COMP2610/6261
Tut 12 Summary

The Bayesian Inference Framework

Bayesian Inference

Bayesian inference provides a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\underbrace{p(Z|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Z)}^{\text{likelihood}} \times \overbrace{p(Z)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

Conditional independence

Definition: Independent Variables

Two variables X and Y are statistically independent, denoted $X \perp\!\!\!\perp Y$, if and only if their joint distribution *factorizes* into the product of their marginals:

$$X \perp\!\!\!\perp Y \leftrightarrow p(X, Y) = p(X)p(Y)$$

Moments for functions of two discrete Random Variables

$$E(X) = \sum x\, p(X = x)$$

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2.$$

$$E(XY) = \sum \sum xy\, p(X = x, Y = y)$$

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 $m \in \mathbb{N}$.

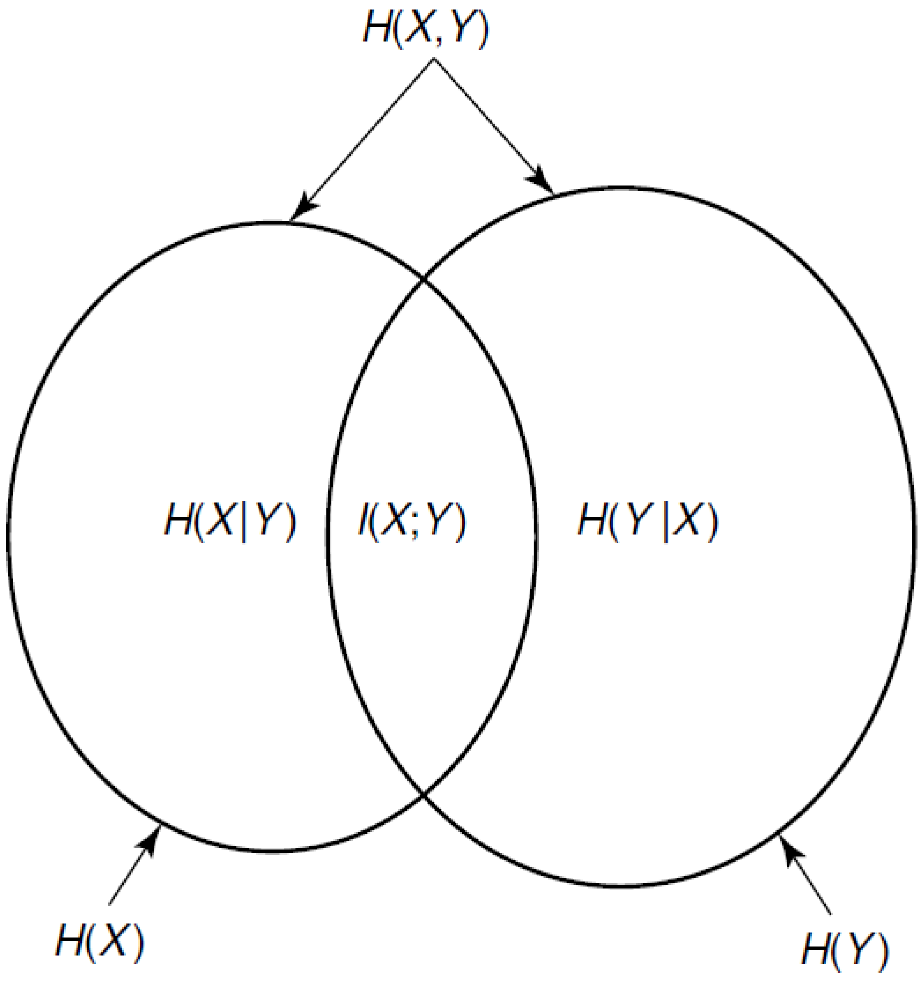
Joint and Conditional Entropy, Visualisation

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

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

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

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



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

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

Australian National University

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 $D(p||q) = 0$ ($\log 1 = 0$).

