Australian National University

COMP2610/6261
Tut 6 Summary

Entropy and its properties

A Measure of Information is Entropy

Entropy: *Average* amount of information in a random variable X with distribution $p(x)$ over alphabet $\mathcal{X}$, defined as

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

$$= -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

Properties of Entropy

- Entropy is non-negative. $H(X) \geq 0$ because
  - $p(x) \geq 0$
  - $\log \frac{1}{p(x)} \geq 0$
- $H(X) = 0$ means X is not random any more, but a sure event.
- Entropy only depends on the probability distribution $p(x)$ and not the alphabet $\mathcal{X}$. So as far as entropy is concerned, we can assume $\mathcal{X} = \{1, 2, \cdots, m\}$ for some integer $m \in \mathbb{N}$.

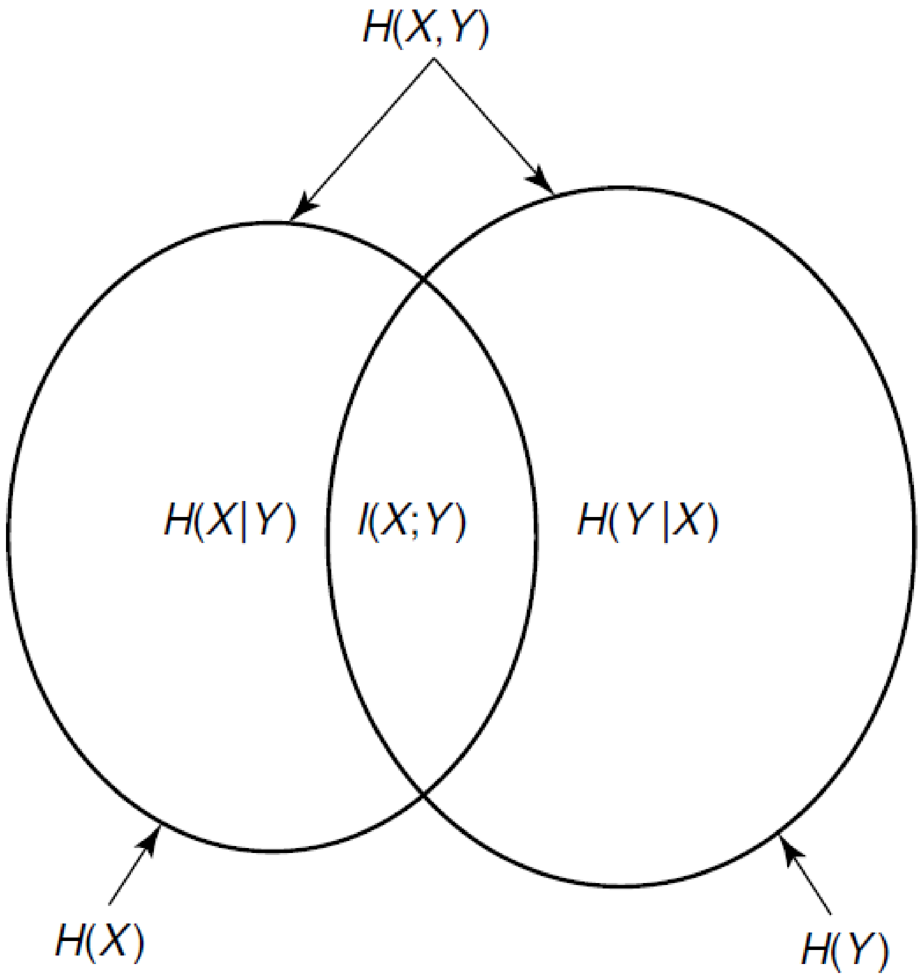
Joint and Conditional Entropy, Visualisation

$$H(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{1}{p(x,y)}$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$$

$$= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$



Entropy Chain Rule

$$H(Y|X) = H(X,Y) - H(X)$$

$$H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y)$$

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

Independent Variables

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)(\log p(y))$$

$$= -\sum_{y\in\mathcal{Y}} p(y)(\log p(y)) \underbrace{\sum_{x\in\mathcal{X}} p(x)}_{=1} = H(Y)$$

$$H(X_1,\cdots,X_n) = \sum_{i=1}^{n} H(X_i)$$

Very Important Entropy Relations

$$H(X,Y) \leq H(X) + H(Y)$$

And

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

Mutual Information Definition

- Mutual Information between two random variables X and Y is denoted by $I(X;Y)$
- It is the amount of information revealed (or amount of uncertainty resolved) about X after observing or knowing Y

$$
\begin{aligned}
I(X;Y) = H(X) - H(X|Y) &= H(Y) - H(Y|X) \\
&= I(Y;X)
\end{aligned}
$$

