ELEC 515
Information Theory

Review

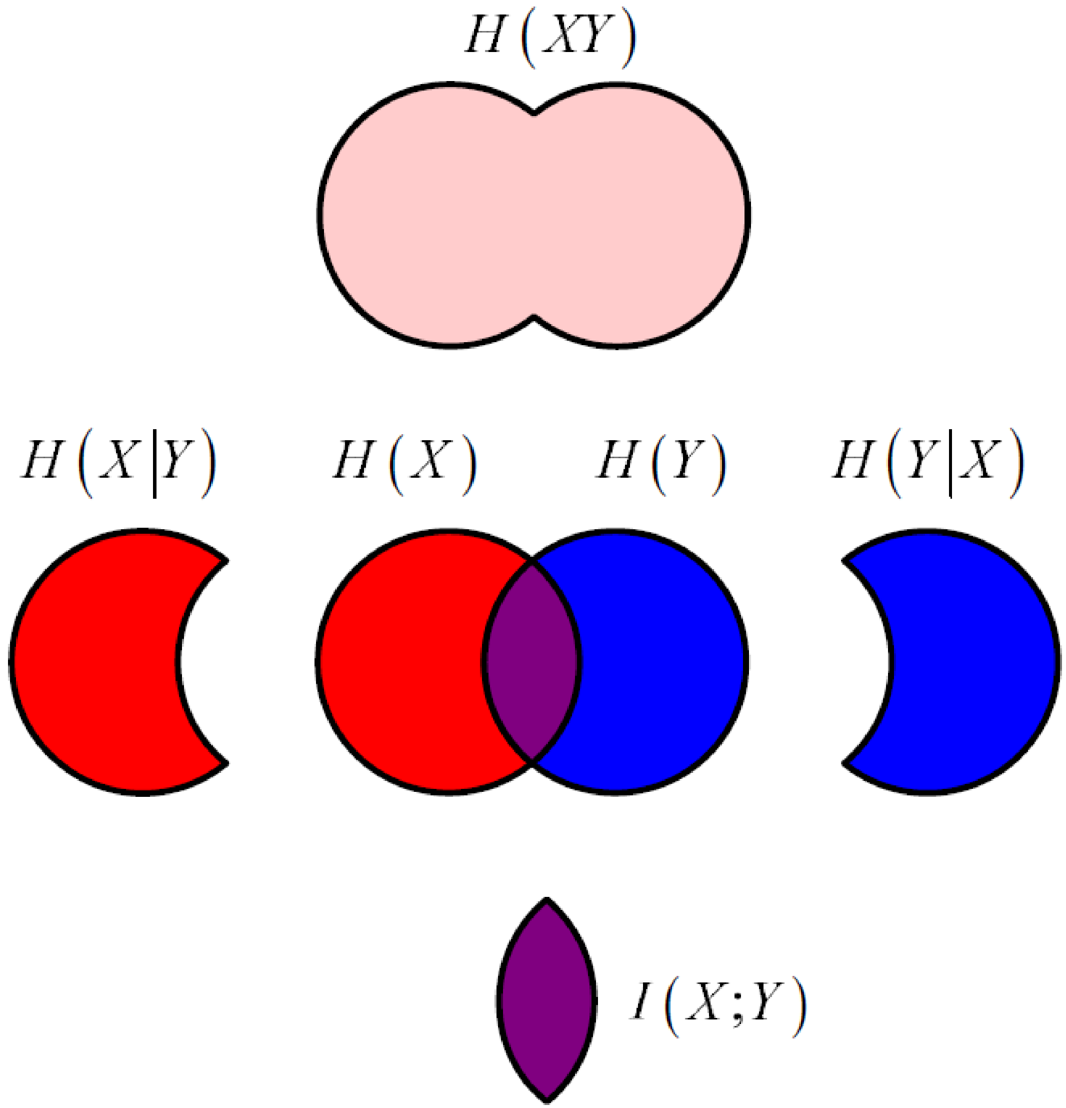
Final Exam

- Friday, December 20, 7:00 PM ECS 116
- 3 hour exam
- ALL course content is covered except
  - logistic regression
  - differential entropy
- Materials Allowed
  - calculator
  - two pages of notes on 8.5” × 11.5” paper

Entropy

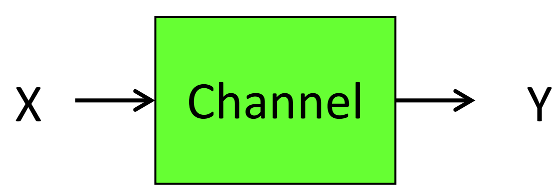
$$H(X) = -\sum_{i=1}^{N} p(x_i)\ \log_b p(x_i)$$

