Information Theory Self Study Notes

Credit: Thomas Cover, Joy Thomas, Raymond Yeung
Reading Course with Professor Steven Evans UC Berkeley

Fangyuan Lin

June 10, 2025

Contents

1 Chapter 1: The Science of Information **1**

2 Chapter 2: Information Measures **2**
2.1 Independence and Markov Chains . . . 2
2.2 2.2 Shannon's Information Measures . . . 3
2.3 Continuity of Shannon's information measures for fixed finite alphabets . 5

3 Chapter 2 Continued **6**
3.1 2.4 Chain Rule . . . 6
3.2 2.5 Informational Divergence . . . 6
3.3 2.6 Basic Information . . . 8
3.4 2.7 More Useful Inequalities . . . 8
3.5 Fano's Inequality . . . 10
3.6 Entropy Rate of a Stationary Source . . . 11
3.6.1 Discrete-time Information Source . . . 11
3.6.2 Entropy Rate . . . 12

4 The I-Measure **15**
4.1 The Second Law of Thermodynamics . . . 15
4.2 The Asymptotic Equipartiton Property AEP . . . 16
4.2.1 One step away from construction of the I-measure μ^* . . . 17
4.2.2 μ^* can be negative . . . 18

5 I-measure for Markov chains **20**
5.1 Information Diagram Applications . . . 21
5.2 Shannon's Perfect Secrecy Theorem . . . 22

6 Zero-Error Data Compression **24**
6.1 The Entropy Bound . . . 24
6.1.1 Expected Length of code . . . 26
6.2 Prefix Codes . . . 27
6.2.1 D-adic Distributions . . . 29

7 Weak Typicality **36**
7.1 The Weak AEP . . . 36
7.2 The Source Coding Theorem . . . 38

8 Strong Typicality **41**
8.1 Strong AEP . . . 41
8.2 Joint Typicality . . . 44
8.3 Conditional Strong AEP . . . 46

9 Channel Coding **49**
9.1 7.1 Definition and Capacity . . . 50
9.2 Discrete Memoryless Channel . . . 51
9.3 The Channel Coding Theorem . . . 54

10 Achievability of the Channel Coding Theorem **60**
10.1 Performance Analysis . . . 62
10.2 Implication of the Channel Coding Theorem . . . 65
10.3 Feedback . . . 65
10.4 Separation of Source and Channel Coding . . . 67

11 Rate Distortion Theory **69**
11.1 Single-Letter Distortion Measures . . . 69
11.2 The Rate Distortion Function . . . 72
11.3 The Rate Distortion Theorem . . . 75
11.4 The "Converse" of the Rate-Distortion Theorem . . . 79
11.5 Achievability of Information Rate-Distortion $R_I(D)$ . . . 81

12 The Blahut-Arimoto Algorithms **86**
12.1 Single-Letter Characterization . . . 86
12.2 Numerical Methods . . . 86
12.2.1 A double supremum . . . 86
12.2.2 An alternating optimization . . . 87
12.3 Computing Channel Capacity . . . 87
12.4 Algorithm for computing the rate-distortion function . . . 90
12.5 Convergence of the algorithm . . . 90
12.5.1 How to prove convergence . . . 90

13 Differential Entropy **93**
13.1 Real random variables . . . 93
13.2 Random Vectors . . . 95
13.3 Gaussian Distribution . . . 95
13.4 Definition . . . 97
13.4.1 Relation with Discrete Entropy . . . 99
13.5 Properties of Differential Entropy . . . 100
13.6 Joint Differential Entropy, Conditional Differential Entropy and Mutual Information . . . 101
13.7 Conditional Differential Entropy . . . 102
13.8 Differential Mutual Information . . . 103
13.9 Interpretation of $I(X;Y)$ . . . 104
13.10 AEP for Continuous Random Variables . . . 105
13.11 Differential Informational Divergence . . . 106
13.12 Maximum Discrete Entropy Distributions . . . 107
13.13 Maximum Differential Entropy Distributions . . . 109

13.14 Differential Entropy and Spread . . . 110
13.15 Continuous-Valued Channels . . . 111
13.16 Discrete-Time Continuous-Valued Channel . . . 111
13.17 The Channel Coding Theorem . . . 114
13.17.1 Assumptions . . . 114
13.17.2 Statement . . . 115
13.17.3 Proof of the Converse of the Channel Coding Theorem . . . 116
13.17.4 Proof of the Achievability part of the Channel Coding Theorem . . . 118

14 Memoryless Gaussian Channels **122**
14.1 Channel Capacity of Memoryless Gaussian . . . 122
14.2 Parallel Gaussian Channels . . . 124
14.3 Correlated Gaussian Channel . . . 127
14.3.1 Decorrelation of the Noise Vector . . . 127

15 Transmission in Continuous Time **129**
15.1 The Bandlimited White Gaussian Channel . . . 129
15.2 Signal Analysis Prelim . . . 129
15.3 Intuitive Treatment of the Bandlimited Channel . . . 133
15.3.1 Power Constraints . . . 135
15.4 The Bandlimited Colored Gaussian Channel . . . 135

Chapter 1: The Science of Information

- Founded by Claude E. Shannon (1916-2001)
- "The Mathematical Theory of Communication," 1948
- Study fundamental limits in communications: transmission, storage.
- information source → transmitter (signal + noise)→ receiver

2 Key Concepts

- Information is uncertainty: modeled as random variables
- Information is digital: Transmission should be 0's and 1's (bits) with no reference to what they represent.

2 Fundamental Theorems

- Source coding theorem: fundamental limit in data compression (zip for general file, especially for text, MP3 for audio, JPEG for image, MPEG for video)
- Channel coding theorem: fundamental limit for reliable communication through a noisy channel (telephone, cell phone, modem, data storage)

Chapter 2: Information Measures

2.1 Independence and Markov Chains

Notations

- X discrete random variable taking values in $\mathcal{X}$.
- $\mathcal{S}_X$: support of X, i.e. $\{x \in \mathcal{X} | p_X(x) > 0\}$
- If $\mathcal{S}_X = \mathcal{X}$, we say the probability measure p is strictly positive.
- Non-strictly positive distributions are dangerous

Definitions

- Independence and mutual independence: joint means multiplication
- Pairwise independent: implied by mutual independence, not vice versa.
- Conditional independence: $X \perp Z|Y$

Proposition 2.5

proposition 2.1. *For X, Y, and Z, and $X \perp Z|Y$ if and only if $p(x,y,z) = a(x,y)b(y,z)$ for any x, y, z such that $p(y) > 0$.*

Proof. Only if: $a(x,y) = \frac{p(x,y)}{p(y)}$, $b(y,z) = p(y,z)$. If: Let $p(x,y,z) = a(x,y)b(y,z)$. $p(x,y,z) = p(x)p(x|y)P(z|x,y)$, therefore $b(y,z)$ must contain $p(z|x,y)$, meaning x doesn't affect $p(z|x,y)$.
Rigorous proof:

$$p(x,y) = \sum_z p(x,y,z) = \sum_z a(x,y)b(y,z) = a(x,y)\sum_z b(y,z)$$

$$p(y,z) = b(y,z)\sum_x a(x,y)$$

$$p(y) = \sum_z p(y,z) = (\sum_x a(x,y))(\sum_z b(y,z))$$

With some algebra, we get $p(x,y,z) = \frac{p(x,y)p(y,z)}{p(y)}$ □

Markov Chain

Markov Chain: $X_1, X_2, ..., p(x_1, x_2, ...) = p(x_1, x_2)p(x_3|x_2)...p(x_n|x_{n-1}$
Remark: the reversed chain is still Markov. (symmetry of independence)

Prop

proposition 2.2. *$X_1, ..., X_n$ is a Markov chain iff $(x_1, ..., x_m), ..., x_n$ is an MC for any m.*

proposition 2.3. *$X_1, ..., X_n$ forms an MC if and only if $p(x_1, , , x_n) = f_1(x_1, x_2), ..., f_{n-1}(x_{n-1}, x_n)$.*
Generalization of Prop 2.5

proposition 2.4. *Any subchain of a Markov chain is a Markov chain!*

Proof. Never forget about law of total probability! □

2.2 2.2 Shannon's Information Measures

There are 4 types of information measures: entropy, conditional entropy, mutual information, and conditional mutual information.

Definition 2.5. The entropy $H(X)$ of a random variable X is defined as

$$H(X) = -\sum_x p(x) \log p(x) \tag{2.1}$$

Convention: summation is taken over $Support_X$.
When the base of the logarithm is a, write H as H_a.
The unit for entropy is bit if $a = 2$, nat if $a = e$, D-it if $a = D$.
The notion of bit in I.T. is very different from the notion of bit in CS.

Remark 2.6. $H(X)$ depends only on the distribution of X but not on the actual values taken by X, hence also write $H(p_X)$.

Entropy as Expectation

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

Binary Entropy Function

Definition 2.7. For $0 \leq \gamma \leq 1$, define the binary entropy function

$$h_b(\gamma) = -\gamma \log \gamma - (1 - \gamma) \log(1 - \gamma)$$

with the convention $0 \log 0 = 0$, as by L'Hopital's rule, $\lim_{a \to 0} a \log a = 0$

For binary random variable with distribution γ, $X \sim \{\gamma, 1 - \gamma\}, H(X) = h_b(\gamma)$.
As a function of γ, the curve looks like a cap.

Interpretation

Consider tossing a coin with $p(Head) = \gamma$, $p(Tail) = 1 - \gamma$. Then $h_b(\gamma)$ measures the amount of uncertainty in the outcome of the toss. When $\gamma = 0$ or 1, the coin is deterministic. When the coin is fail, the uncertainty is 1 bit.

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of random variables X and Y is defined as

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) = -E \log p(X, Y).$$

For random variables X and Y, the conditional entropy of Y given X is defined as $H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) = -E \log p(Y|X)$.

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

Remark 2.8. Conditional expectation is a random variable but conditional entropy is

Proposition 2.16

$$H(X, Y) = H(X) + H(Y|X) H(X, Y) = H(Y) + H(X|Y) \tag{2.2}$$

Proof.

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

□

The total amount of entropy is the same independent of the number of steps you take to reveal the random variables.

mutual information

The mutual information between X and Y is defined as

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = E \log \frac{p(X, Y)}{p(X)p(Y)}$$

Note the mutual information is symmetric.
Alternatively, $I(X;Y) = E \log \frac{p(X|Y)}{P(X)}$

proposition 2.9. *The mutual information between a random variable X and itself is equal to its entropy. (Proof is trivial.)*
The entropy is sometimes called self-information.

proposition 2.10.

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

and

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

if the quantities are finite.
Analogous to de Morgan, where I is like intersection.

Note that you can also have conditional mutual information:

$$I(X;Y|Z) := \sum_{x,y,z} p(x,yz) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = E \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$

The self-information argument also applies to conditional mutual information.

Remark 2.11. All Shannon's information measures are special cases of conditional mutual information. Let a denote a RV taking constant value. Then $H(X) = I(X;X|a)$...

2.3 Continuity of Shannon's information measures for fixed finite alphabets

All Shannon's information measures are continuous when the alphabets are fixed and finite. For countable alphabets, Shannon's information measures are everywhere discontinuous.

Definition 2.12. Let p and q be 2 probability distributions on a common alphabet $\mathcal{X}$. The variational distance (L_1 distance) between p and q is defined as

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

We can show that $\lim_{p' \to p} H(p') = H(\lim_{p' \to p} p') = H(p)$.

An example: Let $\mathcal{X}$ be the set of positive integers. $P_X = \{1, 0, 0, ...\}$. $P_{X_n} = \{1 - \frac{1}{\sqrt{\log n}}, ...\}$. The variational distance tends to 0. However, $H(P_X) = 0$ but $\lim_{n \to \infty} H(P_n) = \infty$.

Chapter 2 Continued

Introduction

Finish up chapter 2 and prove Fano's inequality.

3.1 2.4 Chain Rule

Chain Rule for Entropy

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

Note a special case: $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$. Note we also have chain rule for conditional entropy, mutual information, and conditional mutual information.

Proof. Induction. For the conditional version:

$$\begin{aligned} & H(X_1, X_2, \ldots, X_n | Y) \\ &= H(X_1, \ldots, X_n, Y) - H(Y) \\ &= H((X_1, Y), X_2, \ldots, X_n) - H(Y) \end{aligned}$$

□

3.2 2.5 Informational Divergence

Informational Divergence

The informational divergence between 2 probability distributions p and q on a common alphabet $\mathcal{X}$ is defined as

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)} \tag{3.1}$$

where E_p is the expectation with respect to p.

- Summation is over the support of p.

- $c \log \frac{c}{0} = \infty$ for $c > 0$.
- If $D(p||q) < \infty$, then $p(x) > 0 \implies q(x) > 0$ or $\mathcal{S}_p \subset \mathcal{S}_q$.
- $D(p||q)$ measures the "distance" between p and q, but the informational divergence is not symmetric and does not satisfy the triangle inequality.

The Fundamental Inequality

For any $a > 0$,

$$\ln a \leq a - 1$$

with equality if and only if $a = 1$.
This inequality is very clear.

corollary.

$$\ln a \geq 1 - \frac{1}{a}$$

Divergence Inequality

Theorem. *For any two probability distributions p and q on a common alphabet, $D(p||q) \geq 0$ with equality if and only if $p = q$.*

Proof. (Note also follows from Jensen's inequality as log is concave.)

$$\begin{aligned} D(p||q) &= \sum p(x) \log \frac{p(x)}{q(x)} \\ &= (\log e) \sum p(x) \ln \frac{p(x)}{q(x)} \\ &\geq \log e \sum p(x)(1 - \frac{q(x)}{p(x)}) \\ &= \log e \sum (p(x) - q(x)) \\ &= 0 \end{aligned}$$

□

Log-Sum Inequality

$$\sum_i a_i \log \frac{a_i}{b_i} \geq (\sum_i a_i) \log \frac{\sum a_i}{\sum b_i}$$

Proof. Let $a'_i = a_i / \sum a_j$ and $b'_i = b_i / \sum_j b_j$. So we constructed 2 probability measures, then we can apply the divergence inequality. □

Pinsker's Inequality

$$D(p||q) \geq \frac{1}{2 \ln 2} V^2(p, q)$$

This means if divergence is small, the L_1 distance is also small. Also, convergence in divergence is stronger than convergence in variational distance.

3.3 2.6 Basic Information

Mutual information is non-negative

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

with equality if and only if $X \perp Y|Z$.
A consequence is that all Shannon measures of information are non-negative.

Proof.

$$\begin{aligned} & I(X;Y|Z) \\ &= sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z}{p(x|z)p(y|z)} \\ &= \sum_z \sum_{x,y} p(z)p(x,y|z) \log \frac{p(x,y|z}{p(x|z)p(y|z)} \\ &= \sum_z p(z) D(P_{X,Y|z} || p_{X|z} p_{Y|z}) \end{aligned}$$

The mutual information is 0 if and only if X, Y are independent given Z. □

- $H(X) = 0$ if and only if X is deterministic.

  Proof. If: trivial Only if: If X is not deterministic, then there exists a point ... $H(X) > 0$. □

- $H(Y|X) = 0$ if and only if Y is a function of X. .

3.4 2.7 More Useful Inequalities

Conditioning does not increase entropy

$H(Y|X) \leq H(Y)$ with equality if and only if X and Y are independent.

Proof.

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

□

Warning: conditioning does not necessarily decrese mutual information.

Independence Bound for Entropy

Theorem.

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

with equality if and only if X_i are mutually independent.

Proof. Use chain rule and apply last theorem. □

Mutual information decreases when removing RV

Theorem 3.1.

$$I(X; Y, Z) \geq I(X; Y)$$

with equality if and only if X, Y, Z forms a Markov chain.

Proof. By the chain rule, we have $I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \geq I(X; Y)$. The above inequality is tight if and only if $I(X; Z|Y) = 0$ i.e. the future is independent of the past given present. □

Mutual information and Markov chain

Lemma 3.2. *If X, Y, Z forms a Markov chain, then $I(X; Z) \leq I(X; Y)$ and $I(X; Z) \leq I(Y; Z)$*

If 2 random variables in a chain are close, then they have high mutual info.

Proof. The first inequality is obtained by applying the chain rule: $I(X; Z) = I(X; Y, Z) - I(X; Y|Z) \leq I(X; Y, Z)$
$= I(X; Y) + I(X; Z|Y) = I(X; Y)$ The second inequality follows from the fact that the reversed chain is also markov. □

corollary. *If X, Y, Z Markov, then $H(X|Z) \geq H(X|Y)$.*

Suppose Y is an observation of X. Then further processing of Y can only increase the uncertainty about X on average.

Data Processing Theorem

If $U \to X \to Y \to V$ forms a Markov chain, then

$$I(U; V) \leq I(X; V)$$

Proof. Consider the subchain U, X, Y and U, Y, V. From the first chain, we have $I(U; Y) \leq I(X; Y)$. From the second chain, we have $I(U, V) \leq I(U, Y)$.
Therefore $I(U; V) \leq I(X; Y)$ □

This perspective is better!

Let $Z = g(Y)$, then

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

because $X, Y, g(Y)$ is Markov. We infer X from Y. If you manipulate your data, the inference cannot be improved.

3.5 Fano's Inequality

Upper bound of entropy

Theorem. *For any RV X,*

$$H(X) \leq \log |\mathcal{X}|$$

with equality if and only X is uniform.

Proof. Let $u(x) = \frac{1}{|\mathcal{X}|}$. Then

$$\begin{aligned} &\log |\mathcal{X}| - H(X) \\ &= -\sum_x p(x)u(x) + \sum_x p(x) \log p(x) \\ &= D(p||u) \geq 0 \end{aligned}$$

□

Note

- If alphabet is finite, then entropy is finite.
- The entropy of a RV may take any non-negative value: intermediate value theorem.

Example: Let X be RV st $p(i) = 2^{-i}$. Then $H_2(X) = 2$. Expectation of geometric distribution.

Fano's Inequality

Let X and $\hat{X}$ have the same alphabet. Then

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1)$$

, where $P_e = Pr(\hat{X} \neq X)$.

Proof. Define indicator Y, which is equal to 1 when we have error, i.e. $X \neq \hat{X}$. $P(Y = 1) = P_e$. $H(Y) = h_b P_e$. Since Y is a function of X and $\hat{X}$, $H(Y|X, \hat{X}) = 0$.

Then

$$\begin{aligned}
&H(X|\hat{X}) \\
&= H(X|\hat{X}) + H(Y|X,\hat{X}) \quad \text{addition by 0} \\
&= H(X,Y|\hat{X}) \\
&= H(Y|\hat{X}) + H(X|Y,\hat{X}) \\
&\leq H(Y) + \sum_{\hat{x}}[P(\hat{X}=\hat{x},Y=0)H(X|\hat{X}=\hat{x}Y=0) \\
&+ P(\hat{X}=\hat{x},Y=1)H(X|\hat{X}=\hat{x}Y=1)] \\
&= H(Y) + \sum_{\hat{x}} P(\hat{X}=\hat{x},Y=1)H(X|\hat{X}=\hat{x}Y=1) \\
&\leq H(Y) + \sum_{\hat{x}} P(\hat{X}=\hat{x},Y=1)\log(|X|-1) \\
&= h_b(P_e) + P_e \log(|X|-1)
\end{aligned}$$

corollary. *Weaker:*

$$H(X|\hat{X}) < 1 + P_e \log|\mathcal{X}|.$$

□

- Think of $\hat{X}$ as an estimate of X.
- P_e is the probability of error. If the error probability is small, then $H(X|\hat{X})$ should also be small.

3.6 Entropy Rate of a Stationary Source

3.6.1 Discrete-time Information Source

In most communication systems, communication takes place continually instead of over a finite period of time.
Exampls: TV broadcast, internet, cellular system. The information source can be modeled as a discrete-time random process $\{X_k, k \geq 1\}$.
$\{X_k, k \geq 1\}$ is an infinite collection of random variables indexed by the set of positive integers. The index k is referred as the "time."
Random variables X_k are called letters, and we assume they have finite entropy.

Total Entropy of $\{K_k\}$

For a finite subset A of the index set $\{k : k \geq 1\}$, the joint entropy $H(X_k, k \in A))$ is finite because

$$H(X_k, k \in A) \leq \sum_{k \in A} H(X_k) < \infty$$

by the independence bound.

Example, X_k are iid and $H(X_k) = h$. Then joint entropy $H(X_k, k \geq 1) = \sum_{k=1}^{\infty} H(X_k) = \sum_{k=1}^{\infty} h = \infty$. In general, it's not meaningful to discuss the joint entropy of an information process.

3.6.2 Entropy Rate

We are motivated to define the entropy rate of an information source instead, which gives the average entropy of a letter of the source.

Entropy Rate

The entropy rate of an information source $\{X_k\}$ is defined as

$$H_X = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

when the limit exists.

Entropy Rate May Exist

Let X_k be iid source with generic random variable X. Then

$$\lim_{n\to\infty} H(X_1, \ldots, X_n)/n = \lim_{n\to\infty} \frac{nH(X)}{n} = H(X)$$

, i.e., the entropy rate of an i.i.d. source is the entropy of any one of the letters.

Entropy Rate May Not Exist

Suppose the entropy grows linearly: Let $H(X_k) = k$. Then

$$\begin{aligned} \frac{1}{n} H(X_1, \ldots, X_n) &= \frac{1}{n} \sum_{k=1}^{n} H(X_k) \\ &= \frac{1}{n} \sum_{k=1}^{n} k \\ &= \frac{1}{n} \frac{n(n+1)}{2} \\ &\to \infty \end{aligned}$$

Note it's natural to consider

$$H'_X = \lim_{n\to\infty} H(X_n | X_1, \ldots, X_{n-1})$$

if it exists.

- This quantity H'_X is the conditional entropy of the next letter given all the history of the source.
- Fundamental question: $H'X = H_X$?

Stationary Information Source

A stationary information source is one such that any finite block of random variables and any of its time-shift versions have exactly the same joint distribution. That is $X_1, \ldots, X_m$ has the same joint distribution as $x_{l+1}, \ldots, X_{m+l}$.

Stationarity and H'_X

If a process is stationary, then H'_X exists.

Proof. Conditional entropy is lower bounded by 0. We can show that the sequence H(current letter|past) is non-increasing to show the limit H'_X exists by the monotone-bounded theorem.

$$\begin{aligned} & H(X_n|X_1, \ldots, X_n) \\ & \leq H(X_n|X_2, \ldots, X_{n-1}) \\ & = H(X_{n-1}|X_1, \ldots, X_{n-2}) \quad \text{by stationarity} \end{aligned}$$

□

Cesáro Mean

Consider a sequence $\{a_n\}$. Construct a sequence

$$b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$$

. b_n is the average of the first n terms, called the Ceséro meanmeans

Cesáro Mean goes to the same limit

Let b_n be a_n's cesaro mean, and $a_n \to a$. Then $b_n \to a$.

Proof. For all ϵ, there exists $N(\epsilon)$ so that the distance between a_n and a is less than ϵ once we go beyond the threshold N. Consider

$$\begin{aligned} |b_n - a| &= |\frac{1}{n} \sum_{i=1}^{n} (a_i - a)| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} |a_i - a| \end{aligned}$$

Break the summation into 2 parts: first part summing over $1 \leq i \leq N(\epsilon)$ □

With stationarity H_X is equal to H'_X

Proof. First of all, H'_X exists for stationary process. Let a_n be $H'_n = H(X_n|X_1, \ldots, X_{n-1})$. Let $b_n = \frac{1}{n} H(X_1, \ldots, X_n)$.

By the chain rule,

$$\frac{1}{n}H(X_1,\ldots,X_n)$$
$$=\frac{1}{n}\sum_{k=1}^{n}H(X_k|X_1,\ldots,X_{k-1})$$

Then the left hand side is just the cesaro mean. □

The entropy rate of an information source exists under a fairly general assumption - stationarity.
Note H'_X is an alternative definition for entropy rate for stationary process.

The I-Measure

Information diagram: Joint entropy is like the union. Mutual information is like the intersection. Conditional entropy of X is like the entropy exclusive to X.
Why does this work? The answer lies in I-Measure.

4.1 The Second Law of Thermodynamics

Second law: The entropy of an isolated system is non-decreasing.
In statistical thermodynamics, entropy is defined as the log of the number of microstates in the system: corresponding to information entropy if the states are equally likely.
We model an isolated system as a Markov chain with transitions governed by physical laws. There are 4 interpretations:

- Relative entropy/KL divergence $D(\mu_n||\mu'_n)$ decreases with n, where μ_n and μ'_n are 2 probability distributions on the state space of a Markov chain at time n. Like Cauchy sequence lol. Proof uses chain rule of KL divergence, splitted in 2 different ways, and a key observation is that the transition probabilities are the same because we are in the same system.

- Relative entropy between a distribution on states at time n and a stationary distribution decreases with n. This implies any state distribution gets closer and closer to the stationary distribution. This is a special case of the above.

- Entropy increases if the stationary distribution is uniform: because any distribution of states tend to the stationary distribution, and entropy is maximized by uniform distribution.

- $H(X_n|X_1)$ increases with n for stationary process. Note by stationarity, marginal entropy $H(X_n)$ is constant.

  $$\begin{aligned}
  &H(X_n|X_1)\\
  &\geq H(X_n|X_1, X_2\\
  &= H(X_n|X_2) \quad \text{by Markovianity, and entropy is quantity for distribution}\\
  &= H(X_{n-1}|X_1) \quad \text{by stationarity.}
  \end{aligned}$$

  We can also use the data processing inequality.

- Shuffles increase entropy: $H(TX) \geq H(X)$ where T is a random permutation of $\mathcal{X}$.

4.2 The Asymptotic Equipartiton Property AEP

In information theory, the AEP is analogous to the law of large numbers.

AEP

If $X_1, X_2, \ldots$ are iid $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \to H(X)$$

in measure.

Proof. Note functions of iid RVs are also iid RVs (in our case, the function refers to a probability measure), so by weak law of large numbers,

$$\begin{aligned} -\frac{1}{n}\log p(X_1, ..., X_n) = -\frac{1}{n}\sum_i \log p(X_i) \quad &\text{by independence} \\ \to -E\log p(X) &\text{in probabiilty} \\ = H(X) & \end{aligned}$$

□

Preliminaries: The I-Measure for two random variables

Substitution of Symbols:

- $H/I \iff \mu*$
- , $\iff \cup$
- ; $\iff \cap$
- | $\iff -$ (set subtraction, i.e., intersection with complement)
- This is exactly how we define the signed measure μ^*:
- $\mu^*(\hat{X}_1 - \hat{X}_2) := H(X_1|X_2)$
- $\mu^*(\hat{X}_2 - \hat{X}_1) = H(X_2|X_1)$
- $\mu^*(\hat{X}_1 \cap \hat{X}_2) = I(X_1; X_2)$

$\mu*$ is some measure (set-additive function, satisfy axioms for measure)
Inclusion-Exclusion formulation in set theory:

$$\mu^*(X_1 \cup X_2) = \mu^*(X_1) + \mu^*(X_2) - \mu^*(X_1 \cap X_2)$$

corresponding to $H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1; X_2)$.

Field

The field F_n generated by sets $X_1, ..., X_n$ is the collection os sets which can be obtained by any sequence of usual set operations (union, interaction, complement, and difference) on $X_1, ..., X_n$.
The atoms of F_n are sets of the form $\cap_{i=1}^n Y_i$ where Y_i is either X_i or its complement.

The condition is stronger than necessary.
Examples: Field generated by 2 sets: 4 atoms.
A measure is completely specified by its values on the atoms.

Construction of the I-Measure μ^*

Let $\tilde{X}$ be a set corresponding to a r.v. X.
Fix n and let $N_n = \{1, \dots, n\}$.
Let the universal set be $\omega = \cup_{i \in N_n} \tilde{X}_i$.
The atom $A_0 = \cap_{i \in N_n} \tilde{X}_i^C =$ is called the empty atom of F_n.
Let A denote the set of all other atoms of F_n, called non-empty atoms. $|A| = 2^n - 1$.
A measure μ on F_n is completely specified by its values on the atoms of A.
Let $X_G = (X_i, i \in G)$. $\tilde{X}_G = \cup_{i \in G} \tilde{X}_i$.

3.6

Let $\mathcal{B} = \{\tilde{X}_G : \text{G is a nonempty subset of } N_n\}$. Then a measure μ on F_n is completely specified by $\{\mu(B), B \in \mathcal{B}\}$, which can be any set of real numbers.

Remark 4.1. We saw that μ is determined by its values on A, the atoms. This theorem says μ can also be determined by its values on the unions.

- We can get that $\mu(A_1 \cap A_2) = \mu(A_1) + \mu(A_2) - \mu(A_1 \cup A_2)$, inclusion-exlusion formula (true for any set-additive measure).

Proof. For any nonempty atom $A \in \mathcal{A}$,

$$A = \Cap_{i=1}^n Y_i$$

where Y_i is either $\tilde{X}_i$ pr $\tilde{X}_i^C$. Then there exists at least one i such that $Y_i = \tilde{X}_i$; otherwise, A would be equal to the empty atom.
Next, we split the intersection into 2 parts...
Eventually $\mu(B)$ is related to $\mu(A)$ through an invertible linear transformation. □

4.2.1 One step away from construction of the I-measure μ^*

3.7 3.8 Two Lemmas

- $$\mu(A \cap B - C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C).$$

- $$I(X;Y|Z) = H(X,Z) + H(X,Y) - H(X,Y,Z) - H(Z)$$

Proof.

$$\begin{aligned}\mu(A \cap B - C) &= \mu(A - c) + (B - C) - \mu(A \cup B - C) \quad \text{by inclusion-exclusion} \\ &= (\mu(A \cup C) - \mu(C)) + (\mu(B \cup C) - \mu(C)) - (\mu(A \cup B \cup C) - \mu(C)) \\ &\text{by additivity of measure} \\ &= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup b \cup C) - \mu(C).\end{aligned}$$

□

Construction of the I-measure μ^* on F_n:

- μ^* is meaningful if it is consistent with all Shannon's information measures via the substitution of symbols: for all subsets G, G', G'' of N_n

$$\mu^*(\hat{X_G} \cap \hat{X_{G'}} - \hat{X''_G}) = I(X_G; X_{G'}|X_{G''})$$

Special Cases:

- $G'' = \emptyset$:

$$\mu^*(\hat{X_G} \cap \hat{X'_G} = I(X_G; X'_G)$$

- $G = G'$:

$$\mu * (\hat{X_G} - \hat{X''_G} = H(X_G|X''_G)$$

- $G = G''$ and $G'' =$, $\mu*$ is consistent with entropy.

3.9 Uniqueness of μ^*

μ^* is the unique signed measure on F_N which is consistent with all Shannon's information measures.

