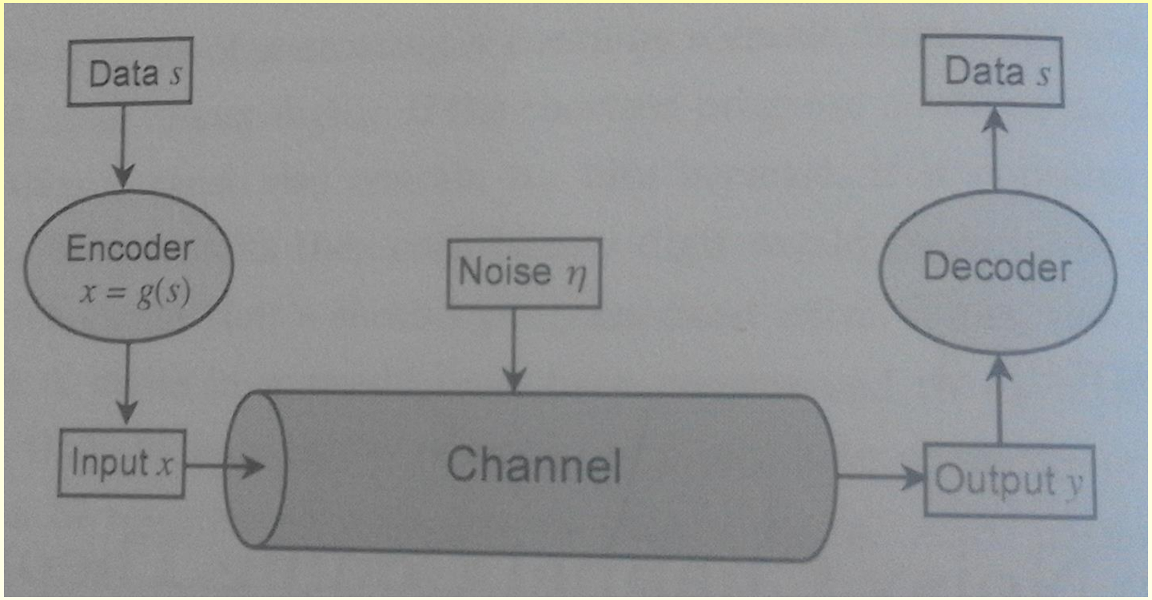
CYBERNETICS

Information in communication:

Shannon Information Theory II

The Noisy Channel Coding Theory

Communication of information in the presence of noise



The concept of Mutual Information

- Is a general measure of association between two variables: (input and output)

- For the variables X and Y, the mutual information I(X,Y) is:

The average information we gain about Y after knowing a single value of X, (x_i)

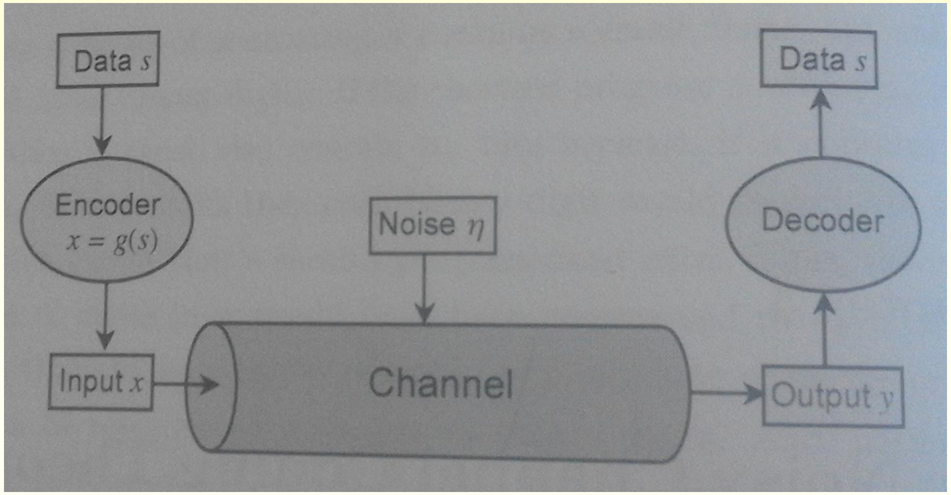
- Symmetrical: I(X,Y) = I(Y,X)

The concept of Mutual Information

- I(X,Y) is the average reduction in <u>uncertainty</u> about Y, H(Y), after knowing a value of X, (x_i) and vice versa

$H(Y) \rightarrow$ reading X $\rightarrow$ residual uncertainty about Y: $H(Y|X)$

