



Information Theory

ELEC 3004: Systems: Signals & Controls

Dr. Surya Singh & Timothy Sherry

 $\begin{tabular}{ll} Lecture & 24 \\ (with material from $\underline{\rm D.\,MacKay}$, B. Lathi, $\underline{\rm K.\,Harris}$, $\underline{\rm D.\,Corrigan}$, and more!) \\ \end{tabular}$

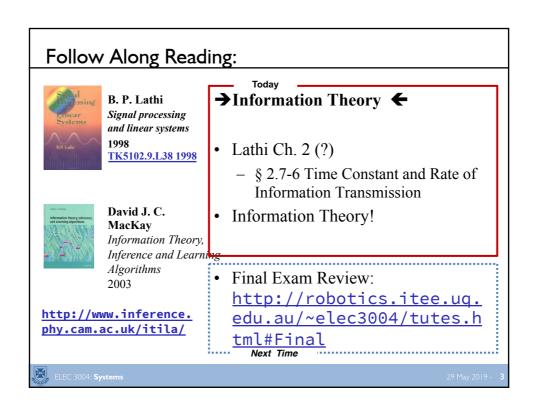
elec3004@itee.uq.edu.au

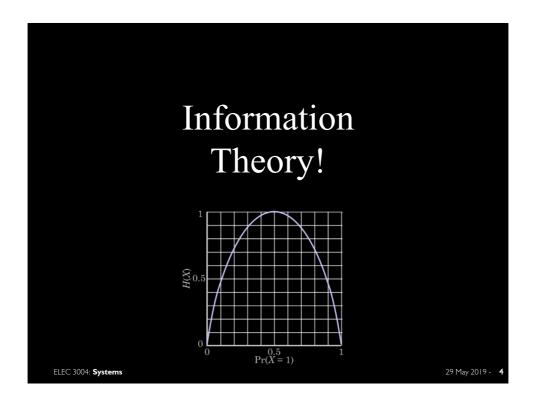
May 29, 2019

http://robotics.itee.uq.edu.au/~elec3004/

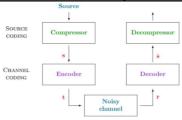
Lecture Schedule:

Week	Date	Lecture Title
1	27-Feb	Introduction
		Systems Overview
2		Systems as Maps & Signals as Vectors
		Systems: Linear Differential Systems
3		Sampling Theory & Data Acquisition
,		Aliasing & Antialiasing
4		Discrete Time Analysis & Z-Transform
		Second Order LTID (& Convolution Review)
5		Frequency Response
'		Filter Analysis
6	3-Apr	Digital Filters (IIR) & Filter Analysis
		PS 1: Q & A
7		Digital Windows
_ ′	12-Apr	Digital Filter (FIR)
8	17-Apr	Active Filters & Estimation
	19-Apr	
	24-Apr	Holiday
	26-Apr	
9		Introduction to Feedback Control
		Servoregulation & PID Control
10		State-Space Control
10		Guest Lecture: FFT
11		Advanced PID & & FFT Processes
		State Space Control System Design
12		Shaping the Dynamic Response
	24-May	Stability and Examples
13	29-May	System Identification & Information Theory & Information Space
	31-May	Summary and Course Review





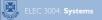
Shannon Information Theory



"The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point."

On the transmission of information over a noisy channel:

- An information source that produces a message
- A transmitter that operates on the message to create a <u>signal</u> which can be sent through a channel
- A **channel**, which is the medium over which the signal, carrying the information that composes the message, is sent
- A receiver, which transforms the signal back into the message intended for delivery
- A destination, which can be a person or a machine, for whom or which the message is intended



29 May 2019 -

Information theory is...

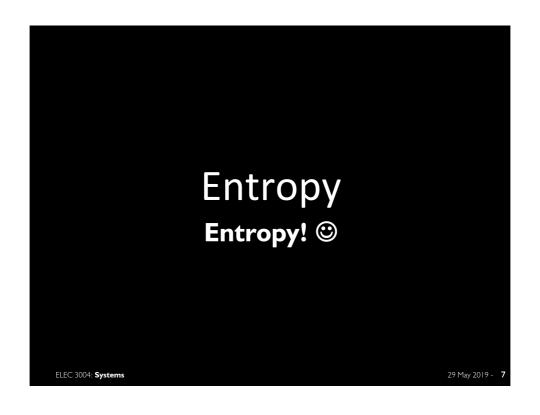
• It all starts with **Probability Theory!**