- Joint Entropy H(XY)
- Conditional Entropy H(X|Y)
- Mutual Information I(X;Y)

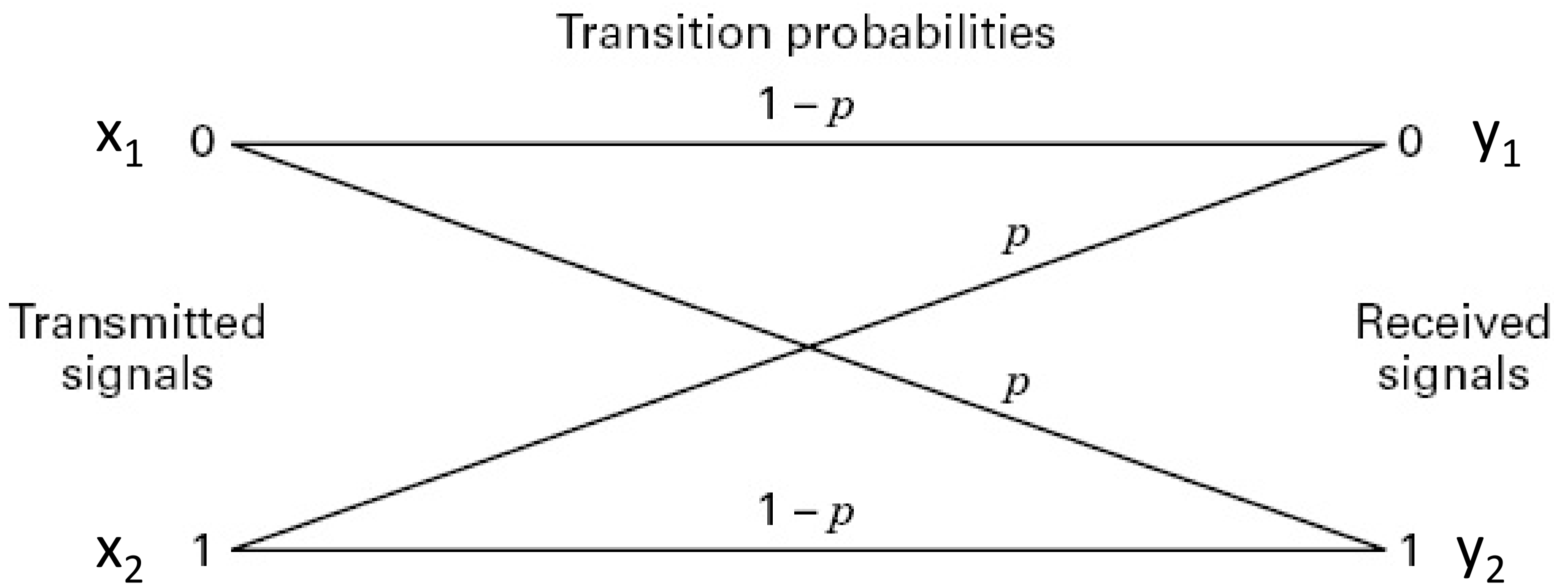

$H(XY)$
$H(X|Y)$
$H(X)$
$H(Y)$
$H(Y|X)$
$I(X;Y)$

Information Channels

- An information channel is described by an
- Input alphabet X
- Output alphabet Y
- Set of conditional probabilities $p(y_j|x_i)$