$H(Y|X)$ is called *conditional entropy*



Because $Y = X + \eta$

then $H(Y|X) = H(\eta)$

$H(\eta)$ is the entropy of a *joint distribution*

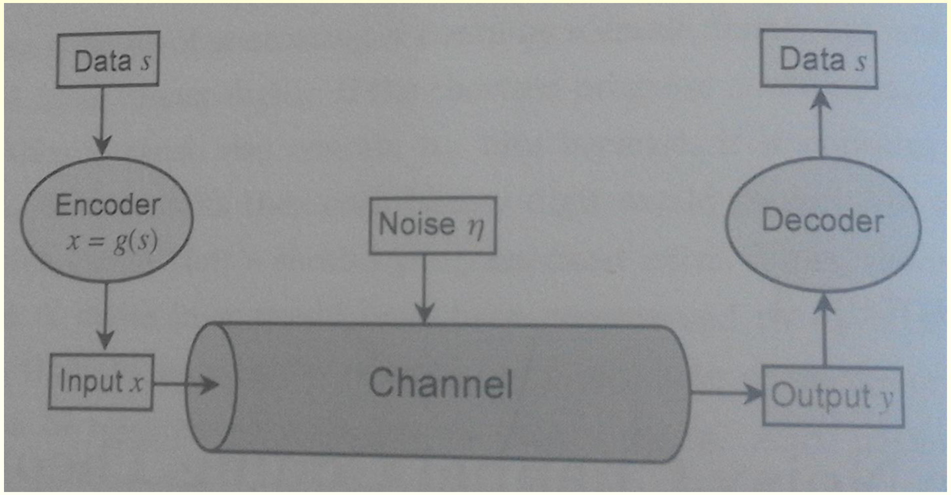
Entropy of Joint Distributions

Let's consider the transmission of 4 possible messages:

$s_1 = 0$, $s_2 = 1$, $s_3 = 2$ and $s_4 = 3$

With a noiseless channel we require log4 = 2 binary digits/message

With a noisy channel we require > 2 binary digits/message



Input:
$p(X) = \{p(x_1),(x_2),(x_3),(x_4)\}$

Output:
$p(Y) = \{p(y_1),(y_2),(y_3),(y_4)\}$

Entropy of Joint Distributions

Let's perform 128 trials and determine the frequencies of outcomes

Y / X	x_1	x_2	x_3	x_4	Sum
y_1	12	15	2	0	29
y_2	4	21	10	0	35
y_3	0	10	21	4	35
y_4	0	2	15	12	29
Sum	16	48	48	16	128

From this we determine the joint probability distribution p(X,Y):

Y / X	x_1	x_2	x_3	x_4	$p(Y)$
y_1	0.094	0.117	0.016	0.000	0.227
y_2	0.031	0.164	0.078	0.000	0.273
y_3	0.000	0.078	0.164	0.031	0.273
y_4	0.000	0.016	0.117	0.094	0.227
$p(X)$	0.125	0.375	0.375	0.125	1

Entropy of Joint Distributions

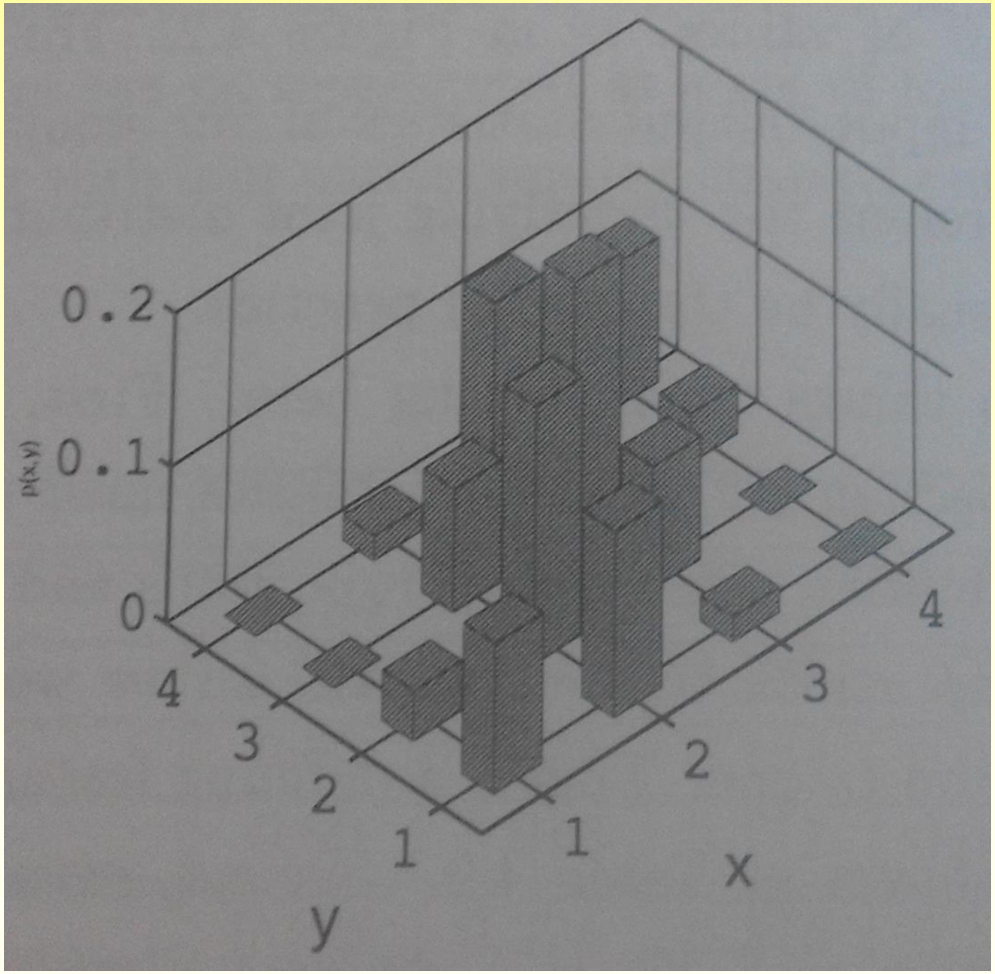
Y / X	x_1	x_2	x_3	x_4	$p(Y)$
y_1	0.094	0.117	0.016	0.000	0.227
y_2	0.031	0.164	0.078	0.000	0.273
y_3	0.000	0.078	0.164	0.031	0.273
y_4	0.000	0.016	0.117	0.094	0.227
$p(X)$	0.125	0.375	0.375	0.125	1