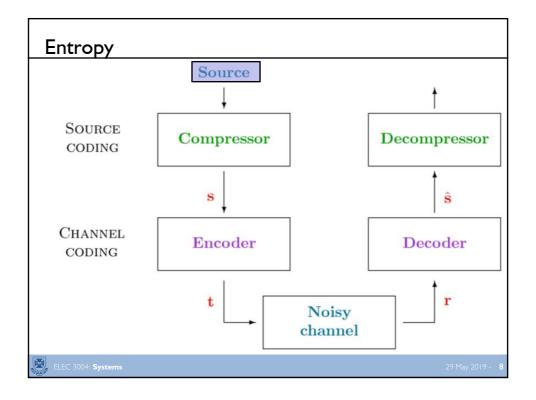


- (Wikipedia) Information theory is a branch of applied mathematics, electrical engineering, and computer science involving the quantification of information.
- 2. Information theory is **probability theory** where you take logs to base 2.



9 May 2019 ₋





Entropy - a measure of randomness

Entropy

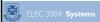
The entropy of a random variable X with a probability mass function p(x) is defined by

$$H(X) = -\sum_{x} p(x) \log_2 p(x).$$
 (1.1)

Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. We can calculate the entropy of the horse race as

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\frac{1}{64}\log\frac{1}{64}$$

= 2 bits. (1.3)



29 May 2019 -

Entropy

- Consider the Random Variable *X*, the outcome of each consecutive horse race.
- Assigning a binary number to each horse in the race would require $8 = 2^k \rightarrow k = 3 \ bits/symbol$
- Shannon states that some encoding scheme on the sequence $X_1, X_2 ... X_n$ can reduce the average number of bits -> H(X) = $2 \ bits/symbol$
- We will show you how soon

ELEC 3004: Systems

Connection to Physics

- A macrostate is a description of a system by large-scale quantities such as pressure, temperature, volume.
- A macrostate could correspond to many different microstates i, with probability p_i .
- Entropy of a macrostate is
- $S = -k_B \sum_i p_i \ln p_i$



29 May 2019 - 1

Entropy: The bent coin

Example 2.1.1 Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$
 (2.4)

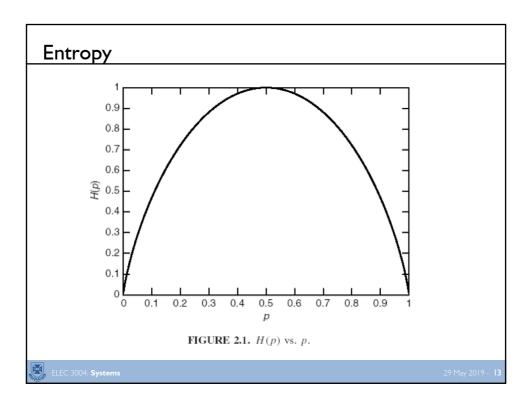
Then

$$H(X) = -p \log p - (1-p) \log(1-p) \stackrel{\text{def}}{=} H(p).$$
 (2.5)

- Consider X to be the result of a coin toss.
- This coin has been modified to land a Head (1) with probability p and Tails (0) with probability 1-p.
- What is the entropy of X as we vary p?

Plot a graph of H(p) against p.





Bent/Unfair Coin

- The most unpredictable coin is the fair coin.
- More generally for a variable of k states, a uniform distribution across all k states exhibits maximum entropy
- The observation of a coin with two tails is not random at all.

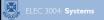


Example: Mystery Text

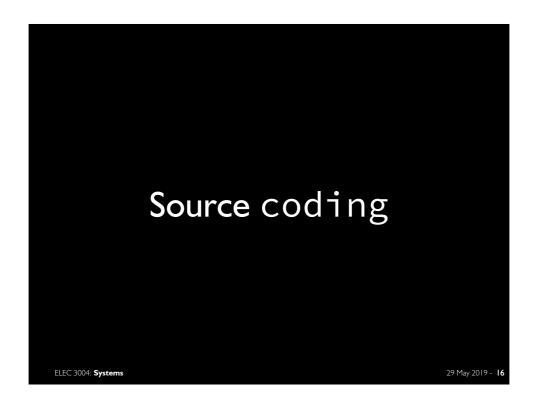
- II. Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty one years in the world with very little to distress or Vex her. She was the youngest of the two daughters of a most affectionate, indolent father; and had, in consequence of her sister's marriage, been mistress of his house from a very early period. Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses; and her place had been supplied by an excellent woman as governess, who had fallen little short of a mother in affection. Sixteen years had Miss Taylor been in Mr Woodhouse's family, less as a eoverness than a friend, very