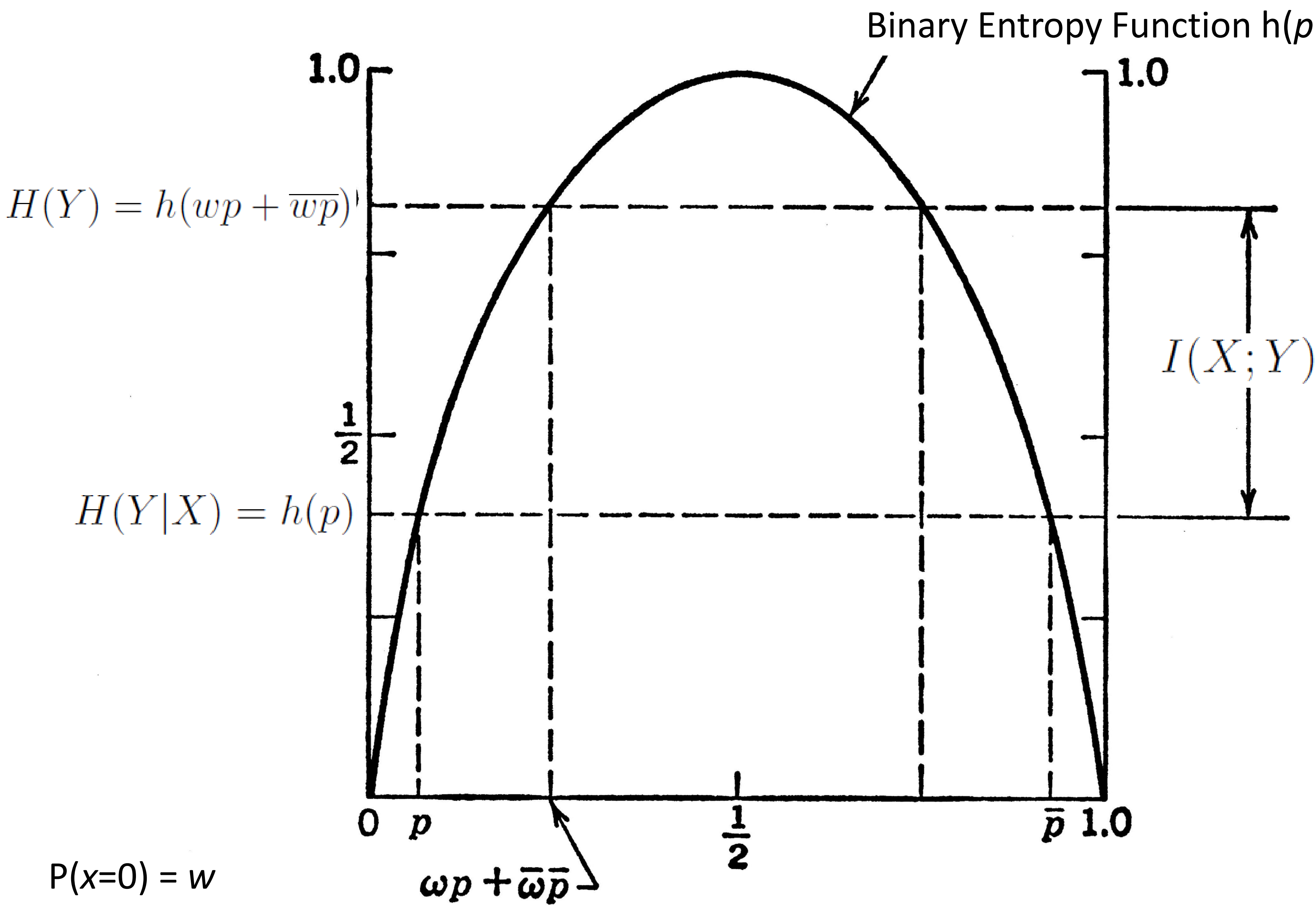


Binary Symmetric Channel



$p(0|1) = p(1|0) = p$
$p(0|0) = p(1|1) = 1\text{-}p$

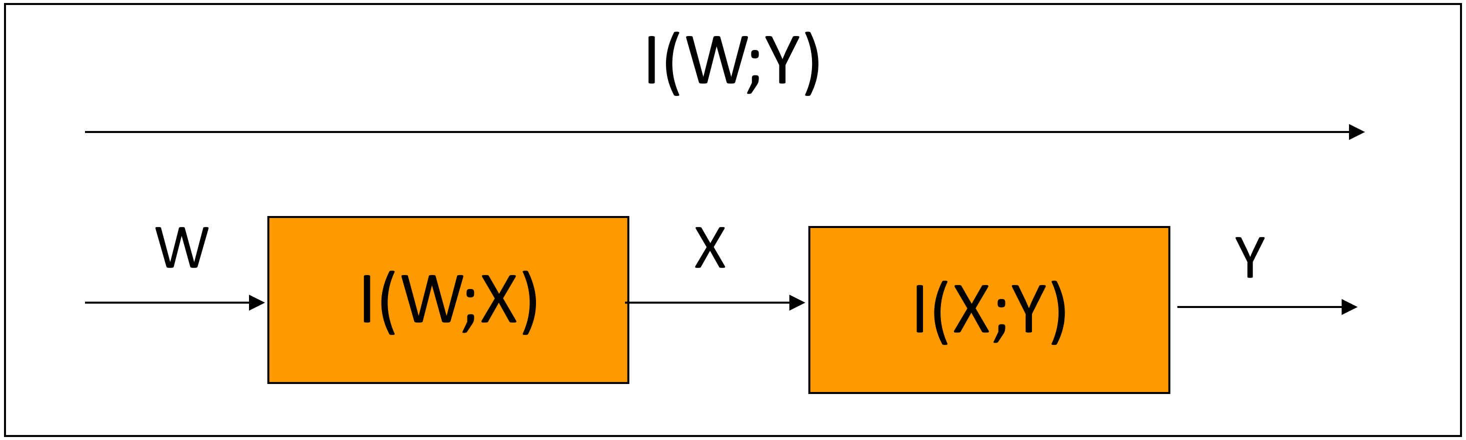
$I(X;Y) = H(Y)\text{-}H(Y|X)$
$= H(Y)\text{-}h(p)$