Markov Chain

- A discrete stochastic process $X_1, X_2, \cdots$ is said to be a Markov chain or a Markov process if for all $n = 1, 2, \cdots$

$$Pr\{(X_{n+1} = x_{n+1} | X_n = x_n, \cdots, X_1 = x_1)\}$$
$$= Pr\{(X_{n+1} = x_{n+1} | X_n = x_n)\}$$

- That is, the process only depends on the immediate past.

Markov Chain

- In general

$$p(x, y, z) = p(x)p(y|x)p(z|x, y)$$

If

$$p(z|x, y) = p(z|y) \Rightarrow p(x, y, z) = p(x)p(y|x)p(z|y)$$

then

Then we say X, Y, Z form a Markov Chain and denote it like $X \rightarrow Y \rightarrow Z$.

Markov Chain Consequence

- Consequence 1:
  - ▸ $X \to Y \to Z$ if and only if (iff) X and Z are independent **given** Y.
- Consequence 2:
  - ▸ $X \to Y \to Z$ iff $Z \to Y \to X$.
  - ❖ Markov Chain is SYMMETRIC
- Consequence 3:
  - ▸ If $Z = f(Y)$, then $X \to Y \to Z$.
  - ❖ NOT a necessary, but a sufficient, condition

Data Processing Inequality

- Data Processing Inequality 1:
  - If $X \to Y \to Z$ then
  $$I(X;Y) \geq I(X;Z)$$
- Data Processing Inequality 2:
  - If $X \to Y \to f(Y)$ then
  $$I(X;Y) \geq I(X;f(Y))$$
- Corollary of Data Processing Inequality:
  - If $X \to Y \to Z$ then
  $$I(X;Y|Z) \leq I(X;Y)$$

Inequalities

- Markov inequality.

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

- Chebyshev's inequality.

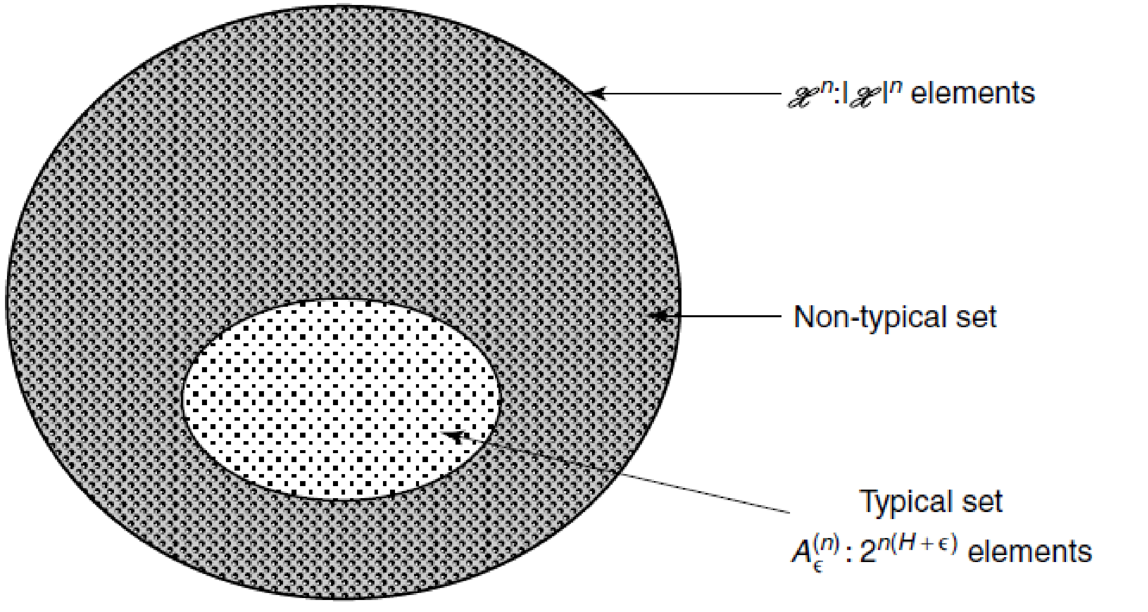
$$p(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\mathbb{V}[X]}{\lambda^2}.$$

