

# METU EE533 INFORMATION THEORY - LECTURE NOTES

ALPER YAZAR

This lecture notes set is based on my personal notes taken for E533 Information Theory course given at METU Electrical and Electronics Engineering Department, Spring 2015 by Prof. Ali Ozgur Yilmaz. I also used second edition of Elements of Information Theory by T. Cover and J. Thomas as a reference book.

This document is published on public domain with permission from Prof. Ali Ozgur Yilmaz with **CC BY-NC-SA 4.0** license.

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

Please consider the license if you use that document for your own purposes.

Please note that this not the official lecture note of this course. These notes are my personal notes and PLEASE ALWAYS BE SUSPICIOUS ABOUT CORRECTNESS. I really appreciate if you find a mistake and contact me for correction. Please feel free to contact me about corrections, comments, etc.

This document is served at the following URL officially.

<http://www.alperyazar.com/r/EE533Notes>

For my other lecture notes available, please check the following page.

<http://www.alperyazar.com/outputs>

Contact: [alperyazar@gmail.com](mailto:alperyazar@gmail.com)

Version:1, 20170225

## CONTENTS

<b>Part 1. Lecture 01 - 23.02.2016</b>	<b>5</b>
1. Introduction	5
<b>Part 2. Lecture 02 - 25.02.2016</b>	<b>7</b>
2. What is Entropy?	7
3. Joint Entropy	9
4. Conditional Entropy	10
5. Chain Rule	11
<b>Part 3. Lecture 03 - 01.03.2016</b>	<b>12</b>
6. Divergence	12
7. Divergence Inequality	13
8. Convex Function	14
9. Jensen's Inequality and Its Proof	15
10. Proof of Divergence Inequality	17
11. Mutual Information	18
12. Entropy and Mutual Information	19
13. Chain Rule for Entropy	20

---

Date: Spring 2015.

14. Conditional Mutual Information	20
15. Chain Rule for Information	20
<b>Part 4. Lecture 04 - 03.03.2016</b>	<b>21</b>
16. Upper Bound for $H(X)$	21
17. Conditioning Reduces Entropy	22
18. Independence Bound on Entropy	22
19. Data-Processing Inequality	23
20. Markov Chain	23
<b>Part 5. Lecture 05 - 08.03.2016</b>	<b>24</b>
21. Fano's Inequality	25
22. Log Sum Inequality	26
23. The Asymptotic Equipartition Property	27
24. Weak Law of Large Numbers (WLLN)	27
<b>Part 6. Lecture 06 - 10.03.2016</b>	<b>28</b>
25. The Law of Frequencies	29
26. Typical Sets	30
<b>Part 7. Lecture 07 - 15.03.2016</b>	<b>30</b>
27. Typical Set	31
28. Shannon - McMillan Theorem	32
29. Negative Statement of AEP	33
30. AEP and Data Compression	33
<b>Part 8. Lecture 08 - 17.03.2016</b>	<b>35</b>
31. Source Coding (Data Compression)	35
32. Kraft's Inequality	37
<b>Part 9. Lecture 09 - 22.03.2016</b>	<b>37</b>
33. Noiseless Coding Theorem	39
34. Bounds on Optimal Code Lengths	41
<b>Part 10. Lecture 10 - 24.03.2016</b>	<b>43</b>
35. Huffman Codes	44
36. Channel Capacity	44
<b>Part 11. Lecture 11 - 29.03.2016</b>	<b>45</b>
37. Example of Channels and Their Capacities	46
37.1. Noiseless Binary Channel	46
37.2. Noisy Channel with Non-overlapping Outputs	47
37.3. Noisy Channel with Overlapping Outputs	48
37.4. Binary Symmetric Channel (BSC)	49
37.5. Binary Erasure Channel	49
<b>Part 12. Lecture 12 - 31.03.2016</b>	<b>50</b>

<b>Part 13. Lecture 13 - 05.04.2016</b>	<b>51</b>
38. Properties of Channel Capacity	52
39. Preview and Definitions	52
39.1. Channel Code	52
39.2. Conditional Probability of Error	52
39.3. Decoding Region	53
39.4. Maximal Probability of Error	53
<b>Part 14. Lecture 14 - 07.04.2016</b>	<b>53</b>
39.5. Average Probability of Error	53
39.6. Code Rate, Achievable Rate and Operational Capacity	54
40. Jointly Typical Sequences	55
<b>Part 15. Lecture 15 - 12.04.2016</b>	<b>56</b>
40.1. Proofs	56
41. The Channel Coding Theorem (Fundamental Theorem in Information Theory)	58
42. Proof of The Channel Coding Theorem	58
42.1. Outline of The Proof	58
42.2. Encoding and Decoding in The Proof	59
<b>Part 16. Lecture 16 - 14.04.2016</b>	<b>60</b>
42.3. Final Interpretation	62
<b>Part 17. Lecture 17 - 19.04.2016</b>	<b>63</b>
43. Fano's Inequality and The Converse	63
43.1. Lemma A (Another form of Fano's Inequality)	63
43.2. Lemma B	63
44. Proof of The Converse	64
45. Properties of Good Codes	65
46. The Joint Source-Channel Coding Theorem	66
<b>Part 18. No Lecture on 21.04.2016</b>	<b>67</b>
<b>Part 19. Lecture 18 - 26.04.2016</b>	<b>67</b>
47. Differential Entropy	67
48. Typical Set	69
49. Volume of a Set	69
50. AEP for Continuous Random Variables	70
51. Differential Entropy vs Discrete Entropy	71
52. Joint Differential Entropy	71
53. Conditional Differential Entropy	72
54. Relative Entropy and Mutual Information	73
<b>Part 20. Lecture 19 - 28.04.2016</b>	<b>73</b>
55. Brief Summary of Continuous Time Relations	74
56. Hadamard's Inequality	76

<b>Part 21. Lecture 20 - 03.05.2016</b>	76
57. The Gaussian Channel	77
58. The Gaussian Channel (DTCA)	78
<b>Part 22. Lecture 21 - 05.05.2016</b>	78
59. The Information Channel Capacity of The Gaussian Channel	78
60. Sphere Packing Argument	79
<b>Part 23. Lecture 22 - 10.05.2016</b>	81
61. Band-Limited Channels	81
62. Sampling (Shannon-Nyquist) Theorem	82
<b>Part 24. Lecture 23 - 12.05.2016</b>	84
63. Normalized Capacity	84
64. Parallel Gaussian Channels	85
<b>Part 25. Lecture 24 - 17.05.2016</b>	87
65. Channels with Colored Gaussian Noise	87
65.1. Part a	87
65.2. Part b	88
65.3. Part c	88
65.4. Part d	88
65.5. If noise is a WSS Gaussian Process	89
66. Rate Distortion Theory (Lossy Source Coding)	90
67. Performance Measures	90
68. Quantization	90
68.1. $R = 0$	90
68.2. $R = 1$	91
68.3. $R = 2$	91
69. Some Distortion Definitions	91
69.1. Distortion Function	91
<b>Part 26. No Lecture on 19.05.2016</b>	92
<b>Part 27. Lecture 25 - 24.05.2016</b>	92
69.2. Rate-Distortion Code	92
69.3. Achievable Rate	93
69.4. Distortion Region	93
69.5. Operational Rate Distortion Function	94
69.6. Operational Distortion Rate Function	94
69.7. The Information Rate Distortion Function	94
69.8. Binary Source with Hamming Distortion	95
<b>Part 28. Lecture 26 - 26.05.2016</b>	97
69.9. Gaussian Source with MSE	97
<b>Part 29. To Do's</b>	99

<b>Part 30. Materials</b>	99
70. Lecture 01	99
71. Lecture 02	100
72. Lecture 03	100
73. Lecture 04	100
74. Lecture 05	100
75. Lecture 06	100
76. Lecture 07	100
77. Lecture 08	100
78. Lecture 09	101
79. Lecture 10	101
80. Lecture 11	101
81. Lecture 12	101
82. Lecture 13	101
83. Lecture 14	101
84. Lecture 15	101
85. Lecture 16	101
86. Lecture 17	102
87. Lecture 18	102
88. Lecture 19	102
89. Lecture 20	102
90. Lecture 21	102
91. Lecture 22	102
92. Lecture 23	102
93. Lecture 24	103
94. Lecture 25	103
95. Lecture 26	103
<b>Part 31. Code Appendix</b>	103

## Part 1. Lecture 01 - 23.02.2016

### 1. INTRODUCTION

Information theory deals with limits of communication.

#### Highlight 1.

**Arbitrarily Small Probability** means that probability goes to 0 but not exactly 0 (I am not sure about this definition with probability 1).

This is the end of the first hour.

How can we measure randomness?

First quantify the randomness.

**Highlight 2.**

**Entropy** is a measure of randomness/uncertainty.

How can we reduce the randomness?

Asking questions and processing the answers reduce the uncertainty.

**Example 1.**

A random variable  $X$  with  $\Omega = A_X = \{1, 2, 3, 4\}$  all equally likely. Yes/No questions can be asked. How many questions needed to reveal outcome?

Let's assume that we use the following set of questions (set #1):

- (1) Is it 1? If yes it is 1, if no continue asking.
- (2) Is it 2? If yes it is 2, if no continue asking.
- (3) Is it 3? If yes it is 3, if no it is 4.

Let's  $N$  denotes the number of questions asked. In that case  $P(N = 1) = \frac{1}{4}$ ,  $P(N = 2) = \frac{1}{4}$ ,  $P(N = 3) = \frac{2}{4}$ .

$$E[N] = 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{2}{4} = \frac{9}{4}$$

Now assume that we use the following set of questions (set #2):

- Is it  $\geq 2$ ?
  - { If yes is it 3?
  - { If no is it 1?

For any case, 2 questions are asked. Then  $E[N] = 2$ .

Now, the question is that: Can we further reduce  $E[N]$ ? What is the lower limit? This kind of questions are related with data compression topic. Notice that Yes/No questions corresponds to binary approach.

**Example 2.**

$X$  with  $\Omega = A_X = \{0, 1\}$

- $p(0) = 1, p(1) = 0 \rightarrow$  No uncertainty
- $p(0) = 0.5, p(1) = 0.5 \rightarrow$  Large uncertainty
- $p(0) = 0.1, p(1) = 0.9 \rightarrow$  Less uncertainty

**Example 3.**

$X$  with  $\Omega = A_X = \{0, 1\}$

- $X_1$  with  $\Omega = A_{X_1} = \{1, 2, 3, 4\} \rightarrow 2$  questions
- $X_2$  with  $\Omega = A_{X_2} = \{1, 2, 3, 4, 5, 6, 7, 8\} \rightarrow 3$  questions  $\rightarrow$  More uncertainty

## Part 2. Lecture 02 - 25.02.2016

## 2. WHAT IS ENTROPY?

**Definition 1** (Entropy).

Entropy of a discrete random variable  $X$  is defined as

$$(2.1) \quad H(X) = - \sum_x P_X(x) \log P_X(x)$$

where  $P_X(x)$  or  $P(X = x)$  or  $p(x)$  is the probability of random variable  $X$  being equal to  $x$ . Unit of entropy is **bits** or **nats** if the logarithm in (2.1) is taken in **base 2** or in **base e**, respectively. Change of base brings a multiplication factor.

**Highlight 3.**

After that point, logarithm will be taken in **base 2** unless explicitly noted.

**Example 4** (Entropy of a fair coin toss).

$$P(X = H) = P(X = T) = 0.5$$

Using (2.1), it can be found as

$$H(X) = - \left( \frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} \right) = 1 \text{ bit}$$

**Highlight 4.**

Yes/No question equally likely to be answered also holds a 1 bit randomness.

**Highlight 5.**

Entropy is a functional of the probability distribution of  $X$ . In other words, not  $f(x)$  but  $f(g(x))$ . It is not related with outcomes of a random variable. It depends on PMF only.

What about the contribution of a value  $x_0$  to entropy if  $P(X = x_0) = 0$ ? Well, it should be zero intuitively. If  $X$  can't be  $x_0$ , how can  $x_0$  affect uncertainty of  $X$ ? Let's calculate using (2.1).

$$-p(x) \log p(x) = -0 \times \log 0 = -0 \times -\infty \rightarrow \text{indeterminate, oooops!}$$

But it should be 0, right?

Observe that

$$x \log x = \frac{\log x}{1/x}.$$

$$\lim_{x \downarrow 0} x \log x = \lim_{x \downarrow 0} \frac{\log x}{1/x} \stackrel{LH}{=} \lim_{x \downarrow 0} \frac{1/x}{-1/x^2} = \lim_{x \downarrow 0} \frac{-x^2}{x} = \lim_{x \downarrow 0} -x = 0.$$

See L'Hopital's Rule for further.  
Remember (2.1).

$$\begin{aligned} H(X) &= - \sum_x P_X(x) (\log P_X(x)) \\ &= -E[\log P_X(X)] \\ &= E \left[ \frac{1}{\log P_X(X)} \right] \end{aligned}$$

**Corollary 1.**

$$H(X) = E \left[ \frac{1}{\log p(X)} \right]$$

**Lemma 1.**

$$H(X) \geq 0$$

**Proof:**

$$0 \leq P_X(x) \leq 1 \rightarrow \log P_X(x) \leq 0 \rightarrow -\log P_X(x) \geq 0 \rightarrow H(X) \geq 0$$

Entropy of a binary random variable, i.e.,

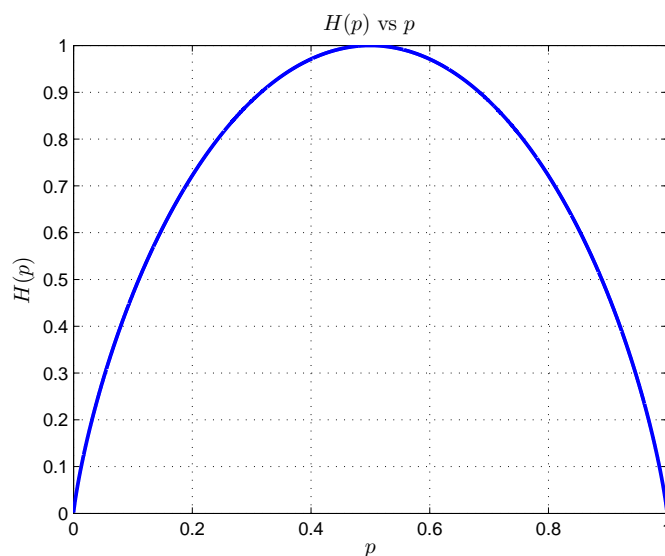
$$X = \begin{cases} 0, & \text{with probability } p \\ 1, & \text{with probability } 1 - p \end{cases}$$

$$H(X) = -p \log p - (1 - p) \log(1 - p)$$

**Definition 2** (Binary Entropy Function).

$$H(p) = -p \log p - (1 - p) \log(1 - p)$$



FIGURE 2.1.  $H(p)$  vs  $p$ 

## 3. JOINT ENTROPY

**Definition 3** (Joint Entropy).

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

Note that  $p(x, y)$  is joint PMF of  $(X, Y)$ .

**Example 5.**

$Y/X$	0	1
0	0	0.5
1	0.5	0

$$H(X, Y) = 1$$

Or, we can construct a *combined (super) random variable*,  $Z$ , as follows

$$A_Z = \{00, 01, 10, 11\}$$

$Z$	00	01	10	11
$p(z)$	0	0.5	0.5	0

Then,

$$H(Z) = 1$$

Let's extend definition of entropy to conditional entropy.

$Y$  conditioned on  $X = a$  is a regular random variable with  $p(y|X = a) = P(Y = y|X = a)$ . Then,

$$H(Y|X = a) = - \sum_y p(y|X = a) \log p(y|X = a)$$

#### 4. CONDITIONAL ENTROPY

**Definition 4** (Conditional Entropy).

$$(4.1) \quad H(Y|X) = \sum_x p(x) H(Y|X = x)$$

From (4.1),

$$\begin{aligned} H(Y|X) &= - \sum_x p(x) \left( \sum_y p(y|x) \log p(y|x) \right) \\ &= - \sum_x \sum_y p(x) p(y|x) \log p(y|x) \\ &= - \sum_x \sum_y p(x, y) \log p(y|x) \end{aligned}$$

Notice that  $p(x, y) = p(x)p(y|x)$  and  $E[g(X)] = \sum g(x)p(x)$ . Let's take  $g(x, y) = \log p(y|x)$ . Then,

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

**Corollary 2.**

$$H(Y|X) = -E[\log p(Y|X)]$$

**Example 6.**

$Y/X$	0	1
0	1/4	1/6
1	1/4	1/6
2	0	1/6

$$H(Y|X = 0) = \log 2 = 1\text{bit}$$

$$H(Y|X = 1) = \log 3\text{bits}$$

Using (4.1)

$$H(Y|X) = \frac{1}{2} \log 2 + \frac{1}{2} \log 3 = 0.5 \log 6\text{bits}$$

## 5. CHAIN RULE

**Theorem 1** (Chain Rule).

$$(5.1) \quad H(X, Y) = H(X) + H(Y|X)$$

**Proof:**

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

Note that  $p(x, y) = p(x)p(y|x)$ .  $\log p(x, y) = \log p(x) + \log p(y|x)$

$$\begin{aligned}
 H(X, Y) &= - \sum_x \sum_y p(x, y) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \\
 &= - \sum_x \log p(x) \sum_y p(x, y) - \sum_x \sum_y p(y|x)p(x) \log p(y|x) \\
 &= - \sum_x \log(p(x))p(x) - \sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
 &= H(X) + \sum_x p(x)H(Y|X = x) \\
 &= H(X) + H(Y|X)
 \end{aligned}$$

## Part 3. Lecture 03 - 01.03.2016

## Corollary 3.

$$(5.2) \quad H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$$

## Highlight 6.

In general,  $H(X|Y) \neq H(Y|X)$  is not equal.

$$-\sum_{x,y} p(x,y) \log p(x|y) \neq -\sum_{x,y} p(x,y) \log p(y|x)$$

Also, from the chain rule (5.1), it can be understood as well.

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

## 6. DIVERGENCE

## Definition 5 (Divergence, Kullback-Leibler Distance).

The divergence (relative entropy, cross entropy, Kullback-Leibler distance) between two PMFs  $p(x)$  and  $q(x)$  **with respect to (Order is important.)**  $p(x)$  is defined as

$$(6.1) \quad D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \left[ \log \frac{p(X)}{q(X)} \right]$$

Notice that,  $E_p$  in (6.1) is important because now, we have two different PMFs.

## Highlight 7.

Some equalities especially for divergence expression:

$$\begin{aligned} 0 \log \frac{0}{q} &= 0 \\ p \log \frac{p}{0} &= \infty \\ 0 \log \frac{0}{0} &= 0 \end{aligned}$$

Notice that  $p$  and  $q$  are a non-zero value from  $p(x)$  and  $q(x)$ , respectively.

**Example 7.**

$X$	1	2	3
$p(x)$	1/3	1/3	1/3
$q(x)$	1/2	1/2	0

Notice that  $D(p||q) = \infty$  (consider  $X = 3$  point) and  $D(q||p) < \infty$ .

**Highlight 8.**

- $D(p||q) \neq D(q||p)$  in general.
- Divergence is not an actual distance due to asymmetry.
- $D(p||q) = 0$  iff  $p$  and  $q$  are exactly same.

**7. DIVERGENCE INEQUALITY****Lemma 2 (Divergence Inequality).**

$$(7.1) \quad D(p||q) \geq 0$$

$D(p||q) = 0$  iff  $p = q \equiv p(x) = q(x)$  for  $\forall x$ . Notice that in this expression,  $q$  and  $p$  are not single values, they are PMFs.

Now, we are going to prove that (7.1) is hold. We will use this inequality more often than divergence itself in this course. Proof is after Jensen's Inequality.

## 8. CONVEX FUNCTION

**Definition 6** (Convex Function).

A function  $f(x)$  is convex over an interval  $(a, b)$  if for every  $x_1, x_2 \in (a, b)$  and  $0 \leq \lambda \leq 1$  if it satisfies

$$(8.1) \quad f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

If LHS is  $<$ , not  $\leq$ , than the RHS, it is called as **strictly convex**.

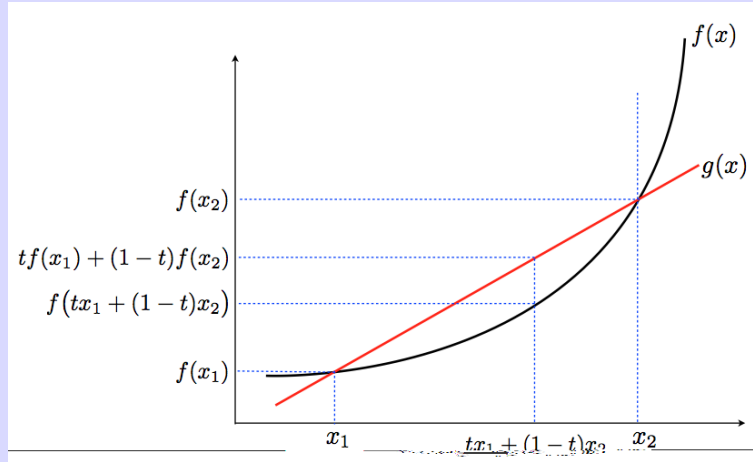


FIGURE 8.1. A Typical Convex Function

Notice that  $\lambda$  in (8.1) is denoted by  $t$  in Figure 8.1 and RHS of (8.1) is denoted by  $g(x)$ .

If  $f(x)$  is a convex function,  $-f(x)$  becomes concave or vice versa.

**Example 8.**

$x^2$  and  $e^x$  are convex and  $\log(x)$  and  $\sqrt{x}$  are concave functions.

**Theorem 2.**

A function is convex or strictly convex if its second derivative is non-negative or positive, respectively.

## 9. JENSEN'S INEQUALITY AND ITS PROOF

**Theorem 3** (Jensen's Inequality).

If  $f(\cdot)$  is a convex function and  $X$  is a random variable then,

$$(9.1) \quad E[f(X)] \geq f(E[X]).$$

In other words,

$$\sum_x p(x)f(x) \geq f\left(\sum_x p(x)x\right).$$

**Proof by Induction:**

First, let's assume that we have a two-point mass function as  $P(X = x_1) = p_1$  and  $P(X = x_2) = p_2 = 1 - p_1$ . Now, let's write  $E[f(X)]$ .

$$\begin{aligned} E[f(X)] &= p_1 f(x_1) + p_2 f(x_2) \\ E[f(X)] &= p_1 f(x_1) + (1 - p_1) f(x_2) \end{aligned}$$

Notice that this expression is very similar to RHS of (8.1), just switch  $\lambda$  by  $p_1$ . But since  $f(\cdot)$  is a convex function, (9.1) is hold. Then,

$$\begin{aligned} f(p_1 x_1 + (1 - p_1) x_2) &\leq p_1 f(x_1) + (1 - p_1) f(x_2) \\ f(p_1 x_1 + (1 - p_1) x_2) &\leq E[f(X)] \\ f(E[X]) &\leq E[f(X)] \end{aligned}$$

Jensen's Inequality is proven for a random variable with two-point PMF.

Now assume that inequality holds for  $K - 1$  point PMF, check whether it holds for  $K$  point PMF.

Let's assume that  $X$  has a  $K$  point PMF take values from  $p_1, p_2, \dots, p_k$ . It can be said that

$$\sum_{i=1}^k p_i = 1.$$

Similarly,

$$\sum_{i=1}^{k-1} p_i = 1 - p_k.$$

Define a new PMF as

$$p_i^0, \quad \frac{p_i}{1 - p_k}$$

$$\begin{aligned}
E_p[f(X)] &= \sum_{i=1}^k p_i f(x_i) \\
&= p_k f(x_k) + \sum_{i=1}^{k-1} p_i f(x_i) \\
&= p_k f(x_k) + \sum_{i=1}^{k-1} p_i^\theta (1 - p_k) f(x_i) \\
(9.2) \quad &= p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p_i^\theta f(x_i)
\end{aligned}$$

Notice that  $p_1^\theta, p_2^\theta, \dots, p_{k-1}^\theta$  forms a valid PMF. Also,

$$E_{p'}[f(X)] = \sum_{i=1}^{k-1} p_i^\theta f(x_i)$$

We assumed that Jensen's inequality is hold for  $K - 1$  point PMF. Then,

$$(9.3) \quad E_{p'}[f(X)] \geq f(E_{p'}[X])$$

Let's use (9.3) in (9.2).

$$\begin{aligned}
E_p[f(X)] &\geq p_k f(x_k) + (1 - p_k) f(E_{p'}[X]) \\
&\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p_i^\theta x_i\right)
\end{aligned}$$

Notice that  $f(\cdot)$  is a convex function, then



$$\begin{aligned}
E_p[f(X)] &\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p_i^\theta x_i\right) \geq f\left(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p_i^\theta x_i\right) \\
&\geq f\left(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p_i^\theta x_i\right) \\
&\geq f\left(p_k x_k + \sum_{i=1}^{k-1} (1 - p_k) p_i^\theta x_i\right) \\
&\geq f\left(p_k x_k + \sum_{i=1}^{k-1} p_i x_i\right) \\
&\geq f\left(\sum_{i=1}^k p_i x_i\right) \\
E_p[f(X)] &\geq f(E_p[X])
\end{aligned}$$

We have showed that Jensen's Inequality is hold for 2-point PMF. Also, we have showed that if it is hold for  $K - 1$  point PMF, it is also hold for  $K$  point PMF. So starting from 2, we may say that it is valid for 3-point PMF. From 3, we can say that it is also valid for 4 point PMF and so on.