4.2.2 μ^* can be negative

For $n = 2$, μ^* is always non-negative. This is because the values of μ^* on the non-empty atoms of F_2 are all Shannon's information measures
For $n = 3$, $mu^*(\hat{X_1} \cap \hat{X_2} \cap \hat{X_3}) = I(X_1; X_2; X_3)$ can be negative (note this is not a Shannon's information measure). There are 7 non-empty atoms. One of the atoms corresponds to $I(X_1; X_2; X_3)$, which does not correspond to any information measure.

Ex 3.10 μ^* can be negative

Let X_1 and X_2 be independent binary random variables with uniform distributions: $P(X_i = 0) = P(X_i = 1) = 0.5$. Let $X_3 = (X_1 + X_2) \mod 2$ (check whether X_1 and X_2 are different).
Then $H(X_i) = 1$ for $i = 1, 2, 3$.
Also, the 3 RVs are pairwise independent. Therefore,

$$H(X_i; X_j) = H(X_i) + H(X_j) = 1 + 1 = 2$$

and

$$I(X_i; X_j) = 0.$$

$H(X_3|X_1; X_2) = 0$. Then by chain rule,

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1, X_2) + (X_3|X_1, X_2) \\ &= 2 + 0 = 2 \end{aligned}$$

Now

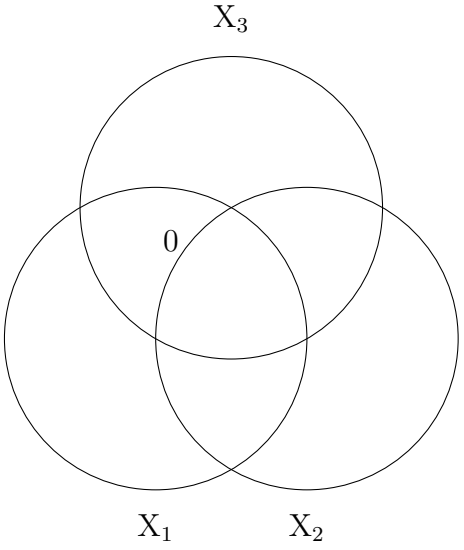
$$\begin{aligned} I(X_i; X_j|X_k) &= H(X_i, X_k) + H(X_j, X_k) - H(X_i, X_j, X_k) - H(X_k) \\ &= 2 + 2 - 2 - 1 = 1 \end{aligned}$$

Then

$$\begin{aligned} &\mu^*(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) \\ &= \mu^*(\tilde{X}_1 \cap \tilde{X}_2) - \mu^*(\hat{X}_1 \cap \hat{X}_2 - \hat{X}_3) \\ &= 0 - 1 = -1 < 0 \end{aligned}$$

I-measure for Markov chains

If $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain, then structure of μ^* becomes much simpler and the information diagram can be simplified.
For example, for $n = 3$, $X_1 \to X_2 \to X_3$ iff $I(X_1; X_3|X_2) = \mu^*(\tilde{X_1} \cap \tilde{X_3} - \tilde{X_2}) = 0$. Therefore, the atom $\tilde{X_1} \cap \tilde{X_3} - \tilde{X_2}$ can be neglected in the information diagram.



We can turn the diagram into 3 mountains now. We can easily demonstrate with the "mountains" that $I(X_1; X_3|X_2) = 0$.
Also $\mu^*(X_1 \cap X_2 \cap X_3) = \mu^*(X_1 \cap X_3)$.
Note μ^* for a Markov chains is a measure (values are all Shannon's information measures)

Markov chain with 5 RVs 5 atoms vanish.

Structure of μ^* for $X_1 \to X_2 \to X_3 \to X_4$

$$0 = I(X_1; X_3|X_2) = I(X_1; X_3; X_4|X_2) + I(X_1; X_3|X_2, X_4)$$

Let $I(X_1; X_3|X_2, X_4) = a \geq 0$. Then

$$I(X_1; X_3; X_4|X_2) = -a/$$

By considering the subchains $X_1 \to X_2 \to X_4$ and $X_1 \to X_3 \to X_4$... we get $a = 0$ so 5 atoms vanish. We can also check the non-negativity of μ^* on markov chain with 5 rvs.
In general, the information diagram for general markov chain can be represented by the "mountain" graph and μ^* is a measure.

5.1 Information Diagram Applications

- To obtain information identities is WYSIWYG What you see is what you get.
- To obtain information inequalities:
  - If μ^* is nonnegative, then $A \subset B \implies \mu^*(A) \leq \mu^*(B)$.
  - If μ^* is a signed measure, then it's more complicated to compare $\mu^*(A)$ and $\mu^*(B)$.

Concavity of Entropy

Let $X_1 \sim p_1$ and $X_2 \sim p_2$, and let X be a convex combination of them. Show that $H(X) \geq \lambda H(X_1) + \hat{\lambda} H(X_2)$.
Let Z be a random variable independent of X_1 and X_2 that takes value 1 with probability λ and value 2 with probability $1 - \lambda$. Let Z be the switch between X_1 and X_2 in $X \sim \lambda p_1(x) + \hat{\lambda} p_2(x)$. μ^* is a measure on 2 random variables X and Z:

$$\begin{aligned} H(X) &\geq H(X|Z) \\ &= P(Z=1)H(X|Z=1) + P(Z=2)H(X|Z=2) = \lambda H(X_) + \hat{\lambda} H(X_2) \end{aligned}$$

Interpretation: The entropy of a mixture of distributions is at least equal to the mixture of the corresponding entropies.

Convexity of Mutual Information in the "channel"

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Show that given $p(X)$, $I(X;Y)$ is a convex function of $p(y|x)$.

1. Let $p_1(y|x)$ and $p_2(y|x)$ be 2 transition matrices representing 2 channels.
2. Consider a system where the switch between 2 transition channels is determined by a random variable Z as in the previous example where Z is independent of X.
3. Let $I(X; Z|Y) = a \geq 0$. Then
$$I(X; Y : Z) = -a$$
because $I(X; Z) = 0$.
4. Then
$$\begin{aligned} I(X;Y) &\leq I(X;Y|Z) \\ &= P(Z=1)I(X;Y|Z=1) + P(Z=2)I(X;Y|Z=2) \end{aligned}$$

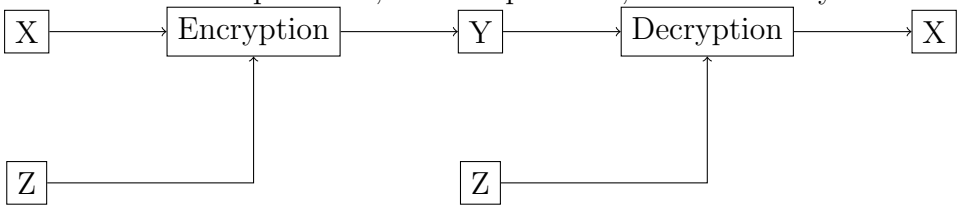
Interpretation: For a fixed input distribution $p(x)$, the mutual information between the input and the output of the system obtained by mixing 2 channels $p_1(y|x)$ and $p_2(y|x)$ is at most the mixture of the 2 mutual informations corresponding to $p_1(y|x)$ and $p_2(y|x)$.x

Concavity of Mutual Information in the "input"

Consider a system where the swtich on the 2 inputs is determined by Z. Then $Z \to X \to Y$ is a Markov chain.
In the information diagram, we see $I(X;Y) \geq I(X;Y|Z)$ because μ^* is a measure for Markov processes.
Now do the conditioning like before to finish the proof.
Interpretation: For a fixed channel, by mixing the input distribution, the mutual information is at least equal to the mixture of the corresponding mutual informations.

5.2 Shannon's Perfect Secrecy Theorem

Let X denote the plaintext, Y the ciphertext, and Z the key.



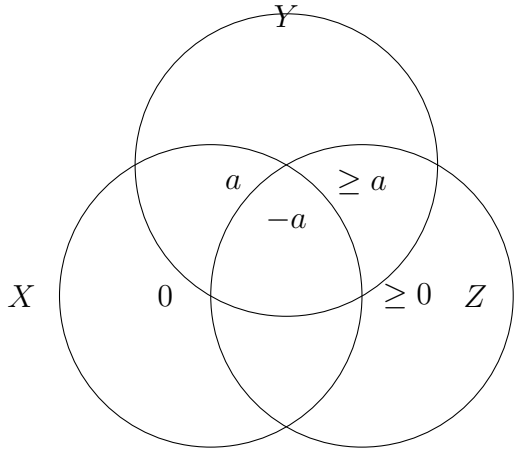
We define perfect secrecy as

$$I(X;Y) = 0.$$

We define decipherability/correctness as

$$H(X|Y,Z) = 0.$$

These requirement implies $H(Z) \geq H(X)$, i.e., the length of the key is at least the same as the length of the plaintext (I think we should note that this is assuming uniform distribution).
Shannon (1949) gave a combinatorial proof.
We can obtain the result using information diagram:



By looking at the atoms, we see that $\mu^*(Z) \geq \mu^*(X)$

Example: Imperfect Secrecy Theorem

Since X can be recovered from Y and Z, we have $H(X|Y,Z) = 0$. Show that this constraint implies $I(X;Y) \geq H(X) - H(Z)$.
Interpretation: $I(X;Y)$ measures the "leakage of information." When $I(X;Y) = 0$, it reduces to Shannon's perfect secrecy theorem.
Fangyuan: I drew the diagram on paper and saw that this is true.

Example: Data Processing Theorem

This theorem can be visualized using the information diagram, providing a more enlightening proof.

Remark 5.1. Many identities and equalities are difficult to discover without an information diagram.

Note Shannon's information inequalities implied by basic inequalities can be verified by ITIP. The system returns ture/false/not provable.

Zero-Error Data Compression

6.1 The Entropy Bound

In this section, we study why entropy is a fundamental bound for how much we can compress data.

D-ary Source Code

A D-ary source code $\mathcal{C}$ for a source random variable X is a mapping from the alphabet of $\mathcal{X} \to D^*$, the set of all finite length sequences of symbols taken from a D-ary code alphabet.

Decodability

A code $\mathcal{C}$ is uniquely decodable if for any finite source sequence, the sequence of code symbols corresponding to this source sequence is different from the sequence of code symbols corresponding to any other finite source sequence.

An example of source code

Let $\mathcal{X}$ be $\{A, B, C, D\}$. Consider the code $\mathcal{C}$ defined by

$$\begin{aligned} \mathcal{C}(A) &= 0 \\ \mathcal{C}(B) &= 1 \\ \mathcal{C}(C) &= 01 \\ \mathcal{C}(D) &= 10 \end{aligned}$$

For example,

- $x = B$ is a source symbol
- AAD is source sequence
- The mapping $\mathcal{C}(X)$ is a source code
- 0010 is a code sequence

Note $AAD \to 0010$, $ACA \to 0010, AABA \to 0010$, so the source code $\mathcal{C}$ is not uniquely decodable because given a code sequence, we cannot certainly recover the source sequence.

Kraft Inequality

Let $\mathcal{C}$ be a D-ary source code, and let $l_1, l_2, \ldots, l_m$ be the lengths of the codewords. If $\mathcal{C}$ is **uniquely decodable**, then

$$\sum_{k=1}^{m} D^{-l_k} \leq 1.$$

Proof Technique

Consider the code $\mathcal{C}$ in the previous example. Let $l_1 = l_2 = 1$ and $l_3 = l_4 = 2$. These correspond to the lengths of the codewords in $\mathcal{C}$.
Consider the polynomial $\sum_{k=1}^{4} 2^{-l_k} = 2^{-1} + 2^{-1} + 2^{-2} + 2^{-2}$.
If we raise the polynomial to the power 2 (this is a combinatorial argument. Think of each source symbol as a term in the polynomial and it quickly makes sense):

$$\begin{aligned}&(2^{-1} + 2^{-1} + 2^{-2} + 2^{-2})^2 \\ &= 4 \cdot 2^{-2} + 8 \cdot 2^{-3} + 4 \cdot 2^{-4} \\ &= A_1 \cdot 2^{-2} + A_2 \cdot 2^{-3} + A_3 \cdot 2^{-4}\end{aligned}$$

Then $A_2 = 4$ is the total number of sequences of $N = 2$ codewords with a toal length of 2 code symbols: $00(AA), 01(AB), 10(BA), 11(BB)$. Similarly, $A_3 = 8$ is the total number of sequnces of 2 codewords with a total length of 3 code symbols.

Proof. Without loss of generality, assume $l_1 \leq l_2 \leq \cdots \leq l_m$.
Let N be an arbitrary positive integer, and consider

$$\begin{aligned}&(\sum_{k=1}^{m} D^{-l_k})^N \\ &= \sum_{k_1}^{m} \sum_{k_2=1}^{m} \cdots \sum_{k_N=1}^{m} D^{-(l_{k_1} + \cdots + l_{k_N})} \quad \text{typical power of sum expansion}\end{aligned}$$

Next, express as

$$(\sum_{k=1}^{m} D^{-l_k})^N = \sum_{i=1}^{N \cdot l_m} A_i D^{-i}$$

Now note that A_i gives the total number of sequences of N codewords with a total length of i code symbols. (generalization of the example above).
Next, since the code is uniquely decodable, these code sequences must be distinct, and therefore

$$A_i \leq D^i$$

because there are D^i distinct sequences of i code symbols.
Then $(\sum_{k=1}^{m} D^{-l_k})^N \leq N \cdot l_m$.
Then $\sum_{k=1}^{m} D^{-l_k} \leq (N \cdot l_m)^{1/N}$ for any N. We obtain the inequality by letting

$N \to \infty$ (the limit is the empty product 1, the log argument).
(There's an interesting proof without heavy machinary that uses the inequality $n^{\frac{1}{n}} \leq 1 + 2\sqrt{\frac{1}{n}}$, though this inequality is hard to discover.) □

6.1.1 Expected Length of code

Remark 6.1. Source code in information theory is not the same as source code in software engineering. In information theory, source is a random variable, code is a mapping from the alphabet of source to some representations of the letters. In software enginnering, source code is the code a programmer writes.

Let $X \sim \{p_1, \cdots, p_m\}$.
The expected length of $\mathcal{C}$ is:

$$L = \sum_i p_i l_i$$

Intuitively, for a uniquely decodable code $\mathcal{C}$,

$$H_D(X) \leq L$$

because for X to be recoverable, and each $D - ary$ symbol can carry at most one $|D|$-it of information.

Entropy Bound

Let $\mathcal{C}$ be a D-ary **uniquely decodable code** for a source random variable X with entropy $H_D(X)$. Then the expected length of $\mathcal{C}$ is lower bounded by $H_D(X)$:

$$L \geq H_D(X).$$

This lower bound is tight if and only if $l_i = -\log_D p_i$ for all i.

Proof. Since $\mathcal{C}$ is uniquely decodable, the lengths of the codewords satisfy the kraft inequality.

$$L = \sum_i p_i l_i = \sum_i p_i \log_D D^{l_i}$$

Recall that

$$H_D(X) = -\sum_i p_i \log_D p_i$$

Then

$$\begin{aligned}
L - H_D(X) &= \sum_i p_i(\log_D p_i + \log_D D^{l_i}) \\
&= \sum_i p_i \log_D(p_i \times D^{l_i}) \\
&= (\ln D)^{-1} \sum_i p_i \ln(p_i \times D^{l_i}) \\
&\geq (\ln D)^{-1} \sum_i p_i(1 - \frac{1}{p_i D^{l_i}}) \quad \text{by the fundamental inequality} \\
&= (\ln D)^{-1} \sum_i (p_i - D^{-l_i}) \\
&= (\ln D)^{-1} \left[1 - \sum_i (-D^{-l_i})\right] \\
&\geq (\ln D)^{-1}(1-1) \\
&= 0
\end{aligned}$$

In order to achieve equality, we need equality in the fundamental inequality: $p_i D^{l_i} = 1$, so $l_i = -\log_D p_i$. If this is true, then

$$\sum_i D^{-l_i} = \sum_i p_i = 1,$$

so the kraft inequality is also tight, automatically. □

corollary.

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

We proved this earlier. The idea is to have a random variable and we represent the its source symbols by themselves (an $|\mathcal{X}|$-ary code). Here $L = 1$ and

$$1 = L \geq H_{|\mathcal{X}|}(X)$$

. We get the result through a change of basis.

Redundancy of a coe

The redundancy R of a D-ary uniquely decodable code is the difference between the expected length of the code and the entropy of the source.

$$R = L - H_D(X) \geq 0.$$

6.2 Prefix Codes

Prefix code is a very important class of uniquely decodable codes.

Prefix-free code

A code is called a prefix-free code if no codeword is a prefix of any other codeword. (For brevity, a prefix-free code is called a prefix code.lol)

Example

$$\mathcal{C}(A) = 0$$
$$\mathcal{C}(B) = 10$$
$$\mathcal{C}(C) = 110$$
$$\mathcal{C}(D) = 1111$$

Here $\mathcal{C}$ is a prefix-code.

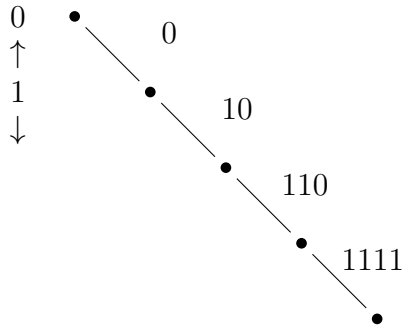
Code Tree for Prefix Code

- The tree representation of a prefix code is called a code tree.
- A D-ary tree is a graphical representation of a collection of finite sequences of D-ary symbols.
- A node is either an internal node or a leaf.

Instantaneous Decoding The the code $\mathcal{C}$ in the previous example (the pink box above), we see that

$$BCDAC\cdots \to 1011011110110\ldots$$

When we concatenate the codewords, their boundaries are not clear.
The stream of coded symbols are then transmitted to the receiver.
Since $\mathcal{C}$ is a prefix code, the codewords can be represented by a code tree!



Convention: Going up is 0 and going down is 1.
Instanteneous decoding: tracing the code tree from the root:

$$1011011110110 \to 10, 110, ...$$

First, we receive a 1, so starting at the root, we need to go down, then we see a 0, and we found a node that matches the code so far (10). For 110, restarting from the root, we see a 1, so go down, then another 1, then go down, then we see a 0, we enter the branch and get 110.
We say that prefix codes are self-punctuating.

Existance of prefix-code

There exists a $D-ary$ prefix code with codeword lengths $l_1, \ldots, l_m$ if and only if the krafy inequality

$$\sum_{k=1}^{m} D^{-l_k} \leq 1$$

is satisfied.

Proof. For the forward direction, we saw from the above algorithm that a prefix-code is uniquely decodable and hence satisfies Krafy inequality.
For the backward direction, assume WLOG $l_1 \leq \cdots \leq l_m$. Consider all the D-ary sequences of lengths less than or equal to l_m and regard them as the nodes of the D-ary tree of depth l_m. We will refer to sequence of lenfth l as a node of order l. There are $D^{l_1} > 1$ nodes of order l_1 can can be chosen as the first codeword. Thus choosing the first codeword is always possible. Assume that the first i codewordss have been chosen successfully where $1 \leq i \leq m-1$ and we want to choose a node of order $li+1$ as the $(i+1)$th codeword such that it is not prefixed by any of the previous chosen codewords.
Since all the previously chosen codewords are not prefixes of each other, their decendants of order l_{i+1} do not overlap. The $(i+1)$th node that we choose cannot be a secendant of any of the previously chosen codeword. For node of depth n, the number of descendants of order l_{i+1} is $D^{l_{i+1}-l_n}$.
Therefore the number of nodes which can be chosen as the $(i+1)$th codeword is

$$D^{l_{i+1}} - \sum_{k=1}^{i} D^{l_{i+1}-l_k}$$

If $l_1, ..., l_m$ satisfy the kraft inequality, we have $D^{-l_1} + \cdots + D^{l_{i+1}} \leq 1$.
Multuply by $D^{l_{i+1}}$ we get

$$\sum_{k=1}^{i+1} D^{l_{i+1}-l_k} \leq D^{i_{i+1}},$$

which implies

$$D^{l_{i+1}} - D^{l_{i+1}-l_1} - \cdots - D^{l_{i+1}-l_i} \geq 1$$

meaning a choice exists, so by induction, a prefix code exists. □

6.2.1 D-adic Distributions

D-adic Distributions

Let $p_i = D^{-t_i}$ for a;; i, where t_i is an integer.
This is called a dyadic distribution when $D = 2$.

corollary. *There exists a D-ary prefix code which achieves the entropy bound for a distribution $\{p_i\}$ if and only if $\{p_i\}$ is D-adic.*

Proof. Consider a D-ary prefix code which achieves the entropy bound for a distribution $\{p_i\}$. Let l_i be the length of the codeword assigned to the probability

p_i. Then by the entropy bound, $l_i = -\log_D p_i$ (condition for entropy bound to be tight), or

$$p_i = D^{-l_i}.$$

This $\{p_i\}$ is D-adic.
For the if direction: suppose $\{p_i\}$ is D-adic, then $t_i = -\log_D p_i$. Let $l_i = t_i$ for all i. verify that $\{l_i\}$ satisfies the Kraft inequality:

$$\sum_i D_{l_i} = \sum_i p_i = 1 \leq 1.$$

Then there exists a prefix code with codeword lengths l_i's. Assign the codeword with length l_i the probability p_i. □

Huffman Codes

A simple constrution of optimal prefix codes.

- Binary case: keep merging the 2 smallest probability masses until one probability mass (1) is left.
- D-ary Case: Insert zero probability masses until there are $D + k(D-1)$ masses, if necessary. Then keep merging the D smallest probability masses until one probability mass is left.
- In general, there can be more than one Huffman code.

Huffman Procedure

$$p = \{0.35, 0.1, 0.15, 0.2, 0.2\}$$

First we merge the 2 smallest probability masses to get

$$p = \{0.35, 0.25, 0.2, 0.2\}$$

. Then

$$p = \{0.35, 0.25, 0.4\}$$

. Then

$$p = \{0.6, 0.4\}.$$

By doing so, we have formed a code tree. Assign the codes by the convention that going up appends 0 and going down appends 1.

Lemmas for Optimality of Huffman Codes

Without loss of generality, assume $p_1 \geq p_2 \geq \cdots \geq p_m$, Denote the codeword assigned to p_i by c_i and its length by l_i.

Lemma 6.2. *In an optimal code, shorter codewords are assigned to larger probabilities.*

Proof. The expected length is improved (shortened) by swapping the a bad pair. □

Lemma 6.3. *There exists an optimal code in which the codewords assigned to the 2 smallest probabilities are siblings (of same length and only differ in the last symbol.)*

Proof. From the above lemma, the codeword c_m has the longest length because it has the smallest probability. Then the sibling of c_m canot be the prefix of another codeword.
The sibling of c_m must be a codeword because: otherwise, we replace c_m by its parent to improve the code, contradicting the optimality of the code. □

Here we introduce the concept of reduced code tree: suppose c_i and c_j are siblings in a code tree, if we merge them into their comman parent called c_{ij} and also merge the probabilities into $p_i + p_j$, we will imrove the expected codeword length by $p_i + p_j$.
Note this improvement only depends on the distribution over the words but not the structure of the reduced code tree.

Optimality of Huffman Procedure

Proof. Recursive proof: Consider an optimal code in which c_m and c_{m-1} are siblings. Existence by lemma.
Let p' be the reduced distribution by merging p_m and p_{m-1}. Let L and L' be the expected word lengths of the original and reduced code respectively.

$$L = L' + (p_{m-1} + p_m).$$

So L is the optimal length for p if and only if L' is the optimal length for p.
Reduce the problem all the way to 2 probability masses, for which optimal code is easy to find. □

The idea for D-ary Huffman Procedure is similar.

Upper Bound of the optimal codeword length!

The expected length of a Huffman code, denoted by L_{Huff}, satisfies the inequality

$$L_{Huffman} < H_D(X) + 1$$

This is the tightest bound among all upper bounds on L_{Huff} which depends only on the source random variable's entropy.

Proof. Construct a code with codeword lengths $l_i = \lceil -\log_D p_i \rceil$ and show that the Kraft inequality is satisfied.

$$-\log p_i \le l_i < -\log p_i + 1$$

$$\log p_i \le l_i > \log p_i - 1$$

$$p_i \ge D^{-l_i} \ge D^{-1} p_i$$

$$\sum_i D^{-l_i} \le \sum_i p_i = 1$$

. So prefix code exists. Then $L = \sum_i p_i l_i < H(X) + 1$.
Lastly, use the fact that $L_{Hugg} \leq L$ by optimality.
To show this is the tightest bound, we have to show that $H_D(X)+1$ can be achieved as a limit of a sequence of L_{Huff}s. Such sequence can be chosen as follows:

$$p_k = \{1 - \frac{D-1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\} \quad \text{(D-1) copies of } \frac{1}{k}.$$

As k goes to infinity, the entropy goes to 0. The Huffman code for each p_k consists of D codewords length 1. □

Asymptotic Achievability of $H(X)$

$$H(X) \leq L_{Huff} < H(X) + 1$$

If we use a Huffman code to encode n i.i.d. copies of X. Then

$$nH(X) \leq L^n_{Huff} < nH(X) + 1$$

Then

$$H(X) \leq \frac{1}{n} L^n_{Huff} < H(X) + \frac{1}{n} \to H(X)$$

as $n \to \infty$ $\frac{1}{n} L^n_{Huff}$ is called the **rate** of the code, measured in D-it per source symbol. This is the average number of $D-ary$ symbols to encode a source symbol(RV)

remark. Therefore, in asymptotic sense, the entropy $H(X)$ measures the amount of information contained in X.

Redundancy of Prefix Codes

In previous section, we have proved the entropy bound for a $D-ary$ uniquely decodable code

$$L \geq H_D(X)$$

We now prove this bound specially for prefix code and this will offer insights about the redundancy of a code. **A D-ary Code Tree**

- Let X be a source random variable with probability distribution $\{p_1, \ldots, p_m\}$
- A D-ary prefix code for X can be represented by a $D-$ary code tree with m leaves, where each leaf corresponds to a codeword.
- Ddenote the leaf (word) corresponding to p_i by c_i and the order of c_i by l_i (length)
- Let the alphabet be $\{0, \ldots, D-1\}$.
- Let I be the index set of all the internal codes (including the root) in the code tree. (a node that has at least one descendant.)

Reaching Probability

- To decode codeword of a prefix code, we can trace the path specified by the codeword from the root of the code tree until it terminates at the leaf corresponding to that codeword.
- Let q_k be the probability of reaching an internal node k during the decoding process.
- q_k is equal to the sum of the probabilities of all the leaves decending from node k.

Branching at an Internal Node

- Let $p_{k,j}$ be the probability that the jth branch of node k is taken during the decoding process.
- The probabilities $p_{k,j}$ where $0 \leq j \leq D-1$, are called the branching probabilities of node k, and
$$q_k = \sum_j p_{k,j}$$
- Once node k is reached, the conditional branching distribution is
$$\{\frac{p_{k,0}}{q_k}, \ldots, \frac{p_{k,D-1}}{q_k}\}$$
, obtained by normalizing $p_{k,\cdot}$.
- Then define the conditional entropy of node k by
$$h_k = H_D(\{\frac{p_{k,0}}{q_k}, \ldots, \frac{p_{k,D-1}}{q_k}\}) =\leq \log_D D = 1$$

4.19 A Lemma

$$H_D(X) = \sum_{k \in I} q_k h_k$$

where

- q_k is the reaching probability of internal node k
- h_k is the conditional entropy of internal node k.

Proof. Induction on the number of internal nodes.
For the base case, if there is only one internal node, it must be the root of the code tree. Then the lemma is trivially true because the reaching probability of the root is 1 and all the codewords have length of 1
Now inductive step: assume the lemma is true for code trees with n internal nodes. Let k be an internal code such that it is the parent of a leaf c with maximum order. Each sibling of c may or may not be a leaf. If it is not a leaf, it cannot be the ascendant of another leaf, otherwise contradicting the maximality of the order of c.
Now consider revealing the outcome of X in 2 steps. In the first step, if the outcome of X is not a leaf descending from k, we identify the outcome exactly, otherwise we identify the outcome to be a child of node k. We call this random variable V.
If we do not identify the outcome exactly in the first step, which happens with

probability q_k, we further identify in the second step which of the children of node k the outcome is. We call this random variable W.
Then $X = (V, W)$. The outcome of V can be represented by a code tree with n internal nodes which is obtained by prunning the original code tree at node k. By the inductive hypothesis, Then

$$H(X) = H(V, W) = H(V) + H(W|V)$$

□

4.20 Expected Length in terms of q_k

$$L = \sum_{k \in I} q_k$$

Proof. Easy to show by drawing a tree.
Rigorously, we can define a quantity

$$a_{ki} = \begin{cases} 1 & \text{if leaf } c_i \text{ is a descendent of internal node } k \\ 0 & \text{otherwise} \end{cases}$$

□

Local Redendancy

- Define the local redundancy of an internal node k by

$$r_k = q_k(1 - h_k)$$

- This quantity is local wrt note k in the sense that it only depends on the branching probabilities of node k.
- $r_k = 0$ if and only if

$$p_{k,j} = \frac{q_k}{D}$$

for all j, i.e. if and only if the internal node k is **balanced**.

Local Redundancy Theorem

Let R be the redundancy of a D-ary prefix code for a source random variable X. Then

$$R = \sum_{k \in I} r_k$$

Proof.

$$\begin{aligned} R &= L - H_D(X) \\ &= \sum_k q_k - \sum_k q_k h_k \\ &= \sum r_k \end{aligned}$$

□

Entropy Bound for Prefix Code

Let R be the redundancy of a prefix code. Then $R \geq 0$ with equality if and only if all the internal nodes in the code tree are balanced.

Proof. Consider

$$R = \sum_{k \in I} r_k.$$

$R \geq 0$ is non-negative because $r_k \geq 0$ for all internal nodes k.
$R = 0$ if and only all r_k are 0, meaning all of the internal nodes are balanced. □

- The entropy bound says that $H_D(X) \leq L$.
- This makes sense because intuitively each $D-ary$ symbol can carry at most 1 D-it of information.
- Therefore, when the entropy bound is tight, each code symbol has to carry exactly one $D-it$ of information.
- InterpretationL Consider revealing a a random variable codeword one symbol at a time. Balanced: as long as the codeword is not completed, the next code symbol to be revealed always carries one $D-it$ of information because it is distributed uniformly on the alphabet.

Now we discuss a lower bound on R. Consider

$$R = \sum_{k \in I} r_k \geq \sum_{k \in I'} r_k$$

for any subset I' of I.
If we can compute r_k for all $k \in I'$, we can compute a lower bound.

