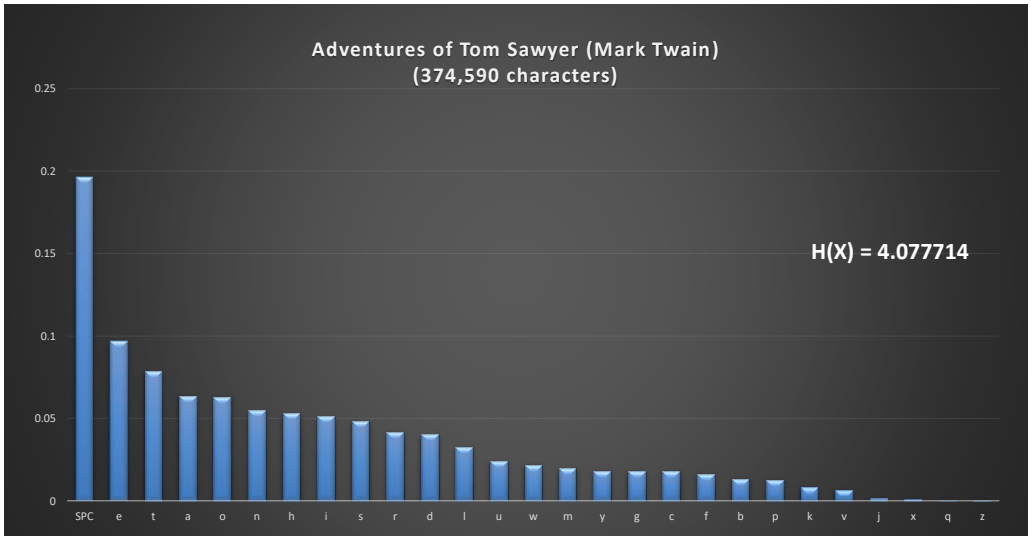
Introduction to Information Theory

Part 4-A

Assignment#2 Results



Quick Code Review

- Using dictionaries in Python
- Using the Counter class

Assignment#2: Huffman Encoding

Symbol	Probability	Code
A	0.75	0
B	0.25	1
	H=0.8118	
	L=1	

Assignment#2: Huffman Encoding

Symbol	Probability	Code
A	0.75	0
B	0.25	1
	H=0.8118	
	L=1	

Symbol	Probability	Code
AA	0.5625	0
AB	0.1875	10
BA	0.1875	110
BB	0.0625	111
	L/2=0.84375	

Assignment#2: Huffman Encoding

Symbol	Probability	Code
A	0.75	0
B	0.25	1
	H=0.8118	
	L=1	

Symbol	Probability	Code
AA	0.5625	0
AB	0.1875	10
BA	0.1875	110
BB	0.0625	111
	L/2=0.84375	

Symbol	Probability	Code
AAA	0.421875	1
AAB	0.140625	001
ABA	0.140625	010
BAA	0.140625	100
ABB	0.046875	00000
BAB	0.046875	00001
BBA	0.046875	00010
BBB	0.015625	00011
	L/3 = 0.8229167	

Assignment#2: Huffman Encoding

Symbol	Probability	Code
A	0.75	0
B	0.25	1
	H=0.8118	
	L=1	

Symbol	Probability	Code
AA	0.5625	0
AB	0.1875	10
BA	0.1875	110
BB	0.0625	111
	L/2=0.84375	

Symbol	Probability	Code
AAA	0.421875	1
AAB	0.140625	001
ABA	0.140625	010
BAA	0.140625	100
ABB	0.046875	00000
BAB	0.046875	00001
BBA	0.046875	00010
BBB	0.015625	00011
	L/3 = 0.8229167	

Notice that as we increase the length of symbols, entropy/letter approaches 0.8118.

Lempel-Ziv Coding

- Sequences of text repeat patterns (words, phrases, etc)
- Construct a dictionary of common patterns
- Send references to patterns as triples (x, y, z)

Lempel-Ziv Coding (LZ77)

