Discrete probability space

The alphabet is $\mathcal{X}$ ($\mathcal{X}$ is discrete).

Random variable X takes its values in $\mathcal{X}$

Probability distribution $p_X(x) = \mathbf{Prob}(X = x), x \in \mathcal{X}$. Generally, this quantity is simply denoted by $p(x)$.

Expectation of the real random variable $V(X)$ (V is a function from $\mathcal{X}$ to $\mathbb{R}$)

$$\mathbb{E}(V) = \sum_{x \in \mathcal{X}} p(x) V(x)$$

Joint Probability Distribution

Alphabet $\mathcal{X} \times \mathcal{Y}$ with probability distribution $p(x, y)$.

Random variables X and Y taking their values in $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Marginals	Conditional probability distribution
$\mathbf{Prob}(X = x) = p_X(x) = \sum_y p(x, y)$	$\mathbf{Prob}[X = x \mid Y = y] = \frac{p(x, y)}{p_Y(y)}$
$\mathbf{Prob}(Y = y) = p_Y(y) = \sum_x p(x, y)$	$\mathbf{Prob}[Y = y \mid X = x] = \frac{p(x, y)}{p_X(x)}$

In general we use the simplified notation $p(x), p(y), p(x|y), p(y|x)$, for respectively $\mathbf{Prob}(X = x)$, $\mathbf{Prob}(Y = y)$, $\mathbf{Prob}(X = x|Y = y)$ and $\mathbf{Prob}(Y = y|X = x)$.
X and Y are independent iff

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, p(x, y) = p(x)p(y)$$

Entropy – Properties

Definition[Entropy]

$$H(X) \stackrel{\mathrm{def}}{=} -\sum_{x} p(x) \log_2 p(x)$$

The entropy of a Bernoulli distribution ($X = 0$ with probability p, $X = 1$ with probability $1 - p$) is equal to

$$h(p) = -p \log p - (1 - p) \log(1 - p)$$

Conditional Entropy, Mutual Information

Definition[entropy of a pair of random variables]

$$H(X,Y) \stackrel{\text{def}}{=} -\sum_{x,y} p(x,y) \log_2 p(x,y)$$

Definition[conditional entropy]

$$\begin{aligned} H(X|Y=y) &\stackrel{\text{def}}{=} -\sum_{x} p(X=x|Y=y) \log_2 p(X=x|Y=y) \\ H(X|Y) &\stackrel{\text{def}}{=} \sum_{y} H(X|Y=y)p(y) \end{aligned}$$

Definition[mutual information]

$$I(X;Y) = H(X) - H(X|Y)$$

Properties

Theorem 1. *1.* $I(X;Y) = \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

2. $I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) = I(Y;X)$.