Weak Typicality

We will discuss

- for $X = (X_1, X_2, \ldots, X_n)$ where X_i are i.i.d. $\sim p(x)$, what a "typical" outcome of X would be.
- How typical sequences are related to data compression.
- Shannon's source coding theorem for data compression.

7.1 The Weak AEP

The Notion of Typical Sequences Consider tossing a fair coin n times. If the outcome is "head" approximately half of the time, the sequence of outcome is "normal" or "typical." How to quantify the typicality of a sequence?

Answer: weak and strong typicality.

The main theorems are weak and strong Asymptotic Equipartition Properties (AEP), consequences of weak law of large numbers. **Setup**

- $\{X_k, k \geq 1\}$, X_k iid $\sim p(x)$.
- X denotes generic rv with finite entropy.
- Let $\mathcal{X} = (X_1, \ldots, X_n)$, then $p(\vec{X} = p(X_1) \ldots p(X_n)$.
- The alphabet $\mathcal{X}$ is allowed to be countably infinite.
- Let the base of log be 2, so the unit of entropy $H(X)$ is bit.

Note that in Cover's textbook, the weak AEP is just called AEP.

Weak AEP I

$$-\frac{1}{n}\log p(X_1, ...X_n) \to H(X)$$

in probability as $n \to \infty$.
For any $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|-\frac{1}{n}\log p(X_1, ..., X_n) - H(X)| > \epsilon\} = 0.$$

Another way to say this is:

$$P\{|-\frac{1}{n}\log p(X_1,...,X_n)-H(X)|\leq\epsilon\}>1-\epsilon$$

Proof. Quick consequence of weak law of large numbers. By independence, $p(X_1,...,X_n)=p(X_1)\dots p(X_n)$. Then log of product is equal to sum of logs. Then sum becomes the sample mean. □

Weakly typical set

The weakly typical set $W^n_{X_\epsilon}$ with respect to $p(x)$ is the set of sequences $\vec{x}=(x_1,\dots,x_n)\in\mathcal{X}^n$ such that

$$|-\frac{1}{n}\log p(\vec{x})-H(X)|\leq\epsilon$$

, or equivalently

$$H(X)-\epsilon\leq-\frac{1}{n}\log p(\vec{x})\leq H(X)+\epsilon.$$

The sequences in W are called wealy $\epsilon-$typical sequences.

Empirical Entropy

The empirical entropy of a sequence $\vec{x}=(x_1,...,x_n)$ is defined as

$$-\frac{1}{n}\log p(\vec{x})=-\frac{1}{n}\log\Pi^n_{k=1}p(x_k)=-\frac{1}{n}\sum_{k=1}^{n}\log p(x_k)$$

A typical sequence is just a sequence whose empirical entropy is close to the true entropy $H(X)$.

Weak AEP II

For any $\epsilon>0$:

- If $\vec{x}\in W$, then
$$2^{-n(H(X)+\epsilon)}\leq p(\vec{x})\leq 2^{-n(H(X)-\epsilon)}$$
- For n sufficiently large,
$$P(\vec{X}\in W)>1-\epsilon.$$
- For n sufficiently large,
$$(1-\epsilon)2^{n(H(X)-\epsilon)}\leq|W|\leq 2^{n(H(X)+\epsilon)}$$

Proof.
1. The first bullet point is simply multiplying both sides of the equation in the definition of by n and then take exponential.
2. Re-expression of the same event.

3. From the first point, we get a lower bound on $p(\vec{x})$, then $|W| \cdot 2^{-n(H(X)+\epsilon)} \leq P(W) \leq 1$

□

Weak AEP says that

- For large n, the probability of occurrence of the sequence drawn is close to $2^{-nH(X)}$ with high probability. The probabilities of the typical sequences are almost the same.
- The total number of weakly typical sequences is approximately equal to $2^{nH(X)}$.

However, Weak AEP does NOT say that most sequences are weakly typical. It also does NOT say that the most likely sequences is weakly typical.

Example

Consider X such that $p(0) = 0.2$ and $p(1) = 0.8$.

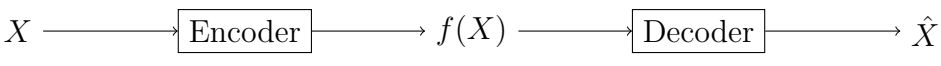
- The most likely sequence is $\vec{1} = (1, 1, \ldots, 1)$ with $p(\vec{1}) = 0.8^n$.
- Then
$$-\frac{1}{n}\log p(\vec{1}) = -\frac{1}{n}\log 0.8^n = -\log 0.8 \neq H(X)$$
, not close to true entropy, so $\vec{1}$ is not typical.
- This seems to be a contradiction because $P(W) \approx 1$ but $\vec{1} \notin W$. This is actually not a contradiction because $p(\vec{1}) \to 0.$x

When n is large, one can think of the sequence $\vec{X}$ as being obtained by choosing a sequence from the weakly typical set uniformly at random! (because the probabilities of the typical sequences are almost the same)
Note $|W| \approx 2^{nH(X)}$.

$$2^{nH(X)} << 2^{n\log|\mathcal{X}|} = |\mathcal{X}|^n = |X^n|$$

Therefore the size of the typical set would be much smaller than the size of all possible sequences., while having almost all the probability mass.

7.2 The Source Coding Theorem



source sequence reproduction sequence

- The encoder maps a random source sequence $\vec{X} \in \mathcal{X}^n$ to an index $f(\vec{X})$ in an index set $\{1, 2, \ldots, M\}$.
- Such a code is called a block code with block length n.

- The encoder sends $f(\mathcal{X})$ to the decoder through a noiseless channel.
- Based on the index, the decoder outputs $\hat{\vec{X}}$ as an estimate on $\vec{X}$.
- The encoder is specified by $f : \mathcal{X}^n \to \{1, ..., M\}$
- The rate of the code is given by $R = n^{-1} \log M$ in bits per source symbol, where M is the size of the index set and n is the block length.
- If $M = |\mathcal{X}^n|$, the rate of the code is $\log |\mathcal{X}|$.
- Typically, we take $R < \log |\mathcal{X}|$ for data compression. Notice that because the number of indices is less than the total number number of sequences, the decoder may not be able to reproduce the source sequence correctly.
- Denote the error probability by $P_e := P(\vec{X} \neq \hat{\vec{X}})$

The Source Coding Theorem

For arbitrarily small P_e, there exists a block code whose coding rate is arbitrarily close to $H(X)$ when n is sufficiently large.

- This direction says that reliable communication can be achieved if the coding rate at least $H(X)$.

For any block code with block length n and coding rate less than $H(X) - \zeta$, where $\zeta > 0$ does not change with n, then $P_e \to 1$ as $n \to \infty$.

- This says it's impossible to achieve reliable communication if the coding rate is less than $H(X)$.

Proof. For the first part, we need to construct a sequence of codes with block length n such that $P_e < \epsilon$ for large n.

Remark 7.1. The idea is to choose a subset A of $\mathcal{X}^n$ and let $M = |A|$. For each sequence $\vec{x}$ in A, assign it to a unique index $f(\vec{x})$. For each source sequence $\vec{x} \notin A$, just let $f(\vec{1})$.
For the decoding, if $\vec{x} \in A$, we are good. If not, then the decoding fails, so

$$P_e = P(\vec{X} \notin A)$$

For $\epsilon > 0$, consider the ϵ-typical set W_ϵ, and we take $M = |W|$
For sufficiently large n, the size of the W is close to $2^{n(H(X))}$
The coding rate satisfies $\frac{1}{n} \log(1 - \epsilon) + H(X) - \epsilon \leq \frac{1}{n} \log M \leq H(X) + \epsilon$, by Weak AEP, $P_e = P(\vec{X} \notin W)$. The coding rate tends to $H(X)$ and P_e tends to 0 as $\epsilon \to 0$.
For the second part, the idea is that the set A covers part of the typical set. □

Source Coding Theorem Converse

For any block code with blcok length n and coding rate less than $H(X) - \zeta$, where $\zeta > 0$ does not change with n, then $P_\epsilon \to 1$ as $n \to \infty$.

Proof.

$$rate = \frac{1}{n}\log M < H(X) - \zeta$$

where $\zeta > 0$ does not change with n. The total number of codewords

$$M \leq 2^{n(H(X)-\zeta)}$$

Some of the indices in our index set I cover $\vec{x} \in W_\epsilon$.
By WAEP, the total probabilty of typical sequences is upper bounded by

$$2^{n(H(X)-\zeta)}2^{-n(H(X))-\epsilon} = 2^{-n(\zeta-\epsilon)}$$

$$P(\vec{X} \in A) \leq 2^{-n(\zeta-\epsilon)} + P(X \notin W_\epsilon) < 2^{-n(\zeta-\epsilon)} + \epsilon$$

Let $\epsilon \to 0$, we see that the probability of not making mistake goes to 0. □

Strong Typicality

8.1 Strong AEP

The setup is as follows:

- $\{X_k\}$ is a sequence of i.i.d. random variables $\sim p(x)$.
- X denotes generic r.v. with entropy $H(X) < \infty$
- $\vec{X} = (X_1, \ldots, X_n)$. Then
$$p(\vec{X}) = p(X_1) \ldots p(X_n)$$
- New assumption: $|\mathcal{X}| < \infty$
- Let the base of the logarithm be 2, so $H(X)$ is in bits.

Notations:

- Let $N(x; \vec{x})$ ne the number of occurrences of x in the sequence $\vec{x}$.
- $n^{-1}N(x; \vec{x})$ is called the relative frequency of x in $\vec{x}$.
- $\{\frac{1}{n}N(x; \vec{x}) : x \in \mathcal{X}\}$ is called the empirical distribution of $\vec{x}$.

Example of Empirical Distribution

Let $\vec{x} = (1, 3, 2, 1, 1)$. Then $N(1; \vec{x}) = 3$ and $empircalDist(1) = \frac{3}{5}$.

Strongly Typical Set

The stronly typical set $T^n_{X^\delta}$ with respect to $p(x)$ is the set of sequences $\vec{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ such that
$$N(x; \vec{x}) = 0$$
for $x \notin S_X$ and
$$\sum_x |\frac{1}{n}N(x; \vec{x}) - p(x)| \leq \delta$$

remark.
- If $\sum_x |\frac{1}{n}N(x; \vec{x}) - p(x)|$ is small, then each term $|\frac{1}{n}N(x; \vec{x}) - p(x)|$ in the summand is also small.
- If $\vec{x}$ is strongly typical, the empirical distribution of $\vec{x}$ is approximately equal

to the true distribution $p(x)$.

Strong AEP

There exists $\eta > 0$ such that $\eta \to 0$ as $\delta \to 0$ and the following is true:

- If $\vec{x}$ is strongly δ-typical, then
$$2^{-n(H(X))+\eta} \leq p(\vec{x}) \leq 2^{-nH(X)-\eta}$$
- For n sufficiently large,
$$Pr(\vec{X} \in T_\delta) > 1 - \delta$$
- For n sufficiently large,
$$(1-\delta)2^{n(H(\vec{X})-\eta)} \leq |T_\delta^n| \leq 2^{n(H(\vec{X})+\eta)}$$

The form of the strong AEP is very similar to the form of the weak AEP.

Proof. Idea: If $\vec{x}$ is strongly typical, then the empirical distribution is "good" (close to the real distribution). If the empirical distribution is "good," the sample entropy should be close to the real entropy, i.e.

$$-\frac{1}{n}\log p(\vec{x}) \approx H(X),$$

meaning $p(\vec{x}) \approx 2^{-nH(X)}$.
For any $\vec{x} \in T_\delta^n$ that is δ typical, we have

$$p(\vec{x}) = p(x_1)\dots p(x_n) = \prod_{x \in \mathcal{S}_X} p(x)^{N(x;\vec{x})} > 0.$$

Then consider

$$\begin{aligned}
\log p(\vec{x}) &= \sum_x N(x;\vec{x}) \log p(x) \\
&= \sum_x (N(x;\vec{x}) - np(x) + np(x)) \log p(x) \\
&= n\sum_x p(x)\log p(x) - n\sum_x (\frac{1}{n}N(x) - p(x))(-\log p(x)) \\
&= -n[H(X) + \sum_x (\frac{1}{n}N(x) - p(x))(-\log p(x))].
\end{aligned}$$

Since $\vec{x} \in T$,

$$\sum_x |\frac{1}{n}N(x;\vec{x}) - p(x)| \leq \delta.$$

Now consider the triangle inequality,

$$\begin{aligned}
&|\sum_x (\frac{1}{n}N(x;\vec{x}) - p(x))(-\log p(x))| \\
&\leq \sum_x |\frac{1}{n}N(x;\vec{x}) - p(x)|(-\log p(x)) \\
&\leq -\log(\min_x p(x)) \sum_x |\frac{1}{n}N(x;\vec{x}) - p(x)| \\
&\leq -\delta \log(\min_x p(x)) \\
&= \eta
\end{aligned}$$

where

$$\eta = -\delta \log(\min_x p(x)).$$

Therefore,

$$-\eta \leq \sum_x (\frac{1}{n}N(x;\vec{x}) - p(x))(-\log p(x)) \leq \eta.$$

Then we get the desired inequality by taking exponential.
Note that η tends to 0 as δ tends to 0.

2): By WLLN, with probability tending to 1, the empirical distribution of $\vec{X}$ is close to $p(x)$, so by definition $\vec{X}$ is strongly typical. Consider

$$N(x;\vec{x}) = \sum_{k=1}^{n} I_k(x)$$

where I_k is an indicator for the event $X_k = x$. Then $I_k(x)$ are iid and $E(B_k(x)) = p(x)$.. By WLLN, for n sufficiently large, $Pr(|\frac{1}{n}\sum_{k=1}^{n} I_k(x) - p(x)| > \frac{\delta}{|\mathcal{X}|}) < \frac{\delta}{|\mathcal{X}|}$.
Then consider the probability that for some $x \in \mathcal{X}$, the empirical probability deviates from the true probability, we use the union bound to show is the this probability is upper bounded by δ.
Note that $\sum_x |\frac{1}{n}N(x;\vec{x}) - p(x)| > \delta$ implies $|\frac{1}{n}N(x;\vec{x}) - p(x)| > \frac{\delta}{||\mathcal{X}}$ for some x (at least one of them is greater than the average). We have the desired inequality.
3): $|T|lower_bound(p(\vec{x})) \leq P(T) \leq 1$ □

6.3 Probability that a sequence is not strongly typical tends to 0 exponentially fast

For sufficiently large n, there exists $\phi(\delta) > 0$ such that

$$P(\vec{X} \notin T_\delta^n)) < 2^{-n\phi(\delta)}.$$

The proof uses the Chernoff bound.

Strong Typicality versus Weak Typicality

remark.

- Weak Typicality: empirical entropy is close to the true entropy

- Strong Typicality: empirical distribution is close to the true distribution
- Strong typicality $\implies$ weak typicality.
- Weak typicality $\not\Rightarrow$ strong typicality
- Strong typicality works only for finite alphabet, i.e. $|\mathcal{X}| < \infty$, but weak typicality works for any countable alphabet.

Strong typicality implies weak typicality

If $\vec{x} \in T_\delta^n$, then $x \in W_\eta^n$ where $\eta \to 0$ as $\delta \to 0$.

Proof. Let $\vec{x}$ be strongly δ-typicality, then $2^{-n(H(X)+\eta)} \leq p(\vec{x}) \leq 2^{-n(H(X)-\eta)}$. Which means $H(X) - \eta \leq -\frac{1}{n}\log p(\vec{x}) \leq H(X) + \eta$. □

Weak Typicality does not imply strong typicality

Consider X with distribution p such that

$$p(0) = 0.5, p(1) = 0.25, p(2) = 0.25.$$

Consider a sequence $\vec{x}$ of length n and let $q(x) = n^{-1}N(x;\vec{x})$ be the relative frequency of occurrence of symbol x in $\vec{x}$, where $x = 0, 1, 2$. Consider

$$\begin{aligned} -\frac{1}{n}\log p(\vec{x}) &= -\frac{1}{n}\sum_k \log p(x_k) \\ &= -\frac{1}{n}[N(0;\vec{x})\log p(0) + N(1;\vec{x})\log p(1) + N(2;\vec{x})\log p(2)] \\ &= q(0)\log p(0) + q(1)\log p(1) + q(2)\log p(2) \end{aligned}$$

While real entropy is

$$p(0)\log p(0) + p(1)\log p(1) + p(2)\log p(2)$$

If we have half 0 and half 1, then the empirical entropy is exactly the same as the true entropy.

8.2 Joint Typicality

Setup:

- $\{(X_k, Y_k)\}$ where (X_k, Y_k) are iid $\sim p(x, y)$
- (X, Y) denotes pair of generic r.v. with entropy $H(X, Y) < \infty$
- $|\mathcal{X}|, |\mathcal{Y}| < \infty$

Notations

- Consider $(x, y) \in \mathcal{X}^n \times \mathcal{Y}^n$, let $N(x, y; \vec{x}, \vec{y})$ be the number of occurrences of (x, y) in the pair of sequences $(\vec{x}, \vec{y})$

- $\frac{1}{n}N(x,y;\vec{x},\vec{y})$ would be the relative frequency of the tuple (x,y)

Strongly jointly typical set

The strongly jointly typical set T_δ^n with respect to $p(x,y)$ is the set of sequence pairs $(\vec{x},\vec{y})$ such that

$$\sum_x \sum_y |\frac{1}{n}N(x,y;\vec{x},\vec{y}) - p(x,y)| \leq \delta$$

Consistency

If $(\vec{x},\vec{y})$ is jointly typical, then $\vec{x}$ and $\vec{y}$ are both marginally typical.

Preservation

Let $Y = f(X)$. If

$$\vec{x} = (x_1,\ldots,x_n) \in T_X$$

, then $f(\vec{x}) = (y_1,\ldots,y_n) \in T_Y$ where $y_i = f(x_i)$

Strong JAEP

Let $(\vec{X},\vec{Y})$ be a pair of sequences of iid tuple, then there exists $\lambda > 0$ such that $\lambda \to 0$ as $\delta \to 0$ and the following hold:

- If $(\vec{x},\vec{y}) \in T$, then

$$2^{-nH(X,Y)+\lambda} \leq p(\vec{x},\vec{y}) \leq 2^{-n(H(X,Y)-\lambda}$$

- For n sufficiently large,

$$Pr((\vec{X},\vec{Y}) \in T) > 1-\delta$$

- For n sufficiently large,

$$(1-\delta)2^{n(H(X,Y)-\lambda)} \leq |T| \leq 2^{nH(X,Y)+\lambda}$$

Stirling's Approximation

$$\ln n! \approx n\ln n$$

Proof.

$$\ln n! = \ln 1 + \cdots + \ln n$$

Since $\ln x$ is monotonically increasing, we have

$$\int_{k-1}^{k} \ln x\,\mathrm{d}x < \ln k < \int_{k}^{k+1} \ln x\,\mathrm{d}x$$

□

Sum over $1 \leq k \leq n$, we have

$$\int_0^n \ln x \, \mathrm{d}x < \ln k < \int_1^{n+1} \ln x \, \mathrm{d}x$$

ie

$$n \ln n - n < \ln n! < (n+1)\ln(n+1) - n$$

Binomial coefficient

$$\binom{n}{np, n(1-p)} \approx 2^{nH_2(p)}$$

Proof.

$$\binom{n}{np, n(1-p)} = \frac{n!}{(np)!(n(1-p))!}$$

Then take logarithm, we have

$$\ln \binom{n}{np, n(1-p)} = \ln n! - \ln(np! - \ln(n(1-p))!$$

This quentity is approximately $nH_e(\{p, 1-p\})$ by Sterling's approximation. Lastly, change the base to 2. □

8.3 Conditional Strong AEP

Conditional Strong AEP

For any $\vec{x} \in T$, define

$$T^n_{Y|X_\delta}(\vec{x}) = \{\vec{y} \in T^n_Y : (\vec{x}, \vec{y}) \in T^n_{XY}$$

If $|T_{Y|X}| > 1$, then

$$2^{n(H(Y|X)-\nu)} \leq |T^n_{Y|X}(\vec{x})| \leq 2^{n(H(Y|X)+\nu)}$$

where $\nu \to 0$ as $n \to \infty$ and $\delta \to 0$

Note that since

$$\frac{|T_{Y|X}|}{|T_X|} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}} = 2^{n(H(X,Y)-H(X))} = 2^{nH(Y|X)}$$

, so the number of $\vec{y}$ that are jointly typical with a typical $\vec{x}$ is approximately equal to $2^{nH(Y|X)}$

Upper bound of conditional SAEP

If $|T^n_{Y|X_\delta}(\vec{x})| \geq 1$, then

$$|T^n_{Y|X_\delta}(\vec{x}) \leq 2^{n(H(Y|X)+\nu)}|$$

where $\nu \to 0$ as $n \to \infty$ and $\delta \to 0$.

Proof. For any $\nu > 0$, consider

$$\begin{aligned}
2^{-n(H(X)-\nu/2)} &\geq p(\vec{x}) \quad \text{by typicality of } \vec{x} \\
&= \sum_y p(x,y) \\
&\geq \sum_{y \in T^n_{Y|X}} p(x,y) \\
&\geq \sum_{y \in T_{Y|X}} 2^{-n(H(X,Y)+\nu/2)} \quad \text{by joint typicality} \\
&= |T^n_{Y|X}| 2^{-n(H(X,Y)+\nu/2)}
\end{aligned}$$

□

The proof of the lower bound is more complicated.

Lower Bound

Consider $\mathcal{X} = \{0,1\}$, $\{\mathcal{Y} = \{a,b,c\}$. There are probably $np(0,a)$ occurrences of the pair $(0,a)$, etc. If we rearrange the components of $\vec{y}$, corresponding to $x_k = 0$, then joint typicality is preserved. Then we can compute the number of arrangements. Then use the binomial coefficient approximation. As long as there's a typical $\vec{y}$, we can get a approx $2^{nH(Y|X)}$.

6.12 Corollary

Let S_X be the set of all sequences $\vec{x} \in T_X$ such that $T^n_{Y|X}$ is nonempty. Then

$$|S_X| \geq (1-\delta)2^{n(H(X)-\psi)},$$

where $\psi \to 0$ as $n \to \infty$ and $\delta \to 0$. This says S_X and T_X grows at the same asymptotic rate.

Proof. By consistentcy of strong typicality, if $(\vec{x},\vec{y})$ is jointly typical, $\vec{x}$ and $\vec{y}$ are marginally typical. Then

$$T^n_{XY} = \bigcup_{\vec{x} \in S_X} \{(\vec{x},\vec{y}) : \vec{y} \in T^n_{Y|X}\}$$

Using the lower bound on $|T^n_{XY}|$ in strong JAEP and upper bound on $T^n_{Y|X}$, we get the desired inequality, where ψ is $\lambda + \delta$ □

proposition. *With respect to a joint distribution $p(x,y)$, for any $\delta > 0$,*

$$Pr\{\vec{X} \in S^n_X\} > 1 - \delta$$

for n sufficiently large.
This says with high probability, we can obtain sequence such that there exists a $\vec{y}$ that is jointly typical with it.

Joint typicality has an "asymptotic quasi-uniform" structure. We can draw an array to visualize. The rows are $2^{nH(X)}$ typical x sequences. The columns are $2^{nH(Y)}$ typical y sequences. The total number of points in this array where $\vec{x}$ and $\vec{y}$ are jointly typical are $2^{nH(X,Y)}$. If we fix a typical x sequence, the number of dots in that row is $2^{nH(Y|X)}$ dots. (every row in a strongly typical array has approx the same number of dots).

Interpretation of the basic inequalities Since the number of dots is less than or equal to the number of cells, we have that

$$2^{nH(X,Y)} \leq 2^{nH(X)}2^{nH(Y)}$$

$$\text{or } H(X,Y) \leq H(X) + H(Y)$$
or

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

The quasi-uniform array provides a combinatorial interpretation of information inequalities.

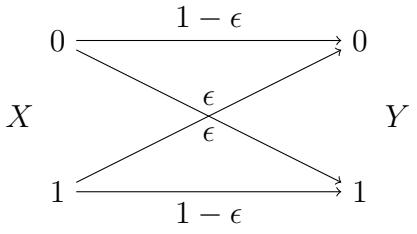
Channel Coding

This chapter discusses communication through a noisy channel reliably at optimal rate.

Discrete Memoryless Channel DMC

Binary Symmetric Channel (BSC): the simplest possible channel.

- input alphabet $\mathcal{X} = \{0, 1\}$
- output alphabet $\mathcal{Y} = \{0, 1\}$



- crossover probability $= \epsilon$, $0 \leq \epsilon \leq 1$.

A simple code:

- Assume $\epsilon < 0.5$.
- Two possible messages A, B are to be sent throug the channel.
- Coding scheme 1:

$$Encoding \begin{cases} A \to 0 \\ B \to 1 \end{cases}$$

$$Decoding \begin{cases} 0 \to A \\ 1 \to B \end{cases}$$

- a decoding error occurs if and only if a crossover occurs. Therefore $P_e = \epsilon$

A more elaborated code: repetition code:

- To improve reliability, use the BSC n ties for a large n.
- Let $N_0 = \#0$'s received and $N_1 = \#1$'s received.

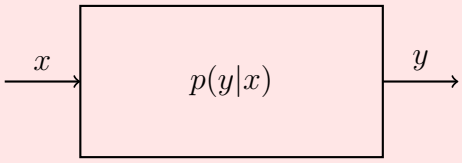
- Coding scheme 2:

$$\textit{Encoding} \begin{cases} A \to 00\ldots0 \\ B \to 11\ldots1 \end{cases} \quad \textit{Decoding} \begin{cases} N_0 > N_1 \to A \\ N_0 > N_1 \to B \end{cases}$$

- If the message if A, by WLLN, $N_0 \to n(1-\epsilon)$ and $N_1 \to n\epsilon$ in probability.
- The decoding rate goes to 1.
- However, coding rate $R = \frac{1}{n}\log 2 \to 0$ as $n \to \infty$. Coding rate is the amount of information transmitted per use of the channel.

9.1 7.1 Definition and Capacity

Discrete Memoryless Channel 1



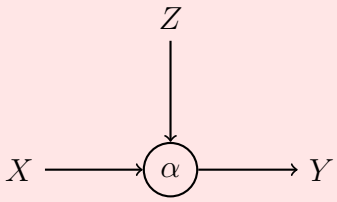
- Single input random variable X takes values in discrete alphabet $\mathcal{X}$.
- Single Output random variable takes values in discrete output alphabet $\mathcal{Y}$.
- The channel is specified by a transition matrix $p(y|x)$ from $\mathcal{X}$ to $\mathcal{Y}$.
- Input-output relation:

$$Pr\{X = x, Y = y\} = Pr\{X = x\}p(y|x)$$

- BSC with cross probability ϵ

$$[p(y|x)] = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$$

Discrete Memoryless Channel 2



- Input random variable X takes values in discrete alphabet $\mathcal{X}$.
- Output random variable takes values in discrete output alphabet $\mathcal{Y}$.
- Noise variable Z takes values in discrete alphabet $\mathcal{Z}$

- Z is independent of X
- α is a function from $\mathcal{X} \times \mathcal{Z}$ to $\mathcal{Y}$
- Input-output relation:
$$Y = \alpha(X, Z)$$
- The channel is specified by (α, Z)
- The channel is specified by a transition matrix $p(y|x)$ from $\mathcal{X}$ to $\mathcal{Y}$.
- Input-output relation:
$$Pr\{X = x, Y = y\} = Pr\{X = x\}p(y|x)$$

Equivalence of Discrete Channel 1 and Discrete Channel 2.
If a channel can be modeled by DS2, then it can also be modeled by discrete channel 1.
Given we can let $p(y|x) = Pr(a(X, Z) = y|X = x) = P(\alpha(x, Z) = y)$, by independence.
If a channel can be modeled by DS1, then it can also be modeled by discrete channel 2.
For $x \in \mathcal{X}$, define Z_x with $\mathcal{Z}_x = \mathcal{Y}$ such that $Pe\{Z_x = y\} = p(y|x)$.
Assume that $Z_x, x \in \mathcal{X}$ are mutually independent and also independent of X.
Define the noise $Z = (Z_x : x \in \mathcal{X})$. Let $Y = Z_x$ if $X = x$ so that $Y = \alpha(X, Z)$.

$$\begin{aligned}
\Pr\{X = x, Y = y\} &= \Pr\{X = x\} \Pr\{Y = y \mid X = x\} \\
&= \Pr\{X = x\} \Pr\{Z_x = y \mid X = x\} \\
&= \Pr\{X = x\} \Pr\{Z_x = y\} \\
&= Pr\{X = x\}p(y|x)
\end{aligned}$$

Equivalence of channels

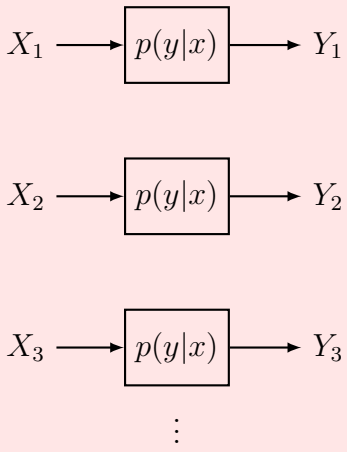
2 channels $p(y|x)$ and (α, Z) are equivalent if
$$Pr\{\alpha(x, Z) = y\} = p(y|x)$$
for all x and y.

9.2 Discrete Memoryless Channel

- A discrete channel can be used repeatedly at every time index $i = 1, 2, \ldots$.
- Assume the noise for th transmission over the channel at different time indices are independent of each other.
- We regard DMC as a subsystem of a discrete-time stochastic system which will be referred to as "the system".
- In such a system, random variables are generated sequentially in discrete-time.

- More than one random variable may be generated instantaneously, but sequentially at a particular time index.

DMC 1

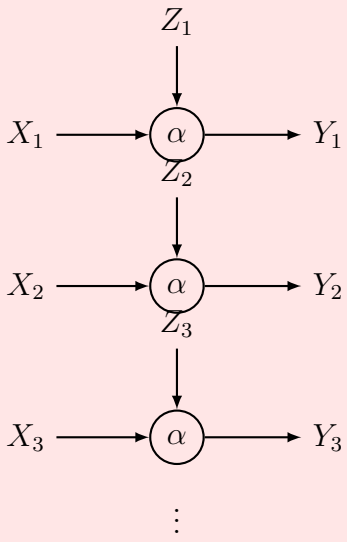


- A DMC specified by $p(y|x)$ is a sequence of replicates of a generic discrete channel $p(y|x)$
- Let T_{i-} be the collection of all random variables in the system generated before X_i.
- Memoryless property (independent noise):

$$Pr\{Y_i = y, X_i = x, T_{i-} = t\} = Pr\{X_i = x, T_{i-} = t\}p(y|x)$$

- note that $p(y|x)$ represents $Pr(Y_i = y|X_i = x|T_{i-} = t)$
- Equivalently, $T_{i-} \to X_i \to Y_i$ forms a Markov chain.