Message	Search Buffer	Look-Ahead Buffer
		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message	Search Buffer	Look-Ahead Buffer
		THIS-THESIS-IS-THE-THESIS.
0 0 T T		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.
0	0	I	I		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.
0	0	I	I		THIS-THESIS-IS-THE-THESIS.
0	0	S	S		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.
0	0	I	I		THIS-THESIS-IS-THE-THESIS.
0	0	S	S		THIS-THESIS-IS-THE-THESIS.
0	0	-	-		THIS-THESIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.
0	0	I	I		THIS-THESIS-IS-THE-THESIS.
0	0	S	S		THIS-THESIS-IS-THE-THESIS.
0	0	-	-		THIS-THESIS-IS-THE-THESIS.
5	2	E	E	THIS-THE	SIS-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H		THIS-THESIS-IS-THE-THESIS.
0	0	I	I		THIS-THESIS-IS-THE-THESIS.
0	0	S	S		THIS-THESIS-IS-THE-THESIS.
0	0	-	-		THIS-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.
0	0	I	I	TH	IS-THESIS-IS-THE-THESIS.
0	0	S	S	THI	S-THESIS-IS-THE-THESIS.
0	0	-	-	THIS	-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-I	S-THE-THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.
0	0	I	I	TH	IS-THESIS-IS-THE-THESIS.
0	0	S	S	THI	S-THESIS-IS-THE-THESIS.
0	0	-	-	THIS	-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-THE	SIS-IS-THE-THESIS.
5	1	I	SI	THIS-THESI	S-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-I	S-THE-THESIS.
10	5	-	S-THE-	THIS-THESIS-IS-THE-	THESIS.

Lempel-Ziv Coding (LZ77)

Message				Search Buffer	Look-Ahead Buffer
					THIS-THESIS-IS-THE-THESIS.
0	0	T	T		THIS-THESIS-IS-THE-THESIS.
0	0	H	H	T	HIS-THESIS-IS-THE-THESIS.
0	0	I	I	TH	IS-THESIS-IS-THE-THESIS.
0	0	S	S	THI	S-THESIS-IS-THE-THESIS.
0	0	-	-	THIS	-THESIS-IS-THE-THESIS.
5	2	E	THE	THIS-TH	ESIS-IS-THE-THESIS.
5	1	I	SI	THIS-THES	IS-IS-THE-THESIS.
7	2	I	S-I	THIS-THESIS-	IS-THE-THESIS.
10	5	-	S-THE-	THIS-THESIS-IS-TH	E-THESIS.
14	6	.	THESIS.	THIS-THESIS-IS-THE	-THESIS.

Lempel-Ziv Coding

- Sequences of text repeat patterns (words, phrases, etc)
- Construct a dictionary of common patterns
- Send references to patterns as triples (x, y, z)
  e.g. $(5, 3, F)$
  go back 5 received chars
  take the next 3 from there
  add F to the end
- Size of Search Buffer and Look-Ahead Buffer is finite.
- Used by ZIP, PKSip, Lharc, PNG, gzip, ARJ
- Extended to LZ78 (uses dictionary), LZW (+Terry Welch)
- Achieves optimal rate of transmission in the long run w/o using probability dist.

Decode

Message

0	0	I	
0	0	-	
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	
1	1	-	
5	5	L	
5	3	Y	

Decode

			Message
0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	
5	5	L	
5	3	Y	

Decode

			Message
0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	
5	3	Y	

Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	I - M I S S - M I S S - L
5	3	Y	

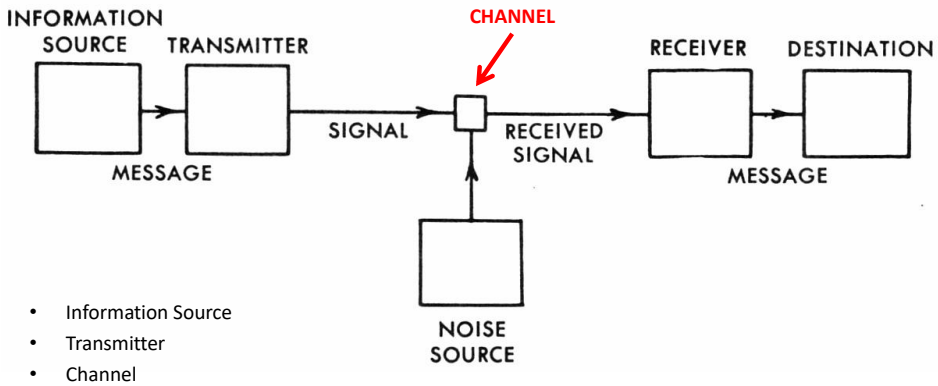
Decode

Message

0	0	I	I
0	0	-	I -
0	0	M	I - M
3	1	S	I - M I S
1	1	-	I - M I S S -
5	5	L	I - M I S S - M I S S - L
5	3	Y	I - M I S S - M I S S - L I S S Y