Mutual Information Properties

- 1. Chain Rule

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)$$

$$I(X_1, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1})$$

- 2. Symmetric and Positive

$$I(X; Y) = I(Y; X)$$

$$I(X; Y) = I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y) \geq 0$$

- 3. Independent

$$I(X; Y) = I(Y; X) = H(Y) - H(Y|X) = H(Y) - H(Y) = 0$$

- 4. Y is a function of X, H(Y|X)=0

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

Relative Entropy and properties

- Relative Entropy: A measure of distance between two probability distributions p and q
- Definition:

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

$$= -H(p) + \sum p(x) \log \frac{1}{q(x)}$$

- Note that $D(p||q) \neq D(q||p)$.
- Also note that if $p(x) = q(x), \forall x$ then $D(p||q) = 0$ ($\log 1 = 0$).

Inequalities

- Markov inequality.

$$p(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}.$$

- Chebyshev's inequality.

$$p(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\mathbb{V}[X]}{\lambda^2}.$$

Typical Set

- Empirical entropy is defined as

$$\tilde{H}(\mathbf{x}) = -\frac{1}{n}\log P(\mathbf{x}) = -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n)$$

for a sequence of i.i.d random variables drawn from $p(x)$.

- That is:

$$-\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) = -\frac{1}{n}\log \prod_{i=1}^{n} p(x_i) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)$$

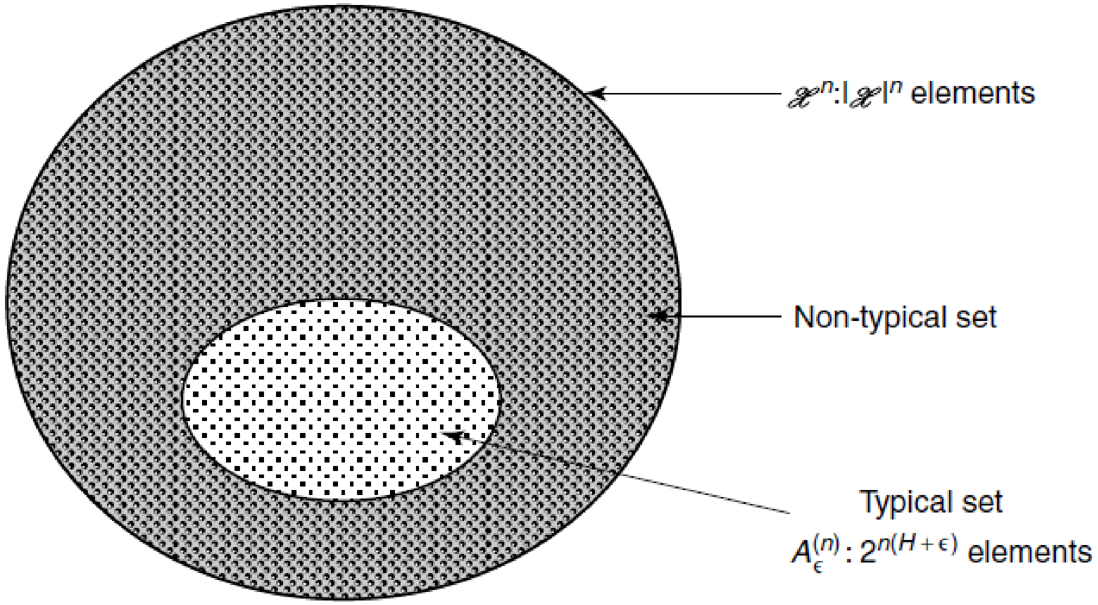
- Based on AEP we can divide the set of all sequences into two sets, the **typical set**, where empirical entropy $\tilde{H}(\mathbf{x})$ is *close enough* to the true entropy $H(X)$, and the **nontypical set**, which contains all other sequences.

- In other words, a sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ belongs to $A_\epsilon^{(n)}$ if it satisfies

$$\left|\tilde{H}(\mathbf{x}) - H(X)\right| \leq \epsilon$$

$$n(H(X) - \epsilon) \leq -\log p(x_1, x_2, \cdots, x_n) \leq n(H(X) + \epsilon)$$

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$



Australian National University

Important Properties of Typical Set

1. If $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$ then

$$n(H(X) - \epsilon) \leq -\log p(x_1, x_2, \cdots, x_n) \leq n(H(X) + \epsilon)$$

2. $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for sufficiently large n.

3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in A.

4. $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$ for sufficiently large n.