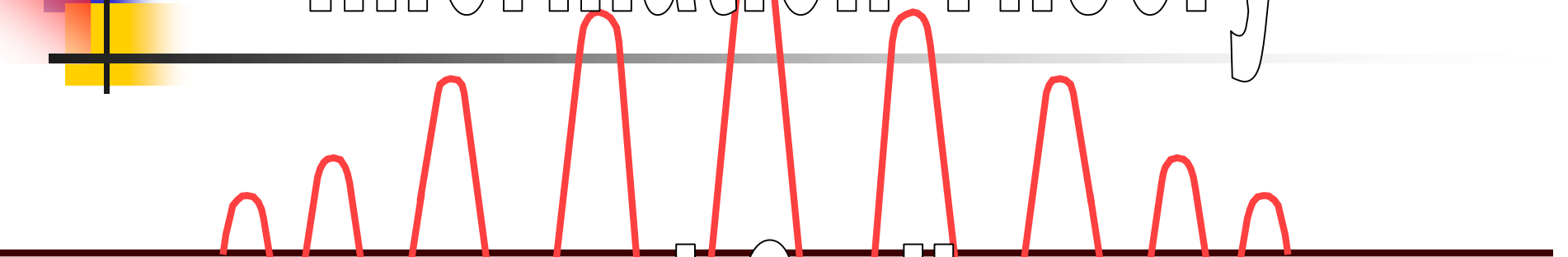


Information Theory



and Coding

# EE670 Information Theory and Coding

- Class home page:

<http://koala.ece.stevens-tech.edu/~utureli/EE670/>

Class information, assignments and other links.

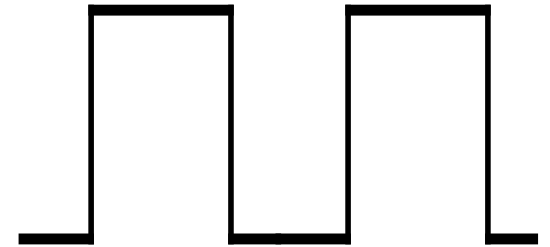
- Text

Elements of Information Theory, Thomas Cover and Joy A Thomas, Wiley, 1991.

- Entropy, Data Compression, Rate Distortion Theory
- Ch. 2- 14
- No class next Thursday 1/24/02,(Monday schedule)

# DIGITAL SIGNAL

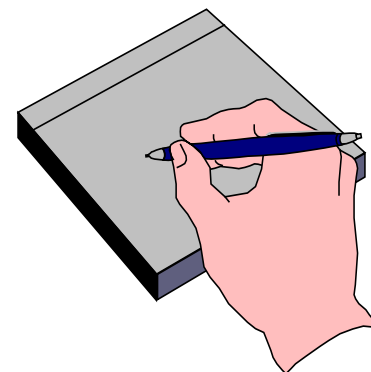
- DISCRETE WAVEFORM
- TWO DISCRETE STATES:
  - 1-BIT & 0-BIT
  - ON / OFF PULSE
- DATA COMMUNICATION
- USES **MODEM** TO TRANSLATE ANALOG TO DIGITAL, DIGITAL TO ANALOG



\*

0110101111010000100010011110111010101001011100101110100

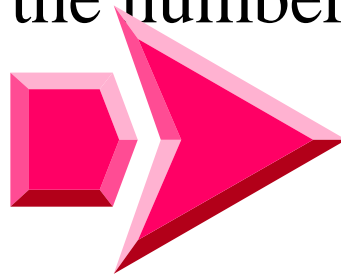
# LEARNING OBJECTIVES



- Information Theory
  - Biotech, Communications, Security, Finance
- Uncertainty, mutual information
  - DNA molecule, Compression , Coding, Investment, Gambling
- Topics
  - Entropy and mutual information
  - Source coding
  - Channel capacity, noisy channel coding theorem
  - Rate-distortion theorem

# Fundamental Limits of Communication

- Purpose of communication system: carry information bearing signals.
- Entropy: irreducible complexity below which a signal cannot be compressed
- Capacity: intrinsic ability of a channel to transmit information.
- If capacity of a channel exceeds entropy rate of the source, than lossless communication is possible!
- Rate Distortion determines accuracy in the reconstructed signal as a function of the number of bits used (i.e. the rate)?