#### Highlight 9.

Jensen's Inequality is also hold for continuous random variables.

### 10. PROOF OF DIVERGENCE INEQUALITY

We are going to prove (7.1) which is  $D(p||q) \geq 0$ .

**Proof of Divergence Inequality:** Define a new random variable  $Y$  which is

$$Y = \frac{q(X)}{p(X)}.$$

Take the logarithm of  $Y$  but notice that logarithm is a concave function so Jensen's Inequality is hold in opposite way.

$$\begin{aligned}
E_p[\log Y] &\leq \log(E_p[Y]) \\
\sum_x p(x) \log \frac{q(x)}{p(x)} &\leq \log \left( \sum_x p(x) \frac{q(x)}{p(x)} \right) \\
\sum_x p(x) \log \frac{q(x)}{p(x)} &\leq \log \left( \sum_x q(x) \right) \\
\sum_x p(x) \log \frac{q(x)}{p(x)} &\leq \log(1) \\
\sum_x p(x) \log \frac{q(x)}{p(x)} &\leq 0 \\
-\sum_x p(x) \log \frac{p(x)}{q(x)} &\leq 0 \\
-D(p||q) &\leq 0 \\
D(p||q) &\geq 0
\end{aligned}$$

## 11. MUTUAL INFORMATION

### Definition 7 (Mutual Information).

$X$  and  $Y$  with joint PMF  $p(x, y)$  and marginal PMF's  $p(x)$  and  $p(y)$ . The **mutual information**  $I(X; Y)$  is defined as

$$\begin{aligned}
I(X; Y) &= D(p(x, y) || p(x)p(y)) \\
&= \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= E_{p(x, y)} \left[ \log \frac{p(X, Y)}{p(X)p(Y)} \right]
\end{aligned}$$

### Highlight 10.

$I(X; Y)$  indicates that "How much  $X$  knows about  $Y$ ?" or vice versa. It is "Knowledge of  $X$  on  $Y$ ." or vice versa. If  $X$  and  $Y$  are independent, mutual information becomes 0. It means that  $X$  or  $Y$  knows nothing about  $Y$  or  $X$ .

## 12. ENTROPY AND MUTUAL INFORMATION

$$\begin{aligned}
I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{p(y)} \\
&= \sum_x \sum_y p(x,y) \log p(y|x) - \sum_y \sum_x p(x,y) \log p(y) \\
&= \sum_x \sum_y p(x,y) \log p(y|x) - \sum_y p(y) \log p(y) \\
&= -H(Y|X) + H(Y) \\
I(X;Y) &= H(Y) - H(Y|X)
\end{aligned}$$

Notice that

$$p(y|x) = \frac{p(x,y)}{p(x)}.$$

**Corollary 4.**

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

Mutual information between two random variable is the **reduction** in the uncertainty of one when the other is known. There is a symmetry.

**Highlight 11.**

$$I(X;X) = H(X)$$

**Corollary 5.**

$$I(X;Y) \geq 0.$$

Because it is a divergence.

## 13. CHAIN RULE FOR ENTROPY

**Theorem 4** (Chain Rule for Entropy).

$$\begin{aligned}
H(X_1, X_2, \dots, X_n) &= H(X_1) \\
&\quad + H(X_2|X_1) \\
&\quad + H(X_3|X_1, X_2) \\
&\quad + \dots \\
&\quad + H(X_n|X_1, X_2, \dots, X_{n-1}) \\
H(X_1, X_2, \dots, X_n) &= \sum_{i=1}^n H(X_i|X_1, X_2, \dots, X_{i-1})
\end{aligned}$$

**Proof:**

We have shown the case when  $n = 2$  in (5.1). Let's start with  $n = 3$  case.

$$\begin{aligned}
H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \quad X_2, X_3 \text{ forms a super r.v.} \\
&= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \quad \text{Remember (5.2).} \\
&= \vdots \\
H(X_1, X_2, \dots, X_n) &= \sum_{i=1}^n H(X_i|X_1, X_2, \dots, X_{i-1})
\end{aligned}$$

## 14. CONDITIONAL MUTUAL INFORMATION

**Definition 8** (Conditional Mutual Information).

$$\begin{aligned}
I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\
&= E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}
\end{aligned}$$

## 15. CHAIN RULE FOR INFORMATION

**Theorem 5** (Chain Rule for Information).

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y|X_1, X_2, \dots, X_{i-1})$$

Here, we can think that  $X_i$ s are noisy observations and  $Y$  is a parameter to be estimated.

## Part 4. Lecture 04 - 03.03.2016

**Proof of Chain Rule for Information:**Let  $Z = X_1, \dots, X_n$ .

$$\begin{aligned}
I(X_1, X_2, \dots, X_n; Y) &= H(Z) - H(Z|Y) \\
&= H(X_1, \dots, X_n) - H(X_1, \dots, X_n|Y) \\
&= \sum_{i=1}^n H(X_i|X_1, \dots, X_{i-1}) - \sum_{i=1}^n H(X_i|X_1, \dots, X_{i-1}, Y) \\
&= \sum_{i=1}^n (H(X_i|X_1, \dots, X_{i-1}) - H(X_i|X_1, \dots, X_{i-1}, Y)) \\
&= \sum_{i=1}^n I(X_i; Y|X_1, X_2, \dots, X_{i-1}) \quad \text{using conditional mutual information}
\end{aligned}$$

16. UPPER BOUND FOR  $H(X)$ **Theorem 6** (Upper Bound for  $H(X)$ ).

$$(16.1) \quad H(X) \leq \log |A_X|$$

where  $A$  means "alphabet" and  $|A_X|$  is the alphabet size of  $X$ . $H(X) = \log |A_X|$  iff  $X$  is uniformly distributed in  $A_X$ .**Proof:**Let  $u(x) = 1/|A_X|$ .  $x \in A$  be the uniformly distributed of  $A_X$ .

$$\begin{aligned}
D(p||u) &= \sum_{x \in A_X} p(x) \log \frac{p(x)}{u(x)} \\
&= \sum_{x \in A_X} p(x) \log p(x) - \sum_{x \in A_X} p(x) \log u(x)
\end{aligned}$$

Notice that  $\log u(x) = -\log |A_X|$  and it is independent from  $x$ , it can be taken out of the summation. Then,

$$\begin{aligned}
D(p||u) &= -H(X) + \log |A_X| \sum_{x \in A_X} p(x) \\
&= -H(X) + \log |A_X| \cdot 1 \\
&= -H(X) + \log |A_X|
\end{aligned}$$

We know that divergence (LHS) is always  $\geq 0$ . It means that

$$H(X) \leq \log |A_X|$$

## 17. CONDITIONING REDUCES ENTROPY

**Theorem 7** (Conditioning Reduces Entropy).

*a.k.a.* Information can't hurt.

$$(17.1) \quad H(X|Y) \leq H(X)$$

$H(X|Y) = H(X)$  iff they are independent.

**Proof:** We know that  $I(X;Y) \geq 0$  means that  $H(X) - H(X|Y) \geq 0$ .

**Example 9.**

$p(x, y)$  is given as in the table.

$Y/X$	1	2
1	0	$3/4$
2	$1/8$	$1/8$

$$H(X) = H(1/8) \approx 0.544 \text{ bits}$$

$$H(X|Y = 1) = 0 \text{ bits}$$

$$H(X|Y = 2) = 1 \text{ bits}$$

Notice that for a single value of condition ( $Y = 2$ ), entropy doesn't have to decrease. But in average, it will do...

From conditional entropy (4.1),

$$\begin{aligned} H(X|Y) &= P(Y = 1)H(X|Y = 1) + P(Y = 2)H(X|Y = 2) \\ &= \frac{3}{4}0 + \frac{1}{4}1 \\ &= \frac{1}{4} \text{ bits.} \end{aligned}$$

## 18. INDEPENDENCE BOUND ON ENTROPY

**Theorem 8** (Independence Bound on Entropy).

Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ .

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$$

LHS and RHS are equal iff  $X_i$  are independent.

**Proof:**

It can be shown using chain rule for entropies.

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1}) \\ &\leq \sum_{i=1}^n H(X_i) \end{aligned}$$

From (17.1), we know that  $H(X_i | X_1, \dots, X_{i-1}) \leq H(X_i)$ .

## 19. DATA-PROCESSING INEQUALITY

Now, assume that we are interested in finding  $X$ . However, we don't have it directly, we have noisy observations  $Y$  for example. We are going to estimate/detect  $X$ . We process  $Y$  somehow and generate a new random variable called  $Z$ . We may use  $Z$  or  $Y$  to estimate  $X$ . Should we use  $Y$  or  $Z$ ?

**Highlight 12.**

No clever manipulation of the data can improve the inferences (detection, estimation) that can be made directly from data. You can't gain further information by processing the observation. So why do we have FFT, signal processing algorithms, etc. ? Well, they may not be optimal but they are practical, we can implement them. Also, we don't gain information but it doesn't mean that we lose. Information may still be same. However, generally we lose information. But we are engineers!

## 20. MARKOV CHAIN

**Definition 9** (Markov Chain).

Random variables  $X, Y, Z$  form a Markov chain in the order  $X \rightarrow Y \rightarrow Z$  if  $p(z|x, y) = p(z|y)$ .

In general (not special to Markov Chain), we can write:

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \dots$$

In Markov chains,  $p(x_3|x_2, x_1)$  becomes  $p(x_3|x_2)$ .

Also in general:

$$p(x, z|y) = p(x|y)p(z|x, y)$$

But for the previous definition:

$$p(x, z|y) = p(x|y)p(z|y)$$

**Highlight 13.**

If  $Z = g(Y)$  then  $X \rightarrow Y \rightarrow Z$  but not vice versa always! In this notation,  $g()$  is a function, it maps a value from  $Y$  to a single value to  $Z$ . For example, there isn't any randomness in  $g()$ . If  $Y$  is known,  $Z$  is fixed. That's the meaning here, it is a deterministic function.

**Part 5. Lecture 05 - 08.03.2016****Theorem 9** (Data Processing Inequality (DPI)).

If  $X \rightarrow Y \rightarrow Z$  then  $I(X; Y) \geq I(X; Z)$ . For equality,  $X \rightarrow Z \rightarrow Y$  should be satisfied.

No processing of  $Y$ , deterministic or random, can increase the information that  $Y$  contains about  $X$ .

**Proof:**

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y) \\ I(X; Z) + I(X; Y|Z) &= I(X; Y) + I(X; Z|Y) \end{aligned}$$

Notice that  $X$  and  $Z$  are conditionally independent ( $Y$ ). Therefore,  $I(X; Z|Y) = 0$ .

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

Since  $I(X; Y|Z) \geq 0$ ,  $I(X; Y) \geq I(X; Z)$ .

For equality,  $I(X; Y|Z) = 0$ . It means that  $X \rightarrow Z \rightarrow Y$  (WHY NOT  $Y \rightarrow Z \rightarrow X$ ???) This (equality case etc.) is related to sufficient statistics.

**Corollary 6.**

If  $Z = g(Y)$  then,  $I(X; Y) \geq I(X; Z)$ .  
Because, if  $Z = g(Y)$  then  $X \rightarrow Y \rightarrow Z$ .

**Corollary 7.**

If  $X \rightarrow Y \rightarrow Z$  then,  $I(X; Y|Z) \leq I(X; Y)$ .  
Notice that "conditioning reduces mutual information." statement is valid if there is a Markov chain! It may not be true in general.  
It can be shown easily using last step of proof of Data Processing Inequality.



**Example 10.**

We have a case where  $X \rightarrow Y \rightarrow Z$  does not hold. Let  $X$  and  $Y$  be independent and fair binary random variables.

$$Z = X \oplus Y$$

$$I(X; Y) = 0 \text{ bits}$$

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(X|Z) \\ &= P(Z = 0)H(X|Z = 0) + P(Z = 1)H(X|Z = 1) \\ &= 0.5 \times 1 + 0.5 \times 1 \\ &= 1 \text{ bit} \end{aligned}$$

**21. FANO'S INEQUALITY**

Suppose that  $X$  is observed as  $Y$ .  $Y$  is related to  $X$ , of course. Fano's inequality relates the probability of error in guessing  $X$  to its conditional entropy  $H(X|Y)$ . Let  $\hat{X}$  be estimate of  $X$ .  $\hat{X} = f(Y)$ .  $X \rightarrow Y \rightarrow \hat{X}$  forms a Markov chain.

$$P_e = P(\hat{X} \neq X)$$

**Theorem 10 (Fano's Inequality).**

$$(21.1) \quad H(P_e) + P_e \log(|A_X| - 1) \geq H(X|Y)$$

Little expansion:

$$H(P_e) + P_e \log(|A_X| - 1) \geq H(X|\hat{X}) \geq H(X|Y)$$

Notice that  $H(P_e)$  is a binary entropy function and its value between 0 and 1.

Notice that  $\hat{X} \in A_X$ . If not, see the proof below.

**Proof:**

(21.1) can be weakened to:

$$\begin{aligned} 1 + P_e \log(|A_X| - 1) &\geq H(X|Y) \\ P_e &\geq \frac{H(X|Y) - 1}{\log |A_X|} \end{aligned}$$

Notice that  $P_e = 0$  implies  $H(X|Y) = 0$ .

**Highlight 14** (Weakened Fano's Inequality (?)).

$$P_e \geq \frac{H(X|Y) - 1}{\log |A_X|}$$

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$$

Then,  $P_e = P(E = 1)$ ,  $H(P_e) = H(E)$ . (Again  $H(P_e)$  is binary entropy function.)

$$\begin{aligned} H(E, X|Y) &= H(X|Y) + H(E|X, Y) \\ &= H(E|Y) + H(X|E, Y) \end{aligned}$$

Notice that  $H(E|X, Y) = 0$  because when  $X$  and  $\hat{X}$  which is function of  $Y$  are known,  $E$  is known. Also,  $H(E|Y) \leq H(E)$  or  $H(E|Y) \leq H(P_e)$  because conditioning reduces the entropy. Also,

$$\begin{aligned} H(X|E, Y) &= P(E = 0)H(X|E = 0, Y) + P(E = 1)H(X|E = 1, Y) \\ &\leq (1 - P_e)0 + P_e \log(|A_X| - 1). \end{aligned}$$

How do we end up with  $\log(|A_X| - 1)$  term? It is related to upper bound for entropy given in (16.1). But since it is known that there is an error, sample space is decreased by one outcome. That's the reason why do we have minus 1 term.

*Note: If we don't want to take  $\hat{X} \in A_X$ , i.e., estimation from  $Y$  values not exist in  $A_X$  then we may drop minus 1 term in logarithm.*

Finally,

$$H(X|Y) \leq H(P_e) + P_e \log(|A_X| - 1)$$

## 22. LOG SUM INEQUALITY

**Theorem 11** (Log Sum Inequality).

For  $n$  positive numbers  $i = 1, \dots, n$ ,  $a_i \geq 0$ ,  $b_i \geq 0$ :

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

with equality iff  $a_i/b_i$  is constant.

THIS IS END OF CHAPTER 2.

## 23. THE ASYMPTOTIC EQUIPARTITION PROPERTY

Basic conclusion at the end of this chapter: When a long sequence of random variables are observed, the frequency of an event approaches the event's probability as the length of the sequence goes to infinity.

Consider sequence of random variables  $X_1, X_2, \dots, X_n, \dots$ . Assume that  $X_i$  is a I.I.D. (independent and identically distributed) with PMF  $p(x)$ . Then,

$$\begin{aligned} p(x_1, x_2, \dots, x_n) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= p(x_1)p(x_2) \dots p(x_n) \\ &= \prod_{i=1}^n P(X_i = x_i) \\ &= \prod_{i=1}^n p(x_i) \end{aligned}$$

Given an IID random sequence, laws of large numbers deal with random variables (summarizing variables let's say) obtained through the sequence like

$$\begin{aligned} S &= \sum_{i=1}^n X_i \\ \tilde{S} &= \frac{S - E[S]}{\sqrt{n}} \\ \hat{S} &= \frac{S}{n} \\ \hat{S} &= \frac{1}{n} \sum_{i=1}^n X_i \end{aligned}$$

Most of the time, we will be interested in  $\hat{S}$ , average. We will see that as  $n$  goes to  $\infty$ ,  $\hat{S}$  will concentrate around a single value. It will "behave" like a deterministic value.

## 24. WEAK LAW OF LARGE NUMBERS (WLLN)

Let  $X_1, X_2, \dots$  be a real-valued IID random sequence with finite mean ( $E[X_i] < \infty$ ) and finite variance ( $\sigma_{X_i}^2 < \infty$ ). Given  $n$ , let  $\hat{S} = 1/n \sum_{i=1}^n X_i$ . Then, for any  $\epsilon > 0$

**Theorem 12** (Weak Law of Large Numbers (WLLN)).

$$(24.1) \quad P\left(\left|\hat{S} - E[X]\right| \leq \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Alternatively,

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{S} - E[X]\right| \leq \epsilon\right) = 1$$

$$P\left(\left|\hat{S} - E[X]\right| \leq \epsilon\right) > 1 - \delta \quad \text{for any } \delta \text{ as } n \rightarrow \infty.$$

#### Part 6. Lecture 06 - 10.03.2016

**Proof:**

It can be proved using Chebyshev's Inequality which is given in (24.2).

$$(24.2) \quad P(|Y - E[Y]| \geq a) \leq \frac{\sigma_Y^2}{a^2}$$

Take  $Y = \hat{S}$ ,  $a = \epsilon$ .

$$\begin{aligned} E[Y] &= E[S] \\ &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= E[X] \end{aligned}$$

$$\begin{aligned} \sigma_Y^2 &= \sigma_{\hat{S}}^2 \\ &= n\sigma_X^2 \frac{1}{n^2} \\ &= \frac{\sigma_X^2}{n} \end{aligned}$$

Then we can write (24.2) can be written as

$$\begin{aligned} P\left(\left|\hat{S} - E[X]\right| \geq \epsilon\right) &\leq \frac{\sigma_X^2}{n\epsilon^2} \\ P\left(\left|\hat{S} - E[X]\right| \geq \epsilon\right) &\geq 1 - \frac{\sigma_X^2}{n\epsilon^2} = 1 - \delta \end{aligned}$$

Since

$$\lim_{n \rightarrow \infty} \frac{\sigma_X^2}{n\epsilon^2} = 0$$

then

$$(24.3) \quad \lim_{n \rightarrow \infty} P\left(\left|\hat{S} - E[X]\right| \leq \epsilon\right) = 1$$

(24.3) can be written as (24.4).

$$(24.4) \quad P\left(\hat{S} = E[X] \pm \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

## 25. THE LAW OF FREQUENCIES

### Corollary 8 (The Law of Frequencies).

Let  $\{X_i\}$  be I.I.D. random sequence with alphabet  $A_X$ . Define  $F$  which is a subset of  $A_X$ . Remember that any subset is called as event. Then  $C_F(\underline{x})$  is the number of times  $F$  occurs in the sequence. Then,  $\frac{C_F(\underline{x})}{n}$  is the frequency which  $F$  occurs in  $\underline{x}$  where  $n$  is the length of sequence. For any  $\epsilon > 0$ ,

$$P\left(\frac{C_F(X)}{n} = P(X \in F) \pm \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

**Proof:**

$$Z_i = \begin{cases} 1 & \text{if } X_i \in F \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} E[Z_i] &= 1 \times P(X_i \in F) + 0 \times P(X_i \notin F) \\ &= P(X_i \in F) \\ &= P(X \in F) \end{aligned}$$

$$\hat{S} = \frac{1}{n} \sum_{i=1}^n Z_i$$

By definition of frequency:

$$\frac{C_F(X)}{n} = \frac{1}{n} \sum_{i=1}^n Z_i = \hat{S}$$

WLLN (Weak Law of Large Numbers) says that

$$(25.1) \quad P\left(\frac{C_F(X)}{n} = P(X \in F) \pm \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Notice that in (25.1),

$$\frac{C_F(\underline{X})}{n} = \hat{S}$$

$$P(X \in F) = E[Z_i].$$

## 26. TYPICAL SETS

We wish to find probabilities of sequence which cause WLLN. Let's consider a sequence  $\underline{x} = [x_1, x_2, \dots, x_n]$ . Each  $x_i$  is an outcome of  $X_i$  and  $X_i$ s are I.I.D. Let's  $A_X$  be the alphabet of any  $X_i$  such that  $A_X = \{a_1, a_2, \dots, a_q\}$ . Probability of each outcome is  $(p_1, p_2, \dots, p_q)$ . Also, let's say that  $C_i(\underline{x})$  denotes the number of occurrence of  $a_i$  in  $\underline{x}$ . Then,

$$\begin{aligned} p(\underline{x}) &= p(x_1)p(x_2) \dots p(x_n) \\ &= (p_1)^{C_1(\underline{x})} (p_2)^{C_2(\underline{x})} \dots (p_q)^{C_q(\underline{x})} \\ (26.1) \quad &\simeq p_1^{np_1} p_2^{np_2} \dots p_q^{np_q} \\ (26.2) \quad &\simeq 2^{np_1 \log p_1} 2^{np_2 \log p_2} \dots 2^{np_q \log p_q} \\ &\simeq 2^{n(\sum_{i=1}^q p_i \log p_i)} \\ p(\underline{x}) &\simeq 2^{nH(X)} \end{aligned}$$

Notice that, transition from (26.1) to (26.2) is done with help of (26.3).

$$(26.3) \quad a^b = 2^{\log a^b} = 2^{b \log a}$$

### Highlight 15.

$$p(\underline{x}) \simeq 2^{nH(X)}$$

Notice that this is "equipartition" part and it is not dependent on exact value of  $\underline{x}$  directly. In other words, there are many  $\underline{x}$ s with this probability.

## Part 7. Lecture 07 - 15.03.2016

If  $X_1, X_2, \dots, X_n$  are IID with PMF  $p(x)$ , then

$$P\left(-^1\right)$$

$$p(X_1, X_2, \dots, X_n) = \prod_{i=1}^n p(X_i)$$

$$-\frac{1}{n} \log \prod_{i=1}^n X_i = -\frac{1}{n} \sum_{i=1}^n \log X_i$$

Now, define a new random variable  $Y_i$ ,  $-\log p(X_i)$  and remember the WLLN (24.1). Then,

$$E[Y] = E[-\log p(X)]$$

$$= H(X)$$

Using WLLN,

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n Y_i - E[Y]\right| \leq \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n -\log p(X_i) - H(X)\right| \leq \epsilon\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

#### Highlight 16.

What do we mean by  $P(X \leq 3)$ ? In probability theory, we talk about sets and probabilities are assigned to set. Where is set in this case? Actually,  $P(X \leq 3) = P(\{x : x \leq 3, x \in A_X\})$ .  $B$  is a subset (event) of  $A_X$  such that  $B \subset A_X$  and  $B = \{x : x \leq 3, x \in A_X\}$ .