Entropy of English - 2.6 Bits/letter || Raw encoding 27+4 = 31sym = 5.1 Bits/letter

Source: MacKay VideoLectures 02, Slide 5

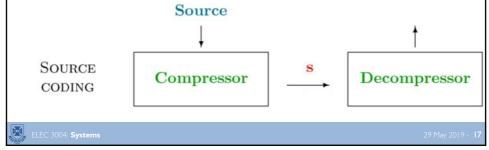


29 May 2019 - 1!



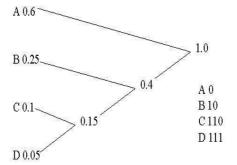
Data Compression/Source coding

- Removes redundancy to reduce bit length of messages
- Save bits on common symbols/sequences
- Spend extra bits on the 'surprises', as they very rarely occur
- Lossless (Huffman, Algorithmic, LZ, DEFLATE)
- Lossy (JPEG, MP3, H.265 etc)
- Lossy techniques exploit perceptual dynamic range to expend bits where a human is sensitive, and save where they're not



Huffman Coding

- Huffman is the simplest entropy coding scheme
 - It achieves average code lengths no more than 1 bit/symbol of the entropy
 - A binary tree is built by combining the two symbols with lowest probability into a dummy node
 - The code length for each symbol is the number of branches between the root and respective leaf

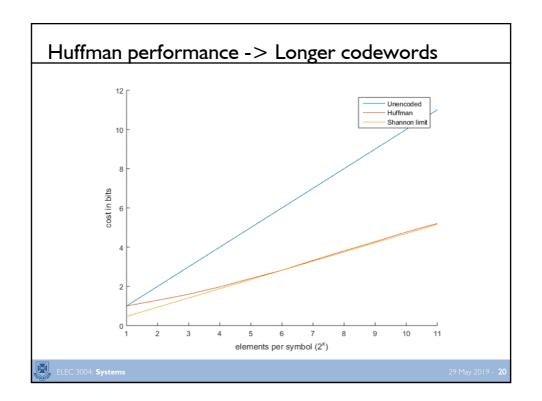




Ex: Bent/Unfair Coin

- Form symbols TT, HT, TH, HH
- P(TT) = 0.81; P(HT) = P(TH) = 0.09; P(HH) = 0.01
- $X_{TT} = 0 \mid p = 0.81$ $X_{TH} = 10 \mid p = 0.09$ $X_{HT} = 110 \mid p = 0.09$ $X_{HH} = 111 \mid p = 0.01$
- Transmission cost $0.81 \cdot 1 + 0.09 \cdot 2 + 0.09 \cdot 3 + 0.01 \cdot 3 = 1.29 \frac{bits}{symbol}$
- $H(X) = 0.4690 \cdot 2 = 0.938 \frac{Bits}{symbol}$ non-optimal code
- Huffman transmission cost $H(X) \le C < H(X) + 1 \frac{bit}{symbol}$
- longer code words = better performance