Binary Entropy Function h(p)
1.0
1.0
$H(Y) = h(wp + \overline{wp})$
$I(X;Y)$
$\frac{1}{2}$
$H(Y|X) = h(p)$
0
p
$\frac{1}{2}$
$\bar{p}$
1.0
$\omega p + \bar{\omega}\bar{p}$

P(x=0) = w

The Data Processing Inequality

Cascaded Channels



The mutual information I(W;Y) for the cascade cannot be larger than I(W;X) or I(X;Y), so that

$$I(W;Y) \leq I(W;X) \qquad I(W;Y) \leq I(X;Y)$$

Relative Entropy and Cross Entropy

$$D\left[p(X) \| q(X)\right] = \sum_{i=1}^{N} p(x_i) \ \log_b \left[\frac{p(x_i)}{q(x_i)}\right]$$

$$H(p, q) = -\sum_{i=1}^{N} p(x_i) \log q(x_i)$$

Shannon's Noiseless Coding Theorem

$$\frac{\mathrm{H(X)}}{\log_b J} \leq \mathrm{L(C)} < \frac{\mathrm{H(X)}}{\log_b J} + 1$$

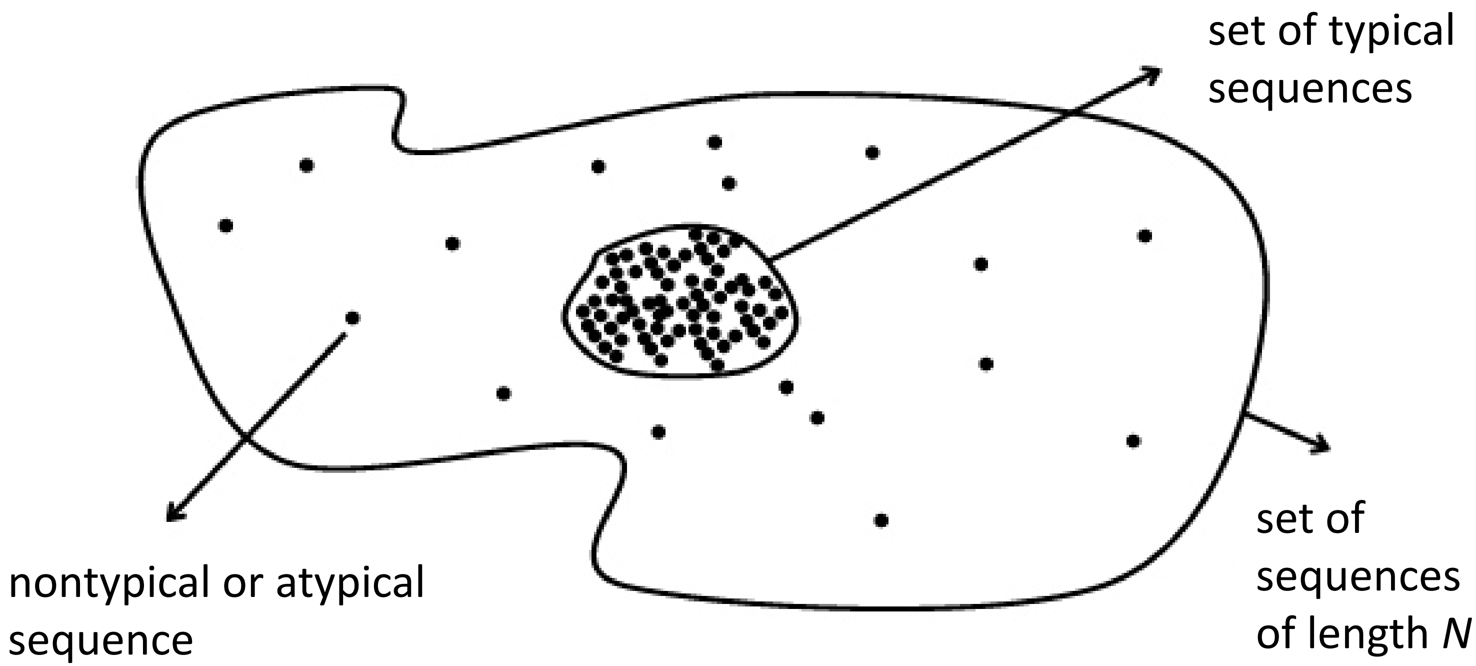
$$\frac{\mathrm{H(X)}}{\log_b J} \leq \frac{\mathrm{L}_N(\mathrm{C})}{N} < \frac{\mathrm{H(X)}}{\log_b J} + \frac{1}{N}$$