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

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

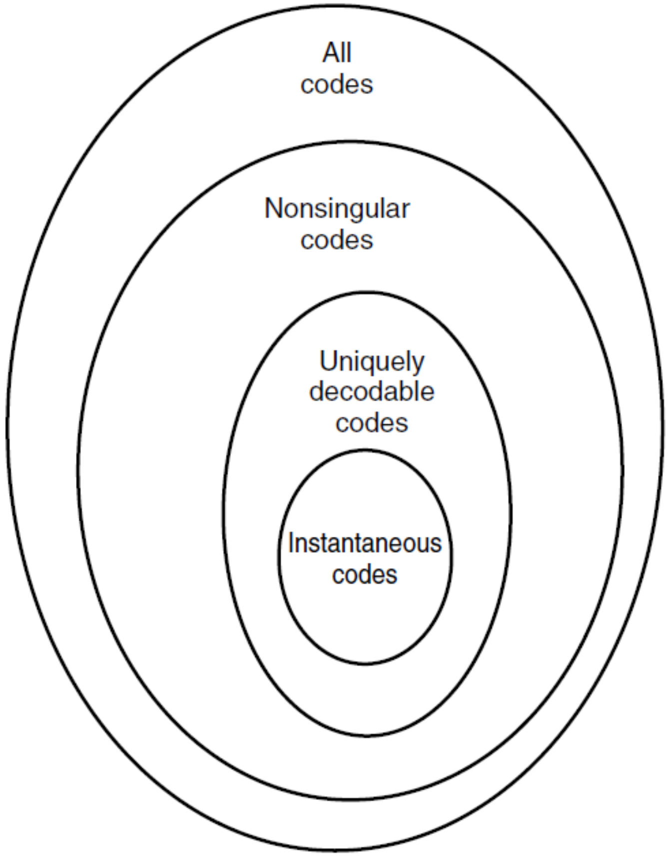
$$n(H(X) - \epsilon) \leq -\log p(x_1, x_2, \cdots, x_n) \leq n(H(X) + \epsilon)$$

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



- A source code C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite-length strings of symbols from a D-ary alphabet (If $D = 2$, then code is binary).

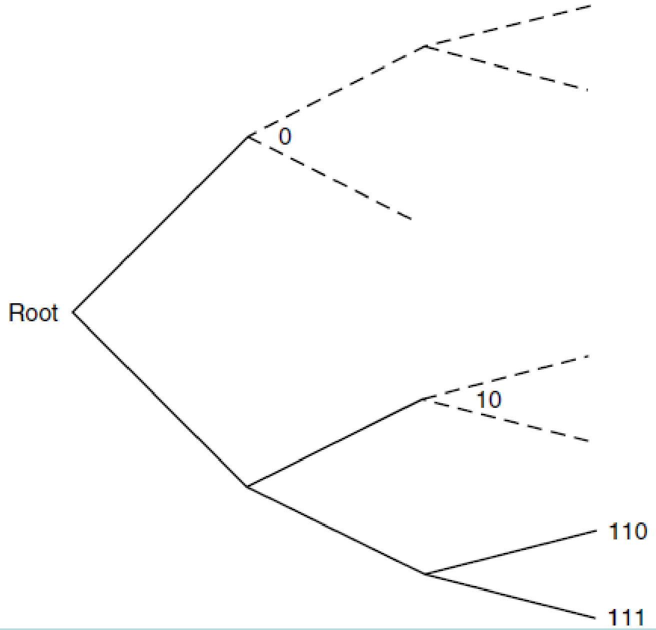
$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$$



Example of Source Code Types

x	Sing.	Non-sing. but not UD	UD but not prefix	Prefix
1	0	0	0	0
2	0	1	01	10
3	1	00	011	110
4	10	11	0111	111

Code Tree and Kraft Inequality



- Branch: each node can have D branches labelled $0, \cdots, D-1$. For binary code, $D = 2$, and each node has two branches
- Leaf: The last node of a branch (no more codes along this branch to ensure prefix-free condition).
- Code: Read along the root node to the leaf.

Kraft Inequality and its Converse

For any instantaneous code (prefix code) over an alphabet of size D, the codeword lengths $l_1, l_2, \cdots, l_m$ must satisfy the inequality

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these codeword lengths.