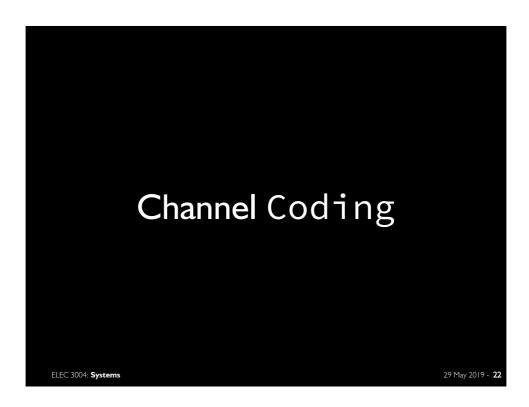
ELEC 3004: Systems

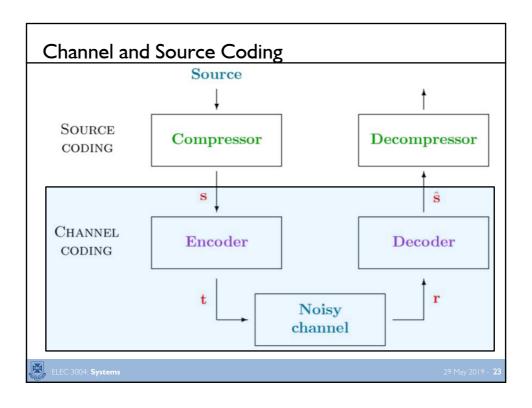


Lossless compression and entropy

- The lower the entropy of the source, the more we can gain via compression.
- The fair coin exhibits no exploitable means of **compressing** a sequence of observations from their direct description, without losing some information in the process
- All forms of compression exploit structure to reduce the average number of bits to describe a sequence.







Channel coding

- All physical transmission is susceptible to noise.
- Noise results in errors when a transmission is converted back to the source space
- Channel coding **introduces** redundancy to reduce the probability an error is made in conversion back to the source.

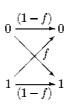


Noisy channels: Binary Symmetric Channel

$$x \xrightarrow{0} \xrightarrow{0} y \xrightarrow{P(y=0 \mid x=0)} = \xrightarrow{1-f}; \begin{array}{ccc} P(y=0 \mid x=1) & = & f; \\ P(y=1 \mid x=0) & = & f; \end{array} \xrightarrow{P(y=0 \mid x=1)} = \xrightarrow{1-f}.$$

- A 1 bit digital channel where each bit experiences a probability, *f*, of being misinterpreted (flipped)
- Ex: A File of N=10,000 bits is transmitted with **f=0.1**







ELEC 3004: Systems

29 May 2019 - **2!**

How many bits are flipped?

Assume a Binomial distribution

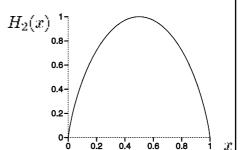
- Mean: $\mu = Np$
- Variance: $\sigma^2 = Npq$

Then:

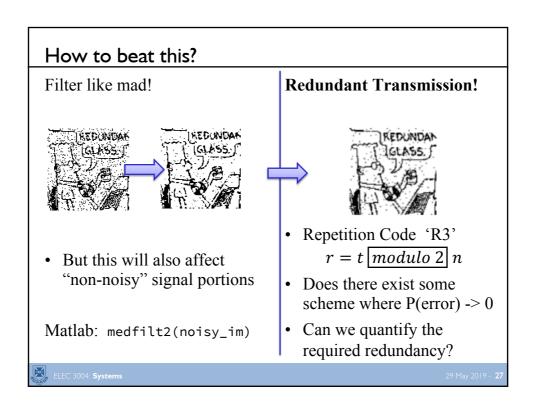
$$\mu = (0.1)10,000$$
 $\sigma^2 = (0.1)(0.9)10,000 = 900$

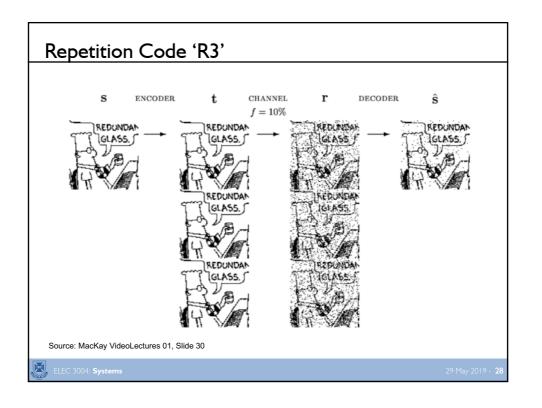
Thus:

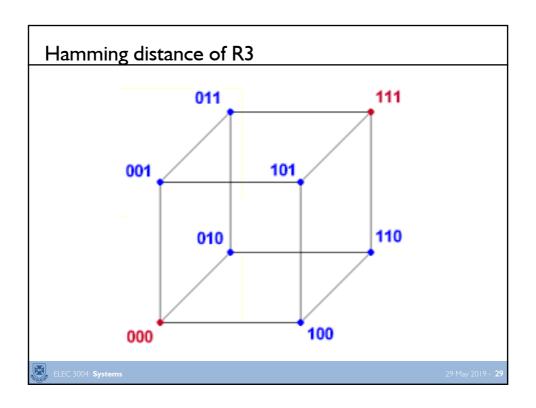
$$1000 \pm 30$$









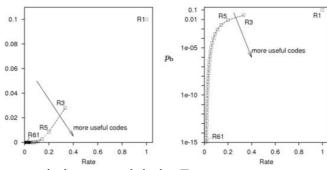


P_{error} of the R_3 code

- $P(r_{000}, s_1) = P(r_{111}, s_0) = 0.1^3 = 0.001$
- $P(r_{100}, s_1) = P(r_{010}, s_1) \dots = 0.1^2 = 0.01$
- $P_{err} = (0.001 + 0.01 \cdot 3) \cdot p(s_1) + (0.001 + 0.01 \cdot 3) \cdot p(s_0)$
- $P_{err} = 0.031$
- The number of errors $\mu = 310 \pm 17.3$
- Approximately 3x reduction in errors, but 3x as many bits required to send the message

ELEC 3004: Systems

Performance of Repetition Codes



- Want to maximise rate, minimise Error
- Software (e.g. Checksums) can identify errors and correct to higher stringency
- Can we do better?

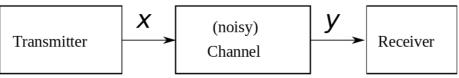
Source: MacKay VideoLectures 01, Slide 45



29 May 2019 - **3**

The Noisy-channel coding theorem

• The channel capacity for the general channel:



• Is the **mutual information** between X and Y:

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

• We will show Capacity of the BSC is:

$$Capacity_{BSC} = I(X_{BSC}; Y_{BSC}) = 1 - H(P_{BSC})$$

• Where H is the **Shannon entropy** of the bent coin with P_{BSC}



Conditional Entropy

- Suppose Alice wants to tell Bob the value of X
 - And they both know the value of a second variable Y.
- Now the optimal code depends on the conditional distribution p(X|Y)
- Code length for X = i has length $-\log_2 p(X = i|Y)$
- Conditional entropy measures average code length when they know Y

$$H(X|Y) = -\sum_{X,Y} p(X,Y) \log_2 p(X|Y)$$



29 May 2019 - **3**

Mutual information

 How many bits do Alice and Bob save when they both know Y?

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

$$= \sum_{X,Y} p(X,Y)(-\log_2 p(X) + \log_2 p(X|Y))$$

$$= \sum_{X,Y} p(X,Y) \log_2 \left(\frac{p(X,Y)}{p(X)p(Y)}\right)$$

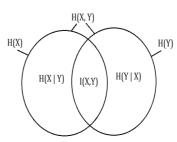
- Symmetrical in X and Y!
- Amount saved in transmitting X if you know Y equals amount saved transmitting Y if you know X.



Properties of Mutual Information

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

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



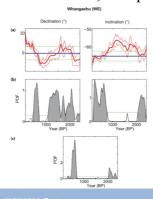
- If X = Y, I(X; Y) = H(X) = H(Y)
- If X and Y are independent, I(X;Y) = 0 $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) H(X|Y) \le H(X)$
- In straightforward terms, information is never negative, we are always at least as certain about X|Y vs X

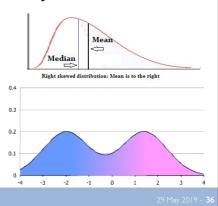


29 May 2019 - 3!

Conditional Entropy & Mutual Information

- Note that the distributions p(X), p(Y) and p(X|Y) may be complex.
- computing p(X|Y) may be intractable
- Mutual Information and Conditional entropy hold for all distributions, but quantifying them may be intractable





Mutual information of the $BSC_{p=0.1}$

$$P(X_0, Y_0) = P(X_1, Y_1) = 0.5 * 0.9 = 0.45$$

 $P(X_0, Y_1) = P(X_1, Y_0) = 0.5 * 0.1 = 0.05$
 $P(X_0) = P(X_1) = P(Y_0) = P(Y_1) = 0.5$

$$I(X;Y) = \sum_{X,Y} p(X,Y) \log_2 \left(\frac{p(X,Y)}{p(X)p(Y)} \right)$$

$$I(Y;Y) = 2 \cdot 0.45 \log_2 \left(\frac{0.45}{0.5^2} \right) + 2 \cdot 0.05 \log_2 \left(\frac{0.05}{0.5^2} \right)$$