Properties of the joint probability distribution p(X,Y):

- They add up to 1
- It is a continuous function

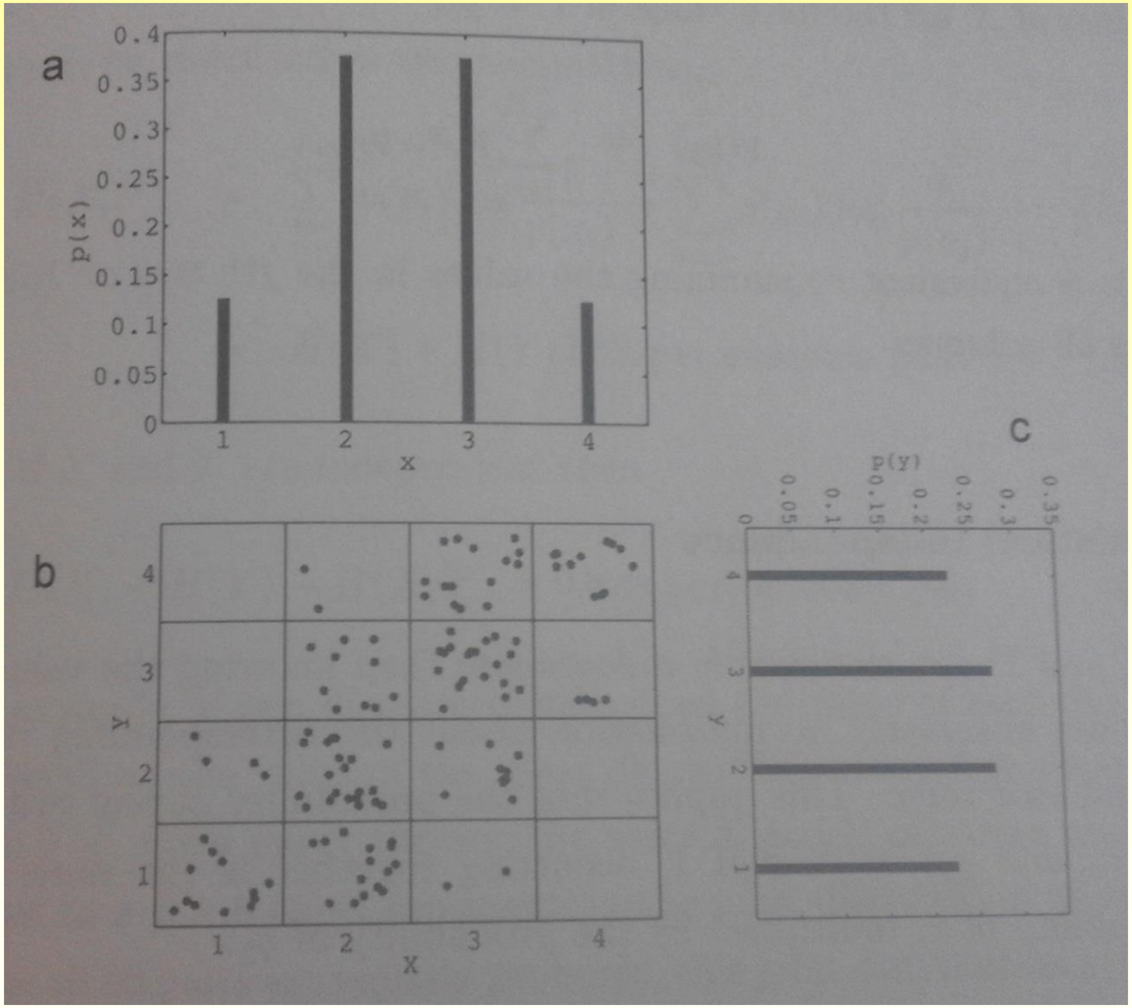
Entropy of Joint Distributions

Graphical representation of the probability distribution p(X,Y):



Entropy of Joint Distributions

Marginal probabilities of p(X,Y):



$$p(x_i) = \sum_{j=1}^{m_y} p(x_i, y_j)$$

$$p(y_j) = \sum_{i=1}^{m_x} p(x_i, y_j)$$

Entropy of Joint Distributions

Entropy of the joint probability distribution p(X,Y):

$$\boldsymbol{H(X) = \sum_{i=1}^{n} p(x_i) log_2 \frac{1}{p(x_i)}}$$

$$\boldsymbol{H(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_j) log \frac{1}{p(x_i, y_j)}}$$

Entropy of Joint Distributions

Statistical independence:

If X and Y are statistically independent, knowing the value of X gives no information about Y, and vice versa

$$p(x_i, y_j) = p(x_i)p(y_j)$$

$$H(X,Y) = \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

$$H(X,Y) = \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i)p(y_j) \log \frac{1}{p(x_i)p(y_j)}$$