Let's use this notation our last theorem.

$$P\left(\left\{x_1, x_2, \dots, x_n : \frac{1}{n} \sum_{i=1}^n -\log p(x_i) = H(X) \pm \epsilon\right\}\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

## 27. TYPICAL SET

### Definition 10 (Typical Set).

The typical set  $A_\epsilon^{(n)}$  with respect to  $p(x)$  is defined as

$$A_\epsilon^{(n)} = \left\{x : \frac{1}{n} \sum_{i=1}^n -\log p(x_i) = H(X) \pm \epsilon\right\} \quad x, x_1, x_2, \dots, x_n$$

## 28. SHANNON - McMILLAN THEOREM

**Theorem 14** (Shannon - McMillan Theorem).

- (1) If  $\underline{x} = (x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$ , then
- $$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots, x_n) \leq H(X) + \epsilon.$$
- (2)  $P\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$
- (3)  $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$
- (4)  $\left|A_\epsilon^{(n)}\right| \geq (1 - \epsilon)2^{n(H(X)+\epsilon)}$

**Proofs:**

- (1) It is actually definition from typical set.
- (2) It is based on the previous theorem.  $P\left(A_\epsilon^{(n)}\right) \rightarrow 1$  as  $n \rightarrow \infty$  is equivalent to  $P\left(A_\epsilon^{(n)}\right) > 1 - \delta$  as  $n \rightarrow \infty$ . Set  $\delta = \epsilon$ .
- (3)

$$\begin{aligned}
 1 &= \sum_{\underline{x} \in A_X^n} p(\underline{x}) \\
 &\geq \sum_{\underline{x} \in A_\epsilon^{(n)}} p(\underline{x}) \\
 (28.1) \quad &\geq \sum_{\underline{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \text{ since } p(\underline{x}) \geq 2^{-n(H(X)+\epsilon)} \\
 &\geq 2^{-n(H(X)+\epsilon)} \sum_{\underline{x} \in A_\epsilon^{(n)}} 1 \\
 &\geq 2^{-n(H(X)+\epsilon)} \left|A_\epsilon^{(n)}\right|
 \end{aligned}$$

*Note: How is (28.1) possible? Since  $p(\underline{x}) \simeq 2^{-nH(X)}$  and take  $\epsilon$  is a non-negative value. Because if we look at previous usage of  $\epsilon$ , non-negative values makes sense.*



(4) Use property 2 and similar proof method for 3.

$$\begin{aligned}
 1 - \epsilon &\leq P\left(A_{\epsilon}^{(n)}\right) \\
 &\leq \sum_{A_{\epsilon}^{(n)}} p(x) \\
 &\leq \sum_{A_{\epsilon}^{(n)}} 2^{-n(H(X) - \epsilon)} \\
 &\leq 2^{-n(H(X) - \epsilon)} \sum_{A_{\epsilon}^{(n)}} 1 \\
 &\leq 2^{-n(H(X) - \epsilon)} \left|A_{\epsilon}^{(n)}\right|
 \end{aligned}$$

## 29. NEGATIVE STATEMENT OF AEP

For any  $\epsilon > 0$ , there is a positively valued sequence  $a_{\epsilon,n}$  that converges to zero as  $n \rightarrow \infty$  such that for any positive integer  $n$  and any set  $S$  containing sequence with length  $n$   $|S| \geq (P(S) - a_{\epsilon,n}) 2^{n(H(X) - \epsilon)}$ . If  $P(S) \rightarrow 1$ ,  $|S| \geq 2^{n(H(X) - \epsilon)}$ . (I am not sure about the  $\&$  symbol. Prof. Yilmaz draws similar but different symbol with equality with meaning "roughly larger than equal to" according to him. I couldn't find it in L<sup>A</sup>T<sub>E</sub>X.)



”typicality bit.” Then, for the remaining bits we use two different representations one for typical sequences and one for non-typical ones.

For typical set, since there are roughly  $2^{nH(X)}$  elements, we can use  $nH(X)$  number of bits + 1 typicality bit to represent a typical sequence. To find an upper bound let’s say that we use  $n(H(X) + \epsilon) + 1$  bits.

For non-typical set, we may think that we should use  $n \log(|A_X|) + 1$  bits. Notice that this is an upper bound and generally number of typical sets are much more smaller than number of non-typical sets. Therefore, we think that all sets are non-typical sets to find an upper bound.

Notice that, from upper bound for entropy, (16.1), we say that typical sets needs less number of bits than non-typical sets for representation.

Let  $l(\underline{X})$  be the length of the representation corresponding to  $\underline{X}$ . Let’s find the average length for per symbol (for single random variable,  $X$ , in a sequence  $\underline{X}$ ).

$$\begin{aligned}
\frac{1}{n} E[l(\underline{X})] &= \frac{1}{n} \sum_{\underline{x}} p(\underline{x}) l(\underline{x}) \\
&= \frac{1}{n} \left( \sum_{\underline{x} \in A_\epsilon^{(n)}} p(\underline{x}) l(\underline{x}) + \sum_{\underline{x} \notin A_\epsilon^{(n)}} p(\underline{x}) l(\underline{x}) \right) \\
&= \frac{1}{n} \left( \sum_{\underline{x} \in A_\epsilon^{(n)}} p(\underline{x}) [n(H(X) + \epsilon) + 1] + \sum_{\underline{x} \notin A_\epsilon^{(n)}} p(\underline{x}) [n \log(|A_X|) + 1] \right) \\
&= \frac{1}{n} \left( [n(H(X) + \epsilon) + 1] \sum_{\underline{x} \in A_\epsilon^{(n)}} p(\underline{x}) + [n \log(|A_X|) + 1] \sum_{\underline{x} \notin A_\epsilon^{(n)}} p(\underline{x}) \right) \\
&= \left( H(X) + \epsilon + \frac{1}{n} \right) P(A_\epsilon^{(n)}) + \left( \log |A_X| + \frac{1}{n} \right) P((A_\epsilon^{(n)})^C)
\end{aligned}$$

Notice that as  $n \rightarrow \infty$ ,  $P(A_\epsilon^{(n)}) \rightarrow 1$  and  $P((A_\epsilon^{(n)})^C) \rightarrow 0$ . Therefore, in limit case,

$$\frac{1}{n} E[l(\underline{X})] \simeq H(X) + \epsilon$$

Interesting, right! The question is that ”Is this the minimum?”. We haven’t done it yet. But we will, we will... We will try to see that the lower bound is  $H(X)$ . Then we can say that the proposed scheme in other words typical set idea works well.

## Part 8. Lecture 08 - 17.03.2016

## 31. SOURCE CODING (DATA COMPRESSION)

**Definition 11** (Source Code).

A source code  $C$  for a random variable  $X$  is a mapping from  $A_X$  to  $\mathcal{D}$ , the set of (finite length)(strings of) symbols from a  $D$ -ary alphabet.  $C(x)$  is codeword corresponding to  $x$ .  $l(x)$  is length of  $C(x)$ .

**Example 13.**

Morse code is a source code.  $A_X = \{a, b, c, \dots, z\}$ .  $D = 2$ .  $\mathcal{D} = \{-, ., --, .-, \dots\}$  and so on.  $l(a) = 2$ ,  $l(b) = 3$ ,  $l(c) = 3$ .

**Definition 12.**

Expected length of a source code  $C$  is

$$L(C) = \sum_{x \in A_X} p(x)l(x).$$

**Example 14.**

$A_X = \{\text{FB, GS, BJK, TS, BS}\}$  with probabilities 0.1, 0.2, 0.4, 0.2, 0.1, respectively. Let  $\text{FB} \rightarrow 00$ ,  $\text{GS} \rightarrow 01$ ,  $\text{BJK} \rightarrow 02$ ,  $\text{TS} \rightarrow 10$ ,  $\text{BS} \rightarrow 2$ . In this case,  $D = 3$ .

$$L(C) = 0.1 \times 2 + 0.2 \times 2 + 0.4 \times 2 + 0.2 \times 2 + 0.1 \times 1 = 1.9 \text{ symbols.}$$

**Definition 13** (Non-Singular Code).

A code is non-singular if every element in  $A_X$  is mapped into a different string  $\mathcal{D}$ .  $x_i \neq x_j$  iff  $C(x_i) \neq C(x_j)$ . Note that non-singularity is related to single values of  $X$ , not a sequence. Can I go back for single codeword to its original value?

**Example 15.**

$A_X = \{x_1, x_2, x_3\}$ ,  $D = 2$ .  $C(x_1) = 0$ ,  $C(x_2) = 01$ ,  $C(x_3) = 1$ . This code is non-singular.

**Definition 14** (Extension of a Code).

The extension  $C^*$  of a source code  $C$  is the mapping from finite length strings of  $A_X$  to finite length strings in  $\mathcal{D}$  defined by

$$C^*(x_1, x_2, \dots, x_n) = C(x_1)C(x_2) \dots C(x_n) \quad \text{concatenation.}$$

**Example 16.**

$$A_X = \{a, b\}. \quad C(a) = 0. \quad C(b) = 11. \quad C(ab) = 011. \quad C(ba) = 110.$$

**Definition 15** (Unique Decodability).

A code is uniquely decodable if its extension is non-singular. Any encoded string is a uniquely decodable code has only one possible source string producing it.

**Example 17** (A Uniquely Decodable Code).

$$A_X = \{a, b\}. \quad C(a) = 0. \quad C(b) = 01.$$

**Definition 16** (Prefix (Instantaneous) Code).

A code is called a prefix (instantaneous) code if no codeword is a prefix of another. For example,  $C(a) = 0$ ,  $C(b) = 01$  is not a prefix code but  $C(a) = 0$ ,  $C(b) = 11$  is.

**Highlight 17.**

Prefix codes (our favourites!)  $\subset$  Uniquely decodable codes  $\subset$  Non-singular codes  $\subset$  All source codes.

## 32. KRAFT'S INEQUALITY

**Theorem 15** (Kraft's Inequality).

For any prefix code with alphabet size  $D$ , the codeword lengths  $l_1, l_2, \dots, l_m$  must satisfy the inequality

$$\sum_{i=1}^n D^{-l_i} \leq 1.$$

Conversely, given  $l_1^0, l_2^0, \dots, l_m^0$  satisfying  $\sum_{i=1}^n D^{-l_i^0} \leq 1$ , there exists a prefix code with code lengths  $l_1^0, l_2^0, \dots, l_m^0$ .

**Example 18.**

$A_X = \{a, b, c, d, e\}$ . Take  $D = 2$  and  $l(a) = l(b) = l(c) = l(d) = l(e) = 2$ . Let assign  $a \rightarrow 00$ ,  $b \rightarrow 01$ ,  $c \rightarrow 10$  and  $d \rightarrow 11$ . We can't assign anything to  $e$ . Check using Kraft's inequality:  $\sum_{i=1}^n D^{-l_i} = 5/4 \not\leq 1$ .

MT1 is up-to here.

## Part 9. Lecture 09 - 22.03.2016

**Proof:**

We will use tree representation of computer sciences for proof. Consider the Figure 32.1.

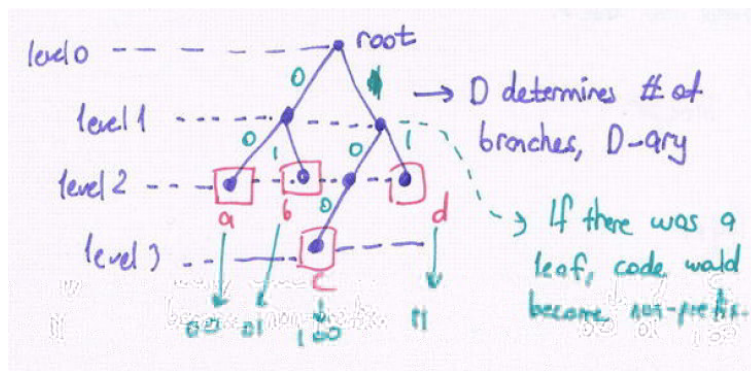


FIGURE 32.1. Tree Representation

In this representation:

- Each node has  $D$  children.
- Branches represent symbol from that  $D$ -ary alphabet.
- Each code word is represented by a leaf.
- The path from the root to a leaf determines the codeword for that leaf.

*Note: In tree terminology, leaf is defined as a node with no children. I think note written in green in the Figure 32.1 which is "If there was a leaf, code would become non-prefix." seems to be wrong. Because it conflicts with definition of leaf. Leaf definition in this section is slightly different than the true leaf definition. Please keep it in mind for the following parts.*

For prefix condition no codeword is prefix of another. This implies that descendants of a leaf can't be a leaf.

Consider the diagram shown in Figure 32.2.

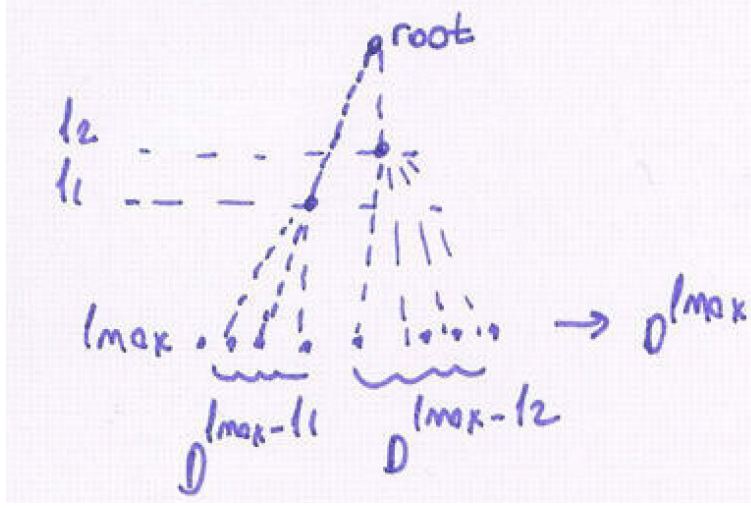


FIGURE 32.2. Levels of Tree

$l_{max}$  denotes the length of the longest code word. No node is a descendant of two codewords (non-singular code or prefix code?), there are 3 possibilities.

A node is,

- (1) descendant of a codeword. (ancestor ? Why do you put a node under a leaf, it should be a prefix code, right?)
- (2) codeword.
- (3) directly connected to root.

At level  $l_{max}$ , we can have at most  $D^{l_{max}}$  codewords. Number of descendants of a codeword with length  $l_i$  is  $D^{l_{max} - l_i}$ . Then,

$$\sum_i D^{l_{max} - l_i} \leq D^{l_{max}}$$

$$\sum_i D^{-l_i} \leq 1$$

**Example 19.**

Let  $l_1 = 1, l_2 = 2, l_3 = 3, l_4 = 3, D = 2$ . Is this possible?

$$2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} \leq 1$$

$$1 \leq 1$$

Yes.

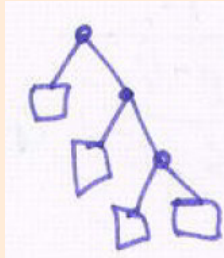


FIGURE 32.3. A Possible Solution

**Theorem 16** (General Kraft's Inequality (?)).

Kraft's inequality also holds for uniquely decodable codes which is a larger set than prefix codes. Proof is in the book (I hope).

## 33. NOISELESS CODING THEOREM

**Theorem 17** (Noiseless Coding Theorem).

The expected length ( $L$ ) of any uniquely decodable  $D$ -ary code for a random variable  $X$  satisfies

$$(33.1) \quad L \geq H_D(X)$$

where  $H_D(X)$  denotes the entropy taken in base  $D$  in (33.1). Equality is hold iff  $D^{-l_i} = p_i$  for all  $i$  where  $l_i$  is integer.  $p_i$  is the probability of  $X$  taking value  $i$ .

**Proof:**

$$\begin{aligned}
L - H_D(X) &= \sum_i p_i l_i - (-1) \sum_i p_i \log_D p_i \\
&= \sum_i p_i \log_D D^{l_i} + \sum_i p_i \log_D p_i \\
&= (-1) \sum_i p_i \log_D D^{-l_i} + \sum_i p_i \log_D p_i
\end{aligned}$$

Let's take a term  $c = \log_D \left( \sum_i D^{-l_i} \right)$  and add and subtract it.

$$\begin{aligned}
L - H_D(X) &= - \sum_i p_i \log_D D^{-l_i} + c + \sum_i p_i \log_D p_i - c \\
&= - \sum_i p_i (\log_D D^{-l_i} - c) + \sum_i p_i \log_D p_i - c \\
&= - \sum_i p_i \log_D \frac{D^{-l_i}}{\sum_j D^{-l_j}} + \sum_i p_i \log_D p_i - c \\
&= - \sum_i p_i \log_D \frac{p_i}{\sum_j \frac{p_j}{D^{-l_j}}} - c
\end{aligned}$$

Let  $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$ . Notice that  $r_i$  is a valid PMF because  $0 \leq r_i \leq 1$  and  $\sum_i r_i = 1$ .  
From definition of divergence,

$$\begin{aligned}
L - H_D(X) &= D(p||r) - c \\
&= D(p||r) - \log_D \left( \sum_i D^{-l_i} \right)
\end{aligned}$$

$D(p||r) \geq 0$ .  $\sum_i D^{-l_i} \leq 1$  is valid from Craft's inequality. Then minus logarithm becomes non-negative. Then,  $L \geq H_D(X)$ .

For equality,  $D(p||r) = 0$  and  $\sum_i D^{-l_i} = 0$ . Then,

$$p_i = r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}} = D^{-l_i}, \forall i$$



**Example 20.**

$X$  is a random variable.  $A_X = \{a, b, c\}$  with probabilities  $1/2, 1/4, 1/4$ , respectively.  $D = 2$ . It can be found that  $H(X) = 3/2$  bits. From the previous theorem, we know that average codeword length of best code is  $3/2$ . Let's try to find this. Do we satisfy  $p_i = D^{-l_i}$  condition? Yes!

$$\begin{aligned}\frac{1}{2} &= 2^{-l_1} \\ l_1 &= 1 \\ \frac{1}{4} &= 2^{-l_2} = 2^{-l_3} \\ l_2 &= l_3 = 2\end{aligned}$$

Let  $c(a) = 0, c(b) = 10, c(c) = 11$ . Check  $L$ .

$$\begin{aligned}L &= \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{4}2 \\ &= \frac{3}{2}\end{aligned}$$

**Definition 17 (D-adic Distribution).**

For D-adic distribution,  $p_i = D^{-l_i} \forall i$  where  $l_i$  is integer.

It is easy to come up with the optimal source code for D-adic distributions. Not so for non-D-adic ones.

**34. BOUNDS ON OPTIMAL CODE LENGTHS**

Optimal codes have smallest code length. Optimal code does not have to be unique.

$l_i - \log_D p_i$ . Notice that  $l_i$  is not integer for an arbitrary  $p_i$ . Shannon says that take  $l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$ . This is an ceiling function. Is Kraft's inequality satisfied with this selection?

$$\sum_i D^{-\left\lceil \log_D \frac{1}{p_i} \right\rceil} \leq D^{-\log_D \frac{1}{p_i}} = \sum_i p_i = 1$$

Kraft's inequality is satisfied for Shannon codelengths.

$$\begin{aligned}
\log_D \frac{1}{p_i} &\leq \left\lceil \log_D \frac{1}{p_i} \right\rceil < 1 + \log_D \frac{1}{p_i} \\
\sum_i p_i \log_D \frac{1}{p_i} &\leq \sum_i p_i \left\lceil \log_D \frac{1}{p_i} \right\rceil < \sum_i p_i \left( 1 + \log_D \frac{1}{p_i} \right) \\
H_D(X) &\leq L < 1 + H_D(X)
\end{aligned}$$

Let's say that a smart person says that he/she finds optimal codes. For optimal codes with length  $L$  means that  $H_D(X) \leq L \leq 1 + H_D(X)$ . Notice that Shannon code does not claim an optimal code. It is an sub-optimal code. It would be optimal if  $H_D(X) = L$ .

There is an overhead of 1 bit with this single symbol construction. If we send a sequence of source symbols,

$$(34.1) \quad H_D(X_1, \dots, X_n) \leq L(C_n) < H_D(X_1, \dots, X_n) + 1$$

Same inequality holds because we can construct a super random variable like  $X$ ,  $X_1, \dots, X_n$ .  $L(C_n)$  is average codelength for optimal  $C_n$  code constructed for length- $n$   $X$  sequence.

$$L_n(C_n) = \frac{L(C_n)}{n}$$

$L_n(C_n)$  is average codelength per source symbol to make a fair compression. From (34.1),

$$\begin{aligned}
\frac{1}{n} H_D(X_1, \dots, X_n) &\leq L_n(C_n) < \frac{1}{n} H_D(X_1, \dots, X_n) + \frac{1}{n} \\
(34.2) \quad H_D(X) &\leq L_n(C_n) < H_D(X) + \frac{1}{n} \quad \text{by taking I.I.D.}
\end{aligned}$$

(34.2) is linked to idea in AEP. Take  $n \rightarrow \infty$ .

## Part 10. Lecture 10 - 24.03.2016

## Example 21.

Let  $A_X = \{a, b\}$ ,  $p(a) = 0.8$ ,  $p(b) = 0.2$ . We can find that  $H_D(X) = 0.72$ . Shannon codeword lengths:

$$\begin{aligned}
 l_i &= \left\lceil \log_D \frac{1}{p_i} \right\rceil \\
 l_a &= \left\lceil \log_2 \frac{1}{0.8} \right\rceil = 1 \\
 l_b &= \left\lceil \log_2 \frac{1}{0.2} \right\rceil = 3 \\
 L_1 &= p(a)l_a + p(b)l_b \\
 &= 0.8 \times 1 + 0.2 \times 3 \\
 &= 1.4 \text{ bits/symbol}
 \end{aligned}$$

Then,

$$\begin{aligned}
 H_D(X) &\leq L_1 < H_D(X) + 1 \\
 0.72 &\leq 1.4 < 1.72
 \end{aligned}$$

For double symbol case:

$$\begin{aligned}
 l_{aa} &= \left\lceil \log_2 \frac{1}{0.64} \right\rceil = 1 \\
 l_{ab} = l_{ba} &= \left\lceil \log_2 \frac{1}{0.16} \right\rceil = 3 \\
 l_{bb} &= \left\lceil \log_2 \frac{1}{0.04} \right\rceil = 5 \\
 L_2 &= \frac{1}{2} (0.64 \times 1 + 0.16 \times 2 \times 3 + 0.04 \times 5) \\
 &= 0.9 \text{ bits/symbol}
 \end{aligned}$$

Then,

$$\begin{aligned}
 H_D(X) &\leq L_2 < H_D(X) + 1 \\
 0.72 &\leq 0.9 < 0.72 + 1/2
 \end{aligned}$$

## 35. HUFFMAN CODES

**Highlight 18.**

This section should be read from the book. It will be asked in exam with probability 1!

Huffman Codes are optimal, i.e., they have the smallest average blocklength for a given source. Note that average length of a source code does not necessarily equal to entropy. For example, for the previous example (single symbol case) minimum (practical) is 1 bit. It is optimum. But it is larger than the lower bound (0.72 bits). We can match the lower bound if source has a D-adic distribution.

Idea is in Huffman Code is assigning shorter codes to more probable sources. We form a tree starting from least probable source.

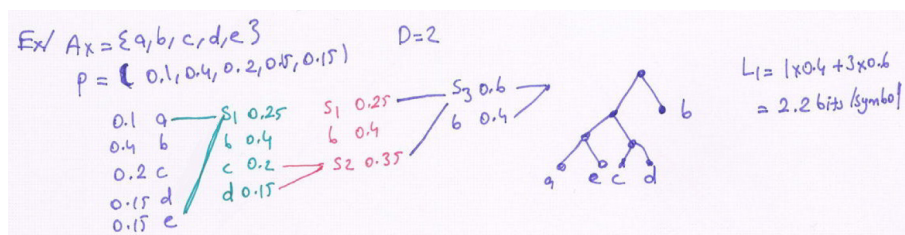


FIGURE 35.1. Binary Example

**Example 22.**

For ternary example, see the book.

## 36. CHANNEL CAPACITY

How does communication between two points: A & B occurs? The transfer of information is a physical process and hence subject to uncontrollable disturbance (noise) and imperfections of the physical signalling process. A generic communication system can be shown as in [36.1](#).

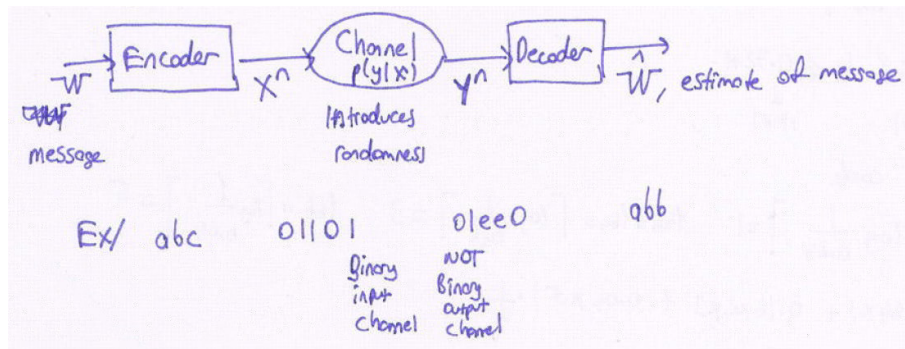


FIGURE 36.1. A General Communication System

## Part 11. Lecture 11 - 29.03.2016

**Definition 18** (Discrete Channel).

Let  $A_X$  and  $A_Y$  be the input and output alphabet, respectively. A discrete channel is a system of probability functions  $p_n(y_1, \dots, y_n | x_1, \dots, x_n)$ ,  $x_1, \dots, x_n \in A_X$ ,  $y_1, \dots, y_n \in A_Y$  where  $n = 1, 2, \dots$

$$p_n(y_1, \dots, y_n | x_1, \dots, x_n) \geq 0$$

$$\sum_{y_1, \dots, y_n} p_n(y_1, \dots, y_n | x_1, \dots, x_n) = 1$$

A discrete channel is memoryless if output  $y_k$  depends on input  $x_k$ .

$$p(y_k | y_{-k}, x_{-k}) = p(y_k | x_k)$$

**Definition 20** ("Information" Channel Capacity).

The "information" channel capacity of a discrete memoryless channel is defined as

$$(36.2) \quad C = \max_{p(x)} I(X; Y).$$

Notice that (36.2) is maximization over a function.

**Definition 21** ("Operational" Channel Capacity).

The "operational" channel capacity of a DMC(discrete memoryless channel) is the highest rate in bits per channel use at which information can be sent with arbitrarily small probability of error.