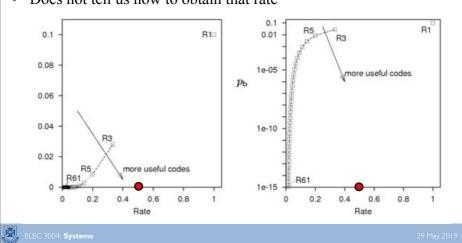
$$I(Y;Y) = 0.7632 - 0.2322 = 0.531$$

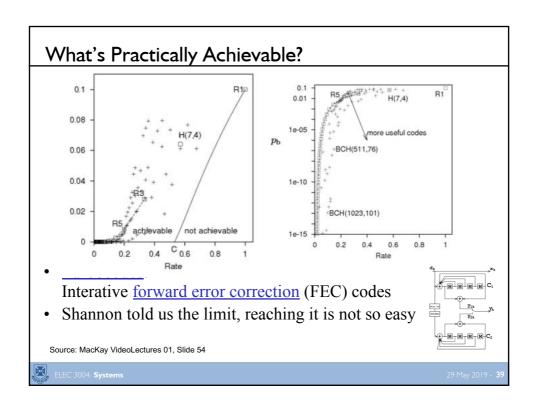
ELEC 3004: Systems

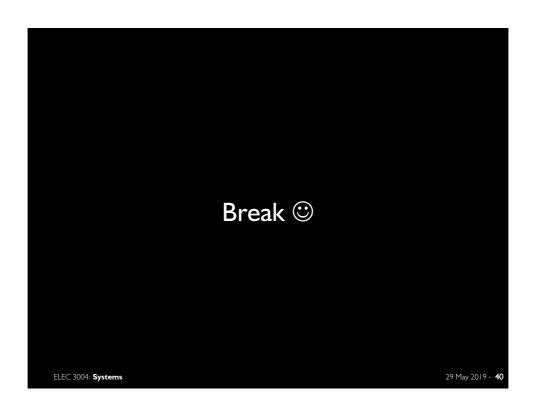
29 May 2019 - **37**

Ultimate error free performance

- H(0.1) = 0.4690
- Capacity = 0.531 Bits per Symbol
- Theoretically can obtain a $p_b = 0$ at that rate
- Does not tell us how to obtain that rate







Communications LEC 3004: Systems 29 May 2019 - 41

Physical layer – Modulation Schemes

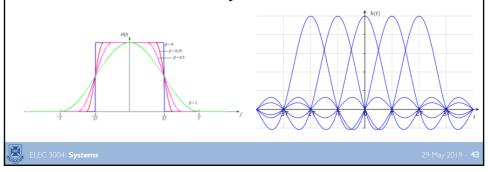
- The final piece to transmitting over real communication channels
- Analog modulation (AM (SSB),FM, PM, QAM)
- Digital modulation(ASK,PSK,FSK,QAM)
- Both encode the signal on the amplitude (A) or phase (F,P,Q) of a carrier sine frequency
- Will only cover digital Quadrature Amplitude Modulation (QAM)

ELEC 3004: Systems

.9 May 2019 - **42**

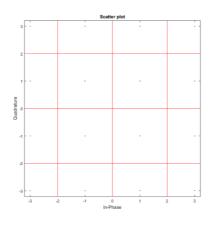
QAM

- A pulse of the carrier frequency encompasses a symbol.
- The symbol can take on a number of states, encoded as finite levels of amplitude and phase (quadrature)
- Here we need to window our sine carrier to form a pulse
- A rectangular window would produce spectral leakage, and heavy interference use a raised cosine window
- A RCW results in no inter-symbol-interference



Constellations

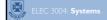
- For QAM 16 16 discrete locations on the inphase / quadrature plane
- Red lines denote decision boundary

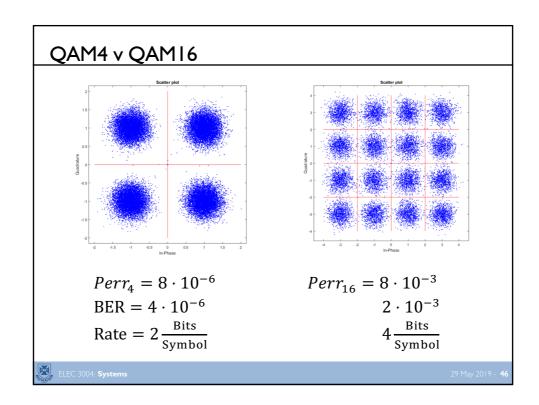


ELEC 3004: Systems

Doubling data rate - QAM4 to QAM16

- Bits per symbol of QAM4 = 2, QAM16 = 4
- Constellation diagrams
- Assuming receive power of $\frac{Eb}{No} = 10dB$, what is the probability that a message sent crosses the descision boundary?
 - $-\frac{Eb}{No}$ is the energy per bit received, over the average noise energy at the receiver.