DMC 2 alternative definition



- A DMC specified by (α, Z) is a sequence of replicates of a generic discrete

channel (α, Z). The output of the DMC at time i is given by $Y_i = \alpha(X_i, Z_i)$

- Z_i is the noise variable for transmission at time i, and has the same distribution as Z.
- Memoryless Property (Independent noise): Z_i is independent of (X_i, T_{i-})
- This channel is equivalent to the DMC 1 above.

Capacity of a channel

The capacity of a discrete memoryless channel $p(y|x)$ is defined as

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

where X and Y are respectively the input and th output of the generic discrete channel, and the maximum is taken over all input distributions $p(x)$.

For each input didtribution $p(x)$, we have

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

From $p(x,y)$, we can compute $I(X;Y)$. Maximize $I(X;Y)$ over all input distributions.

Remark 9.1. Since $I(X;Y)$ is a continuous functional of $p(x)$ and the set of all $p(x)$ is a compact set, closed and bounded in $\mathcal{R}^{|\mathcal{X}|}$, the maximum value of $I(X;Y)$ can always be attained.

We will see that C is the maximum rate at which information can be reliably communicate through a DMC.

We will see that we can communicate through a channel with positive rate while $P_e \to 0$.

Example BSC

Consider $Y = X + Z \mod 2$, with

$$P(Z=0) = 1-\epsilon, P(Z=1) = \epsilon$$

and Z is independent of X. (If $Z = 1$, crossover occurs.)

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) h_b(\epsilon) \\ &= H(Y) - h_b(\epsilon) \\ &\le 1 - h_b(\epsilon) \quad \text{because } Y \text{ carries at most 1 bit of information} \end{aligned}$$

Therefore

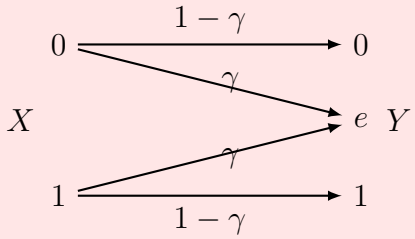
$$C = \max_{p(x)} I(X;Y) \le 1 - h_b(\epsilon)$$

The upper bound on $I(X;Y)$ is tight if $H(Y=1)$, and this can be achieved by taking the uniform input distribution.
We say that the channel has capacity $C = 1 - h_b(\epsilon)$ bit per use.

Remark 9.2.

- when $\epsilon = 0$, $C(0) = 1$. Intuitively, this means the channel doesn't error and we can communicate information at 1 bit per use.
- When $\epsilon = 1$, $C(0) = 1$ as well. This is symmetric to the case $\epsilon = 0$ where the decoder just flips the information.
- X and Y are independent for $\epsilon = 0.5$, in this case, we cannot communicate information. Also Y follows unfirom distribution in this case.

Example: Binary Erasure Channel



- The output symbol e denotes "erasure."
- γ: Erasure probability $0 \leq \gamma \leq 1$.
- With probability $1 - \gamma$, $Y = X$ (regardless of whether $X = 0$ or $X = 1$)
- With probability γ, $Y = e$ (erasure)
- Note there is either no error, or there is an erasure.

$$\begin{aligned} C &= \max_{p}(x) I(X;Y) \\ &= \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(Y) - h_b(\gamma) \\ &= (1-\gamma) \max_{p(0)} h_b(p(0)) \end{aligned}$$

The channel capacity is $1 - \gamma$ per use.

9.3 The Channel Coding Theorem

- Direct direction: Information can be communicated through a DMC with an arbitrarily small probability of error at any rate less than the channel capacity.
- Converse: If information is communicated through a DMC at a rate higher than the capacity, then the probability of error is bounded away from zero.

Channel Code

An (n, M) code for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by an encoding function

$$f : \{1, \ldots, M\} \to \mathcal{X}^n$$

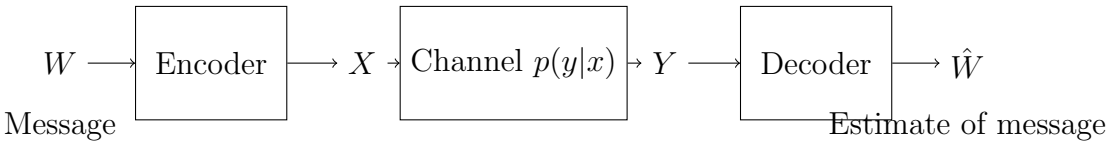
and a decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}.$$

- n: block length
- $\mathcal{W} = \{1, \ldots, M\}$: Message Set
- $f(1), \ldots, f(M)$: codewords
- the set of all codewords: codebook

Assumptions:

- W is uniformly chosen from the message set $\mathcal{W}$, so $H(W) = \log M$
- $\vec{X} = (X_1, ..., X_n), \vec{Y} = (Y_1, \ldots, Y_n)$.
- Thus $X = f(W)$, i.e. the transmitted sequence is a codeword for the chosen message.
- Let $\hat{W} = g(\vec{Y})$ denote the estimate on the message W by the decoder.



Performance Measures of the channel code

Error Probability

For all $1 \leq w \leq M$, let

$$\lambda_w = Pr\{\hat{W} \neq w | W = w\} = \sum_{\vec{y} \in Y^n : g(\vec{y}) \neq w} Pr\{\vec{Y} = \vec{y} | \vec{X} = f(w)\}$$

be the conditional probability of error given the message is w.
The maximal probability of error of an (n, M) code is $\lambda_{max} = \max_w \lambda_w$. It's the error probability of the worst codeword.
The average probability of error of an (n, M) code is defined as

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

P_e and λ_{max} Consider

$$\begin{aligned} P_e &= \Pr\{\hat{W} \neq W\} \\ &= \sum_w \Pr(W = w) \Pr(\hat{W} \neq W | W = w) \\ &= \sum_w \frac{1}{M} \Pr(\hat{W} \neq w | W = w) \\ &= \frac{1}{M} \sum_w \lambda_w \end{aligned}$$

So

$$P_e \leq \max_w \lambda_w \leq \lambda_{max}$$

Rate of a Channel Code

The rate of an (n, M) channel code is $n^{-1} \log M$ in bits per use.
A rate R is asymptotically **achievable** for a discrete memoryless channel if for any $\epsilon > 0$, for sufficiently large n there exists an (n, M) code such that

$$\frac{1}{n} \log M > R - \epsilon$$

and

$$\lambda_{max} < \epsilon$$

The following is our main theorem of the chapter.

Channel Coding Theorem

A rate R is achievable for a discrete memoryless channel if and only if $R \leq C$, the capacity of the channel.

Now we prove the converse of the channel coding theorem.

- The communication system consists of the r.v.'s $W, X_1, Y_1, \ldots, X_n, Y_n, \hat{W}$.
- The memorylessness of the DMC imposes the following Markov constraint for each i :
$$(W, X_1, Y_1, \ldots, X_{i-1}, Y_{i-1}) \to X_i \to Y_i$$
- Let q denote the joint distribution and marginal distribution of the rvs. Then for all $(w, \vec{x}, \vec{y}.\hat{w}) \in W \times \mathcal{X}^n \times \mathcal{Y}^n \times W$ such that $q(\vec{x}) > 0$ and $q(\vec{y}) > 0$,
$$q(w, \vec{x}, \vec{y}, \hat{w}) = q(w)(\prod_{i=1}^{n} q(x_i|w))(\prod_{i=1} p(y_i|x_i))q(\hat{w}|\vec{y})$$
- $W \to \vec{X} \to \vec{Y} \to \hat{W}$ forms a Markov chain.

Propositions

$$q(\vec{y}|\vec{x}) = \prod_{i=1}^{n} p(y_i|x_i)$$

Proof. Apply law of total probability to sum over w and $\hat{w}$. □

$$H(\vec{Y}|\vec{X}) = \sum_{i=1}^{n} H(Y_i|X_i)$$

$$-E\log q(\vec{Y}|\vec{X}) = -E\log \prod_{i=1}^{n} p(Y_i|X_i) = \sum_{i=1}^{n} -E\log p(Y_i|X_i)$$

Why is C related to $I(X;Y)$?

- Encoder is deterministic: $H(\vec{X}|W) = 0$
- Decoder is deterministic: $H(\hat{W}|\vec{Y}) = 0$
- Since W and $\hat{W}$ are almost identical for reliable communication, assume $H(\hat{W}|W) = H(W|\hat{W}) = 0$
- Then we see that $H(W) = I(\vec{X};\vec{Y})$
- This suggests that the channel capacity is obtained by maximizing $I(X;Y)$. This says that the information conveyed through the channel is essentially the mutual information between the input and output sequence.

Setup for the proof of the converse of channel coding thm:

- $$I(X_i;Y_i) \le C = \max_{p(x)} I(X;Y)$$
- Then $$\sum_{i=1}^{n} I(X_i;Y_i) \le nC$$
- Prove that $$I(\vec{X};\vec{Y}) \le \sum_{i=1}^{n} I(X_i;Y_i)$$
- $$\frac{1}{n}\log M = \frac{1}{n}\log|W| = \frac{1}{n}H(W)$$
$$\approx \frac{1}{n}I(\vec{X};\vec{Y}) \le \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) \le C$$

Lemma 7.16

$$I(\vec{X};\vec{Y}) \leq \sum_{i=1}^{n} I(X_i, Y_i)$$

Proof. From the previous proposition, we have

$$H(\vec{Y}, \vec{X}) = \sum_{i=1}^{n} H(Y_i|X_i)$$

Then

$$\begin{aligned} I(\vec{X};\vec{Y}) &= H(\vec{Y}) - H(\vec{Y}|\vec{X}) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{by independence bound} \\ &= \sum_{i=1}^{n} I(X_i; X_i) \end{aligned}$$

□

Proof of the converse of the Channel coding theorem

Proof. Let R be an achievable rate, i.e. for $\epsilon > 0$, there exists (n, M) code for sufficiently large n such that

$$\frac{1}{n} \log M > R - \epsilon \quad \text{and} \quad \lambda_{max} < \epsilon$$

Consider

$$\log M = H(W)$$

, because we assume each message is chosen uniformly.

$$\begin{aligned} \log M &= H(W) \\ &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq H(W|\hat{W}) + I(\vec{X};\vec{Y}) \quad \textbf{by data processing inequality} \\ &\leq H(W|\hat{W}) + \sum_{i=1}^{n} I(X_i; Y_i) \\ &\leq H(W|\hat{W}) + nC \end{aligned}$$

By Fano's inequality,

$$\begin{aligned} H(W|\hat{W}) &< 1 + P_e(\log |\mathcal{W}|) \\ &= 1 + P_e \log M \end{aligned}$$

Then,

$$\begin{aligned}\log M &\leq H(W|\hat{W}) + nC \\ &< 1 + P_e \log M + nC \\ &\leq 1 + \lambda_{max} \log M + nC \\ &\leq 1 + \epsilon \log M + nC\end{aligned}$$

So

$$(1-\epsilon)\log M < 1 + nC$$

meaning

$$\frac{1}{n}\log M < \frac{\frac{1}{n} + C}{1-\epsilon}$$

Therefore

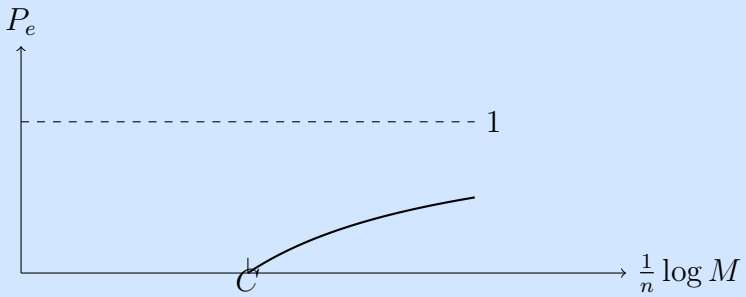
$$R - \epsilon < \frac{\frac{1}{n} + C}{1-\epsilon}$$

Then let $n \to \infty$ and then $\epsilon \to 0$, we get that $R \leq C$ □

A corollary on probability of error

$$P_e \geq 1 - \frac{C}{\frac{1}{n}\log M}$$

$$1 - \frac{C}{\frac{1}{n}\log M}$$



If we take the code rate $\frac{1}{n}\log M$ to be greater than the channel capacity, P_e is bounded away from 0 for large n.

Achievability of the Channel Coding Theorem

To prove achievability:

- Consider a DMC (discrete memoryless channel) $p(y|x)$.
- For every input distribution $p(x)$, prove that the rate $I(X;Y)$ is achievable by showing for large n the existence of a channel code such that
  - the rate of the code is arbitrarily close to $I(X;Y)$
  - the maximal probability of error λ_max is arbitrarily small.
- Choose the input distribution $p(x)$ to be the one that achieves the channel capacity, i.e. $I(X;Y) = c$

Lemma: upper bound of the probability of generation being jointly typical

Let $(\vec{X}', \vec{Y}')$ be n iid copies of a pair of generic random variables (X', Y') and X' and Y' are independent. Then

$$\Pr\{(\vec{X}', \vec{Y}') \in T^n_{XY_\delta}\} \leq 2^{-n(I(X;Y)-\tau)}$$

where $\tau \to 0$ as $\delta \to 0$.

Proof. Consider

$$Pr((\vec{X}', \vec{Y}') \in T^n_{XY_\delta}) = \sum_{\vec{x},\vec{y} \in T} p(\vec{x})p(\vec{y}).$$

By consistentcy of strong typicality, for each $(\vec{x}, \vec{y}) \in T_{XY}$, $\vec{x}$ and $\vec{y}$ are both marginally typical.
By the strong AEP,

$$p(\vec{x}) \leq 2^{-n(H(X)-\eta)}$$

and

$$p(\vec{y}) \leq 2^{-n(H(Y)-\zeta)}$$

where $\eta, \zeta \to 0$ as $\delta \to 0$.
By the strong JAEP,

$$|T^n_{XY_\delta}| \leq 2^{n(H(X,Y)+\xi)}$$

where $\xi \to 0$ as $\delta \to 0$. Then

$$\begin{aligned}
&\Pr\{(\vec{X}', \vec{Y}') \in T^n_{XY_\delta}\} \\
&= \sum_{(\vec{x},\vec{y})\in T} p(\vec{x})p(\vec{y}) \\
&\leq 2^{n(H(X,Y)+\xi)} \times 2^{-n(H(X)-\eta)} \times 2^{-n(H(Y)-\zeta)} \\
&= 2^{-n(I(X;Y)-\tau)}
\end{aligned}$$

where $\tau = \zeta + \eta + \zeta \to 0$ as $\delta \to 0$ □

This lemma seems counter-intuitive because the probability of a generated sequence being jointly typical is supposed to be large, but this makes sense because: we generate the sequence X' iid wrt the distribution $p(x)$ and generate the other sequence Y' iid wrt $p(y)$. The 2 generation processes are independent. However, X and Y are not necessarily independent. The less independent they are, the less likely that $(\vec{X}', \vec{Y}')$ generated in this fasion is typical.

Interpretation of the above lemma

Consider the quasi-uniform array where the rows are typical x sequences and the columns are typical y sequences.

- Randomly choose a row with uniform distribution and randomly choose a column with uniform distribution.
- Then

$$\Pr\{\text{obtaining a jointly typical pair}\} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}$$

Random Coding Scheme Parameter settings

- Fix $\epsilon > 0$ and input distribution $p(x)$. Let δ be specified later.
- Let M be an even integer such that

$$I(X;Y) - \frac{\epsilon}{2} < \frac{1}{n}\log M = \text{rate of the code} < I(X;Y) - \frac{\epsilon}{4}$$

  where n is sufficiently large, i.e. $M \approx 2^{nI(X;Y)}$
- The lower bound guarantees that the rate is close to $I(X;Y)$. The upper bound guarantees that the rate is not too close to $I(X;Y)$.

The Random Coding Scheme:

- 1. Construct the codebook $\mathcal{C}$ of an (n, M) code by generating M codewords in $\mathcal{X}^n$ independently and identically accoding to $p(x)^n$. Denote these codewords by $\vec{X}(1), \vec{X}(2), \ldots, \vec{X}(M)$
- Generate each component accoding to $p(x)$

- There are a total of $|\mathcal{X}|^{Mn}$ possible codebooks that can be constructed.
- Regard two codebooks whose sets of codewords are permutations of each other as two different codebooks.
- 2. Reveal the codebook $\mathcal{C}$ to both the encoder and the decoder.
- 3. A message W is chosen from $\mathcal{W}$ according to the uniform distribution.
- 4. Transmit $\vec{X} = \vec{X}(W)$ through the channel.
- The channel outputs a sequence Y according to

$$\Pr(\vec{Y} = \vec{y}|\vec{X}(W) = \vec{x}) = \prod_{i=1}^{n} p(y_i|x_i).$$

- 6. The sequence $\vec{Y}$ is decoded to the message w if $(\vec{X}(w), \vec{Y}) \in T^n_{XY_\delta}$ and there does not exists $w' \neq w$ such that $(\vec{X}(w'), \vec{Y}) \in T^n_{XY_\delta}$. Otherwise, $\vec{Y}$ is decoded to some constant message in $\mathcal{W}$. Denote the message to which $\vec{Y}$ is decoded to by $\hat{W}$.

10.1 Performance Analysis

- We need to show that $\Pr\{Error\} = \Pr\{\hat{W} \neq W\}$ can be made arbitrarily small.
- Consider

$$\Pr\{Error\} = \sum_{w=1}^{M} \Pr\{Error|W = w\}p(W = w)$$

  This is equal to

$$\Pr\{Error|W = 1\} \sum_{w=1}^{M} \Pr\{W = w\} = \Pr\{Error|W = 1\}$$

  by the construction of the Random Coding Scheme. All codewords are generated randomly in the same fashion.
- Assume without loss of generality that message 1 is chosen.
- For $1 \leq w \leq M$, define the event

$$E_w := \{(\vec{X}(w), \vec{Y}) \in T^n_{XY_\delta}\}$$

- If E_1 occurs but E_w does not occur for all $2 \leq w \leq M$, then no decoding error, because according to our decoding rule, the decoded message $\hat{W}$ will be 1. Therefore

$$\Pr\{Error^C|W = 1\} \geq \Pr\{E_1 \cap E_2^C \cap \cdots \cap E_M^C|M = 1\}$$

- Consider
$$\begin{aligned}\Pr\{Err|W=1\} &= 1-\Pr\{Err^C|W=1\}\\ &\leq 1-\Pr\{E_1\cap E_2^C\cap\cdots\cap E_M^C|W=1\}\\ &= \Pr\{E_1^C\cup E_2\cup\cdots\cup E_M|W=1\}\end{aligned}$$

- By the union bound,
$$\Pr\{Err|W=1\}\leq \Pr\{E_1^C|W=1\}+\sum_{w=2}^{M}\Pr\{E_w|W=1\}$$

- By strong JAEP,
$$\Pr\{E_1^C|W=1\}=\Pr\{(\vec{X}(1),\vec{Y})\notin T^n_{XY_\delta}|W=1\}<\nu$$
where $\nu\to 0$ as $\delta\to 0$

- Conditioning on $\{W=1\}$, for $2\leq w\leq M$, $(\vec{X}(w),\vec{Y})$ are n iid copies of the pair of generic random variables (X',Y') where $X'\sim X$ and $Y'\sim Y$.

- Since DMC is memoryless, X' and Y' are independent because $\vec{X}(1)$ and $\vec{X}(w)$ are independent and the generation of $\vec{Y}$ depends only on $\vec{X}(1)$.

- For $2\leq w\leq M$,
$$\begin{aligned}\Pr\{E_w|W=1\} &= \Pr\{(\vec{X}(w),Y)\in T^n_{XY_\delta}|W=1\}\\ &\leq 2^{-n(I(X;Y)-\tau)}\end{aligned}$$
where $\tau\to 0$ as $\delta\to 0$.

- Note that
$$\frac{1}{n}\log M<I(X;Y)-\frac{\epsilon}{4}\iff M<2^{n(I(X;Y)-\frac{\epsilon}{4}}$$

- Therefore,
$$\begin{aligned}\Pr\{Err\} &< \nu+2^{n(I(X;Y)-\frac{\epsilon}{4})}\cdot 2^{-n(I(X;Y)-\tau)}\\ &= \nu+2^{-n(\frac{\epsilon}{4}-\tau)}\end{aligned}$$

- ϵ is fixed. Since $\tau\to 0$ as $\delta\to 0$, we can choose δ to be sufficiently small so that
$$\frac{\epsilon}{4}-\tau>0$$

- Then
$$2^{-n(\frac{\epsilon}{4}-\tau)}\to 0$$
as $n\to\infty$

- Let $\nu<\frac{\epsilon}{3}$, we get that
$$Pr(Err)<\frac{\epsilon}{2}$$
for sufficiently large n.

Idea of Performance Analysis

- Let n be large.
- $\Pr(\vec{X}(1)\text{jointly typical with}\vec{Y}) \to 1$.
- For $w \neq 1$, $\Pr\{\vec{X}(w)$ jointly typical with $\vec{Y}\} \approx 2^{-nI(X;Y)}$.
- If $|\mathcal{C}| = M$ grows at a rate $< I(X;Y)$, then

$$\Pr\{\vec{X}(w) \text{ jointly typical with } \vec{Y} \text{ for some } w \neq 1\}$$

  can be made arbitrarily small.
- Then $\Pr\{\hat{W} \neq W\}$ can be made arbitrarily small.

Existence of Deterministic Code

- According to the random coding scheme,

$$\Pr\{Err\} = \sum_{C} \Pr(C) \Pr\{Err|C\}$$

- Then there exists at least one codebook C^* such that

$$P_e = \Pr\{Err|C^*\} \leq \Pr\{Err\} < \frac{\epsilon}{2}$$

- By construction, this codebook has rate

$$\frac{1}{n} \log M > I(X;Y) - \frac{\epsilon}{2}$$

  , which can be arranged to be arbitrarily close to the channel capacity.

Code with $\lambda_{max} < \epsilon$

- We want a code with $\lambda_{max} < \epsilon$, not just $P - e < \frac{\epsilon}{2}$.
- Without loss of generality, assume $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M$. (ascending order) Consider

$$\frac{1}{M} \sum_{w=1}^{M} \lambda_w < \frac{\epsilon}{2} iff \sum_{w=1}^{M} \lambda_w < \frac{M}{2}\epsilon$$

  Since M is even, $M/2$ is an integer. Then

$$\sum_{w=M/2+1}^{M} \lambda_w < (\frac{M}{2})\epsilon \iff \frac{1}{M/2} \sum_{M/2+1}^{M} \lambda_w < \epsilon$$

- Conclusion: If $P_e < \frac{\epsilon}{2}$, then $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{M/2} < \epsilon$
- By discarding the worst half of the codewords in C^*, we have that $\lambda_{max} < \epsilon$.
- Then, the rate of the code becomes

$$\frac{1}{n} \log \frac{M}{2} = \frac{1}{n} \log M - \frac{1}{n} > (I(X;Y) - \frac{\epsilon}{2}) - \frac{1}{n} > I(X;Y) - \epsilon$$

  for sufficiently large n.

10.2 Implication of the Channel Coding Theorem

The channel coding theorem says that an indefinitely long message can be communicated reliably through the channel when the block length $n \to \infty$. This is stronger than the requirement that the Bit Error Rate tending to 0, because for long block length, even though BER can be small, the error rate of the whole message can be very large.

- The direct part of the channel coding theorem is an existential proof as opposed to a constructive proof.
- A randomly constructed code has the following issues: encoding and decoding are computationally prohibitive; high storage requirements for encoder and decoder.
- Nevertheless, the direct part implies that when n is large, if the codewords are chosen randonly, most likely the code is good.
- The repetition code is not a good code because the numbers of 0 and 1 in the codewords are not roughly the same.

Illustration of good code

- We have approximately $2^{nI(X;Y)}$ codewords in $T^n_{X_\delta}$
- There are approximately $2^{nH(Y)}$ sequences in $T^n_{Y_\delta}$.
- For each codeword that we choose from the x-typical set, it is jointly typical with approximately $2^{nH(Y|X)}$ y sequences
- This set of y sequences is represented by a cone. We want to pack as many codewords as possible such that these cones do not overlap.
- Therefore, the number of codewords cannot exceed about

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)} = 2^{nC}$$

Construction of codes with efficient encoding and decoding algorithms falls in the domain of channel coding theory. Performance of a code is measured by how far the rate is away from the channel capacity. All channel codes used in practice are linear in terms of computation and storage. Channel coding is widely used in all communications.

10.3 Feedback

Feedback is common in practical communication systems for **correcting** possible errors which occur during transmission. Daily examples include phone call, classroom teaching. For data communication, the receiver may request a packey to be retransmitted if the parity check bits received are incorrect (Automatic Repeat-reQuest). The transmitter can decide what to transmit next based on the feedback on far.

- Question: Can feedback increase the channel capacity?

- Not for DMC, even with complete feedback.

Feedback Code

An (n, M) code with **complete feedback** for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by encoding functions

$$f_i : \{1, 2, \ldots, M\} \times \mathcal{Y}^{i-1} \to \mathcal{X}$$

for $1 \leq i \leq n$ and a decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}.$$

- Notation: $\vec{Y}^i = (Y_1, Y_2, \ldots, Y_i)$, $X_i = f_i(W, \vec{Y}^{i-1})$
- As we encode, we take information about the packets received by the receiver.

Dependency graph:

$$q(w, \vec{x}, \vec{y}, \hat{w}) = q(w)(\prod_{i=1}^{n} q(x_i|w, \vec{y}^{i-1}))(\prod_{i=1}^{n} p(y_i|x_i))q(\hat{w}|\vec{y})$$

Achievable Rate with complete feedback

A rate R is achievable with complete feedback for a discrete memoryless channel $p(y|x)$ if for any $\epsilon > 0$, there exists for sufficiently large n and (n, M) code with complete feedback such that

$$\frac{1}{n} \log M > R - \epsilon$$

and

$$\lambda_{max} < \epsilon.$$

Feedback Capacity

The feedback capacity C_{FB} of a discrete memoryless channel is the supremum over all the rates achievable by codes with complete feedback.
Note that a channel code without feedback is a speicial case of a channel code with complete feedback, so $C_{FB} \geq C$.

Lemma: Markov property of channel with feedback

For all $1 \leq i \leq n$,

$$(W, \vec{Y}^{i-1}) \to X_i \to Y_i$$

forms a Markov chain.

Proof. The Markov chain

$$(W, \vec{X}^{i-1}, \vec{Y}^{i-1}) \to X_i \to Y_i$$

holds because the channel is memoryless.

Then

$$0 = I(W, \vec{X}^{i-1}, \vec{Y}^{i-1}; Y_i, X_i)$$
$$= I(W, \vec{Y}^{i-1}; Y_i|X_i) + I(\vec{X}^{i-1}; Y_i|W, X_i, \vec{Y}^{i-1})$$

By non-negative of Shannon's information measure, $I(W, \vec{Y}^{i-1}; Y_i|X_i) = 0$, proving the Markov property. □

$C_{FB} \leq C$ for DMC

Consider any code with complete feedback.
Consider

$$\log M = H(W) = I(W; \vec{Y}) + H(W|\vec{Y})$$

$$\begin{aligned}
I(W; \vec{Y}) &= H(\vec{Y}) - H(\vec{Y}|W) \\
&= H(\vec{Y}) - \sum_{i=1}^{n} H(Y_i|\vec{Y}^{i-1}, W) \quad \text{by the chain rule} \\
&= H(\vec{Y}) - \sum_{i=1}^{n} H(Y_i|\vec{Y}^{i-1}, W, X_i), \quad X_i \text{ does not provide new information} \\
&= H(\vec{Y}) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{by Markov property} \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{by independence bound} \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\leq nC
\end{aligned}$$

$$\begin{aligned}
H(W|\vec{Y}) &= H(W|\vec{Y}, \hat{W}) \quad \text{since } \hat{W} \text{ is a function of } \vec{Y} \\
&\leq H(W|\hat{W}) \approx 0
\end{aligned}$$

Then

$$\log M \leq nC$$

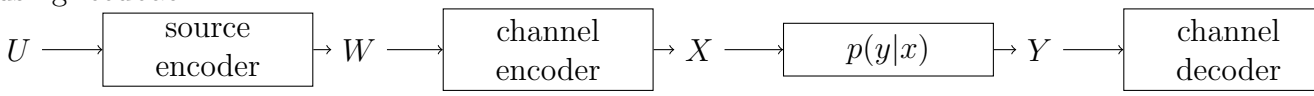
Rigorously, we can apply Fano's inequality to upper bound $H(W|\hat{W})$.
Lastly, we conclude that $R \leq C$ for any rate R achievable with complete feedback.

Even though feedback does not increase the capacity of a DMC, the availability of feedback can make coding simpler.
Also if the channel has memory, feedback can increase the capacity.

10.4 Separation of Source and Channel Coding

- Consider transmitting an information source with entropy rate H reliably through a DMC with capacity C.

- If $H < C$, this can be achieved by separating source and channel coding without using feedback.



- Choose R_s and R_c such that

$$H < R_s < R_c < C$$

- it can be shown that even with complete feedback, reliable communication is impossible if $H > C$
- The source code and the channel code can be designed separately without losing asymptotic optimality
- The source coding removes redundancy while the channel coding adds redundancy.

Rate Distortion Theory

Now we discuss information transmission with distortion.

- Consider compressing an information source with entropy rate H at rate $R < H$ (bad)
- By the **source coding theorem**, $P_e \to 1$ as the block length $n \to \infty$.
- Under this this constraint, information must be transmitted with distortion.
- What is the best tradeoff then?

Objectives

- Single-letter distortion measures
- The rate-distortion function $R(D)$.
- The rate-distortion theorem for an iid information source.

11.1 Single-Letter Distortion Measures

The setup is as follows:

- Let $\{X_k, k \geq 1\}$ be an iid information source with generic random variable $X \sim p(x)$ where $|\mathcal{X}| < \infty$ is finite.
- Consider a source sequence $\vec{x} = (x_1, x_2, \ldots, x_n)$ and a reproduction sequence $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$.
- The components of $\vec{\hat{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$.
- The component of $\vec{\hat{x}}$ take values in **reproduction alphabet** $\hat{\mathcal{X}}$, where $|\hat{\mathcal{X}}| < \infty$.
- In general, $\hat{\mathcal{X}}$ maybe different from $\mathcal{X}$.
- For example, $\vec{\hat{x}}$ can be a quantized version of $\vec{x}$.