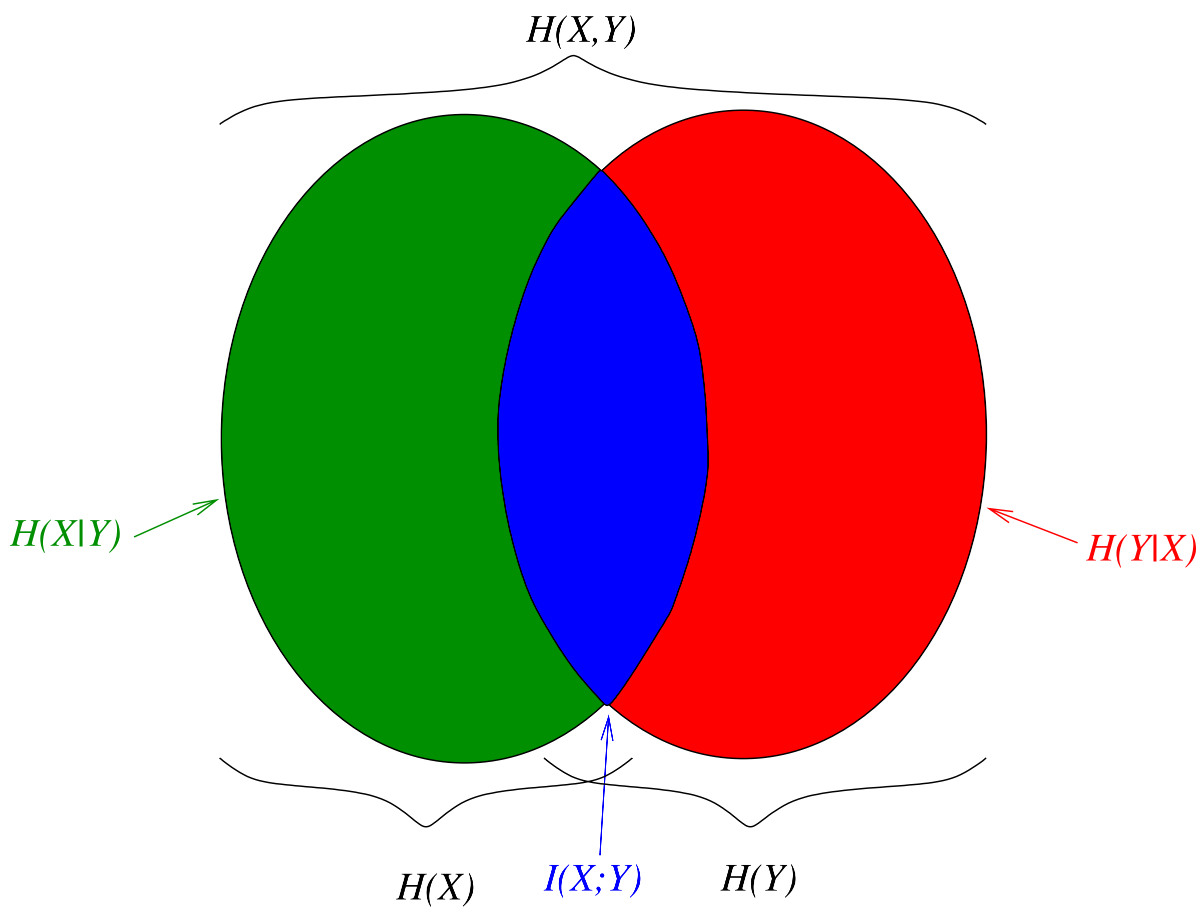
3. $I(X;Y) \geq 0$

4. $I(X;Y) = 0$ *iff* X *et* Y *are independent.*

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

6. $H(X) \leq \log |\mathcal{X}|$ *for* X *taking its values in* $\mathcal{X}$.

$$I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) = I(Y;X)$$



Proof

1.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= -\sum_{x} p(x)\log p(x) + \sum_{y} p(y) \sum_{x} p(x|y)\log p(x|y) \\
&= -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(x,y)\log p(x|y) \\
&= \sum_{x,y} p(x,y)\log \frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)}.
\end{aligned}
$$

2. By symmetry $I(X;Y) = I(Y;X) = H(Y) - H(Y|X)$.

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y) \\ &= -H(X,Y) + H(X) + H(Y). \end{aligned}$$

3. $I(X;Y) = D(p(x,y)||p(x)p(y))$.

4. $D(p(x,y)||p(x)p(y)) = 0 \Rightarrow p(x,y) = p(x)p(y)$.

5. $H(X) - H(X|Y) = I(X;Y)$.

6.

$$\begin{aligned}
-H(X) + \log|\mathcal{X}| &= \sum_x p(x)\log p(x) + \sum p(x)\log|\mathcal{X}| \\
&= \sum_x p(x)\log\frac{p(x)}{\frac{1}{|\mathcal{X}|}} \\
&= D(p||u)
\end{aligned}$$

Kullback Divergence

Definition[Kullback Divergence] For two probability distributions p and q over a same alphabet $\mathcal{X}$:

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

Theorem 2.

$$D(p||q) \geq 0$$

with equality iff $p = q$.

Proof

$$
\begin{aligned}
D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in \mathcal{X}} p(x) \left(-\log \frac{q(x)}{p(x)} \right) \\
&\geq -\log \left(\sum_{x} p(x) \frac{q(x)}{p(x)} \right) \\
&= 0.
\end{aligned}
$$

Entropy of an n-tuple of r.v.

Theorem 3. [Chain Rule for entropy]

$$H(X_1X_2\ldots X_N) = H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X_1\ldots X_{N-1})$$

Corollary 1.

$$H(X_1X_2\ldots X_N) \leq H(X_1) + H(X_2) + \cdots + H(X_N)$$

with equality iff the X_i's are independent.

Estimating a r.v. X with another r.v. Y

Theorem 4. [Fano's lemma] *Let*

- X *and* Y *be two random variables*
- X *taking its values in an alphabet of size* a
- $\hat{X}$ *an estimator for* X *computed from the knowledge of* Y
- P_e *the error probability of the estimator, that is* $P_e = p(\hat{X} \neq X)$.

Then

$$h(P_e) + P_e \log_2(a-1) \geq H(X|Y).$$