Lempel-Ziv Compression: Class Exercise

A General Communication System

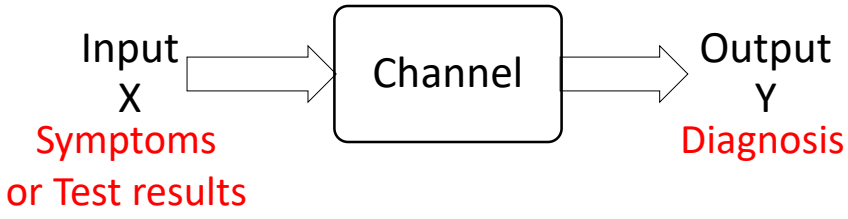


- Information Source
- Transmitter
- Channel
- Receiver
- Destination

Information Channel

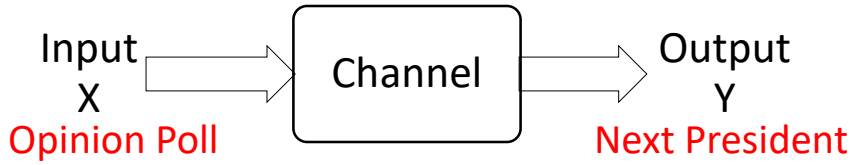
Information Channel

Information Channel

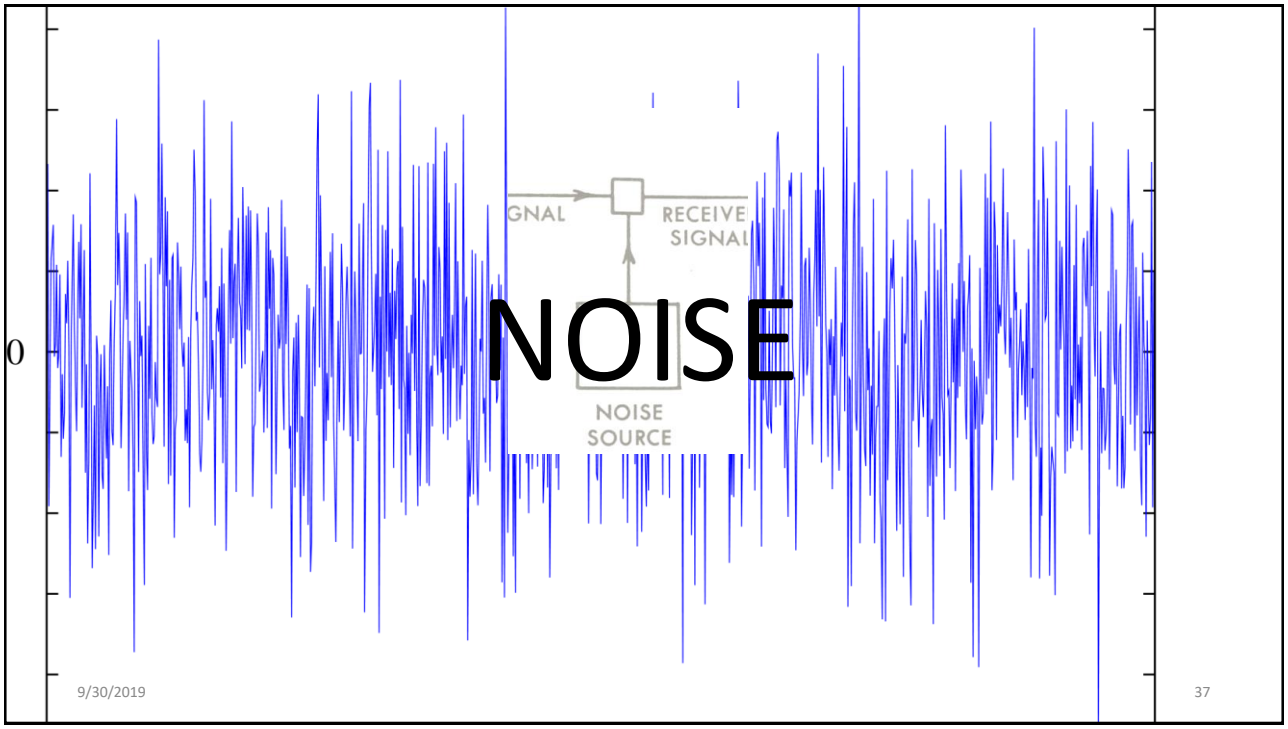


Information Channel

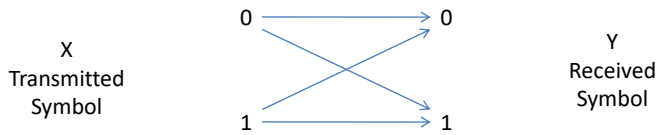
Information Channel



Perfect Communication (*Discrete Noiseless Channel*)

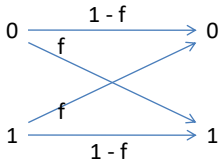


Motivating Noise…



Motivating Noise…

f = 0.1, n = ~10,000



Motivating Noise…

Message: $5213.75
Received: $5293.75

1. Detect that an error has occurred.
2. Correct the error.
3. Watch out for the overhead.

Error Detection by Repetition

In the presence of 20% noise…

Message : $ 5 2 1 3 . 7 5

Transmission 1: $ 5 2 9 3 . 7 5
Transmission 2: $ 5 2 1 3 . 7 5
Transmission 3: $ 5 2 1 3 . 1 1
Transmission 4: $ 5 4 4 3 . 7 5
Transmission 5: $ 7 2 1 8 . 7 5

There is no way of knowing where the errors are.