**Highlight 19.**

We will show that two definitions are the same.

## 37. EXAMPLE OF CHANNELS AND THEIR CAPACITIES

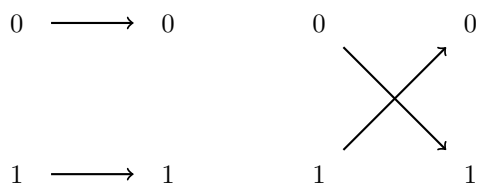


FIGURE 37.1. Noiseless Binary Channel

**37.1. Noiseless Binary Channel.** Operational channel capacity is 1 bit/channel use.

$$\begin{aligned}
 C &= \max_{p(x)} I(X; Y) \\
 &= \max_{p(x)} H(X) - H(X|Y) \\
 &= \max_{p(x)} H(X) \\
 &= 1 \text{ bit, when } p(x) \text{ is uniform.}
 \end{aligned}$$

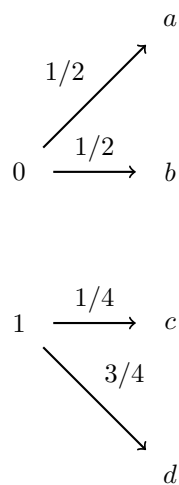


FIGURE 37.2. Noisy Channel with Non-overlapping Outputs

**37.2. Noisy Channel with Non-overlapping Outputs.** Again, 1 bit (since they are non-overlapping) can be reliably transmitted.

$$\begin{aligned}
 C &= \max_{p(x)} I(X; Y) \\
 &= \max_{p(x)} H(X) - H(X|Y) \\
 &= \max_{p(x)} H(X) \\
 &= 1 \text{ bit, when } p(x) \text{ is uniform.}
 \end{aligned}$$

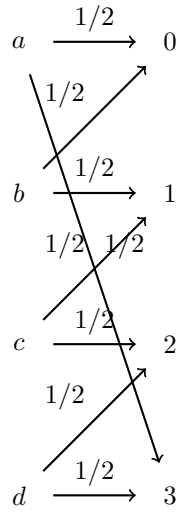


FIGURE 37.3. Noisy Channel with Overlapping Outputs

**37.3. Noisy Channel with Overlapping Outputs.** We can still transmit 1 bit reliably. For example, we can use non-overlapping pairs like  $a$  and  $c$ . But is this the maximum.

$$\begin{aligned}
 C &= \max_{p(x)} I(X; Y) \\
 &= \max_{p(x)} H(X) - H(X|Y) \\
 &= \max_{p(x)} H(Y) - H(Y|X) \\
 &= \max_{p(x)} H(Y) - \sum_x p(x) H(Y|X = x) \\
 &= \max_{p(x)} H(Y) - \sum_x p(x) 1 \\
 &= \max_{p(x)} H(Y) - 1 \\
 &= 2 - 1 = 1 \text{ bit, when } p(x) \text{ is uniform (not unique)(also } Y).
 \end{aligned}$$



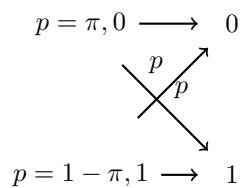


FIGURE 37.4. Binary Symmetric Channel

**37.4. Binary Symmetric Channel (BSC).** In this case,  $p$  denotes the probability of error.

$$\begin{aligned}
 C &= \max_{p(x)} I(X; Y) \\
 &= \max_{p(x)} H(Y) - H(Y|X) \\
 &= \max_{p(x)} H(Y) - H(p) \\
 &= 1 - H(p) \text{ bits, when } X \text{ is uniform.}
 \end{aligned}$$

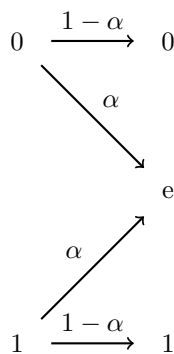


FIGURE 37.5. Binary Erasure Channel

**37.5. Binary Erasure Channel.** In Figure 37.5, "e" means erasure. This bit is erased, we have no idea about it.

$$\begin{aligned}
 C &= \max_{p(x)} I(X; Y) \\
 &= \max_{p(x)} H(Y) - H(Y|X) \\
 &= \max_{p(x)} H(Y) - H(\alpha)
 \end{aligned}
 \tag{37.1}$$

For example, can  $H(Y)$  be  $\log 3$ ? No! Let's try to maximize it.

$$E = \begin{cases} 1, & Y = e \\ 0, & \text{otherwise} \end{cases}$$

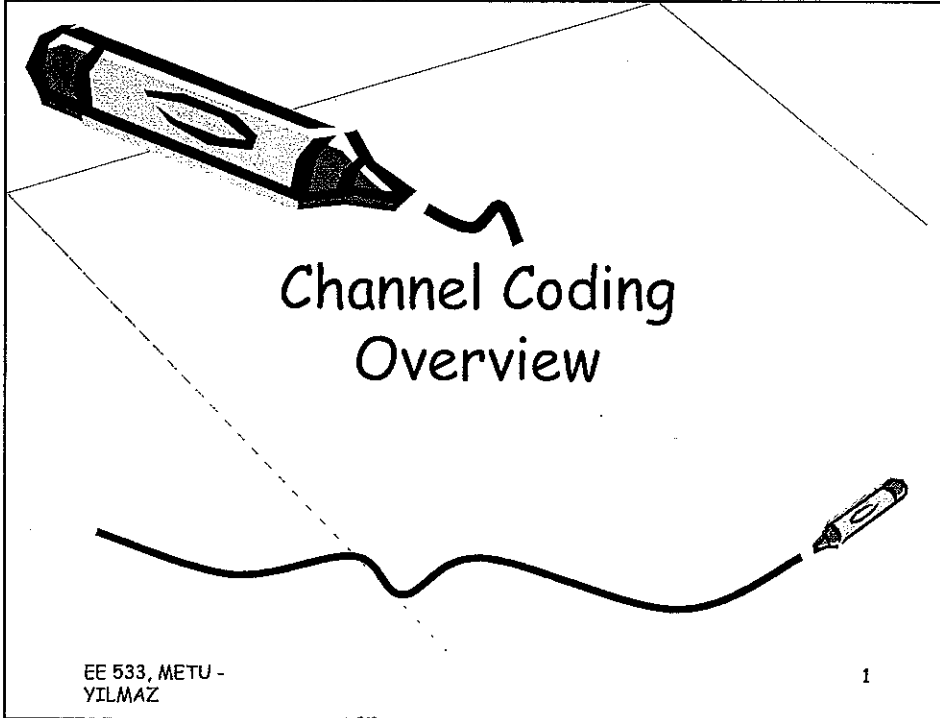
$$\begin{aligned} H(Y) &= H(Y) + 0 \\ &= H(Y) + H(E|Y) \\ &= H(E, Y) \\ &= H(E) + H(Y|E) \\ &= H(\alpha) + P(E = 0)H(Y|E = 0) + P(E = 1)H(Y|E = 1) \\ &= H(\alpha) + P(E = 0)H(Y|E = 0) + P(E = 1) \times 0 \\ &= H(\alpha) + P(E = 0)H(Y|E = 0) \\ &= H(\alpha) + P(E = 0)H(X) \\ &= H(\alpha) + (1 - \alpha)H(X) \end{aligned}$$

Putting final expression into (37.1) leads to:

$$\begin{aligned} C &= \max_{p(x)} H(Y) - H(\alpha) \\ &= \max_{p(x)} H(\alpha) + (1 - \alpha)H(X) - H(\alpha) \\ &= \max_{p(x)} (1 - \alpha)H(X) \\ &= 1 - \alpha \text{ bits, when } X \text{ is uniform.} \end{aligned}$$

## Part 12. Lecture 12 - 31.03.2016

Now add 12th slide.



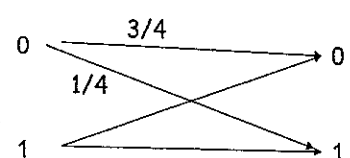
# Channel Coding Overview

EE 533, METU - YILMAZ

1

## SUMMARY

- The information channel capacity of BSC is  $1-H(p)$ .
- How to achieve that?
- Answer: Coding
- Consider a BSC



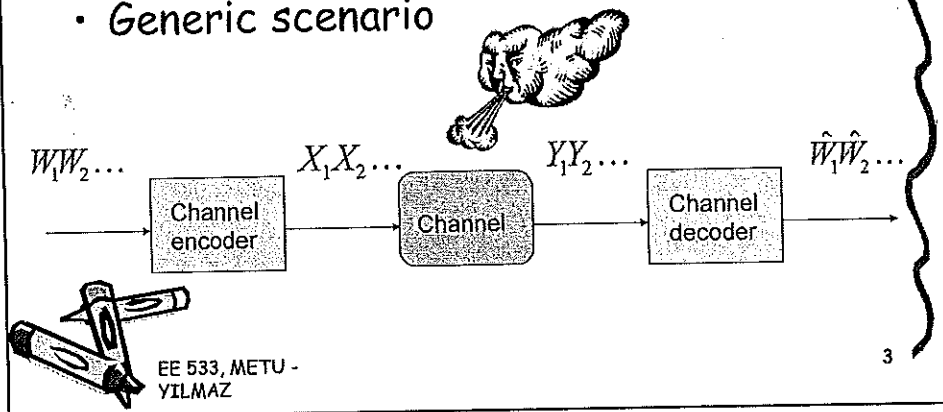
- Redundancy brings reliability at the cost of efficiency? TRADEOFF

EE 533, METU - YILMAZ

2

## Task of a Channel Code

- Transmitting data through a noisy channel in an accurate and efficient manner
- Generic scenario



## Our Assumptions

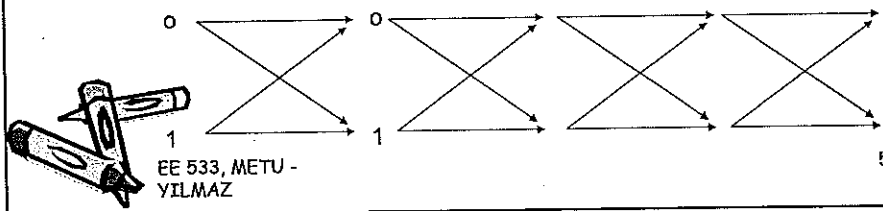
- The data  $W_1, W_2, \dots$  are i.i.d. and binary.
- The channel
  - Discrete: accepts input at discrete time and produces output at discrete time

- Memoryless: output only depends on the current input.

$$p(y_n | x_1, \dots, x_n, y_1, \dots, y_{n-1}, \dots) = p(y_n | x_n)$$

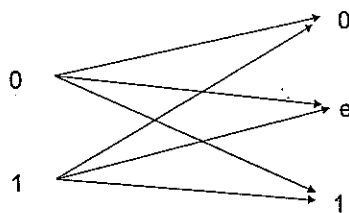
- Stationary: probabilistic characteristics do not change.

$$p(y_n | x_n) = p(y | x), \text{ for all } n, x, y$$



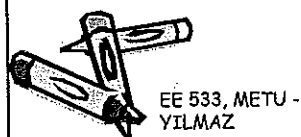
## Examples

- Binary erasures and errors channel



- Additive Gaussian channel

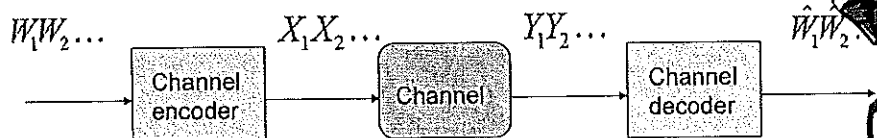
$$y = x + n$$



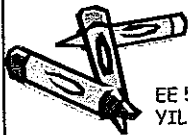
EE 533, METU - YILMAZ

6

# Channel Codes



- An encoder and a decoder
- Encoder
  - represents data with symbols from the channel input alphabet  $A_X$



EE 533, METU -  
YILMAZ

7

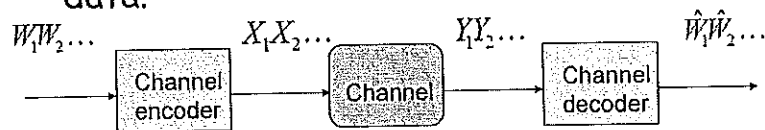
- Decoder
  - Recovers the message from the channel observation



# Channel Code Performance

- Accuracy

- An accurate code usually recovers the data.



$$P_e = \Pr(\hat{W} \neq W)$$

$$P_{BER} = \frac{1}{K} \sum_{i=1}^K \Pr(\hat{W}_i \neq W_i)$$

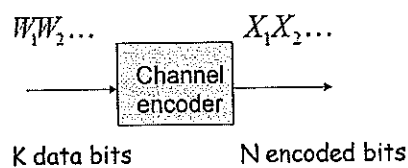


EE 533, METU -  
YILMAZ

9

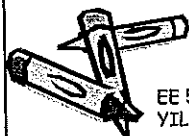
- Efficiency

- An efficient code has a relatively small number of encoded bits per data bit.



$$R = \frac{K}{N} \rightarrow \text{Code rate}$$

complexity

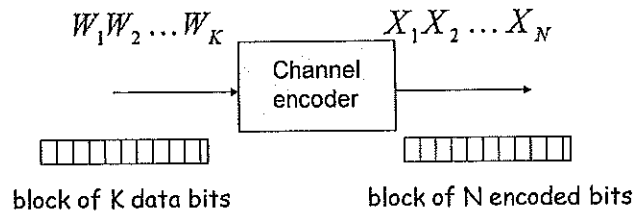


EE 533, METU -  
YILMAZ

10

# Block Codes

- Fixed-length to fixed-length codes



- K: data block length
- N: encoded block length
- The codebook is an ordered set.

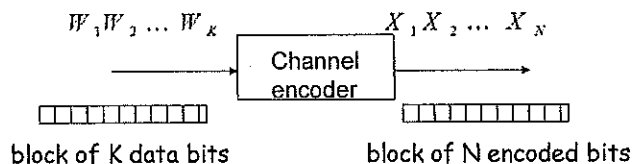
$$C = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{2^K}\}$$

$$C \subset A_X^N$$

*code word*

EE 533, METU -  
YILMAZ

11



- A one-to-one encoding rule

$$f_e : A_{\underline{W}}^{(K)} = \{0,1\}^{(K)} \rightarrow C$$

$$\underline{X} = f_e(\underline{W})$$

EE 533, METU -  
YILMAZ

12



**Part 13. Lecture 13 - 05.04.2016**

Now add from 12th slide.

block of N channel observations      block of K estimated data bits

- A many-to-one decoding rule

$$f_d : A_Y^N \rightarrow A_W^K = \{0,1\}^K$$

$$\underline{\hat{W}} = f_d(\underline{Y})$$

EE 533, METU - YILMAZ

Repetition

code:

Encoding 0 → 000  
1 → 111

4-to-1

Decoding

000	→ 0
001	→ 0
010	→ 0
011	→ 0
100	→ 1
101	→ 1
110	→ 1
111	→ 1

many to one

## • Error probability

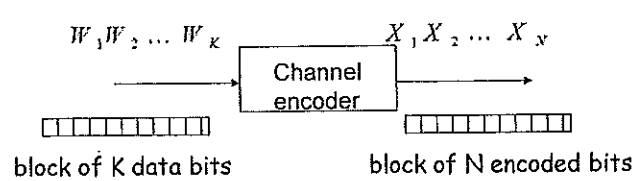
$$P_{BER} = \frac{1}{K} \sum_{i=1}^K \Pr(\hat{W}_i \neq W_i) \rightarrow \text{single symbol}$$

$$P_{PER} = \Pr(\underline{\hat{W}} \neq \underline{W}) \rightarrow \text{Sequence}$$

$$\frac{1}{K} P_{PER} \leq P_{BER} \leq P_{PER}$$

↓  
Block error probability

EE 533, METU - YILMAZ



The diagram shows a 'Channel encoder' block. An input arrow from the left is labeled  $W_1, W_2, \dots, W_K$  and is accompanied by a horizontal bar divided into 10 segments, labeled 'block of K data bits'. An output arrow to the right is labeled  $X_1, X_2, \dots, X_N$  and is accompanied by a horizontal bar divided into 10 segments, labeled 'block of N encoded bits'.

• Code rate

$$R = \frac{K}{N}$$

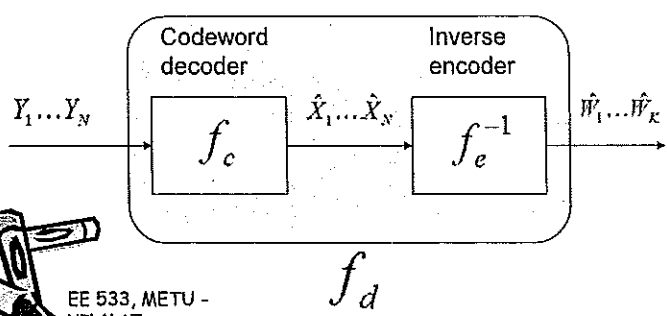
$$R = \frac{1}{N} \log_2 |C|, \text{ C is the codebook.}$$

EE 533, METU - YILMAZ

15

## The codeword decoding rule

- First, guess the codeword
- Then, find the corresponding data



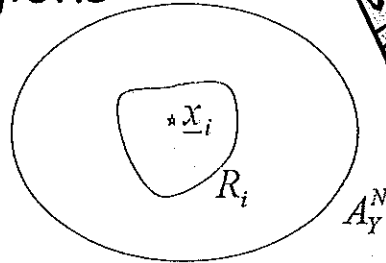
The diagram shows a 'Codeword decoder' block. An input arrow from the left is labeled  $Y_1, \dots, Y_N$ . Inside the block, the signal passes through a box labeled  $f_c$ , then an arrow labeled  $\hat{X}_1, \dots, \hat{X}_N$ , then a box labeled  $f_e^{-1}$ , and finally an output arrow labeled  $\hat{W}_1, \dots, \hat{W}_K$ . The entire block is labeled  $f_d$  at the bottom.

EE 533, METU - YILMAZ

16

## Decoding regions

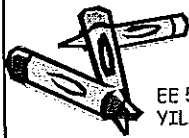
$$R_i = \{\underline{y} \in A_Y^N : f_c(\underline{y}) = \underline{x}_i\}$$



- Error probability using decoding regions

$$P_{PER} = \sum_{i=1}^{2^K} 2^{-K} P_{PER}(\underline{x}_i) \rightarrow \text{equally likely assumption}$$

$$P_{PER}(\underline{x}_i) = \sum_{\underline{y} \notin R_i} p(\underline{y} | \underline{x}_i)$$



EE 533, METU -  
YILMAZ

17

$\rightarrow f_c(\underline{y}) = \underline{x}_i$  if  $p(\underline{x}_i | \underline{y}) \geq p(\underline{x}_j | \underline{y})$ , for all  $j$

- Decoding regions are determined by the decoding rule.
- The optimal (w.r.t. minimum error probability) decoding rule is the maximum a posteriori (MAP) rule.



Equally likely data

$$f_c(\underline{y}) = \underline{x}_i \text{ if } p(\underline{y} | \underline{x}_i) \geq p(\underline{y} | \underline{x}_j), \text{ for all } j$$

Maximum likelihood (ML) rule



EE 533, METU -  
YILMAZ

18

$$\rightarrow p(\underline{x} | \underline{y}) = \frac{p(\underline{y} | \underline{x}) p(\underline{x})}{p(\underline{y})}$$

$$f_c(\underline{y}) = \underline{x}_i \text{ if } p(\underline{y} | \underline{x}_i) \geq p(\underline{y} | \underline{x}_j), \text{ for all } j$$

- ML rule determines the decoding regions and thus the error performance.

$$C = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{2^k}\}$$

- Performance of a code is determined by its codebook.
- The essential problem in channel coding: choosing a good codebook



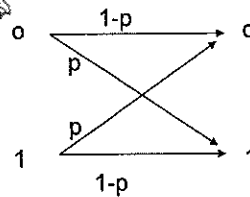
EE 533, METU -  
YILMAZ

19

## Block coding for BSC

$$f_c(\underline{y}) = \underline{x}_i \text{ if } p(\underline{y} | \underline{x}_i) \geq p(\underline{y} | \underline{x}_j), \text{ for all } j$$

$$p(\underline{y} | \underline{x}) = \prod_{i=1}^N p(y_i | x_i) \rightarrow \text{memoryless}$$



$$p(\underline{y} | \underline{x}) = p^{d_H(\underline{y}, \underline{x})} (1-p)^{N-d_H(\underline{y}, \underline{x})}$$

$d_H(\underline{y}, \underline{x})$  = number of places they differ  
Hamming distance



EE 533, METU -  
YILMAZ

20

Hamming distance = 1  
↑  
000 → 010  
probability  
= (1-p) p (1-p)

$$p(\underline{y} | \underline{x}) = \left( \frac{p}{1-p} \right)^{d_H(\underline{y}, \underline{x})} (1-p)^N$$



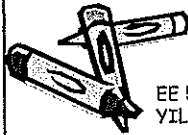
• The ML rule simplifies to  $(0 \leq p < 1/2)$

if  $p > 1/2$   
we can invert  
"

## Error correcting capability

- A code corrects an error pattern  $\underline{n}$  if  $\underline{x} + \underline{n}$  is decoded to  $\underline{x}$  for every  $\underline{x}$ .
- A code is capable of correcting  $e$  errors if all error patterns having  $e$  or fewer errors can be corrected.
- A code has error correcting capability  $t$  if it is  $t$  error correcting but not  $t+1$  error correcting.

$$P_{PER} \leq \sum_{e=t+1}^N \binom{n}{e} p^e (1-p)^{N-e}$$



EE 533, METU -  
YILMAZ

What happens when  $t$  is large?

23

- Relation between the minimum distance and error correcting capability

$$d_{H,min} = \min_{\substack{\underline{w}, \underline{v} \in C \\ \underline{w} \neq \underline{v}}} d_H(\underline{w}, \underline{v})$$

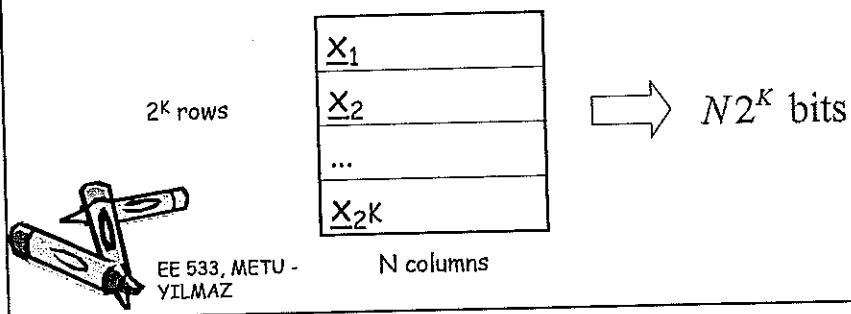
$$t = \left\lfloor \frac{d_{H,min} - 1}{2} \right\rfloor$$

$$\text{large } t \Leftrightarrow \text{large } d_{H,min}$$

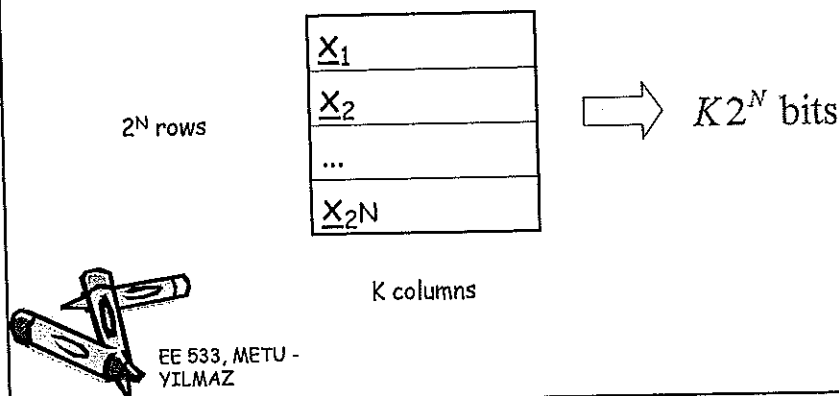


## Complexity

- A conceptually simple way to do encoding and decoding is the use of tables.
- Encoding



- Decoding
  - Checking the encoding table to find the closest codeword
  - Another table for decoding





- $N=100, K=50$

- Encoding table  $100.2^{50}$  bits

- Decoding table  $50.2^{100}$  bits

- Truly astronomical numbers!!!
- Rather than this brute force method, we need structure in codes which allows for computation.



EE 533, METU -  
YILMAZ

27

## Linear codes

- A codebook  $C$  is said to be linear if the modulo-2 sum (bit-by-bit) of any of its codewords is also a codeword.