Minimum Code Length

A prefix code C is optimal if the average code length $L(C)$ is as small as possible.

The optimal code length satisfies

$$L(C) \geq H_D(X)$$

where $H_D(X)$ is entropy in *logarithm base D*.

The equality is achieved if and only if $D^{-l_i} = p_i$ or $-\log_D p_i = l_i$ is an integer for all i.

Entropy is the fundamental limit of "lossless" data compression.

Shannon Code and its Properties

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$$

- Property 1:

They satisfy Kraft's inequality (code should be found from the construction method in Kraft converse proof):

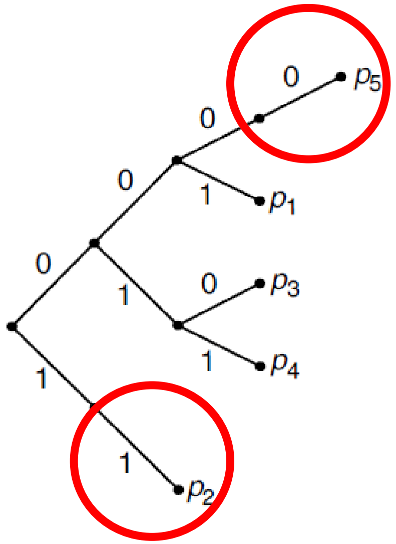
$$\sum D^{-l_i} = \sum D^{-\left\lceil \log_D \frac{1}{p_i} \right\rceil} \leq \sum D^{-\log_D \frac{1}{p_i}} = \sum p_i = 1$$

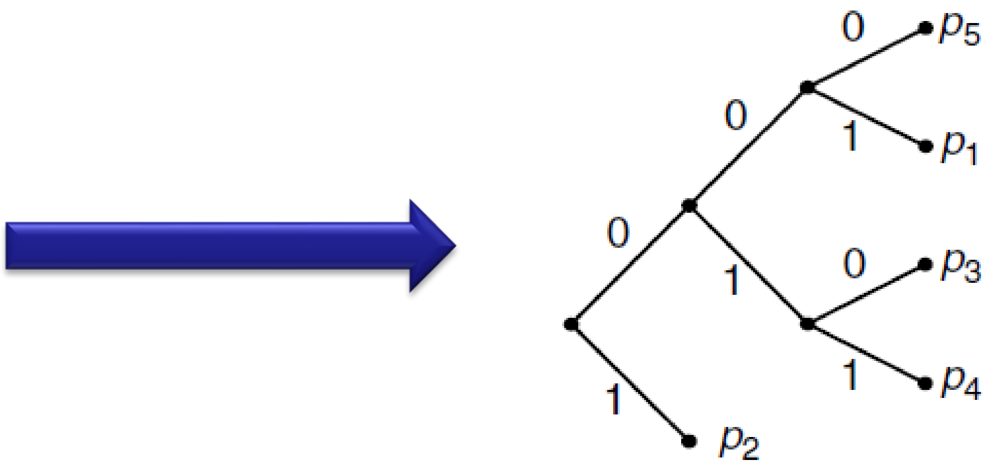
- Property 2:

$$H(X) \leq L(C) < H(X) + 1$$

Optimal Source Code

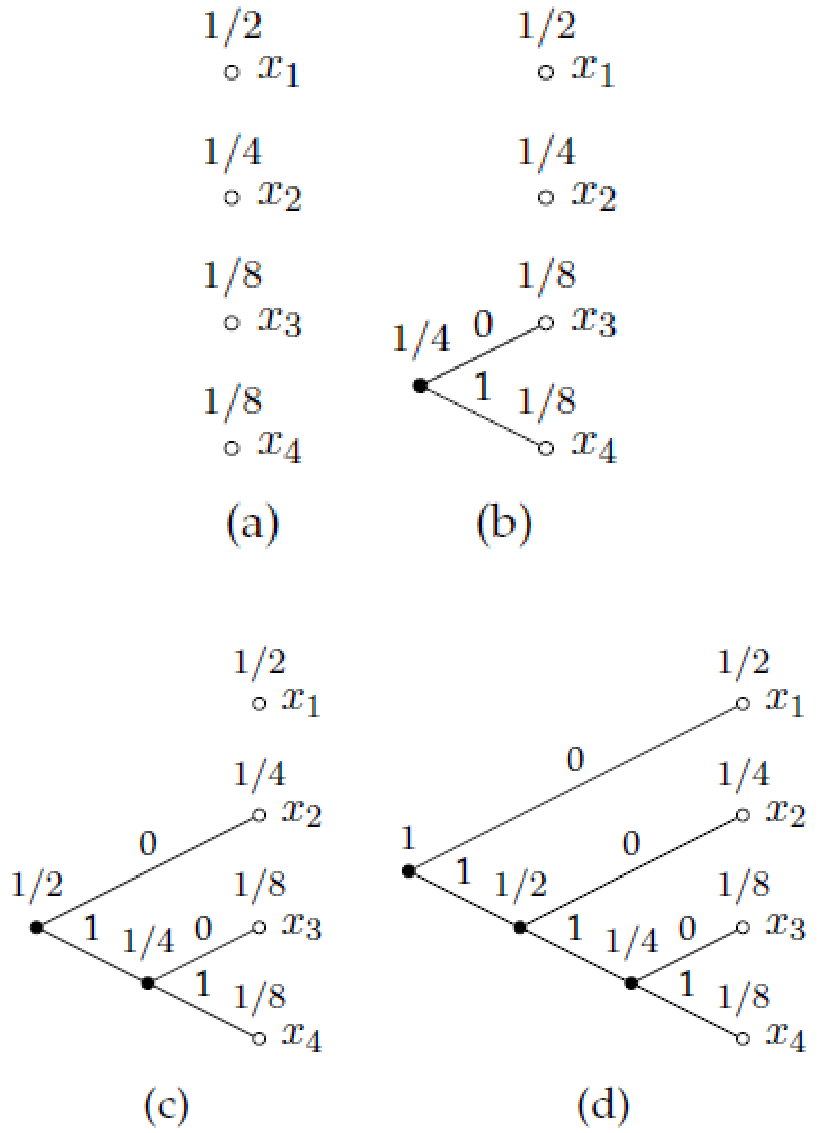
- Requirements:
  - Codeword lengths are inversely ordered with probabilities: $p_j > p_k \Rightarrow l_j \leq l_k$ (else swap codewords)
  - In a tree corresponding to an optimal code there are no unused leaves (otherwise remove the single branch)



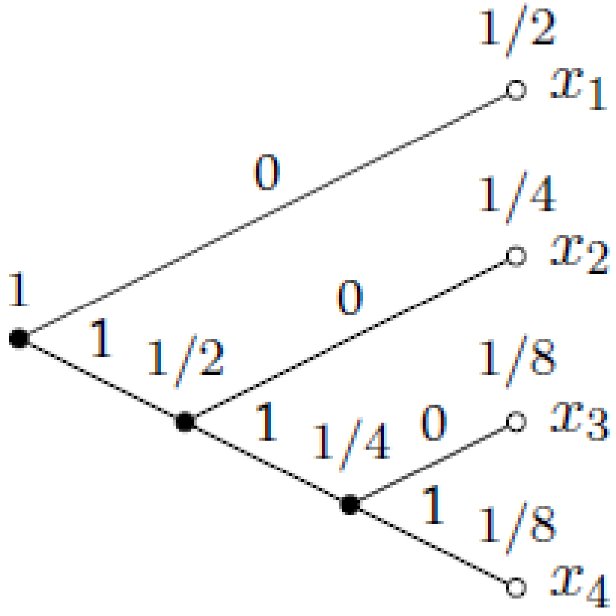


Huffman Code

1. Sort the symbols according to their decreasing probabilities.
2. Let x_j and $x_{j'}$ be the least two probable symbols in the list with probabilities p_j and $p_{j'}$, respectively.
3. Remove x_j and $x_{j'}$ from the list and connect them in a binary tree.
4. Add the root node $\{x_j, x_{j'}\}$ as one symbol with probability $p_j + p_{j'}$
5. If there is only one symbol in the list, stop, otherwise go to step 2