Entropy of Joint Distributions

Statistical independence:

$$H(X,Y) = \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i)p(y_j)log\frac{1}{p(x_i)p(y_j)}$$

$$= \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i)p(y_j)log\frac{1}{p(x_i)} + \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i)p(y_j)log\frac{1}{p(y_j)}$$

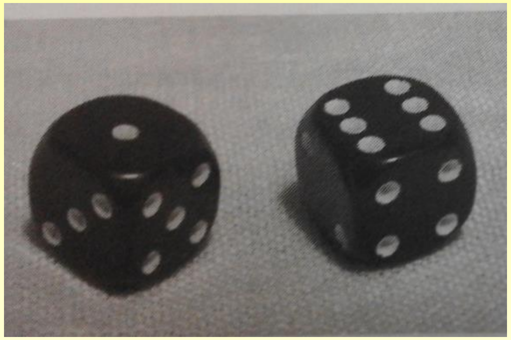
$$= \sum_{j=1}^{m_y} p(y_j)\sum_{i=1}^{m_x} p(x_i)log\frac{1}{p(x_i)} + \sum_{i=1}^{m_x} p(x_i)\sum_{j=1}^{m_y} p(y_j)log\frac{1}{p(y_j)}$$

$$= \sum_{i=1}^{m_x} p(x_i)log\frac{1}{p(x_i)} + \sum_{j=1}^{m_y} p(y_j)log\frac{1}{p(y_j)} = H(X) + H(Y)$$

Entropy of Joint Distributions

Statistical independence:

$$\boldsymbol{H(X,Y) = H(X) + H(Y)} \qquad \boldsymbol{H(X) + H(Y) - H(X,Y) = 0}$$



For the through of two dice:
H(X,Y) = log(36) = 5.17 bit / outcome

For the through of die X:
H(X) = log(6) = 2.585 bit / outcome

For the through of die Y:
H(Y) = log(6) = 2.585 bit / outcome

H(X,Y) = H(X) + H(Y)

If X and Y are independent the entropy of the joint distribution p(X,Y) is the sum of the entropies of the marginal distributions

Mutual Information

For a given channel:

- What portion of the entropy in the outcome reflects information in the input?
- How much of the output entropy is telling about the input, and how much is noise?

The rate of information is given by:

- The entropy of the input, H(X)
- The entropy of the output, H(Y)
- The relationship between X and Y

- H(X) should be high
- H(Y) should be high
- H(noise) should be low

Mutual Information

The mutual information is defined by:

$$I(X,Y) = \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

$$= \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log[p(x_i,y_j)] + \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log\frac{1}{p(x_i)p(y_j)}$$

$$= -\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log\frac{1}{p(x_i,y_j)} + \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log\frac{1}{p(x_i)}$$

$$+ \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i,y_j)\log\frac{1}{p(y_j)}$$

Mutual Information

The mutual information is defined by:

$$= -\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i, y_j)\, log \frac{1}{p(x_i, y_j)} + \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i, y_j) log \frac{1}{p(x_i)}$$

$$+ \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p(x_i, y_j) log \frac{1}{p(y_j)}$$

$$\boldsymbol{I}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbf{H}(\mathbf{X}) + \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{X}, \mathbf{Y})$$

Calculating Mutual Information

Y / X	x_1	x_2	x_3	x_4	*p(Y)*
y_1	0.094	0.117	0.016	0.000	0.227
y_2	0.031	0.164	0.078	0.000	0.273
y_3	0.000	0.078	0.164	0.031	0.273
y_4	0.000	0.016	0.117	0.094	0.227
p(X)	0.125	0.375	0.375	0.125	1

$$\boldsymbol{H(X) = \sum_{i=1}^{m_x} p(x_i) log \frac{1}{p(x_i)} = 1.81\ bits}$$

$$\boldsymbol{H(Y) = \sum_{j=1}^{m_y} p(y_j) log \frac{1}{p(y_j)} = 1.99\ bits}$$

$$\boldsymbol{H(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_j) log \frac{1}{p(x_i, y_j)} = 3.3\ bits}$$

Calculating Mutual Information

Y / X	x_1	x_2	x_3	x_4	$p(Y)$
y_1	0.094	0.117	0.016	0.000	0.227
y_2	0.031	0.164	0.078	0.000	0.273
y_3	0.000	0.078	0.164	0.031	0.273
y_4	0.000	0.016	0.117	0.094	0.227
$p(X)$	0.125	0.375	0.375	0.125	1

$$\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{H}(\mathbf{X}) + \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{X},\mathbf{Y}) = \mathbf{0.509\ bits}$$

$$\frac{\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y})}{\boldsymbol{H}(\boldsymbol{Y})} = \mathbf{0.256}$$

- Only 25.6% of the output entropy is information about the input

- 74.4% is just channel noise

Conditional Entropy

- I(X,Y) is the average reduction in <u>uncertainty</u> about Y, H(Y), after knowing a value of X, (x_i) and vice versa

H(Y) → reading X → residual uncertainty about Y: H(Y|X)

H(Y|X) is called *conditional entropy*

As mutual information is defined by:

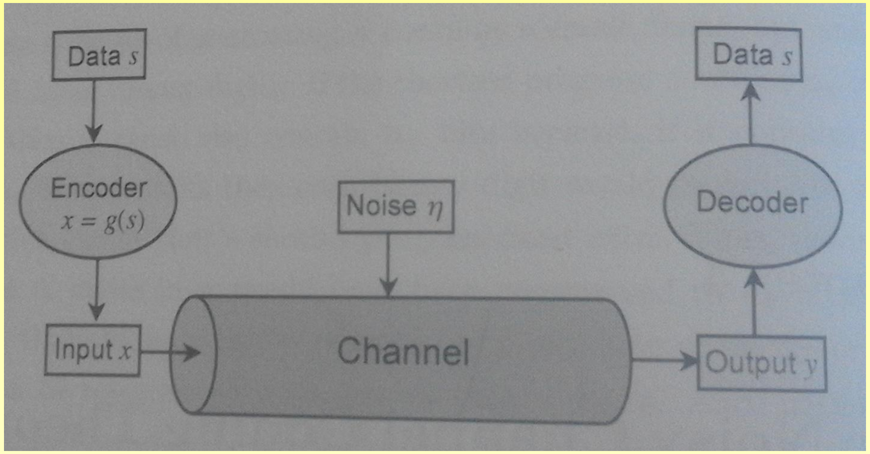
$$I(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

$$I(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_j) log \frac{p(y_j|x_i)}{p(y_j)}$$

$$I(X,Y) = \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{Y}|\mathbf{X})$$

$$I(X,Y) = \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}|\mathbf{Y})$$

Conditional Entropy and Noise



$$Y = X + \eta$$

$$I(X,Y) = \mathbf{H(Y)} - \mathbf{H(Y|X)}$$

$$I(X,Y) = \mathbf{H(Y)} - \mathbf{H([X+\eta]|X)}$$

$$I(X,Y) = \mathbf{H(Y)} - \mathbf{H}(X|\mathbf{X}) - \mathbf{H(\eta|X)}$$

$$I(X,Y) = \mathbf{H(Y)} - \mathbf{H(\eta|X)}$$

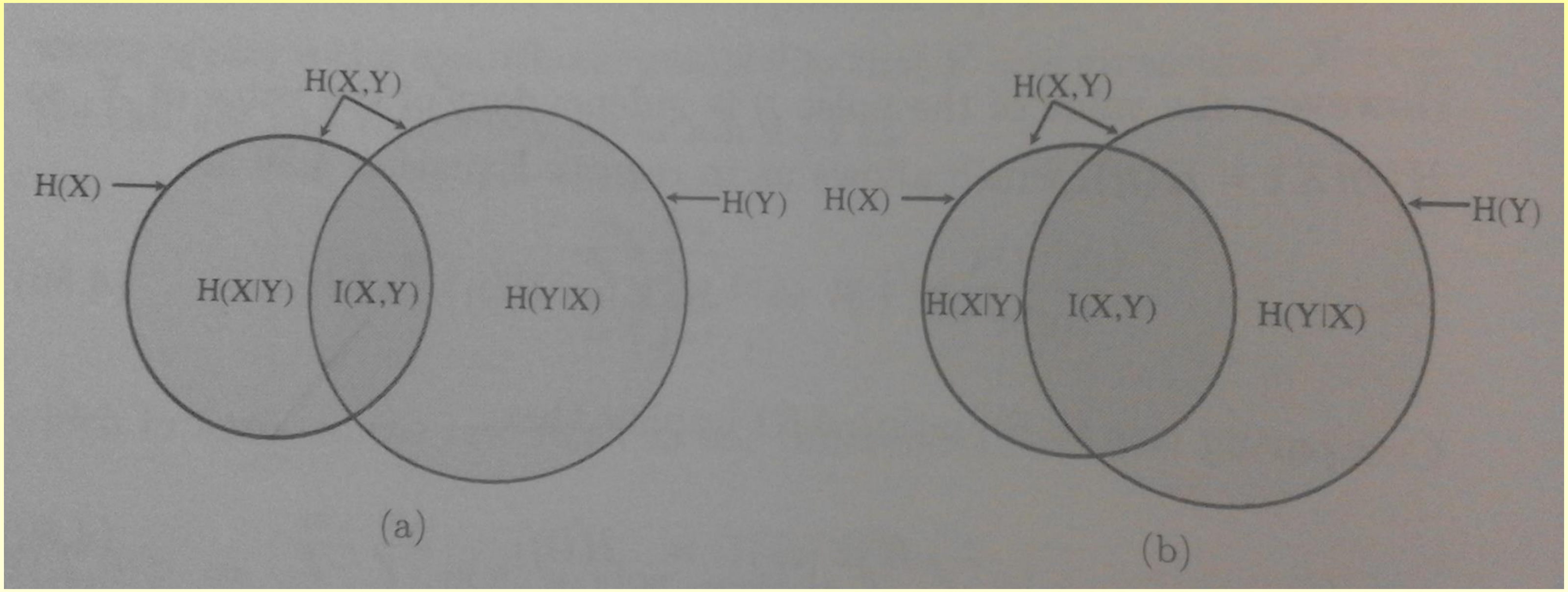
But the noise is independent of X, thus H(η|X) = H(η)