Here comes the definition.

Single-letter distortion measure

A single-letter distortion measure is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+.$$

The value $d(x, \hat{x})$ denotes the distortion incurred when a source symbol x is reproduced as $\hat{x}$. This is intuitive.

Average distortion / distortion between sequences

The average distortion between a source sequence $\vec{x} \in \mathcal{X}^n$ and a reproduction sequence $\hat{\vec{x}} \in \hat{\mathcal{X}}^n$ induced by a single-letter distortion measure d is defined as

$$d(\vec{x}, \hat{\vec{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_k, \hat{x}_k)$$

Examples of a Distortion Measure

Let $\hat{X} = X$.

- **Square-error**:
$$d(x, \hat{x} = (x - \hat{x})^2$$
where the source and reproduction RVs are real-valued.
- **Hamming distortion**:
$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases}.$$

Let $\hat{X}$ be an estimate of X,

- If d is the square-error distortion measure, $\mathbb{E}d(X, \hat{X})$ is called the mean square error.
- d is the Hamming distortion measure, then
$$\mathbb{E}d(X, \hat{X}) = \Pr(X \neq \hat{X}),$$
which is the probability of error. For a source sequence $\vec{x}$ and a reproduction sequence $\hat{\vec{x}}$, the average distortion $d(\vec{x}, \hat{\vec{x}})$ gives the frequency of error in $\hat{\vec{x}}$.

The distortion measure can be normalized.

Normalization of a Distortion Measure

For a distortion measure d, for each $x \in X$, let $\hat{x}^*(x) \in \hat{\mathcal{X}}$ minimize $d(x, \hat{x})$ over all

$\hat{x} \in \hat{\mathcal{X}}$. A distortion measure is said to be nomral if

$$c_x := d(x, \hat{x}^*(x)) = 0$$

for all $x \in \mathcal{X}$.

Remark 11.1. This says that for normal distortion measure, for each source symbol x, there exists some reproduction symbol $\hat{x}$ such that the distortion between them is 0.

Computation of $\hat{x}^*$

Let d be a distortion measure defined by

$d(x, \hat{x})$	**a**	**b**	c
1	2	7	5
2	4	3	8

Then $\hat{x}^*(1) = a$ and $\hat{x}^*(2) = b$.

- A normal distortion measure is one which allows a source X to be reproduced with zero distortion.
- The square-error distortion measure and the Hamming distortion measure are normal distortion measures.
- **To normalize a distortion measure** d

  $$\tilde{d} := d(x, \hat{x}) - c_x$$

  for all $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$
- It suffices to only consider normal distortion measure as we will see.

Example of Normalization

Consider the example above again

$d(x, \hat{x})$	a	b	c
1	2	7	5
2	4	3	8

$$c_1 = 2, c_2 = 3$$

Then to normalize the measure d, we subtract 2 from the first row and subtract 3 from the second row to get $\tilde{d}$:

$\tilde{d}(x, \hat{x})$	a	b	c
1	0	5	3
2	1	0	5

Let $\hat{X}$ ne any estimate of X which takes values in $\hat{\mathcal{X}}$. Then

$$\begin{aligned}
\mathbb{E}d(X,\hat{X}) &= \sum_x \sum_{\hat{x}} p(x,\hat{x}) d(x,\hat{x}) \\
&= \sum_x \sum_{\hat{x}} p(x,\hat{x}) \tilde{d}(x,\hat{x}) + c_x \\
&= \mathbb{E}\tilde{d}(X,\hat{X}) + \sum_x \sum_{\hat{x}} p(x) p(\hat{x}|x) c_x \\
&= \mathbb{E}\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \sum_{\hat{x}} p(\hat{x}|x) \\
&= \mathbb{E}\tilde{d}(X,\hat{X}) + \sum_x p(x) c_x \\
&= \mathbb{E}\tilde{d}(X,\hat{X}) + \Delta
\end{aligned}$$

where $\Delta := \sum_x p(x) c_x$ is a constant which depends only on $p(x)$ and d but not on the conditional distribution $p(\hat{x}|x)$

Distortion minimizing point

Let $\hat{x}^*$ minimizes $\mathbb{E}d(X,\hat{x})$ over all $\hat{x} \in \hat{\mathcal{X}}$, and define

$$D_{max} = \mathbb{E}d(X,\hat{x}^*)$$

Note: $\hat{x}^*$ is not the same as $\hat{x}^*(x)$, and it depends on $p(x)$.

- If we know nothing about a source variable X except for $p(x)$, then $\hat{x}^*$ is the best estimate of X, and D_{max} is the minimum expected distortion between X and a constant estimate of X
- It's confusing that D_max is called "max", because it should be minimum over $\mathbb{E}d(X,\hat{x})$. We'll see how this makes sense.
- Specifically, by taking $\vec{\hat{x}}^* = (\hat{x}^*, \ldots, \hat{x}^*)$ to be the reproduction sequence, D_{max} can be asymptotically achived, because by WLLN,

$$d(\vec{X}, \vec{\hat{x}}^*) = \frac{1}{n} \sum_{k=1} d(X_k, \hat{x}^*) \to \mathbb{E}d(X,\hat{x}^*) = D_{max}$$

- We don't need to consider $D \geq D_{max}$ for the reproduction sequence. This is also why D_{max} is called "D-max" because it's the maximum distortion we would ever care about.

11.2 The Rate Distortion Function

We first consider the definition for rate-distortion code. The setup is as follows:

- All the discussions are with respect to an i.i.d. information source $\{X_k, k \geq 1\}$ with generic random variable X and a distortion measure d.

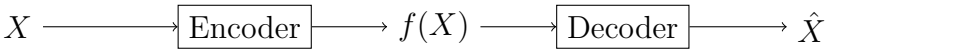
Rate-Distortion Code

An (n, M) rate-distortion code is defined by an encoding function

$$f : \mathcal{X}^n \to \{1, 2, \ldots, M\}$$

and a decoding function

$$g : \{1, 2, \ldots, M\} \to \hat{X}^n$$

- Index set: $\mathcal{I} = \{1, 2, \ldots, M\}$.
- Codewords: the reproduction sequence $g(1), g(2), \ldots, g(M)$.
- Codebook: the set of all codewords.
- 

source sequence reproduction sequence

Rate of an (n M) rate-distortion code

The **rate** of an (n, M) rate-distortion code is

$$\frac{1}{n} \log M$$

in bits per symbol.

Achievable rate-distortion pair

A rate-distortion pair (R, D) is asymptotically achievable if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) rate-distortion code such that

$$\frac{1}{n} \log M \leq R + \epsilon$$

and

$$\Pr\{d(\vec{X}, \vec{\hat{X}}) > D + \epsilon\} \leq \epsilon$$

where $\vec{\hat{X}} = g(f(\vec{X}))$.

Remark 11.2. If (R, D) is achievable, then any (R', D) and (R, D') are also achievable for $R' \geq R$ and $D' \geq D$. i.e. (R', D') are achievable for all $R' \geq R$ and $D' \geq D$.

Rate-distortion region

A rate-distortion region is the subset of $\mathbb{R}^2$ containing all achievable pairs (R, D).

Closedness and Convexity of rate-distortion region

The rate-distortion region is closed and convex.

Proof. The closedness follows from the definition of achievability since we had non-strict inequalities. The convexity is proved by a technique called *time-sharing*. Time sharing between two codes: one code achieves $(R^{(1)}, D^{(1)})$ for λ fraction of the time, and the other code achieves (R^2, D^2) for $(1-\lambda)$ fraction of the time. □

Rate-Distortion Function

The rate-distortion function $R(D)$ is the minimum of all rates R for a given distortion D such that (R, D) is achievable.
i.e. we fix D and minimize R.

Distortion-Rate Function

The distortion-rate function $D(R)$ is the minimum of all distortions D for a given rate R such that (R, D) is achievable.
i.e. we fix R and minimize D.

Remark 11.3.
- Most of the time, we will be using $R(D)$ instead of $D(R)$.
- If (R, D) is achievable, then $R \geq R(D)$ by definition.

Properties of rate-distortion function $R(D)$

- $R(D)$ is non-increasing in D.
- $R(D)$ is convex.
- $R(D) = 0$ for $D \geq D_{max}$
- $R(0) \leq H(X)$

Proof.
- Let $D' \geq D$, $(R(D), D)$ achievable implies that $R(D), D'$ is also achievable. Then $R(D) \geq R(D')$ by definition of $R(D')$.
- Follows from the convexity of the rate-distortion region.
- $(0, D_{max}$ is achievable implies $R(D_{max}) = 0$ because $R(D_{max})$ is non-negative. Then $R(D) = 0$ for $D \geq D_{max}$ because $R(\cdot)$ is non-increasing.
  Note that $R(D_{max}) = 0$ because there exists a rate-distortion code which has only one codeword - $\hat{x}^*$, which achieves D_{max} asymptotically.
- $(H(X), 0)$ is achievable by the source coding theorem. The expected distortion can be made arbitrarily small, since d is normal. Therefore, $R(0) \leq H(X)$ by definition of $R(0)$.
  Rigorously, $(H(X), 0)$ is achievable because by using a rate no more than $H(X) + \epsilon$, we can describe the source sequence with error probability less than ϵ by the source coding theorem. Then let the code $\hat{X}_k = \hat{x}^*(X_k)$, so that whenever an error does not occur, $d(\vec{X}, \hat{\vec{X}}) = 0$.

□

11.3 The Rate Distortion Theorem

Information Rate-Distortion Function

For $D \geq 0$, the information rate-distortion function is defined by

$$R_I(D) = \min_{\hat{X}:\mathbb{E}d(X,\hat{X})\leq D} I(X;\hat{X}).$$

- The minimization is taken over the set of all transition matrices $p(\hat{x}|x)$ such that $\mathbb{E}d(X,\hat{X}) \leq D$:

$$\{p(\hat{x}|x) : \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D\}$$

- Since this is a topologically compact set (closed and bounded) in $\mathbb{R}^{|\mathcal{X}||\hat{\mathcal{X}}|}$, and $I(X,\hat{X})$ is a continuous functional of the conditional distribution $p(\hat{x}|x)$, the minimum value of $I(X,\hat{X})$ can be attained, so R_I is well defined.
- Equivalently, the minimization can be taken over the set of all joint distributions $p(x,\hat{x})$ with marginal distribution $p(x)$, the given source distribution, such that $\mathbb{E}d(X,\hat{X}) \leq D$.
- Since
$$\mathbb{E}\tilde{d}(X,\hat{X}) = \mathbb{E}d(X,\hat{X}) - \Delta$$
, where Δ does not depend on $p(\hat{x},x)$, we can replace d by $\tilde{d}$ and D by $D-\Delta$ in the definition of $R_I(D)$ without changing the minimization problem.
- Without loss of generality, we can assume d is normal.

The following theorem is the most important theorem of this chapter.

The Rate-Distortion Theorem

$$R(D) = R_I(D)$$

i.e., the rate-distortion function is actually equal to the information-rate-distortion function.

Properties of the information rate-distortion function $R_I(D)$

1. $R_I(D)$ is non-increasing in D.
2. $R_I(D)$ is convex in D.
3. $R_I(D) = 0$ for $D \geq D_{max}$
4. $R_I(0) \leq H(X)$

Proof. 1. The minimization problem is now over a larger set for larger D.

2. Consider any $D^1, D^2 \geq 0$, and $0 \leq \lambda \leq 1$. Let $\hat{X}^i$ achieves $R_I(D^i)$, i.e.,

$$R_I(D^i) = I(X; \hat{X}^i)$$

where $\mathbb{E}d(X, \hat{X}^i) \leq D^i$.
Let $\hat{X}^\lambda$ be jointly distributed with X defined by

$$p_\lambda(\hat{x}|x) = \lambda p_1(\hat{x}|x) + (1-\lambda)p_2(\hat{x}|x).$$

Then

$$\mathbb{E}d(X, \hat{X}^\lambda) = \lambda \mathbb{E}d(X, \hat{X}^1) + (1-\lambda)\mathbb{E}d(X, \hat{X}^2)$$

by the linearity of expectation. Then we let

$$D^\lambda := \lambda D^1 + (1-\lambda)D^2$$

Finally consider

$$\lambda R_I(D^1) + (1-\lambda)R_i(D^2) = \lambda I(X; \hat{X}^1) + I(X; \hat{X}^2) \geq I(X; \hat{X}^\lambda)$$

because the mutual information $I(X;Y)$ is a convex functional of $p(y|x)$ for fixed $p(x)$.
Lastly, $I(X; \hat{X}^\lambda) \geq R_I(D^\lambda)$ by definition of R_I.

3. Let $\hat{X} = \hat{x}^*$ w.p. 1 to show that $(0, D_{max})$ is achievable. Then for $R_I(D) \leq I(X; \hat{X}) = 0$ (mutual information with a point mass).

4. Let $\hat{X} = \hat{x}^*(X)$, so that $\mathbb{E}d(X, \hat{X}) = 0$ since d is normal. Then

$$R_I(0) \leq I(X; \hat{X}) \leq H(X)$$

because mutual information is always less than self-information by information measure diagram.

□

A corollary: stronger properties

If $R_1(0) > 0$, then $R_I(D)$ is strictly decreasing for $0 \leq D \leq D_{max}$, and the inequality constraint in the definition of $R_I(D)$ can be replaced by an equality constraint.

Proof. Assume by contradiction that $R_I(D') = 0$ for some $0 \leq D' < D_{max}$, and let $R_I(D')$ be achieved by some $\hat{X}$. Then

$$R_I(D') = I(X; \hat{X}) = 0,$$

meaning X and $\hat{X} = 0$.
Then note that the estimate $\hat{X}$ of X does not do a better job than the constant estimate $\hat{x}^*$ (which minimizes $\mathbb{E}d(X, \hat{x})$ over $\hat{x}$). Consider

$$D' \geq \mathbb{E}d(X, \hat{X}) = \sum_x \sum_{\hat{x}} p(x, \hat{x}) d(x, \hat{x})$$

By independence, we have

$$\sum_x \sum_{\hat{x}} p(x,\hat{x})d(x,\hat{x}) = \sum_x \sum_{\hat{x}} p(x)p(\hat{x})d(x,\hat{x})$$

$$= \sum_{\hat{x}} p(\hat{x}) \sum_x p(x)d(x,\hat{x})$$

$$= \sum_{\hat{x}} p(\hat{x})\mathbb{E}d(X,\hat{x})$$

$$\geq \sum_{\hat{x}} p(\hat{x})\mathbb{E}d(X,\hat{x}^*) = D_{max}$$

This is contradiction because we assumed $D' < D_{max}$.
Next we show that $R_I(D)$ must be strictly decreasing for $0 \leq D \leq D_{max}$ because $R_I(0) > 0$, $R_I(D_{max}) = 0$, and $R_(D)$ is non-increasing and convex. Note that non-strict decreasing messes up with the convexity. (It's helpful draw a curve to visualize.)
Next, show that the inequality constraints in $R_I(D)$ can be replaced by an equality constraint by contradiction. Assume by contradiction that $R_I(D)$ is achieved by some $\hat{X}^*$ such that $\mathbb{E}d(X,\hat{X}^*) = D'' < D$. Then

$$R_I(D'') = \min_{\hat{X}:\mathbb{E}d(X,\hat{X})\leq D''} I(X;\hat{X}) \leq I(X;\hat{X}^*) = R_I(D),$$

where the above inequality is true because $I(X;\hat{X}^*)$ is the minimum over a larger set. This is a contradiction because $R_I(D)$ is strictly decreasing for $0 \leq D \leq D_{max}$. Therefore, $\mathbb{E}d(X,\hat{X}^*)$ □

Remark 11.4. $R_I(0) > 0$ is not a very strong assumption. In all problems of interest,

$$R(0) = R_I(0) > 0$$

because otherwise, $R(D) = 0$ for all $D \leq 0$ because $R(D)$ is non-negative and non-increasing. Therefore,

$$R_I(D) = \min_{\hat{X}:\mathbb{E}d(X,\hat{X})=D} I(X;\hat{X})$$

(Simplest Example) Binary Information Source

Let X be a binary random variable with

$$\Pr\{X = 0\} = 1 - \gamma$$

and

$$\Pr\{X = 1\} = \gamma.$$

Let $\hat{\mathcal{X}}$ be $\{0,1\} = \mathcal{X}$ and d be the Hamming distortion measure.
$R_I(D)$:
First, consider $0 \leq \gamma \leq \frac{1}{2}$. We will show that

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } 0 \leq D < \gamma \\ 0 & \text{if } D \geq \gamma \end{cases}$$

Proof. 1. Since $\gamma \leq \frac{1}{2}$, $\hat{x}^* = 0$ and $D_{max} = \mathbb{E}d(X,0) = \Pr(X=1) = \gamma$.

2. Consider any $\hat{X}$ and let $Y = d(X, \hat{X})$.

3. $H(X|\hat{X}) = H(Y|\hat{X})$. because given $\hat{X}$, X and Y can determine each other. Then for any $\hat{X}$, such that $\mathbb{E}d(X,\hat{X}) \leq D$, where $D < \gamma = D_{max}$,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= h_b(\gamma) - H(Y|\hat{X}) \\ &\geq h_b(\gamma) - H(Y) \\ &= h_b(\gamma) - h_b(\Pr(X \neq \hat{X})) \\ &\geq h_b(\gamma) - h_b(D), \end{aligned}$$

because $\Pr\{X \neq \hat{X}\} = \mathbb{E}d(X,\hat{X}) \leq D$ and $h_b(a)$ is increasing for $0 \leq a \leq \frac{1}{2}$. Therefore,

$$R_I(D) = \min_{\hat{X}:\mathbb{E}d(X,\hat{X})\leq D} I(X;\hat{X}) \geq h_b(\gamma) - h_n(D).$$

Now we need to construct $\hat{X}$ which gives tightness for the above inequalities so that the above bound can be achieved. We need:

- Y independent of $\hat{X}$
- $P(X \neq \hat{X}) = D$

The required $\hat{X}$ can be constructed by a reverse BSC, i.e. with probability D, $\hat{X} \neq X$ (Here, D is the cross-over probability). This satisfies the above 2 conditions.
Then let $P(\hat{X}=0) = \frac{1-\gamma-D}{1-2D}$, and $P(\hat{X}=1) = \frac{\gamma-D}{1-2D}$.
We can verify that:

$$P(X=1) = \frac{1-\gamma-D}{1-2D} \times D + \frac{\gamma-D}{1-2D} \times (1-D) = \gamma$$

4. Therefore, we see that for $0 \leq \gamma \leq \frac{1}{2}$,

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } 0 \leq D < \gamma \\ 0 & \text{if } D \geq \gamma = D_{max} \end{cases}$$

5. Finally, do the same thing to the case where $\frac{1}{2} \leq \gamma \leq 1$ to get

$$R_I(D) = \begin{cases} h_b(\gamma) - h_b(D) & \text{if } 0 \leq D < \min(\gamma, 1-\gamma) \\ 0 & \text{if } D \geq \min(\gamma, 1-\gamma) \end{cases}$$

□

Remark 11.5.
- $R_I(D)$ is the minimum possible mutual information between X and its estimate when the single-letter error probability is D.
- By the rate-distortion theorem, $R_I(D)$ is also the minimum achievable rate when a single-letter error probability D can be tolerated.

Remark 11.6. The source coding theorem is **NOT** a special case of the rate-distortion theorem.
In the above example, $R_I(0) = h_b(\gamma) = H(X)$. By the rate-distortion theorem, if $R > H(X)$, the average Hamming distortion (i.e. the error probability per symbol), can be made arbitrarily small.
However, by the source coding theorem, if $R > H(X)$, the message error probability can be made arbitrarily small, which is a much stronger statement, because the message error probability is small implies that the error probability per symbol is small.

11.4 The "Converse" of the Rate-Distortion Theorem

In this section, we provide a proof for the "converse" of the rate-distortion theorem.

"Converse" of the Rate-Distortion Theorem

For any achievable rate-distortion pair (R, D), $R \geq R_I(D)$

Proof. Let (R, D) be any achievable rate-distortion pair, i.e., for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\text{rate of the code} = \frac{1}{n} \log M \leq R + \epsilon$$

and

$$\Pr\{d(\vec{X}, \vec{\hat{X}}) > D + \epsilon\} \leq \epsilon$$

where $\vec{\hat{X}} = g(f(\vec{X}))$ Then

$$\begin{aligned}
n(R+\epsilon) &\geq \log M \\
&\geq H(f(\vec{X})) \\
&\geq H(g(f(\vec{X}))) \\
&= H(\vec{\hat{X}}) \\
&= H(\vec{\hat{X}}) - H(\vec{\hat{X}}|\vec{X}) \quad \text{because } \vec{\hat{X}} = g(f(\vec{X})) \\
&= I(\vec{\hat{X}}; \vec{X}) \\
&= H(\vec{X}) - H(\vec{X}|\vec{\hat{X}}) \\
&= \sum_{i=1}^{n} H(X_k) - \sum_{k=1}^{n} H(X_k|\vec{\hat{X}}, X_1, X_2, \ldots, X_{k-1}) \\
&\text{because the source is iid, chain rule} \\
&\geq \sum_{k=1}^{n} H(X_k) - \sum_{k=1}^{n} H(X_k|\hat{X}_k) \\
&= \sum_{k=1}^{n} I(X_k, \hat{X}_k) \\
&\geq \sum_{k=1}^{n} R_I(\mathbb{E}d(X_k, \hat{X}_k)) \\
&= n(\frac{1}{n}\sum_{k=1}^{n} R_I(\mathbb{E}d(X_k, \hat{X}_k))) \\
&\geq nR_I(\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}d(X_k, \hat{X}_k)) \quad \text{by convexity} \\
&= nR_I(\mathbb{E}d(\vec{X}, \vec{\hat{X}}))
\end{aligned}$$

Now let $d_{\max} = \max_{x,\hat{x}} d(x, \hat{x})$. Then

$$\begin{aligned}
&\mathbb{E}d(\vec{X},\ vec\hat{X}) \\
&= \mathbb{E}[d(\vec{X}, \vec{\hat{X}})|d(\vec{X}, \vec{\hat{X}}) > D + \epsilon] \Pr(d(\vec{X}, \vec{\hat{X}}) > D + \epsilon) \\
&+ \mathbb{E}[d(\vec{X}, \vec{\hat{X}})|d(\vec{X}, \vec{\hat{X}}) \leq D + \epsilon] \Pr(d(\vec{X}, \vec{\hat{X}}) \leq D + \epsilon) \\
&\leq d_{max} \times \epsilon + (D + \epsilon) \times 1 \\
&\to D \quad \text{as } \epsilon \to 0
\end{aligned}$$

Therefore

$$\begin{aligned}
R + \epsilon &\geq R_I(\mathbb{E}d(\vec{X}, \vec{\hat{X}})) \\
&\geq R_I(D + (d_{max} + 1)\epsilon)
\end{aligned}$$

because $R_I(D)$ is non-increasing in D.

Then by convexity of $R_I(D)$, it is continuous in D, so by letting $\epsilon \to 0$, we have

$$\begin{aligned} R &\geq \lim_{\epsilon \to 0} R_I(D + (d_{max} + 1)\epsilon) \\ &= R_I(\lim_{\epsilon \to 0} D + (d_{max} + 1)\epsilon) \quad \text{by continuity} \\ &= R_I(D) \end{aligned}$$

□

11.5 Achievability of Information Rate-Distortion $R_I(D)$

Recall:

The Rate-Distoriton Theorem

$$R(D) = R_I(D)$$

How to prove achievability:

- An iid source $\{X_k\}$ with generic random variable $X \sim p(x)$ is given.
- For every random variable $\hat{X}$ taking values in $\hat{\mathcal{X}}$ with $\mathbb{E}d(X, \hat{X}) \leq D$, where $0 \leq D \leq D_{max}$, prove that the rate-distortion pair $(I(X; \hat{X}), D)$ is achievable by showing for large n, there exists a rate-distortion code such that
  1. The rate of the code is not more than $I(X; \hat{X}) + \epsilon$
  2. $d(X, \mathcal{X}) \leq D + \epsilon$ with probability close to 1.
- Then minimize $I(X; \hat{X})$ over all such $\hat{X}$ to conclude that $(R_I(D), D)$ is achievable.
- This implies that $R_I(D) \geq R(D)$

We use the random coding scheme technique like before, which uses the properties of joint typicality. Random Coding Scheme Parameter Settings

1. Fix $\epsilon > 0$ and $\hat{X}$ with $\mathbb{E}(X, \hat{X}) \leq D$, where $0 \leq D \leq D_{max}$. Let δ be specified later.
2. Let M be an integer satisfying

$$I(X; \hat{X}) + \frac{\epsilon}{2} \leq \frac{1}{n} \log M \leq I(X; \hat{X}) + \epsilon$$

where n is sufficiently large for such M to exist.

The Random Coding Scheme

1. Construct a codebook $\mathcal{C}$ of an (n, M) code by randomly generating M codewords in $\mathcal{X}^n$ independently and identically according to $p(\hat{x})^n$. Denote these codewords by $\hat{X}(1), \ldots, \hat{X}(M)$

2. Reveal the codebook $\mathcal{C}$ to both the encoder and the decoder.
3. The source sequence $\vec{X}$ is generated according to $p(x)^n$.
4. The encoder encodes the source sequence $\vec{X}$ into an index K in the set $I = \{1, 2, \ldots, M\}$. The index K takes the value i if
   (a) $(\vec{X}, \vec{\hat{X}}(i)) \in T^n_{[X\hat{X}]_\delta}$
   (b) for all $i' \in I$, if $(\vec{X}, \vec{\hat{X}}(i')) \in T^n_{[X\hat{X}]_\delta}$, then $i' \le i$.
   That is, if there exists more than one i satisfying the first condition, let K be the largest one. Otherwise K takes the constant value 1. (Give up encoding it.)
5. The index K is delivered to the decoder.
6. The decoder outputs $\vec{\hat{X}}(K)$ as the reproduction sequence.

Remark 11.7.
- The event $\{K = 1\}$ occurs in one of the following two scenarios:
  1. $\hat{X}(1)$ is the only codeword in $\mathcal{C}$ which is jointly typical with $\vec{X}$.
  2. No codeword in $\mathcal{C}$ is jointly typical with $\vec{X}$.
- If $K \neq 1$, then $\vec{\hat{X}}(K)$ is always jointly typical with $\vec{X}$.

Performance Analysis

1. As remarked above, the event $\{K = 1\}$ occurs in one of the following two scenarios:
   (a) $\hat{X}(1)$ is the only codeword in $\mathcal{C}$ which is jointly typical with $\vec{X}$.
   (b) No codeword in $\mathcal{C}$ is jointly typical with $\vec{X}$.

   So if $K = 1$, then $\vec{X}$ is jointly typical with none of the codewords $\vec{X}(2), \vec{X}(3), \ldots, \vec{X}(M)$. We will show that $\Pr\{K = 1\}$ can be made arbitrarily small.
2. Define the event
$$E_i = \{(\vec{X}, \vec{\hat{X}}(i) \in T^n_{[X\hat{X}]_\delta})\}$$
3. Then
$$\{K = 1\} \subset E_2^C \cap E_3^C \cap \cdots \cap E_M^C$$
4. Since the codewords are generated iid, conditioning on $\{\vec{X} = x\}$ for any $\vec{x} \in \mathcal{X}^n$, the events E_i are mutually independent and have the same probability.
5. Then for any $\vec{x} \in \mathcal{X}^n$,
$$\begin{aligned}\Pr\{K = 1|\vec{X} = \vec{x}\} &\le \Pr\{E_2^C \cap E_3^C \cap \cdots \cap E_M^C|\vec{X} = \vec{x}\} \\ &= \prod_{i=2}^{M} \Pr\{E_i^C|\vec{X} = \vec{x}\} \\ &= (1 - \Pr\{E_1|\vec{X} = \vec{x}\})^{M-1}\end{aligned}$$

6. We will focus on $\vec{x} \in S^n_{[X]_\delta}$ where

$$S^n_{[X]_\delta} = \{\vec{x} \in T^n_{[X]_\delta} : |T^n_{[\hat{X}|X]_\delta}| \geq 1\}$$

because the probability $\Pr\{\vec{X} \in S^n_{[X]_\delta}\}$ is close to 1 for large n.

7. For $\vec{x} \in S^n_{[X]_\delta}$, obtain the following lower bound on $\Pr\{E_1|\vec{X} = \vec{x}\}$:

$$\begin{aligned} P(E_1|\vec{X} = \vec{x}) &= P\{(\vec{x}, \vec{\hat{X}}(1)) \in T^n_{X\hat{X}_\delta}\} \\ &= \sum_{\hat{x} \in T^n_{[\hat{X}|X]_\delta}} p(\hat{x}) \\ &\geq \sum_{\hat{x} \in T^n_{[\hat{X}|X]_\delta}} 2^{-n(H(\hat{X})+\eta)} \\ &\geq 2^{n(H(\vec{X}|X)-\xi)} \times 2^{-n(H(\hat{X})+\eta)} \\ &= 2^{-nI(X;\hat{X})+\zeta} \end{aligned}$$

where $\zeta = \xi + \eta \to 0$ as $n \to \infty$ and $\delta \to 0$

8. Therefore,

$$\begin{aligned} &\Pr\{K = 1|\vec{X} = \vec{x}\} \\ &\leq (2^{-nI(X;\hat{X})+\zeta})^{M-1} \end{aligned}$$

9. Now note that

$$\frac{1}{n} \log M \geq I(X;\hat{X}) + \frac{\epsilon}{2} \iff M \geq 2^{n(I(X;\hat{X})+\epsilon/2)}$$

Then

$$\begin{aligned} &\ln \Pr\{K = 1|\vec{X} = \vec{x}\} \\ &\leq (M-1) \ln[1 - 2^{-n(I(X;\hat{X})+\zeta)}] \\ &\leq -(2^{n(I(X;\hat{X})+\frac{\epsilon}{2})} - 1)2^{-n(I(X;\hat{X})+\zeta)} \quad \text{by the fundamental inequality} \\ &\to -\infty \quad \text{as } n \to \infty \end{aligned}$$

if we let δ be sufficiently small so that

$$\frac{\epsilon}{2} - \zeta > 0.$$

Therefore

$$\Pr\{K = 1|\vec{X} = \vec{x}\} \to 0$$

as $n \to \infty$.