Typical Sequences

$$\mathcal{T}_X(\delta) \equiv \{\mathbf{x}: \ |-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\,| < \delta\}$$

$$\mathcal{T}_X^c(\delta) \equiv \{\mathbf{x}: |-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\,| \geq \delta\}$$

Typical Sequences



Shannon-McMillan Theorem

a) The probability that a particular sequence $\mathbf{x}$ of blocklength N belongs to the set of atypical sequences $\mathcal{T}_X^c(\delta)$ is upperbounded as:

$$\boxed{Pr[\mathbf{x} \in \mathcal{T}_X^c(\delta)] < \epsilon}$$

b) If a sequence $\mathbf{x}$ is in the set of typical sequences $\mathcal{T}_X(\delta)$ then its probability of occurrence $p(\mathbf{x})$ is approximately equal to $b^{-NH(X)}$, that is:

$$\boxed{b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}}$$

c) The number of typical, or likely, sequences $\|\mathcal{T}_X(\delta)\|$ is bounded by:

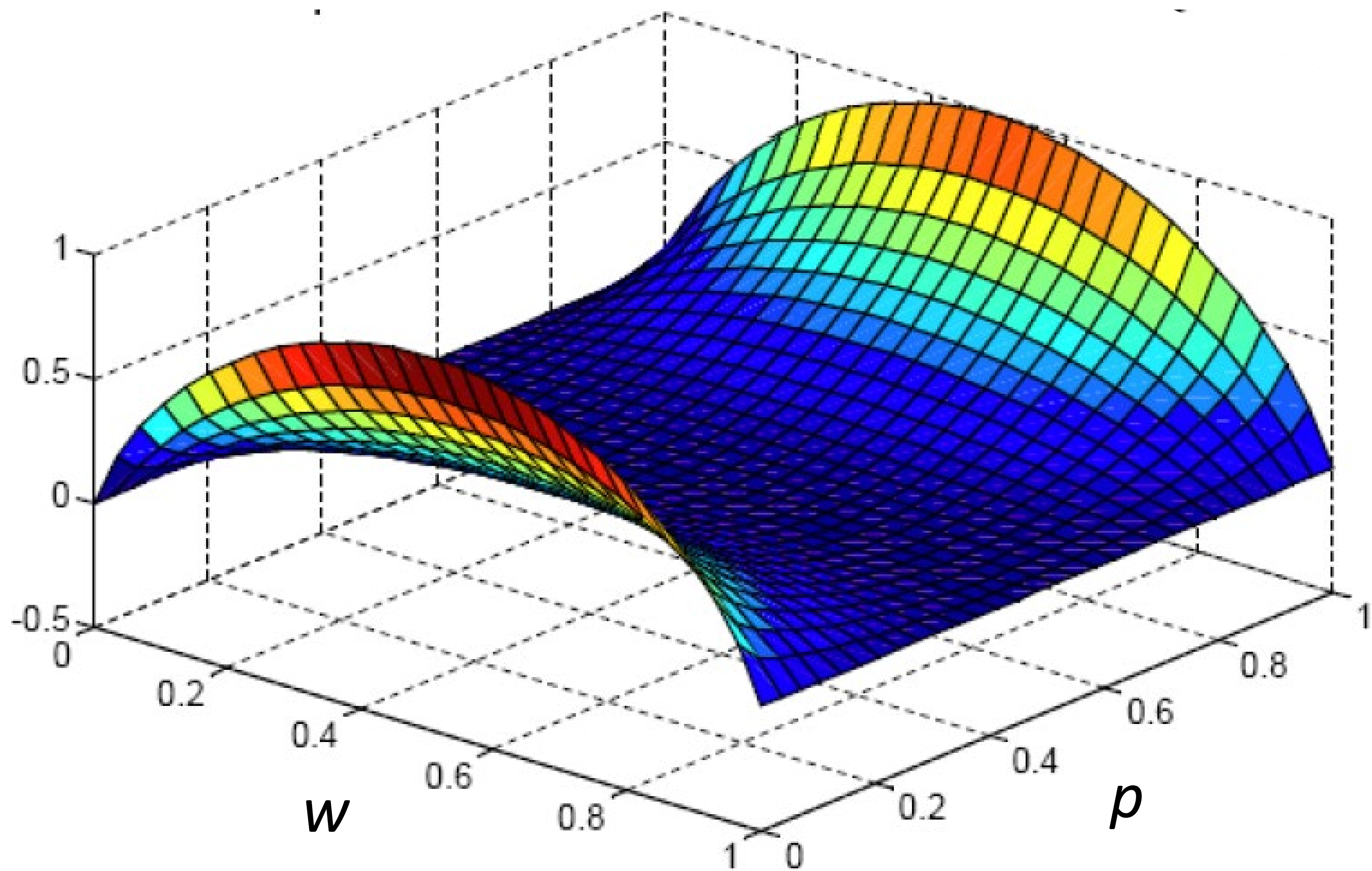
$$\boxed{(1-\epsilon)b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]}}$$