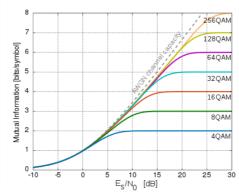
QAM4 v QAM16

- Double the data rate
- 500× more likely to flip a bit!
- Generally BER is a specification of the communication protocol
- For example Gigabit Ethernet adjusts the constellation to maintain a raw BER below 10⁻¹⁰ (one per 10 gigabits)
 - In other words, your Ethernet connection will produce \sim 7-8 one bit errors per hour of 4k Netflix streaming (\sim 7GB/Hr)
- Source coding reduces this to 10^{-14} (one per 100 terabits)

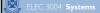


29 May 2019 - **47**

QAM vs Theoretical maximum



- Using a smaller constellation restricts the channel capacity
- Using too high a constellation requires complex channel coding to exploit



Shannon Capacity of the AWGN channel

- $C = B \log_2(1 + \frac{S}{N})$
- B is the bandwidth S the signal power in real terms, N noise in real terms, and C the capacity in bits.
- Upper bound on error free transmission
- By intelligently selecting QAM constellation and Channel coding schema in tandem capacity can be maximised.
- Requires some coordination between transmitter and sender

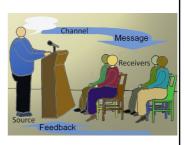


29 May 2019 - **4**9

Shannon and Weaver: Models of Communication

Three "problems" in Communication:

- The technical problem: how accurately can the message be transmitted?
- The <u>semantic</u> problem: how precisely is the meaning "conveyed"?
- The effectiveness problem: how effectively does the received meaning affect behaviour?



Source: http://en.wikipedia.org/wiki/File:Transactional_comm_model.jpg

ELEC 3004: Systems

What is the "optimal code"?

- X is a random variable
- Alice wants to tell Bob the value of X (repeatedly)
- What is the best binary code to use?
- How many bits does it take (on average) to transmit the value of X?



Optimal code lengths

- In the optimal code, the word for X = i has length
- $\log_2 \frac{1}{p(X=i)} = -\log_2 p(X=i)$
- For example:

Value of X	Probability	Code word
A	1/2	0
В	1/4	10
C	1/4	11

- ABAACACBAACB coded as 010001101110001110
- If code length is not an integer, transmit many letters together



29 May 2019 - **5**3

Kullback-Leibler divergence

- Measures the difference between two probability distributions
 (Mutual information was between two random variables)
- Suppose you use the wrong code for *X*. How many bits do you waste?

 Length of Length of optimal

codeword

codeword

• $D_{KL}(p||q) = \sum_{x} p(x) \left[\log_2 \frac{1}{q(x)} - \log_2 \frac{1}{p(x)} \right]$ = $\sum_{x} p(x) \log_2 \frac{p(x)}{q(x)}$

- $D_{KL}(p||q) \ge 0$, with equality when p and q are the same.
- $I(X;Y) = D_{KL}(p(x,y)||p(x)p(y))$

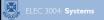


Continuous variables

- X uniformly distributed between 0 and 1.
- How many bits required to encode X to given accuracy?

Decimal places	Entropy	
1	3.3219	
2	6.6439	
3	9.9658	
4	13.2877	
5	16.6096	
Infinity	Infinity	

• Can we make any use of information theory for continuous variables?



29 May 2019 - 55

K-L divergence for continuous variables

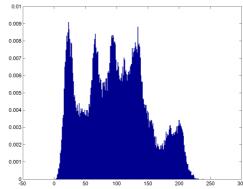
- Even though entropy is infinite, K-L divergence is usually finite.
- Message lengths using optimal and non-optimal codes both tend to infinity as you have more accuracy. But their difference converges to a fixed number.

$$\sum_{x} p(x) \log_2 \frac{p(x)}{q(x)} \to \int p(x) \log_2 \frac{p(x)}{q(x)} dx$$