$$I(X,Y) = \mathbf{H(Y)} - \mathbf{H(\eta)}$$

Thus, $\mathbf{H(Y|X)} = \mathbf{H(\eta)}$

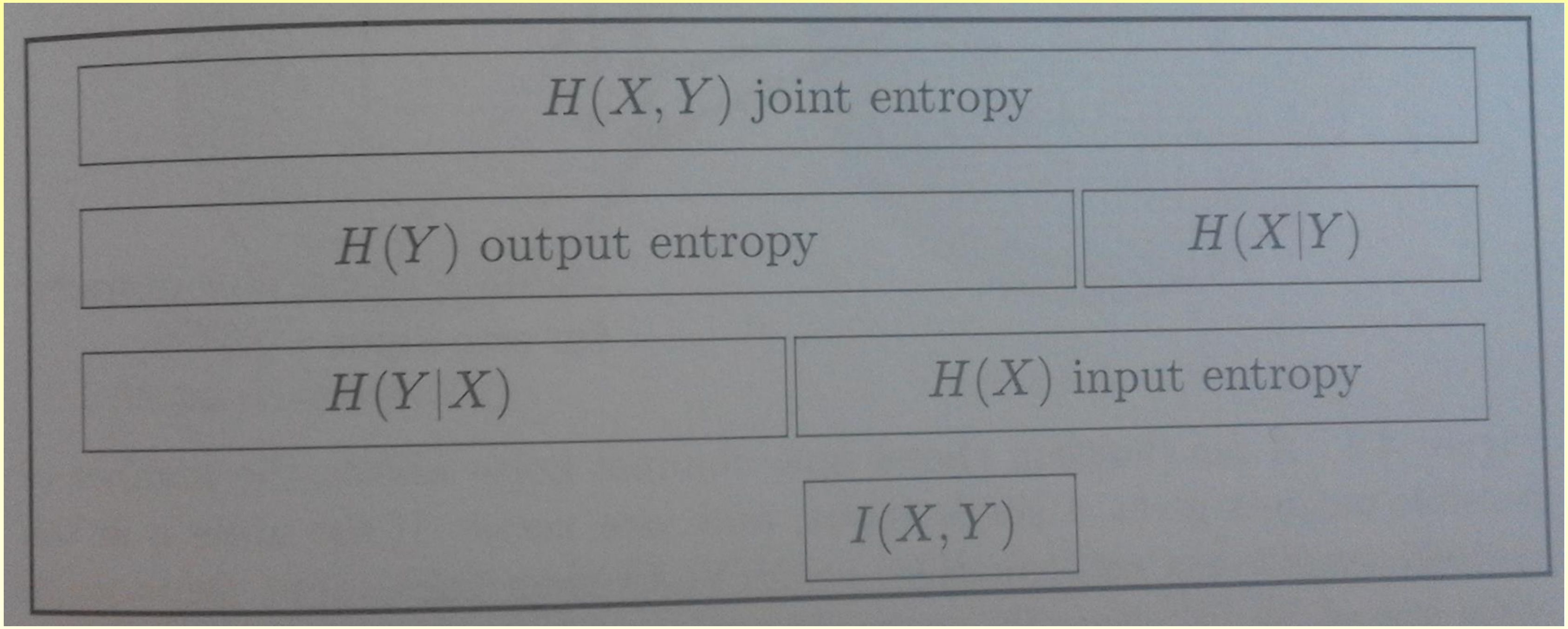
The conditional entropy is the entropy of the channel's noise

Conditional Entropy and Noise



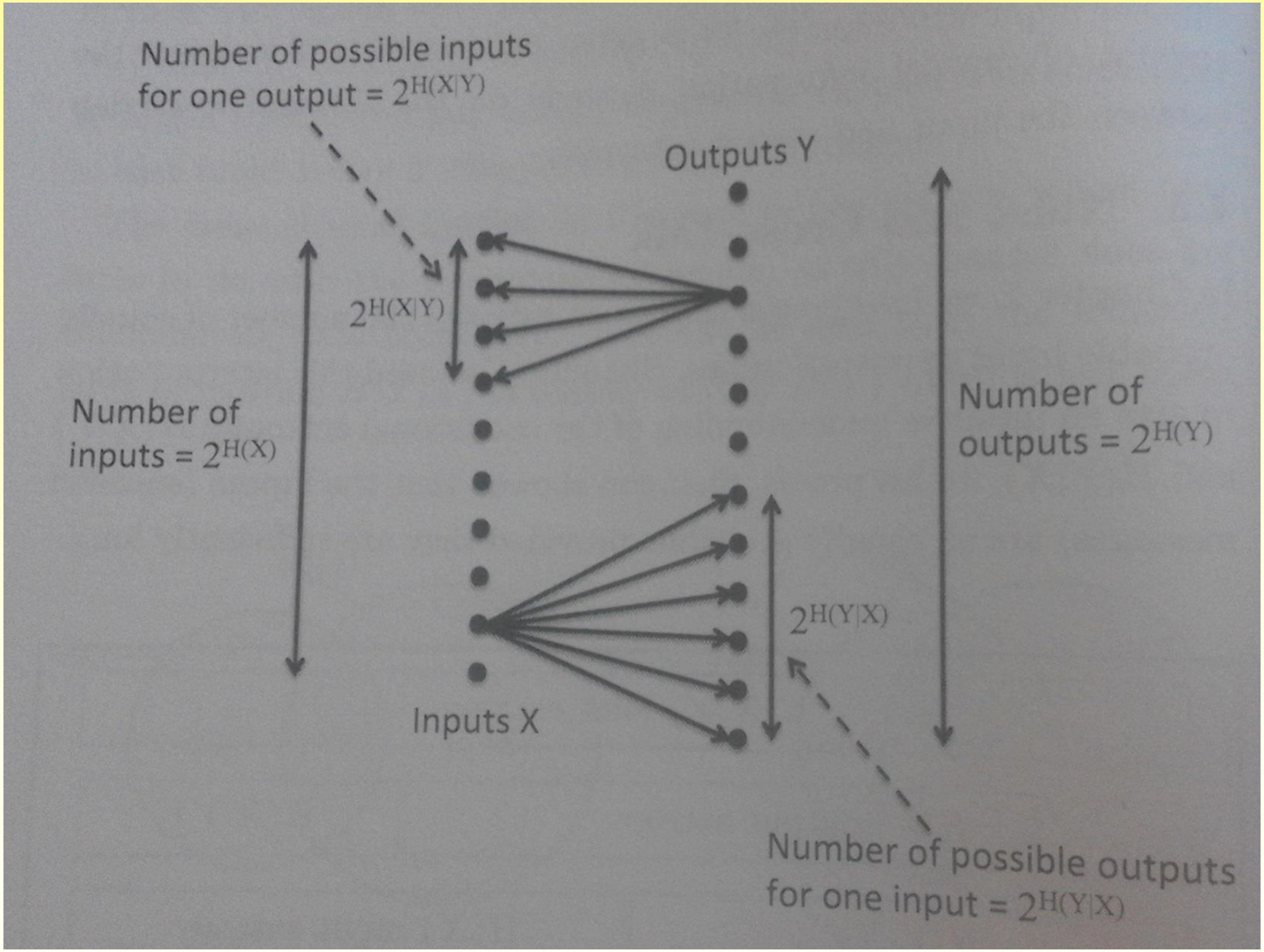
$$\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{Y}|\mathbf{X})$$