- The essence of source coding or data compression is that as $N \to \infty$, atypical sequences almost never appear as the output of the source.
- Therefore, one can focus on representing typical sequences with codewords and ignore atypical sequences.
- Since there are only about $2^{N\mathrm{H}(\mathrm{X})}$ typical sequences of length N, and they are approximately equiprobable, it takes about $N\mathrm{H}(\mathrm{X})$ bits to represent them.
- On average it takes $\mathrm{H}(\mathrm{X})$ bits to represent a source symbol.

Source Coding Algorithms

- Shannon
- Fano
- Huffman
- Tunstall
- Arithmetic
- Fixed Length Source Compaction
- Lempel-Ziv

I(X;Y) for the BSC

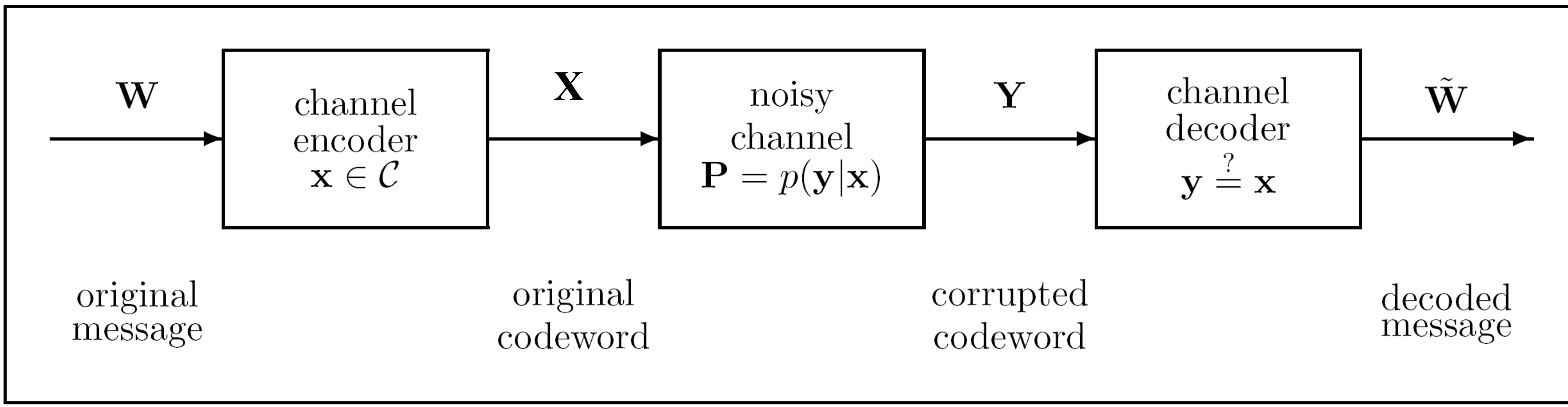


Channel Capacity

The *maximum* value of $I(X;Y)$ as the input probabilities $p(x_i)$ are varied is called the Channel Capacity