10. Then for $\vec{x} \in S$, for sufficiently large n,

$$\Pr\{K = 1|\vec{X} = \vec{x}\} \leq \frac{\epsilon}{2}$$

11. Lastly,

$$\begin{aligned}
&\Pr\{K=1\} \\
&= \sum_{\vec{x}\in S} P(K=1|\vec{X}=\vec{x})\Pr(\vec{X}=\vec{x}) + \sum_{x\notin S} \Pr(K=1|\vec{X}=\vec{x})\Pr(\vec{X}=\vec{x}) \\
&\le \sum_{\vec{x}\in S} \frac{\epsilon}{2}\Pr(\vec{X}=\vec{x}) + \sum_{x\notin S} 1\times\Pr(\vec{X}=x) \\
&\le \frac{\epsilon}{2}\times 1+\delta
\end{aligned}$$

12. Arrange that $\Pr(K=1)<\epsilon$.

Proof Outline

1. Randomly generate M codewords in $\hat{\mathcal{X}}^n$ according to the product measure $p(\hat{x})^n$, where n is large.
2. $\vec{X}\in S^n_{[X]_\delta}$ with high probability.
3. For $\vec{x}\in S^n_{[X]_\delta}$, by conditional strong AEP,

$$\Pr\{(\vec{X},\vec{\hat{X}}(i))\in T^n_{[X\hat{X}]_\delta}|\vec{X}=\vec{x}\}\approx 2^{-nI(X;\hat{X})}.$$

4. If M grows with n at a rate higher than $I(X;\hat{X})$, then the probability that there exists at least one $\vec{\hat{X}}(i)$ which is jointly typical with the source sequence $\vec{X}$ with respect to $p(x,\hat{x})$ is high.
5. Such an $\vec{\hat{X}}(i)$, if exists, would have $d(\vec{X},\vec{\hat{X}})\approx \mathbb{E}d(X,\hat{X})\le D$, because the joint relative frequency of $(\vec{x},\vec{\hat{X}}(i))\approx p(x,\hat{x})$. i.e. if the source sequence and a codeword are jointly typical, then the distortion of the codeword is approximately equal to the distortion between X and $\hat{X}$. See the proposition below.
6. Lastly, use this $\vec{\hat{X}}(i)$ to represent $\vec{X}$ to satisfy the distortion constraint.

Proposition

For $\hat{X}$ such that $\mathbb{E}d(X,\hat{X})\le D$, if $(\vec{x},\vec{\hat{x}})\in T^n_{X\hat{X}}\delta$, then

$$d(\vec{x},\vec{\hat{x}})\le D+d_{max}\delta$$

Proof. For $(\vec{x},\vec{\hat{x}})$ that are jointly typical, consider

$$d(\vec{x},\vec{\hat{x}})=\frac{1}{n}\sum_{k=1}^{n} d(x_k,\hat{x}_k)$$

This summation can be written as

$$\frac{1}{n}\sum_{x,\hat{x}} d(x,\hat{x})N(x,\hat{x}|\vec{x},\vec{\hat{x}}) \quad \text{just a notation for counting the occurrences}$$

$$= \frac{1}{n}\sum_{x,\hat{x}}^{n} d(x,\hat{x})(np(x,\hat{x})N(x,\hat{x}|\vec{x},\vec{\hat{x}}) - np(x,\hat{x}))$$

Then we distribute the terms.

$$= \sum_{x,\hat{x}}^{n} d(x,\hat{x})p(x,\hat{x} + \sum_{x,\hat{x}} d(x,\hat{x})(\frac{1}{n}N(x,\hat{x}|\vec{x},\vec{\hat{x}}) - p(x,\hat{x}))$$

$$\begin{aligned}
&= \mathbb{E}d(X,\hat{X}) + \sum_{x,\hat{x}} d(x,\hat{x})(\frac{1}{n}N(x,\hat{x}|\vec{x},\vec{\hat{x}}) - p(x,\hat{x})) \\
&\leq \mathbb{E}d(X,\hat{X}) + \sum_{x,\hat{x}} d(x,\hat{x})|\frac{1}{n}N(x,\hat{x}|\vec{x},\vec{\hat{x}}) - p(x,\hat{x})| \\
&\leq \mathbb{E}d(X,\hat{X}) + d_{max}\sum_{x,\hat{x}} |\frac{1}{n}N(x,\hat{x}|\vec{x},\vec{\hat{x}}) - p(x,\hat{x})| \\
&\leq \mathbb{E}d(X,\hat{X}) + d_{max}\delta \quad \text{by definition of strong typicality} \\
&\leq D + d_{max}\delta \quad \text{by assumption}
\end{aligned}$$

□

To finish up the proof of achievability, we upper bound $\Pr\{d(\vec{X},\vec{\hat{X}}) > D + \epsilon\}$ by conditioning on whether K is 1.
After showing that for any $\hat{X}$ such that $\mathbb{E}d(X,\hat{X}) \leq D$, $(I(X;\hat{X}), D)$ is achievable. Finally, we minimize $I(X;\hat{X})$ over all $\hat{X}$ with acceptable distortion to conclude that $(R_I(D), D)$ is achievable, i.e.,

$$R_I D \geq R(D)$$

by definition of $R(D)$.

Quasi-uniform array interpretation

We have $\approx 2^{nH(X)}$ sequences in $T^n_{[X]_\delta}$ and $\approx 2^{nI(X;\hat{X})}$ codewords in $T^n_{[\hat{X}]_\delta}$.

- For each codeword that is a typical $\hat{X}$ sequence, it's jointly typical with approximately $2^{nH(X|\hat{X})}$ typical X sequences.
- Therefore, the number of codewords must be at least

$$\frac{2^{nH(X)}}{2^{nH(X|\hat{X})}} \approx 2^{nI(X,\hat{X})} \geq 2^{nR_I(D)}$$

The Blahut-Arimoto Algorithms

We will now discuss how to evaluate channel capacity and rate distortion function numerically.

12.1 Single-Letter Characterization

- For a DMC $p(y|x)$, the capacity

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

  where $r(x)$ is the input distribution, gives the maximum asymptotically achievable rate for reliable coomunication as blocklength $n \to \infty$

- This characterization of C, in the form of an optimization problem, is called a **single-letter characterization** because it involves only $p(y|x)$ but not n.

- Similarly, the rate-distortion function

$$R(D) = \min_{Q(\hat{x}|x):\mathbb{E}d(X,\hat{X})\le D} I(X;\hat{X})$$

  for an i.i.d. information source $\{X_k\}$ is a single-letter characterization because it does not involve the blocklength n.

12.2 Numerical Methods

- When the alphabets are finite, C and $R(D)$ are given as solutions of finite-dimensional optimization problems.

- In general, we are not able to express these quantities in closed-forms. In fact, we can only do it in very special cases.

- The Blahut-Arimoto algorithms are iterative algorithms devised for this purpose.

12.2.1 A double supremum

Consider the double supremum

$$\sup_{u_1\in A_1}\sup_{u_2\in A_2} f(u_1,u_2)$$

- A_i is a convex subset of R^{n_i}
- $f : A_1 \times A_2 \to \mathbb{R}$ is bounded from above, such that
  - f is continuous and has continuous partial derivaties on $A_1 \times A_2/$
  - For all $u_2 \in A_2$, there exists a unique $c_1(u_2) \in A_1$ such that

$$f(c_1(u_2), u_2) = \max_{u_1' \in A_1} f(u_1', u_2)$$

    and for all $u_1 \in A_1$, there exists a unique $c_2(u_1) \in A_2$ such that

$$f(u_1, c_2(u_1)) = \max_{u_2' \in A_2} f(u_1, u_2')$$

- The supremum of f is taken over the Cartesian product $A := A_1 \times A_2$ and we denote $f^* := \sup_{\vec{u} \in A} f(\vec{u})$

12.2.2 An alternating optimization

Here is a very intuitive iterative optimization algorithm.

- Let $\vec{u}^{(k)} = (u_1^(k), u_2^(k))$ for $k \geq$ be defined as follows:
- Let $u_1^{(0)}$ be an arbitrarily chosen vector is A_1, and let $u_2^{(0)} = c_2(u_1^{(0)})$.
- For $k \geq 1$, $\vec{u}^{(k)}$ is defined by

$$u_1^{(k)} = c_1(u_2^{(k-1)})$$

  and

$$u_2^{(k)} = c_2(u_1^{(k)})$$

  Let $f^{(k)} = f(\vec{u}^{(k)})$. Then

$$f^{(k)} \geq f^{(k-1)}$$

  because the points are getting "better and better."
- This algorithm must converge by the monotone-bounded theorem. f is bounded above by assumption.
- We will show that $f^{(k)} \to f^*$ if f is concave.
- We can solve the minimization problem by replacing f with $-f$ so that double sup becomes double inf.
- This alternating optimization algorithm will be used to compute C and $R(D)$.
- This algorithm can be visualized if we draw a contour plot. We start at some point and we first move horizontally to find an optimum and then move vertically and stop at a new optimum. Then move horizontally again... Hopefully we can reach the top of the mountain (we certainly will if the mountain is concave.)

12.3 Computing Channel Capacity

Lemma about conditional distribution

Let $r(x)p(y|x)$ be a given joint distribution on $\mathcal{X} \times \mathcal{Y}$ such that $r > 0$. Let q be a transition matrix from $\mathcal{Y} \to \mathcal{X}$. Then

$$\max_q \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} = \sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{r(x)}$$

where the maximization is taken over all q such that

$$q(x|y) = 0 \quad \text{if and only if} \quad p(y|x) = 0$$

and

$$q^*(x|y) = \frac{r(x)p(y|x)}{\sum_{x'} r(x')p(y|x)} = \quad \text{the conditional distribution } q_{X|Y}$$

- Intuitively, we think of $r(x)$ as the input distribution and $p(y|x)$ as the rule of transmission.

Proof. Let

$$w(y) = \sum_{x'} r(x')p(y|x')$$

Then

$$r(x)p(y|x) = w(y)q^*(x|y)$$

For any reverse transition matrix q, consider

$$\begin{aligned}
&\sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{r(x)} - \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} \\
&= \sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{q(x|y)} \\
&= \sum_y \sum_x w(y)q^*(x|y) \log \frac{q^*(x|y)}{q(x|y)} \\
&= \sum_y w(y) \sum_x q^*(x|y) \log \frac{q^*(x|y)}{q(x|y)} \\
&= \sum_y w(y) D(q* || q) \\
&\leq 0 \quad \text{by non-negativity of KL divergence/relative entropy}
\end{aligned}$$

□

Recharacterization of Channel Capacity

For a discrete memoryless channel $p(y|x)$,

$$C = \sup_{r>0} \max_q \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} = \max_{r \geq 0} I(r, p)$$

Here we write $I(X;Y)$ as $I(r,p)$.

Proof. 1. We first prove

$$C = \max_{r \geq 0} I(x,p) = \sup_{r>0} I(r,p)$$

let r^* achieve C. If $r^* > 0$, then no problem. If $r^* \geq 0$. Since $I(r,p)$ is continuous with respect to r, for any $\epsilon > 0$, $\exists \delta > 0$ such that if

$$\|r - r^*\| < \delta$$

then

$$|C - I(r,p)| < \epsilon$$

In particular, there exists $\hat{r} > 0$ strictly positive such that $\|\hat{r} - r\| < \delta$. Visualization: r^* is on the boundary of the probability simplex as there are some zero probability masses in r^*. Within $Ball(r^*, \delta)$, we have a strictly positive distribution.

Then

$$C = \max_{r \geq 0} I(x,p) = \sup_{r>0} I(r,p) \geq I(\hat{r},p) > C - \epsilon$$

Let $\epsilon \to 0$, we have that $C = \sup_{r>0} I(r,p)$ □

The BA Algorithm for computing C

1. Let
$$f(r,q) = \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)}$$
where r plays the role of u_1 and q plays the role of u_2 in the double supremum problem.

2. Let
$$A_1 = \{(r(x), x \in \mathcal{X}) : r(x) > 0 \quad \text{and} \sum_x r(x) = 1\} \subset \mathbb{R}^{|\mathcal{X}|}$$
and
$$A_2 = (q(x|y), (x,y) \in \mathcal{X} \times \mathcal{Y}) : q(x|y) \geq 0,$$
$$q(x|y) > 0 \quad \text{iff} \quad p(y|x) > 0, \sum_x q(x|y) = 1 \forall y \in \mathcal{Y}\}$$
$$\subset \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$$

Note that

- Both A_1 and A_2 are convex.
- f is bounded above
- By property of A_2, all the probabilites involved in the double summation are positive.

- Therefore, f is continuous and has continuous partial derivatives on $A = A_1 \times A_2$.
- By the above theorem, $f^* = C$. (supremum over $q \in A_2$ is in fact a maximum by the above lemma)

Algorithm details

1. By the above lemma, for any given $r \in A_1$, the unique $q \in A_2$ that maximizes f is given by
$$q(x|y) = \frac{r(x)p(y|x)}{\sum_{x'} r(x')p(y|x')}$$
2. Using Lagrange multipliers, we can show that for any given $q \in A_2$, the unique input distribution r that maximizes f is given by
$$r(x) = \frac{\prod_y q(x|y)p(y|x)}{\sum_{x'} \prod_y q(x'|y)p(y|x')}$$
where $\prod_y$ is over all y such that $p(y|x) > 0$
3. Let $r^{(0)}$ be arbitrary positive input distribution in A_1. Then we can compute $q^{(0)}$

Maximizing f given a fixed q

Last week, we learned how to use BA algorithm to compute channel capacity numerically. Now we compute the rate-distortion function. But to be honest, I cannot wait to get started with differential entropy.

12.4 Algorithm for computing the rate-distortion function

The idea is similar to the previous section and I don't have enough interest to go through the technical details again.

12.5 Convergence of the algorithm

12.5.1 How to prove convergence

- Consider the double supremum optimization problem
- First prove that in general that if f is concave, then $f^{(k)} \to f^*$.
- Then apply this sufficient condition to prove the convergence of the BA algorithm for computing the channel capacity.

Recall that

- $$\vec{u}^{(k+1)} = (u_1^{(k+1)}, u_2^{(k+1)}) = (c_1(u_2^{(k)}), c_2(c_1(u_2^{(k)})))$$
- Define
$$\Delta f(\vec{u}) = f(c_1(u_2), c_2(c_1(u_2)) - f(u_1, u_2)$$
- Then
$$f^{(k+1)} - f^{(k)} = \Delta f(u^{(k)})$$

If f is concave, the algorithm doesn't get trapped

Let f be concave. Then:
If $f^{(k)} < f^*$, then $f^{(k+1)} > f^{(k)}$.

Proof. It suffices to prove that $\Delta f(\vec{u}) > 0$ for any $\vec{u}$ such that $f(\vec{u}) < f^*$.

1. Suppose $\Delta f(\vec{u}) = 0$. We show that convergence has been achieved already. Consider
$$f(c_1(u_2), c_2(c_1(u_2))) \geq f(c_1(u_2)) \geq f(u_1, u_2)$$
If $\Delta f(\vec{u}) = 0$, then the above inequalities are equalities. Due to the uniquess of $c_2(\cdot)$ and $c_1(\cdot)$, equalities above imply that
$$u_1 = c_1(u_2), u_2 = c_2(c_1(u_2)) = c_2(u_1)$$
meaning the algorithm has converged.

2. Second, consider any $\vec{u} \in A$ such that $f(\vec{u}) < f^*$. Prove $\Delta f(\vec{u}) > 0$ by contradiction. Suppose by contradiction that $\Delta f(\vec{u}) = 0$. Then $u_1 = c_1(u_2)$ and $u_2 = c_2(u_1)$, i.e. u_1, u_2 maximize f by fixing the other.
Since $f(\vec{u}) < f^*$, there exists $\vec{v} \in A$ such that $f(\vec{u}) < f(v)$. Let
   - $\hat{z}$ be the unit vector in the direction of $\vec{v} - \vec{u}$
   - $z_1 = (1, 0)$, $z_2 = (0, 1)$

   Then $\hat{z} = \alpha_1 z_1 + \alpha_2 z_2$ where $\alpha_i = \frac{|v_i - u_i|}{\|\vec{v} - \vec{u}\|}$. Since f is continuous and has continuous partial derivatives, the directional derivative of f at $\vec{u}$ in the direction of z_1 is given by $\nabla f \cdot z_1$.
   f attains its maximum at $\vec{u}$ when u_2 is fixed.
   Consider the line passing through (u_1, u_2) and (v_1, u_2), u_1 is the optimal value along this line so f attains its maximum at $\vec{u}$ along this line. Therefore, $\nabla f \cdot z_1 = 0$. Similarly $\nabla f \cdot z_2 = 0$.
   Then $\nabla f \cdot \hat{z} = 0$.
   Since f is concave, this implies $f(\vec{u}) \geq f(v)$, a contradiction. Therefore $\Delta f(\vec{u}) > 0$.

□

Even if the algorithm does not get trapped, the increment can be arbitrarily small. We now show that this is not a problem.

Theorem: if f is concave, then the algorithm converges

If f is concave, then $f^{(k)} \to f^*$

Proof. 1. First of all, $f^{(k)}$ is necessarily convergent by the monotone-bounded theorem. Denote the limit by f'.

2. Hence, for any $\epsilon > 0$ and sufficiently large k,

$$f' - \epsilon \leq f^{(k)} \leq f'$$

3. Let

$$\gamma = \min_{\vec{u} \in A'} \Delta f(\vec{u})$$

where $A' = \{\vec{u} \in A; f' - \epsilon \leq f^{(k)}(\vec{u}) \leq f'\}$. γ is well defined because:

4. $\Delta f(\vec{u})$ is continuous because we assumed that f has continuous partial derivatives.

5. A' is compact because it is the inverse image of a closed interval under a continuous function and A is bounded. Therefore, γ is well defined.

6. If $f' < f^*$, since f is concanve, $\Delta f(\vec{u}) > 0$ for all $\vec{u} \in A'$. Then $\gamma > 0$, by the above lemma.

7. Therefore, for sufficiently large k, $\Delta f(\vec{u}^{(k)}) \geq \gamma$. So the algorithm will converge to the optimum.

□

To wrap up this section, we need to show that f is concave in the case of computing the channel capacity. We can use the log-sum inequality for the proof.

Differential Entropy

So far, we have been talking about discrete random variables and discrete-time information transmission. In this chapter, we will discuss

- Real-valued random vectors
- Symmetric, positive definite, and covariance matrices
- Differential entropy and mutual information
- AEP and informational divergence/relative entropy
- Gaussian distribution: usually used to model noise.

13.1 Real random variables

Here is a quick review of some (probably too) basic facts.

- A real r.v. X with cumulative distribution function (CDF) $F_X(x) = \Pr\{X \leq x\}$ is continuous if $F_X(x)$ is continuous.
- If the CDF increases only at a countable number of x, then X would be discrete.
- We say that X is mixed if its CDF is neigher discrete nor continuous. (I think in most probability books, this case is also considered continuous.)
- Support of X:

$$S_X = \{x \in \mathbb{R} : F_X(x) > F_X(x - \epsilon) \quad \text{for all } \epsilon > 0\}.$$

- $$\mathbb{E}g(X) := \int_{S_X} g(x) dF_X(x)$$

  where the RHS is the Lebesgue-Stieltjes integral (where the the CDF (the measure) can be discrete continuous or mixed)
- A nonnegative function $f_X(x)$ is called a probability density function (pdf) of X if

$$F_X(x) = \int_{-\infty}^{x} f_X(u) du$$

  for all x.

- By the fundamental theorem of calculus,

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_X(u)du = f_X(x)$$

- If X has a pdf, then X is continuous, but not vice versa. We can think of $dF_X(x)$ as $f_X(x)dx$ but $dF_X(x)$ is something more general.

Jointly Distributed Random Variables

- Let X and Y be two real random variables with joint CDF $F_{XY}(x,y) = \Pr\{X \leq x, Y \leq y\}$.

- Marginal CDF of X: $F_X(x) = F_{XY}(x,\infty)$

- A non-negative function $f_{XY}(x,y)$ is called a joint pdf of X and Y if

$$F_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)dydu$$

- Conditional pdf of Y given $\{X = x\}$:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Variance and Covariance

- Variance of X:

$$varX = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

- Covariance between X and Y:

$$cov(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$$

-

$$var(X+Y) = varX + varY + 2cov(X,Y)$$

- $IfX \perp Y$, then $cov(X,Y) = 0$ and we say that X and Y are uncorrelated. However, the converse if not true.

- If $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} varX_i$$

13.2 Random Vectors

- Let $\vec{X} = [X_1, \ldots, X_n]^T$.
- Covariance matrix:

$$K_{\vec{X}} = \mathbb{E}(\vec{X} - \mathbb{E}\vec{X})(\vec{X} - \mathbb{E}\vec{X})^T = [cov(X_i, X_j)]_{i,j=1}^n$$

- The ith diagonal element is $var(X_i)$
- Correlation matrix:

$$\tilde{K}_{\vec{X}} = \mathbb{E}\vec{X}\vec{X}^T = [\mathbb{E}X_iX_j]$$

- Relations between $K_{\vec{X}}$ and $\tilde{K}_{\vec{X}}$:
  - $$K_{\vec{X}} = \tilde{K}_{\vec{X}} - (\mathbb{E}\vec{X})(\mathbb{E}\vec{X})^T$$
  - $$K_{\vec{X}} = \tilde{K}_{\vec{X}-\mathbb{E}\vec{X}} \quad \text{normalize the expectation to 0}$$
- The above is the generalization of

$$varX = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

and

$$var(X, Y) = \mathbb{E}(X - \mathbb{E}X)^2$$

respectively.

13.3 Gaussian Distribution

Review: symmetric, pd

- A square matrix K is symmetric if $K^T = K$
- An $n \times n$ matrix is positive definite if $x^T Kx > 0$ for all nonzero vector x. In 54, we know that this is equivalent to all e-vals being positive.

- $\mathcal{N}(\mu, \sigma^2)$: Gaussian distribution with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}, \infty < x < \infty$$

- $\mathcal{N}(\vec{\mu}, K)$: multivariate Gaussian distribution with mean μ and covariance matrix K, i.e. the joint pdf of the distribution is given by

$$f(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n (\det K)^{1/2}} \exp(-\frac{1}{2}(\vec{x} - \vec{u})^T K^{-1}(\vec{x} - \vec{\mu})), x \in \mathbb{R}^n$$

where K is a symmetric positive definite matrix (classical result).

- Note that both K and K^{-1} are symmetric positive definite matrices.
- Note that this implies that $(\vec{x}-\vec{\mu})^T K^{-1}(\vec{x}-\mu)$ is always positive.

Diagonalization

- A symmetric matrix K can be diagonalized as

$$K = Q\Lambda Q^T$$

where Λ is a diagonal matrix and Q is an orthogonal matrix (i.e. $Q^{-1} = Q^T$, note $\det Q = \det Q^T = \pm 1$) (spectral thm)

- $$KQ = Q\Lambda$$

so

$$kq_i = \lambda_i q_i,$$

i.e. q_i is an eigenvector of K with eigenvalue λ_i

- This implies that the e-vals of a psd matric are non-negative.

Proposition: linear transformation of a random vector

Let $\vec{Y} = A\vec{X}$, where A is an $n \times n$ matrix. Then

$$K_{\vec{Y}} = AK_{\vec{X}}A^T$$

and similarly

$$\tilde{K}_{\vec{Y}} = A\tilde{K}_{\vec{X}}A^T$$

Proof.

$$\begin{aligned} K_{\vec{Y}} &= \mathbb{E}YY^T - (\vec{E}Y)(\vec{E}Y)^T \\ &= \mathbb{E}(AX)(X^TA^T) - (\vec{E}AX)(\mathbb{E}X^TA^T) \\ &= A\mathbb{E}(XX^T)A^T - A(\mathbb{E}X)(\mathbb{E}X^T)A^T \\ &= AK_{\vec{X}}A^T \end{aligned}$$

□

Decorrelation

Let $\vec{Y} = Q^T\vec{X}$ where $K_{\vec{X}} = Q\Lambda Q^T$. Then

$$K_{\vec{Y}} = \Lambda$$

i.e.

- The random variables in $\vec{Y}$ are uncorrelated.
- $VarY_i = \lambda_i$ for all i.

Proof. By the above proposition,

$$K_{\vec{Y}} = Q^T K_{\vec{X}} Q = Q^T Q\Lambda Q^T Q = \Lambda$$

Note $cov(Y_i, Y_j) = 0$ for $i \neq j$. □

Remark 13.1. As a corollary: any random vector $\vec{X}$ can be written as a linear transformation of an uncorrelated vector, i.e. $\vec{X} = Q\vec{Y}$ where $K_{\vec{X}} = Q\Lambda Q^T$.

The following is a generalization of the fact that for independent RV X and Y, we have $var(X + Y) = varX + varY$.

Proposition

Let $\vec{X}$ and $\vec{Z}$ be independent and $\vec{Y} = \vec{X} + \vec{Z}$. Then

$$K_{\vec{Y}} = K_{\vec{X}} + K_{\vec{X}}$$

Preservation of Energy

Let $\vec{Y} = Q\vec{X}$, where Q is an orthogonal matrix. Then

$$\mathbb{E}\sum_{i=1}^{n} Y_i^2 = \mathbb{E}\sum_{i=1}^{n} X_i^2$$

Proof. Consider

$$\begin{aligned}\sum_{i=1}^{n} Y_i^2 &= Y^T Y \\ &= X^T Q^T Q X^T \\ &= \sum_{i=1}^{n} X_i^2\end{aligned}$$

□

13.4 Definition

Differential Entropy

The differential entropy $h(X)$ of a continuous random variable X with pdf $f(x)$ is defined as

$$h(X) = -\int_S f(x)\log f(x)dx = -\mathbb{E}\log f(X)$$

Remark 13.2.

- Differential entropy is not a measure of the amount of information contained in a continuous r.v. despite they have similar forms.

- A continuous random variable generally contains an infinite amount of information.

Example

Let X be uniformly distributed on $[0, 1)$. Then we can write

$$X = 0.X_1X_2X_3\ldots$$

the dyadic expansion of X where X_i are fair bits.
Then

$$\begin{aligned} H(X) &= H(X_1, X_2, X_3, \ldots) \\ &= \sum_{i=1}^{\infty} H(X_i) \\ &= \sum_{i=1}^{\infty} 1 \\ &= \infty \end{aligned}$$

13.4.1 Relation with Discrete Entropy

Relation with Discrete Entropy

- Consider a continuous r.v. X with a continuous pdf $f(x)$.
- Define a discrete r.v. $\hat{X}_\Delta$ by

$$\hat{X}_\Delta = i \quad \text{if } X \in [i\Delta, (i+1)\Delta]$$

- $\hat{X}_\Delta$ is called a quantization of X with resolution Δ
- Since $f(x)$ is continuous,

$$p_i = \Pr\{\hat{X}_\Delta = i\} \approx f(x_i)\Delta$$

where $x_i \in [i\Delta, (i+1)\Delta]$

- Then for small Δ

$$\begin{aligned} H(\hat{X}_\Delta) &= -\sum_i p_i \log p_i \\ &\approx -\sum_i (f(x_i)\Delta)\log(f(x_i)\Delta) \\ &= -\sum_i (f(x_i)\Delta)(\log f(x_i) + \log\Delta) \\ &= -\sum_i [f(x_i)\log f(x_i)]\Delta - (\log\Delta)\sum_i f(x_i)\Delta \\ &\approx -\int f(x)\log f(x)dx - \log\Delta \int f(x)dx \\ &= h(X) - \log\Delta \end{aligned}$$

- The argument above can be formalized by chasing the definition of continuity.

Example: uniform random variable

Let X be uniformly distributed on $[0, a)$. Then

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

Remark 13.3.
- $h(X) < 0$ if $a < 1$ so $h(\cdot)$ cannot be a measure of information.

Gaussian Distribution

Let $X \sim N(0, \sigma^2)$. Then

$$h(X) = \frac{1}{2}\log(2\pi e\sigma^2)$$

Proof. Use e as the base of the logarithm. Then

$$\begin{aligned} h(X) &= -\int f(x) \ln f(x) dx \\ &= -\int f(x) \left(-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right) dx \\ &= \frac{1}{2\sigma^2} \int x^2 f(x) dx + \ln \sqrt{2\pi\sigma^2} \int f(x) dx \\ &= \frac{\mathbb{E}X^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \\ &= \frac{\sigma^2 + 0}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \\ &= \frac{1}{2} \ln(2\pi e\sigma^2) \end{aligned}$$

□

Fangyuan: it's kinda like variance isn't it

13.5 Properties of Differential Entropy

Property: Invariance under Translation

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

Proof. Let $Y = X + c$. Then

$$f_Y(y) = f_X(y - c)$$

and $S_Y = \{x + c : x \in S_X\}$ Let $x = y - c$

$$\begin{aligned} h(X) &= -\int_{S_X} f_X(x) \log f_X(x) dx \\ &= -\int_{S_Y} f_X(y - c) \log_X(y - c) dy \\ &= -\int_{S_Y} f_Y(y) \log f_Y(y) dy \\ &= h(Y) \end{aligned}$$

□

Property: Scaling

For $a \neq 0$,

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

Remark 13.4. The differential entropy is

- increased by $\log|a|$ if $|a| > 1$
- decreased by $-\log|a|$ if $|a| < 1$
- unchanged if $a = \pm 1$
- related to the "spread" of the pdf. The more spread out the pdf is, the larger the differential entropy is.

Proof. Let $Y = aX$. Then

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$$

Let $x = \frac{y}{a}$.

$$\begin{aligned} h(X) &= -\int_{S_X} f_X(x) \log f_X(x) dx \\ &= -\int_{S_Y} f_X(\frac{y}{a}) \log f_X(\frac{y}{a}) \frac{dy}{|a|} \\ &= -\int_{S_Y} \frac{1}{|a|} f_X(\frac{y}{a}) \left[\log \frac{1}{|a|} f_X(\frac{y}{a}) + \log|a|\right] dy \\ &= h(Y) - \log|a| \end{aligned}$$

□

13.6 Joint Differential Entropy, Conditional Differential Entropy and Mutual Information

Joint differential entropy

The joint differential entropy $h(\vec{X})$ of a random vector $\vec{X}$ with joint pdf $f(\vec{x})$ is defined as

$$h(\vec{X}) = s\int_S f(\vec{x}) \log f(\vec{x}) d\vec{x} = -\mathbb{E} \log f(\vec{X})$$

Proposition

If $X_1, \ldots, X_n$ are mutually independent, then

$$h(\vec{X}) = \sum_{i=1}^{n} h(X_i)$$

Invariance under translation

$$h(\vec{X} + c)$$

Scaling

$$h(A\vec{X}) = h(\vec{X}) + \log|\det A|$$

Theorem: Joint Differential Entropy of Multivariate Gaussian

Let $\vec{X} \sim \mathcal{N}(\vec{\mu}, K)$. Then

$$h(\vec{X}) = \frac{1}{2}\log\left((2\pi e)^n |K|\right).$$

Proof.