Huffman Code Example



(d)

Root to Leaf

x	$p(x)$	y
x_1	1/2	0
x_2	1/4	10
x_3	1/8	110
x_4	1/8	111

- Average Code Length: $L = \left(\frac{1}{2} * 1\right) + \left(\frac{1}{4} * 2\right) + \left(\frac{1}{8} * 3\right) + \left(\frac{1}{8} * 3\right) = 1.75 \text{ bits}$

- Entropy: $H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + 2 * \frac{1}{8}\log_2 8 = 1.75 \text{ bits}$

Information and Operation Channel Capacity

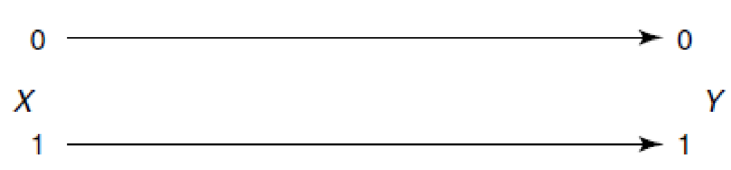
- **Operational** definition of channel capacity is the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.
- **Information** channel capacity of a discrete memoryless channel is defined as

$$\boxed{C = \max_{p(x)} I(X;Y)}$$

- **Shannon proved that operational and information channel capacities are equal.**

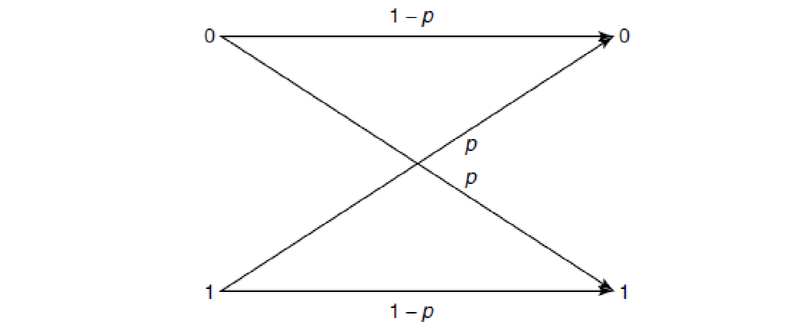
Channel Capacity Examples

- Error Free Channel



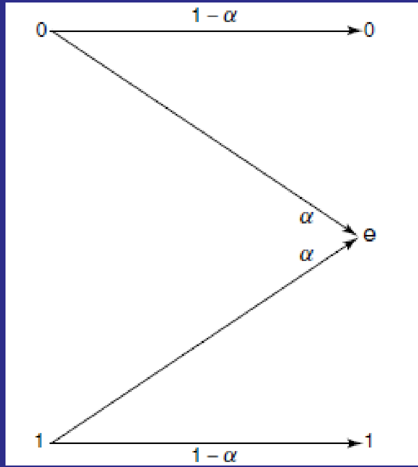
$$I(X;Y) = H(X) - H(X|Y) = H(X)$$

- Binary Symmetric Channel



$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum p(x)H(Y|X=x) \\ &= H(Y) - H(p) \\ &\le 1 - H(p) \end{aligned}$$

- Binary erasure channel



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

$$C = 1 - \alpha.$$

Empirical Joint Entropy

- We can compute its **empirical joint** entropy as

$$\tilde{H}(\mathbf{x},\mathbf{y}) = -\frac{1}{n}\log Pr(\mathbf{x},\mathbf{y}) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i = x_i, Y_i = y_i)$$

Joint Typical Set

- That is, a sequence $(\mathbf{x},\mathbf{y})$ belongs to $A_\epsilon^{(n)}$ if we have all the following

$$\boxed{|\tilde{H}(\mathbf{x}) - H(X)| < \epsilon}$$

$$\boxed{|\tilde{H}(\mathbf{y}) - H(Y)| < \epsilon}$$

$$\boxed{|\tilde{H}(\mathbf{x},\mathbf{y}) - H(X,Y)| < \epsilon}$$