The entropy of lena is = 7.57 bits/pixel approx



20 May 2010 **57**

Huffman Coding of Lenna

Symbol	Code Length
0	42
1	42
2	41
3	17
4	14

Average Code Word Length = $\sum_{k=0}^{255} p_k l_k = 7.59 \ bits/pixel$

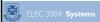
So the code length is not much greater than the entropy

ELEC 3004: Systems

But this is not very good

- Why?
 - Entropy is not the minimum average codeword length for a source with memory
 - If the other pixel values are known we can predict the unknown pixel with much greater certainty and hence the effective (ie. conditional) entropy is much less.
- Entropy Rate
 - The minimum average codeword length for any source.
 - It is defined as

$$H(\chi) = \lim_{n \to \infty} \frac{1}{N} H(X_1, X_2, ..., X_n)$$



29 May 2019 - **5**9

Coding Sources with Memory

- It is very difficult to achieve codeword lengths close to the entropy rate
 - In fact it is difficult to calculate the entropy rate itself $P(X_1|X_2...X_n)$ is described in Rⁿ Space for lenna n = 65536
- We looked at LZW as a practical coding algorithm
 - Average codeword length tends to the entropy rate if the file is large enough
 - Efficiency is improved if we use Huffman to encode the output of LZW
 - LZ algorithms used in lossless compression formats (eg. .tiff, .png, .gif, .zip, .gz, .rar...)

ELEC 3004: Systems

Efficiency of Lossless Compression



- Lenna (256x256) file sizes
 - Uncompressed tiff 64.2 kB
 - LZW tiff 69.0 kB
 - Deflate (LZ77 + Huff) 58 kB



- Green Screen (1920 x 1080) file sizes
 - Uncompressed 5.93 MB
 - LZW 4.85 MB
 - Deflate 3.7 MB



29 May 2019 - **6**

Differential Coding

• Key idea – code the differences in intensity.

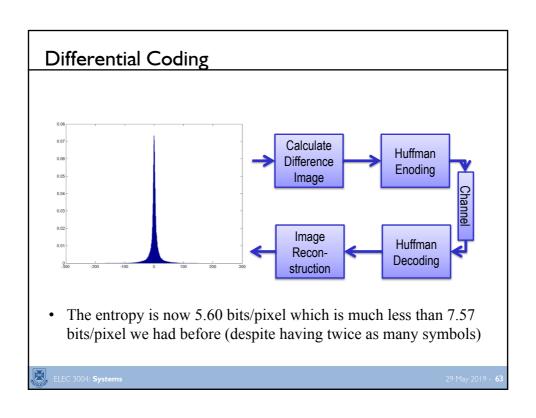


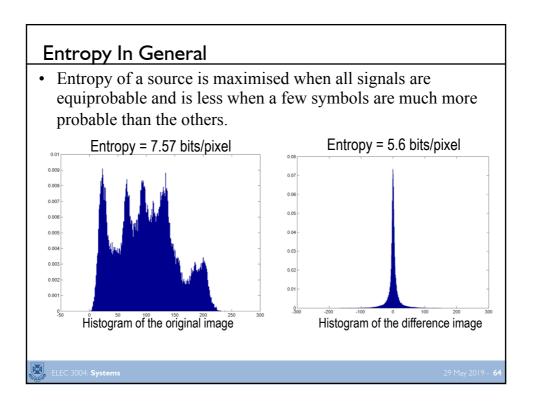


G(x,y) = I(x,y) - I(x-1,y)



.9 May 2019 - **62**





Lossy Compression

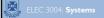
- But this is still not enough compression
 - Trick is to throw away data that has the least perceptual significance



Effective bit rate = 8 bits/pixel



Effective bit rate = 1 bit/pixel (approx)



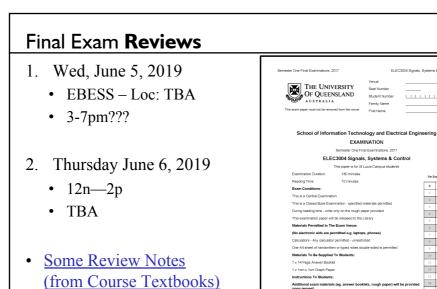
10 Maii 2010 - 4E

Next Time...



- Exam Review!!
- Review:
 - Chapter 6 of FPW
 - Chapter 13 of Lathi
- Deeper Pondering??





Fun Fact: With this, 42% of the exam is already "public" ©

→ http://robotics.itee.uq.edu.au/

~elec3004/tutes.html

ELEC 3004: Systems