$C$  is linear



$$\underline{w} \oplus \underline{v} \in C \text{ for all } \underline{w}, \underline{v} \in C$$

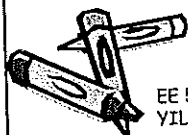


EE 533, METU -  
YILMAZ

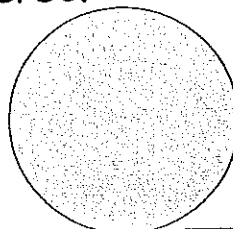
28

All zeros here for all linear codes → Add it by itself = 0

- Example:  $C = \{00000, 00111, 11100, 11011\}$
- Linearity implies that if one of the codewords is subtracted from all the codewords in the codebook, the obtained set is the codebook again.
- This is like you are in such a universe that the universe looks the same from each point of the universe.



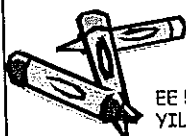
EE 533, METU -  
YILMAZ



29

- Linear spaces and thus linear algebra is used for the construction of these codes.
- $C$  is considered as a subspace.
- Let  $\{\underline{g}_1, \underline{g}_2, \dots, \underline{g}_K\}$  be a basis for  $C$ .
- Given  $\underline{w} = (w_1, \dots, w_K)$

$$f_e(\underline{w}) = w_1 \underline{g}_1 \oplus w_2 \underline{g}_2 \dots \oplus w_K \underline{g}_K$$



EE 533, METU -  
YILMAZ

30

$$f_e(\underline{w}) = w_1 \underline{g}_1 \oplus w_2 \underline{g}_2 \dots \oplus w_K \underline{g}_K$$

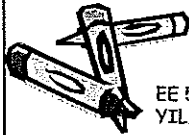
$$G = \begin{bmatrix} \underline{g}_1 \\ \vdots \\ \underline{g}_K \end{bmatrix}$$

generator matrix

$$f_e(\underline{w}) = \underline{w}G$$



KN multiplications and N(K-1) additions



EE 533, METU -  
YILMAZ

31

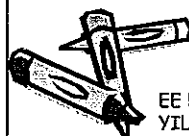
## Linear Block Code Example

- Hamming code

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$w_1 \oplus w_2 \oplus w_3$        $w_1 \oplus w_3 \oplus w_4$

$w_1 \oplus w_2 \oplus w_4$



EE 533, METU -  $\underline{w} = [w_1 \ w_2 \ w_3 \ w_4]$   $f_e(\underline{w}) = \underline{w}G$  32

- There exist efficient algorithms to decode block codes.
- However, most of these algorithms can work with binary inputs (hard-decision decoding).

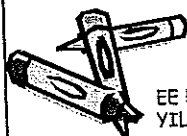
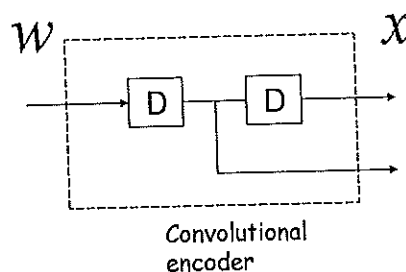


EE 533, METU -  
YILMAZ

33

## Convolutional Codes

- Codewords are obtained by the use of linear finite-state shift registers.



EE 533, METU -  
YILMAZ

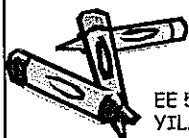
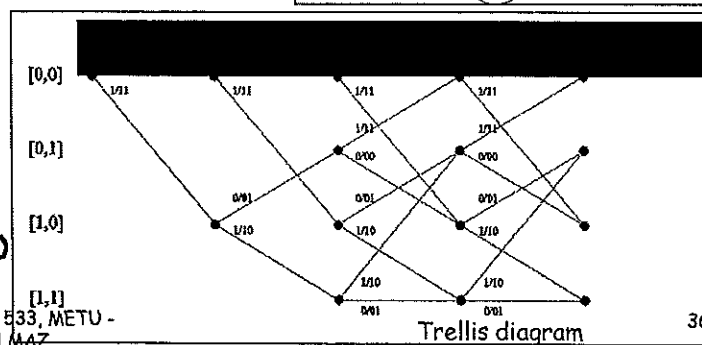
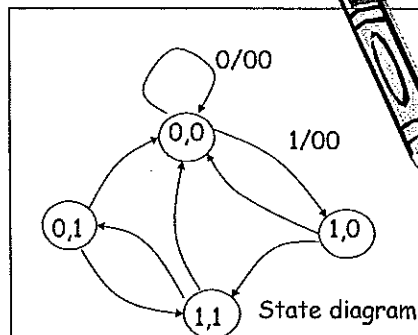
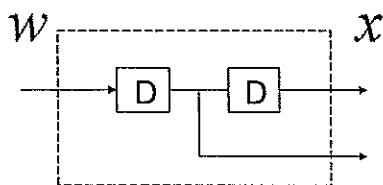
34

- Why convolutional codes?
  - Simple encoding implementation (even simpler than matrix multiplication)
  - Good code properties
  - Efficient algorithms, especially soft-decision
  - Easier to play with the rates
- Properties
  - Linear (since formed by linear shift registers)
  - All the properties regarding minimum distance and  $t$  hold.



EE 533, METU -  
YILMAZ

35



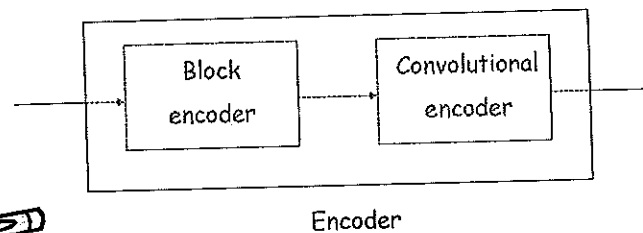
EE 533, METU -  
YILMAZ

Trellis diagram

36

## Concatenated Codes

- Strengths of different codes are sometimes combined to obtain more powerful codes.



EE 533, METU -  
YILMAZ

37

## Random-Like Codes

- As we will see, good codes should have a high degree of randomness.
- There have been great advances in the study of random-like codes in the last two decades.
  - Turbo codes
  - LDPC codes
  - Repeat accumulate codes



EE 533, METU -  
YILMAZ

38

## 38. PROPERTIES OF CHANNEL CAPACITY

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

- (1)  $C \geq 0$  since  $I(X; Y) \geq 0$
- (2)  $C = \max_{p(x)} I(X; Y) = \max_{p(x)} H(X) - H(X|Y) \leq \max_{p(x)} H(X) \leq \log |A_X|$
- (3)  $C \leq \log |A_Y|$ .
- (4)  $I(X; Y)$  is a continuous function of  $p(x)$ .
- (5)  $I(X; Y)$  is a concave function of  $p(x)$ . For example,  $f(x) = 10 - (x - 2)^2$  is a concave function. It means that you can find a maximum. Notice that  $C$  is a functional of  $p(x)$ .

## 39. PREVIEW AND DEFINITIONS

Consider a DMC channel meaning that (36.1) holds and we use the representation shown in Figure 36.1.

## 39.1. Channel Code.

**Definition 22** (Channel Code).

An  $(M, n)$  code for the channel consists of the followings. Note that a channel is characterized by  $(A_X, p(y|x), A_Y)$ .

- (1) An index set for the messages. We have  $M$  messages:  $\{1, 2, \dots, M\}$ .
- (2) An **encoding function**  $x^n : \{1, 2, \dots, M\} \rightarrow A_X^n$  yielding codewords  $x^n(1), x^n(2), \dots, x^n(m)$ . The set of codeword is called **codebook**.
- (3) A **decoding function**  $g : A_Y^n \rightarrow \{1, 2, \dots, M\}$  which is a deterministic rule that assigns a message estimate for each received vector.

**Example 23.**

Let  $n = 3$ ,  $M = 3$ . Let index set be  $\{1, 2, 3\}$ . An example encoding function is  $1 \rightarrow 000$ ,  $2 \rightarrow 011$ ,  $3 \rightarrow 101$ . Decoding function is  $000 \rightarrow 1$ ,  $001 \rightarrow 3$ ,  $\dots$ ,  $111 \rightarrow 2$ .

## 39.2. Conditional Probability of Error.

**Definition 23** (Conditional Probability of Error).

Conditional probability of error given message  $w$  is sent.

$$\lambda_w = P(g(Y^n) \neq w | X^n = x^n(w))$$

Let's do some manipulations.

$$\begin{aligned} P(g(Y^n) \neq w) &= \sum_{y^n: g(y^n) \neq w} p(y^n) \\ (39.1) \quad &= \sum_{y^n} p(y^n) I\{g(y^n) \neq w\} \end{aligned}$$

Note that we use indicator function in (39.1).  $I\{x\}$  is 1 if  $x$  is correct, otherwise it is 0.

$$\lambda_w = \sum_{y^n} p(y^n | x^n(w)) I\{g(y^n) \neq w\}$$

**39.3. Decoding Region.****Definition 24** (Decoding Region).

Decoding region for message  $w$  is defined as the following. Then, we can add another expression for conditional probability of error definition.

$$\begin{aligned} D_w &= \{y^n : g(y^n) = w\} \\ \lambda_w &= P(Y^n \notin D_w | X^n = x^n(w)) \end{aligned}$$

**39.4. Maximal Probability of Error.****Definition 25** (Maximal Probability of Error).

$$\lambda^{(n)} = \max_{w \in \{1, 2, \dots, M\}} \lambda_w$$



**Definition 26** (Average Probability of Error).

$$P_e^{(n)} = \frac{1}{M} \sum_{w=1}^M \lambda_w.$$

When  $W$  is uniformly distributed then,

$$P(W \neq g(Y^n)) = P_e^{(n)}$$

*Proof:*

$$\begin{aligned} (39.2) \quad P(W \neq g(Y^n)) &= \sum_w P(W \neq g(Y^n), X^n = x^n(w)) \\ &= \sum_w P(W \neq g(Y^n) | X^n = x^n(w)) P(X^n = x^n(w)) \\ &= \sum_w \lambda_w \frac{1}{M} \\ &= \lambda_w \end{aligned}$$

Notice that we do marginalization in (39.2). Also  $(X^n = x^n(w)) \equiv (W = w)$ .

**Highlight 20.**

$$\lambda^{(n)} \geq P_e^{(n)}$$

### 39.6. Code Rate, Achievable Rate and Operational Capacity.

**Definition 27** (Code Rate).

The rate of an  $(M, n)$  code,

$$R = \frac{\log_2 M}{n}$$

bits per transmission channel use.

If we have 8 messages,  $M = 8$  before coding each message occupy 3 bits.  
 $M$  is number of messages, not number of bits.

**Definition 28** (Achievable Rate).

A rate  $R$  is said to be achievable if there exists a sequence of  $(\lceil 2^{nR} \rceil, n)$  codes such that  $\lambda^n$  tends to 0 (reliable transmission) as  $n \rightarrow \infty$ . Notice that there is a sequence of codes,  $C_n$ s, not scalar sequence like  $a_n = 1 + 1/n$ .

**Definition 29** ("Operational" Capacity).

The ("operational") capacity of a DMC is the supremum ( $\sim$ maximum) of all achievable rates.

As a side note, basic difference between "supremum" and "maximum" term is as follows: Consider a sequence  $a_n = 10 - 1/n$ . Then,  $a_1 = 10$ ,  $a_2 = 9.5$ . Maximum  $a_n$  should be in set of  $a_n$ 's, this is the definition. With limit idea, we say that its maximum is 10 but it is not because 10 is not in the set. Its supremum is 10 and supremum means *least upper bound*.

## 40. JOINTLY TYPICAL SEQUENCES

Now, we are back to AEP ideas.

**Definition 30** (Jointly Typical Sequences).

$$A_{\epsilon, XY}^{(n)} = \{(x^n, y^n) \in A_X^n \times A_Y^n : \\ \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon\}.$$

**Theorem 18** (Joint AEP).

Let  $(X^n, Y^n)$  be sequences of length- $n$  drawn according to  $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$ . We assume that  $x_i$ 's are independent and we have DMC. Then,

- (1)  $P\left((X^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) \rightarrow 1$  as  $n \rightarrow \infty$ .
- (2)  $|A_{\epsilon, XY}^{(n)}| \leq 2^{n(H(X, Y) + \epsilon)}$ .
- (3) Let  $(\tilde{X}^n, Y^n) \sim p(x^n)p(y^n)$ , i.e.,  $\tilde{X}^n$  and  $Y^n$  are independent but  $\tilde{X}^n$  has the same marginal distribution  $p(x^n)$  as  $X^n$ . It corresponds to someone next to generator of  $X^n$  generates another sequence  $\tilde{X}^n$ .
  - (a)  $P\left((\tilde{X}^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) \leq 2^{-n(I(X; Y) - 3\epsilon)}$
  - (b)  $P\left((\tilde{X}^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) \geq (1 - \epsilon)2^{-n(I(X; Y) + 3\epsilon)}$

Notice that at the last statements probabilities go to 0 since mutual information is non-negative as  $n$  goes to  $\infty$ . We say that  $\tilde{X}^n$  and  $X^n$  are independent (can we say this from independence of  $\tilde{X}^n$  and  $Y^n$ ?). If  $(X^n, Y^n)$  are in typical set and since two  $X$ s are independent,  $(\tilde{X}^n, Y^n)$  are not in typical set.

**Part 15. Lecture 15 - 12.04.2016****40.1. Proofs. Proof 1:**

First, prove the first theorem which is  $P\left((X^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) \rightarrow 1$  as  $n \rightarrow \infty$ .

Recall that if we say  $P(A) > 1 - \epsilon_1$ ,  $P(B) > 1 - \epsilon_2$  and  $P(C) > 1 - \epsilon_3$  then,  $P(A \cap B \cap C) > 1 - (\epsilon_1 + \epsilon_2 + \epsilon_3)$ .

$$\begin{aligned}
 T_1 &= \left\{ (x^n, y^n) : p(x^n) = 2^{-n(H(X) - \epsilon)} \right\} \\
 T_2 &= \left\{ (x^n, y^n) : p(y^n) = 2^{-n(H(Y) - \epsilon)} \right\} \\
 T_3 &= \left\{ (x^n, y^n) : p(x^n, y^n) = 2^{-n(H(X, Y) - \epsilon)} \right\} \\
 P(T_1) &= P\left(\left\{ (x^n, y^n) : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \right\}\right) \\
 (40.1) \quad &= P\left(\left| -\frac{1}{n} \log p(X^n) - H(X) \right| < \epsilon\right) > 1 - \epsilon_1, \text{ as } n \rightarrow \infty
 \end{aligned}$$

(40.1) can be shown similarly for  $T_2$  as in (40.2).

$$(40.2) \quad P(T_2) > 1 - \epsilon_2, \text{ as } n \rightarrow \infty$$

Let's form a super random variable as  $Z = (X^n, Y^n)$ , then we can write as follows

$$P(T_3) = P\left(\left|-\frac{1}{n}\log p(Z) - H(Z)\right| < \epsilon\right) > 1 - \epsilon_3, \text{ as } n \rightarrow \infty$$

$$P(T_3) = P\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X, Y)\right| < \epsilon\right) > 1 - \epsilon_3, \text{ as } n \rightarrow \infty$$

Then, we can write

$$\begin{aligned} A_{\epsilon, XY}^{(n)} &= T_1 \cap T_2 \cap T_3 \\ &> 1 - (\epsilon_1 + \epsilon_2 + \epsilon_3) \\ &> 1 - \epsilon \end{aligned}$$

**Proof 2:**

$$\begin{aligned} |A_{\epsilon, XY}^{(n)}| &\leq 2^{n(H(X, Y) + \epsilon)} \\ 1 &= \sum_{x^n, y^n} p(x^n, y^n) \\ &\geq \sum_{x^n, y^n \in A_{\epsilon, XY}^{(n)}} p(x^n, y^n) \end{aligned}$$

We know that  $p(x^n, y^n) \geq 2^{-n(H(X, Y) + \epsilon)}$ .

$$\begin{aligned} 1 &\geq 2^{-n(H(X, Y) + \epsilon)} \sum_{x^n, y^n \in A_{\epsilon, XY}^{(n)}} 1 \\ &\geq 2^{-n(H(X, Y) + \epsilon)} |A_{\epsilon, XY}^{(n)}| \\ |A_{\epsilon, XY}^{(n)}| &\leq 2^{n(H(X, Y) + \epsilon)} \end{aligned}$$

**Proof 3:**

Remember that  $\tilde{X}^n$  is independent from  $X^n$  and  $Y^n$ . Notice that PMF of  $\tilde{X}^n$  is same as PMF of  $X^n$ .

$$\begin{aligned}
P\left((\tilde{X}^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) &= \sum_{\tilde{x}^n, y^n \in A_{\epsilon, XY}^{(n)}} p_{\tilde{X}^n, Y^n}(\tilde{x}^n, y^n) \\
&= \sum_{\tilde{x}^n, y^n \in A_{\epsilon, XY}^{(n)}} p_{\tilde{X}^n}(\tilde{x}^n) p_{Y^n}(y^n) \\
&\leq \sum_{\tilde{x}^n, y^n \in A_{\epsilon, XY}^{(n)}} 2^{-n(H(X) - \epsilon)} 2^{-n(H(Y) - \epsilon)} \\
&\leq 2^{-n(H(X) + H(Y) - 2\epsilon)} \sum_{\tilde{x}^n, y^n \in A_{\epsilon, XY}^{(n)}} 1 \\
&\leq 2^{-n(H(X) + H(Y) - 2\epsilon)} \left| A_{\epsilon, XY}^{(n)} \right|
\end{aligned}$$

Since  $\left| A_{\epsilon, XY}^{(n)} \right| \leq 2^{n(H(X, Y) + \epsilon)}$ ,

$$\begin{aligned}
P\left((\tilde{X}^n, Y^n) \in A_{\epsilon, XY}^{(n)}\right) &\leq 2^{-n(H(X) + H(Y) - 2\epsilon - H(X, Y) - \epsilon)} \\
&\leq 2^{-n(I(X; Y) - 3\epsilon)}
\end{aligned}$$

Notice that in general  $I(X; Y) \geq 0$ . We can say that  $I(X; Y) > 0$  for a meaningful channel, otherwise channel input and output becomes completely independent. This type of channel is meaningless for us obviously. Also note that channel capacity,  $C$ , is maximum of the mutual information. So, we talk about a non-zero value.

#### 41. THE CHANNEL CODING THEOREM (FUNDAMENTAL THEOREM IN INFORMATION THEORY)

##### **Theorem 19** (The Channel Coding Theorem).

This theorem, a.k.a. Fundamental Theorem in Information Theory, states that all rates below capacity  $C$  are achievable. That is, for every rate  $R < C$ , there exists a sequence of  $(\lceil 2^{nR} \rceil, n)$  codes with  $\lambda^{(n)} \rightarrow 0$ . Conversely, which is very important and makes this theorem powerful, any sequence of  $(\lceil 2^{nR} \rceil, n)$  codes with  $\lambda^{(n)} \rightarrow 0$  must have  $R < C$ .

#### 42. PROOF OF THE CHANNEL CODING THEOREM

##### 42.1. Outline of The Proof.

- (1) Fix  $p(x)$  (arbitrarily) and assume that  $p(y|x)$  is given, in other words we know the channel.
- (2) Find the expected error probability,  $P_e^{(n)} = E_C \left[ P_e^{(n)}(C) \right]$  over all possible codebooks. We will show  $P_e(n) < \epsilon$  if  $R < I(X; Y) - \Delta$ ,  $\Delta > 0$ .
- (3) A few more steps in order to distinguish a sequence of codes with  $\lambda^{(n)} \rightarrow 0$ .
- (4) Utilize  $p(x)$  maximizing  $I(X; Y)$ .

## 42.2. Encoding and Decoding in The Proof.

- (1) **Random Coding:** A random code  $c$  is generated by randomly generating codewords. We don't know whether is a good code or not. Even, it may not be a one-to-one code which is meaningless for our purposes. They are randomly generating. But we will see that occurrence of a codeword more than once does not hurt us at the end of the day.

$$p(x^n) = \prod_{j=1}^n p(x_j), \quad x^n \in A_X^n$$

Generate  $2^{nR}$  codewords independently and write them into a matrix.

$$c = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & \dots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix}_{2^{nR} \times n}$$

How many different  $c$ , codebooks, exists? We can generate  $|A_X|^{n2^{nR}}$  different codewords. If we take  $n = 1000$  which larger in practical systems and take  $R = 1/4$  arbitrarily there are  $2^{1000 \cdot 2^{1000 \times 1/4}}$  different possibilities which is a very HUGE number. Notice that we are not saying that all codebooks are equally likely. Probability of a codeword is

$$\begin{aligned} p(c) &= \prod_{w=1}^{2^{nR}} \prod_{j=1}^n p(x_j(w)) \\ &= \prod_{w=1}^{2^{nR}} p(x^n(w)) \end{aligned}$$

**Example 24.**

$A_X = \{0, 1\}$ ,  $p(0) = 1/6$ ,  $n = 2$ ,  $R = 1/2$ .

$$c = \begin{bmatrix} \dots & \dots \\ \dots & \dots \end{bmatrix}_{2 \times 2}$$

$$P\left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right) = \left(\frac{1}{6}\right)^4$$

$$P\left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right) = \left(\frac{1}{6}\right)^3 \frac{5}{6}$$

As we can see, they are not equally likely.

- (2) The codebook is revealed to the sender and receiver. They also know  $p(y|x)$ .  
 (3) A message  $W$  is chosen according to uniform distribution meaning that

$$P(W = w) = 2^{-nR}, \quad w = 1, \dots, 2^{nR}.$$

**Highlight 21.**

We assume that a "perfect" source coding block compress the original message, to its entropy, prior to the channel coder. If messages do not have uniform distribution, it means that message can be compressed further.

- (4) The codeword is  $x^n(w)$ .  
 (5) A sequence  $Y^n$  is received with distribution:

$$p(y^n|x^n(w)) = \prod_{j=1}^n p(y_j|x_j(w))$$

which corresponds to a DMC.

- (6) Receiver performs typical set decoding rather than the optimal MAP decoding due to avoid difficulty in error probability analysis. Notice that we are away from the optimal case. Here is the decoding rule: The estimate is  $\hat{W}$  if
- $(x^n(\hat{W}), y^n) \in A_{\epsilon, XY}^{(n)}$  AND
  - There is no other index  $k$  such that  $(x^n(k), y^n) \in A_{\epsilon, XY}^{(n)}$
- A flag is raised if
- No such  $k$  such that  $(x^n(k), y^n) \in A_{\epsilon, XY}^{(n)}$  OR
  - More than one  $k$  such that  $(x^n(k), y^n) \in A_{\epsilon, XY}^{(n)}$
- (7) There is a correct decision if  $\hat{W} = W$  and incorrect if  $\hat{W} \neq W$  OR flag is raised.

**Part 16. Lecture 16 - 14.04.2016**

We evaluate the probability of error over all possible codebooks.

$$(42.1) \quad \begin{aligned} P_r(\epsilon) &= E_C \left[ P_e^{(n)}(C) \right] \\ &= \sum_c p(c) P_e^{(n)}(c) \end{aligned}$$

Notice that in (42.1),  $p(c)$  stands for probability of a specific code  $c$  and  $P_e^{(n)}$  stands for average probability of error for code  $c$  and we sum for all codes  $(c)$ .

From (42.1),

$$(42.2) \quad P_r(\epsilon) = \sum_c p(c) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(c)$$

$$(42.3) \quad = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_c p(c) \lambda_w(c)$$

In (42.2), we assume that all messages have equal probability (from  $\frac{1}{2^{nR}}$  term).

Let's consider inner summation in (42.3) and expand as in (42.4).

$$(42.4) \quad \sum_c p(c) \lambda_w(c) = E_C [\lambda_W(C)]$$

$$(42.5) \quad = \sum_c p(c) P(g(Y^n) \neq w | c, W = w)$$

$$(42.6) \quad = P(g(Y^n) \neq 1 | W = 1)$$

How transition from (42.5) to (42.6) is possible? Here is the answer: Does  $w$  matters? We are investigating all the possible codes. Let's consider the following example:

**Example 25.**

Let's consider two different codebooks:  $c_1$  and  $c_2$ . Notice that their probabilities are equal.

$$c_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, c_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Now let's consider that there is an error in first codeword. For  $c_1$  case, we transmit 00 but receive 01 and similarly for  $c_2$  we transmit 01 but receive 00. Now consider that there is an error in second codeword. For  $c_1$  case, we transmit 01 but receive 00 and similarly for  $c_2$  we transmit 00 but receive 01. It turns out to be that probability of error in first codeword is equal to probability of error in second codeword. Everything is symmetric is here. That's why we are fine with considering a single codeword.