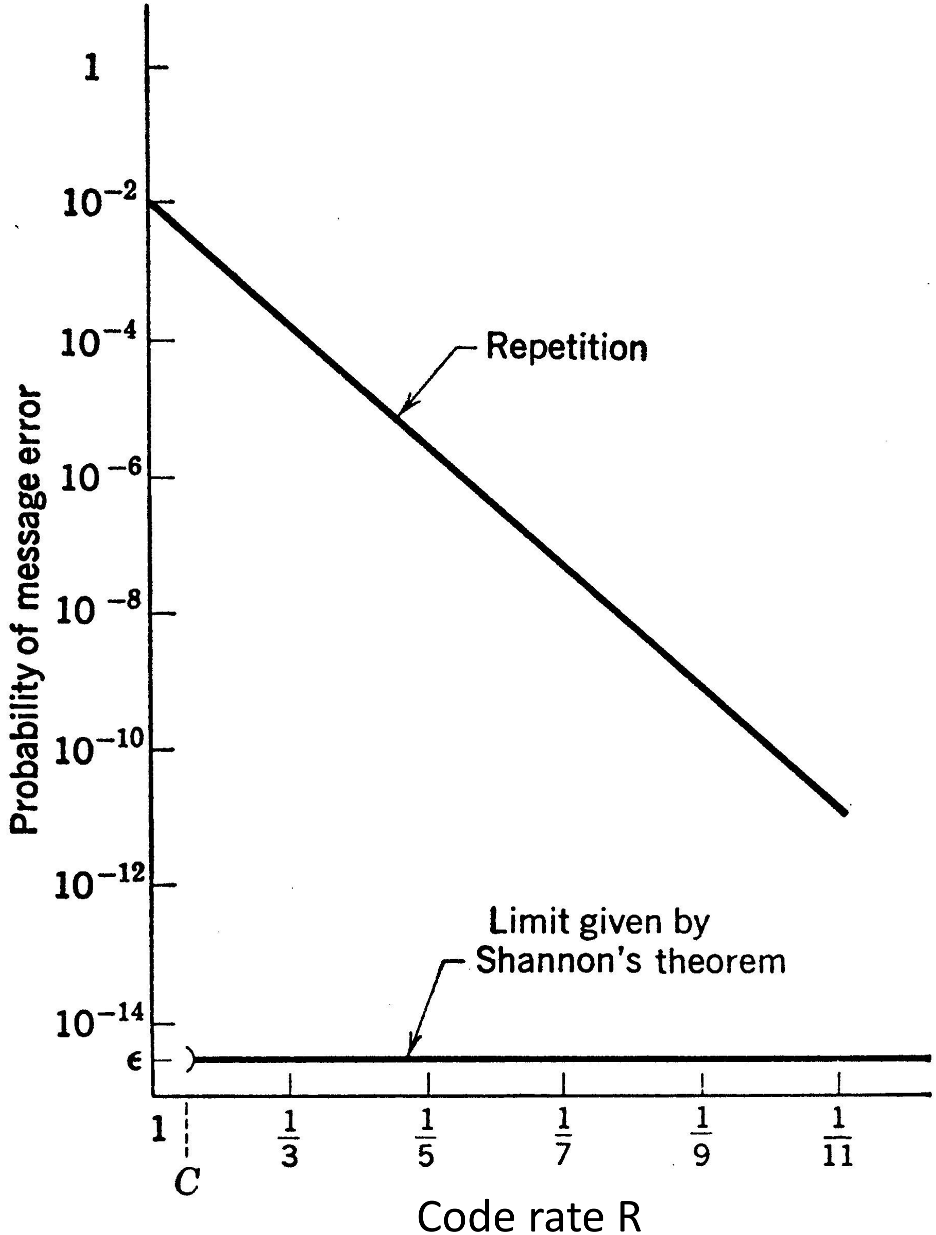
$$C = \max_{p(x_i)} I(X;Y)$$

Communication over Noisy Channels



Shannon's Noisy Coding Theorem

For any $\varepsilon > 0$ and for any rate *R* less than the channel capacity C, there is an encoding and decoding scheme that can be used to ensure that the probability of decoding error is less than ε for a sufficiently large block length *N*.