Error Detection by Repetition

In the presence of 20% noise…

Message : $ 5 2 1 3 . 7 5

Transmission 1: $ 5 2 9 3 . 7 5
Transmission 2: $ 5 2 1 3 . 7 5
Transmission 3: $ 5 2 1 3 . 1 1
Transmission 4: $ 5 4 4 3 . 7 5
Transmission 5: $ 7 2 1 8 . 7 5
Most common: $ 5 2 1 3 . 7 5

1. Guesswork is involved.
2. There is overhead.

Error Detection by Repetition

In the presence of 50% noise…

Message : $ 5 2 1 3 . 7 5

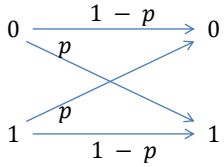
…
Repeat 1000 times!

1. Guesswork is involved.
   But it will almost never be wrong!

2. There is overhead.
   A LOT of it!

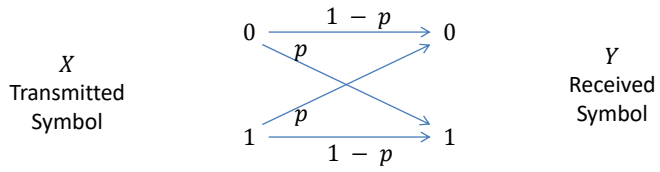
Binary Symmetric Channel (BSC)
(Discrete Memoryless Channel)

X
Transmitted
Symbol



Y
Received
Symbol

Binary Symmetric Channel (BSC) (Discrete Memoryless Channel)

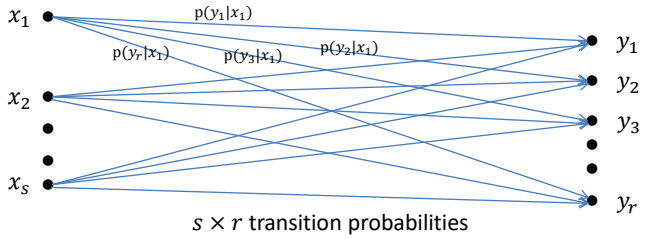


Defined by a set of **conditional probabilities** (*aka* **transitional probabilities**)

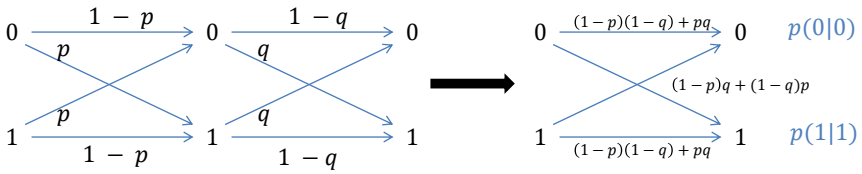
$$p(y|x) for\ all\ x \in X\ and\ y\ \in Y$$

The probability of y occurring at the output when x is the input to the channel.

A General Discrete Channel



Channel With Internal Structure



References

- Eugene Chiu, Jocelyn Lin, Brok Mcferron, Noshirwan Petigara, Satwiksai Seshasai: *Mathematical Theory of Claude Shannon: A study of the style and context of his work up to the genesis of information theory. MIT* 6.933J / STS.420J The Structure of Engineering Revolutions
- Luciano Floridi, 2010: *Information: A Very Short Introduction*, Oxford University Press, 2011.
- Luciano Floridi, 2011: *The Philosophy of Information*, Oxford University Press, 2011.
- James Gleick, 2011: *The Information: A History, A Theory, A Flood*, Pantheon Books, 2011.
- Zhandong Liu , Santosh S Venkatesh and Carlo C Maley, 2008: *Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples*, *BMC Genomics* 2008, **9:**509
- David Luenberger, 2006: *Information Science*, Princeton University Press, 2006.
- David J.C. MacKay, 2003: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
- Claude Shannon & Warren Weaver, 1949: *The Mathematical Theory of Communication*, University of Illinois Press, 1949.
- W. N. Francis and H. Kucera: *Brown University Standard Corpus of Present-Day American English*, Brown University, 1967.
- Edward L. Glaeser: A Tale of Many Cities, New York Times, April 10, 2010. Available at: http://economix.blogs.nytimes.com/2010/04/20/a-tale-of-many-cities/
- Alan Rimm-Kaufman, The Long Tail of Search. Search Engine Land Website, September 18, 2007. Available at: http://searchengineland.com/the-long-tail-of-search-12198