Now combine (42.3) and (42.6) as (42.7).

$$(42.7) \quad \begin{aligned} P_r(\epsilon) &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P(g(Y^n) \neq 1 | W = 1) \\ &= P(g(Y^n) \neq 1 | W = 1) \end{aligned}$$

Let's say that we send  $X^n(1)$  and receive  $Y^n$ . Notice that  $X^n(1)$  is still random variable.

$$C = \begin{bmatrix} X^n(1) \\ X^n(2) \\ \vdots \\ X^n(2^{nR}) \end{bmatrix}$$



$$\begin{aligned}
P_r(\epsilon) &= P(g(Y^n) \neq 1 | W = 1) \\
&= P(((X^n(1), Y^n) \notin A_{\epsilon, XY}^{(n)}) \text{ OR} \\
&\quad ((X^n(2), Y^n) \in A_{\epsilon, XY}^{(n)}) \text{ OR} \\
&\quad ((X^n(3), Y^n) \in A_{\epsilon, XY}^{(n)}) \text{ OR} \\
&\quad \vdots \\
&\quad ((X^n(2^{nR}), Y^n) \in A_{\epsilon, XY}^{(n)}))
\end{aligned}$$

Let's remember union bound which is  $P(A) + P(B) \leq P(A \cup B)$ .

$$(42.8) \quad P_r(\epsilon) \leq P((X^n(1), Y^n) \notin A_{\epsilon, XY}^{(n)}) + \sum_{w=2}^{2^{nR}} P((X^n(w), Y^n) \in A_{\epsilon, XY}^{(n)})$$

Remember that  $P((X^n(1), Y^n) \notin A_{\epsilon, XY}^{(n)}) \leq \epsilon$  and  $P((X^n(w), Y^n) \in A_{\epsilon, XY}^{(n)}) \leq 2^{-n(I(X;Y) - 3\epsilon)}$ .  
(42.8) can be continued as (42.9)

$$(42.9) \quad P_r(\epsilon) \leq \epsilon + 2^{-n(I(X;Y) - 3\epsilon)} 2^{nR}$$

$$(42.10) \quad \leq \epsilon + 2^{-n(I(X;Y) - R - 3\epsilon)}$$

Notice that (42.10) becomes  $< \epsilon$  as  $R < I(X;Y)$ .

**42.3. Final Interpretation.** IF  $R < I(X;Y)$ ,  $P_r(\epsilon) = P(\epsilon | W = 1) \leq 2\epsilon$ . When  $p(x)$  maximizing  $I(X;Y)$  is used,  $C = \max_{p(x)} I(X;Y)$ , then  $p(\epsilon) \rightarrow 0$  as  $R < C$ ,  $n \rightarrow \infty$ . We have found average error. But we are interested in maximum probability of error.

If  $P(\epsilon) \leq 2\epsilon$ , then there exists at least one codebook with  $P_e^{(n)}(c) \leq 2\epsilon$ . For example think that mean of 10 different, non-negative value is 7. One of them should be at least smaller than or equal to 7. Otherwise, mean can't be 7. Proof by contradiction.

Now for this codebook which is  $P_e^{(n)}(c) \leq 2\epsilon$ , at least half of the codewords should have codeword error probability  $\leq 4\epsilon$ . Think like the previous case. If mean value of 10 different non-negative value is less than equal to 7, their sum is less than equal to 70. Therefore, sum of 5 of them should be less than equal to 70 meaning than they should be less than equal to 14 individually. Let's put these codewords (half of them) into a new codebook,  $c^\theta$ . The rate of code  $c^\theta$  is denoted by  $R^\theta$  which is

$$\begin{aligned}
R^\theta &= \frac{\log(2^{nR}/2)}{n} \\
&= R - \frac{1}{n}
\end{aligned}$$

But  $R^\theta \rightarrow R$  as  $n \rightarrow \infty$  meaning that  $R^\theta < C$ . In this code,  $c^\theta$ , maximal error probability is less than equal to  $4\epsilon$ .

Overall, we showed that there exists a code with rate  $R^\theta = R - 1/n$  and  $\lambda^{(n)} \leq 4\epsilon$  for  $R < C$ . The supremum of  $R^\theta$  equals to  $C$ . This is why operational capacity was defined as the supremum of achievable rate.

## Part 17. Lecture 17 - 19.04.2016

### 43. FANO'S INEQUALITY AND THE CONVERSE

43.1. **Lemma A (Another form of Fano's Inequality)**. For a DMC with a codebook  $C$  with rate  $R$  and input messages equally likely. Let  $P_e^{(n)} = P(W \neq g(Y^n))$ , then

$$H(X^n|Y^n) \leq 1 + P_e^{(n)} nR$$

**Proof:**

Consider a black box system with input  $S$  and output  $T$ . Then, Fano's inequality says that

$$(43.1) \quad H(S|T) \leq H(P_e) + P_e \log(|A_s| - 1).$$

Let  $S = W$  and  $T = Y^n$ ,

$$H(W|Y^n) \leq H(P_e^{(n)}) + P_e^{(n)} \log(|A_w| - 1)$$

We know that  $H(P_e^{(n)}) \leq 1$  and  $|A_w| = M = 2^{nR}$ , then  $\log(|A_w| - 1) \leq nR$ . Let's consider DPI (Data Processing Inequality), we can say that

$$(43.2) \quad H(X^n|Y^n) \leq H(W|Y^n).$$

From (43.1) and (43.2),

$$H(X^n|Y^n) \leq 1 + P_e^{(n)} nR.$$

43.2. **Lemma B.**  $Y^k$  is result of passing  $X^k$  through a DMC. Then,

$$I(X^k; Y^k) \leq kC \quad \forall p(x^k)$$

**Proof:**

$$(43.3) \quad I(X^k; Y^k) = H(Y^k) - H(Y^k | X^k)$$

$$(43.4) \quad = H(Y^k) - \sum_{i=1}^k H(Y_i | X_i)$$

$$(43.5) \quad \begin{aligned} &\leq \sum_{i=1}^k H(Y_i) - \sum_{i=1}^k H(Y_i | X_i) \\ &\leq \sum_{i=1}^k H(Y_i) - H(Y_i | X_i) \\ &\leq \sum_{i=1}^k I(Y_i; X_i) \\ &\leq kC \end{aligned}$$

Notice that transition from (43.3) to (43.4) is property of DMC. (43.5) comes from independence bound for entropy.

**Highlight 22.**

The capacity per transmission does not increase if a DMC is used many times as opposed to source coding. We are already using the channel many times to achieve capacity.

#### 44. PROOF OF THE CONVERSE

Show that any sequence of  $(2^{nR}, n)$  codes with  $\lambda^{(n)} \rightarrow 0$  must have  $R \leq C$ . We are assuming that messages are equally likely. It means that if  $\lambda^{(n)} \rightarrow 0$ , then  $P_e^{(n)} \rightarrow 0$ .

$$\begin{aligned} H(W) &= H(W | Y^n) + I(W; Y^n) \\ nR &= H(W | Y^n) + I(W; Y^n) \end{aligned}$$

Notice that  $I(W; Y^n) \leq I(X^n; Y^n)$ , then

$$nR \leq H(W | Y^n) + I(X^n; Y^n)$$

From lemma A,  $H(W | Y^n) \leq 1 + P_e^{(n)} nR$  and from lemma B,  $I(X^n; Y^n) \leq nC$ . Then,

$$\begin{aligned}
nR &\leq 1 + P_e^{(n)}nR + nC \\
R &\leq \frac{1}{n} + P_e^{(n)}R + C \\
(44.1) \quad P_e^{(n)} &\geq 1 - \frac{C}{R} - \frac{1}{nR}
\end{aligned}$$

From (44.1), it can be said that as  $n \rightarrow \infty$ ,  $P_e^{(n)} \geq 1 - C/R$ .

In general,  $P_e^{(n)} \simeq 0$  if rate is under the capacity and  $P_e^{(n)}$  increases rapidly as rate becomes larger than the capacity. This is the weak converse. Strong converse says that  $P_e^{(n)} \rightarrow 1$  exponentially fast with  $n$  as  $R > C$ .

#### 45. PROPERTIES OF GOOD CODES

Good code means that we have reliable transmission.

$$\begin{aligned}
nR &= H(W) \\
&= H(W|\hat{W}) + I(W; \hat{W})
\end{aligned}$$

Notice that  $H(W|\hat{W})$  is 0 if  $P_e^{(n)} = 0$ .

$$nR = I(W; \hat{W})$$

$I(W; \hat{W}) \leq I(X^n; Y^n)$  in general. But equality holds if one-to-one encoding is used and we have an informational lossless system, i.e.,  $\hat{W} = g(Y^n)$ .

$$\begin{aligned}
nR &= I(X^n; Y^n) \\
(45.1) \quad &= H(Y^n) - H(Y^n|X^n)
\end{aligned}$$

$$(45.2) \quad = H(Y^n) - \sum_{i=1}^n H(Y_i|X_i)$$

$$(45.3) \quad \leq \sum_{i=1}^n H(Y^i) - \sum_{i=1}^n H(Y_i|X_i)$$

Transition from (45.1) to (45.2) is done using properties of DMC. Equality in (45.3) holds if  $Y_i$ 's are independent.

$$\begin{aligned}
nR &= \sum_{i=1}^n I(X_i; Y_i) \\
(45.4) \quad &\leq nC
\end{aligned}$$

Equality in (45.4) holds if  $p(x)$  is capacity achievement distribution, i.e., maximizes  $I(X_i; Y_i)$ .

**Highlight 23.**

In summary, to have a good code:

- (1) Distinct messages have distinct codewords.
- (2) Capacity achieving distribution  $p(x)$  should be used.
- (3)  $Y_i$ s should be (seem) independent which is possible by independent  $X_i$ s.

**46. THE JOINT SOURCE-CHANNEL CODING THEOREM**

Consider the system shown in Figure 46.

**Theorem 20** (The Joint Source-Channel Coding Theorem).

If  $V_1, V_2, \dots, V_n$  is an IID (not have to be independent indeed, it is valid for more general case.) random sequence, then there exists a source-channel code with  $P_e^{(n)} \rightarrow 0$  if  $H(V) < C$ . Conversely, for any sequence if  $H(V) > C$ , then  $P_e^{(n)} > \alpha > 0$ .

Idea is the combination of source coding + channel coding into single encoder. Source encoding compresses and channel coding decompresses generally. It says that we can think source coding and channel coding separately as  $n$  goes to  $\infty$ .

**Proof:** There is a typical set with  $|A_{\epsilon, V}^{(n)}| < 2^{n(H(V)+\epsilon)}$ . We will only encode the sequences in  $A_{\epsilon, V}^{(n)}$ . Then,

$$\begin{aligned} R &= \frac{1}{n} \log 2^{n(H(V)+\epsilon)} \\ &= H(V) + \epsilon \end{aligned}$$

If  $R < C$  ( $H(V) + \epsilon < C$ ) then,

$$\begin{aligned} P_e^{(n)} &= P(V^n \neq \hat{V}^n) \\ &= P(V^n \neq \hat{V}^n | V^n \in A_{\epsilon, V}^{(n)}) P(V^n \in A_{\epsilon, V}^{(n)}) \\ &\quad + P(V^n \neq \hat{V}^n | V^n \notin A_{\epsilon, V}^{(n)}) P(V^n \notin A_{\epsilon, V}^{(n)}) \\ (46.1) \quad &\leq P(V^n \neq \hat{V}^n | V^n \in A_{\epsilon, V}^{(n)}) + P(V^n \neq \hat{V}^n | V^n \notin A_{\epsilon, V}^{(n)}) \\ (46.2) \end{aligned}$$

Notice that transition from (46.1) to (46.2) is very simple since a probability is always less than or equal to 1 (second multiplicands in (46.1)).

Notice that since we have a reliable transmission  $P(V^n \neq \hat{V}^n | V^n \in A_{\epsilon, V}^{(n)}) < \epsilon$  and having a non-typical set is similar to  $P(V^n \neq \hat{V}^n | V^n \notin A_{\epsilon, V}^{(n)}) < \epsilon$ . Then for sufficiently large  $n$ ,

$$P_e^{(n)} \leq \epsilon.$$

**Proof of The Converse:** Show that if  $P_e^{(n)} \rightarrow 0$  then,  $H(V) < C$  for any  $X^n(V^n) : A_V^n \rightarrow A_X^n$  and  $g(Y^n) : A_Y^n \rightarrow A_V^n$ .

$$(46.3) \quad \frac{H(V^n)}{n} = \frac{1}{n} H(V^n | \hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n)$$

$$(46.4) \quad \leq \frac{1}{n} \left( 1 + P_e^{(n)} n \log |A_V| \right) + \frac{1}{n} I(V^n; \hat{V}^n)$$

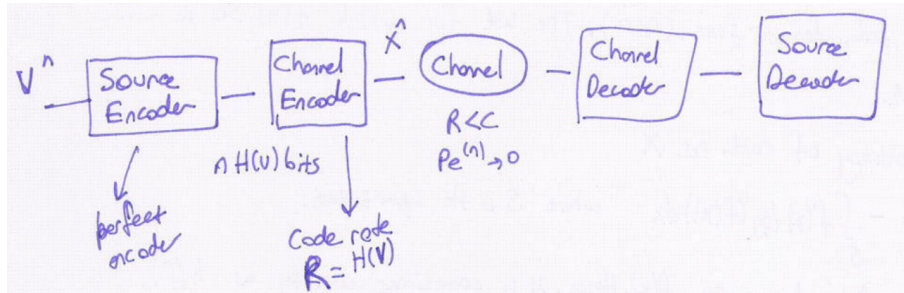
$$(46.5) \quad \leq \frac{1}{n} \left( 1 + P_e^{(n)} n \log |A_V| \right) + \frac{1}{n} I(X^n; Y^n)$$

$$(46.6) \quad \leq \frac{1}{n} \left( 1 + P_e^{(n)} n \log |A_V| \right) + C$$

$$(46.7) \quad H(V) \leq \frac{1}{n} + P_e^{(n)} \log |A_V| + C$$

Transition from (46.3) to (46.4) is by Lemma A, from (46.4) to (46.5) is by DPI, from (46.5) to (46.6) is by Lemma B and from (46.6) to (46.7) is by independence bound for entropy.

There for as  $n \rightarrow \infty$  and  $P_e^n \rightarrow 0$ , then  $H(V) < C$ . Notice that we didn't specify any coding (random coding etc.). This result shows that source coding and channel coding are separable. The condition is always  $H(V) < C$ .



Part 18. No Lecture on 21.04.2016

Part 19. Lecture 18 - 26.04.2016

#### 47. DIFFERENTIAL ENTROPY

It is extension of entropy to continuous random variable.

**Definition 31** (Continuous Random Variable).

$X$  is a random variable with CDF,  $F(x) = P(X < x)$ . If  $F(x)$  is continuous, then  $X$  is said to be continuous. Let  $f(x)$  be derivative of  $F(x)$ . When the derivative is defined,  $f(x)$  is called probability density function (PDF).

**Definition 32** (Support Set).

The set for which  $f(x) > 0$  is called the support set ( $S$ ).

**Definition 33** (Differential Entropy).

Differential entropy of a continuous random variable  $X$  is defined as

$$h(X) = - \int_S f(x) \log f(x) dx.$$

**Highlight 24.**

Differential entropy depends only on  $f(x)$ . Hence, it is sometimes written as  $h(f)$ .

**Example 26** (Uniform Distribution).

Let  $U$  be uniform over  $(0, a)$ . In other words,

$$f_U(x) = \begin{cases} 1/a & 0 \leq x \leq a \\ 0 & \text{other.} \end{cases}$$

$$h(U) = - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a$$

Notice that  $h(U)$  is **negative (it is entropy!)** if  $a < 1$  and positive otherwise. As  $a \rightarrow 0$ ,  $h(U) \rightarrow -\infty$ ,  $f(x) \rightarrow \delta(x)$  (non-random).

**Highlight 25.**

Although minimum value of entropy is 0, differential entropy may go to  $-\infty$ .

**Example 27** (Gaussian Distribution).

$$\begin{aligned}
X &\sim \Phi(x) = N(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \\
h(X) &= - \int_{-\infty}^{\infty} \Phi(x) \ln \Phi(x) dx \\
&= - \int_{-\infty}^{\infty} \Phi(x) \left[ -\frac{1}{2} \ln 2\pi\sigma^2 + \left( \frac{-x^2}{2\sigma^2} \right) \right] dx \\
&= \frac{1}{2} \ln 2\pi\sigma^2 \int_{-\infty}^{\infty} \Phi(x) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} x^2 \Phi(x) dx \\
&= \frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2\sigma^2} E[X^2] \\
&= \frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2} \\
&= \frac{1}{2} \ln 2\pi e \sigma^2 \text{ nats} \\
&= \frac{1}{2} \log 2\pi e \sigma^2 \text{ bits}
\end{aligned}$$



**Highlight 26.**

If for single random variable volume corresponds to length, for bivariate case corresponds to area and so on.

**Highlight 27.**

WLLN holds for continuous random variables. When  $X_i$ s are drawn I.I.D it can be written as

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - E[g(X)]\right| \leq \epsilon\right) \geq 1 - \epsilon$$

## 50. AEP FOR CONTINUOUS RANDOM VARIABLES

**Theorem 21** (AEP for Continuous Random Variables).

- (1)  $P\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$  for  $n$  sufficiently large.
- (2)  $\text{Vol}\left(A_\epsilon^{(n)}\right) \leq 2^{n(h(X)+\epsilon)}$  for all  $n$ .
- (3)  $\text{Vol}\left(A_\epsilon^{(n)}\right) \geq (1 - \epsilon)2^{n(h(X) - \epsilon)}$

These are similar to discrete case.

**Proofs:**

- (1) It can be proven using WLLN.
- (2) If  $(x_1, \dots, x_n) \in A_\epsilon^{(n)}$  then

$$2^{-n(h(X)+\epsilon)} \leq f(x_1, x_2, \dots, x_n) \leq 2^{-n(h(X) - \epsilon)}$$

Remember that

$$P\left(A_\epsilon^{(n)}\right) \leq 1$$

$$\int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \leq 1$$

$$2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \dots dx_n \leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \leq 1$$

$$2^{-n(h(X)+\epsilon)} \text{Vol}\left(A_\epsilon^{(n)}\right) \leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \leq 1$$

$$2^{-n(h(X)+\epsilon)} \text{Vol}\left(A_\epsilon^{(n)}\right) \leq 1$$

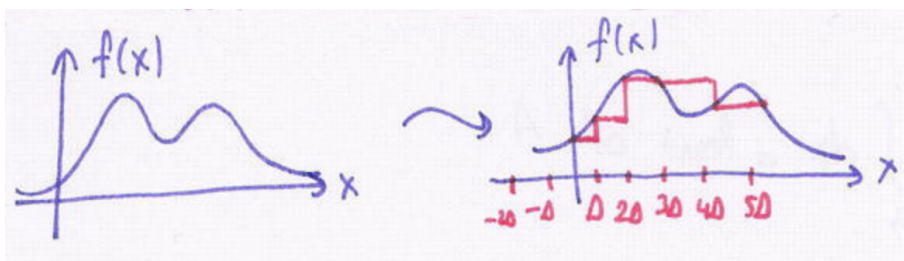
$$\text{Vol}\left(A_\epsilon^{(n)}\right) \leq 2^{n(h(X)+\epsilon)}$$

**Highlight 28.**

$\text{Vol}(A_\epsilon^{(n)}) \simeq 2^{nh}$  as  $n \rightarrow \infty$ .

A hypercube of dimension  $n$  has volume  $a^n$  with side length  $a$ . The side length of a hypercube with volume  $2^{nh}$  :  $l = (2^{nh})^{1/n} = 2^h$ ,  $h = \log l$ . Low entropy means that probability is confined in a small volume.

## 51. DIFFERENTIAL ENTROPY VS DISCRETE ENTROPY



$$p(k) = f(k\Delta)\Delta$$

$$\sum_{k=-1}^1 p(k) = 1, \quad \Delta \rightarrow 0$$

$$\lim_{\Delta \rightarrow 0} \sum_{k=-1}^1 f(k\Delta)\Delta = \int_{-1}^1 f(x) dx = 1$$

Let,

$$X^\Delta : P_{X^\Delta}(k) = f(k\Delta)\Delta$$

then,

**Highlight 29.**

$$h(X) = H(X^\Delta) + \log \Delta \quad \text{as } \Delta \rightarrow 0$$

It can be used to compute differential entropy using discrete approach.

Proof is in the book.

## 52. JOINT DIFFERENTIAL ENTROPY

Let  $X^n = X_1 X_2 \dots X_n$ . Then,

$$X^n \sim f(x_1, \dots, x_n) = f(x^n)$$

$$\begin{aligned}
h(X^n) &= - \int \dots - \int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n \\
&= - \int f(x^n) \log f(x^n) dx^n \\
&= E[-\log f(X^n)]
\end{aligned}$$

**Highlight 30.**

$$h(X^n) = E[-\log f(X^n)]$$

**53. CONDITIONAL DIFFERENTIAL ENTROPY**

Let  $X, Y \sim f(x, y)$

$$\begin{aligned}
h(X|Y) &= - \int \int f(x, y) \log f(x|y) dx dy \\
&= - \int \int f(x, y) \log \frac{f(x, y)}{f(y)} dx dy \\
&= h(X, Y) - h(Y)
\end{aligned}$$

**Highlight 31.**

$$h(X|Y) = h(X, Y) - h(Y)$$

**Theorem 22.**

Let  $X_1, \dots, X_n$  be multivariate normal distribution with mean  $\underline{\mu}$  and covariance matrix  $K$ .  $X^n \sim N(\underline{\mu}, K)$ .

$$h(X^n) = \frac{1}{2} \log ((2\pi e)^n |K|) \text{ bits}$$

where  $|\cdot|$  is the determinant operator.

## 54. RELATIVE ENTROPY AND MUTUAL INFORMATION

**Definition 36** (Relative Entropy (Divergence)).Relative entropy (divergence) between densities  $f$  and  $g$  is defined as

$$D(f||g) = \int_{S_f} f(x) \log \frac{f(x)}{g(x)} dx$$

**Note that**  $D(f||g)$  is finite only if  $S_f \subseteq S_g$ .

## Part 20. Lecture 19 - 28.04.2016

**Definition 37** (Mutual Information).

$$\begin{aligned} I(X; Y) &= \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy \\ &= D(f(x, y) || f(x)f(y)) \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned}$$

**Highlight 32.**

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta|Y^\Delta) \\ &\simeq h(X) + \log \Delta - [h(X|Y) + \Delta] \\ &\simeq h(X) - h(X|Y) \\ &\simeq I(X; Y) \end{aligned}$$

**Theorem 23.**

$$D(f||g) \geq 0$$

and equality holds iff  $f(x) = g(x)$  almost everywhere. "Almost everywhere" means that they are same except one point, for example. In other words having  $f(x) \neq g(x)$  for a set of probability zero is not a problem.

**Proof:**Recall that  $\ln x \leq x - 1$ .

$$\begin{aligned}
-D(f||g) &= \int_{S_f} f(x) \ln \frac{g(x)}{f(x)} dx \\
&\leq \int_{S_f} f(x) \left( \frac{g(x)}{f(x)} - 1 \right) dx \\
&\leq \int_{S_f} f(x) \frac{g(x)}{f(x)} dx - \int_{S_f} f(x) dx \\
&\leq \int_{S_f} g(x) dx - 1 \\
&\leq (\leq 1) - 1
\end{aligned}$$

## 55. BRIEF SUMMARY OF CONTINUOUS TIME RELATIONS

**Highlight 33.**

$$I(X; Y) \geq 0$$

**Highlight 34.**

$$h(X) \geq h(X|Y)$$

**Highlight 35.**

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

**Highlight 36.**

$$h(X_1, \dots, X_n) \leq \sum_{i=1}^n h(X_i)$$

and equality is hold iff  $X_i$ 's are independent.

**Theorem 24.**

$$h(X + c) = h(X)$$

**Theorem 25.**

$$h(aX) = h(X) + \log |a|$$

Notice that entropy does not change in discrete case if  $X$  is scaled.

**Proof:**

Let  $Y = aX$  then,  $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$ .

$$\begin{aligned}
 h(Y) &= - \int_{S_y} f_Y(y) \log f_Y(y) dy \\
 &= - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log \left[ \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \right] dy \\
 &= - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \left[ \log \frac{1}{|a|} + \log f_X\left(\frac{y}{a}\right) \right] dy \\
 &= - \log \frac{1}{|a|} \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) dy - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log f_X\left(\frac{y}{a}\right) dy \\
 &= - \log \frac{1}{|a|} 1 - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log f_X\left(\frac{y}{a}\right) dy \\
 (55.1) \quad &= \log |a| - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log f_X\left(\frac{y}{a}\right) dy \\
 (55.2) \quad &= \log |a| - \int_{S_x} f_X(x) \log f_X(x) dx \\
 &= \log |a| + h(X)
 \end{aligned}$$

Notice that for transition from (55.1) to (55.2),  $x = y/a$  also  $dx = dy/a$

## 56. HADAMARD'S INEQUALITY

**Example 28** (Hadamard's Inequality).

Consider a non-negative definite symmetric matrix  $K$ . Let  $\underline{X} \sim N(0, K)$

$$\begin{aligned}
 h(X_1, \dots, X_n) &\leq \sum_{i=1}^n h(X_i) \\
 \frac{1}{2} \log(2\pi e)^n |K| &\leq \sum_{i=1}^n \frac{1}{2} \log(2\pi e) \sigma_{X_i}^2 \\
 \frac{1}{2} \log(2\pi e)^n |K| &\leq \frac{1}{2} \log \prod_{i=1}^n (2\pi e) K_{ii} \\
 (2\pi e)^n |K| &\leq (2\pi e)^n \prod_{i=1}^n K_{ii} \\
 |K| &\leq \prod_{i=1}^n K_{ii}
 \end{aligned}$$

## Part 21. Lecture 20 - 03.05.2016

**Theorem 26.**

Let the random vector  $\underline{X} \in R^n$  have zero mean covariance matrix  $K = E[\underline{X}\underline{X}^T]$  meaning  $K_{ij} = E[X_i X_j]$ . Then

$$h(\underline{X}) \leq \frac{1}{2} \log((2\pi e)^n |K|).$$

with equality iff  $\underline{X} \sim N(0, K)$ .