Conditional Entropy and Noise



$$\boldsymbol{I(X, Y)} = \mathbf{H(Y) - H(Y|X)}$$

Conditional Entropy and Noise



Noisy pictures

Input Output

In the Input picture 72.4% of the pixels are black (0) and the rest is white (1)

Because of noise 10% of the pixels in the output are changed

H(X) = p(0)log(1/p(0)) + p(1)log(1/p(1)) H(X) = 0.851 bits/pixel

H(Y) = p(0)log(1/p(0)) + p(1)log(1/p(1)) H(Y) = 0.906 bits/pixel

State	Input=0	Input=1
Output=0	$p(0,0) = 0.651$	$p(0,1) = 0.028$
Output=1	$p(1,0) = 0.073$	$p(1,1) = 0.249$

H(X,Y) = 1.32 bits/pixel

I(X,Y) = 0.437 bits/pixel

Noisy pictures

Input Output

In the Input picture 72.4% of the pixels are black (0) and the rest is white (1)

Because of noise 10% of the pixels in the output are changed

$H(Y|X) = H(Y) - I(X,Y) = 0.47$ bits

$H(\eta) = p\log(1/p) + (1-p)\log(1/(1-p))$

$H(\eta) = 0.1\log(1/0.1) + (0.9)\log(1/0.9)$

$H(\eta) = 0.469$ bits

$H(X) = 0.851$ bits/pixel

$H(Y) = 0.906$ bits/pixel

$H(X,Y) = 1.32$ bits/pixel

$I(X,Y) = 0.437$ bits/pixel

Noisy pictures

Input Output

In the Input picture 72.4% of the pixels are black (0) and the rest is white (1)

Because of noise 10% of the pixels in the output are changed

Transmission efficiency:

$$\frac{I(X,Y)}{H(Y)} = 0.481$$

H(X) = 0.851 bits/pixel

H(Y) = 0.906 bits/pixel

H(X,Y) = 1.32 bits/pixel

I(X,Y) = 0.437 bits/pixel

Error correcting codes

Input Output

Why do we still see the image,
even in the presence of a lot of noise?

The image has a lot of Redundancy

Redundancy could be used to correct for transmission errors

Error correcting codes

s = [1101001101011000]

Arrange in a grid:

$\mathbf{s} =$

1	1	0	1
0	0	1	1
0	1	0	1
1	0	0	0

Add binary digit for parity check

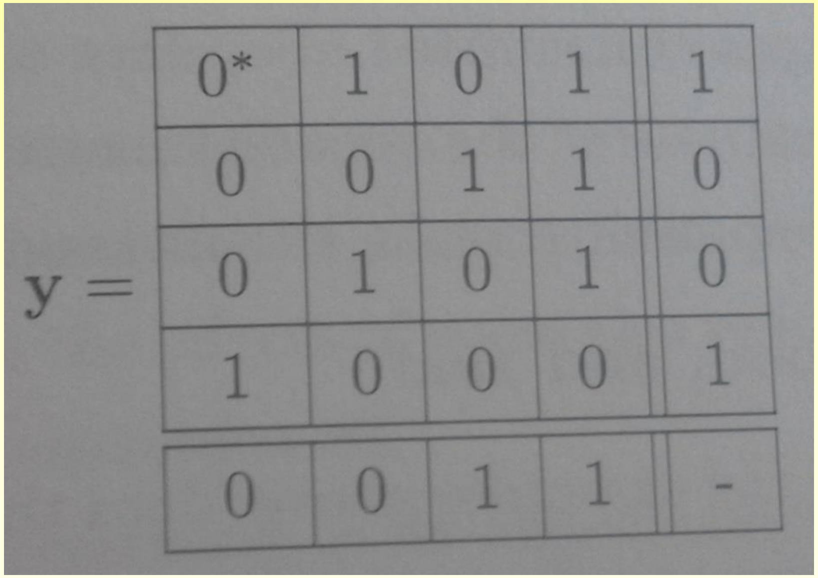
$\mathbf{x} =$

1	1	0	1	1
0	0	1	1	0
0	1	0	1	0
1	0	0	0	1
0	0	1	1	-

Error correcting codes

s = [1101001101011000]

