boxed{H(X) \leq \log k}$$

The distribution that maximizes the entropy is the uniform distribution:

$$H_U(X) = -\sum_{i}^{k} \frac{1}{k} \log \frac{1}{k} = \log k$$

2 Mutual information

We consider a pair of discrete random variables (X_1, X_2) defined over a finite set $\mathcal{X}_1 \times \mathcal{X}_2$. We denote the joint distribution $p_{1,2}$, and the respective marginal distributions p_1 and p_2.

The mutual information is defined as

$$I(X_1, X_2) \equiv \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2) \log \frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)} \tag{2.1}$$

a) $I(X_1, X_2) \geq 0$

Proof. We rewrite the mutual information in term of the expectation operator over X_1 and X_2 and rearrange the sum over all pairs (x_1, x_2):

$$\begin{aligned}
I(X_1, X_2) &= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} p_{1,2}(x_1,x_2)\log\frac{p_{1,2}(x_1,x_2)}{p_1(x_1,x_2)} \\
&= \sum_{x_1\in\mathcal{X}_2} p_1(x_1) \sum_{x_2\in\mathcal{X}_2} p_2(x_2)\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\log\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)} \\
&= \mathbb{E}_{p_1}\left[\mathbb{E}_{p_2}\left[\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\log\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right]\right]
\end{aligned}$$

The function $z\log z, z \geq 0$ is convex. The Jensen's inequality for a convex function states that

$$\varphi\left(\mathbb{E}\left[X\right]\right) \leq \mathbb{E}\left[\varphi(X)\right]$$

Therefore,

$$\mathbb{E}\left[Z\right]\log\mathbb{E}\left[Z\right] \leq \mathbb{E}\left[Z\log Z\right]$$

This allow us to simplify the expression for I:

$$\begin{aligned}
I(X_1, X_2) &\geq \mathbb{E}_{p_1}\left[\mathbb{E}_{p_2}\left[\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right]\log\mathbb{E}_{p_2}\left[\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right]\right] \\
&\geq \mathbb{E}_{p_1}\left[\mathbb{E}_{p_2}\left[\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right]\right]\log\mathbb{E}_{p_1}\left[\mathbb{E}_{p_2}\left[\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right]\right]
\end{aligned}$$

We now recover the definition of the expectations,

$$\begin{aligned}
I(X_1, X_2) &\geq \left(\sum_{x_1\in\mathcal{X}_1} p_1(x_1)\sum_{x_2\in\mathcal{X}_2} p_2(x_2)\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right)\log\left(\sum_{x_1\in\mathcal{X}_1} p_1(x_1)\sum_{x_2\in\mathcal{X}_2} p_2(x_2)\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right) \\
&\geq \underbrace{\left(\sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2)\right)}_{=1}\log\underbrace{\left(\sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2)\right)}_{=1} \\
&\geq 0
\end{aligned}$$

By definition of a probability measure of the joint.

□

b) Mutual Information in Term of Entropy

We let $Z = (X_1, X_2)$ be a random variable. The mutual information can therefore be expressed as

$$\begin{aligned}
I(X_1, X_2) &= \sum_z p_{1,2}(z)\log\frac{p_{1,2}(z)}{p_1(x_1)p_2(x_2)} \\
&= \sum_z p_{1,2}(z)\big(\log p_{1,2}(z) - \log p_1(x_1) - \log p_2(x_2)\big) \\
&= \sum_z p_{1,2}(z)\log p_{1,2}(z) - \sum_{x_1,x_2} p_{1,2}(x_1,x_2)\log p_1(x_1) - \sum_{x_1,x_2} p_{1,2}(x_1,x_2)\log p_2(x_2) \\
&= -H(Z) - \sum_{x_1}\log p_1(x_1)\sum_{x_2} p_{1,2}(x_1,x_2) - \sum_{x_2}\log p_2(x_2)\sum_{x_1} p_{1,2}(x_1,x_2)
\end{aligned}$$

Since the single variable sum over the joint is just the marginal,

$$I(X_1, X_2) = -H(Z) - \sum_{x_1} p_1(x) \log p_1(x_1) - \sum_{x_2} p_2(x_2) \log p_2(x_2)$$
$$\implies \boxed{I(X_1, X_2) = -H(Z) + H(X_1) + H(X_2)}$$

c) Joint of Maximal Entropy

Since the mutual information cannot be negative, we get an upper bound for the joint entropy

$$H(Z) \leq H(X_1) + H(X_1) \leq \log |\mathcal{X}_1| + \log |\mathcal{X}_2|$$

The maximal entropy joint in general is thus

$$\max H(Z) = H(X_1) + H(X_2)$$

This happens only when X_1 and X_2 are marginally independent such that the joint can be factored in term of the marginals:

$$p_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2),\ \forall (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$$

Indeed,

$$\begin{aligned}
H(Z) &= -\sum_{(x_1,x_2)\in\mathcal{X}_2\times\mathcal{X}_2} p_1(x_1)p_2(x_2)\big(\log p_1(x_1) + \log p_2(x_2)\big) \\
&= -\sum_{x_1\in\mathcal{X}_1} p_1(x_1)\log p_1(x_2) \underbrace{\sum_{x_2\in\mathcal{X}_2} p_2(x_2)}_{=1} - \sum_{x_2\in\mathcal{X}_2} p_2(x_2)\log p_2(x_2) \underbrace{\sum_{x_1\in\mathcal{X}_1} p_1(x_1)}_{=1} \\
&= H(X_1) + H(X_2)
\end{aligned}$$