## Random Variables and Probability

- ***Random variable***  $X$  assumes a value as a function from outcomes of a process which can not be determined in advance.
- ***Sample space***  $S$  of a random variable is the set of all possible values of the variable  $X$ .
- $\Omega$ : set of all outcomes and divide it into elementary events, or ***states*** 
$$\sum_{\{x\}} p(x) = 1 \quad 1 \geq p(x) \geq 0$$

## Expectation, Variance and Deviation

The moments of a random variable define important characteristics of random variables:

The first moment is the *expectation*  $\mu$ :

**Note:** The expectation has a misleading name and is not always the value we expect to see most. In the case of the number on a dice the expectation is 3.5 which is not a value we will see at all!. The expectation is as a weighted average.

The *variance* is defined by  $\text{Var}[x] = \langle x^2 \rangle - \langle x \rangle^2 = M_2 - M_1^2$ .

The standard deviation  $\sigma = \text{Var}[x]^{1/2}$  evaluates the “spread factor” or  $x$  in relation to the mean.

## Conditional probability and Bayes' Rule

The knowledge that a certain event occurred can change the probability that another event will occur.  $P(x|y)$  denotes the *conditional probability* that  $x$  will happen given that  $y$  has happened. Bayes' rule states that:

The *complete probability formula* states that

$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$  or in the more general case

**Note:**  $P(A) = \alpha P(A|B) + (1-\alpha)P(A|\neg B)$  mirrors our intuition that the unconditional probability of an event is somewhere between its conditional probability based on two opposing assumptions.

# Information Theory: Entropy

We want a fundamental measure that will allow us to gauge the amount of missing information in a distribution, or a set of possible states. A measure which gives us such an estimate is the entropy:

- ◆  $H[P] \geq 0$  (and is 0 when there is one possible outcome)
- ◆  $H[P]$  is concave:  $H[\alpha P_1 + (1-\alpha)P_2] \geq \alpha H[P_1] + (1-\alpha)H[P_2]$
- ◆ The entropy get it's maximal value when there is the most uncertainty. This intuitively occurs where  $p(x)$  is uniform (we do not prefer any state to the other).

## Entropy (2)

The *joint entropy* is defined by:

$$H[X, Y] = -\sum_x \sum_y p(x, y) \log p(x, y) = -E[\log p(X, Y)]$$

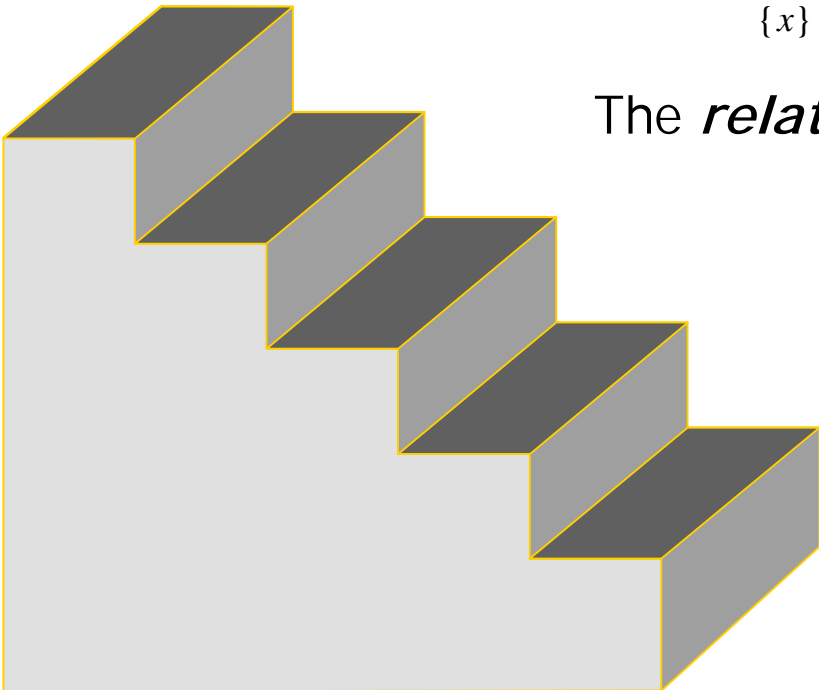
The *conditional entropy* is defined by:

$$\begin{aligned} H[Y | X] &= \sum_{\{x\}} p(x) H(Y | X = x) = -\sum_{\{x\}} p(x) \sum_{\{y\}} p(y | x) \log p(y | x) \\ &= -\sum_{\{x\}} \sum_{\{y\}} p(x, y) \log(p(y | x)) \end{aligned}$$

The *relative entropy* (or the “distance between two distributions) is:

$$D[p | q] = \sum_{\{x\}} p(x) \log \frac{p(x)}{q(x)}$$

equals 0 only if  $p(x)$  and  $q(x)$  are the same.



## Mutual Information

The *mutual information* between two random variables is:

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

- ◆ The information of X on Y is same as of Y on X.
- ◆  $I(X;Y)=0$  if X and Y are independent
- ◆ The mutual missing information  $H(X,Y)$  is the missing information of X and the conditional missing information of Y given that X is known
- ◆ The information of Y on X is the improvement of missing information about X once Y is known.

## Information

The *mutual information* between two random variables is:

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

- ◆ The information of X on Y is same as of Y on X.
- ◆  $I(X;Y)=0$  if X and Y are independent
- ◆ The mutual missing information  $H(X,Y)$  is the missing information of X and the conditional missing information of Y given that X is known
- ◆ The information of Y on X is the improvement of missing information about X once Y is known.

# Chain Rule

$$H(X_1, X_2) = H(X_1) + H(X_2 | X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3 | X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1)$$

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_{n-1}, \dots, X_1)$$

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

## Jensen's Inequality

A function  $f(x)$  is convex over  $(a,b)$ , if  $\forall x_1, x_2 \in (a,b)$

$$0 \leq \lambda \leq 1, \quad f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

◆ A function is *concave* if  $-f$  is convex.

◆ If  $f$  convex:  $x^2, e^x, \log(x)$

$$Ef(X) \geq f(EX).$$

◆ Proof:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2),$$

◆ Assume true for  $k-1$  mass points, let  $p_i' = p_i / (1-p_k), i = 1, 2, \dots, k-1$

$$\sum_{i=1}^k p_i f(x_i) = p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i)$$

$$\sum_{i=1}^k p_i f(x_i) \geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

$$\sum_{i=1}^{k-1} p_i' f(x_i) \geq f(p_k x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

## Jensen's Inequality

- ◆ Prove  $EX^2 \geq (EX)^2$ .
- ◆ Hint:  $f(X) = X^2$
- ◆ Assume true for  $k-1$  mass points, let

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2),$$

$$p_i' = p_i / (1 - p_k), \quad i = 1, 2, \dots, k-1$$

$$\sum_{i=1}^k p_i f(x_i) = p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p_i' f(x_i)$$

$$\sum_{i=1}^k p_i f(x_i) \geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

$$\sum_{i=1}^{k-1} p_i' f(x_i) \geq f(p_k x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

## Data Processing Inequality

If  $X \rightarrow Y \rightarrow Z$  forms a Markov chain, then  $I(X;Y) \geq I(X;Z)$

- ◆ Proof: Chain rule, expand mutual information  $I(X;Y,Z)$  in two ways:

$$I(X;Y,Z) = I(X;Y) + I(X;Z | Y) = I(X;Y) + 0 = I(X;Y)$$

$$I(X;Y,Z) = I(X;Z) + I(X;Y | Z) \geq I(X;Z)$$