1. Let K be diagonzalized as $Q\Lambda Q^T$.
2. Let $\vec{X} = Q\vec{Y}$ where the components in $\vec{Y}$ are uncorrelated with $varY_i = \lambda_i$, the diagonal elements of Λ
3. Since $\vec{X}$ is Gaussian, so is its linear transformation $\vec{Y}$.
4. Then the random variables in $\vec{Y}$ are mutually independent because they are uncorrelated.
5. Consider

$$\begin{aligned}
h(\vec{X}) &= h(Q\vec{Y}) \\
&= h(\vec{Y}) + \log|\det Q| \\
&= h(\vec{Y}) + 0 \\
&= \sum_{i=1^n h(\vec{Y}_i)} \\
&= \sum_{i=1}^{n} \frac{1}{2}\log(2\pi e\lambda_i) \\
&= \frac{1}{2}\log\left[(2\pi e)^2 \prod_i \lambda_i\right] \\
&= \frac{1}{2}\log\left[(2\pi 2)^n \det K\right]
\end{aligned}$$

□

13.7 Conditional Differential Entropy

Now we generalize the model of channel to have arbitrary input.

Generalized Channel

The output random variable Y (continuous/discrete) is related to the (general) input random variable X through a conditional pdf $f(y|x)$ or conditional pmf $p(y|x)$ defined for all x.

Conditional Differential Entropy

Let X and Y be jointly distributed random varaibles where Y is continuous and is related to X through a conditional pdf $f(y|x)$ defined for all x. The conditional differential entropy of Y given $\{X = x\}$ is defined as

$$h(Y|X = x) = -\int_{S_Y(x)} f(y|x) \log f(y|x) dy$$

and the conditional differential entropy of Y given X is defined as

$$h(Y|X) = -\int_{S_X} h(Y|X = x) dF(x) = -\mathbb{E} \log h(Y|X)$$

Again, $dF(x)$ is measure-theoretical generalization of $f(x)dx$.

Here is a review a basic fact.

- If Y relates to X through $f(y|x)$, then $f(y) = \int f(y|x) dF(x)$. The proof uses Fubini's theorem to exchange the order of integration.

13.8 Differential Mutual Information

Mutual Information

The mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &:= \mathbb{E} \log \frac{f(Y|X)}{f(Y)} \\ &= \int_{S_X} \int_{S_Y} f(y|x) \log \frac{f(y|x)}{f(y)} \, \mathrm{d}y \, \mathrm{d}F(x) \end{aligned}$$

If both X and Y are continuous and $f(x, y)$ exists, then

$$I(X;Y) = \mathbb{E} \log \frac{f(X,Y)}{f(X)f(Y)}$$

Conditional Mutual Information

The mutual information between X and Y given T is defined as

$$I(X;Y|T) = \int_{S_T} I(X;Y|T = t) dF(t) = \mathbb{E} \log \frac{f(Y|X,T)}{f(Y|T)}$$

where

$$I(X;Y|T = t) = \int_{S_X(t)} \int_{S_Y(x,t)} f(y|x,t) \log \frac{f(y|x,t)}{f(y|t)} \, \mathrm{d}y \, \mathrm{d}F(x|t)$$

13.9 Interpretation of $I(X;Y)$

1. Assume $f(x,y)$ exists and is continuous.
2. For a fixed Δ, for all integer i and j, define the intervals
$$A_x^i = [i\Delta, (i+1)\Delta]$$
on the X-axis and
$$A_y^j = [j\Delta, (j+1)\Delta]$$
on the Y-axis. Define the rectangles
$$A_x^i A_y^j = A_x^i \times A_y^j$$
3. Define discrete r.v.'s
$$\begin{cases} \hat{X}_\Delta = i & \text{if in } A_x^i \\ \hat{Y}_\Delta = j & \text{if } Y \in A_y^j \end{cases}$$
4. $\hat{X}_\Delta$ and $\hat{Y}_\Delta$ are called quantizations of X and Y.
5. For all i and j, pick any $(x_i, y_j) \in A_x^i \times A_y^j$.
6. Then
$$\begin{aligned} & I(\hat{X}_\Delta, \hat{Y}_\Delta) \\ &= \sum_i \sum_j \Pr\{(\hat{X}_\Delta, \hat{Y}_\Delta) = (i,j)\} \log \frac{\Pr\{(\hat{X}_\Delta, \hat{Y}_\Delta) = (i,j)\}}{\Pr\{\hat{X}_\Delta = i\} \Pr\{\hat{Y}_\Delta = j\}} \\ &\approx \sum_i \sum_j f(x_i, y_j)\Delta^2 \log \frac{f(x_i, y_j)\Delta^2}{f(x_i)\Delta(f(y_j)\Delta)} \quad \text{by continuity of } f_{XY} \\ &\approx \int\int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \, \mathrm{d}x \, \mathrm{d}y \\ &= I(X;Y) \end{aligned}$$
7. Therefore, $I(X;Y)$ can be interpreted as the limit of $I(\hat{X}_\Delta, \hat{Y}_\Delta)$ as the resolution $\Delta \to 0$

Proposition

For two random variables X and Y,

1. If Y is continuous,
$$h(Y) = h(Y|X) + I(X;Y)$$
2. If Y is discrete,
$$H(Y) = H(Y|X) + I(X;Y)$$

Chain Rule

The chain rule has the same form as the discrete analogue:

$$h(X_1, \ldots, X_n)$$

Non-negativity of mutual information

$I(X;Y) \geq 0$ with equality if and only if X is independent of Y.

- Same goes to conditional mutual information.
- As a corollary, conditioning does NOT increase differential entropy (natural).
- As a corollary, we have independence bound for differential entropy:

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

with equality if and only X_i are mutually independent.

13.10 AEP for Continuous Random Variables

AEP I

$$-\frac{1}{n} \log f(\vec{X}) \to h(X)$$

in probability as $n \to \infty$. i.e. for any $\epsilon > 0$, for sufficiently large n, we have

$$\Pr\{|-\frac{1}{n} \log f(\vec{X}) - h(X)| < \epsilon\} > 1 - \epsilon$$

Proof. This is a consequence of the weak law of large numbers. □

Typical Sequence

The typical set with respect to pdf $f(x)$ is the set of sequences $\vec{x}$ such that

$$|-\frac{1}{n} \log f(\vec{x}) - h(X)| < \epsilon$$

i.e. the empirical differential entropy is close to the true differential entropy.

Volume

The volumn of a set $A \subset \mathbb{R}^n$ is defined as

$$Vol(A) = \int_A d\vec{x}$$

AEP II for continuous random variables

The following hold for any $\epsilon > 0$:

1. If $\vec{x} \in W^n_{[X]_\epsilon}$, then
$$2^{-n(h(X)+\epsilon)} < f(\vec{x}) < 2^{-n(h(\vec{X})-\epsilon)}$$
2. for sufficiently large n,
$$\Pr\{\vec{X} \in W^n_{[X]_\epsilon} > 1 - \epsilon$$
3. for sufficiently large n,
$$(1-\epsilon)2^{n(h(X)+\epsilon)} < Vol(W^n_{[X]_\epsilon}) < 2^{n(h(\vec{X})+\epsilon)}$$

- If the differential entropy is large, then the volume of the typical set is also large.

13.11 Differential Informational Divergence

Information Divergence

Let f and g be two pdf's defined on $\mathbb{R}^n$ with supports $\mathcal{S}_f$ and $\mathcal{S}_g$. The information diverfence between the two distributions f and g is defined as
$$D(f||g) := \int_{\mathcal{S}_f} f(x) \log \frac{f(x)}{g(x)} \, \mathrm{d}x = \mathbb{E}_f \log \frac{f(X)}{g(X)}$$
where E_f is expectation with respect to the distribution f.

Remark 13.5. If $D(f||g) < \infty$, then
$$\mathcal{S}_f \backslash \mathcal{S}_g = \{x : f(x) > 0 \text{ and } g(x) = 0\}$$
has zero Lebesgue measure, i.e. S_f is essentially a subset of $\mathcal{S}_g$, except for countably many points.

Divergence Inequality

Let f and g be two pdf's defined on $\mathbb{R}^n$. Then
$$D(f||g) \geq 0$$
with equality if and only if $f = g$ almost surely.

13.12 Maximum Discrete Entropy Distributions

Maximization of entropy

Consider the problem of maximizing over all probability distributions p defined on a countable subset $\mathcal{S}$ of the set of real numbers, subject to

$$\sum_{x \in \mathcal{S}_p} p(x) r_i(x) = a_i (= \mathbb{E}_p r_i(X))$$

for $1 \leq i \leq m$, where $\mathcal{S}_p \subset \mathcal{S}$ and $r_i(x)$ is defined for all $x \in \mathcal{S}$.

- Theorem: $p^*(x) = 2^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(x)}$ for all $x \in \mathcal{S}$, where λ_i are chosen such that the constraint on the expectation of r_i are satisfied. Then p^* maximizes $H(p)$ over all probability distribution p on $\mathcal{S}$ subject to the constraints.
- Let $q_i = e^{-\lambda_i}$. Then we can write

$$\begin{aligned} p^*(x) &= e^{-\lambda_0} e^{-\lambda_1 r_1(x)} \dots e^{-\lambda_m r_m(x)} \\ &= q_0 q_1^{r_1(x)} \dots q_m^{r_m(x)} \end{aligned}$$

  where q_0 is called the normalization constant.

Proof. Consider

$$\begin{aligned} & H(p^*) - H(p) \\ &= -\sum_{x \in \mathcal{S}} p^*(x) \ln p^*(x) + \sum_{x \in S_p} p(x) \ln p(x) \\ &= -\sum_{x \in \mathcal{S}_p} p(x) \ln p^*(x) + \sum_{x \in S_p} p(x) \ln p(x) \quad \text{Non-trivial: see below} \quad = D(p||p^*) \\ &\geq 0 \end{aligned}$$

Detail:

$$
\begin{aligned}
p^*(x) &= e^{-\lambda_0 - \sum_i \lambda_i r_i(x)} \\
\ln p^*(x) &= -\lambda_0 - \sum_i \lambda_i r_i(x) \\
-\sum_{x\in\mathcal{S}} p^*(x) \ln p^*(x) &= -\sum_{x\in\mathcal{S}} p^*(x)(-\lambda_0 - \sum_i \lambda_i r_i(x)) \\
&= -\lambda_0 \sum_{x\in\mathcal{S}} p^*(x) + \sum_i \lambda_i \left(\sum_{x\in\mathcal{S}} p^*(x) r_i(x) \right) \\
&= \lambda_0 + \sum_i \lambda_i a_i \\
&= \lambda_0 \left(\sum_{x\in\mathcal{S}_p} p(x) \right) + \sum_i \lambda_i \left(\sum_{x\in\mathcal{S}_p} p(x) r_i(x) \right) \\
&= -\sum_{x\in\mathcal{S}_p} p(x) \left(-\lambda_0 - \sum_i \lambda_i r_i(x) \right) \\
&= -\sum_{x\in\mathcal{S}_p} p(x) \ln p^*(x)
\end{aligned}
$$

□

Example 2.53

Let $\mathcal{S} = \{0, 1, 2, \dots\}$ and let the set of constraints be

$$\sum_x p(x) x = a \geq 0$$

Then, let $q_i = e^{-\lambda_i}$ for $i = 0, 1$. Then

$$p^*(x) = e^{-\lambda_0} e^{-\lambda_1 x} = q_0 q_1^x$$

Note that p^* is then a geometric distribution, so we have

$$q_1 = 1 - q_0$$

Then by the constraint

$$q_0 = (a+1)^{-1}$$

13.13 Maximum Differential Entropy Distributions

Maximizing differential entropy subject to constraints

Consider the following maximization problem:

- Maximize $h(f)$ over all pdf f defined on a subset $\mathcal{S}$ of $\mathbb{R}^n$, subject to

$$\int_{\mathcal{S}_f} r_i(\vec{x}) f(\vec{x})\,\mathrm{d}\vec{x} = a_i (= \mathbb{E}_f r_i(X))$$

  for $1 \leq i \leq m$ where $\mathcal{S}_f \subset \mathcal{S}$ and $r_i(\vec{x})$ is defined for all $\vec{x} \in \mathcal{S}$.

- Theorem: Let

$$f^*(x) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(\vec{x})}$$

  for all $\vec{x} \in \mathcal{S}$, where λ_i are chosen so that the constraints are satisfied. Then f^* maximizes $h(f)$ over all pdf f defined on $\mathcal{S}$, subject to the constraints.

Upper bound on differential entropy: Gaussian

Let X be a continuous random variable with $\mathbb{E}X^2 = \kappa$. Then

$$h(X) \leq \frac{1}{2}\log(2\pi e \pi \kappa),$$

with equality if and only if $X \sim \mathcal{N}(0, \kappa)$.

Proof. Consider maximizing $h(f)$ subject to the constraint

$$\int x^2 f(x)\,\mathrm{d}x = \mathbb{E}X^2 = \kappa$$

Then by the above theorem, $f^*(X) = ae^{-bx^2}$, which is Gaussian distribution with zero mean.
To satisfy the constraint on the second moment, the only choices are a and b are:

$$a = \frac{1}{\sqrt{2\pi\kappa}}, b = \frac{1}{2\kappa}$$

□

Remark 13.6. Let X be continuous random variable with mean μ and variance σ^2. Then

$$h(X) \leq \frac{1}{2}\log(2\pi e \sigma^2)$$

with equality if and only if $X \sim \mathcal{N}(\mu, \sigma^2)$.

13.14 Differential Entropy and Spread

- From the above theorem, we have that

$$h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2) = \log\sigma + \frac{1}{2}\log(2\pi e)$$

  where $\sigma^2 = varX$

- $h(X)$ is at most equal to the logarithm of the standard deviation (spread) plus a constant.

- $h(X) \to -\infty$ as $\sigma \to 0$.

Generalization of upper bound

1. Let $\vec{X}$ be a vector of n continuous random variables with correlation matrix $\hat{K}$. Then

$$h(\vec{X}) \leq \frac{1}{2}\log\left[(2\pi e)^n \det\hat{K}\right]$$

   with equality if and only if $\vec{X} \sim \mathcal{N}(0, \hat{K})$

2. Let $\vec{X}$ be a vector of n continuous random variables with mean $\vec{\mu}$ and covariance matrix K. Then

$$h(\vec{X}) \leq \frac{1}{2}\log\left[(2\pi e)^n \det K\right]$$

   with equality if and only if $\vec{X} \sim \mathcal{N}(\vec{\mu}, K)$

Proof. 1. Let $r_{ij}(\vec{x}) = x_i x_j$ and $\hat{K} = [\hat{k}_{ij}]$.

2. Then the constraints on $f(\vec{x})$ are equivalent to

$$\begin{aligned}
&\hat{k}_{ij} \\
&= \int_{\mathcal{S}_f} r_{ij}(\vec{x}) f(\vec{x})\,\mathrm{d}\vec{x} \\
&= \int_{\mathcal{S}_f} x_i x_j(\vec{x}) f(\vec{x})\,\mathrm{d}\vec{x} \\
&= \mathbb{E}X_i \mathbb{E}X_j
\end{aligned}$$

3. By the above theorem, the joint pdf that maximizes $h(\vec{X})$ has the form

$$f^*(\vec{x}) = e^{\lambda_0 - \sum_{i,j}\lambda_{i,j}x_i x_j} = e^{-\lambda_0 = \vec{x}^T\Lambda\vec{x}}$$

   where $\Lambda = [\lambda_{ij}]$

4. f^* is the joint pdf of a multivariate Gaussian distribution with zero mean.

□

13.15 Continuous-Valued Channels

- In a communication system, the input and output of a channel not always take discrete values and transmittion is not always in discrete time.
- For example, a waveform channel.

13.16 Discrete-Time Continuous-Valued Channel

We first consider discrete-time continuous-valued channel. Here is a very natural definition.

discrete-time Continuous Channel I

Let $f(y|x)$ be a conditional pdf defined for all x, where

$$h(Y|X=x) = -\int_{S_Y(x)} f(y|x)\log f(y|x)\,\mathrm{d}y < \infty$$

for all x. A discrete-time continuous channel $f(y|x)$ is a system where input random variable X and the output random variable Y are related through $f(y|x)$.

Like before, we will a definition of the channel using noise instead of transition probabilities.

Continuous Channel II

et $\alpha : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and Z be a real random variable, called the noise variable. A (discrete-time) continuous channel (α, Z) is a system with a real input and a real output. For any input random variable X, the noise random variable Z is independent of X, and the output random variable Y is given by

$$Y = \alpha(X, Z)$$

i.e. you know what the output is given the input and the noise.

Equivalence of the two channels

Two continuous channels $f(y|x)$ and (α, Z) are equivalent if for every input distribution $F(x)$,

$$\Pr\{X\le x, \alpha(X,Z)\le y\} = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{Y|X}(v|u)\,\mathrm{d}v\,\mathrm{d}F_X(u) = F_{XY}(X\le x, Y\le y)$$

$=$ for all x and y.

- The second definition is more general than the first definition because it does not require the existence of $f(y|x)$.
- In this section we assume $f(y|x)$ exists and will use the first definition.

Continuous Memoryless Channel I

A continuous memoryless channel (CMC) $f(y|x)$ is a sequence of replicates of a generic continuous channel $f(y|x)$. These continuous channels are indexed by a discrete-time index i. Transmission through a channel is assumed to be instantaneous. Let X_i nad Y_i be the input and the ouput of the CMC at time i, and let T_{i-} denote all the random variables that are generated in the system before the ith input X_i.

$$T_{i-} \to X_i \to Y_i$$

is a Markov chain, and

$$\Pr\{X_i \le x, Y_i \le y\} = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{Y|X}(v|u)\,\mathrm{d}v\,\mathrm{d}F_{X_i}(u).$$

Remark 13.7. I think it's just a bunch of independent copies of a channel.

CMC II

A continuous memoryless channel (α, Z) is a sequence of replicates of a generic continuous channel (α, Z). These continuous channels are indexed by a discrete-time index i. Transmission through a channel is assumed to be instantaneous. Let X_i nad Y_i be the input and the ouput of the CMC at time i, and let T_{i-} denote all the random variables that are generated in the system before the ith input X_i. *The noise variable Z_i for the transmission at time i is a copy of the generic noise variable Z and is independent of (X_i, T_{i-}).* The output of the CMC at time i is given by

$$Y + i = \alpha(X_i, Z_i)$$

Average input constraint

Let κ be a real function. An average input constraint (κ, P) for CMC is the requirement that for any codeword $(x_1, x_2, \ldots, x_n)$ transmitted over the channel,

$$\frac{1}{n}\sum_{i=1}^{n} \kappa(x_i) \le P$$

1. For a fixed value of x, we think of $\kappa(x)$ as the cost of transmitting x.
2. For example, if $\kappa(x) = x^2$, then $\kappa(x)$ is the energy and P is the power.

Channel Capacity

The capacity of a continuous memoryless channel $f(y|x)$ with input constraint (κ, P) is defined as

$$C(P) = \sup_{F(x):\mathbb{E}\kappa(X)\le P} I(X;Y)$$

Property of Channel Capacity

$C(P)$ is

1. non-decreasing
2. concave:
$$C(\lambda P_1 + (1-\lambda)P_2) \geq \lambda C(P_1) + (1-\lambda)C(P_2)$$
3. left-continuous

Proof. 1. immediate

2. consequence of the concavity of mutual information with respect to the input distribution.
   (a) Let $j = 1, 2$. For any P_j, for all $\epsilon > 0$, by the definition of $C(P_j)$, there exists distribution $F_j(x)$ such that
   $$\mathbb{E}\kappa(X_j) \leq P_j$$
   and
   $$I(X_j; Y_j) \leq C(P_j) - \epsilon$$
   (b) Define $X^\lambda \sim \lambda F_1(x) + (1-\lambda)F_2(x)$.
   (c) The $\mathbb{E}\kappa(X^\lambda) \leq \lambda P_1 + (1-\lambda)P_2$.
   (d) Also by the the concavity of mutual information iwth respect to the input distribution, we have
   $$I(X^\lambda, Y^\lambda) \leq \lambda I(X_1; Y_1) + (1-\lambda) I(X_2; Y_2) \geq \lambda C(P_1) + (1-\lambda) C(P_2) - \epsilon$$
   (e) Then
   $$C(\lambda P_1 + \bar{\lambda} P_2) \geq I(X^\lambda, Y^\lambda) \geq \lambda C(P_1) + \bar{\lambda} C(P_2) - \epsilon$$
   for all $\epsilon > 0$.
3. left-continuous: consequence of concavity.
   (a) Let $P_1 < P_2$, so that $P_2 \geq \lambda P_1 + (1-\lambda)P_2$. Since $C(P)$ is non-decreasing, we have
   $$C(P_2) \geq \lambda C(P_1) + \bar{\lambda} C(P_2)$$
   (b) Let $\lambda \to 0$, we have
   $$C(P_2) \geq \lim_{\lambda \to 0} C(\lambda P_1 + \bar{\lambda} P_2) \geq C(P_2)$$
   (c) This implies
   $$\lim_{\lambda \to 0} C(\lambda P_1 + \bar{\lambda} P)2) = C(P_2)$$

□

13.17 The Channel Coding Theorem

Encoding function

An (n, M) code for a continuous memeoryless channel with input constraint (κ, P) is defined by an encoding function

$$e : \{1, 2, \ldots, M\} \to \mathbb{R}^n$$

and a decoding function

$$g : \mathbb{R}^n \to \{1, 2, \ldots, M\}$$

- Message Set: $W = \{1, 2, \ldots, M\}$
- Codewords: $e(1), e(2), \ldots e(M)$, where $e(w) = (x_1(w), x_2(w), \ldots, x_n(w))$
- Codebook: the set of all codewords.
- Also,

$$\frac{1}{n}\sum_{i=1}^{n} \kappa(x_i(\omega)) \leq P$$

for $1 \leq w \leq M$, i.e. each codeword satisfiies the input constraint.

13.17.1 Assumptions

- W is randomly chosen from the set of messages $\mathcal{W}$, so

$$H(W) = \log M$$

- input and output of the channel:

$$\vec{X} = (X_1, X_2, \ldots, X_n); \vec{Y} = (Y_1, Y_2, \cdots Y_n)$$

- $\vec{X} = e(W)$
- Let $\hat{W} = g(\vec{Y})$ be the estimate on the message W by the decoder.

Error Probabilites

For all messages $1 \leq w \leq M$, let

$$\lambda_w = \Pr\{\hat{W} \neq w | W = w\} = \int_{\vec{y}:g(y)\neq w} f_{\vec{Y}|\vec{X}}(\vec{y}|e(w))\, \mathrm{d}\vec{y}$$

be the conditional probability of error given that the message is w.

- The maximal probability of error of an (n, M) code is defined as

$$\lambda_{max} = \max_w \lambda_w$$

- The average probability of error is defined as

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

13.17.2 Statement

Achievability of rate

A rate R is (asymptotically) achievable for a continuous memoryless channel if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that if for any $\epsilon > 0$, there exists for sufficiently large n an (n,M) code such that (coding rate good/big enough)

$$\frac{1}{n} \log M > R - \epsilon$$

and (error small enough)

$$\lambda_{max} < \epsilon.$$

Channel Coding Theorem

A rate R is achievable for a continuous memoryless channel with input constraint (κ, P) if and only if $R \leq C(P)$, the capacity of the channel.

Lemma: Data processing inequality

$$I(W, \hat{W}) \leq I(\vec{X}, \vec{Y})$$

- We first remark that $W \to \vec{X} \to \vec{Y} \to \hat{W}$ forms Markov chain
- $W, \hat{W}$ are discrete
- X is discrete and Y is continuous.
- So we have that $I(W, \hat{X}) \leq I(\vec{X}, \hat{W})$ because everything here is discrete.

Proof.

$$
\begin{aligned}
I(W, \hat{W}) &\leq I(\hat{W}; \vec{X}) \\
&\leq I(\hat{W}; X) + I(X; Y|\hat{W}) \quad \text{by non-negativity of mutual information} \\
&= \mathbb{E} \log \frac{p(\hat{W}, \vec{X})}{p(\hat{W})p(\vec{X})} + \mathbb{E} \log \frac{f(\vec{Y}|\vec{X}, \hat{W})}{f(\vec{Y}|\hat{W})} \\
&= \mathbb{E} \log \frac{p(\hat{W}, \vec{X}) f(\vec{Y}|\vec{X}, \hat{W})}{p(\hat{W})p(\vec{X})f(\vec{Y}|\hat{W})} \\
&= \mathbb{E} \log \frac{f(\vec{Y})p(\vec{X}, \hat{W}|\vec{Y})}{p(\hat{X})f(\vec{Y})p(\hat{W}|\vec{Y})} \\
&= \mathbb{E} \log \frac{p(\vec{X}, \hat{W}|\vec{Y})}{p(\hat{X})p(\hat{W}|\vec{Y})} \\
&= \mathbb{E} \log \frac{p(\vec{X}|\vec{Y})p(\hat{W}|\vec{X}, \vec{Y})}{p(\hat{X})p(\hat{W}|\vec{Y})} \\
&= \mathbb{E} \frac{p(\vec{X}|\vec{Y})}{p(\vec{X})} + \mathbb{E} \log \frac{p(\hat{W}|\vec{X}, \vec{Y})}{p(\hat{W}|\vec{Y})} \\
&= \mathbb{E} \frac{f(\vec{Y}|\vec{X})}{f(\vec{Y})} + 0 \\
&= I(\vec{X}; \vec{Y})
\end{aligned}
$$

□

13.17.3 Proof of the Converse of the Channel Coding Theorem

1. Let R be achievable rate, i.e. for any $\epsilon > 0$, for sufficiently large n, there exists (n, M) code such that the coding rate is good enough

$$\frac{1}{n} \log M > R - \epsilon$$

and error is small

$$\lambda_{max} < \epsilon.$$

2. Consider

$$\begin{aligned}
\log M &= H(M) \quad \text{because of message is chosen uniformly at random.} \\
&= H(W|\hat{W}) + I(W;\hat{W}) \quad \text{by the chain rule} \\
&\leq H(W|\hat{W}) + I(\vec{X};\vec{Y}) \quad \text{by the data processing inequality} \\
&= H(W|\hat{W}) + h(\vec{Y}) - h(\vec{Y}|\vec{X}) \\
&\leq H(W|\hat{W}) + \sum_{i=1}^{n} h(Y_i) - h(\vec{Y}|\vec{X}) \quad \text{by independence bound} \\
&= H(W|\hat{W}) + \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Y_i|X_i) \quad \text{by Chain rule and Markov property} \\
&= H(W|\hat{W}) + \sum_{i=1}^{n} I(X_i;Y_i)
\end{aligned}$$

3. Let V be a mixing random variable distributed uniformly on $\{1,\ldots,n\}$ independent of X_i.

4. Let $X = X_V$ and Y be the output of the channel with X as the input.

5.

$$\begin{aligned}
\mathbb{E}\kappa(X) &= \mathbb{E}\mathbb{E}[\kappa(X)|V] \quad \text{by towering property} \\
&= \sum_{i=1}^{n} \Pr\{V=i\}\mathbb{E}[\kappa(X)|V=i] \\
&= \sum_{i=1}^{n} \Pr\{V=i\}\mathbb{E}[\kappa(X_i)|V=i] \\
&= \sum_{i=1}^{n} \Pr\{V=i\}\mathbb{E}[\kappa(X_i)] \\
&= \sum_{i=1}^{n} \frac{1}{n}\mathbb{E}[\kappa(X_i)] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\kappa(X_i)\right]
\end{aligned}$$

We assume that each codeword satisfies the input constraint,

$$\frac{1}{n}\sum_{i=1}^{n}\kappa(x_i(w)) \leq P \quad \text{for } 1 \leq w \leq M.$$

then the randomly chosen codeword X, the input, satisfies the input constraint with probability 1. Therefore the above expectation

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\kappa(X_i)\right] \leq P$$

6. By the concavity of mutual information with respect to the input distribution,

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) \le C(P)$$

because

$$X \sim \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

7. Then

$$\sum_{i=1}^{n} I(X_i;Y_i) \le nC(P)$$

8. Then

$$n(R-\epsilon) < \log M \le H(W|\hat{W}) + nC(P)$$

9. Apply Fano's inequality on $H(W|\hat{W}) \to 0$ as the upper bound on error probability $\epsilon \to 0$. Then we conclude that

$$R \le C(P).$$

13.17.4 Proof of the Achievability part of the Channel Coding Theorem

- In the definition

$$C(P) = \sup_{F(x):\mathbb{E}_\kappa(X)\le P} I(X;Y)$$

Since X can be a mixed random variable, so it can be difficult to consider sequences typical w.r.t. the distribution of X.

- Now we introduce a new notion of joint typicality: mutual typicality.
- Note that it's difficult to formulate the notion of joint typicality like we did in discrete case because the input distribution $F(x)$ may not have a pdf.

Mutually typical set

The mutually typical set $\Psi^n_{[XY]_\delta}$ with respect to the joint distribution $F(x,y)$ is the set of pairs of sequences $(\vec{x},\vec{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that the empirical mutual information is close to the true mutual information:

$$|\frac{1}{n}\log\frac{f(\vec{y}|\vec{x})}{f(\vec{y})} - I(X;Y)| \le \delta,$$

where

$$f(\vec{y}|vecx) = \prod_{i=1}^{n} f(y_i|x_i) \quad \text{and} \quad f(\vec{y}) = \prod_{i=1}^{n} f(y_i)$$

and δ is an arbitrarily small positive number. A pair of sequences $(\vec{x},\vec{y})$ is called mutually δ-typical if it is in $\Psi^n_{[XY]_\delta}$.

Lemma: sequences are mutually typical in the long run

For any $\delta > 0$, for sufficiently large n,

$$\Pr\{(\vec{X}, \vec{Y}) \in \Psi^n_{[XY]_\delta}\} \leq 1 - \delta$$

Proof. Consider

$$\begin{aligned}\frac{1}{n}\log\frac{f(\vec{Y}|\vec{X})}{f(\vec{Y})} &= \frac{1}{n}\log\prod_{i=1}^{n}\frac{f(Y_i|X_i)}{f(Y_i)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f(Y_i)} \quad \text{which is the sample average}\end{aligned}$$

Then by WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f(Y_i)} \to \mathbb{E}\log\frac{f(Y|X)}{f(Y)} = I(X;Y)$$

in measure. □

Lemma

Let $(\vec{X}', \vec{Y}')$ be n i.i.d. copies of a pair of generic random variables (X', Y') where X' and Y' are independent (this is a strong assumption) and have the same marginal distributions as X and Y, respectively. Then

$$\Pr\{(\vec{X}', \vec{Y}') \in \Psi^n_{[XY]_\delta}\} \leq 2^{-n(I(X;Y)-\delta)}$$

Proof.