Error detection:



y =

0*	1	0	1	1
0	0	1	1	0
0	1	0	1	0
1	0	0	0	1
0	0	1	1	-

This allows detecting 1 error in 4x4 + 2x4 binary digits

If using an nxn grid we can detect 1 error in $n^2 + 2n$ binary digits

This increases the number of binary digits in a factor:
$(n^2 + 2n)/n^2 = 1+2/n$ binary digits

The investment in parity digits improves with n, but allows correcting only 1 binary digit

Redundancy: Good and Bad

Input Output

Redundant data is more resistance to errors

but

large data processing to recover a small amount of original information

Capacity of a Noisy Channel

$$\boldsymbol{C} = \max_{\boldsymbol{p}(\boldsymbol{X})} \boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y})$$

$$\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}|\mathbf{Y})$$

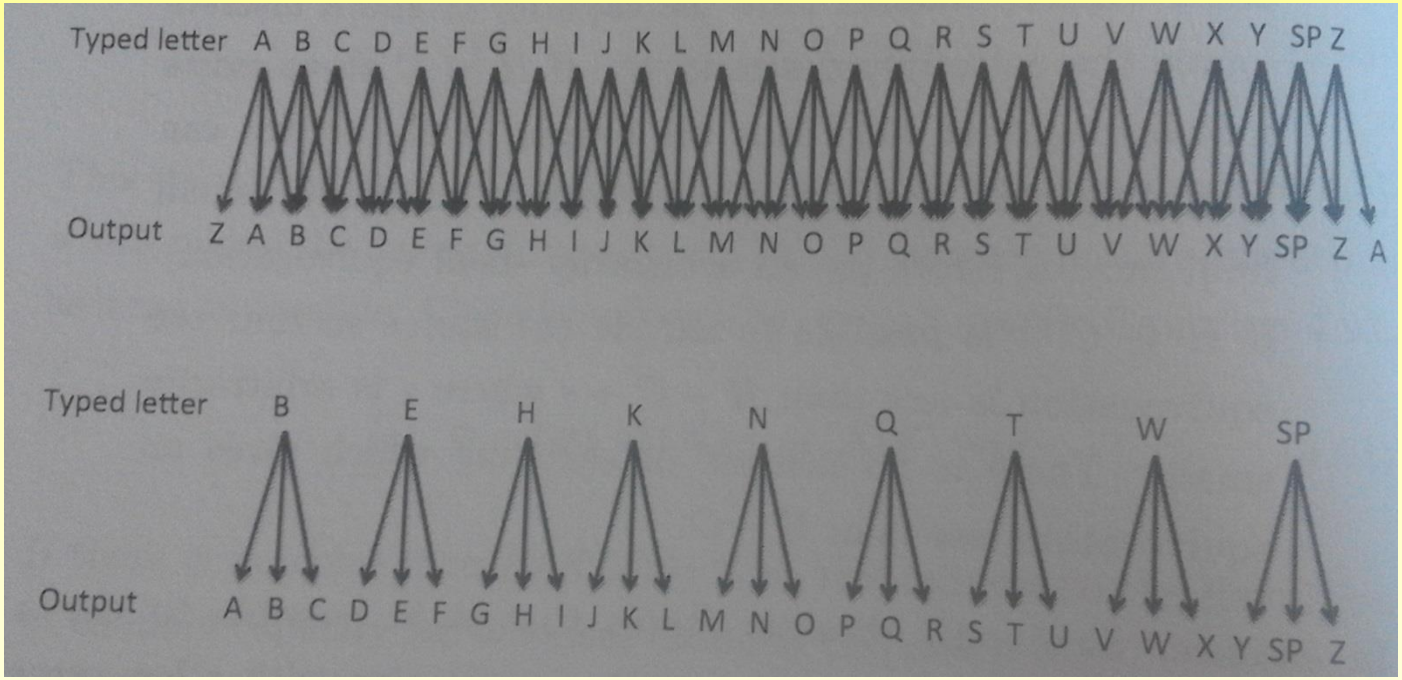
Shannon’s theorem for noisy channels

For a channel with capacity C and source of entropy H,

- If H < C there is a code allowing transmission with an arbitrarily small error
- If H > C there is a code allowing transmission with an error close to H – C

- It is possible to communicate information with a low error at a rate close to the channel capacity.
- It is not possible to communicate with no error at a rate higher than C

Noisy typewriter



XFZAEYXDU = WE BE WET

H(X) = log9 = 3.17 bits

H(Y) = log27 = 4.76 bits

H(Y|X) = log3 = 1.59 bits

I(X,Y) = H(Y) – H(Y|X) = 3.17 bits