**Proof:**

$g$  is an arbitrary PDF with covariance matrix  $K$  and  $\Phi_K$  is PDF of Gaussian random vector with covariance  $K$ .

$$\begin{aligned}
 0 &\leq D(g||\Phi_K) \\
 &\leq \int g \ln \frac{g}{\Phi_K} \\
 &\leq \int g \ln g - \int g \ln \Phi_K \\
 (56.1) \quad &\leq -h(g) - \int g \ln \Phi_K
 \end{aligned}$$

$$\begin{aligned}
\Phi_K(\underline{x}) &= \frac{1}{|2\pi K|^{1/2}} \exp\left(-\frac{1}{2}\underline{x}^T K^{-1}\underline{x}\right) \\
\ln \Phi_K(\underline{x}) &= -\frac{1}{2} \ln |2\pi K| - \frac{1}{2}\underline{x}^T K^{-1}\underline{x} \\
&= -\frac{1}{2} \ln |2\pi K| - \frac{1}{2} \sum_i \sum_j x_i K_{ij}^{-1} x_j \\
(56.2) \quad &= -\frac{1}{2} \sum_i \sum_j K_{ij}^{-1} x_i x_j - \frac{1}{2} \ln |2\pi K|
\end{aligned}$$

$$\begin{aligned}
\int g(\underline{x}) \ln \Phi_K(\underline{x}) d\underline{x} &= \int g(\underline{x}) \left(-\frac{1}{2}\right) \sum_i \sum_j K_{ij}^{-1} x_i x_j d\underline{x} - \int \frac{1}{2} \ln |2\pi K| g(\underline{x}) d\underline{x} \\
(56.3) \quad &= -\frac{1}{2} \sum_i \sum_j K_{ij}^{-1} \int g(\underline{x}) x_i x_j d\underline{x} - \int g(\underline{x}) \frac{1}{2} \ln |2\pi K| d\underline{x} \\
&= -\frac{1}{2} \sum_i \sum_j K_{ij}^{-1} \int \Phi_K(\underline{x}) x_i x_j d\underline{x} - \int \Phi_K(\underline{x}) \frac{1}{2} \ln |2\pi K| d\underline{x}
\end{aligned}$$

$$(56.4) \quad = \int \Phi_K(\underline{x}) \left[ -\frac{1}{2} \ln |2\pi K| - \frac{1}{2} \sum_i \sum_j x_i K_{ij}^{-1} x_j \right] d\underline{x}$$

$$\begin{aligned}
(56.5) \quad &= \int \Phi_K(\underline{x}) \ln \Phi_K(\underline{x}) d\underline{x} \\
&= h(\Phi_K)
\end{aligned}$$

Notice that in (56.3),  $\int g(\underline{x}) x_i x_j d\underline{x}$  is  $E[x_i x_j]$ . Since covariance matrices of both distributions are same replace  $g(\underline{x})$  by  $\Phi_K(\underline{x})$ . Similarly  $\int g(\underline{x}) d\underline{x}$  is 1 and we can replace it too.

Transition from (56.4) to (56.5) is possible with help of (56.2).

Using (56.1),

$$\begin{aligned}
0 &\leq -h(g) + h(\Phi_K) \\
h(g) &\leq h(\Phi_K)
\end{aligned}$$

#### Highlight 37.

For fixed variance(power) random variable, maximum entropy is obtained from Gaussian random variable. Check the HW5-Q7.

## 57. THE GAUSSIAN CHANNEL

Continuous channels can be split into two categories:

- (1) **Discrete Time, Continuous Amplitude (DTCA)** Codewords are in  $R^n$  like 2.5,  $\sqrt{2}$ ,  $-1$  etc. Noise is has similar values like  $n_1, n_2, \dots$



- (2) **Continuous Time, Continuous Amplitude (CTCA)** In that case both codeword and noise are function of time like  $x(t)$  or  $n(t)$ , respectively.

CTCA channels may be converted to DTCA channels. We focus on DTCA channels since CTCA (real-life) channels can be converted to DTCA channels. Conversion is based on orthonormal expansion idea.

## 58. THE GAUSSIAN CHANNEL (DTCA)

$$Y_i = X_i + Z_i$$

$Z_i$ 's are I.I.D. and  $Z_i \sim N(0, N)$ .  $X_i$  and  $Z_i$  are assumed to be independent. Notice that in this case  $Z_i$  is AWGN and it is the worst disturbing noise due to entropy limit actually.

In case there is no restriction on  $X_i$ 's, capacity is infinite.

- (1) If  $N = 0$ , any real number can be transmitted hence capacity is  $\infty$ .
- (2) If  $N > 0$  then  $X$  values can be chosen to be infinitely support from each other. Capacity is  $\infty$ .

## Part 22. Lecture 21 - 05.05.2016

The most common input restriction is the average power constraint. For any codeword  $(x_1, \dots, x_n)$  we should have

$$(58.1) \quad \frac{1}{n} \sum_{i=1}^n x_i^2 \leq P$$

and as  $n \rightarrow \infty$  from WLLN,

$$(58.2) \quad E[X_i^2] \leq P$$

*Note: For (58.1) and (58.2), I am not sure about case of  $X$  or  $x$ .*

## 59. THE INFORMATION CHANNEL CAPACITY OF THE GAUSSIAN CHANNEL

**Definition 38** (The Information Channel Capacity of The Gaussian Channel).

$$\begin{aligned} C &= \max_{p(x): E[X^2] \leq P} I(X; Y) \\ &= \frac{1}{2} \log \left( 1 + \frac{P}{N} \right) \end{aligned}$$

where  $N$  is variance of the noise.

**Proof:**

$$\begin{aligned}
(59.1) \quad I(X; Y) &= h(Y) - h(Y|X) \\
(59.2) \quad &= h(Y) - h(Z|X) \\
(59.3) \quad &= h(Y) - h(Z) \\
(59.4) \quad &\leq \frac{1}{2} \log(2\pi e \sigma_Y^2) - \frac{1}{2} \log(2\pi e N)
\end{aligned}$$

Transition from (59.1) to (59.2) is possible since  $Y = X + Z$  and transition from (59.2) to (59.3) is possible since  $X$  and  $Z$  are independent.

(59.4) is valid due to upper bound for differential entropy found in previous lectures. There is equality iff  $Y \sim N(0, \sigma_Y^2)$ . This is enabled by  $X \sim N(0, \sigma_X^2)$  because since  $Z$  is zero mean, to have  $\mu_Y = 0$ ,  $\mu_Z = 0$  and addition of two independent Gaussian random variables make a Gaussian random variable. Therefore to maximize the mutual information  $X$  should be a zero mean Gaussian random variable.

$\sigma_X^2 = P$  since  $X$  is a zero-mean and  $\sigma_Y^2 = P + N$ . Continuing from (59.4),

$$\begin{aligned}
C &= \frac{1}{2} \log(2\pi e(P + N)) - \frac{1}{2} \log(2\pi e N) \\
&= \frac{1}{2} \log \left( 1 + \frac{P}{N} \right)
\end{aligned}$$

**Definition 39** (( $M, n$ ) Code Satisfying The Power Constraint).

- (1) An index set  $\{1, 2, \dots, M\}$
- (2) An encoding function  $\{1, 2, \dots, M\} \rightarrow R^n$ .  
 $X^n(1), X^n(2), \dots, X^n(M)$  are codewords.

$$\sum_{i=1}^n (x_i(w))^2 \leq nP$$

$$\|X^n(w)\|^2 \leq nP$$

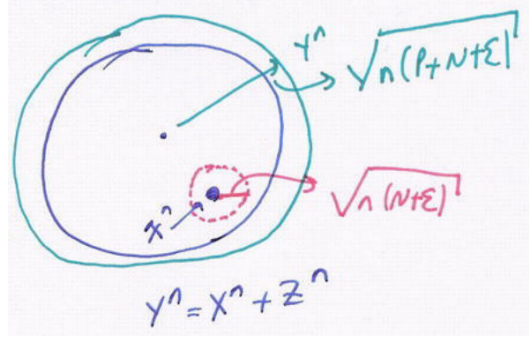
- (3) A decoding function  $g, g : R^n \rightarrow \{1, 2, \dots, M\}$ .

$$R \leq \frac{\log_2 M}{n}$$

$$P_e \leq P(X^n \neq g(Y^n))$$

## 60. SPHERE PACKING ARGUMENT

Notice that the expression  $\|x^n\|^2 \leq nP$  defines a sphere (n-dimensional, hypersphere). Consider Figure 60.



A codeword  $x^n$  is set. Then,

$$Y^n = x^n + Z^n$$

$$E[Y^n] = x^n$$

By WLLN,

$$\frac{1}{n} \sum_{i=1}^n Z_i \rightarrow N \quad \text{as } n \rightarrow \infty$$

$$(60.1) \quad \|Z^n\|^2 = \sum_{i=1}^n Z_i^2 \leq n(N + \epsilon) \quad \text{as } n \rightarrow \infty$$

Notice that (60.1) defines a sphere too.

$$\begin{aligned} \|Y^n\|^2 &= \|X^n + Z^n\|^2 \\ &= \langle X^n, X^n \rangle + \langle X^n, Z^n \rangle + \langle Z^n, X^n \rangle + \langle Z^n, Z^n \rangle \\ &= \|X^n\|^2 + \|Z^n\|^2 + 2 \langle X^n, Z^n \rangle \\ &= \|X^n\|^2 + \|Z^n\|^2 + 2 \sum_{i=1}^n X_i Z_i \\ &= \|X^n\|^2 + \|Z^n\|^2 + 2E[X_i Z_i] \\ &= (\leq nP) + \left( \leq n \left( N + \frac{\epsilon}{2} \right) \right) + 2 \left( \leq n \left( 0 + \frac{\epsilon}{4} \right) \right) \\ &\leq nP + n \left( N + \frac{\epsilon}{2} \right) + 2n \left( 0 + \frac{\epsilon}{4} \right) \\ (60.2) \quad &\leq n(P + N + \epsilon) \end{aligned}$$

Notice that (60.2) defines a sphere too.

How many non-intersecting spheres can we fit?

$$\begin{aligned}
 V_n &= A_n r^n \\
 \frac{A_n (n(P + N + \epsilon))^{n/2}}{(n(N + \epsilon))^{n/2}} &= \left( \frac{P + N + \epsilon}{N + \epsilon} \right)^{n/2} \\
 &= 2^{\frac{n}{2} \log \left( \frac{P + N + \epsilon}{N + \epsilon} \right)} \\
 &= 2^{\frac{n}{2} \log \left( 1 + \frac{P}{N} \right)} \quad \text{as } n \rightarrow \infty
 \end{aligned}
 \tag{60.3}$$

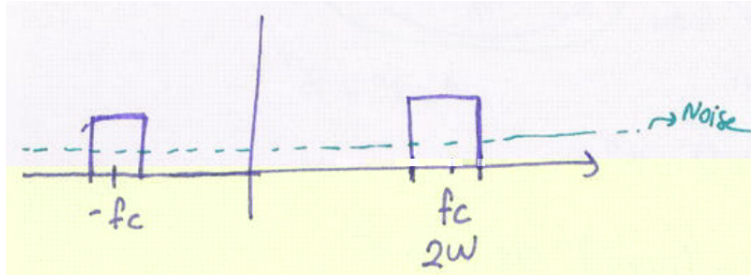
Notice that in (60.3),

$$C = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right)$$

## Part 23. Lecture 22 - 10.05.2016

### 61. BAND-LIMITED CHANNELS

A general band-limited channel is shown in Figure 61. This is a common model when a finite bandwidth is available for communication.



$$Y(t) = (X(t) + Z(t)) * h(t) \tag{61.1}$$

In (61.1),  $X(t)$ ,  $Z(t)$  and  $h(t)$  denote signal waveform, noise waveform and impulse response of an ideal band-pass filter. Later on, we will work in baseband and it will be denoting its baseband version which is an ideal low-pass filter. Ideal low-pass filter response is expressed in (61.2).

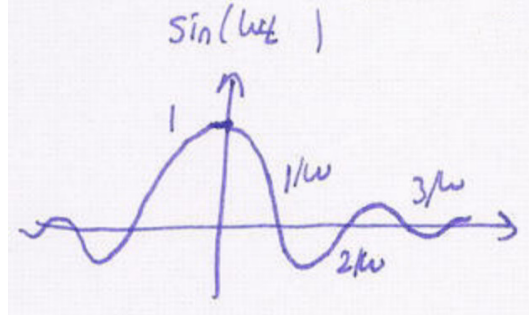
$$H(f) = \begin{cases} 1, & \text{if } |f| < W \\ 0, & \text{otherwise} \end{cases} \tag{61.2}$$

## 62. SAMPLING (SHANNON-NYQUIST) THEOREM

Consider a function band-limited to  $W$ , i.e  $G(f) = 0$  for  $|f| > W$ . Then,  $g(t)$  is completely specified by taking  $1/2W$  seconds apart.

$$\begin{aligned}
 g(t) &= \int_{-W}^W G(f) e^{j2\pi ft} dt \\
 g(t) &= \int_{-W}^W G(f) e^{j2\pi fn} dt \\
 g\left(n \frac{1}{2W}\right) &= \int_{-W}^W G(f) e^{j2\pi fn/2W} dt \\
 (62.1) \quad g(t) &= \sum_{n=-\infty}^{\infty} g\left(n \frac{1}{2W}\right) \text{sinc}(Wt - n)
 \end{aligned}$$

Let's analyze (62.1).  $g\left(n \frac{1}{2W}\right)$  implies that each  $g()$  can be chosen independently. It means that in overall, a band-limited signal has  $2W$  degrees of freedom per second.  $\text{sinc}(Wt - n)$  are orthonormal basis functions for the set of band-limited signals. An example sinc function is drawn in Figure 62.



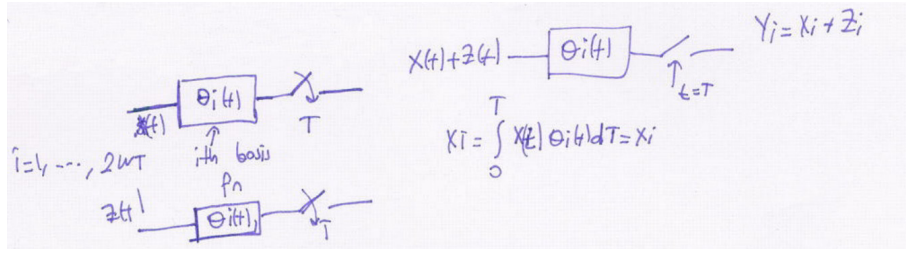
In practical systems, pulses are time limited to  $[0, T]$ . We want to have band-limited signals which is not possible for time-limited signals. Hence, we consider almost time-limited almost band-limited signals: Most energy is confined in  $[0, T]$  and  $[-W, W]$ . In this case, basis functions are called as prolate spheroidal wave functions (PSWF).

There are about  $2WT$  orthonormal basis functions for the set of almost time-limited almost band-limited signals when  $WT$  is large.

Consider Figure 62. Let

$$X_i = \int_0^T X(t) \Theta_i(t) dt.$$

We can define  $Y_i$  and  $Z_i$  similarly.



In this setting,  $Z_i$ 's are Gaussian since output of an LTI system is Gaussian process when the input is. It is also white due to orthogonality because  $Z_i$ 's are become uncorrelated.

$$(62.2) \quad E[X_i^2] = \frac{PT}{2WT}$$

$$(62.3) \quad \begin{aligned} &= \frac{P}{2W} \\ E[Z_i^2] &= \frac{N_0/2WT}{2WT} \\ &= \frac{N_0}{2} \end{aligned}$$

Capacity for  $Y_i = X_i + Z_i$  is calculated as

$$\begin{aligned} C_i &= \frac{1}{2} \log \left( 1 + \frac{P/2W}{N_0/2} \right) \\ &= \frac{1}{2} \log \left( 1 + \frac{P}{N_0 W} \right) \\ C &= \frac{1}{T} 2WT \frac{1}{2} \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bps} \\ &= W \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bps} \end{aligned}$$

(62.2) and (62.3) are valid since  $\Theta_i$ s are normalized. We can think (62.2) as the following. Power is  $P$  and we are transmitting over interval  $T$  therefore total energy is  $PT$  and this energy is shared by  $2WT$  orthogonal and identical channels.

**Highlight 38** (Famous Capacity Relation).

$$(62.4) \quad C = W \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bps}$$

## Part 24. Lecture 23 - 12.05.2016

## 63. NORMALIZED CAPACITY

$$(63.1) \quad \begin{aligned} \frac{C}{W} &= \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bps/Hz} \\ P &= C E_b \end{aligned}$$

$$(63.2) \quad \begin{aligned} \frac{C}{W} &= \log \left( \frac{E_b}{N_0} + \frac{C}{W} \right) \text{ bps/Hz} \\ \eta &= \frac{C}{W} \end{aligned}$$

$$(63.3) \quad \begin{aligned} \eta &= \log \left( 1 + \frac{E_b}{N_0} \eta \right) \\ 2^\eta &= 1 + \frac{E_b}{N_0} \eta \\ \frac{E_b}{N_0} &= \frac{2^\eta - 1}{\eta} \end{aligned}$$

In (63.1),  $C$  is in bps and  $E_b$  denotes energy per message bit. **Spectral efficiency** in bps/Hz is defined in (63.2). (63.3) can be interpreted as the following.

$$(63.4) \quad \begin{aligned} \lim_{\eta \rightarrow 0} \frac{2^\eta - 1}{\eta} &= \ln 2 \\ &= -1.6 \text{ dB} \end{aligned}$$

Notice that (63.4) can be obtained by using L'Hospital rule.

**Example 29.**

Let  $\eta = 1$  bps/Hz then,

$$\left( \frac{E_b}{N_0} \right)_{\min} = \frac{2 - 1}{1} = 0 \text{ dB.}$$

**Example 30.**

Let  $\eta = 2$  bps/Hz then,

$$\left( \frac{E_b}{N_0} \right)_{\min} = \frac{2^2 - 1}{1} = 1.8 \text{ dB.}$$

Notice that from (62.4)  $C$  depends on  $W$  both proportionally and inversely proportionally. So what happens at the end of the day?

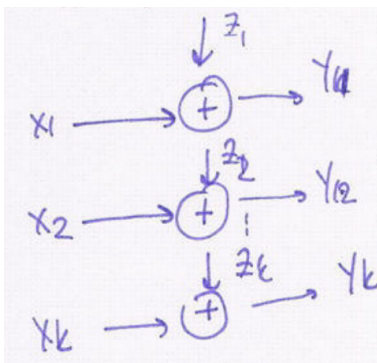
$$(63.5) \quad \lim_{W \rightarrow \infty} C = \lim_{W \rightarrow \infty} \frac{\log \left( 1 + \frac{P}{N_0 W} \right)}{1/W}$$

$$(63.6) \quad = \frac{P}{N_0 \ln 2}$$

Again, L'Hospital rule is used from transition to (63.5) to (63.6).

#### 64. PARALLEL GAUSSIAN CHANNELS

In this problem, there are  $K$  independent AWGN channels as shown in Figure 64.



$$E[Z_i] = 0$$

$$E[Z_i^2] = N_i$$

$$\sum_{i=1}^k P_i \leq P$$

Let  $X^k$  be the vector of  $X_1, X_2, \dots, X_k$ . Then,

$$C = \max_{p(X^k): \sum_{i=1}^k E[X_i^2] \leq P} I(X^k; Y^k)$$



$$\begin{aligned}
I(X^k; Y^k) &= h(Y^k) - h(Y^k | X^k) \\
&= h(Y^k) - h(Z^k) \\
(64.1) \quad &\leq \sum_{i=1}^k h(Y_i) - h(Z^k)
\end{aligned}$$

$$\begin{aligned}
(64.2) \quad &\leq \sum_{i=1}^k h(Y_i) - \sum_{i=1}^k h(Z_i) \\
&\leq \sum_{i=1}^k h(Y_i) - \sum_{i=1}^k \frac{1}{2} \log 2\pi e N_i
\end{aligned}$$

$$(64.3) \quad h(Y_i) \leq \frac{1}{2} \log 2\pi e (P_i + N_i)$$

$$I(X^k; Y^k) \leq \sum_{i=1}^k \frac{1}{2} \log 2\pi e (P_i + N_i) - \sum_{i=1}^k \frac{1}{2} \log 2\pi e N_i$$

Equality holds in (64.1) if  $Y_i$ 's are independent. Transition from (64.1) to (64.2) is due to independence of  $Z_i$ s. Equality holds in (64.3) if  $Y_i$ , i.e.,  $X_i$  is Gaussian. Maximum of mutual information can be found as in (64.4).

$$\begin{aligned}
(64.4) \quad I(X^k; Y^k) &= \sum_{i=1}^k \frac{1}{2} \log \left( 1 + \frac{P_i}{N_i} \right) \\
C &= \max_{P_1, \dots, P_k: \sum_{i=1}^k P_i = P} \sum_{i=1}^k \frac{1}{2} \log \left( 1 + \frac{P_i}{N_i} \right)
\end{aligned}$$

Lagrange multipliers can be used to solve this maximization problem.

$$\mathcal{L}(P_1, \dots, P_k) = \sum_{i=1}^k \frac{1}{2} \log \left( 1 + \frac{P_i}{N_i} \right) + \lambda \left( \sum_{i=1}^k P_i - P \right)$$

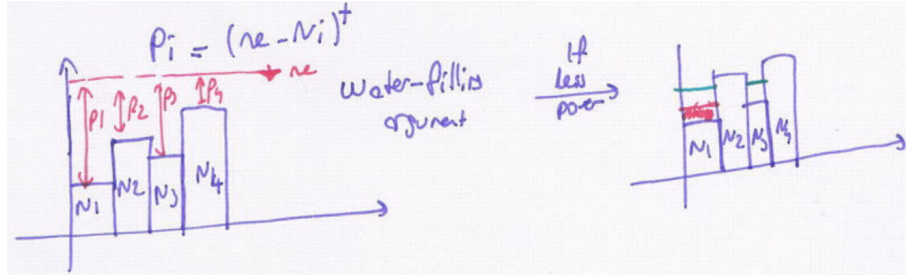
By taking derivative with respect to  $P_i$ ,

$$\begin{aligned}
(64.5) \quad &\frac{\ln 2}{2} \frac{1}{1 + \frac{P_i}{N_i}} \frac{1}{N_i} + \lambda = 0 \\
&c \frac{1}{N_i + P_i} + \lambda = 0
\end{aligned}$$

(64.5) means that  $N_i + P_i = v$  (I don't know why!). Notice that summation of  $P_i$  and  $N_i$  can't be negative since they are powers. So, suggest the following  $P_i$  selection

$$P_i = (v - N_i)^+ = \begin{cases} v - N_i, & v > N_i \\ 0, & \text{otherwise.} \end{cases}$$

Check the KarushKuhnTucker (KKT) conditions, they are satisfied. So we have our solution. Solution is visualized in Figure 64. It is called as "water filling solution."



## Part 25. Lecture 24 - 17.05.2016

### 65. CHANNELS WITH COLORED GAUSSIAN NOISE

$$Y^k = X^k + Z^k$$

In that case,  $K_Z$  is not diagonal. Then,

$$h(Z^k) \neq \sum_{i=1}^k h(Z_i).$$

There is a total power constraint such that

$$\sum_{i=1}^k E[X_i^2] \leq P$$

$$\text{tr}(K_X) \leq P$$

#### 65.1. Part a.

$$\begin{aligned} I(X^k; Y^k) &= h(Y^k) - h(Y^k | X^k) \\ &= h(Y^k) - h(Z^k) \\ &= h(Y^k) - \frac{1}{2} \log((2\pi e)^k |K_Z|) \\ K_Y &= E[Y^k (Y^k)^T] \\ &= E[X^k (X^k)^T] + E[X^k (Z^k)^T] + E[Z^k (X^k)^T] + E[Z^k (Z^k)^T] \\ &= K_X + \mathbf{0} + \mathbf{0} + K_Z \\ &= K_X + K_Z \end{aligned}$$

$$(65.1) \quad h(Y^k) \leq \frac{1}{2} \log((2\pi e)^k |K_X + K_Z|)$$

$$(65.2) \quad I(X^k; Y^k) \leq \frac{1}{2} \log((2\pi e)^k |K_X + K_Z|) - \frac{1}{2} \log((2\pi e)^k |K_Z|)$$

(65.2) holds with equality iff  $X^k \sim N(\mathbf{0}, K_X)$ . We should find  $K_X$ .