1
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
ϵ
Probability of message error
Repetition
Limit given by
Shannon's theorem
1
C
$\frac{1}{3}$
$\frac{1}{5}$
$\frac{1}{7}$
$\frac{1}{9}$
$\frac{1}{11}$
Code rate R

Best Known Codes Comparison

- BSC $p = 0.01$ R = 2/3 $M = 2^{NR}$

N	P_e	$\log_2 M$
3	1.99×10^{-2}	2
12	6.17×10^{-3}	8
30	3.32×10^{-3}	20
51	1.72×10^{-3}	34
81	1.36×10^{-3}	54
105	6.92×10^{-4}	70
126	2.99×10^{-4}	84

- For fixed R, P_e can be decreased by increasing N

Code Matrix

$$\mathcal{C} = \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,n} & \cdots & c_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,n} & \cdots & c_{M,N} \end{bmatrix}$$

Binary Codes

- For given values of M and N, there are

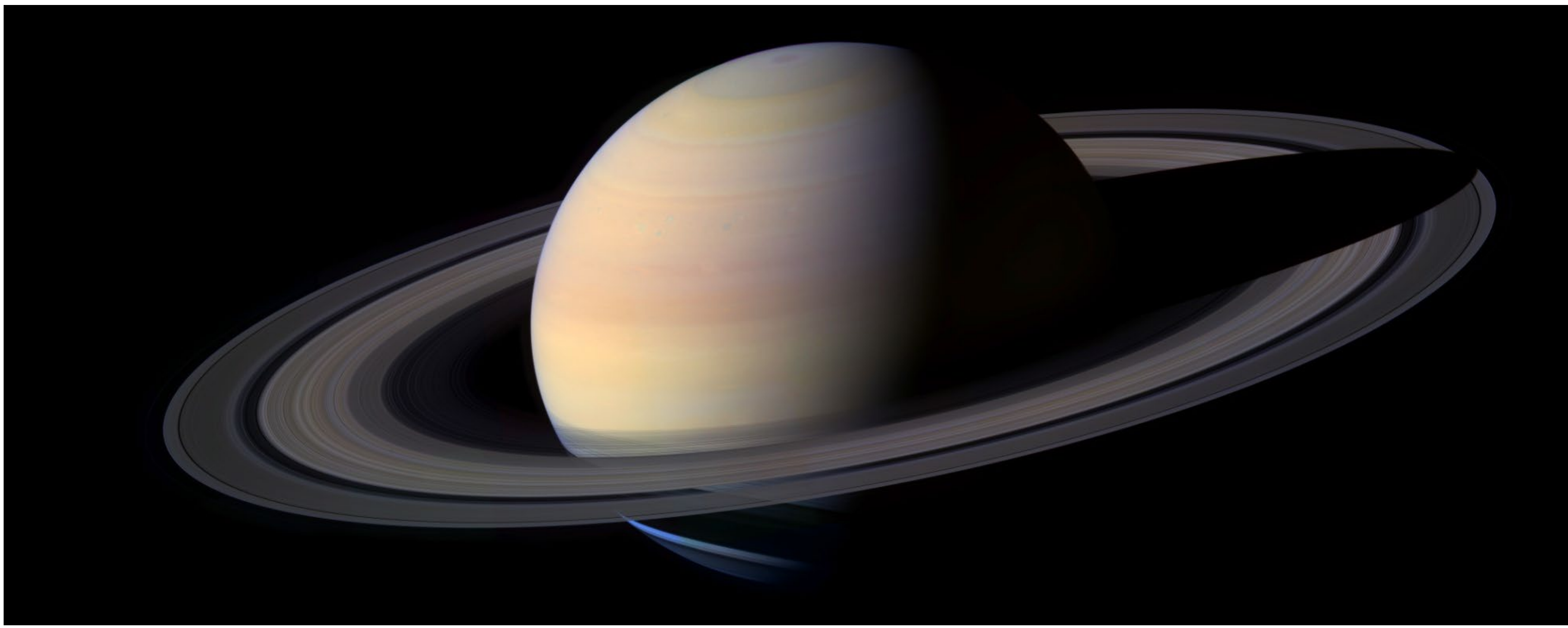
  $$2^{MN}$$

  possible binary codes.
- Of these, some will be bad, some will be best (optimal), and some will be good, in terms of P_e
- An average code will be good.

There are many classes of practical codes

- Hamming codes
- Convolutional codes
- Reed-Muller codes
- Cyclic codes (CRC codes)
- Reed-Solomon codes
- Product codes
- BCH codes
- LDPC codes
- Turbo codes
- Repeat-accumulate codes
- Polar codes
- …

Deep Space Communications



Mars Rover 2021