1. For any $(\vec{x}, \vec{y}) \in \Psi^n_\delta$, by definition
$$|\frac{1}{n}\log\frac{f(\vec{y}|\vec{x})}{f(\vec{y})} - I(X;Y)| \leq \delta$$
2. Then
$$\frac{1}{n}\log\frac{f(\vec{y}|\vec{x})}{f(\vec{y})} \leq I(X;Y) - \delta,$$
i.e.
$$\frac{f(\vec{y}|\vec{x})}{f(\vec{y})} \leq 2^n(I(X;Y) - \delta),$$
i.e.
$$f(\vec{y}|\vec{x}) \leq f(\vec{y})2^n(I(X;Y) - \delta)$$
3. Then
$$\begin{aligned}1 &\geq \Pr\{(\vec{X}, \vec{Y}) \in \Psi^n_\delta\} \\ &= \int\int_{\Psi^n_\delta} f(\vec{y}|\vec{x})\,\mathrm{d}F(\vec{x})\,\mathrm{d}\vec{y} \\ &\geq 2^{n(I(X;Y)-\delta)}\int\int_{\Psi^n_\delta} f(\vec{y})\,\mathrm{d}F(\vec{x})\,\mathrm{d}\vec{y} \\ &= 2^{n(I(X;Y)-\delta)}\Pr\{(\vec{X}', \vec{Y}') \in \Psi^n_\delta\} \quad \text{because the measure of the integral is a product form}\end{aligned}$$

□

Random Coding Scheme

Parameters

1. Fix $\epsilon > 0$ and input distribution $F(x)$. Let δ be specified later.
2. Since $C(P)$ is left-continuous, there exists $\gamma > 0$ such that
$$C(P - \gamma) > C(P) - \frac{\epsilon}{6}$$
3. By the definition of $C(P - \epsilon)$, there exists an input random variable X such that
$$\mathbb{E}\kappa(X) \leq P - \gamma$$
and
$$I(X;Y) \geq C(P - \gamma) - \frac{\epsilon}{6}$$
4. Choose for a sufficiently large n an even integer M such that
$$I(X;Y) - \frac{\epsilon}{6} < \frac{1}{n} \log M < I(X;Y) - \frac{\epsilon}{8}$$
5.
$$\frac{1}{n} \log M > I(X;Y) - \frac{\epsilon}{6} \leq C(P - \gamma) - \frac{\epsilon}{3} > C(P) - \frac{\epsilon}{2}$$

The Random Coding Scheme

1. Construct the codebook $\mathcal{C}$ of an (n, M) code randomly by generating M codewords in $\mathbb{R}^n$ independently and identically according to $F(x)^n$.
2. Denote the codewords by $\vec{X}(1), \ldots, \vec{X}(M)$.
3. Reveal the codebook $\mathcal{C}$ to both the encoder the decoder
4. A message W is chosen from the message set uniformly.
5. The sequence $\vec{X} = \vec{X}(W)$ is transmitted through the channel.
6. The channel outputs a sequence $\vec{Y}$ according to
$$\Pr\{Y_i \leq y_i, i \leq i \leq n | X(\vec{W}) = \vec{x}\} = \prod_{i=1}^{n} \int_{-\infty}^{y_i} f(y|x_i)\, \mathrm{d}y$$
7. The sequence $\vec{Y}$ is decoded to the message w if
   - $(\vec{X(w)}, \vec{Y}) \in \Psi_\delta^n$ are mutually typical and
   - There does not exist $w' \neq w$ such that $(\vec{X(w')}, \vec{Y}) \in \Psi_\delta^n$
   - Otherwise $\vec{Y}$ is decoded to a garbage constant message.

We now analyze the performance the coding scheme proposed above.

- Let $vecX\tilde{(}w)$ be the encoded message.
- Define the error event $Err = E_e \cup E_d$, where

$$E_e = \{\frac{1}{n}\sum_{i=1}^{n} \kappa(X_i\tilde{(}W)) > P\}$$

  and

$$E_d = \{\tilde{W} \neq W\}.$$

- By the union bound $\Pr\{Err\} = \Pr\{Err|W = 1\} \leq \Pr\{E_e|W = 1\}+\Pr\{E_d|W = 1\}$
- By letting n be large enough, we can upper bound the probability of E_d by $\epsilon/4$ by the typicality, similar to the discrete case.
- By WLLN, for sufficiently large n, $\Pr\{E_e|W = 1\}$ can be upper bounded by $\frac{\epsilon}{4}$

Memoryless Gaussian Channels

14.1 Channel Capacity of Memoryless Gaussian

- Why Gaussian Channel? The Gaussian channel is the most commonly used model for a noisy channel with real input and output, because:
  1. the Gaussian channel is analytically tractable
  2. the Gaussian noise can be regarded as the WORST kind of additive noise subject to a constraint on a noise power.

Gaussian Channel

A Gaussian channel with noise energy N is a continuous channel with the following two equivalent specifications:

1. $$f(y|x) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-x)^2}{2N}},$$
   a Gaussian pdf with mean x and variance N
2. or it can be specified by a random noise variable Z where
   $$Z \sim \mathcal{N}(0, N) \quad \text{and} \quad Y = \alpha(X, Z) = X + Z$$

Memoryless Gaussian Channel

A memoryless Gaussian channel with noise power N and input power constraint P is a memoryless continuous channel with the Gaussian channel with noise energy N as the generic channel. The input power constraint refers to the input constraint (κ, P) where $\kappa(x) = x^2$

The following is the main theorem in this section.

Capacity of a memoryless Gaussian Channel

The capacity of a memoryless Gaussian channel with noise power N and input power constraint P is
$$\frac{1}{2}\log(1 + \frac{P}{N})$$

The capacity is achieved by the input distribution $\mathcal{N}(0, P)$

- Note that the capacity of a memoryless Gaussian channel only depends on the signal-to-noise ratio P/N.
- The capacity is strictly positive no matter how small P/N is.
- The capacity is infinite if there is no input power constraint, meaning we can transmit information at any finite rate if there's no input power constraint.

Lemma

Let $Y = X + Z$. Then $h(Y|X) = h(Z|X)$ provided that $f_{Z|X}(z|x)$ exists.

Proof. 1. Assume $f_{Z|X}$ exists.

2. Consider

$$\begin{aligned} f_{Y|X}(y|x) &= f_{X+Z|X}(y|x) \\ &= f_{x+Z|X}(y|x) \\ &= f_{x+Z-x|X}(y-x|x) \quad \text{shifting the variable and parameter by the same amount.} \\ &= f_{Z|X}(y-x|x) \end{aligned}$$

Thus $f_{Y|X}$ exists.

3. $h(Y|X = x)$ is defined, and

$$\begin{aligned} h(Y|X) &= \int h(Y|X=x)\,\mathrm{d}F_X(x) \\ &= \int h(X+Z|X=x)\,\mathrm{d}F_X(x) \\ &= \int h(x+Z|X=x)\,\mathrm{d}F_X(x) \\ &= h(Z|X=x)\,\mathrm{d}F_X(x) \quad \text{by invariance of differential entropy under translation} \quad = h(\end{aligned}$$

□

Proof of theorem

Proof. 1. Let $F(x)$ be the CDF of the input random variable X such that $\mathbb{E}X^2 \leq P$, where X is not necessarily continuous.

2. Since $Z \sim \mathcal{N}(0, N)$, f_Z exists. Then $f_{Z|X}(z|x)$ exists and is equal to f_Z because Z is independent of X.

3. As we noted before,

$$f_{Y|X}(y|x) = f_{Z|X}(y-x|x) = f_Z(y-x).$$

Also

$$f_Y(y) \int f_{Y|X}(y|x)\,\mathrm{d}F_X(x).$$

Therefore f_Y exists and $h(Y)$ is defined.

4.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z|X) \quad \text{by the above lemma} \\ &= h(Y) - h(Z) \end{aligned}$$

5. Since $Z \perp X$, and Z is zero-mean,

$$\begin{aligned} \mathbb{E}Y^2 &= \mathbb{E}(X+Z)^2 \\ &= \mathbb{E}X^2 + \mathbb{E}Z^2 + 2(\mathbb{E}XZ) \\ &= \mathbb{E}X^2 + \mathbb{E}Z^2 + 2\mathbb{E}X\mathbb{E}Z \\ &= \mathbb{E}X^2 + \mathbb{E}Z^2 \\ &\leq P + N \end{aligned}$$

6. Then

$$h(Y) \leq \frac{1}{2}\log[2\pi 2(P+N)]$$

with equality if $Y \sim \mathcal{N}(0, P+N)$. This is achieved when $X \sim \mathcal{N}(0,P)$ by additive property of Gaussian variables.

7.

$$\begin{aligned} C(P) &= \sup_{F(x):\mathbb{E}X^2<P} h(Y) - h(Z) \\ &= \frac{1}{2}\log[2\pi e(P+N)] - \frac{1}{2}\log(2\pi eN) \\ &= \frac{1}{2}\log(1+\frac{P}{N}) \end{aligned}$$

□

14.2 Parallel Gaussian Channels

Parallel Gaussian Channels

- We have k copies of memoryless Gaussian Channels.
- $Z_i \sim \mathcal{N}(0, N_i)$ and Z_i are independent, $1 \leq i \leq k$.

- The total input power constraint:

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

- $$C(P) = \sup_{F(\vec{x}):\mathbb{E}\sum_i X_i^2 \leq p} I(\vec{X}, \vec{Y})$$

Channel Capacity of Parallel Gaussian Channels

$$C(P) = \max_{P_1,P_2,\ldots,P_k:\sum_i P_i = P} \frac{1}{2}\sum_{i=1}^{k} \log(1 + \frac{P_i}{N_i})$$

where the input random variables $X_i \sim \mathcal{N}(0, P_i)$ and are mutually independent.

Proof. 1. Let $P_i = \mathbb{E}X_i^2$ be the input power of the ith channel. Consider

$$\begin{aligned}
I(\vec{X};\vec{Y}) &= h(\vec{Y}) - h(\vec{Z}) \\
&\leq \sum_{i=1}^{k} h(Y_i) - \sum_{i=1}^{k} h(Z_i) \quad \text{by independence bound} \\
&\leq \frac{1}{2}\sum_{i=1}^{k} \log[2\pi e(\mathbb{E}Y_i^2)] - \frac{1}{2}\sum_{i=1}^{k} \log(2\pi e N_i) \\
&= \frac{1}{2}\sum_i \log \mathbb{E}Y_i^2 - \frac{1}{2}\sum_i \log N_i \\
&= \frac{1}{2}\sum_i \log(\mathbb{E}X_i^2 + \mathbb{E}Z_i^2) - \frac{1}{2}\sum_i \log N_i \\
&= \frac{1}{2}\sum_i \log(P_i + N_i) - \frac{1}{2}\sum_i \log N_i \\
&= \frac{1}{2}\sum_{i=1}^{k} \log\left(1 + \frac{P_i}{N_i}\right)
\end{aligned}$$

2. The inequalities above are tight when X_i are independent and $X_i \sim \mathcal{N}(0, P_i)$

3. Therefore, maximizing $I(\vec{X}, \vec{Y})$ is a matter of maximizing $\frac{1}{2}\sum_i \log(P_i + N_i) - \frac{1}{2}\sum_i \log N_i$, i.e. we maximize

$$\sum_i \log(P_i + N_i)$$

4. The capacity of the system of parallel Gaussian channels is equal to the sum of the capacities of the individual Gaussian channels with the input power optimally allocated.

□

Maximization Problem

We want to maximize $\sum_i \log(P_i + N_i)$ subject to $\sum_i P_i \leq P$ and $P_i \geq 0$.

Proof. 1. Apply the Lagrange multipliers method.

2. Observe that the inequalities constraint becomes an equality constraint because $\log(P_i + N)$ is increasing in P_i.

3. Let
$$L = \sum_{i=1}^{k} \log(P_i + N_i) - \mu \sum_{i=1}^{k} P_i$$

4. Differentiating with repect to P_i, we get
$$\frac{\partial L}{\partial P_i} = \frac{\log e}{P_i + N_i} - \mu$$

5. Setting $\frac{\partial L}{\partial P_i} = 0$, we have
$$P_i = \frac{\log e}{\mu} - N_i$$

6. Let $\nu = \frac{\log 2}{\mu}$, we have
$$P_i = \nu - N_i$$
and ν is chosen such that the power constraint is satisfied:
$$\sum_{i=1}^{k} P_i = \sum_i (\nu - N_i) = P$$

□

This solution has a water-filling interpretation. For each channel, if the power has not reached ν, P_i is the amount we fill in to reach $\frac{\log e}{\mu}$ from N_i.

Remark 14.1. By an application of the KKarush-Kuhn-tucker (KKT) condition, we obtain that in general,
$$C(P) = \frac{1}{2} \sum_{i=1}^{k} \log\left(1 + \frac{P_i^*}{N_i}\right)$$
where $\{P_i^*, 1 \leq i \leq k\}$ is the optimal input power allocation among the channels given by
$$P_i^* = (\nu - N_i)^+, 1 \leq i \leq k$$
where ν satisfies
$$\sum_{i=1}^{k} (\nu - N_i)^+ = P$$

- The capacity of the parallel Gaussian channels system is attained when:
- The inputs are independent of each other.

- The inputs are Gaussian.
- The input powers are allocated according to water-filling.

An optimization problem

Given $\lambda_i \geq 0$, maximize $\sum_{i=1}^{k} \log(\alpha_i + \lambda_i)$ subject to $\sum_i \alpha_i \leq P$ and $\alpha_i \geq 0$. The solution is

$$\alpha_i^* = (\nu - \alpha_i)^+, 1 \leq i \leq k$$

where ν satisfies

$$\sum_{i=1}^{k} (\nu - \lambda_i)^+ = P.$$

Proof. 1. Prove that the proposed solution satisfies the KKT condition. □

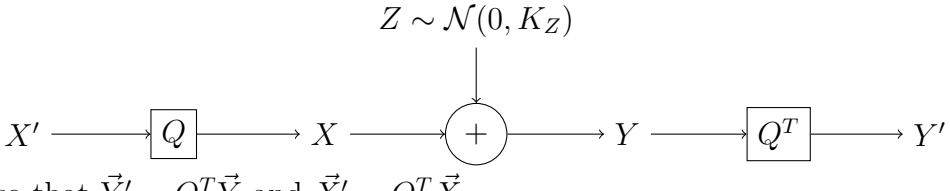
14.3 Correlated Gaussian Channel

- In the setup of parallel Gaussian Channels, we have k copies of independent memoryless Gaussian channels.
  1. The channels are governed by noise vector $\vec{Z} \sim \mathcal{N}(\vec{0}, N)$.
  2. $$N = \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & N_k \end{bmatrix}$$
     by independence.
- We can generalize to correlated Gaussian channels where
  $$\vec{Z} \sim \mathcal{N}(0, K_{\vec{Z}})$$
  where $K_{\vec{Z}}$ is not necessarily a diagonal matrix.

14.3.1 Decorrelation of the Noise Vector

Here is a very useful technique to decorrelate a noise vector:

1. Since the Gaussian correlation matrix is positive semidefinite, we can diagonalize it as $Q\lambda Q^T$ by the spectral theorem.
2. Then apply Q to the input sequence and Q^T to the output sequence:



so that $\vec{Y}' = Q^T\vec{Y}$ and $\vec{X}' = Q^T\vec{X}$.

3. Let $\vec{Z} = Q^T Z$ and so $\vec{Z}'$ is also Gaussian. Then
$$\vec{Y}' = Q^T \vec{Y} = Q^T(\vec{X} + \vec{Z}) = Q^T \vec{X} + Q^T \vec{Z} = \vec{X}' + \vec{Z}'.$$

4. This shows that $\vec{Z}'$ forms a new Gaussian channel where $\vec{X}'$ and $\vec{Y}'$ are the input and output sequences.

5. Since $\vec{Z}' = Q^T \vec{Z}$,
$$K_{\vec{Z}'} = Q^T K_{\vec{Z}} Q = \Lambda.$$
So $\vec{Z}'_i \sim \mathcal{N}(0, \lambda_i)$ and are mutually independent. So we now have a parallel Gaussian channel system.

Decorrelated system has the same capacity

- For the power constraint. since $\vec{X}' = Q^T \vec{X}$ and and Q^T is an orthogonal matrix, then as proved previously
$$\mathbb{E}\sum_i (X'_i)^2 = \mathbb{E}\sum_i X_i^2$$

For the equivalence of capacity, we claim that $I(\vec{X}', \vec{Y}') = I(\vec{X}, \vec{Y})$:

Proof.
$$\begin{aligned}
I(\vec{X}', \vec{Y}) &= h(\vec{Y}') - h(\vec{Y}'|\vec{X}') \\
&= h(\vec{Y}') - h(\vec{Z}'|\vec{X}') \quad \text{by a previous lemma} \\
&= h(\vec{Y}') - h(\vec{Z}') \\
&= h(Q^T \vec{Y}) - h(Q^T \vec{Z}) \\
&= \left[h(\vec{Y}) + \log|\det Q^T|\right] - \left[h(\vec{Z}) + \log|\det Q^T|\right] \\
&= h(\vec{Y}) - h(\vec{Z}) \\
&= I(\vec{X}; \vec{Y}) \quad \text{by repeating what we did in reverse order.}
\end{aligned}$$

□

Transmission in Continuous Time

We have will now study continuous-time transmission, which is true for all real channels at the physical layer.

15.1 The Bandlimited White Gaussian Channel

- We now consider transmission of information in continuous time.
- Consider input variable (process) $X(t)$ and noise process $Z(t)$ added to $X(t)$.
- The sum of $X(t)$ and $Z(t)$ is passed through a filter, with bandwidth W. The output of the filter $Y(t)$ is the output of the channel.
- Here, $Z(t)$ is a zero-mean white Gaussian nosie process with power spectral density $S_Z(f) = \frac{N_0}{2}$ (f is from minus infinity to infinity), called an *additive white Gaussian noise AWGN*

15.2 Signal Analysis Prelim

Fourier Transform

The Fourier transform of a signal $g(t)$ is defined as

$$G(f) = \int_{\infty}^{\infty} g(t) e^{-j2\pi ft} \, \mathrm{d}t$$

- The signal $g(t)$ can be recovered from the Fourier transform $G(f)$ by:

$$g(t) = \int_{\infty}^{\infty} G(f) e^{j2\pi ft} \, \mathrm{d}f$$

  and $g(t)$ is called the inverse Fourier transform of $G(f)$. The functions $g(t)$ and $G(f)$ are said to be a transform pair $g(t) \leftrightarrow G(f)$.

- The variables t and f are considered as time and frequency respectively.

Below is a special signal of interest.

Energy Signal

A signal $g(t)$ is called an energy signal if

$$\int_{-\infty}^{\infty} |g(t)^2| \, \mathrm{d}t < \infty,$$

i.e. the energy of the signal is finite.

- The Fourier transform of an energy signal always exists.

Cross correlation function

Let $g_1(t)$ and $g_2(t)$ be a pair of energy signals. The cross correlation function for $g_1(t)$ and $g_2(t)$ is defined as

$$R_{12}(\tau) = \int_{-\infty}^{\infty} g_1(t) g_2(t-\tau) \, \mathrm{d}t,$$

where τ here is a dummy time variable.

Proposition

For a pair of energy signals $g_1(t)$ and $g_2(t)$,

$$R_{12}(\tau) \rightleftharpoons G_1(f) G_2^*(f)$$

where $G_2^*(f)$ denotes the complex conjugate of $G_2(f)$.

Wide-sense Stationary Process

A process $\{X(t), -\infty < t < \infty\}$ is wide-sense stationary if $\mathbb{E}X(t)$ does not dependent on t and $\mathbb{E}[X(t+\tau)X(t)]$ (autocorrelation) only depends on τ.

Autocorrelation function and spectral density

For a wide-sense stationary process $\{X(t), -\infty < t < \infty\}$, the autocorrelation function is defined as

$$R_X(\tau) = \mathbb{E}[X(t+\tau)X(t)]$$

which does not dependent on t, and the power spectral density is defined as the Fourier transform of the autocorrelation function

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f \tau} \, \mathrm{d}\tau.$$

$$R_X(\tau) \rightleftharpoons S_X(f)$$

Bivariate wide-sense stationary process

Let $\{(X(t), Y(t)), -\infty < t < \infty\}$ be a bivariate wide-sense stationary process. Their cross-correlation functions are defined as

$$R_{XY}(\tau) = \mathbb{E}[X(t+\tau)Y(t)]$$

and

$$R_{YX}(\tau) = \mathbb{E}[Y(t+\tau)X(t)]$$

which do not dependent on t. The corresponding cross-spectral densities are defined as the Fourier transform of the above functions.

An equivalent model for the channel

- Our original model of the channel is that input $X(t)$ and noise $Z(t)$ are summed together and passed through a filter.
- Alternatively, we can pass $X(t)$ through a filter to get $X'(t)$ and pass $Z(t)$ through a filter to get $Z'(t)$. Then we sum up $X'(t)$ and $Z'(t)$ to get $Y(t)$.
- Both X and Z are bandlimited to $[0, W]$. $Z'(t)$ is a bandlimited white Gaussian noise with power spectral density $S'_Z(f)$ equal to $\frac{N_0}{2}$ for $f \in [-W, W]$ and 0 otherwise.

The following is the celebrated sampling theorem.

Nyquist-Shannon Sampling Theorem

Let $g(t)$ be a signal with Fourier transform $G(f)$ that vanishes for $f \in [-W, W]$, i.e. the highest frequency that $g(t)$ can have is W. Then $g(t)$ can be reconstructed from the sample as

$$g(t) = \sum_{i=-\infty}^{\infty} g(\frac{i}{2W}) sinc(2Wt - i)$$

for $-\infty < t < \infty$, where

$$sinc(t) = \frac{\sin(\pi t)}{\pi t}.$$

By continuity, the sinc function is defined to be 1 at $t = 0$.

- The signal $g(t)$ is sampled at a rate equals to $2W$, called the Nyquist rate.
- The sinc function $= 0$ for every integer $t = i \neq 0$.
- $$sinc(2Wt - i) = sinc(2W(t - \frac{i}{2W})) = \begin{cases} 1 & t = \frac{i}{2W} \\ 0 & t = \frac{j}{2w}, j \neq i \end{cases},$$
  i.e. the func function vanishes at every sample point except for $t = \frac{i}{2w}$.
- Let $g_i = \frac{1}{\sqrt{2W}} g(\frac{i}{2W})$ and
  $$\psi_i(t) = \sqrt{2W} sinc(2Wt - i).$$

Then

$$g(t) = \sum_{i=-\infty}^{\infty} g_i \psi_i(t)$$

Remark 15.1. Proposition: The functions

$$\psi_i(t) = \sqrt{2W} sinc(2Wt - i), -\infty < i < \infty$$

form an orthonormal basis for signals which are bandlimited to $[0, W]$.

Proof. 1. Consider

$$\psi_i(t) = \sqrt{2W} sinc(2W(t - \frac{i}{2W}))$$

and $psi_0(t) = \sqrt{2W} sinc(2Wt)$. Therefore,

$$\psi_i(t) = \psi_0(t - \frac{i}{2W}),$$

and so $\psi_i(t)$ and $\psi_0(t)$ have the same energy for all i because they are translations of one another.

2. Consider $sinc(2Wt) \rightleftharpoons \frac{1}{2W} rect(\frac{f}{2W})$ where

$$recf = \begin{cases} 1 & -\frac{1}{2} \leq f \leq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

3. Then by Rayleigh's energy theorem, the energy of the signal $sinc(2Wt)$ in the time domain and the energy of the signal $\frac{1}{2W} rect(\frac{f}{2W})$ are the same, so we have

$$\int_{-\infty}^{\infty} sinc^2(2Wt)\,\mathrm{d}t = (\frac{1}{2W})^2 \int_{-\infty}^{\infty} rect^2(\frac{f}{2W})\,\mathrm{d}f$$

which is equal to

$$(\frac{1}{2W})^2(2W) = \frac{1}{2W}.$$

4. The integral of $sinc^2$ is difficult to evaluate directly but Rayleigh's theorem provides an easy shortcut.

5. Then it follows that

$$\int_{-\infty}^{\infty} \psi_i^2(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \psi_0^2(t)\,\mathrm{d}t = 1$$

6. For $i \neq i'$, we have that both $sinc(2Wt - i)$ and $sinc(2Wt - i')$ have finite energy. consider their cross-correlation function

$$R_{ii'}(\tau) = \int_{-\infty}^{\infty} sinc(2Wt - i) sinc(2W(t - \tau) - i')\,\mathrm{d}t.$$

7. Now $sinc(2Wt - i) \rightleftharpoons e^{-j2\pi f(\frac{i}{2W})}(\frac{1}{2W})rect(\frac{f}{2W}) := G_i(f)$.
8. Then $R_{ii'}(\tau) \rightleftharpoons G_i(f)G_{i'}^*(f)$.
9. $R_{ii'}(0) = 0$ for $i \neq i'$. (computation omitted)
10. Therefore, for $i \neq i'$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \psi_i(t)\psi_{i'}(t)\,\mathrm{d}t \\
&= 2W \int_{-\infty}^{\infty} sinc(2Wt - i)sinc(2Wt - i')\,\mathrm{d}t \\
&= (2W)R_{ii'}(0) \\
&= 0
\end{aligned}$$

11. Therefore, $\psi_i(t), i \in (-\infty, \infty)$ has energy 1 and each pair has cross-correlation function 0 at $\tau = 0$.

□

Our next step objective is to compute the capacity of the bandlimited white Gaussian channel.

15.3 Intuitive Treatment of the Bandlimited Channel

- Assume the input process $X'(t)$ has a Fourier transform, so that

$$X'(t) = \sum_{i=-\infty}^{\infty} X_i' \psi_i(t)$$

- There is a one-to-one correspondence between the continuous-time process $\{X'(t)\}$ and the discrete-time process $\{\mathcal{X}_i'\}$.
- Similarly, assume the output process $Y(t)$ can be written as

$$Y(t) = \sum_{i=-\infty}^{\infty} Y_i \psi_i(t).$$

- With these assumptions, the waveform channel can be regarded as a discrete-time channel defined at $t = \frac{i}{2W}$, with the ith input and output of the channel being X_i' and Y_i' respectively.
- We want to
  1. understand the effect of the noise process $Z'(t)$ on $Y(t)$ at the sampling points.
  2. Relate the power constraint on $\{X_i'\}$ to the power constraint on $X'(t)$.

The noise process at the sampling points is Gaussian

$Z'(\frac{i}{2W})$, $-\infty < i < \infty$ are i.i.d. Gaussian random variables with zero mean and variance N_0W.

Proof.

1. $Z'(t)$ is a filtered version of $Z(t)$, so $Z'(t)$ is also a zero-mean Gaussian process.
2. $Z'(\frac{i}{2W}), -\infty < i < \infty$ are zero-mean Gaussian random variables.
3. The power spectral density of $Z'(t)$ is a rectrangular function:

$$S_{Z'}(f) = \begin{cases} N_0/2 & -W \le f \le W \\ 0 & \end{cases} = \frac{N_0}{2} rect(\frac{f}{2W})$$

4. $R_{Z'}(\tau)$, the autocorrelation function of $Z'(t)$, is a sinc function:

$$S_{Z'}(f) \rightleftharpoons R_{Z'} = N_0 W sinc(2W\tau)$$

5. $R_{Z'}(\tau)$ vanishes at $\tau = \frac{i}{2W}$ for every $i \neq 0$:

$$R_{Z'}(\frac{i}{2W}) = \begin{cases} N_0W & i = 0 \\ 0 & i \neq 0. \end{cases}$$

6. Then for all t and all $i \neq 0$, $Z'(t + \frac{i}{2W})$ are uncorrelated because

$$\mathbb{E}\left[Z'(t + \frac{i}{2W})Z'(t)\right] = R_{Z'}(\frac{i}{2W}) = 0.$$

7. In particular, letting $t = \frac{i}{2W}$, we see that $Z'(\frac{i}{2W})$ and $Z'(\frac{j}{2W} + \frac{i}{2W}) = Z'(\frac{j+i}{2W})$ are uncorrelated, meaning the values of $Z'(t)$ are uncorrelated at any 2 sample points and hence independent.
8. Since $Z'(\frac{i}{2W})$ has zero mean, its variance is given by $R_{Z'}(0) = N_0W$, because

$$R_{Z'}(0) = \mathbb{E}\left[Z'(\frac{i}{2W} + 0)Z'(\frac{i}{2W})\right] = var(Z'(\frac{i}{2W}))$$

□

- Recall that $Y(t) = \sum_i Y_i\psi_i(t)$ and $X'(t) = \sum_i X'_i\psi_i(t)$.
- Let $Z'(t) = \sum_i Z'_i\psi_i(t)$, where $Z'_i = \frac{1}{\sqrt{2W}} Z'(\frac{i}{2W})$.
- Then $Y(t) = X'(t) + Z'(t)$ implies that

$$Y_i = X'_i + Z'_i$$

because $\psi_i(t)$ are orthonormal for $-\infty < i < \infty$.

- Since $Z'(\frac{i}{2W})$ are i.i.d. $\sim \mathcal{N}(0, N_0W)$, $Z_i' \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{N_0}{2})$.
- So the bandlimited white Gussian channel is essentially equivalent to a memoryless Gaussian channel with noise power $\frac{N_0}{2}$.

15.3.1 Power Constraints

- Denote the average energy of X_i' by P' (i.e. the second moment).
- Since $\psi_i(t), -\infty < i < \infty$ are orthonomral, each has unit energy and their energy adds up:

$$\begin{aligned}\int (\psi_i(t) + \psi_j(t))^2 \, \mathrm{d}t &= \int \psi_i^2(t) + \psi_j^2(t) + 2\psi_i(t)\psi_j(t) \, \mathrm{d}t \\ &= \int \psi_i^2(t) \, \mathrm{d}t + \int \psi_j^2(t) \, \mathrm{d}t\end{aligned}$$

- Therefore, $X'(t)$ accumulates energy from the samples at a rate equal to $(2W)P'$.
- Consider

$$(2W)P' \leq P,$$

  where P is the average power constraint on the input process $X'(t)$. We then conclude that $P' \leq \frac{P}{2W}$.
- Fianlly, the capacity of the discrete-time channel is

$$\frac{1}{2} \log(1 + \frac{P}{N_0W}) \quad \text{bits per sample}$$

- Since there are $2W$ samples per unit time, the capacity is

$$W \log(1 + \frac{P}{N_0W}) \quad \text{bits per unit time.}$$

15.4 The Bandlimited Colored Gaussian Channel

To be continued later.