65.2. **Part b.** Whitening of  $Z^k$  is meaning that finding a orthonormal basis  $Q$ .

$$(65.3) \quad K_Z = Q\Lambda Q^T$$

In (65.3),  $Q$  is a unitary matrix, i.e.,  $QQ^T = I$  or  $Q^TQ = I$  and  $\Lambda$  is a diagonal matrix. Let's observe the effect of whitening as an exercise

$$\begin{aligned} \tilde{Z}_k &= Q^T Z_k \\ K_{\tilde{Z}_k} &= E[\tilde{Z}_k(\tilde{Z}_k)^T] \\ &= E[Q^T Z^k (Z^k)^T Q] \\ &= Q^T E[Z^k (Z^k)^T] Q \\ &= Q^T K_Z Q \\ &= Q^T Q \Lambda Q^T Q \\ &= I \Lambda I \\ (65.4) \quad &= \Lambda \end{aligned}$$

Consider (65.1)

$$\begin{aligned} |K_X + K_Z| &= |K_X + Q\Lambda Q^T| \\ &= |QQ^T K_X QQ^T + Q\Lambda Q^T| \\ &= |Q(Q^T K_X Q) + \Lambda)Q^T| \end{aligned}$$

Remember that  $|AB| = |A||B|$  then,

$$\begin{aligned} |K_X + K_Z| &= |Q||Q^T K_X Q + \Lambda||Q^T| \\ &= 1|Q^T K_X Q + \Lambda|1 \\ (65.5) \quad &= |Q^T K_X Q + \Lambda| \end{aligned}$$

65.3. **Part c.** Remember that  $\text{tr}(AB) = \text{tr}(BA)$

$$\begin{aligned} \text{tr}(Q^T K_X Q) &= \text{tr}(K_X QQ^T) \\ (65.6) \quad &= \text{tr}(K_X) \end{aligned}$$

$\text{tr}(K_X) \leq P$  means that  $\text{tr}(Q^T K_X Q) \leq P$

65.4. **Part d.** Let  $A = Q^T K_X Q$ . Our problem is now maximization of  $|A + \Lambda|$  with  $\text{tr}(A) \leq P$ .

From Hadamard's inequality,

$$(65.7) \quad |A + \Lambda| \leq \prod_{i=1}^k (A_{ii} + \Lambda_{ii})$$

(65.7) holds with equality iff  $A + \Lambda$  is diagonal. Since  $\Lambda$  is diagonal, it means that  $A$  should be diagonal.

Remember the equality in (65.5) and (65.6)

$$|K_X + K_Z| \leq \prod_{i=1}^k (A_{ii} + \Lambda_{ii})$$

Also note that

$$|K_Z| = |\Lambda|.$$

$$\begin{aligned} C &= \max_A \frac{1}{2} \log \left( (2\pi e)^n \prod_{i=1}^k (A_{ii} + \Lambda_{ii}) \right) - \frac{1}{2} \log \left( (2\pi e)^n \prod_{i=1}^k \Lambda_{ii} \right) \\ &= \max_{A: \text{tr}(A)} \frac{1}{2} \log \prod_{i=1}^k \frac{A_{ii} + \Lambda_{ii}}{\Lambda_{ii}} \\ (65.8) \quad &= \max_{A: \text{tr}(A)} \sum_{i=1}^k \frac{1}{2} \log \left( 1 + \frac{A_{ii}}{\Lambda_{ii}} \right) \end{aligned}$$

Notice that in (65.8)  $A$  is a diagonal matrix and it turns out to be same as the previous water filling problem. Then,

$$A_{ii} = (v - \Lambda_{ii})^+$$

Finally,

$$K_X = Q A Q^T$$

where  $A$  is diagonal and  $K_X$  is not.

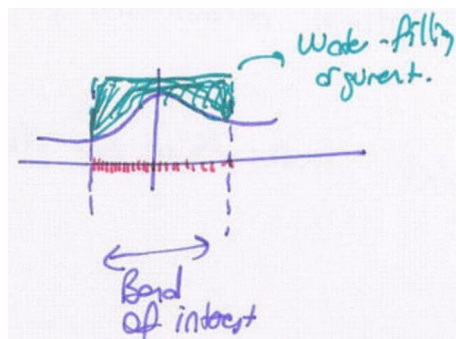
We use whitening and solution is same as the previous uncorrelated problem.

**65.5. If noise is a WSS Gaussian Process.** Now we talk about continuous time channels directly.

$$Y(t) = X(t) + Z(t)$$

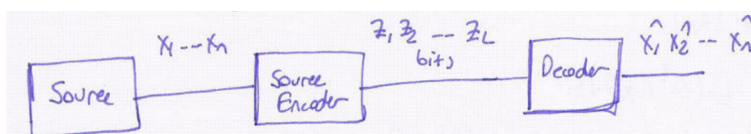
where  $Z(t)$  is WSS. It will have autocorrelation function  $K_Z(\tau)$  and power spectral density  $S_X(f)$  which is related to autocorrelation function obviously. It turns out to be that water filling idea is valid in this scenario too. We work in a frequency domain. We can think that it consists of very small sub-bands as shown in Figure 65.5. They are like independent Gaussian channels.

So, water filling argument almost is valid as long as there are parallel channels.



## 66. RATE DISTORTION THEORY (LOSSY SOURCE CODING)

The idea is representation of continuous random variable with discrete numbers (bits).



Consider a Gaussian random variable. It takes values from  $-\infty$  to  $\infty$ . Therefore, it is not possible to represent a value drawn from a normal distribution and reconstruct it without any error. So, coding is lossy. A continuous source can never be represented with finite number of bits.

## 67. PERFORMANCE MEASURES

One measure is number of used bits and another one is representation error.

Rate  $R$  is defined as the ratio of number of bits / number of source symbols. The other measure is distortion  $D$  which is some cost function which shows the difference between the original and reconstructed source symbol.

## 68. QUANTIZATION

We want to minimize  $MSE = E[(X - \hat{X})^2]$ . In quantization, we choose points and map values to these points. These reconstruction points are used later.

Let's consider case where  $X \sim N(0, \sigma^2)$ .

68.1.  $R = 0$ . In that case, we use 0 bits,  $2^0 = 1$  reconstruction points. A logical reconstruction point is 0 for this specific example. We can find the same answer (taking 0) by using MMSE estimator

$$\begin{aligned}\hat{X}_{MMSE} &= E[X] \\ &= 0\end{aligned}$$

68.2.  $R = 1$ . We have  $2^1 = 2$  reconstruction points. How should we choose them? They should be symmetric, right? Let's say that they are  $-a$  and  $a$  then use MMSE rule.

$$(68.1) \quad a = E[X|X > 0]$$

68.3.  $R = 2$ . We have 4 points:  $-c, -b, b, c$ . What are these numbers now? It is not obvious as the previous examples.

Let's say that we know the points. For an observation, which point should be used for assignment to minimize MSE? We can easily find regions for each reconstruction points (divide at the middles?). Similarly, a given region we can find optimal reconstruction point as in (68.1). However when a reconstruction point is changed, boundary of region is also changed. It is an iterative approach.

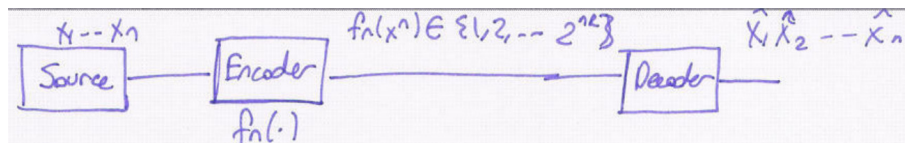
**Highlight 39** (Lloyd's algorithm).

This algorithm is used to find regions and reconstruction points iteratively. Here is the approach:

- (1) Start with some reconstruction points.
- (2) Determine the region for each point.
- (3) Determine the optimal reconstruction point for each region.
- (4) Repeat the previous steps until convergence.

## 69. SOME DISTORTION DEFINITIONS

Let's we have a source generating  $X_1, X_2, \dots$ . These are continuous I.I.D. random variables. Interesting thing that, rate distortion theory will be hold for discrete random variables too.



### 69.1. Distortion Function.

**Definition 40** (Distortion Function).

A distortion function (measure) is a mapping

$$d : A_X \times A_{\hat{X}} \rightarrow R^+$$

where  $A_X$  is source alphabet,  $A_{\hat{X}}$  is representation alphabet and  $R^+$  is non-negative real numbers.

$d(x, \hat{x})$  is a measure of the cost of representing  $x$  with  $\hat{x}$ .

69.1.1. *Hamming (Probability of Error) Distortion.*

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & \text{otherwise} \end{cases}$$

This seems to be logical for discrete valued random variables.

69.1.2. *Squared Error Distortion.*

$$d(x, \hat{x}) = (x - \hat{x})^2$$

69.1.3. *Distortion for Sequences.*

**Definition 41** (Distortion for Sequences).

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i)$$

or

$$d(x^n, \hat{x}^n) = \max d(x_i, \hat{x}_i)$$

Part 26. No Lecture on 19.05.2016

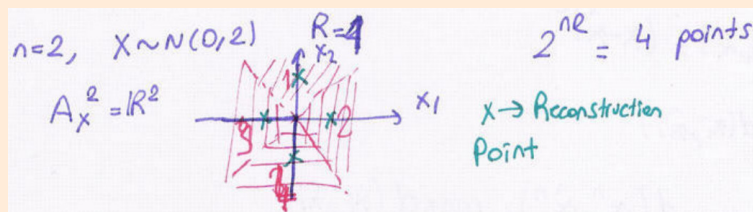
Part 27. Lecture 25 - 24.05.2016

69.2. *Rate-Distortion Code.*

**Definition 42** (Rate-Distortion Code).

A  $(2^{nR}, n)$  rate distortion code has an encoding function  $f: A_X^n \rightarrow \{1, 2, \dots, 2^{nR}\}$  and a decoding function  $g: \{1, 2, \dots, 2^{nR}\} \rightarrow A_{\hat{X}}^n$ .

$$\begin{aligned} D &= E[d(X^n, \hat{X}^n)] \\ &= \begin{cases} \sum_{x^n} p(x^n) d(x^n, \hat{x}^n) & \text{if } X \text{ is discrete.} \\ \int_{A_X^n} f(x^n) d(x^n, \hat{x}^n) dx^n & \text{if } X \text{ is continuous.} \end{cases} \end{aligned}$$

**Example 31.****Highlight 40.**

Notice that  $f_n^{-1}(i)$  associate to assignment region. Similarly,  $g_n(i)$  gives as actual reconstruction sequence with length  $n$ . It can be also shown as  $\hat{X}^n(i)$ . It is referred as codewords, vector quantization, estimate of  $X^n$  or reconstruction of  $X^n$ .

**69.3. Achievable Rate.****Definition 43 (Achievable Rate).**

A rate distortion pair  $(R, D)$  is said to be achievable if there exists a sequence of  $(2^{nR}, n)$  rate distortion codes  $(f_n, g_n)$  with

$$\lim_{n \rightarrow \infty} E[d(X^n, g_n(f_n(X^n)))] \leq D$$

Notice that in lossless case  $D$  was zero.

**69.4. Distortion Region.**



**Definition 44** (Distortion Region).

The rate distortion region for a source is the closure of the set of achievable rate distortion pairs  $(R, D)$ . A general plot is shown in Figure 44.

**69.5. Operational Rate Distortion Function.****Definition 45** (Operational Rate Distortion Function).

$$R(D) = \inf_{C: E_C(d(X^n, \hat{X}^n)) \leq D} R(C)$$

**69.6. Operational Distortion Rate Function.****Definition 46** (Operational Distortion Rate Function).

$$D(R) = \inf_{C: \text{rate}(C) \geq R} D(C)$$

**Highlight 41.**

Operational Rate Distortion Function and Operational Distortion Rate Function determine the boundaries of the rate distortion region.

**69.7. The Information Rate Distortion Function.****Definition 47** (The Information Rate Distortion Function).

$$(69.1) \quad R^I(D) = \min_{p(\hat{X}/X): \sum_{x, \hat{x}} p(x)p(\hat{x}/x)d(x, \hat{x}) \leq D} I(X; \hat{X})$$

In (69.1), it is assumed that  $p(x)$  is given.

**Theorem 27.**

Operational rate distortion function is equal to information rate distortion function.

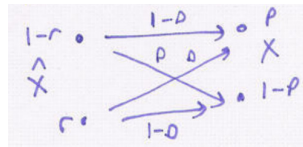
$$R(D) = R^I(D)$$

**69.8. Binary Source with Hamming Distortion.** Let's consider a binary source with  $p$  and use Hamming distortion criteria.

$$\begin{aligned}
 E[d(X, \hat{X})] &\leq D \\
 E[d(X, \hat{X})] &= 0 \times P(X = \hat{X}) + 1 \times P(X \neq \hat{X}) \\
 0 \times P(X = \hat{X}) + 1 \times P(X \neq \hat{X}) &\leq D \\
 (69.2) \quad P(X \oplus \hat{X} = 1) &\leq D \\
 I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\
 (69.3) \quad &= H(p) - H(X \oplus \hat{X}|\hat{X}) \\
 (69.4) \quad &\geq H(p) - H(X \oplus \hat{X}) \\
 (69.5) \quad H(X \oplus \hat{X}) &\leq H(D) \\
 I(X; \hat{X}) &\geq H(p) - H(D)
 \end{aligned}$$

Transition from (69.3) to (69.4) is done by using the fact that conditioning reduces the entropy. For (69.5) let's assume that  $D < 0.5$ . If it was greater than 0.5, we can always switch it. (69.5) is found by using (69.2).

Can we find  $p(\hat{X}|X)$  that achieves  $I(X; \hat{X}) = H(p) - H(D)$ ? This problem can be modeled as a BSC as shown in Figure 69.8.



$$\begin{aligned}
 I(X; \hat{X}) &= H(p) - H(D) \\
 P(X = 0) &= p \\
 &= rD + (1-r)(1-D) \\
 r &= \frac{p - (1-D)}{2D - 1}
 \end{aligned}$$

We know that  $0 \leq r \leq 1$  and  $0 \leq 1 - r \leq 1$  and assumed that  $D \leq 0.5$

$$(69.6) \quad 0 \leq \frac{p + D - 1}{2D - 1} \leq 1$$

$$(69.7) \quad 0 \leq \frac{p - D}{1 - 2D} \leq 1$$

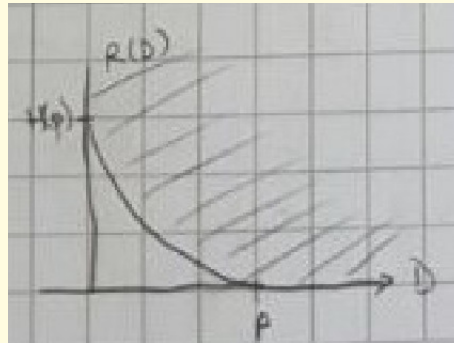
From (69.6) and (69.7), we can find that  $D \leq p$  and  $D \leq 1 - p$ . Therefore if  $D \leq \min(p, 1 - p)$ ,  $I(X; \hat{X}) = H(p) - H(D)$  with

$$p(x|\hat{x}) = \begin{cases} D, & x \neq \hat{x} \\ 1 - D, & \text{otherwise} \end{cases}$$

**Highlight 42** (Binary Source with Hamming Distortion).

Let's consider a binary source with  $p$  and use Hamming distortion criteria.

$$(69.8) \quad R(D) = \begin{cases} H(p) - H(D), & D < \min(p, 1 - p) \\ 0, & \text{otherwise} \end{cases}$$



**Example 32.**

Let  $P(X = 0) = 1/3$  and we want  $D \leq 3/4$ . Select  $R = 0$  meaning that there is a single reconstruction point. Since  $P(X = 1) > P(X = 0)$ , selecting  $\hat{X} = 1$  makes sense. Therefore

$$\begin{aligned} D &= P(X \neq \hat{X}) \\ &= \frac{1}{3} \\ &\leq \frac{3}{4} \end{aligned}$$

Rate 0 satisfies this distortion. Also from (69.8), since the desired  $D$  is not smaller than  $1/3$ ,  $R(3/4) = 0$ .

**Part 28. Lecture 26 - 26.05.2016**

**69.9. Gaussian Source with MSE.** Let  $X \sim N(0, \sigma^2)$ . Use MSE distortion.

$$\begin{aligned} E[(X - \hat{X})^2] &\leq D \\ R(D) &= \min_{f(\hat{X}|X): E[(X - \hat{X})^2] \leq D} I(X; \hat{X}) \\ I(X; \hat{X}) &= h(X) - h(X|\hat{X}) \\ &= \frac{1}{2} \log(2\pi e \sigma^2) - h(X|\hat{X}) \\ &= \frac{1}{2} \log(2\pi e \sigma^2) - h(X - \hat{X}|\hat{X}) \\ h(X - \hat{X}|\hat{X}) &\leq h(X - \hat{X}) \\ I(X; \hat{X}) &\geq \frac{1}{2} \log(2\pi e \sigma^2) - h(X - \hat{X}) \\ h(X - \hat{X}) &\leq h(N(0, E[(X - \hat{X})^2])) \\ h(N(0, E[(X - \hat{X})^2])) &= \frac{1}{2} \log(2\pi e E[(X - \hat{X})^2]) \\ h(N(0, E[(X - \hat{X})^2])) &\leq \frac{1}{2} \log(2\pi e D) \\ I(X; \hat{X}) &\geq \frac{1}{2} \log(2\pi e \sigma^2) - \frac{1}{2} \log(2\pi e D) \\ (69.9) \quad R(D) &\geq \frac{1}{2} \log \left( \frac{\sigma^2}{D} \right) \end{aligned}$$

Can we reach the RHS of (69.9)? We can model this problem as a Gaussian channel shown in Figure 69.9.

$$\begin{array}{c}
 \hat{X} \xrightarrow{\oplus} X \sim N(0, \sigma^2) \\
 \downarrow z \sim N(0, D) \\
 \sim N(0, \sigma^2 - D)
 \end{array}$$

$$\begin{aligned}
 I(X; \hat{X}) &= h(X) - h(X|\hat{X}) \\
 &= h(X) - h(Z) \\
 &= \frac{1}{2} \log 2\pi e \sigma^2 - \frac{1}{2} \log 2\pi e D \\
 &= \frac{1}{2} \log \frac{\sigma^2}{D}
 \end{aligned}$$

If  $\sigma^2 - D \geq 0$ , then

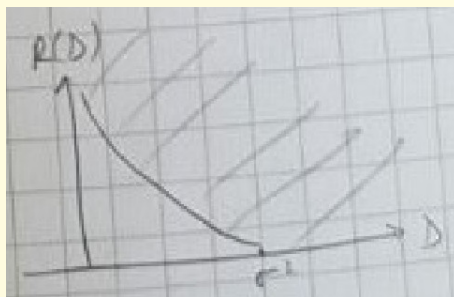
$$R(D) = \frac{1}{2} \log \left( \frac{\sigma^2}{D} \right)$$

If not,  $R(D) = 0$ .

**Highlight 43** (Gaussian Source with MSE).

Let  $X \sim N(0, \sigma^2)$ . Use MSE distortion.

$$R(D) = \begin{cases} \frac{1}{2} \log \left( \frac{\sigma^2}{D} \right), & D < \sigma^2 \\ 0, & \text{otherwise} \end{cases}$$



$$R = \frac{1}{2} \log \frac{\sigma^2}{D(R)}$$

$$\frac{\sigma^2}{D(R)} = 2^{2R}$$

$$D(R) = 2^{-2R} \sigma^2$$

Increasing representation by 1 bit decreases distortion by 1/4 (6 dB).

**Part 29. To Do's**

- Some figures do not have captions and visible figure numbers since no caption is defined for them. Solve this.
- Check  $K_Z$  and  $\tilde{Z}_k$  and terms in (65.4). Subscript or superscript?
- Case of  $X$ s in (69.1).

**Part 30. Materials**

This part is added for author, not reader. Stop reading!

**70. LECTURE 01**

- 205-120-1455656445-77
- 205-120-1455656447-98
- 209-120-1456254230-12
- 209-120-1456254306-57

## 71. LECTURE 02

- 205-120-1455656449-73
- 205-120-1455656451-23
- 205-120-1455656453-98
- 209-120-1456422159-69

## 72. LECTURE 03

- 205-120-1455656455-56
- 205-120-1455656457-26
- 205-120-1455656459-74
- 205-120-1455656461-38
- 205-120-1455656463-82
- 209-120-1456861030-36
- 209-120-1456861391-18

## 73. LECTURE 04

- 205-120-1455656465-41
- 205-120-1455656467-84
- 205-120-1455656469-95
- 209-120-1457159259-18

## 74. LECTURE 05

- 205-120-1455656471-96
- 205-120-1455656473-56
- 205-120-1455656475-56
- 209-120-1458065534-18
- 209-120-1458065588-94

## 75. LECTURE 06

- 205-120-1455656477-94
- 205-120-1455656479-23
- 209-120-1458151335-59

## 76. LECTURE 07

- 205-120-1455656481-45
- 205-120-1455656483-11
- 205-120-1455656485-90
- 205-120-1455656487-47
- 209-120-1458151512-62
- 209-120-1458151561-49

## 77. LECTURE 08

- 205-120-1455656489-54
- 205-120-1455656491-93
- 209-120-1458327865-37

## 78. LECTURE 09

- 205-120-1455656497-84
- 205-120-1455656499-35
- 205-120-1455656501-70
- 205-120-1455656503-46
- 209-120-1459188716-82

## 79. LECTURE 10

- 205-120-1455656505-93
- 205-120-1455656507-47
- 209-120-1459589607-65

## 80. LECTURE 11

- 205-120-1455656509-44
- 205-120-1455656511-67
- 205-120-1455656513-70
- 209-120-1459598832-40

## 81. LECTURE 12

No related material.

## 82. LECTURE 13

- 205-120-1455656515-43
- 205-120-1455656517-33
- 209-120-1460147110-10
- 209-120-1460147184-52

## 83. LECTURE 14

- 205-120-1455656519-95
- 205-120-1455656521-18
- 209-120-1460147220-40

## 84. LECTURE 15

- 205-120-1455656523-70
- 205-120-1455656525-78
- 205-120-1455656527-58
- 205-120-1455656529-55
- 209-120-1460475526-05
- 209-120-1460475560-61

## 85. LECTURE 16

- 205-120-1455656531-73
- 205-120-1455656533-54
- 209-120-1460660764-71



## 86. LECTURE 17

- 205-120-1455656535-11
- 205-120-1455656537-30
- 205-120-1455656539-49
- 205-120-1455656541-33
- 209-120-1461870723-02
- 209-120-1461870799-44

## 87. LECTURE 18

- 205-120-1455656543-75
- 205-120-1455656545-60
- 205-120-1455656547-24
- 205-120-1455656549-23
- 209-120-1461873830-92
- 209-120-1461873843-13

## 88. LECTURE 19

Lecture notes of Emre.

## 89. LECTURE 20

- 205-120-1455656551-14
- 205-120-1455656553-17
- 209-120-1462633861-33
- 209-120-1462633892-71

## 90. LECTURE 21

- 205-120-1455656556-49
- 205-120-1455656558-20
- 209-120-1462633966-28

## 91. LECTURE 22

- 205-120-1455656560-01
- 205-120-1455656562-83
- 209-120-1463299563-95
- 209-120-1463299501-67

## 92. LECTURE 23

- 205-120-1455656564-40
- 205-120-1458504560-15
- 209-120-1463300839-51

## 93. LECTURE 24

- 205-120-1458504562-70
- 205-120-1458504564-57
- 205-120-1458504566-08
- 205-120-1458504568-34
- 209-120-1463655368-46
- 209-120-1463655419-93

## 94. LECTURE 25

- 205-120-1458504570-01
- 205-120-1458504572-12
- 205-120-1458504574-64
- 209-120-1464899057-02
- 209-120-1464899092-37

## 95. LECTURE 26

Lecture notes of Emre.

## Part 31. Code Appendix

## . MATLAB Code Figure 2.1

```

1 p = linspace(0,1,10000);
2 h = -1 * p.* log2(p) - 1 * (1-p) .* log2(1-p);
3 close all;
4 plot(p,h,'linewidth',2)
5 grid on
6 xlabel('$p$', 'Interpreter','Latex','FontSize',12)
7 ylabel('$H(p)$', 'Interpreter','Latex','FontSize',12)
8 title('$H(p)$ vs $p$', 'Interpreter','Latex','FontSize',12)

```

*E-mail address:* alperyazar@gmail.com